Predictive Analysis for COVID-19 Cases in India Based on LSTM Algorithms

Zixuan Zhou*
*Corresponding author: cathyrc.zhou@mail.utoronto.ca

Finance and Economics, Bcom, Faculty of Arts and Science University of Toronto Toronto, Canada

Abstract—COVID-19 continues to hurt the global economy and living standards. As the number of confirmed cases rises, it is crucial to predict COVID-19 transmission and treatment effectively. This paper aims to compare the results under different machine learning methods and improve the efficiency of predictive analysis. Using valid data provided by Indian government, machine learning regression and the random forest model are implemented to predict the correlation between confirmed cases, deaths, and cured cases. The training and testing dataset scale is 80% to 20% for both models. Comparing the numerical and graphical results of the linear regression and random forest models, while both models generate highly accurate predictions compared with the actual data, the later one contains smaller errors. Another aspect of this paper uses the Long Short-Term Memory (LSTM) model, a popular neural network in artificial intelligence, to predict the future trend of COVID-19 cases, and 70% of the dataset is used for training. As a result, LSTM could make relatively accurate predictions on the overall trend of the confirmed cases, while there appear to be some statistically significant errors compared with actual data.

Keywords-Machine learning; Linear regression; Random forest model; Long short term memory; Coronavirus-related predictions

1 INTRODUCTION

The coronavirus pandemic, also known as the COVID-19 pandemic, has been the most serious social concern for humankind in recent years. The virus was first identified in China in December 2019, then spread globally, and soon became a "Public Health Emergency of International Concern" (PHEIC) declared by WHO in January 2020 [1]. As of May 2022, the pandemic has caused over 500 million cases and 6 million deaths globally, which is one of the viruses that causes the most deaths in human history [1]. Under the massive impact of the pandemic on the national economy and stability, countries worldwide tried their best to develop vaccines. Finally, COVID-19 vaccines have been recognized and spread globally since December 2020 [1]. Although vaccines like Pfizer/BioNTech, Modera, and Johnson & Johnson are widely taken, their effectiveness against coronavirus is always questionable. For example, a study in England has found Pfizer/BioNTech to be around 88% effective against the Delta variant [2], which is significantly different from real-world data from Israel, suggesting only an effectiveness of 41%. To efficiently implement policies under the pandemic to minimize loss in different areas, countries need to figure out the relationship between vaccination, cumulative cases, and morality, based on the data collected by regulated apartments. In this paper, the dataset of COVID-19 in

FFIT 2022, October 28-30, Shenzhen, People's Republic of China

DOI 10.4108/eai.28-10-2022.2328465

India will be focused, and the correlations and trends of different variables within the dataset will be explored and discussed using the machine learning method.

To improve data processing and selection efficiency, machine learning could be one of the best choices since they have achieved satisfactory performance in many tasks [3-5]. Machine learning is one form of artificial intelligence which allows the software to be faster and more accurate at predictions, such as detecting patterns in data sets. Using machine learning, larger data sets could be analyzed and predicted with even more variables. Thereinto, machine learning regression is one of the most popular methods to forecast trends and make predictive analysis. It has been not only a calculation-based function during the COVID-19 pandemic, but also a risk predictor in healthcare in the future pandemic environment [6]. Many papers have developed models to analyze various areas regarding COVID-19. For example, one research article by Zhao et al. uses the Poisson distribution to model observed incidence cases and a gamma distribution for the series interval and then to estimate future cases based on posterior distributions [7]. The result produces a reasonably accurate result as they apply the method to a single state in the U.S, while their method could be improved in terms of accuracy and effectiveness. For instance, when they estimate their future incidence cases, they assume that the current transmission rate will stay the same [7]; however, the virus transmission rate will change due to change in local policies. There is also a study that compares different mathematical models that predict COVID-19 cases, moralities, and recoveries [8], but the result indicates that the prediction efficiency is not directly related to the best fit [8]. The efficiency and accuracy might be improved by using methods related to artificial intelligence, which can deal with multidimensional data in a faster and more sustainable way.

This paper will focus on COVID-19 data in India, as India contains the world's largest population and has suffered massive losses from the pandemic. The dataset contains two main parts: state-wise data in India that updates the total cases and deaths, and the other is testing and vaccination data. The state-level data comes from the Ministry of Health and Family Welfare (MoHFW), an Indian government ministry charged with health policy. The testing and vaccination data is collected from "covid19india", an official website that updates covid-related data in all states in India, including individual-level details, testing, and vaccination details. This paper uses linear regression and random forest models to make predictions on the correlation between the number of confirmed cases, the number of deaths, and the number of cured cases. The predicted result generally shows an accurate trend, which means both models are relatively efficient for predictive analysis of the selected dataset. This paper also uses Long Short-term Memory (LSTM) to predict the future trend of confirmed COVID-19 cases. The results also show an accurate general trend, while some errors appear.

2 METHODS

2.1 Dataset Preparation

The dataset contains three data files. The first one reports COVID-19 cases on a daily basis in each state in India, including cases confirmed within and outside the nation, daily cursed, death, and confirmed cases. The second file reports the vaccination condition in different states in India,

including details in terms of gender, the number of doses taken, and the type of vaccines taken. Finally, the third file provides national and state-level data on total samples tested and confirmed cumulative positive cases. The sample data can be found in Table 1.

Table 1 The sample data of the collected Indian dataset.

No	Date	Time	State/Union Territory	Confirmed Indian National	Confirmed Foreign National	Cured	Death	Confirmed
1	2020-01-30	6PM	Kerala	1	0	0	0	1
2	2020-01-31	6PM	Kerala	1	0	0	0	1
3	2020-02-01	6PM	Kerala	2	0	0	0	2
4	2020-02-02	6PM	Kerala	3	0	0	0	3
5	2020-02-03	6PM	Kerala	3	0	0	0	3
6	2020-02-04	6PM	Kerala	3	0	0	0	3
7	2020-02-05	6PM	Kerala	3	0	0	0	3
8	2020-02-06	6PM	Kerala	3	0	0	0	3

2.2 Machine learning algorisms for correlation analysis

In this paper, two popular methods under machine learning, namely the linear regression and the random forest model, are used to make a critical and predictive analysis of the correlation between the number of deaths, confirmed cases, and cured cases.

1)Linear regression: Linear regression is a predictive modeling method in machine learning to analyze relationships between different independent variables and a dependent. The target outcome could be predicted using unseen testing data after the relationship between independent and dependent variables is estimated based on the training data. First, the best fit line is generated in machine learning regression, where the distance between the line and the data points is minimized. Then the gap between the predictions and the historical data could be seen to evaluate the accuracy of the regression.

2)Random forest: Alternatively, another method, the random forest model, is also implemented to make predictions in this paper. In machine learning, the random forest model acts like a classifier where various decision trees on the subset of the selected dataset are averaged to improve the accuracy of the predicted outcome [9]. The training data is put into each decision tree, and then the majority votes by the trees would be used to predict the outcome based on testing data. The more decision trees, the higher accuracy the prediction would generate. Random forest is commonly used to make predictive analysis for disease trends [9].

2.3 Implementation details for linear regression and random forest

The reason these two methods are chosen is not only for their capabilities in dealing with large datasets with high dimensionalities, but also that they could offset each other's disadvantages in some contexts. One big concern of machine learning regression is that the training data must represent the overall data; otherwise, data overfitting would be a risk. Random forest could enhance the efficiency and accuracy of predictions and prevent overfitting. While the random forest model might not be more suitable for regression tasks than classification, the linear regression model could be much better for regression-related predictions. In this paper, both methods will be implemented in the chosen dataset, and the results will be compared to evaluate their efficiency and accuracy.

First, machine learning regression is implemented to predict the number of cured cases, given the number of death and confirmed cases, using daily data in each state in India. Before setting up the regression model, the data is read using a package called pandas, and missing data is filtered out. In order to predict the cured cases given the daily death and confirmed cases, the columns "Deaths" and "Confirmed" are set as the x variable, and the data column "Cured" is set as y. After that, "Sklearn" is used to split data into testing set and training set, where this study set the testing size to be 20% of the whole data. Next, the Linear Regression model could be launched under "Sklearn", and the predicted value could be carried out using the testing data after the model is fitted by the training data. It can be evaluated the accuracy of the result in various ways. The error is generated by taking the difference between the test y values and the predicted y values. This paper used error to calculate the Root Mean Square Error (RMSE) and Mean Absolute Error (MAE).

By "Sklearn", this study then implements the random forest regressor. Then, using the same method to filter and split data, the model can be carried out the same way as the linear regression model.

2.4 Deep learning algorithm for predicting the trend of the confirmed cases

1)Introduction of LSTM: LSTM, also known as the long short term memory, has been recognized as the most popular neural network for artificial intelligence and deep learning in the 21st century. It could be well-suited for single data points like images and a sequence of data, such as videos. As a recurrent neural network, LSTM is very well implemented for classification, processing, and predictive analysis of various AI areas such as robot control, time series predictions, and market predictions [10]. Using LSTM, this paper could make predictions on the number of confirmed COVID cases.

2)Model Structure: Setting up the model, the study imports the layer called LSTM and Dense from Tensorflow to build the neural network. As previously done for regression and random forest models, the missing data would be filtered out first, then split into training and testing datasets. This paper uses the dataset in the column named "confirmed" to predict the trend of daily confirmed cases in different states in India. There are 18, 109 rows of valid data, and this study sets 70% of them to be the training dataset, which is 12, 677 rows. The rest, 30% of the data, would be used to test the dataset for prediction to test the model's generalization ability. In order to make it more convenient for the model built by Tensorflow to process, both the training

and testing datasets will be expanded with one dimension. Moving forward, in order to speed up the training process, both the testing and training data are re-scaled from the range of 0 to 1.

Building up the model, the study uses two layers of LSTM, and both contain 100 neurons. The details of the model can be found in Table 2. Before training the dataset, the study compiles "adam" as the optimizer and the Mean Square Error (MSE) as the loss. The number of the epoch is set to be 50.

Table 2. The arcihtecture of the employed LSTM model.

Layer (type)	**Output Shape**	**Param Number #**
1stm (LSTM)	(None, 4,100)	40800
1stm1 (LSTM)	(None, 50)	30200
Dense (Dense)	(None, 1)	51
Total params: 71051		
Trainable params: 71051		
Non-trainable params: 0		

3 RESULT AND DISCUSSION

3.1 Results of correlation analysis between deaths, confirmed and cured cases

The result of the linear regression model and the random forest model is shown in Table 3. Both the linear regression model and the random forest model split 80% of the data to be the training set and the rest 20% to be the testing set. Root mean square error and mean absolute error are used to evaluate the accuracy and efficiency of the two models. Table 3 shows that the random forest model has produced significantly more minor errors than the linear regression model. The RMSE and MAE of linear regression are around four times and two times larger than that of random forest, respectively. The accuracy of the random forest model is much higher in this study, and this might happen because of the averaging function of the decision trees, which prevents the overfitting risk of linear regression to some extent.

Table 3. The performance based on linear regression and random forest model

Model	**Performance**	
	RMSE (Root Mean Square Error)	***MAE (Mean Absolute Error)***
Linear Regression (Training data 80%, Testing data 20%)	43671.5	18139.3
Rabdom Forest Model (Training data 80%, Testing data 20%)	15247.2	5010.4

Additionally, Figure 1 and Figure 2 show the linear regression and random forest models' illustrated results, respectively. Again, the blue graph represents the actual value, and the red shows the predicted value.

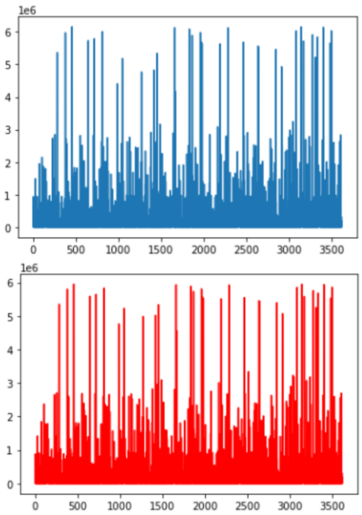

Figure 1. Real Value vs. Predicted Value under Linear Regression Model.

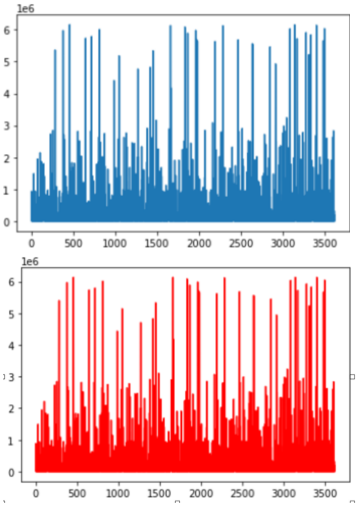

Figure 2. Real Value vs. Predicted Value under Random Forest Model.

Comparing the graphs of actual and predicted values in Figure 1 and Figure 2, it could be seen that both the linear regression model and the random forest model could produce an overall accurate correlation between the number of deaths, confirmed cases and cured cases, while the random forest model contains less obvious error, as it shows from the graph. Furthermore, the overall correlation is predicted to be relatively accurate by both models, indicating an existing correlation between the x and y variables selected by this paper. Intuitively, the confirmed cases

and new deaths from COVID-19 could further help with the development of vaccinations, COVID-adjusted drugs, and other protective methods against coronavirus.

3.2 LSTM-predictive analysis on the trend of coonfirmed cases

Besides predicting the correlation between confirmed cases, deaths, and cured cases, it is also worth studying the trend of upcoming confirmed cases for advanced preparation of future COVID-19 policies and medical facilities. Figure 3 below shows the graphical result of LSTM predictions for the trend of future cases in India. It could be seen from the graph that there are some errors in the prediction that are not reasonable for real-life conditions, where the predicted number of future cases appears to be negative. Also, the model contains a clear fall of confirmed cases, while the actual data shows an overall increasing trend with no obvious declines. Therefore, predicting the confirmed cases based on the machine learning algorithm seems to be a possible route, even though the prediction result should be further improved.

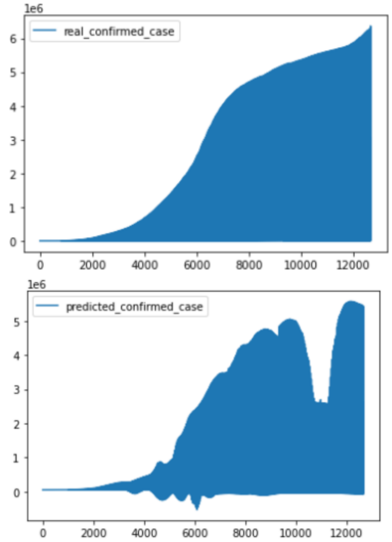

Figure 3. Real Trend vs. Predicted Trend of confirmed cases under LSTM.

4 CONCLUSION

Rather than use mathematical algorisms or single-dimensional models, three machine learning methods are implemented in this paper to improve the accuracy and efficiency of coronavirus-related predictions. Machine learning regression and random forest model are used to predict the correlation between COVID-19 cases, death, and curing, and the long short term memory (LSTM) model is used to predict the trend of confirmed cases. As a result, the random forest model shows a higher accuracy and efficiency than the regression model in correlation predictions. LSTM produces a generally accurate trend for confirmed cases, while coming errors in predictions occur compared to historical data. In the future, more machine learning models should be implemented and tested to improve the efficiency and accuracy of predictive analysis

on COVID-19 issues. Hopefully, artificial intelligence could be more widely used in health and medical areas.

REFERENCES

[1] Wikipedia contributors. "COVID-19 pandemic." Wikipedia, 2022. https://en.wikipedia.org/wiki/COVID-19_pandemic

[2] C. Baraniuk, "Covid-19: How effective are vaccines against the delta variant? " BMJ, n1960. https://doi.org/10.1136/bmj.n1960, 2021.

[3] C. Schuldt, L. Ivan, and C. Barbara. "Recognizing human actions: a local SVM approach." Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004. Vol. 3. IEEE, 2004.

[4] Y. Qiu, et al. "Improved denoising autoencoder for maritime image denoising and semantic segmentation of USV." China Communications, vol. 17, 2020, pp. 46-57.

[5] Y. Qi. "Random forest for bioinformatics." Ensemble machine learning. Springer, Boston, MA, 2012, pp. 307-323.

[6] S. Kushwaha, S. Bahl, A. Bagha, K. Parmar, M. Javaid, A. Haleem, & R. Singh. "Significant Applications of Machine Learning for COVID-19 Pandemic." Journal of Industrial Integration and Management, vol. 5, pp. 453–479. https://doi.org/10.1142/s2424862220500268, 2020.

[7] H. Zhao, N. Merchant, A. McNulty, T. Radcliff, M. Cote, R, S. B., Sang, H., & M. Ory, "COVID-19: Short term prediction model using daily incidence data." PLOS ONE, vol. 16, e0250110. https://doi.org/10.1371/journal.pone.0250110, 2021.

[8] M. Lounis, R. Torrealba, O., & R. Conde-Gutiérrez, "Predictive models for COVID-19 cases, deaths and recoveries in Algeria." Results in Physics, 30, 104845. https://doi.org/10.1016/j.rinp.2021.104845, 2021.

[9] Javatpoint. (n.d.). "Machine Learning Random Forest Algorithm" Www.Javatpoint.Com. https://www.javatpoint.com/machine-learning-random-forest-algorithm

[10] Wikipedia contributors, "*Long short-term memory*." Wikipedia, 2022 https://en.wikipedia.org/wiki/Long_short-term_memory#Applications.