2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)
October 11-14, 2020. Toronto, Canada

A Comparative Study of Predictive Machine Learning Algorithms for COVID-19 Trends and Analysis

Ajinkya Kunjir
Department of Computer Science
Lakehead University
Thunder Bay, Canada
akunjir@lakeheadu.ca

Dishant Joshi
Department of Computer Science
Lakehead University
Thunder Bay, Canada
djoshi2@lakeheadu.ca

Ritika Chadha
Department of Computer Science
Lakehead University
Thunder Bay, Canada
rchadha1@lakeheadu.ca

Tejas Wadiwala
Department of Computer Science
Lakehead University
Thunder Bay, Canada
twadiwal@lakeheadu.ca

Vikas Trikha
Department of Computer Science
Lakehead University
Thunder Bay, Canada
vtrikha@lakeheadu.ca

***Abstract*- This paper attempts to conduct analysis on the WHO dataset to produce predictive analysis applying different machine learning regression approaches such as decision trees, LSTM, and CNN regressor. The primary data has 91 entries, which consists of data of various countries with respect to dates along with confirmed cases, confirmed deaths, and recovered cases. The dataset has been divided into 70:30 in which 70 percent is used for training and validation, and 30 percent is used for testing. The coronavirus disease outbreak started in 2019, arising in Wuhan, China. The key objective is to exercise different artificial intelligence approaches, we ought to predict the confirmed cases, confirmed deaths, and recovered cases, and further, various visualization techniques have been used to deduce the meaningful inferences from the model's prediction and perform specific analytics on the results concluded. The prediction models such as LSTM and CNN are evaluated on the basis of several loss functions such as R2 score and Mean Squared Error.**

Keywords - Coronavirus Disease 2019 (COVID-19), Convolutional Neural Network (CNN), Long-Short Term Memory (LSTM), data visualization, Loss function

I. INTRODUCTION

The explosion of COVID-19 befell in December 2019 dawning from Wuhan, China, and spread throughout at a very rapid rate. In a tiny fraction of time, COVID-19 infected almost every country and caused thousands of deaths, which initiated the WHO to declare this a worldwide pandemic on 11 March 2020. As a matter of time, the spread just increased 13 folds in January 2020 in China, and the number of affected countries got tripled. Since then, a lot of mathematical methods have been used to measure the impact of this pandemic worldwide to check the numbers of infected, recovered, and total cases that happened so far. A much alike novel coronavirus has emerged back in 2002, which was coined as SARS-CoV, and that spread to around 37 countries. SARS-CoV caused more than 8000 infections and 800 deaths, whereas, in the case of COVID-19, the numbers are much worse, more than 2.95 million people got infected, out of which 861K got recovered, and 205K died of this deadly virus. The pandemic can be directly related to sustainable development, which covers social, economic, and environmental goals inclusive of political aims. The assumption of sustainability looks achievable, but one of the biggest hindrances which were observed in the COVID-19 outbreak was the anticipation of the epidemic and the emergency problems leading to the disease. It does not just lag down the economy but additionally produces social issues resulting in damage to the fundamental pillar of sustainable development. Numerous data scientists developed a bunch of statistical techniques to deduce the meaningful inferences striving to reduce the impact of COVID-19, one of which was to introduce the lockdown of the overall country until it achieves some flat curve or any reduction. These measures yet proved to be productive, however, caused social and economic disruption globally. Therefore, it is imperative to examine the consequences of pandemic produced, and considering everything, it is indeed required to assess the pre-arrival of an epidemic. We tried to construct a meaningful machine learning model that is capable of performing predictions when fed with the dataset provided by the WHO to check the impact of the same and tried to perform basic analytic and visual operations to seek more a clear picture of the pandemic.

II. LITERATURE REVIEW

In this section, we discuss about the other simultaneous researches related to the COVID-19 global outbreak and its scientific applications. Scientists and practitioners all over the world have pursued experimentation's on the global real time COVID19 data and have discovered valuable insights, patterns and knowledge from the data. From the paper published in Journal of thoriac disease (2020), the authors showed how various government policies to control COVID-19 affected the

© IEEE 2020. This article is free to access and download, along with rights for full text and data mining, re-use and analysis

Authorized licensed use limited to: IEEE Xplore. Downloaded on January 06,2023 at 10:26:01 UTC from IEEE Xplore. Restrictions apply.

numbers. The authors have used the most updated COVID-19 data with an amalgamation of the number of people migrating before and after January 23rd. The LSTM model was used to process time-series data to predict infections over time. It used the SEIR model for deriving the epidemic curve, where the authors modified the model by adding move in and move out parameters to it. SEIR models the flow of data between four states [1]: Susceptible(S), Exposed (E), Infected (I), and Resistant(R). The paper used data of Hubei, Guangdong, and Zhejiang provinces, as they had the highest number of cases in Mainland China. The results predicted that the epidemic would reach its peak by the end of February and would gradually decrease by the end of April. The authors used various migration index to predict if the migration would have affected the numbers. The results predicted that if the government would have delayed the lockdown by even 5 days it would have increased the size of the epidemic by three times.

The main objective of the following paper is to predict the use of health services and deaths of every state in the United States of America over the next four months. The dataset was extracted from local and national government sites and the WHO website. The modeling and approach was a four-step process: (i) identification and processing of COVID-19 data; (ii) statistical model estimation for population death rates as a function of time since the death rate exceeds a threshold in a location; (iii) predicting time to exceed a given population death threshold in states early in the pandemic; and (iv) modeling health service utilization as a function of deaths. In the end, the results predicted a rise in demand for the health services in the last weeks of March and the early weeks of April and slowly declining with a steady demand until June. The mean forecast of deaths would exceed 2300 by the second week of April. States like New York which had an early sign of epidemic would see a sharp increase in deaths compared to other states like Washington and California which would have a relatively slower death rate [2].

The authors in their research article published in the MDPI journal of sustainability highlights the points on how the corona virus affected sustainable development process. The authors decided on taking Hubie province data for the proposed model as it had the highest number of confirmed COVID-19 cases. The prediction was done for 30 days in future using a binary classification model, mostly GMDH (Group Method of Data Handling) type of neural network inheriting artificial intelligence methods. Extending the experimentation, the authors also did a regressoion analysis on the trends of confirmed cases and compared it with the variations of daily parameters such as wind, humidity and average temperature. The regression analysis was also executed on other provinces of China incluign Hubei such as Guandong, Henan, Zhejiang and Hunan against the daily fluctuating parameters mentioned before. The case study using 20 GMDH methods on total confirmed cases achieved a testing accuracy ranging from 71 % to 85 %. The training accuracy obtained was much higher and ranged from 73 % to 95 %. The methods were ranked according to their performance and the best one amongst them was taken into consideration for further computation. The study was concluded with a statement saying that there would have been a tremendous spike in confirmed cases due to corona virus if the lockdown would have been pushed 5 days farther [3].

Joseph TWu, Kothy Leung and Gabriel M Leung in their research paper 'Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study' mentioned the impact and spread of covid-19 from its epicenter 'Wuhan' to other cities in China. The data used for modelling study is gathered from 31 Dec 2019 to Jan 28 2020 on the basis of total number of cases exported from Wuhan exponentially. The authors explained the reason of using susceptible-exposed-infectious-recovered model for epidemics across all major cities in China. The basic reproductive number estimated using markov chain Monte Carlo model and 95% credible interval (CrI). In the later part of the paper, transmissibility and efficiency is explained in the later sections of the paper. The authors have displayed a graphical comparison between SARS and n-COV19 in Wuhan using bar graphs and table plots [4].

III. DATASET AND PREPROCESSING

The raw dataset for this meticulous project has been extracted from the WHO in which information has been classified into three different spreadsheets. Each spreadsheet consisted of the number of cases along with dates that are organized concerning their countries. Furthermore, these spreadsheets contain the values of global confirmed, recovered, and deaths occurred worldwide. The designated datasets were stripped down to distinct countries, which in our case are India, Canada, and China. In the first stage of data pre-processing, the dates along with the data were transposed and separated using individual CSV files and then embedded directly into the predictive machine learning model for analyzing and solving this regression problem to check the trend observed in COVID-19. After transforming the global data into single columnar time-series data of 91 dates ranging from 22nd January 2020 to 22nd April for confirmed, recovered, and death stats in different CSV files, the exclusive data files were later clubbed to represent a single file for a specific country within the code.

IV. COMPARATIVE STUDY OF ALGORITHMS

For solving this particular time series regression problem, we have taken into consideration the following three algorithms:

- LSTM
- CNN
- Decision Tree

These algorithms would be explained in the section below.

Authorized licensed use limited to: IEEE Xplore. Downloaded on January 06,2023 at 10:26:01 UTC from IEEE Xplore. Restrictions apply.

A. Long Short Term Memory (LSTM)

For handling sequence dependencies, a powerful type of neural network is designed and is called as recurrent neural network (RNN). LSTM is a type of RNN which can be used to solve time series prediction problems. It overcomes the vanishing gradient problem and is trained using backpropagation through time. It is used to address complicated sequence problems in machine learning. LSTM has memory blocks that are connected through layers. A block consists of gates that manage the block's state and output. Each gate within a block operates on input sequence uses a sigmoid activation unit to control whether they are trigerred or not making the change of state and addition of information flowing through the block conditional. LSTM consists of the following three types of gates within a unit:

- Forget Gate: It is used to decide what information to keep and what to throw away from the block.
- Input Gate: It is using a condition to decide which values from the input to update the memory.
- Output Gate: It is using a condition to decide what to output based on input and the memory of the block.

Each unit is like a machine where the gates have weights that are learned during the training procedure [5].

B. Convolutional Neural Network (CNN)

CNNs were traditionally designed for two-dimensional data, but they can be used to model univariate time series forecasting problems. Univariate time series are dataset is composed of single series of observations in which there is a temporal ordering, and the model created is required to learn and understand from the past observation's series to predict the next value in the sequence.

Preparing a univariate series is necessary before it can be modelled. For preparing it we have to divide the sequence of our data into multiple input/output patterns called samples. In our case, we are using three time steps as input and one time step is used as output for one step prediction that is being learned. Next step is developing a CNN model, a one-dimensional CNN is a CNN model that has a convolutional hidden layer that operates over a 1D sequence. This is followed by another convolutional layer if we have long input sequences, and then followed by a pooling layer whose main feature is to extract the most significant feature from the data. It is then followed by dense fully connected layer which is used to interpret the features extracted from the convolutional part of the model. A flatten layer is then used between the convolutional layers and the dense layer to reduce the feature maps to a single one-dimensional vector. We can add or subtract the layers to see what kind of effect it have on the training data using some performance metrics like R squares. The testing data is then passed through this model to predict the next number in the sequence [6].

C. Decision Tree

Decision tree falls under the category of supervised machine learning algorithm. It can be used for both continuous as well as categorical output variables. It uses a decision-making flowchart like tree structure to make decisions. The branches/edges usually represent the result of the node and the nodes have either:

- Conditions (Decision Nodes)
- Results (End Nodes)

Decision tree regression observes features of an object and trains a model in the structure of a tree to predict data in the future to produce meaningful continuous output. Output that is not discrete is known as continuous output, i.e., it is not represented just a discrete set of numbers or values [7].

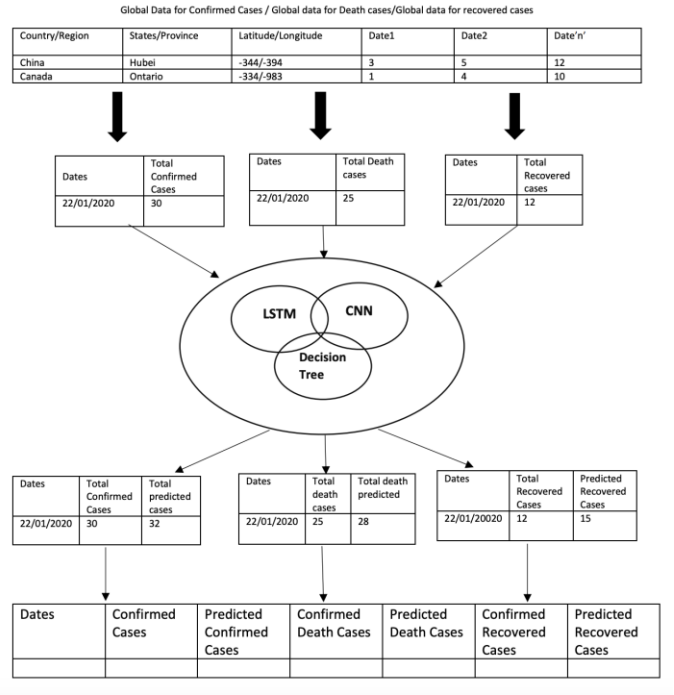

Figure 1: Preprocessing the Dataset

V. EXPERIMENTAL ANALYSIS

In this research, we have used Deep Convolutional Neural Network, Recurrent Neural Network and Decision tree Regressor from Python's sklearn library. The countries selected for 'International COVID19 Spread Analysis' are China and India as China being the most populated country in this world, has the highest number of confirmed COVID19 cases and India ranking second on the charts has less than expected cases. ''Facts''. All the confirmed cases, confirmed deaths, confirmed recovered cases have been analysed, compared and ran through the prediction methods for each Country and State involved in this matrix.

Authorized licensed use limited to: IEEE Xplore. Downloaded on January 06,2023 at 10:26:01 UTC from IEEE Xplore. Restrictions apply.

The International COVID19 Spread comparison and analysis for China using the above mentioned algorithms have been shown below:

- Prediction and Analysis for China using CNN Regressor:

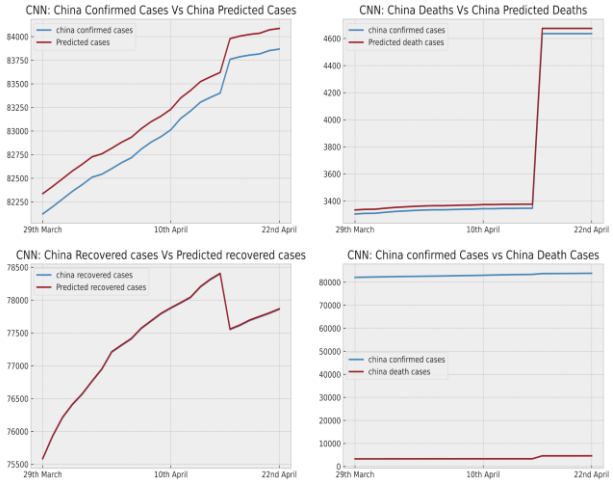

Figure 2: Prediction and Analysis for China using CNN Regressor

The data selected for China COVID19 Analysis starts from 22nd January to 22nd April 2020. The comparison shown in the figure above is from the test data. Making Observations based on CNN Regressor Analysis:

1. The predictions for the 25 days period in unknown future is slightly higher than the current values. The R2 score achieved for actual and predicted values for total confirmed cases in China is
2. The death predictions for 25 days in unknown future ranges from high- low-high as shown in the graph above. The R2 Score achieved for actual and predicted values for death cases in China is
3. The recovered cases predictions for 25 days in unknown future lies in the same line with minor fluctuations between the values. The R2 Score achieved for this category is

- Prediction and Analysis for China using Sklearn's Decision Tree:

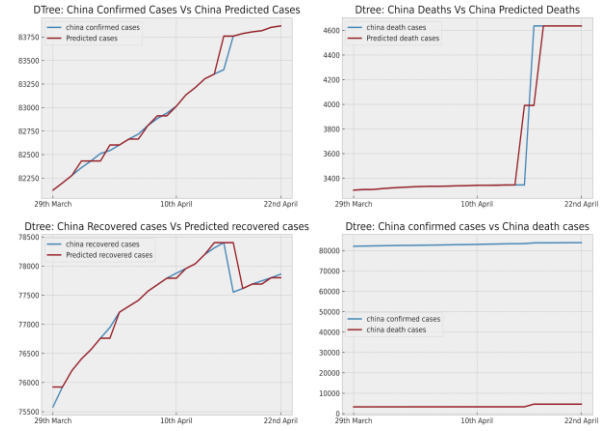

Figure 3: Prediction and Analysis for China using Sklearn's Decision Tree

Making Observations based on Decision Tree Analysis:

1. Confirmed Cases Predictions – Prediction values spikes fluctuate before 10th April and meets the original values in the 22nd April Zone.
2. Death Cases Predictions – The death predictions spike high and low after 10th April.
3. Recovered Cases Predictions – The recovered cases spike up in between 16th – 19th April.
4. The fourth subplot shows the difference between confirmed cases and confirmed deaths in china from 29th March to 22nd April.

- Prediction and Analysis for China using LSTM (Long Short Term Memory):

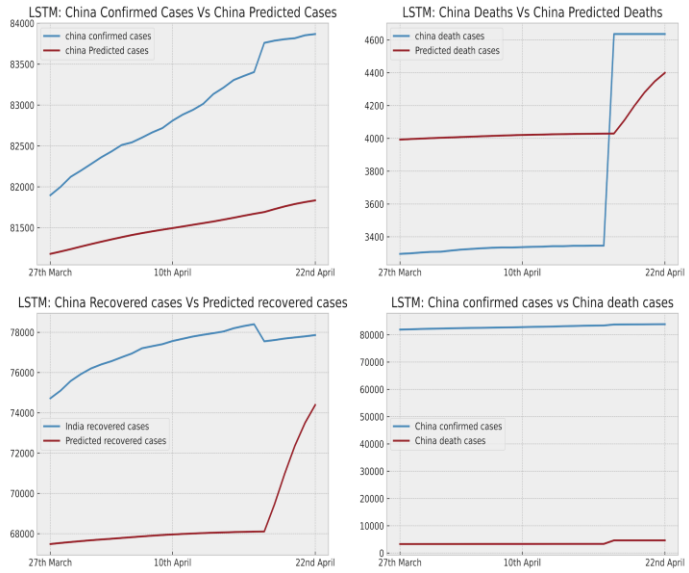

Figure 4: Prediction and Analysis for China using LSTM (Long Short Term Memory)

Making Observations based on Long Short Term Memory Analysis:

1. Confirmed Cases Predictions – Prediction values spikes lower than confirmed cases from 27th march to 22nd April.
2. Death Cases Predictions – The death predictions spike high from the very first day of testing data.
3. Recovered Cases Predictions – The recovered cases prediction descends throughout the graph if compared with confirmed recovered cases.
4. The fourth subplot shows the difference between confirmed cases and confirmed deaths in china from 27th March to 22nd April.

International COVID19 Spread comparison and analysis for India

- Prediction and Analysis for India using CNN Regressor:

Authorized licensed use limited to: IEEE Xplore. Downloaded on January 06,2023 at 10:26:01 UTC from IEEE Xplore. Restrictions apply.

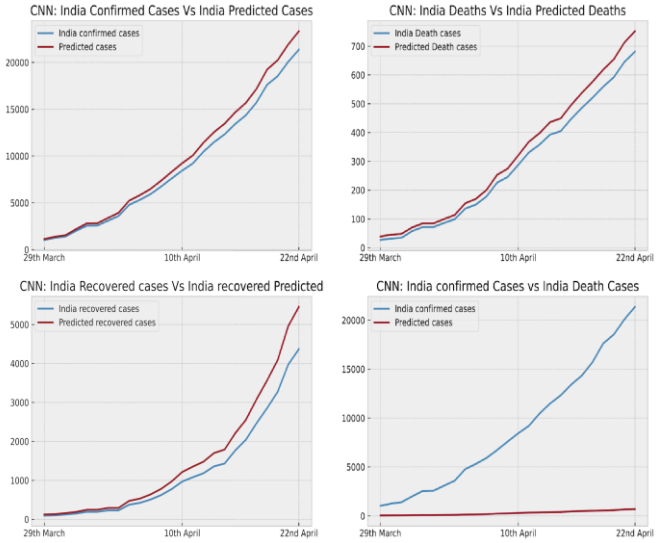

Figure 5: Prediction and Analysis for India using CNN Regressor

COVID19 Spread comparison and analysis for India using the above mentioned algorithms have been shown below:

Making Observations based on CNN Regressor Analysis:

1. Confirmed Cases Predictions – Prediction values spikes rise before 29th March and meets the original values in the 22nd April Zone.
2. Death Cases Predictions – The death predictions spike high after 10th April.
3. Recovered Cases Predictions – The recovered cases spike up in between 10th – 22nd April.
4. The fourth subplot shows the difference between confirmed cases and confirmed deaths in India from 29th March to 22nd April.

- Prediction and Analysis for India Using Decision Tree:

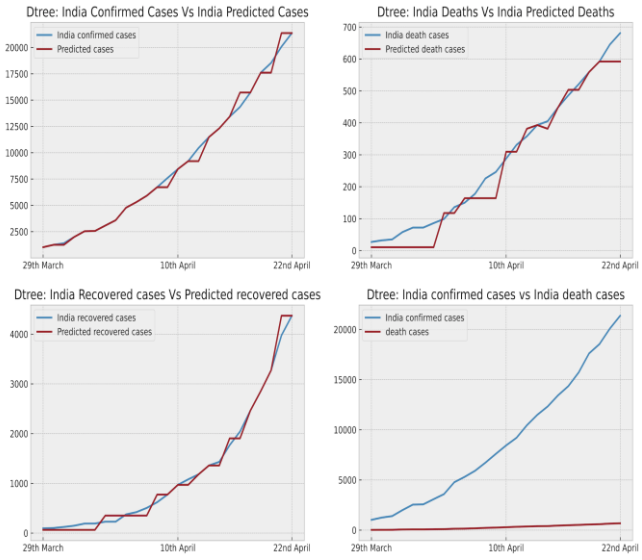

Figure 6: Prediction and Analysis for India Using Decision Tree

Making Observations based on Decision Tree Analysis:

1. Confirmed Cases Predictions – Prediction values spikes rise up and down from 29th March to end of the data.
2. Death Cases Predictions – The death predictions starts with a low number and drops after 20th April.
3. Recovered Cases Predictions – The recovered cases rise up after 20th April.
4. The fourth subplot shows the difference between confirmed cases and confirmed deaths in India from 29th March to 22nd April.

- Prediction and Analysis for India using LSTM (Long short Term Memory):

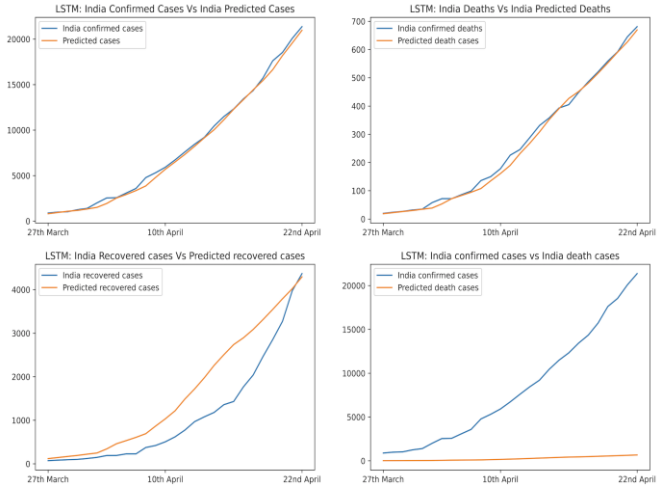

Figure 7: Prediction and Analysis for India using LSTM (Long short Term Memory)

The subplots grid for Comparing confirmed cases and predictions for India's Confirmed cases, death cases, recovered cases has been shown below the observation description.

Making Observations based on LSTM (Long Short Term Memory):

1. Confirmed Cases Predictions – Prediction values spikes rise up and down from 29th March to end of the data.
2. Death Cases Predictions – The death predictions starts with a low number and drops after 20th April.
3. Recovered Cases Predictions – The recovered cases rise up after 20th April.
4. The fourth subplot shows the difference between confirmed cases and confirmed deaths in India from 29th March to 22nd April.

VI. RESULTS & EVALUATION

Moving on to the next phase, the developed model was tested based on r2 score, which is imported using sklearn library's metrics package. The equation of r2 score is shown below:

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

where
r = The correlation coefficient

Authorized licensed use limited to: IEEE Xplore. Downloaded on January 06,2023 at 10:26:01 UTC from IEEE Xplore. Restrictions apply.

n = number in the given dataset
x = first variable in the context
y = second variable

The Table II shows the r2 score observations made when the model was executed on China's dataset:

Algorithm used and R2 Score for China's dataset

Algorithm	R2 Score
CNN	0.9998
LSTM	0.9949
Decision Tree	-0.8212

Table II clearly shows that the performance of the model is the best when the algorithm used is CNN, and it's r2 score is 0.9998.

The Table III shows the r2 score observations made when the model was executed on Ontario's dataset:

Algorithm used and R2 Score for Ontario's dataset

Algorithm	R2 Score
CNN	0.9904
LSTM	0.7113
Decision Tree	0.9477

While performing the analytics, an exceptional downfall was observed in confirmed cases after the lockdown was announced on 25th May in Ontario, Canada. We have calculated the percentage rise in confirmed cases from 15th March to 24th March (i.e., before lockdown) & from 25th March to 3rd April (i.e., after lockdown). The percentage rise in confirmed cases before the lockdown was 82 percent, and it dipped to 78 percent after the government took the essential step. Assuming the statistics to be real, if the lockdown was announced on 15th March instead of 25th March, it can be estimated that a total of 484 cases could have been prevented. There would have been a dip of 4 percent in the rise of confirmed cases.While exercising the same analytics operation on the predicted confirmed cases model, the dip was found to be 2 percent which is very close to the real model. Through this analysis, it can be concluded that if this model is used for any imminent pandemic, it can give results which can be used to save lives.

The same statistical analysis have been performed even on the country level. We have selected India for this, and it has been found out that 92 cases would have been prevented if the lockdown would have been initiated on 15th March rather than 25th March.

VII. CONCLUSION

In this empirical research, we attempted to discover knowledge, insights, and patterns from the global COVID-19 real-time data. After the successful integration of the data with our algorithms, the analytics were visualised using Python libraries such as matplotlib, seaborn, plotly, bokeh, and altair. The diagrammatic representations such as line graphs and bar plots explain the trend detected in the country/state-level data. For ongoing COVID-19 pandemic issue, it can be concluded that these prediction models can be helpful in providing a concrete idea and useful statistics to support our real-time model. Furthermore, the same model can be utilized to determine the pre-arrival and steps to be taken for reducing the impact of any imminent epidemic or pandemic. After performing rigorous training and analysis on the data, the predicted and the real-time results were much alike, and a mutual trend was observed between the two, which led to the above-mentioned conclusive inference. This model contains a thousand capacities in the field of machine learning artificial intelligence and holds the potential for improvements that can be introduced down in the future. The scope of this research can be expanded by adding more country and state-level data for increasing the efficiency of this comparative study. The algorithms used for this study can be evaluated on more than one performance metric which can be mean square error (MSE), variance, etc. Other impressive regressors such as Random Forest, XGBoost, and ARIMA can be added to extend our comparative study.

REFERENCES

[1] Z. Yang, Z. Zeng, K. Wang, S.-S. Wong, W. Liang, M. Zanin, P. Liu, X. Cao, Z. Gao, Z. Mai, et al., "Modified seir and ai prediction of the epidemics trend of covid-19 in china under public health interventions," *Journal of Thoracic Disease*, vol. 12, no. 3, p. 165, 2020.

[2] I. COVID, C. J. Murray, et al., "Forecasting covid-19 impact on hospital bed-days, icu-days, ventilator-days and deaths by us state in the next 4 months," *medRxiv*, 2020.

[3] B. Pirouz, S. Shaffiee Haghshenas, S. Shaffiee Haghshenas, and P. Piro, "Investigating a serious challenge in the sustainable development process: analysis of confirmed cases of covid-19 (new type of coronavirus) through a binary classification using artificial intelligence and regression analysis," *Sustainability*, vol. 12, no. 6, p. 2427, 2020.

[4] J. T. Wu, K. Leung, and G. M. Leung, "Nowcasting and forecasting the potential domestic and international spread of the 2019-ncov outbreak originating in wuhan, china: a modelling study," *The Lancet*, vol. 395, no. 10225, pp. 689–697, 2020.

[5] J. Brownlee, "Time series prediction with lstm recurrent neural networks in python with keras*," Available at: machinelearningmastery.com*, p. 18, 2016.

[6] J. Brownlee, "How to develop convolutional neural network models for time series forecasting*," Available at: machinelearningmastery.com*, 2018.

[7] A. Das, "Python — decision tree regression using sklearn," *Available at:geeksforgeeks.org*.

Authorized licensed use limited to: IEEE Xplore. Downloaded on January 06,2023 at 10:26:01 UTC from IEEE Xplore. Restrictions apply.