SPREADING OF COVID-19 IN INDIA, ITALY, JAPAN, SPAIN, UK, US: A Prediction Using ARIMA and LSTM Model

MUNISH KUMAR, Department of Computational Sciences, Maharaja Ranjit Singh Punjab Technical University, Bathinda, Punjab, India

SURBHI GUPTA, Department of Computer Science & Engineering, Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad, India

KRISHAN KUMAR, Department of Information Technology, University Institute of Engineering and Technology, Panjab University, Chandigarh, India

MONIKA SACHDEVA, Department of Computer Science and Engineering, I. K. G. Punjab Technical University, Kapurthala, Punjab, India

The spread of COVID-19 around the entire world has placed humankind in an unprecedented situation. Because of the rise of the number of cases and its subsequent load on the organizations and wellbeing experts, some immediate strategies are required to envision the number of cases in the future. In this article, the authors have presented two data-driven estimation techniques, namely, Auto-Regressive Integrated Moving Average (ARIMA) and Long Short-Term Memory (LSTM) for prediction of the cumulative number of COVID-19 cases and the cumulative number of deaths due to COVID-19. Various measures of goodness of fit such as Akaike Information Criteria (AIC), Mean Absolute Percentage Error (MAPE), and Root Mean Square Error (RMSE) are computed for the ARIMA model. For LSTM, two parameters such as Mean Absolute Percentage Error (MAPE) and Root Mean Squared Error (RMSE) are computed. 80% of the available data are used to create the model and the remaining 20% are used for testing. The predicted and actual figures are compared to test the prediction accuracy. This article aims to aid India, Italy, Japan, Spain, the UK, and the US administration by providing a statistical tool to predict the future figures of upcoming suspected and death cases because of COVID-19.

CCS Concepts: • **Computing methodologies** → *Modeling and simulation; Model development and analysis; Model verification and validation;*

Additional Key Words and Phrases: COVID-19, regression model, ARIMA predictions, LSTM

ACM Reference format:

Munish Kumar, Surbhi Gupta, Krishan Kumar, and Monika Sachdeva. 2020. SPREADING OF COVID-19 IN INDIA, ITALY, JAPAN, SPAIN, UK, US: A Prediction Using ARIMA and LSTM Model. *Digit. Gov.: Res. Pract.* 1, 4, Article 24 (July 2020), 9 pages.

<https://doi.org/10.1145/3411760>

Authors' addresses: M. Kumar (corresponding author), Department of Computational Sciences, Maharaja Ranjit Singh Punjab Technical University, Bathinda, Punjab, India; email: munishcse@gmail.com; S. Gupta, Department of Computer Science & Engineering, Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad, India; email: royal_surbhi@yahoo.com; K. Kumar, Department of Information Technology, University Institute of Engineering and Technology, Panjab University, Chandigarh, India; email: k.salujaiuet@gmail.com; M. Sachdeva, Department of Computer Science and Engineering, I. K. G. Punjab Technical University, Kapurthala, Punjab, India; email: monasach1975@gmail.com.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

© 2020 Copyright held by the owner/author(s).

2639-0175/2020/07-ART24

<https://doi.org/10.1145/3411760>

1 INTRODUCTION

The coronavirus disease 2019 (COVID-19) spread is persistent and relentless. It has already been declared a pandemic, as its impact is visible around the globe. It has already infected more than 3,083,453 people with 212,498 deaths in 210 countries as of April 28, 2020 (Worldometers.info). A coordinated global response is the need of the hour. The administrations and health systems need to meet an exceptional challenge. It all commenced on December 30, 2019, when several patients with symptoms of pneumonia were observed in Wuhan, China. On January 7, 2020, these patients were isolated with new symptoms called coronavirus. It was named as novel coronavirus 2019 (2019-nCoV) but the official name of COVID-19 was given by the WHO on February 11, 2020. Coronaviruses are surrounded, connected, single-stranded, positive-sense RNA viruses. They have a crown-like characteristic appearance [Guarner et al. 2020], as seen in Figure 1.

There are still many uncertainties regarding COVID-19, as the spread and impact are different in different countries [Lenzer, 2020]. The most affected countries in the world today are the USA, Spain, Italy, France, Germany, and the UK. Surprisingly, having the best in medical care, the USA suffered 56,803 deaths until now. The most impacted country until mid-March was Italy. Italy has 199,414 cases, 26,977 deaths, with 66,624 recoveries to date (worldometers.info). Remuzzi and Remuzzi [2020] observed that the most difficult task is to predict the number of infected patients and those in critical condition. Although Italy seems to be most impacted once, the spread hit the US badly. Recently, the outbreak led to maximum impact and casualties in the US. The matter of concern is the availability of medical care [Lenzer, 2020]. Figure 2 depicts the growth of cases in some major countries and over the globe.

Over the past two years, many authors used traditional ARIMA model and its ensemble for prediction of diseases such as human brucellosis, pulmonary tuberculosis, influenza, infectious diarrhoea, and dengue fever [Chen et al. 2020; Fang et al. 2020; Cao et al. 2020]. But NN-based LSTM model has not been explored much for disease prediction. So, in this article, three data-driven models based on LSTM and ARIMA are proposed for the prediction of the number of patients and casualties to be expected in the upcoming days for six different countries. The proposed model can be very helpful in the efficient management and administration of COVID-19. This article has been organized as follows: The methodology used for prediction models, i.e., ARIMA and LSTM models, is discussed in Section 2. The experimental results and conclusions obtained are subsequently presented in Sections 3 and 4, respectively.

2 METHODOLOGY

2.1 ARIMA (Auto-Regressive Integrated Moving Average)

ARIMA is a common and extensively used statistical method for time series analysis and forecasting. ARIMA means the Auto-Regressive Integrated Moving Average, model. It uses the temporal structures from time series and is applied in some cases where data are of non-stationarity nature. The representative equation for ARIMA is obtained after combining AR+MA equations. It is Equation (1):

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \dots + \dots + \beta_p Y_{t-p} + \phi_1 \varepsilon_{t-1} + \phi_2 \varepsilon_{t-2} + \dots + \phi_q \varepsilon_{t-q}, \quad (1)$$

where I_t is a function of the “lags of I_t ” where

$$\text{predicted } y_{et} = \text{Constant } (\alpha) + \beta \text{ Lags of } Y \text{ up to } p \text{ lags} + \text{Lagged forecast errors } (\varepsilon) \text{ up to } q \text{ lags}.$$

The standard notation used for this model is ARIMA (p, d, q) where p is the lag order, d is the degree of differencing, and q is the order of moving average terms [Zhang, 2003].

2.2 LSTM (Long Short-Term Memory)

LSTM is a recurrent neural network that provided a good answer to the “short-term memory” problem. These networks constitute intelligent gates that can differentiate important information from unnecessary. The LSTM

Fig. 1. Electron microscopy photograph of the negative stain of a coronavirus (image from the Public Health Image Library; <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7109697/figure/11>).

Fig. 2 Growth of cases in some major countries (Source: ourworldindata.org).

network was first presented by Hochreiter and Schmidhuber in 1997. Many LSTM-based forecasting applications such as short-term traffic forecasting, residential load forecasting, and solar irradiance prediction using weather forecasts have been presented by Zhao et al. 2017; Kong et al. 2017; and Qing and Niu et al. 2018, respectively.

A simple LSTM network illustrated in Figure 3 consists of four different gates, i.e., input gate, forget gate, control gate, and output gate.

3 EXPERIMENTAL RESULTS AND DISCUSSION

In the current research, we have used the data of confirmed and death cases from source ourworldindata.org from 22 Feb., 2020, to 23 April, 2020. Other parameters such as social distancing, masks, and isolation are not considered, as it is difficult to quantify them. Moreover, these factors remained unaffected for the period we have considered in our study. All these practices have been started across the globe by the end of February 2020.

Fig. 3. LSTM structure (<https://adventuresinmachinelearning.com/keras-lstm-tutorial/>).

Three different models based on these data are proposed and fitted to predict future cases. 80%–20% of data partition is used for training and testing purposes. RMSE, MAPE, and AIC are used for the evaluation of the proposed models. RMSE measures the rooted mean of the squared difference between the predicted and the original value. The Mean Absolute Percentage Error (MAPE) is another measure for prediction accuracy and it is expressed in percentage. A value lower than 10 is said to be a fit ARIMA model. Another parameter used is the Akaike Information Criteria (AIC), which is an extensively used measure for evaluating an ARIMA model. It quantifies the goodness of fit for the model as well as the simplicity of the model. It should be as low as possible. The lower the AIC, the better it is [Paul et al. 2013].

The evaluation of the ARIMA model is done by preparing a model on a training dataset and by making predictions on a test dataset. The main steps followed for the ARIMA model fitting are as follows:

1. First, the dataset is split into 80%, 20% training, and test sets, respectively.
2. An ARIMA model is trained by considering the time steps dataset.
3. First, one-step prediction is done, and the error score is calculated for predictions.
4. For the next iteration, each time step of the test set is iterated and then the obtained model is used to make estimations on fresh data.
5. A new ARIMA model is trained for each time step using the iterative approach.
6. Parameters are tuned using Grid search. Performance is evaluated for each parameter, and the configurations for the lowest error are determined.
7. To check the goodness of fit of the ARIMA model, we have considered AIC, MAPE, and RMSE values for optimized p , d , and q values.

In the current research, we have used two variations of the ARIMA model, i.e., ARIMA with AIC optimized and simple forecasting and ARIMA with AIC optimized and model retraining.

Table 1 shows the tuned values for ARIMA parameters for the prediction of confirmed and death cases. Parameters were tuned for best predictions for every country and case.

ANALYSIS OF PREDICTION MODELS FOR GOODNESS OF FIT

Table 2 and Table 3 present the country-wise fitted model evaluation using simple ARIMA, ARIMA with retraining, and LSTM models for the prediction of confirmed cases and deaths, respectively. One can observe that the

Table 1. ARIMA Parameters for the Prediction of Confirmed and Death Cases

	Country	Simple ARIMA Model			ARIMA Retraining Model		
		p	d	q	p	d	q
Prediction of confirmed cases	India	0	2	2	0	2	2
	Italy	0	2	0	0	2	0
	Japan	4	2	1	4	2	0
	Spain	0	2	0	0	2	0
	UK	2	2	0	2	2	0
	US	2	2	3	4	2	1
Prediction of deaths	India	2	2	2	2	2	2
	Italy	2	2	2	2	2	2
	Japan	4	2	3	4	2	2
	Spain	0	2	0	0	2	0
	UK	4	2	0	4	2	0
	US	4	2	0	4	2	1

Table 2. Prediction Modelling for Confirmed Cases

	Country	MAPE	RMSE	AIC
Simple ARIMA model	India	1.36	289.78	432.79
	Italy	4.2	9,708.12	748.72
	Japan	21.75	2,133.13	518.13
	Spain	7.39	16,692.07	789.73
	UK	3.04	3,266.49	739.56
	US	4.53	38,868.55	871.68
ARIMA with Retraining	India	1.01	210.67	556.3
	Italy	0.33	623.59	937.25
	Japan	2.4	1,496.72	758.24
	Spain	0.35	892.68	987.61
	UK	1.2	1,446.62	971
	US	1.8	38,868.55	1,082.26
LSTM	India	8.72	309.54	-
	Italy	8.43	3,867.68	-
	Japan	4.39	250.61	-
	Spain	9.53	5,174.86	-
	UK	15.92	2,610.22	-
	US	16.67	16,408.89	-

MAPE rate is low as desired in most of the cases. The only outlier is Japan for the ARIMA model for cases and Spain and the US for LSTM for death predictions.

Figure 4(a) and Figure 4(b) show the valuation of MAPE of proposed models to predict confirmed cases and deaths for the six countries, respectively. Figure 5(a) and Figure 5(b) show the valuation of RMSE of proposed models to predict confirmed cases and deaths for the six countries, respectively. ARIMA_retraining is having

Table 3. Prediction Modelling for Death Cases

	Country	MAPI	RMSE	AIC
Simple ARIMA model	India	5.7	33.87	220.56
	Italy	3.62	1,072.29	560.53
	Japan	22.49	55.75	181.56
	Spain	7.82	1,943.67	545.67
	UK	10.65	2,106.94	581.38
	US	9.93	4,551.08	587.9
ARIMA with Retraining	India	1.75	11.2	289.54
	Italy	0.31	81.07	698.71
	Japan	4.17	23.84	270.26
	Spain	0.48	112.89	700.48
	UK	1.08	178.76	748.3
	US	1.18	386.14	1,352.75
LSTM	India	6.11	4.08	-
	Italy	8.34	495.27	-
	Japan	7.78	4.29	-
	Spain	23.68	233.45	-
	UK	15.78	440.93	-
	US	22.87	1,059.74	-

Fig. 4. Valuation of MAPE for proposed models to predict (a) confirmed cases (b) death cases.

the lowest MAPE as well as RMSE as compared to the other two models. Figure 6 illustrates that the US shows higher AIC value as compared to other counties for the ARIMA_retraining model for the death cases.

ANALYSIS OF PREDICTION MODELS *w.r.t.* ACTUAL FIGURES

Figure 7 illustrates the rise of patients and death cases predictions v/s actual scenario in columns (a) and (b) respectively. If we observe the trends for confirmed cases, the ARIMA_retraining model certainly seems to be fitted very well for all the countries. It has almost overlapped the actual cases trendline. Simple ARIMA model predictions are also close to the actual figures in comparison to LSTM. LSTM is found to be deviating from the actual trend and is predicting a higher figure. When there is a sudden rise in casualties in Japan, the ARIMA_retraining deviated for one day. In case of prediction of death cases, again ARIMA_retraining has been the closest except for the UK. Simple ARIMA predicted better figures for death cases in the UK. LSTM predictions

Fig. 5. Valuation of RMSE for proposed models to predict (a) confirmed cases (b) death cases.

Fig. 6. Valuation of AIC for proposed ARIMA models to predict future cases and deaths.

follow the same trendline but predicted figures are higher than actual figures. In conclusion, we can say that the ARIMA_retraining model has given the best fit for the COVID-19 prediction for six countries.

4 CONCLUSION

In this article, two data-driven models are used to predict the mean absolute percentage error and mean squared errors of the possible number of cumulative COVID-19 suspected cases and cumulative numbers of COVID-19 death cases in India, Italy, Japan, Spain, UK, and the US. The prediction for these countries is done by using two techniques, namely, ARIMA and LSTM. The prediction of various parameters (the number of cumulative cases and the number of death cases) estimated by the proposed method will be a beneficial tool for administrators and health officials. This prediction will help them to make the necessary arrangement for beds, medicine, and other equipment like ventilators in hospitals. Moreover, it can help in deciding the need and protocol for lockdown in various countries. The presented Covid-19 prediction model and its use may have serious implications on the health, social, and economic aspects of a country if used practically.

Fig. 7. Comparison of model-predicted figures with actual figures for different countries.

DECLARATION OF COMPETING INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this article.

REFERENCES

- Long-ting Cao, Hong-hui Liu, Juan Li, Xiao-dong Yin, Yu Duan, and Jing Wang. 2020. Relationship of meteorological factors and human brucellosis in Hebei province, China. *Sci. Total Environ.* 703 (2020), 135491.
- Ye Chen, Kunkun Leng, Ying Lu, Lihai Wen, Ying Qi, Wei Gao, Huijie Chen, Lina Bai, Xiangdong An, Baijun Sun, Ping Wang, and Jing Dong. 2020. Epidemiological features and time-series analysis of influenza incidence in urban and rural areas of Shenyang, China, 2010–2018. *Epidem. Infect.* 148 (2020).
- Xinyu Fang, Wendong Liu, Jing Ai, Mike He, Ying Wu, Yingying Shi, Wenqi Shen, and Changjun Bao. 2020. Forecasting incidence of infectious diarrhea using random forest in Jiangsu Province, China. *BMC Infect. Dis.* 20, 1 (2020), 1–8.
- Jeannette Guarner. 2020. Three emerging coronaviruses in two decades: The story of SARS, MERS, and now COVID-19. *Amer. J. Clin. Pathol.* 153, 4, (2020), 420–421. DOI : [10.1093/ajcp/aqaa029](https://doi.org/10.1093/ajcp/aqaa029)
- Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.* 9, 8 (1997), 1735–1780. Retrieved from <https://adventuresinmachinelearning.com/keras-lstm-tutorial/>.
- Weicong Kong, Zhao Yang Dong, David J. Hill, Fengji Luo and Yan Xu. 2017. Short-term residential load forecasting based on resident behaviour learning. *IEEE Trans. Power Syst.* 33, 1 (2017), 1087–1088.
- Jeanne Lenzer. 2020. Covid-19: US gives emergency approval to hydroxychloroquine despite lack of evidence. *Brit. Med. J.* 369 (2020), m1335.
- Jiban Chandra Paul, Md. Shahidul Hoque, and Mohammad Morshedur Rahman. 2013. Selection of Best ARIMA Model for Forecasting Average Daily Share Price Index of Pharmaceutical Companies in Bangladesh: A Case Study on Square Pharmaceutical Ltd. *Global J. Manag. Bus. Res.* 13(3(1)) (2013), 14–25.
- Xiang yun Qing and Yugang Niu. 2018. Hourly day-ahead solar irradiance prediction using weather forecasts by LSTM. *Energy* 148 (2018) 461–468.
- Andrea Remuzzi and Giuseppe Remuzzi. 2020. COVID-19 and Italy: What next? *The Lancet* 395 (10231) (Mar. 2020).
- G. Peter Zhang. 2003. Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing* 50 (2003), 159–175.
- Zheng Zhao, Weihai Chen, Xingming Wu, Peter C. Y. Chen, and Jingmeng Liu. 2017. LSTM network: A deep learning approach for short-term traffic forecast. *IET Intell. Transport. Syst.* 11, 2 (2017), 68–75.

Received May 2020; revised June 2020; accepted June 2020