Journal of Theoretical and Applied Information Technology
15th January 2021. Vol.99. No 1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

INTEGRATING DATA WAREHOUSE AND MACHINE LEARNING TO PREDICT ON COVID-19 PANDEMIC EMPIRICAL DATA

[1]HASAN HASHIM, [1,2] EL-SAYED ATLAM, [1]MALIK ALMALIKI, [1,2] RASHA EL-AGAMY, [2,3] M. M. EL-SHARKASY, [4]GUESH DAGNEW, [2]IBRAHIM GAD, [5]OSAMA GHONEIM

[1]College of Computer Science and Engineering, Taibah University, Yanbu, Saudi Arabia.
[2]Faculty of Science, Tanta University, Tanta, Egypt.
[3]Department of Mathematics and Statistic, Faculty of Science, Yanbu, Taibah University, KSA.
[4]Department of Computer Science, Institute of Technology, Dire Dawa University, Ethiopia.
[5]Faculty of Computers and informatics, Tanta University, Tanta, Egypt.

E-mail: satlam@taibahu.edu.sa, satlam@yahoo.com

ABSTRACT

The world has recently been plagued by the pandemic of Corona Virus Disease 2019 (COVID-19). Since it is reported in Wuhan city of China, on the 8th of December 2019, the COVID-19 invaded every country around the world. As of October 24th, 2020, a total of 42,549,383 confirmed cases of COVID-19 were officially announced and the death toll was 1,150,163. Globally, huge volumes of datasets are generated regarding COVID-19 pandemic to open new research arena for machine learning and artificial intelligence researchers. In this work, an integration of data warehouse with deep learning approach, namely LSTM model, is introduced to predict the spread of the COVID-19 in selected countries. We present the design and development of COVID-warehouse, a data warehouse that integrates and stores the COVID-19 data made available daily by different countries. The basic idea of the framework is to use a COVID19 time-series dataset for analysis by machine learning models to make forecasting of future trend based on present values. Ultimately, the proposed prediction model can be applied to predict for other countries as the nature of the virus is the same everywhere. In terms of R2 metric, the experimental results of the decision tree model outperforms other models for recovery cases compared with confirmed and death cases. Recovery cases have a R2 of 0.996011, death cases have a R2 of 0.993124 and confirmed cases have a R2 of 0.991676. Finally, our results emphasize the importance of enforcing the public health advice of social distancing as well as applying the infection control measures to combat COVID-19 before it becomes too late.

Keywords: *COVID-19 Virus, Infection control, Artificial intelligence, Data Warehouse, Deep Learning Model, Prediction.*

1. INTRODUCTION

The pandemic of Corona Virus Disease 2019 (COVID-19) that started in Wuhan, China 2019 became a global pandemic and critical health issue. COVID19 has quickly spread worldwide within a few months. Although its real cause of spread is undetermined, COVID-19 has a direct impact on the physical wellbeing of the human, and psychological health as well. According to recent studies, the way the infection transmit is via human-to-human transmissions such as through droplets or direct contact [1].

The World Health Organization (WHO) has officially declared COVID-19 as an infectious disease caused by new corona virus. Since this COVID-19 virus is discovered recently there are too much to be researched about it. Generally, people of all age are affected, but results have shown that the elderly are more liable to be influenced to the disease. In addition, people who have medical problems such as diabetes, cardiovascular disease, cancer, and chronic respiratory disease are more susceptible COVID-19 [2]. At present time, we know about COVID-19 that it spreads from one person to another by occurring

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

direct interaction and is highly contagious. In addition, the WHO has specified that the main human-to-human transmission mechanism varies, but still can be generalized as direct contact with an infected person through shaking hands, exposure to droplets coming out during coughing or sneezing and by travelling to an affected area and attaining the virus in one or other way.

The core symptoms of COVID-19 highly vary, ranging from being severely affected to being asymptomatic. In most cases, the common symptoms are fatigue, general weakness, sore throat, high fever, cough and muscular pain. While, in severe cases, micro coagulopathies, pneumonia, acute respiratory distress syndrome, sepsis and septic shock could occur, and in lots of patients, it lead to death [3, 4].

COVID-19 is a family of Severe Acute Respiratory Syndrome Corona-virus-2 (SARS-CoV-2). Since it has been officially announced in Wuhan city of China, in December 8th, 2019 the COVID-19 invaded 210 countries and territories around the world [5]. As for October 24th, 2020, in just ten months, a total number of 42,549,383 confirmed cases of COVID-19 were recorded and the death number was 1,150,163 deaths [6]. So that, COVID-19 is a reason for a major worry because it spreads at an alarmingly exponential rate. This gives the government's insufficient time for the right planning in terms of appropriate medical supplies and other measures that need to be taken to fight this pandemic or decrease the number of people affected or killed by it.

However, by now, some information is known about COVID-19 but, its full features are still obscure. One of the characteristics of this virus is that it can change its nature frighteningly quickly due to its accelerated genetic mutations. The research scientists have not yet been able to know the full characteristics of it. Consequently, they are continuously carrying out studies just to found the typically facts about COVID-19 that will help them to reduce its spreads or ending [7, 8].

Many recent studies on COVID-19 have employed data science and analytics to accurately provide the number of confirmed and death cases. This helps the governments for devising better and informed public policies such as the time and duration of lockdowns and focusing on important and critical regions in the countries. As a result of the effective public policy, the health sector can be supported and equipped more efficiently in terms of preparation and procurement of masks, ventilators and others medical tools. Using data warehouse to store the COVID-19 dataset and machine learning to predict the confirmed cases and deaths, governments can raise awareness among the public sector which in turn helps in the prevention of a large number of potential future deaths.

The main contributions of this research paper are as follows:

• Using data science and analytics to accurately provide the number of confirmed and death cases.

• providing an LSTM approach to predict the spread of the virus in a number of selected countries.

• devising prediction model can be applied to predict the spread of COVID-19 infections in any other country since the nature of the virus is nearly the same everywhere.

This paper is organized as follows: Section 2 presents the related works. Section 3 introduces our new methodology and the proposed approaches. Section 4 presents the experimental observations. Finally, conclusion and possible future work are introduced in section 5.

2. RELATED WORK

The WHO has declared a global public health emergency of COVID-19. Following this many countries declared lock-down and ban local and international flights [9].

Machine learning and artificial intelligence models are essentially used to improve the performance in the accuracy of prediction for diagnosis and screening infectious and non-infectious diseases [10]. Moreover, machine learning approaches are also widely used in the analyze and prediction of COVID-19 survival rate and discharge-time of patients based on clinical data[11].

Lai et al. [12] studied the epidemic nature of COVID19 regarding every day total record, death rate, and affiliated status of the nations medical care assets and economy. With the disastrous outbreak of COVID-2019 around the world, a tremendous volume of information is created in a split second that opens a hot exploration subject for AI and artificial intelligence researchers.

Punn et al. [13] have proposed the utilization of machine learning and deep learning models to understand the behavior of the virus based on the data taken from Johns Hopkins dashboard.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Dandekar and Barbastathis [14] suggested a hybrid model consists of first-principles epidemiological equations and data-driven neural network to forecast the halting of the spread of the COVID-19 infection. They have used a neural network model to predict for four locations namely Wuhan city, Italy, South Korea, and the US. Finally, for the US, they predicted the currently infected growth curve and predicted a halting of infection by 20 April 2020.

Giuseppe et al. [15] introduced the design and advancement of COVID-WAREHOUSE, a data warehouse that models, incorporates and stores the COVID19 information made accessible daily by the Italian Protezione Civile Department and several pollution and climate data and made it accessible by the Italian Regions.

Zohair et al. [16, 17] proposed different regressor machine learning models that extracted the connection between various factors with the spreading rate of COVID-19 and analysis the risk of second rebound of COVID19 pandemic. The machine learning algorithms utilized in this work estimate the effect of climate factors, such as temperature and humidity on the transmission of COVID-19 by extracting the connection between the number of confirmed cases and the climate factors on specific regions.

According to the WHO report on guidelines to protect against COVID-19, [18], it enters the human body via different parts such as eyes, nose and/or mouth. Hence, it is important to avoid touching the face with unwashed hands. Washing of hands with soap and water for at least 20 seconds, or cleaning hands thoroughly with alcohol-based solutions, gels or tissues is recommended in all settings. It is also recommended to stay at least one meter or more away from one another to reduce the risk of infection through respiratory droplets. COVID-19 spreads rapidly in droplets and on surfaces.

Previous researchers focused on developing techniques to accomplish precise and time-efficient for forecasting of the spread of COVID-19. The main drawbacks of the previous works was, they utilize a predictive model that gives less accurate outcomes in some cases. Concerning the above-related work on COVID-19, there were good thoughts to improve indicates an ascending trend for the cases in the coming days. Previous works lack some promising features that could enable us to predict the highest possible accuracy of the COVID-19 infections, thus slowdown the spread of virus.

3. METHODOLOGY

Figure 1 gives a conceptual view of the proposed framework for the prediction of potential COVID-19 cases. The basic idea of the framework is to apply multiple machine learning algorithms on a time-series dataset of COVID-19 to predict potential future cases.

Every rectangle in the diagram represents a different step in the proposed framework. At each block, the technology used in the development is mentioned alongside open source and third party tools. For the first block, datasets from different sources were collected and merged and then pushed on to the data warehouse block. The machine learning algorithms used are mentioned in the third block which are developed using built-in python libraries such as TensorFlow, Keras, and Stat. Finally, the predictions and forecasting are done in the fourth block, the results are shown on interactive dashboards developed in Tableau.

The main components of COVID-19 data-warehouse are: fact tables, dimension tables, and their relationships. A dimension table contains a number of samples and their corresponding attributes/columns. The main operations that can be applied to the high-dimensional table are grouping, filtering and labeling. Moreover, the samples in the dimension tables are identified by a unique Key and each column represents a range of values.

The available data set is collected for each day from Github, and Kaggle websites. Figure 2 demonstrates the proposed star schema on COVID-19 datasets that contains two categories of tables and it presents one fact table and 5-dimensional tables. The day fact table has five dimension tables, namely location, date, weather, population and tests performed. In general, the fact table consists of a number of primary keys and many keys that refer to their corresponding dimension table. On the other hand, the primary key in the dimension table has the same name as the table that contains observations collected from a particular country.

In the subsequent subsections, we have described the datasets used to validate the proposed method in subsection 3.1 and the detail description of the proposed method incorporated in subsection 3.2.

3.1 Dataset Description

To validate our work, we have used the data from official data repositories such as WHO and Johns Hopkins University and Worldometer official website [6, 19, 20, 21]. The main variables that the dataset includes are Date, Country, Confirmed

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

cases and Deaths. The data shows total COVID-19 confirmed positive cases daily, total death rate, and total and daily recoveries. Usually, increasing the number of variables of a dataset improves the performance of machine learning algorithms in terms of the metric of choice such as loss, accuracy and RMSE. Since the COVID-19 is new, the dataset is small, and also the number of features available is limited. To increase the size of the dataset, more relevant features such as Population, and Tests performed were added to the base data source. Table 1 presents a sample of the top countries sorted by the number of confirmed cases. The Table depicts the time-series summary for confirmed, death and recovered cases of COVID-19 from the following countries US, India, Brazil, Russia, Spain, Argentina, France, Colombia, Peru, and Mexico.

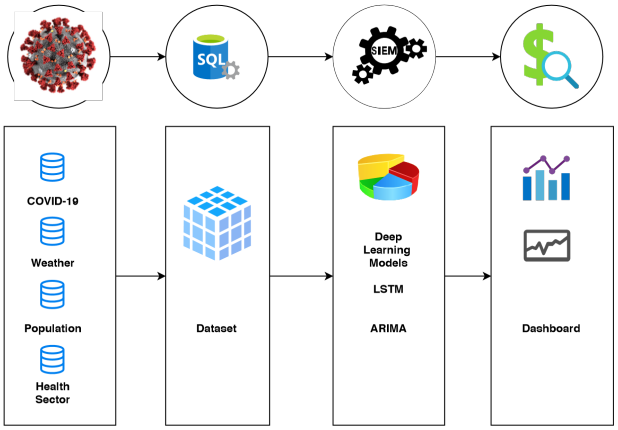

Figure 1: Architucher Of COVID-19 Prediction Framework

3.2 The Proposed Models

Table 2: The description of the proposed LSTM model.

Layer (type)	Output Shape	Param #
lstm_4 (LSTM)	(None, 200)	161600
lstm_5 (LSTM)	(None, 7, 200)	320800
time_distributed_3 (TimeDistributed	(None, 7, 100)	20100
time_distributed_4 (TimeDistributed	(None, 7, 1)	101
Total params		502,601
Trainable params		502,601

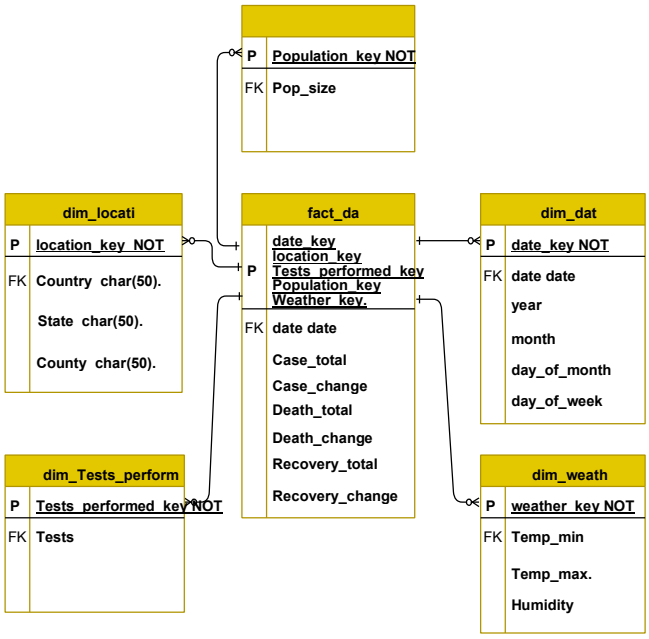

Figure 2: Star schema of COVID-19 data warehouse.

The main models developed and tested for the framework are Long Short-Term Memory (LSTM), decision tree and ARIMA.

In this work, we have used a deep learning model namely LSTM which is a variant of a Recurrent Neural Network (RNN) and it is a powerful tool in predicting sequence and time-series data-related problems. LSTM model was selected in this work as COVID-19 dataset is also time-bound. The LSTM has a special feature provides the ability to store previous information and predict the future trend of the virus under consideration. In predicting the actual span of the COVID-19, we have applied the LSTM and Decision Tree models that enable to predict if the span of COVID-19 can slow down or rise-up. The LSTM model takes in the last three day's features and predicts the figures for the next 7 days. When it reaches a point where the target value ceases to exist, it takes into account its predictions.

The LSTM model that we have used in this study has the following configurations. It uses 2000 epochs and a loss function namely mean squared logarithmic error with *adam optimizer*. The Relu activation function is used in all the input, hidden and output layers. The total number of tunable parameters used in the proposed model is summarized in Table 2.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

4. EXPERIMENTAL RESULTS AND ANALYSIS

In this section, we present and discuss the experimental results of the proposed method. The experimental results are presented visually and tabular.

4.1. Experimental Results

Currently, it is feasible to predict how long the outbreak of COVID-19 will last and how the epidemic will unfold, this is because of the new features exhibited by COVID-19 and a lot of uncertainties remains problematic. Some domain experts remain optimistic during summer in the northern hemisphere, as they consider that COVID-19 will be like the ordinary seasonal influenza. With the help of machine learning, we have developed a predictive model using the available data of COVID-19 taken from famous data repository websites around the globe.

According to the WHO, the first corona-virus that was detected in the Chinese city of Wuhan last December has infected more than 25,409,299 people in at least 210 countries and territories globally, 6,396,551 of which are in US only as shown in Table 1. Out of the infected cases, more than 829,226 people had died. China was the first country that has more than 80,000 reported infections as shown in Table 1. To contain the COVID-19 outbreak, Chinese authorities locked down cities, restricted movements of millions and suspended business operations.

Almost every country and union territory have declared a lock-down time to prevent the outbreak of COVID-19. Figure 3 shows the Lock-down days imposed in the USA and China versus the number of confirmed cases. As it is presented the Figure indicate that China had an effective lock down following the outbreak of COVID-19. China has declared to put Wuhan City, the centre of the outbreak, on lock down on January 23. Before the lockdown time, the growth rate of the pandemic was 0.054 and after imposing the lock-down policy the rate decrease to 0.001. Therefore, China's lock-down considered as an effect model as the spread curve of the virus is getting flatten over time. Although it is not as effective as mainland China, the US growth rate of the COVID-19 has declined after the lock-down. The growth rate for the US is 0.156 before the lock down and 0.014 after the lock down.

The proposed method has forecasted the possible confirmed cases for the upcoming 7 days in the US. Experimental results show that the confirmed cases are exponentially increasing from a few hundreds of thousands to nearly six and a half million.

Our observation at this particular point is that the prediction is not so optimal as we have used few numbers of records in our deep learning model that is a challenging problem to train deep learning models using few datasets. To validate the performance of the proposed model, we have used root mean square error on each of the four attributes namely confirmed case, deaths, recoveries and growth rate.

4.2. Comparison of the Proposed LSTM and Decision Tree Methods

Unlike other diseases, COVID-19 is still spreading worldwide [16]. Moreover, this virus has infected more people than recent outbreaks such as SARS or Ebola and it does not hit the scale of the most massive modern pandemics such as H1N1 or the seasonal flu. Every year the seasonal flu infects millions of people, and it's not life-threatening for most people who gets infected. In contrast, total reported cases of Ebola is less than 30,000, but it was treated as a crisis due to the large number of reported deaths.

Currently, COVID19 is more deadly than the normal flu, but its mortality rate 6.87% of is low compared to the mortality rates of other recent outbreaks such as MERS or Ebola which recorded 34.40% and 39.53%, respectively.

Almost, many common key symptoms exist in all pandemics such as cough, fever, and shortness of breath. Moreover, people of all ages are prone to infection of COVID-19. However, pandemics are normally deadliest among older patients with the weaker immune system. The mortality rate multiplied rapidly as patients got older reaching its highest rate among patients over 65.

Table 3 presents the results of the proposed LSTM model for Confirmed, Recoveries, and Deaths cases. In comparison to Confirmed, Death and Recovery cases, which are collected in the past nine months, the proposed model showed a slightly higher prediction performance for Recovery cases than Confirmed and Death cases. Recoveries have a R2 of 0.989648, Deaths has a R2 of 0.977564 and Confirmed has a R2 of 0.986723.

Similarly, Table 4 presents the results of the proposed decision tree model for Confirmed, Recoveries, and Deaths cases. In terms of R2 metric, the decision tree model outperforms for Recovery cases compared with Confirmed, Death cases. Recoveries have a R2 of 0.996011, Deaths

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

has a R2 of 0.993124 and Confirmed has a R2 of 0.991676. Moreover, Table 5 shows that the predicted values and the actual values are very close based decision tree model.

4.3 Estimation of slowdown of the COVID-19

To contain the fast-spreading COVID-19, many countries around the world applied an effective shutdown. The protective measures have placed many restrictions on the daily lives of millions of people, such as school closures, large scale social distancing, and bans on public gatherings. Because it is not easy to know exactly when a vaccine for COVID-19 becomes available, these protective measures will be extended for the next few months. However, health experts are much more cautious. That's because lifting restrictions to alleviate the economic and social damage of a prolonged period of lock down could risk a second wave of COVID-19 cases.

Table 6 shows the prediction of COVID-19's deadlines using the collected data. The experimental results show that the expected number of globally confirmed cases will be 82516 on March 1, 2021, and after a two weeks, on March 17, 2021, the number of confirmed cases will remain 43835. Likewise, the expected deadline to stop the pandemic globally is forecasted in this work and the experimental result is presented in the same table. The results also show that the number of confirmed cases on March 01, 2021, is 100386 followed by 38913 confirmed cases on March 27, 2021.

Table 6 presents the estimated duration for the virus to decrease its infectiousness nature in the US. The experimental results of the predictive model, the US will have 1210960 confirmed cases on February 12, 2021, and two months later (on April 24, 2021), the expected number of positive cases is 16 patients. Moreover, the expected deadlines to stop the pandemic in Brazil and India are forecasted in this work and the experimental results are presented in the same table.

In this work, we have experimentally proved that the model parameters vary from country to country as the data for each country substantially differs. Considering the relationship within the data, the SARIMA model (p,d,q) × (P, D, Q)s is successfully applied to different time-series data. The period value of time-series of time-series s (seasonality) is considered based on the dataset. Since the daily data for a few months have been used, the value of s is assigned to be 3,7,12.

The best forecasting SARIMA model parameters are selected based on the minimum values of AIC, and P-values that are less than 0.05. Table 7 presents the AIC values of different forecasting models. The following SARIMA(9,0,8) × (0,0,0,3) model has the lowest AIC values as shown in Table 8. The best combination of the parameters (9,0,8) × (0,0,0,3) is considered to be the best for the corresponding model.

To train and validate the SARIMA model,the COVID-19 data was divided into training and testing dataset on the basis of 70% and 30% ratio for training and validation for testing for each country. The training set comprises from 2020-01-22 to 2020-07-26 and the testing set is from 2020-07-27 to 2020-09-30.

Table 9 presents the forecasting values of the confirmed cases with lower and upper confidence limits that are calculated using the SARIMA model for the period from 2020-10-02 to 2020-11-15. Figure 4 shows the observed (marked in blue line) or training set from 22-Jan-2020 to 15-Jun-2020 and the testing set from 15-Jun-2020 to present-day and values for one step ahead forecasting is presented by the red line. Figure 5 depicts the forecasting of confirmed cases for the next days in the Global. Figure 6 presents the possible time period that the virus can slowdown from being infectious in the US

5. CONCLUSION

Based on LSTM, we developed a predictive model that can estimate the spread of the daily COVID-19 infection in top countries that are highly affected by the pandemic and the expected period after which the virus can stop. We have presented the design and development of COVID-WAREHOUSE, a data warehouse that models, integrates and stores the COVID-19 data made available daily by different countries. Globally, our results forecasted that the COVID-19 infections will greatly decline during the first week of April 2021 when they will be going to an end shortly afterwards. Moreover, we can apply the proposed model to other countries that are affected by COVID-19. Additionally, the proposed model could also, evaluate the effectiveness of the public health guidelines, infection control measures and the lock-down decisions that were taken to control the COVID-19 pandemic.

Future work could focus on improving the performance of our model by using big data as training data. Moreover, our proposed model can be applied to other countries cases to evaluate the spread rate and decline rate of the virus by adapting to their respective data with respect to the COVID-19 pandemic.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

REFERENCES

[1]C.-C. Lai, T.-P. Shih, W.-C. Ko, H.-J. Tang, P.-R. Hsueh, Severe acute respiratory syndrome coronavirus 2 (sars-cov-2) and corona virus disease-2019 (covid-19): the epidemic and the challenges, International journal of antimicrobial agents (2020) 105924.

[2]W. H. O. (WHO), Coronavirus, 2020 (accessed April 13, 2020).URL:https://www.who.int/health-topics/ coronavirus.

[3]H. Qiu, J. Wu, L. Hong, Y. Luo, Q. Song, D. Chen, Clinical and epidemiological features of 36 children with coronavirus disease 2019 (covid-19) in zhejiang, china: an observational cohort study, The Lancet Infectious Diseases (2020).

[4]J. Wu, J. Liu, X. Zhao, C. Liu, W. Wang, D. Wang, W. Xu, C. Zhang, J. Yu, B. Jiang, et al., Clinical characteristics of imported cases of covid-19 in jiangsu province: A multicenter descriptive study., Clinical infectious diseases: an official publication of the Infectious Diseases Society of America (2020).

[5]C. Huang, Y. Wang, X. Li, L. Ren, J. Zhao, Y. Hu, L. Zhang, G. Fan, J. Xu, X. Gu, et al., Clinical features of patients infected with 2019 novel coronavirus in wuhan, china, The Lancet 395 (2020) 497–506.

[6]Worldometer, COVID-19 CORONAVIRUS PANDEMIC, 2020 (accessed April 13, 2020). URL: https://www.worldometers.info/coronavirus/.

[7]P. Yang, P. Liu, D. Li, D. Zhao, Corona virus disease 2019, a growing threat to children?, The Journal of Infection (2020).

[8]K. Naeem, M. Riaz, X. Peng, D. Afzal, Pythagorean mpolar fuzzy topology with TOPSIS approach in exploring most effectual method for curing from COVID-19, International Journal of Biomathematics (2020). URL: https://doi. org/10.1142%2Fs1793524520500758. doi:10.1142/ s1793524520500758.

[9]C. Sohrabi, Z. Alsafi, N. O'Neill, M. Khan, A. Kerwan, A. AlJabir, C. Iosifidis, R. Agha, World health organization declares global emergency: A review of the 2019 novel coronavirus (covid-19), International Journal of Surgery (2020).

[10]S. Lalmuanawma, J. Hussain, L. Chhakchhuak, Applications of machine learning and artificial intelligence for covid-19 (sarscov-2) pandemic: A review, Chaos, Solitons & Fractals (2020) 110059.

[11]M. Nemati, J. Ansary, N. Nemati, Machine-learning approaches in covid-19 survival analysis and discharge-time likelihood prediction using clinical data, Patterns 1 (2020) 100074.

[12]C.-C. Lai, C.-Y. Wang, Y.-H. Wang, S.-C. Hsueh, W.-C. Ko, P.R. Hsueh, Global epidemiology of coronavirus disease 2019: disease incidence, daily cumulative index, mortality, and their association with country healthcare resources and economic status, International Journal of Antimicrobial Agents (2020) 105946.

[13]N. S. Punn, S. K. Sonbhadra, S. Agarwal, Covid-19 epidemic analysis using machine learning and deep learning algorithms, medRxiv (2020).

[14]R. Dandekar, G. Barbastathis, Quantifying the effect of quarantine control in covid-19 infectious spread using machine learning, medRxiv (2020).

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[15]G. Agapito, C. Zucco, M. Cannataro, COVID-WAREHOUSE: A data warehouse of italian COVID-19, pollution, and climate data, International Journal of Environmental Research and Public Health 17 (2020) 5596. URL: https://doi.org/10.3390%2Fijerph17155596. doi:10.3390/ijerph17155596.

[16]Z. Malki, E.-S. Atlam, A. E. Hassanien, G. Dagnew, M. A. Elhosseini, I. Gad, Association between weather data and COVID-19 pandemic predicting mortality rate: Machine learning approaches, Chaos, Solitons & Fractals 138 (2020) 110137. URL: https://doi.org/10.1016%2Fj.chaos.2020.110137. doi:10.1016/j.chaos.2020.110137.

[17]Z. Malki, E.-S. Atlam, A. Ewis, G. Dagnew, A. R. Alzighaibi, G. ELmarhomy, M. A. Elhosseini, A. E. Hassanien, I. Gad, ARIMA models for predicting the end of COVID-19 pandemic and the risk of second rebound, Neural Computing and Applications (2020).URL: https://doi.org/10.1007%2Fs00521-02005434-0. doi:10.1007/s00521-020-05434-0.

[18]W. H. Organization, et al., Rational use of personal protective equipment for coronavirus disease (COVID-19): interim guidance, 27 February 2020, Technical Report, World Health Organization, 2020.

[19]Kaggle, covid19 global weather data, Kaggle (2020).URL: https://www.kaggle.com/winterpierre91/

[20]C. GIS, Data, Covid-19, github (2020). URL: https:// github.com/CSSEGISandData/COVID-19.

[21]Kaggle, corona virus report, Kaggle (2020). URL: https://www.kaggle.com/imdevskp/coronavirus-report.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 4: The performance results for decision tree model.

	R2	MAPE	ME	MAE	MPE	MSE	RMSE	Corr	Minmax
Confirmed	0.991676	0.003901	-15029.6	84173	-0.000795	2.46831e+10	157109	0.99608	0.003885
Deaths	0.993124	0.002140	-445.28	1598.56	-0.000552	9.83675e+06	3136.36	0.996703	0.002134
Recoveries	0.996011	0.002633	-6728.92	40104	-0.000488	8.15934e+09	90329.1	0.998015	0.002619

Table 5: The predicted values and the actual values for decision tree model.

Date	Confirmed	Predicted Confirmed	Recoveries	Predicted Recoveries	Deaths	Predicted Deaths
2020-09-06	24475721	24475721.0	16222510	16222510.0	804217	813329.0
2020-09-07	24668692	24475721.0	16411265	16411265.0	813329	813329.0
2020-09-08	24874598	24874598.0	16585974	16411265.0	817777	817777.0
2020-09-09	25141647	25141647.0	16826342	17028134.0	823785	823785.0
2020-09-10	25409299	25409299.0	17028134	17028134.0	829226	829226.0

Table 6: Expected deadline with/without forecasting for Global and top countries.

	The First Rebound									
Country	First confirmed case	Estimation without Forecasting					Estimation with Forecasting				
		Peak point	Start Date %	End Date %	Start Value	End Value	Peak point	Start Date %	End Date %	Start Value	End Value
Global	2020-01-22	2020-09-10	2021-03-01	2021-03-17	82516	43835	2020-09-17	2021-03-01	2021-03-27	100386	38913
US	2020-01-22	2020-09-23	2021-02-12	2021-04-24	1210960	16	2020-10-24	2021-04-01	2021-06-19	1492437.0	17.0
Brazil	2020-02-26	2020-09-23	2020-12-19	2021-02-18	374898.0	2554.0	2020-10-24	2021-02-04	2021-04-15	691758.0	4256.0
India	2020-01-30	2020-09-23	2021-01-31	2021-04-09	62808.0	5.0	2020-10-24	2021-03-19	2021-06-04	124794.0	31.0

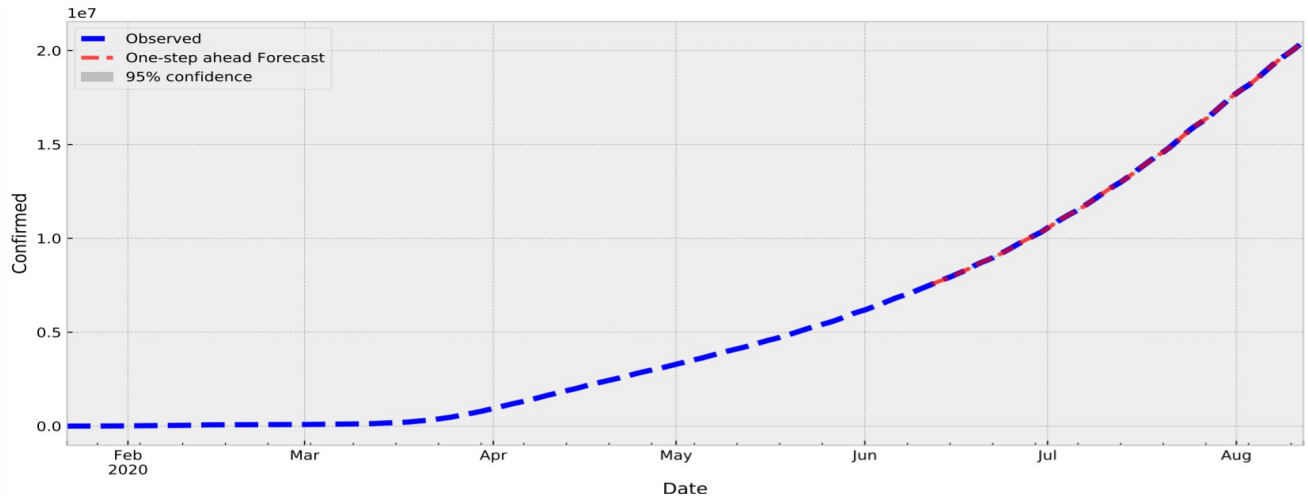

Figure 4 Comparison between the observed and predicted values (one-step ahead result) for SARIMA model on COVID-19 dataset

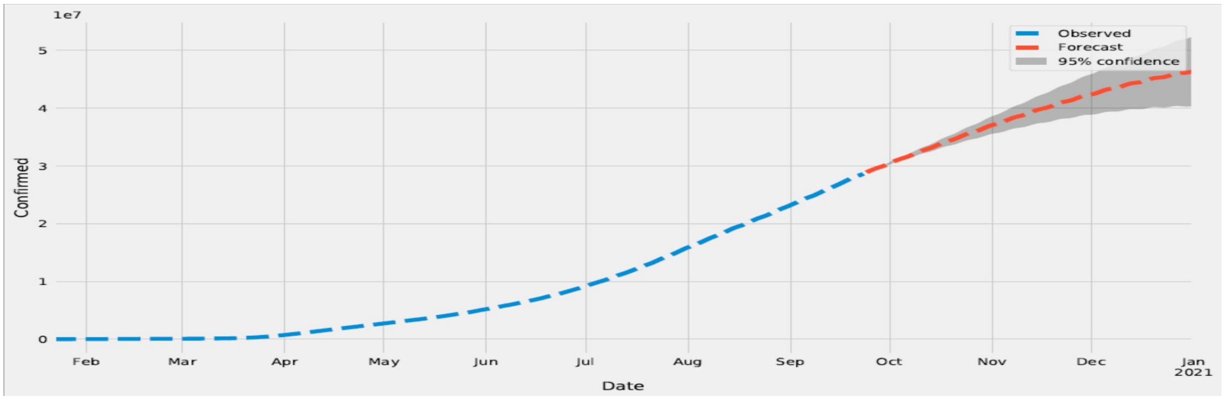

Figure 5: The forecasting of confirmed cases for Global

Table 7: Experimental results of the diagnostics test for SARIMA models that have p-values less than 0.05 for India.

(p,d,q)	(P,D,Q,s)	AIC	MAPE	MAE	MPE	MSE	RMSE	Corr	MinMax
(9, 0, 0)	(0, 0, 1, 3)	-2091.82	15.0456	1.88012	4.2979	3.5401	1.88151	0.990667	1.0466
(9, 0, 0)	(0, 0, 1, 3)	-2091.82	15.0456	1.88012	4.2979	3.5401	1.88151	0.990667	1.0466
(9, 0, 0)	(0, 0, 1, 3)	-2091.82	15.0456	1.88012	4.2979	3.5401	1.88151	0.990667	1.0466
(9, 0, 0)	(0, 0, 1, 3)	-2091.82	15.0456	1.88012	4.2979	3.5401	1.88151	0.990667	1.0466

Table 8: Experimental results of the diagnostics test for SARIMA models that have p-values less than 0.05 for Global.

(p,d,q)	(P,D,Q,s)	AIC	MAPE	MAE	MPE	MSE	RMSE	Corr	MinMax
(9, 0, 9)	(0, 0, 0, 3)	-2677.59	11.4808	1.16488	0.641154	1.36114	1.16668	0.998178	0.789146
(9, 0, 9)	(0, 0, 0, 7)	-2677.59	11.4808	1.16488	0.641154	1.36114	1.16668	0.998178	0.789146
(9, 0, 9)	(0, 0, 0, 12)	-2677.59	11.4808	1.16488	0.641154	1.36114	1.16668	0.998178	0.789146
(9, 0, 8)	(0, 0, 0, 3)	-2676.17	11.4721	1.16237	0.635728	1.35553	1.16427	0.998082	0.788679

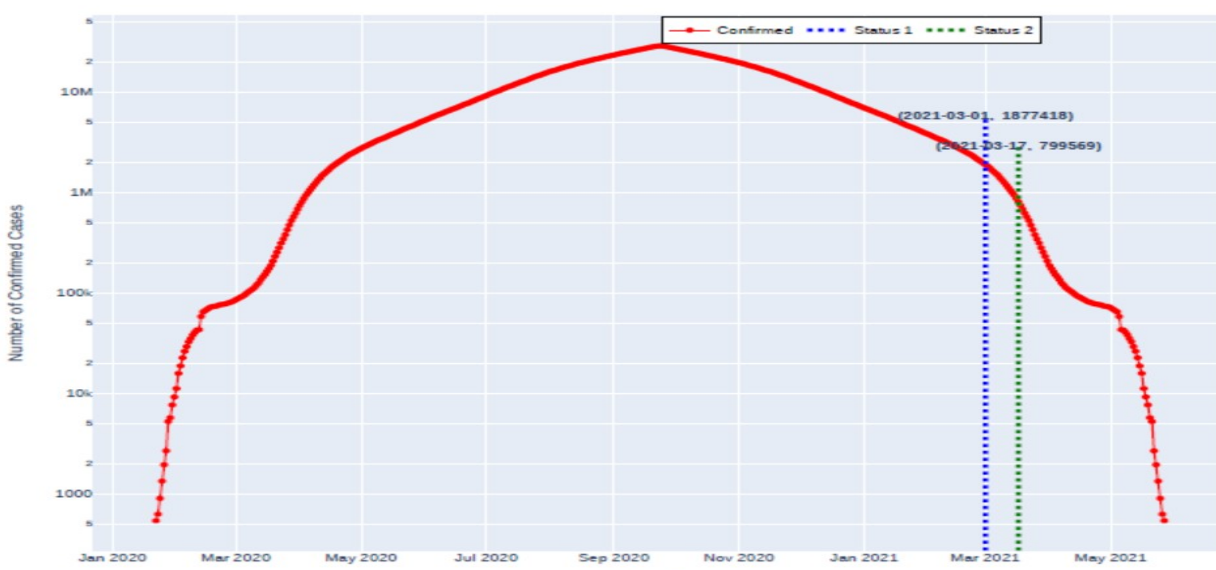

Figure 6: Expected Dead line for Global without forecasting COVID-19

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 9: The forecasted values and the lower, upper values for LSMT model

	USA			Brazil			India		
Date	predict	lower	upper	predict	lower	upper	predict	lower	upper
2020-10-02	7,335,951	7,323,804	7,348,097	4,896,882	4,874,609	4,919,156	6,479,419	6,471,034	6,487,804
2020-10-03	7,382,116	7,364,044	7,400,189	4,922,774	4,893,761	4,951,786	6,567,945	6,555,746	6,580,145
2020-10-04	7,418,254	7,393,612	7,442,897	4,936,708	4,902,063	4,971,352	6,654,014	6,637,263	6,670,765
2020-10-05	7,459,981	7,427,596	7,492,366	4,952,415	4,913,070	4,991,761	6,731,522	6,709,704	6,753,340
2020-10-06	7,501,985	7,461,578	7,542,392	4,978,739	4,935,484	5,021,994	6,811,098	6,784,085	6,838,112
2020-10-07	7,538,712	7,488,998	7,588,425	5,013,786	4,965,400	5,062,173	6,900,030	6,866,932	6,933,129
2020-10-08	7,585,963	7,525,382	7,646,545	5,057,321	4,999,738	5,114,905	6,989,016	6,948,352	7,029,680
2020-10-09	7,642,361	7,569,384	7,715,339	5,090,082	5,022,742	5,157,421	7,074,866	7,025,543	7,124,189
2020-10-10	7,686,964	7,600,899	7,773,030	5,113,533	5,036,640	5,190,426	7,165,648	7,106,534	7,224,763
2020-10-11	7,721,577	7,621,285	7,821,868	5,127,608	5,041,978	5,213,237	7,254,680	7,184,811	7,324,549
2020-10-12	7,763,552	7,648,251	7,878,854	5,143,089	5,049,768	5,236,410	7,334,530	7,253,409	7,415,651
2020-10-13	7,803,453	7,672,700	7,934,206	5,166,735	5,066,281	5,267,190	7,415,690	7,323,343	7,508,036
2020-10-14	7,836,641	7,689,654	7,983,628	5,199,870	5,091,086	5,308,653	7,507,904	7,403,698	7,612,110
2020-10-15	7,881,992	7,717,348	8,046,636	5,237,538	5,118,274	5,356,802	7,600,155	7,483,019	7,717,291
2020-10-16	7,937,696	7,754,363	8,121,029	5,267,929	5,137,419	5,398,438	7,687,777	7,556,826	7,818,728
2020-10-17	7,978,403	7,775,749	8,181,057	5,289,174	5,147,340	5,431,008	7,779,258	7,633,782	7,924,734
2020-10-18	8,010,224	7,787,357	8,233,091	5,302,829	5,150,281	5,455,377	7,870,414	7,709,674	8,031,153
2020-10-19	8,051,093	7,807,353	8,294,833	5,317,328	5,154,939	5,479,718	7,951,887	7,775,540	8,128,234
2020-10-20	8,087,894	7,823,065	8,352,724	5,338,881	5,166,937	5,510,824	8,034,244	7,842,329	8,226,158
2020-10-21	8,115,847	7,829,330	8,402,364	5,368,906	5,186,465	5,551,348	8,128,955	7,920,994	8,336,915
2020-10-22	8,159,126	7,849,745	8,468,507	5,401,852	5,207,543	5,596,162	8,224,313	7,999,361	8,449,266
10-23	8,213,093	7,880,089	8,546,097	5,429,256	5,222,346	5,636,165	8,313,094	8,070,370	8,555,817
2020-	8,249,085	7,891,969	8,606,201	5,448,302	5,228,649	5,667,955	8,404,757	8,143,738	8,665,775

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

10-24									
2020-10-25	8,277,364	7,895,372	8,659,355	5,460,970	5,229,057	5,692,884	8,497,594	8,217,679	8,777,509
2020-10-26	8,317,198	7,909,787	8,724,610	5,474,158	5,230,655	5,717,661	8,580,638	8,281,530	8,879,745
2020-10-27	8,349,981	7,917,106	8,782,855	5,493,517	5,238,555	5,748,480	8,663,903	8,345,615	8,982,191
2020-10-28	8,372,083	7,913,247	8,830,918	5,520,058	5,252,924	5,787,193	8,760,647	8,422,770	9,098,523
2020-10-29	8,413,281	7,927,487	8,899,075	5,548,747	5,268,460	5,829,034	8,858,922	8,500,545	9,217,299
2020-10-30	8,465,235	7,951,904	8,978,566	5,572,888	5,278,802	5,866,974	8,948,441	8,568,843	9,328,040
2020-10-31	8,495,676	7,954,451	9,036,902	5,589,607	5,281,580	5,897,634	9,039,768	8,638,529	9,441,008
2020-11-01	8,520,277	7,950,441	9,090,113	5,600,853	5,279,269	5,922,436	9,133,866	8,710,494	9,557,238
2020-11-02	8,559,310	7,960,432	9,158,188	5,612,452	5,277,813	5,947,091	9,218,504	8,772,707	9,664,300
2020-11-03	8,587,372	7,959,526	9,215,218	5,629,387	5,281,746	5,977,028	9,302,358	8,834,123	9,770,593
2020-11-04	8,603,155	7,945,915	9,260,395	5,652,330	5,291,146	6,013,513	9,400,687	8,909,635	9,891,738
2020-11-05	8,642,559	7,955,035	9,330,084	5,676,956	5,301,477	6,052,434	9,501,573	8,986,811	10,016,33
2020-11-06	8,692,335	7,974,109	9,410,561	5,697,699	5,307,376	6,088,021	9,591,500	9,052,332	10,130,66
2020-11-07	8,716,372	7,967,176	9,465,568	5,711,919	5,306,630	6,117,209	9,681,893	9,117,985	10,245,80
2020-11-08	8,737,436	7,956,576	9,518,296	5,721,401	5,301,450	6,141,353	9,776,850	9,187,801	10,365,89
2020-11-09	8,775,979	7,963,115	9,588,843	5,731,137	5,296,894	6,165,379	9,863,081	9,248,589	10,477,57
2020-11-10	8,798,634	7,953,943	9,643,326	5,745,427	5,296,915	6,193,940	9,947,231	9,307,252	10,587,20
2020-11-11	8,807,655	7,930,750	9,684,559	5,764,735	5,301,533	6,227,936	10,046,620	9,380,793	10,712,44
2020-11-12	8,845,813	7,935,879	9,755,746	5,785,385	5,306,897	6,263,873	10,149,750	9,457,194	10,842,30
2020-11-13	8,893,169	7,949,937	9,836,402	5,802,649	5,308,419	6,296,880	10,239,780	9,519,815	10,959,74
2020-11-14	8,909,939	7,933,203	9,886,675	5,814,205	5,304,131	6,324,278	10,328,612	9,580,974	11,076,24
2020-11-15	8,927,758	7,916,828	9,938,689	5,821,640	5,295,963	6,347,318	10,424,010	9,648,382	11,199,63