**Final Report: COVID-19 Forecast**

**Team Members:**

Jared Cao Jeremy Heist Nik Sworen Dean Wang

jkc5781@psu.edu jmh828@psu.edu nas5816@psu.edu cfw5278@psu.edu

Applied Statistical Modeling Computational Computational

Data Science Data Science Data Science Data Science

**Abstract**

The COVID-19 pandemic has been a major issue for the entire world since early 2020. Creating accurate models to predict future COVID-19 cases and deaths could make people more prepared and mitigate the harm of future spikes of the virus. SARIMA, Prophet, Exponential Smoothing, and NBEATS models were implemented on data from the United States, Brazil, India, France, and Germany dating back to March 2020 to predict future case and death counts. The results show that the univariate time series Exponential Smoothing model was the most accurate on average for the five countries at predicting both cases and deaths.

**Introduction**

The outbreak of COVID-19 was officially declared a pandemic in March 2020 and changed the world in drastic ways. The virus has accounted for over 500 million cases and 6 million deaths. Throughout the pandemic, policies regarding masks, vaccines, and stay-at-home orders have been constantly changed and debated. Accurate COVID-19 forecasts for future cases and deaths could help people better prepare for spikes of the virus, and enforce the proper policies at that time to reduce the potential harm as much as possible. However, the problem of creating accurate COVID-19 prediction models comes with many challenges. The most difficult problem to overcome is the variability of the COVID-19 trends depending on the location and time. The pandemic is worldwide, and the rate of cases and deaths can be vastly different from country to country. Additionally, the virus itself and the ways we combat it have both been constantly changing over the past two years. Many different variants of the virus have come and gone and there have been several vaccines and boosters created to try to protect people from the new variants. This creates the challenge of figuring out which data to train the models on to get an accurate representation of the virus’s current state to make accurate predictions for the future. These challenges were kept in mind throughout the process of creating our solution. Firstly, we decided to focus on only the five countries that have experienced the most COVID-19 cases. We trained and tested each model five times each, once for each country. Our models consist of both univariate and multivariate time series models. Our multivariate models used features relating to vaccinations and boosters to account for that challenge. After comparing the results of our models, the final step was to train and test the best performing model on a smaller training set. This will prevent the model from using past data that may no longer relate to the current situation, to see if it improves the accuracy of the predictions.

**Related Work**

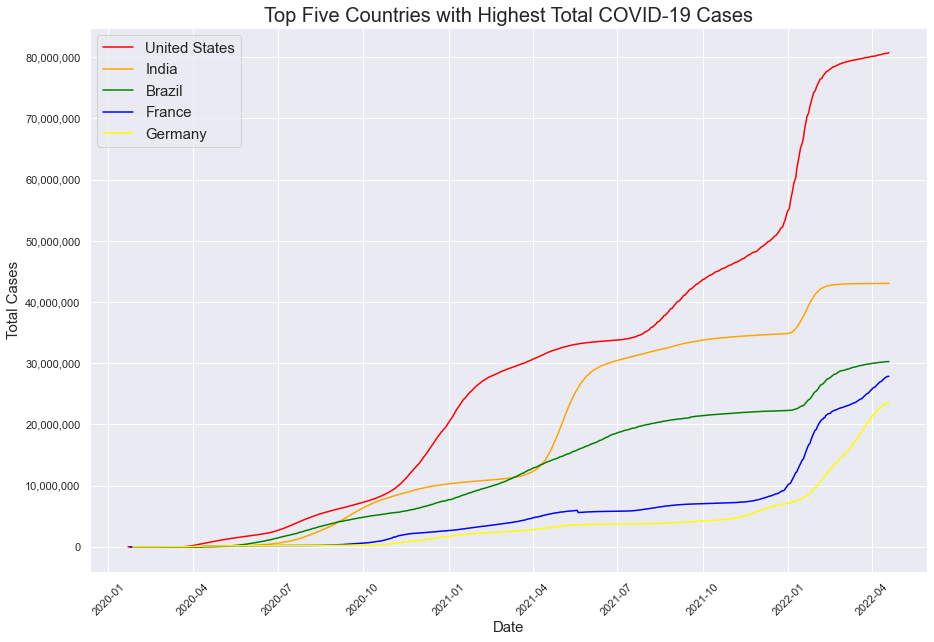
There have been a variety of approaches used in past work to solve the problem of COVID-19 prediction. [2] Rasjid et al. predicts the cases and deaths in Indonesia using Sovitsky Golay Smoothing and LSTM (Long Short Term Memory) Neural Network. They found that the neural network produced better results than the smoothing. [3] Majhi et al. researched a different approach, comparing three different machine learning models for predicting the number of positive COVID-19 cases. These models were nonlinear regression, decision tree based regression, and random forest. They concluded that random forest was the most accurate approach, which had a MAPE of 0.02%. [4] Zisad et al. compared four different models. These were SIR, SEIR, SIR with neural networks, and SEIR with neural networks. They found that the SEIR with NN model performed the best of the four.

We took this past work into account when choosing our methods to solve this problem. The past work that we studied used a variety of models, but used similar datasets. The models utilized time series data of past cases and deaths. Some models used a few other features related to infections and recoveries. Our goal was to compare a variety of models, utilizing both univariate and multivariate models. For the multivariate models we used additional features that were not accounted for in the past work that we studied. This could potentially lead to more accurate models. Some additional features that we used are related to vaccinations, which could be correlated to cases and deaths.

**Methods**

Before implementing the models, our first step was conducting the proper data exploration and data engineering. Our dataset contained data for most countries in the world, but we decided to only use data from the countries that are in the top five of total COVID-19 cases, which can be seen in Figure 1. Additionally, we split the data for each country into train and test splits. All four of the models were trained and tested on data dating back to March 2020. The training data spanned from March 2020 to January 1st, 2022, with the testing data spanning from January 2nd, 2022 to the present. After comparing the results of our models, we implemented an alternate training set on the best performing model to see if it would improve the accuracy. This alternate training set was smaller, containing data from August 2021 to January 1st, 2022.

**Figure 1.** Top Five Covid-19 Cases

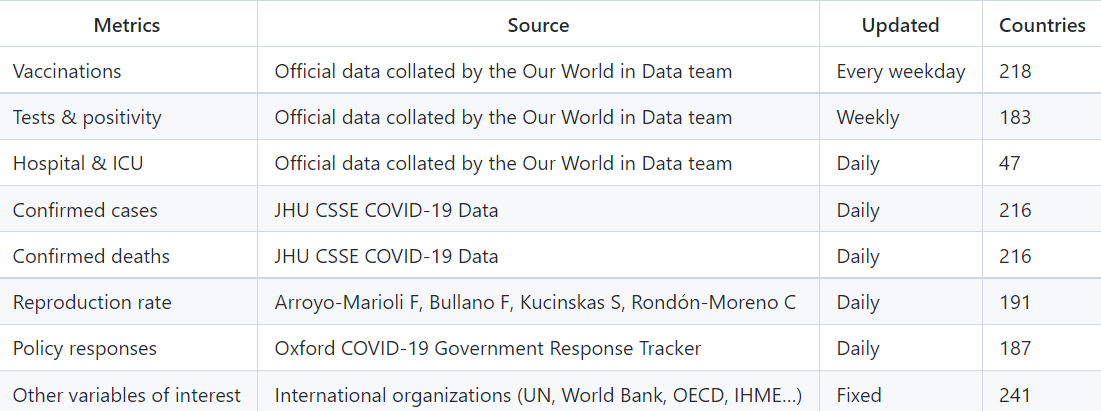


The four models we implemented were SARIMA, Prophet, Exponential Smoothing, and NBEATS. SARIMA, or Seasonal Autoregressive Integrated Moving Average, is a univariate time series model that takes seasonality into account. It also takes the difference of the target values rather than using the raw target values, and also takes lagged prediction errors as inputs. We used grid search to find and use the optimal parameters for the model. Prophet is an open-source library for univariate time series problems that was created by Facebook. It is a type of additive time series model, and is known for being easy to use. Exponential smoothing is a weighted sum of past observations where the weight is exponentially decreasing. NBEATS, or Neural Basis Expansion Analysis for Time Series, is a type of neural network. It utilizes feedforward networks with stacked residual blocks of forecasts and backcasts. We chose this combination of models for a few reasons. SARIMA, and other models in the ARIMA family are very commonly used for time series problems, so we chose that as the first model to implement. Exponential smoothing is also a commonly used model and is considered an alternate choice from something like SARIMA. Prophet is known for being very automatic and easy to use so we wanted to see how that model would fair in a complicated problem such as this one. NBEATS is a newer model that became popular a few years ago when it shockingly performed better than the winner of the M4 forecasting competition, which is a highly regarded time series forecasting contest. We wanted to see how this model would perform against the older, trusted models such as SARIMA and exponential smoothing.

**Evaluation**

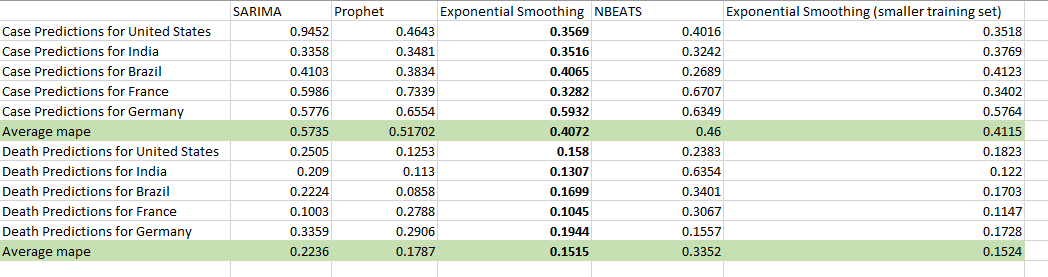
The dataset we used is publicly available on Github, and contains a collection of data maintained by *Our World in Data.* It is a worldwide dataset that contains 67 features, and has been consistently updated since the start of the pandemic in early 2020. We were looking for three things in a potential dataset, and this satisfied all of them. It has continually updated time series data for the entire pandemic, a large variety of features, and contains data for most of the countries in the world. It has many features that could prove useful in creating multivariate time series models, such as vaccination data.

**Table 1.** Dataset Information



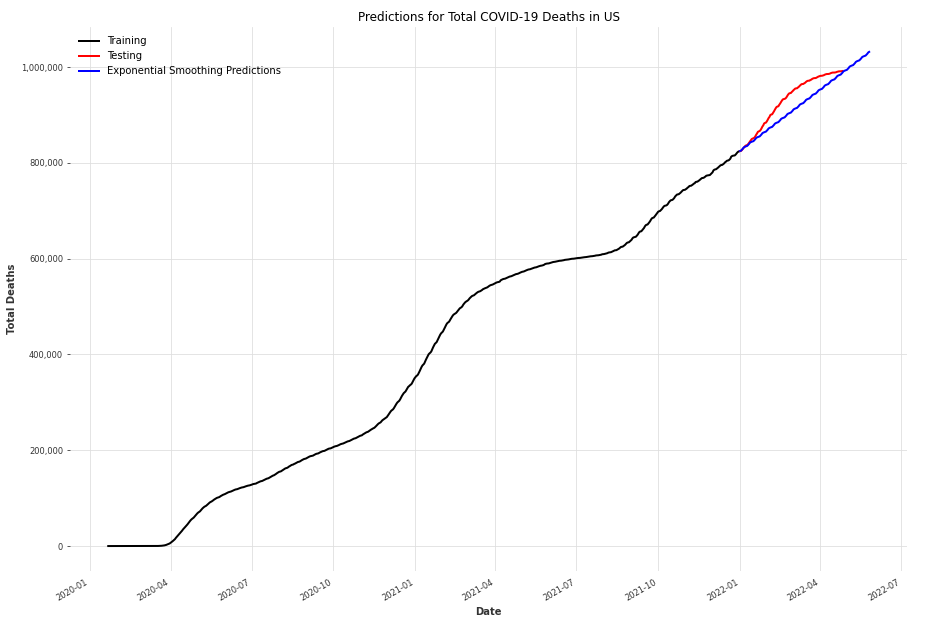
To evaluate our models, we trained and tested our models on each of the top five countries in total COVID-19 cases. These countries are the United States, India, Brazil, France, and Germany. This means that we implemented each of our models ten times each, once for cases and once for deaths in each of the five countries.

**Table 2. Comparison of Models using MAPE**



The performance metric we used to evaluate the models was Mean Absolute Percent Error. There was some variability in which models performed best on different countries. Prophet performed the best on United States deaths, India deaths, and Brazil deaths. NBEATS was the most accurate on India cases, Brazil cases, and Germany deaths. SARIMA was the most accurate on France deaths. Exponential Smoothing was the most accurate model on everything else. Although it was not consistently the best performing model in each scenario, Exponential Smoothing was overall the best performing model on average for both case prediction and death prediction. The Exponential Smoothing MAPE was 0.4072 on case prediction and 0.1515 on death prediction. Decreasing the size of the training set did not improve the average MAPE of Exponential Smoothing.

**Figure 2.** 30 Day Forecast of US Deaths



**Figure 3.** 1 Year Forecast of US Deaths

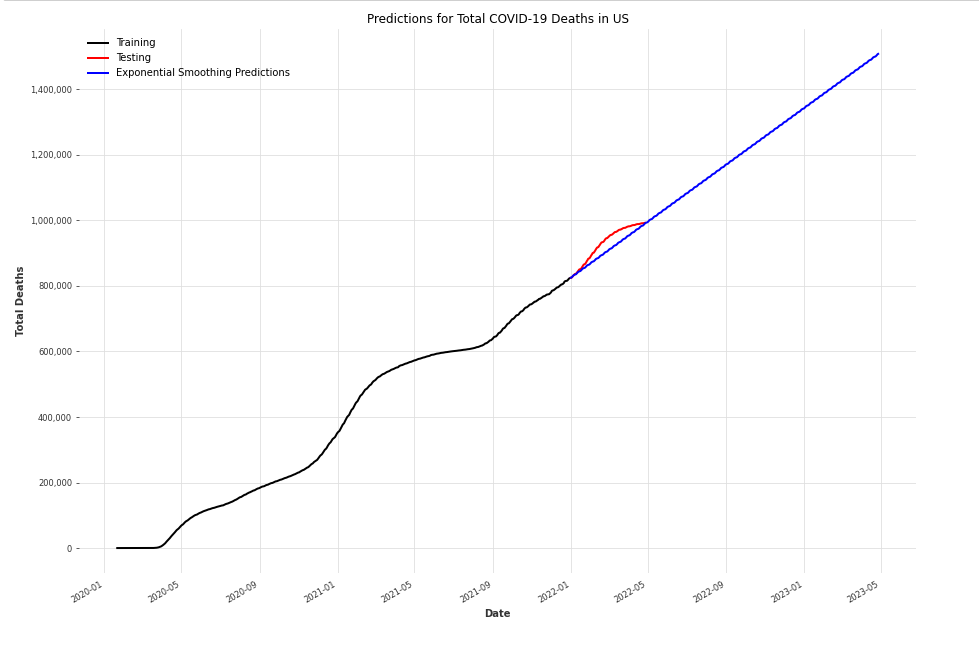
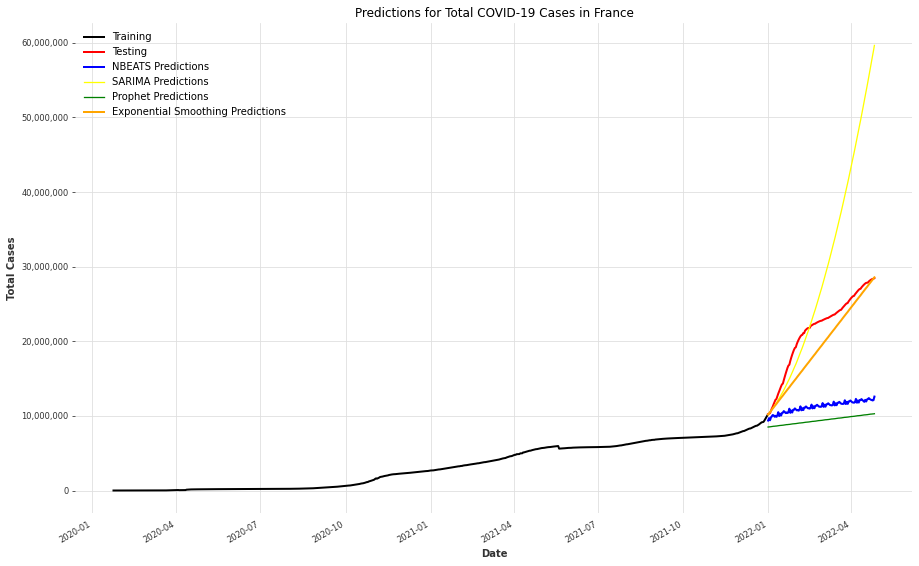


Figure 2 and Figure 3 display the final thirty day and one year forecast for United States deaths using Exponential Smoothing. Obviously, a thirty day forecast is much more practical than a one year forecast.

**Discussion**

Several things worked well for our project. For one, the dataset we used suited this problem very well and allowed us to explore the research questions that we had in mind. It is a very large dataset, and having access to constantly updated data that is consistently formatted for many countries was extremely valuable to us. Additionally, the final model comparisons using MAPE worked well for our project. The final results table gave clear results on what models worked best on each country we tested. We also developed visualizations for each country that show clearly how each model’s predictions compare to the testing set. Figure 3 is an example of one of these visualizations and shows the COVID-19 case predictions for France. We have graphs for the rest of the countries that can be found in our Github.

**Figure 3.** Predictions for Cases in France



All of the models that we tried proved somewhat successful for COVID-19 prediction. SARIMA, Prophet, Exponential Smoothing, and NBEATS all had certain countries where they performed better than the rest of the models. One thing that we were not able to implement but would like to in the future is SEIR or SIR models. There were several complications with those types of models that prevented us from having the time to implement them in this project. The biggest problem was the lack of data available on the number of people who have recovered from COVID-19. At the start of the pandemic, that data was kept track of much more closely because a higher percentage of people were dying and we did not know much about the virus. However as COVID-19 became less deadly, updated data on recovered counts have become more rare, since most people are expected to recover. Additionally, the existence of vaccinations and the possibility of reinfection are two things that would also need to somehow be accounted for in these types of models. Two other things we could work on in the future are testing our models on more countries and incorporating data on COVID-19 variants.

**Conclusion**

The best performing model on average was the Exponential Smoothing model. Although all of the models were accurate in some scenarios, Exponential Smoothing was the most consistent. At the beginning of the project we had believed that adding additional features into a multivariate time series model would improve the results, but that was not the case. We learned that in order for additional features to improve a model, they need to be highly correlated with the target variable. Although our dataset had 67 features, very few actually showed correlation with cases and deaths. The features we used for our multivariate model, total vaccinations and total boosters, were correlated with cases and deaths but still did not improve the accuracy enough to beat the best performing univariate model. Additionally, decreasing the size of the training set did not work as well as we hypothesized. The large training set still outperformed the shortened one, showing that the model benefited from having as much data to train and learn from as possible. Despite the training data dating back to early 2020 when we knew less about the virus, the data was still valuable for the models to learn the trends of COVID-19 as a whole.

**Contribution**

Throughout the semester our group regularly communicated in class, over zoom, and through text to share ideas and remain on the same page. All group members spent extensive time researching, coding, and writing to provide value to the project and reach our end goal. Many steps of this project were discussed and executed as a group. Additionally, we each had individual topics of research that we conducted during the first half of the semester to ensure we knew exactly how we wanted to tackle the problem. Some of that research does not tangibly show up in our final report, as we decided they were not the best way to tackle the problem, but the work was valuable nonetheless. Below is a summary of more specific actions that each group member took.

Jared Cao - He was named the project manager at the start of the semester. He was in charge of keeping the group on track with assignments, and would create the documents for the progress reports each week. He did a majority of writing the new sections for the progress reports each week. His initial research and coding was on models we were familiar with and had worked with in the past, such as random forest and XGBoost. We chose to not pursue those further.

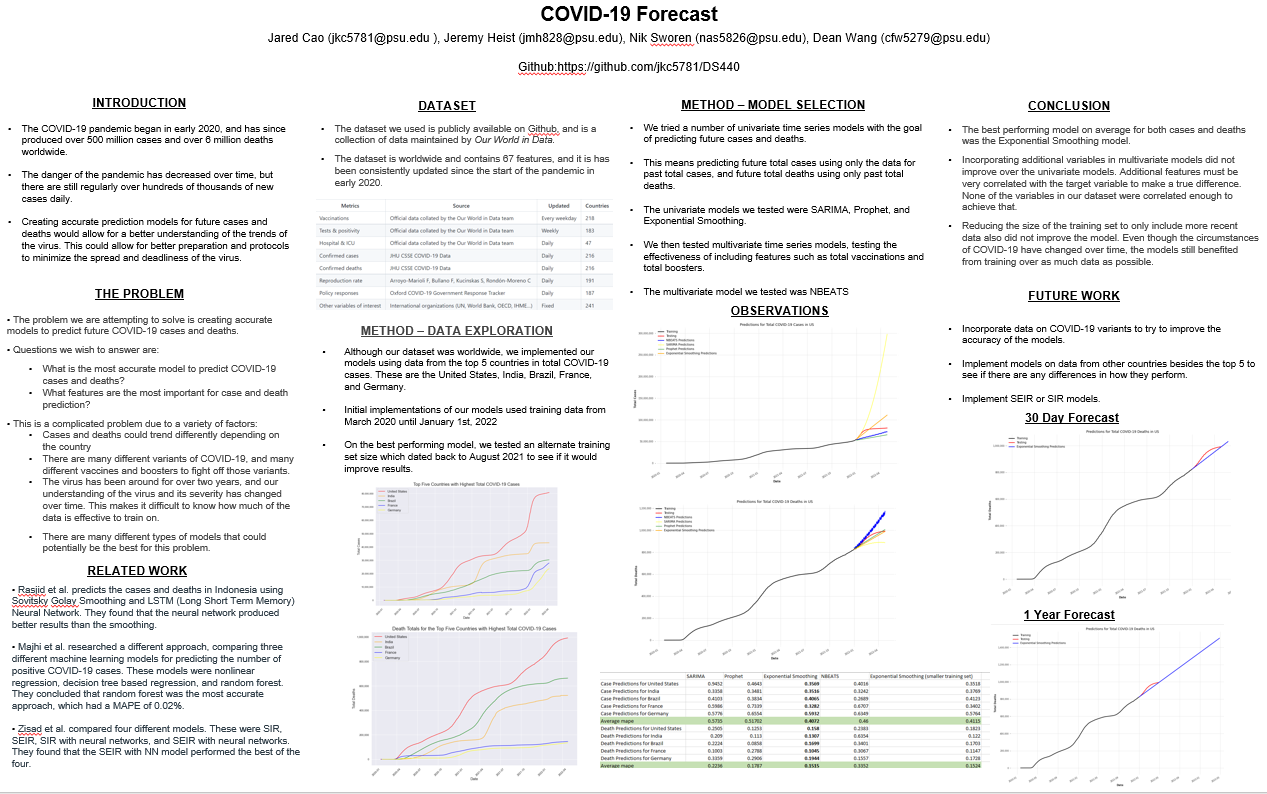
Jeremy Heist - He was in charge of writing and editing the majority of the final report and creating the final presentation. He additionally researched the papers used in the related work section. His initial research and coding was on SEIR models, which we chose not to pursue further for reasons mentioned in our future work section.

Nik Sworen - He was in charge of updating and maintaining the GitHub. He did the majority of the coding, and was responsible for writing all of the final code that was uploaded to our Github. His initial research and coding was on univariate time series models, which ended up being a major approach to solve our problem.

Dean Wang - He was in charge of creating the poster for our project. Additionally, he did the majority of updating past sections of the progress reports each week to keep it consistent with what we were doing as our project progressed. His initial research was on multivariate time series models, which became another approach we ended up using in our final report.

**Appendix**

Github: https://github.com/jkc5781/DS440

Poster: 

PDF of Poster: <https://drive.google.com/file/d/1tHFQPStvGV-rTiuZklJTqeLstPDhcZ2w/view>

**References**

[1] Our World in Data, COVID-19 Data, https://github.com/owid/covid-19-data.

[2] Zulfany Erlisa Rasjid, Reina Setiawan, Andy Effendi,

A Comparison: Prediction of Death and Infected COVID-19 Cases in Indonesia Using Time Series Smoothing and LSTM Neural Network, Procedia Computer Science,Volume 179,

2021, Pages 982-988,ISSN 1877-0509, https://doi.org/10.1016/j.procs.2021.01.102.

[3] Majhi R, Thangeda R, Sugasi RP, Kumar N. Analysis and prediction of COVID-19 trajectory: A machine learning approach. J Public Affairs. 2021;21:e2537. https://doi.org/10.1002/pa.25378.

[4] Zisad, S.N.; Hossain, M.S.; Hossain, M.S.; Andersson, K. An Integrated Neural Network and SEIR Model to Predict COVID-19. Algorithms 2021, 14, 94. https://doi. org/10.3390/a14030094