HIGH SCHOOL EDITION
Journal of Student Research
Volume 11 Issue 3 (2022)

Predicting COVID Cases: Effectiveness of the Vaccine

Ritesh Kasamsetty[1], Alex Hellman[1], Derek Xu1 and Ryan Solgi[#]

[1] James Logan High School, Union City, CA, USA
[#]Advisor

ABSTRACT

The COVID-19 pandemic has been one of the most devastating events in recent history, resulting in millions of deaths and destruction on the global economy. While vaccines have been developed to slow the spread of the virus and achieve global herd immunity, millions of US citizens refuse to take them due to speculation regarding their safety and effectiveness. Our project goal is to reassure people that COVID vaccines are effective at controlling the spread of the virus and reducing the number of people infected. To demonstrate this, we use an autoregressive (AR) model and long short-term memory (LSTM) network to represent the spread of COVID over time. Using data from various US states, we display COVID trends over the last year and make predictions on how the disease will spread in the future (beyond the scope of our data set) with and without vaccines. In the end, our predictions show that vaccines are effective at reducing cases and slowing the spread of the disease. By comparing results from both models for each state, we were able to choose the more accurate model and use it for our graphs and predictions. After comparing sources of error in our models (root mean square error and coefficient of determination), our results indicated that the LSTM neural network was much more accurate than the autoregressive model.

Introduction

COVID was first identified in December 2019. With an origin in Wuhan (a city in the Hubei province of China), this respiratory illness primarily spread through infected people coughing or sneezing near others. As people interacted with infected individuals and engaged in travel and commerce, COVID quickly made its way to nearly every country worldwide. While each nation implemented different precautions to halt the spread of this disease (such as requiring the use of face masks in public, issuing total lockdown, and creating other social distancing practices), COVID has killed 4 million and infected over 150 million people worldwide since its initial outbreak [1]. Luckily, vaccines have been created and approved since the beginning of 2021 to combat the virus. Currently, 3.7 billion total doses have been administered, with three different types (Pfizer, Johnson & Johnson, and Moderna), approved for distribution in the US by the Centers for Disease Control and Prevention (CDC) [2]. Even though Pfizer and Moderna require two vaccine doses to be fully protective, the US has made great progress toward reaching herd immunity. As of now, 55 percent of the US population has received one dose of the vaccine, while 48 percent of people are fully vaccinated [2].

While the vaccines have proved to be very effective in preventing people from contracting the virus, many people are hesitant to receive or don't have access to them. For example, a lack of transportation and the uneven distribution of doses to specific cities makes it challenging for certain communities to be treated. Some also believe vaccines aren't safe and that potential side effects outweigh the benefits of being protected. Many people are wary of the creators of the vaccines and the healthcare system as a whole, while misinformation from politics and media further discourages people from trusting them [3][4]. In addition, many people hate government intervention and view receiving the vaccine as a loss of their rights and personal freedom [5]. This

problem is continuing to get worse, as people continue to downplay the significance of this disease, refuse to believe that millions of people have died from the virus, and still believe vaccines are ineffective. Our goal is to show the safety of herd immunity by demonstrating how quickly the virus can spread without the aid of vaccines, which will hopefully dissuade more people from adopting these negative mindsets.

In 2020, it was crucial to have the resources to predict the number of cases in each country, specific states, and individual cities. These predictions made it possible for governments and countries to determine which communities needed more financial assistance and resources, making it easier to decide where the development of new hospitals and treating of patients would be most effective. While many types of research have been conducted in the field of time series forecasting, not as much information has been collected on forecasting COVID cases. The authors of [6] were able to predict the number of confirmed cases, recovered cases, and deaths in different countries and compare them. They were also able to predict the trend of the disease in different Australian regions. In other COVID research fields, people found ways to predict the spread of coronavirus to different countries and places inside a country. For instance, in source [7], four phenomenological models, as well as a SIR (susceptible, infected, recovered) model were used to predict COVID dynamics using only 223 pieces of data. A computational method called the Particle Swarm Optimization (PSO) algorithm was used to estimate the parameters. The results indicate that the Generalized Richards Model was precise enough with a low margin of error compared to the other analytical methods. It was also able to fit the data much better. In other COVID research fields, people found ways to predict the spread of coronavirus to different countries and within them. Even though the dataset was small and very limited, these models were very effective in predicting the epidemiology and spread in many countries.

These machine learning models were initially used for predicting COVID cases in the future because they provided countries with the knowledge on where to distribute financial aid most effectively to lower COVID infection rates. While many believe that these models are no longer relevant today (as vaccines have already been administered to billions of people and cases are much lower), we disagree. Models and data can still be shown to people who doubt the effectiveness of the vaccine, as the number of new infections has dropped significantly since the beginning of vaccine distribution.

Unfortunately, the spread of the new delta variant and low vaccination rates are the main contributing factors. Out of the 24 US states experiencing an increase in COVID cases, 21 have less than half their population fully vaccinated [8]. Our project goal is to determine the effectiveness of COVID vaccines in reducing the total number of positive cases throughout various states in the US. We compare the progression of COVID with and without vaccines in our models to convince readers that the vaccine is very effective and everyone eligible should sign up to receive it.

Methods / Methodology

Data preprocessing

In the beginning, we reviewed data from various sources that included information on the number of COVID cases in different states over time. After comparing articles to see which one included the most useful COVID case statistics, we decided to use data from the COVID Tracking Project at The Atlantic. This organization compiles COVID data by gathering online information published by public health authorities of different states. The data includes the number of COVID cases in each state every day from March 4, 2020, to March 7, 2021. To use this information for our model predictions, we first downloaded data from each state as CSV files. We cleaned this data by removing all columns that contained unnecessary data (including repetitive information on COVID cases and irrelevant statistics in each state). Afterward, we were left with three columns: date, state name abbreviation, and COVID cases for each date. The data was then organized by least to most recent, as it allowed the algorithm to properly process the data. Next, Nan values were replaced with the value 0. Nan values

are unrepresented data points. They were present in some of the data for March 2020, as they represented no recorded cases. Finally, a MinMaxScaler was used to scale the inputs of our testing and training sets to more easily visualize our predictions as well as achieve more accurate results.

Model Architecture

The first model we use is an autoregressive model (AR). An AR model is a time series model that can predict future values using past values (lags). It is a forecasting tool used when there is a correlation between data in a time series model and values that precede and succeed them [9]. In our model, the previous inputs are the total number of COVID cases for each day from the last 14 months. To make predictions for the future, our model uses a lag of 2. Hence, it takes the two most recent pieces of data from our data set to make a prediction one day ahead. Afterward, it uses the predicted value and the second most recent value to make another prediction two days ahead. This process continues until the model predicts a predetermined amount of time in the future. The equation used to create an AR model is shown in figure 1 below [10].

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2}$$

Figure 1. Equation used by an autoregressive model

- Y_t is the output of the equation
- α is the intercept estimated by the model
- β_1 is the coefficient of the first lag estimated by the model
- Y_{t-1} is the first lag of the series
- β_2 is the coefficient of the second lag estimated by the model
- Y_{t-2} is the second lag of the series

To use lags, we first make an array which we call the Generator array. The array is initialized with the last two lags, or inputs, that were fed into the model. Let's call these two values Beta-1 and Beta-2, respectively. The first thing we do in our generator function is to create a prediction using the two lags inside the generator array. We add this value to our Predictions array which stores the predictions for creating our model. After that, we shift Beta-2 to the first index of the array and move our prediction to the Beta-2 index. The Beta-1 value is now Beta-2 and the Beta-2 value is now the generated prediction. We repeat this process for the length of our testing set. In figure 2, we do this process three times.

Beta-1	Beta-2	Prediction
6	7	8
7	8	9
8	9	10

Figure 2. Generator Array. The values are shifted to the left every time a new prediction is made.

The second model we use is a long short-term memory (LSTM) network. LSTM networks are an extension of recurrent neural networks (RNN). RNNs are algorithms that can remember previous inputs due to having an internal memory. The hidden layer of an RNN contains a loop that enables the algorithm to feed its output into itself, in addition to producing an output [11]. To make a decision, it considers the current input and

the output that it previously fed itself [12]. Hence, they are effective for dealing with machine learning problems that involve sequential data (including time series), which use previous data to make predictions. LSTMs are similar, yet they are better at remembering inputs over long periods. As a result, they are used more frequently than RNNs when dealing with large data sets [13].

In our LSTM network since we used a batch size of one, every time a data sample goes through the algorithm, the model's parameters are updated. We also used twenty epochs which means the learning algorithm worked through the entire training dataset twenty times. To build our LSTM Neural Network, we used a Sequential Model. Our first layer, or input layer, was an LSTM layer that contained 100 neurons. It used a ReLU activation function to introduce non-linearity. Our hidden layer contained one neuron for its hidden loop. The last layer is the output layer which contains one neuron as well because we will be getting one output which is the number of COVID cases for the next day.

Evaluation

After importing our data and writing the code, we represent our findings with graphs. We have two AR models, which each consists of an x and y-axis. The x-axis represents a range of dates, while the y-axis displays the number of new daily COVID cases. Our first AR graph consists of 3 different lines. One line shows the current number of new daily COVID cases from March 2020 to January 2021, and another displays the predicted number of COVID cases from March 2021 to April 2021 (while we are past this point in time, we are making predictions since our data entries end in early March). Our final line predicts the number of COVID cases assuming no vaccines were ever administered. This graph predicts cases from early January (right before the first vaccines were administered) until April. Our second graph is simpler and very similar to our first. There are two lines, each with a range from early January to March. One line shows daily COVID cases, and the other predicts cases assuming the vaccine wasn't distributed.

We used two commonly applied values in data analysis to determine the accuracy of our models called the root mean square error, or RMSE, and coefficient of determination or R2. RMSE values measure the standard deviation of residuals, which are differences between real data points and a regression line measured on a vertical scale. Therefore, they provide a metric for how concentrated data points are around the line of best fit. It takes the sum of the squared differences between actual and predicted values, divides the result by the total number of observations, and takes the square root of this value.

Similarly, R2 values measure how well a regression model fits within a data set. A value of 1 indicates a perfect fit, while anything between 0 and 1 implies the model is less accurate. On the other hand, values greater than 1 indicate an abnormal case or that the data set used was too small. The residual sum of squares is calculated by the summation of squares a perpendicular distance between data points and the regression line. This value is then divided by the total summation of squares, which sums the squares between data points and the average line for data values. After the quotient is taken, this value is subtracted from 1 to obtain the R2 value.

There are a couple of research limitations that we may encounter while we create our models. One factor that affects the accuracy of our data is that most states experience COVID spikes in fall 2020, which immediately precedes when the first set of vaccines is administered. Since most parts of our graphs start during this time (early January), this data significantly affects the accuracy of our model by overestimating the number of COVID cases. In addition, our dataset only runs until early March. While we have 14 months of COVID case statistics from each state, there is no recent data (from the past few months) to help us create our models. As a result, our predictions in our models technically never extend past the present date. Rather, they model COVID cases after our data set ends (early March 2021) and beyond.

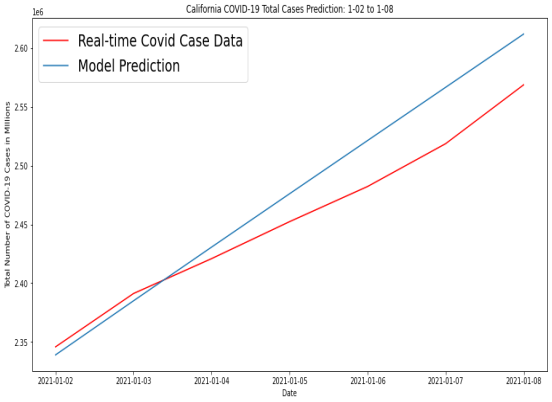
Results / Analysis

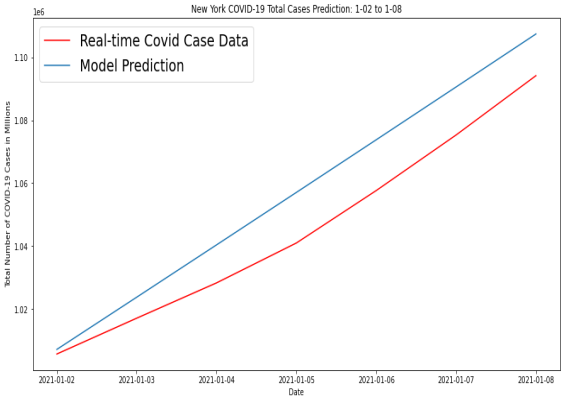
Autoregressive (AR) Model Predictions

We first predicted COVID cases with an AR model for the states of California, New York, Washington, Kansas, and Connecticut. Our testing data ran for 10 months (March 2020 to January 2021), while our testing data was week-long (January 2nd to January 8th). Table 1 displays the RMSE and R2 values that were recorded for each state from the AR model. Figures 2 and 3 are graphs of California and New York. They are examples of the AR Model Prediction of one week before the vaccine was administered. The blue line represents the model prediction while the red line shows the real total number of COVID cases.

Table 1. RMSE and R2 Values of our AR Model Predictions (Before Vaccine)

	California	**New York**	**Washington**	**Kansas**	**Connecticut**
RMSE	30311.5033	12630.3163	3178.8517	5891.6085	2576.2446
R2	0.8153	0.8168	0.7327	-0.3837	0.7889





The predictions of COVID cases without vaccinations and real-time case data for California, Connecticut, and New York are shown in figures 5-7, respectively. This is because they had the highest R2 scores as shown in table 1. Predictions showed that in each state, COVID cases will increase at an approximately constant rate, while the actual data showed that the number of COVID cases began to plateau when vaccinations started being administered. The blue line is the model prediction if there were no vaccines and the red line is the real number of total COVID cases.

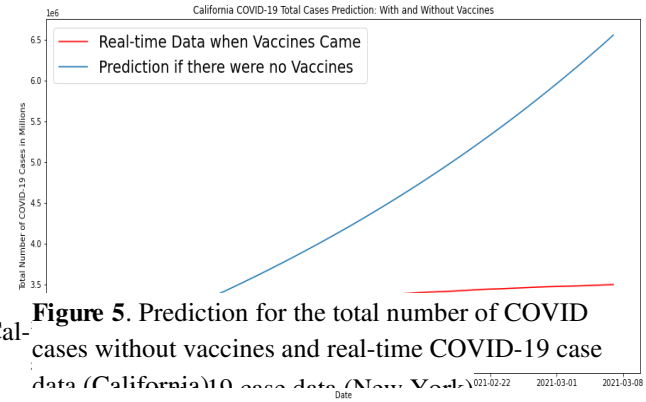
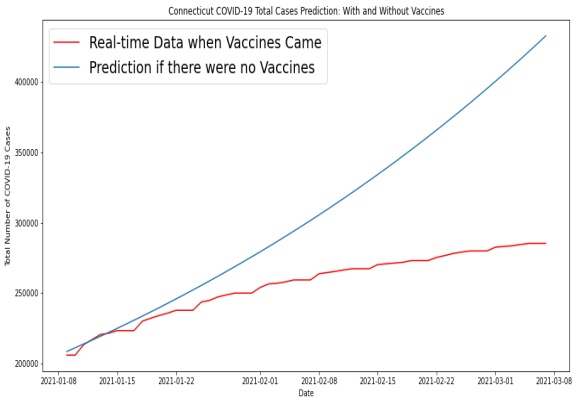
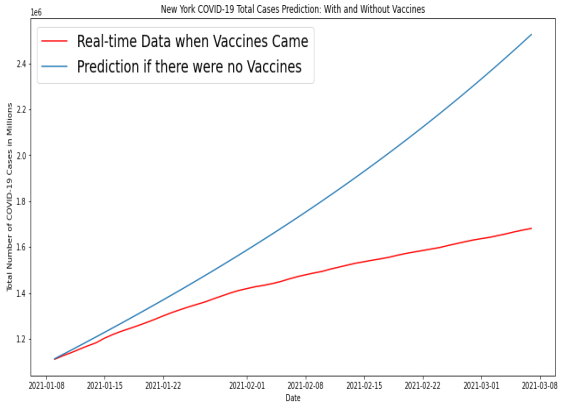


Figure 3. AR Model Prediction vs Actual Data fr... Ja... Cal-

Figure. 6. Prediction for the total number of COVID cases without vaccines and real-time COVID-19 case data (Connecticut)

Figure 5. Prediction for the total number of COVID cases without vaccines and real-time COVID-19 case data (California)19 case data (New York)





Long Short Term Memory (LSTM) Recurrent Neural Network Predictions

After predicting with an AR model, we decided to use an LSTM network. We predicted the number of COVID cases for California, New York, Washington, Kansas, and Connecticut. Our testing data ran from March 2020 to January 2021. Similar to the AR model, our testing data was week-long and lasted from January 2nd to January 8th. Table 2 displays the RMSE and R2 values that were recorded for each state from the LSTM. Figures 7 and 8 of Washington and Kansas are examples of the LSTM network one week before the vaccine was administered. The orange line is the model prediction and the blue line is the real number of total COVID cases.

Table 2. RMSE and R2 Values of our LSTM Network Predictions (Before Vaccine)

	California	New York	Washington	Kansas	Connecticut
RMSE	34688.1765	5803.3196	2031.2507	2022.4028	2403.7648
R2	0.75818	0.9613	0.8908	0.8369	0.8162

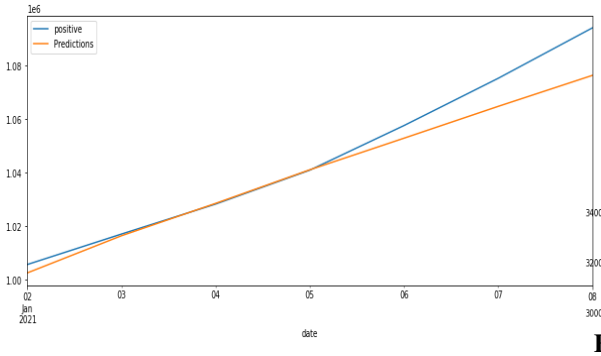


Figure 8. LSTM Network Prediction vs Actual Data from January 2nd to January 8th (Before Vaccine - New York)

The predictions of COVID cases without vaccinations and real-time case data for Washington, Kansas, and New York are shown in figures 10-12, respectively.

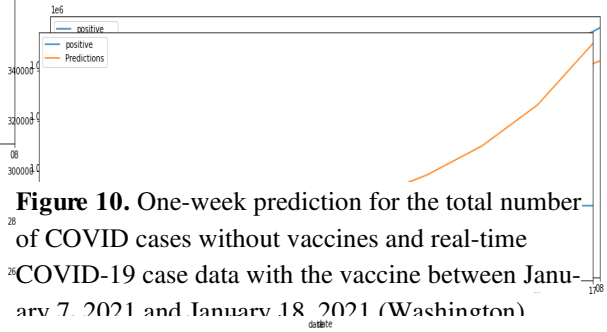


Figure 10. One-week prediction for the total number of COVID cases without vaccines and real-time COVID-19 case data with the vaccine between January 7, 2021 and January 18, 2021 (Washington)

This is because they had the greatest R2 scores as shown in Table 2. For each state, predictions showed that COVID cases would have kept growing at a stable rate if there were no vaccines. In contrast, we can see that the actual COVID data w ith vaccines showed the number of cases increasing at a much slower rate than when there were no vaccines administered. The orange line is the model's prediction and the blue line is the real number of total COVID cases.

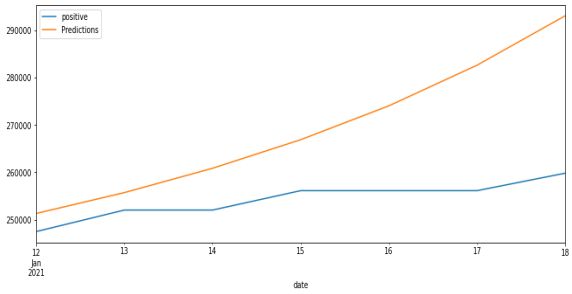


Figure 11. One-week prediction for the total number of COVID cases without vaccines and real-time COVID-19 case data with the vaccine between January 7, 2021 and January 18, 2021

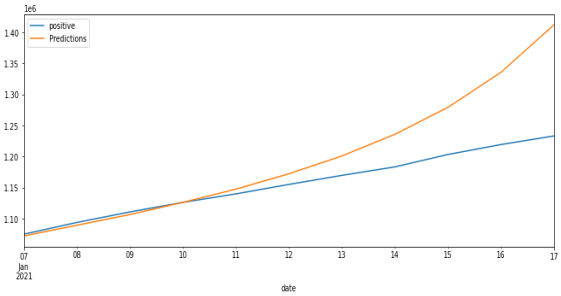


Figure 12. One-week prediction for the total number of COVID cases without vaccines and real-time COVID-19 case data with the vaccine be-

k			**Wa**		
RMSE	16906.5238	5603.0922	7482.1087	8714.1770	6311.4943
R2	-2.9332	0.8919	-18.4693	-120.2777	-12.1771

Predicting Future COVID numbers for States (After our Dataset ended)

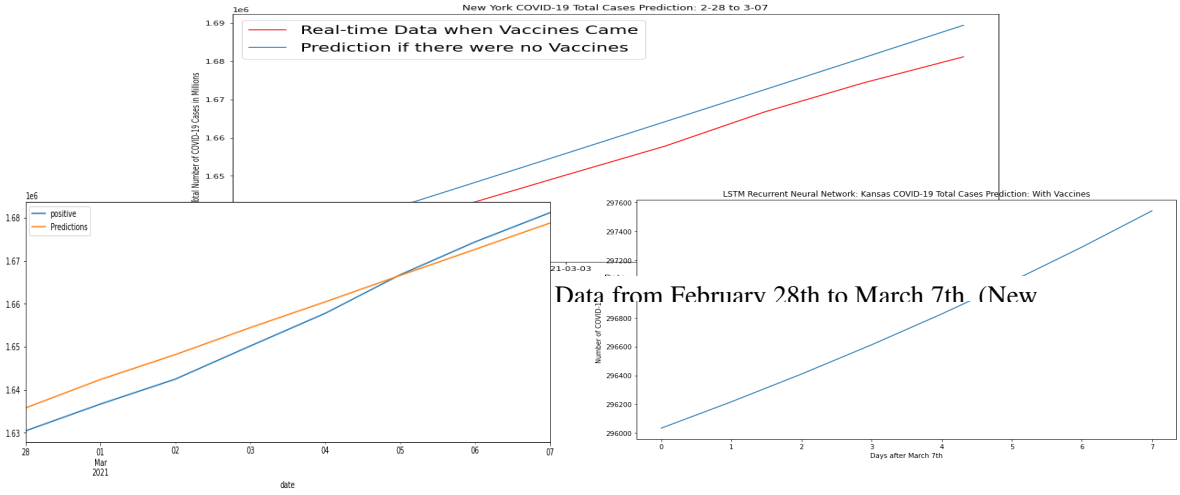
We predicted the number of COVID cases after the vaccine was delivered with both the AR model and the LSTM network for California, New York, Washington, Kansas, and Connecticut. The RMSE values and the R2 scores for the AR and RNN models are shown in Tables 3 and 4, respectively.

Table 3: *RMSE and R2 Values of our AR Model Predictions (After Vac-*

(LSTM)	California	New York	Washington	Kansas	Connecticut
RMSE	3479.2064	3992.0149	1265.9979	306.8200	1667.3495
R2	0.8334	0.9451	0.4425	0.8496	0.0803

Table 4: *RMSE and R2 Values of our LSTM Network Predictions (Af-*

Figure 12 is the AR model prediction for New York after the vaccines were issued. This is because the R2 score was high (roughly 0.89), as shown in table 3. The same goes for Kansas but for the LSTM network. The calc ulated R2 score for Kansas was about 0.63, as shown in table 4.



Data from February 28th to March 7th. (New

Because the predictions for New York and Kansas had a relatively high R2 score compared to the other states, we decided to predict a week into the future for the number of cases with the vaccine. Figure 15 is the prediction for cases using the LSTM one week after our dataset ends. Figure 16 is an AR model, with the blue line representing the number of future cases if there were no vaccines and the orange line representing if vaccines were used. Figure 17 is the same AR model but with the blue line representing the number of cases if there was no vaccine, the black line representing the real-time data, and the orange line representing the one-

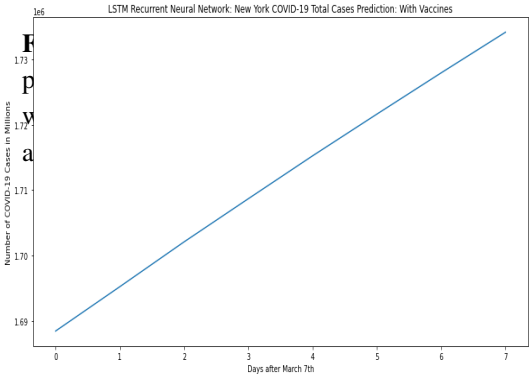


Figure. 16. Autoregressive Model one week prediction for the total number of cases since March 7, 2021 (New York)

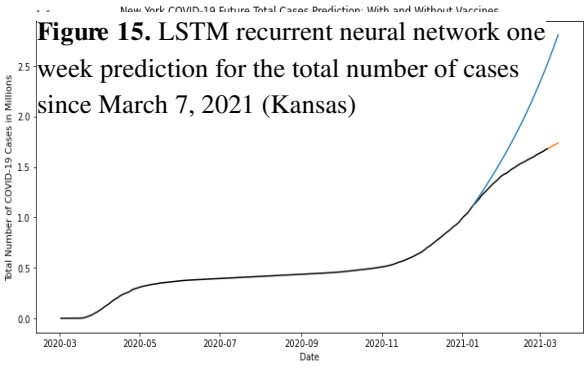


Figure 15. LSTM recurrent neural network one week prediction for the total number of cases since March 7, 2021 (Kansas)

Figure. 17. Autoregressive Model one week prediction for the total number of cases since March 7, 2021, as well as if there were no vaccines (New York)

week prediction. Both models predict a week ahead.

DISCUSSION / CONCLUSION

COVID-19 is a deadly disease that has impacted people's lives on a global scale. By putting significant pressure on healthcare systems and disrupting the global economy, it has negatively affected the quality of life for billions of people. The introduction of this paper presented a brief analysis of other COVID prediction studies using various models, in addition to revealing concerns people had about receiving the vaccine. We then used an autoregressive (AR) model and a long short-term memory (LSTM) network to predict future COVID cases with and without the distribution of vaccines. The goal of the project was to persuade people to make a data-driven decision on whether or not they want to get vaccinated to help reach herd immunity. However, a few predictions we made weren't accurate, resulting in low correlation coefficients.

Our results indicated that the LSTM Network showed better results. This was most likely because of the loop encoded within the hidden layer inside a recurrent neural network. Due to this, the LSTM can exhibit temporal dynamic responses and can easily change as more inputs are added. The LSTM was able to better respond to the changes with the inputs while the AR model wasn't. From our observations, the AR model was not able to adapt to the sudden stagnation of cases when vaccines came, resulting in a low R2 score. In addition, some states had a smaller R2 score than other states, despite having lower RMSE values. One example of this is the AR model prediction for Kansas and California. If California, which contains almost 40 million people, had an RMSE value of around 17000, the R2 score wouldn't be as affected as much when compared to Kansas, whose population is under 3 million.

One limitation of the study is that our data set was limited in size. While anywhere from one to ten thousand pieces of data will be used to construct typical autoregression and neural network models, we were limited to only 360 pieces of data. This played a major role in reducing the accuracy of our models, as they weren't trained with enough information to create accurate equations and predict future values. In addition, our model only accounted for the number of daily COVID cases to make predictions. We didn't take into account any other variables that influence and determine COVID trends, including population density in each state, lockdown restrictions, government regulations, and general human behavior. Furthermore, most states experienced sharp increases in cases throughout the fall of 2020. According to a research paper [5], this was due to other studies predicting UV rays as a factor influencing COVID spread, which we weren't able to consider. The lack of variables in our models severely impacted their accuracy, as the predictions either overestimated or underestimated actual daily by a few thousand.

Since our predictions and actual COVID data showed that there could have been many more cases in different US states if people did not get vaccinations, they could have prevented many COVID cases in the United States. If vaccines were not used, the number of COVID cases will continue to increase at a nearly constant rate as shown in the graphs. Furthermore, the actual COVID data showed that the increase in the total number of cases slowed down significantly in the first few weeks that vaccines were administered. This shows that COVID vaccines were highly effective in containing the spread of the virus and that they are an effective measure in slowing down the spread of the disease. As more Americans continue to be vaccinated, we can expect fewer COVID spikes and a decrease in total cases.

Other machine learning and forecasting tools can be examined and used to improve existing predictions. With the COVID delta variant infecting many people, machine learning models may need to be used to predict the number of cases, the spread of the virus, and even the safety of the state or country. Using the many different types of features that can affect the growth and spread of the virus, a much better and more accurate prediction performance can be achieved.

Acknowledgments

We would like to express gratitude to Ryan Solgi for guiding us throughout this project and for recommending models to use for our project to successfully do time series forecasting. We would also like to thank Laboni Sarker and S. Shailja for providing feedback on our research paper and models. We also appreciate Dr. Lina Kim for giving us this opportunity to research with peers in SRA and receive guidance from instructors.

References

1. https://www.cdc.gov/vaccines/covid-19/health-departments/breakthrough-cases.html
2. Holder J. Tracking Coronavirus Vaccinations Around the World. The New York Times. https://www.nytimes.com/interactive/2021/world/covid-vaccinations-tracker.html. Published January 29, 2021. Accessed July 7, 2021.
3. Suttie J. Five Ways to Respond to People Who Don't Want the COVID-19 Vaccine. Greater Good. https://greatergood.berkeley.edu/article/item/five_ways_to_respond_to_people_who_dont_want_the_covid_19_vaccine. Published May 26, 2021. Accessed July 7, 2021.
4. Lopez G. The 6 reasons Americans aren't getting vaccinated. Vox. https://www.vox.com/2021/6/2/22463223/covid-19-vaccine-hesitancy-reasons-why. Published June 2, 2021. Accessed July 7, 2021.
5. Harrison, Emily & Wu, Julia. (2020). Vaccine confidence in the time of COVID-19. European Journal of Epidemiology. 35. 10.1007/s10654-020-00634-3. https://pubmed.ncbi.nlm.nih.gov/32318915/
6. Rahimi I, Gandomi AH, Asteris PG, Chen F. Analysis and Prediction of COVID-19 Using SIR, SEIQR, and Machine Learning Models: Australia, Italy, and UK Cases. Information. 2021; 12(3):109. https://doi.org/10.3390/info12030109
7. Zreiq R, Kamel S, Boubaker S, Al-Shammary AA, Algahtani FD, Alshammari F. Generalized Richards model for predicting COVID-19 dynamics in Saudi Arabia based on particle swarm optimization Algorithm. AIMS Public Health. 2020;7(4):828-843. Published 2020 Nov 2. doi:10.3934/publichealth.2020064
8. https://www.ny1.com/nyc/all-boroughs/health/2021/07/02/after-long-decline--covid-19-cases-on-the-rise-again-in-the-u-s-
9. Glen S. Autoregressive Model: Definition & The AR Process. Statistics How To. https://www.statisticshowto.com/autoregressive-model/. Published August 19, 2015. Accessed July 20, 2021.
10. Brownlee J. Autoregression Models for Time Series Forecasting With Python. Machine Learning Mastery. https://machinelearningmastery.com/autoregression-models-time-series-forecasting-python/. Published January 2, 2017. Accessed July 20, 2021.
11. Bushkovskyi, O. (2019, May 30). Recurrent Neural Networks Applications Guide [8 Real-Life RNN Applications]. https://theappsolutions.com/blog/development/recurrent-neural-networks/.
12. Mittal A. Understanding RNN and LSTM. Aditi Mittal. https://aditi-mittal.medium.com/understanding-rnn-and-lstm-f7cdf6dfc14e. Published October 12, 2019. Accessed July 20, 2021.
13. Donges N. A Guide to RNN: Understanding Recurrent Neural Networks and LSTM. Builtin. https://builtin.com/data-science/recurrent-neural-networks-and-lstm. Published July 13, 2021. Accessed July 20, 2021.
14. KumarI, A. (2021, March 21). Autoregressive (AR) models with Python examples. https://vitalflux.com/autoregressive-ar-models-with-python-examples/.

15. Nachiketa Hebbar. Time Series Forecasting With RNN(LSTM)| Complete Python Tutorial| [Video]. YouTube. https://www.youtube.com/watch?v=S8tpSG6Q2H0. Published May 19, 2021. Accessed July 18, 2021.

The Data. March 2021. https://covidtracking.com/data. Accessed July 13, 2021.