Evaluating the Infection of COVID-19 Using Machine Learning

1 Abstract

COVID-19 has ravaged the entire world, instead of case numbers declining the COVID-19 infection has been increasing daily. It took around 10 months for 50 million people to get infected and only another 2 months for another 50 million to get infected. The rate of infection is on the rise. Unlike other types of infections, COVID-19 has been the most widespread which has now spread and mutated with more transmissible variants and variants with higher death rate. To manage this crisis governments have put unprecedented plans but the effectiveness of the plan's result can only be seen after a week or two [1]. We can use simulations to test in, what these policies effect the COVID-19 cases and Machine-learning to predict new infection numbers using our world data.

2 Background and Research problem

We aim to determine how different measures can prevent infection numbers to surge and how different the new UK variant (B1.1.7) is compared to the original virus. As the case numbers increase around the world so does new variants, due to millions of people getting infected, countries must prevent the infection of more infectious variants than the original COVID-19 strain. The null hypothesis for this study is the new variant of COVID-19 or UK variant (B1.1.7) is more infectious than original strain of the virus and there will not be an effect predicting the covid-19 cases due to variants. The alternative hypothesis for this study is there will not be significant difference and cause the same number of infection numbers whilst making it harder to predict COVID-19 cases in the future. We will check if other factors such as lockdowns make big changes. We will also try to model and predict the next 7 days of COVID-19 infections from the past COVID-19 cases.

3 Approach Description

Firstly, we will need to confirm what happens if no preventive measures are taken place for a population of 200, for the normal COVID-19 or SARS-COV-2 (Figure 1), the population entered medical emergency at day 10 (frame 100) as seen in Figure 3, whilst the B1.1.7 variant or UK variant (Figure 2) took only half that of day 5 (frame 50) as seen in figure 4. From the fist simulation, we can see how dangerous the new variant is compared to the first confirmed COVID-19.

To compare how dangerous the new variant is, we will simulate one population with both normal and B1.1.7 UK mutant simulated as in figure 5 and compare it with a population with the two normal variants simulated in figure 6. When the infection starts, the B1.1.7 infects most of susceptible people and dwarfs the original COVID-19 strain which puts the population of 200 into medical emergency before day 5(frame 50) as seen in Figure 7. In the other hand, the population of 200 with 2 normal covid-19 infections, they only reach medical emergency at around day 9 as seen in figure 8.

Further on, we will look at preventive measures taken place by governments which are Curfews and Lockdowns. With curfew at 11p.m., prevention of movement till 6am will result in 30% of movement reduction, the results are shown in figure 9 where the population enters medical emergency only from day 12 which does not show enough potential to prevent the infection of COVID-19 for the normal variant.

Compared to curfews, lockdowns are harsher and are known to prevent sudden infection surge like in other simulations. In lockdowns only essential service workers were allowed travel meaning the risk to spread or contract the virus itself. As in Figure 10, the simulation shows the population of 200 entered medical emergency at day 25 which is further than any other simulation showing the most

effective infection prevention method. With the help of quick testing and quarantining ones who were infected will show the true capability of the measure.

From these simulations, we have found out that COVID-19 cases are directly associated with movement, the more movement is reduced, the better is for the population. COVID-19 variants will spread faster but if movement is reduced, it seems unlikely to spread widely. We can confidently say that from our simulation and conclusions from stage 1. the most important factors of predicting new COVID-19 cases movement, testing, positive rate etc.

For us to predict the covid-19 case numbers, we have to pick which features affect the number of new cases the most, for example if most cases are confirmed in places of worship and restaurants we can feed the model with information that are useful. In this case, we will be using mobility data from google [2] which are retail and recreation, grocery and pharmacy, residential, stations, parks, workplaces and covid-19 data such as new cases, new deaths, new tests, positive rate, total vaccination and population density.

As the infection count grows the higher the chances for the new variant to emerge proving new problems. New variants were first spotted around April and December 2020, which means if we evaluate the data entire dataset, the dataset with and without variants will show similar results. We need to train the model so it knows that new variants will be there and make predictions assuming them. Therefore, providing the model with the information of the new variants will not be feasible as there will be a blank time where researchers will need to research about the variant and then only we can feed our model with it. Our data of 1+ year already includes dates where variants spread within the population and the model shall learn from these experiences.

In the first stage of the projected we assembled about 18k bytes or 76k rows of “our world in data” dataset. The dataset contains each line for every country and every date since 2020 March when the pandemic started. We further on collected google mobility data to track how much movement is going on a country. This Google mobility dataset is divided into Country, date, movement in retail, grocery, residential, transit, parks and workplaces. To determine what features are needed to predict COVID-19 infection numbers we will make simulations for different scenarios such as Lockdowns, curfews and variants.

For this problem we will use time-series prediction method which is implemented by giving the model past 7 days of data (infections, tests, etc..) and the model will predict the next 7 days of covid-19 predictions. We will fit our data to two different countries with two very distinct situation which are United Kingdom and Japan. UK had very high number of COVID-19 cases with lockdown but UK has a swift vaccine roll where 60% have received the first vaccine shot, and Japan where the case numbers have been relative low but the vaccine shots have not been anywhere close to UK’s with mobility changes such as soft curfew and lockdowns.

About the model:

- Fully connected layer (Dense Model
- Input layer (Batch Size, 7, 12)
- Output (7,1)
- Total of 21 thousand parameters
- Optimizer: Common Adam optimizer

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense (Dense)                (None, None, 32)          416
_________________________________________________________________
dense_1 (Dense)              (None, None, 64)          2112
_________________________________________________________________
dense_2 (Dense)              (None, None, 128)         8320
_________________________________________________________________
dense_3 (Dense)              (None, None, 64)          8256
_________________________________________________________________
dense_4 (Dense)              (None, None, 32)          2080
_________________________________________________________________
dense_5 (Dense)              (None, None, 1)           33
=================================================================
Total params: 21,217
Trainable params: 21,217
Non-trainable params: 0
_________________________________________________________________
```

4 Evaluation Setup

As seen in Figure 13, the dataset is divided into 8:1:1 division where 8 is training, 1 is testing and 1 is DevTest. To determine how accurate our model is, we use the model's evaluate function to evaluate the correctness of model. For validation for our model we will use the evaluation method of mean accuracy error of which the formula can be seen in Figure 15, which means it is the absolute mean of Expected value subtracted by the predicted value.

For better evaluation technique we will need to look at its actual prediction to truth data differences on a graph to visually evaluate the model. We can try predicting the next 7 days COVID-19 cases using the past 7 days.

5 Results and Analysis

As from the simulation, we concluded not to have variant data to be inserted for training because of new variants will emerge in the future making it more unpredictable as variant information will not be present at that time. For example, the UK variant was first confirmed in December 2020, if the model were trained before December, the model would not perform well for data after December due to variant's surge in infections. If we do not include that data, we would not need to worry about the variant as the model would expect new variants and make predictions accordingly.

Lockdown and curfews would directly affect the movement of people's daily life, also meaning less movement. For this case, we included data from Google's mobility dataset which include movement in retail, grocery, residential, transit, parks and workplaces. This would tell the model the likelihood of contact between people and increase the risk of infections. As from the curfew and lockdowns we understood that there is a very high correlation between people's movement mobility and infection numbers. For example, in Japan a soft lockdown which requests and not orders restaurant to close at 8 p.m. gave massive results that reduces Tokyo's infection number from 2500 a day to just 108 a day within a month. These targeted mobility data is what Google's mobility dataset provides us.

On 1000 epochs of training on the Japanese data, our model's evaluation has concluded mean accuracy error of just 38.314 as seen in figure 14 which is low with testing loss as low as 60 and training loss about 100 as seen in figure 11 but the model tested better for DevTest dataset. However, the model's training loss higher than testing is questionable but at the end of the training, the model's training and testing loss became normal.

After 1000 epochs of training the Japan model showed significant decrease in training and testing loss as seen in figure 11. The evaluation on the DevTest dataset gave a mean accuracy error of just 38 as seen in figure 14.

Also, after 1000 epochs our UK model, unlike our Japan model's evaluation showed mean accuracy error of 42 as seen in figure 18. The training and testing loss does not differ from Tokyo, but the data seems very inconsistent with Japan's as Japan maintains consistent waves of covid-19 cases but UK has sudden surges and sudden plunges due to lockdowns and COVID-19 vaccine roll outs. Our prediction in Figure 17 for DevTest data shows only a little correctness compared to Japan's model.

From COVID-19 UK variant simulation, variants have a very big impact on the surge of COVID-19 case numbers in the population. Variants cause significant change as the population will enter the state of medical emergency faster than the normal COVID-19. Also, with Lockdown and curfew simulation which was designed to reduce the mobility of people in the population gave us important insights that indicate that there is massive correlation between infection numbers and mobility. By using the mobility information, we can clearly have mobility as one feature to train our model.

From training our model for two very different countries, we found out that infection numbers can be predicted with the help of different factors. Even without keeping the COVID-19 variant information, we can accurately predict the COVID-19 case numbers by an average error of ±40 meaning as seen in figures 12 and 17 where the model predicts the COVID-19 cases. In most instances, the model perfectly captures the COVID-19 case's trends giving the correct indicating if there will be surge or the case numbers will plunge. When these COVID-19 case numbers are smoothened like our inputs, it would give a very accurate prediction.

As looking at the infection numbers decrease in the last couple of months in the UK shows that the vaccine rollout has proven effective. The model has a mean accuracy error for both training and testing of about 42 meaning that the model can predict accurate prediction but due to vaccination and not due to variants, the model seems to prove not reliable for only the last couple of months or when the vaccine roll out got more intensive. The Japan model proves us without vaccine rollout, predicting COVID-19 cases is not that hard.

We can retain our null hypothesis as from the simulation we have proven that COVID-19 UK variant is more infectious than the original SARS-COV-2 when compared side by side. When predicting, we also have proven that without a proper vaccine rollout which slows down the infections in population, we can predict COVID-19 cases for population like Japan. But in exceptional cases like UK, where the vaccine rollout has been going well, the model is not reliable since when the vaccine rolls out began.

Even looking at our Figure 19 and 20 we can see that vaccinations reduce the susceptibility in population such as United States and Israel where vaccination speeds have been the highest in the world. This increases the hurdle for prediction therefore our model may not

In conclusion, due to less data on factors that will decrease the COVID-19 infections, we find it hard to predict COVID-19 cases. For example, the vaccinations we do not have enough data whereas lockdowns we do, both have the same effects to the population but the model seems to predict bad only on dates where more vaccinations meaning that the model was yet to learn that vaccinations make a change in infection numbers as seen in the UK variant. We could correctly predict in cases where vaccination roll out was not that up to date. Therefore, we can conclude our research and analysis on COVID-19 cases and prediction.

Datasets

1. owid-covid-data.csv: https://ourworldindata.org/coronavirus-source-data
2. changes-visitors-covid.csv: https://www.google.com/covid19/mobility/

Works Cited

[1] E. Hersh, "Health Line," 13 March 2020. [Online]. Available: https://www.healthline.com/health/coronavirus-incubation-period#:~:text=Most%20people%20who%20develop%20COVID,call%20your%20doctor%20for%20advice.. [Accessed 15 05 2021].

[2] "Our World In Data," Oxford Martin, 17 5 2021. [Online]. Available: COVID-19 . [Accessed 13 5 2021].

Appendices

A Covid-19 SARS-COV-2 Simulation

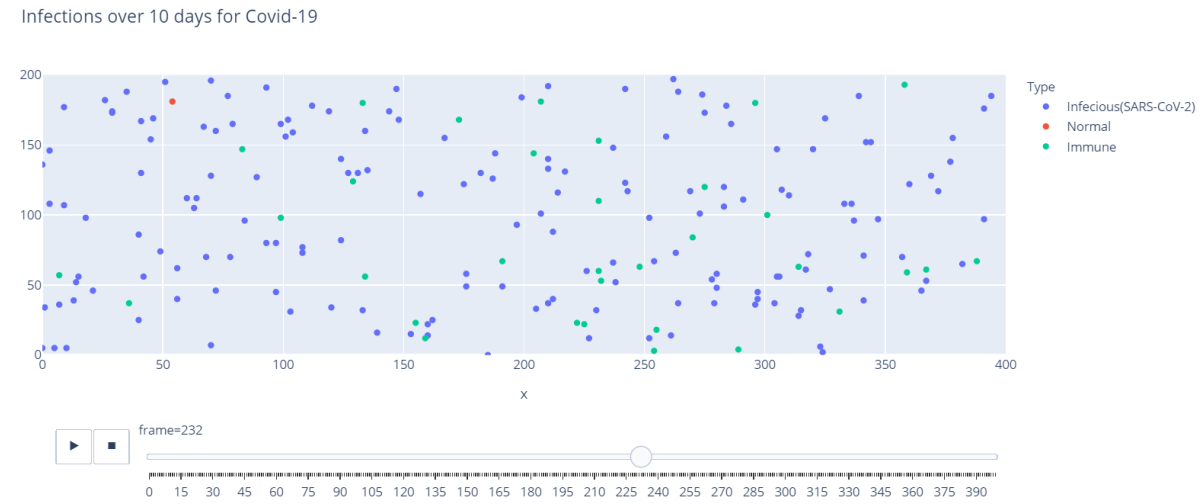

Figure 1: Normal SARS-COV-2 infection simulation on hypothetical population of 200

B Covid-19 B 1.1.7 variant Simulation

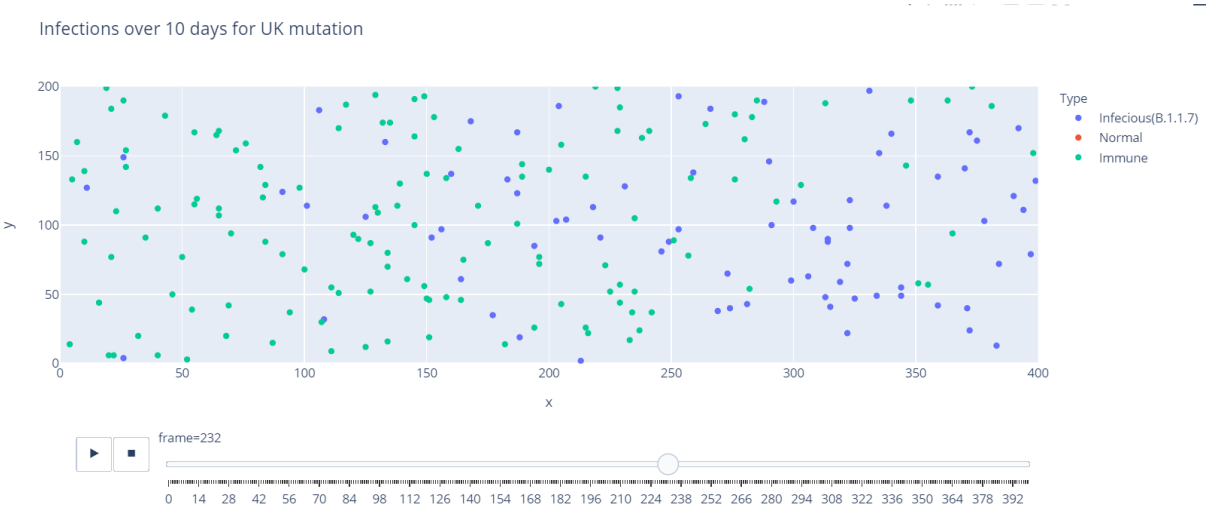

Figure 2: Covid-19 B1.1.7 variant infection Simulation on hypothetical population of 200

C Normal Covid-19 SARS-COV-2 infection count

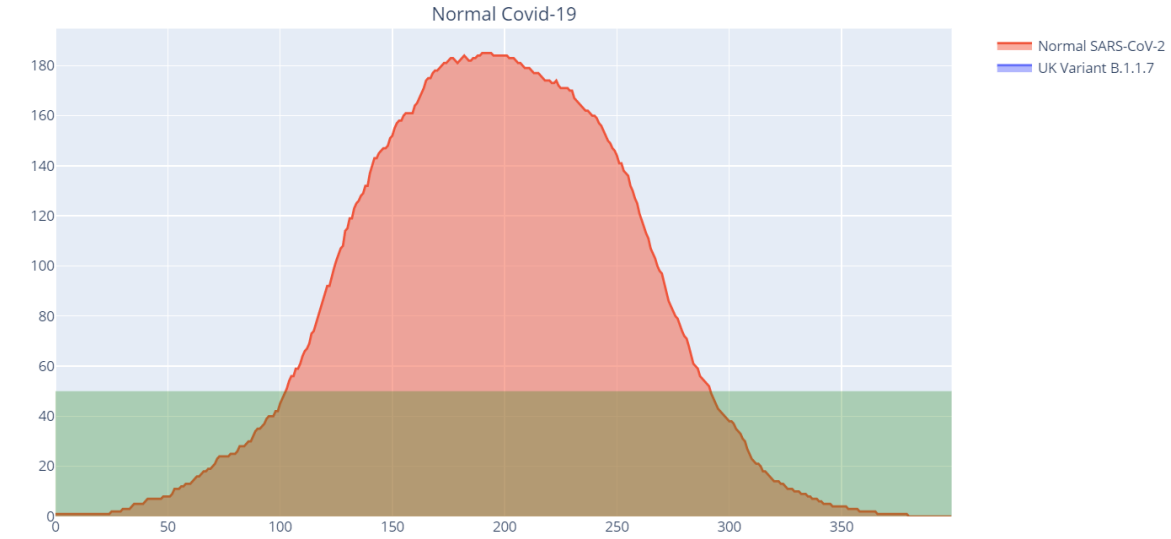

Figure 3: Normal SARS-COV-2 infection number, the green zone indicates the 25% and within medical capacity area.

D Covid-19 B1.1.7 infection count

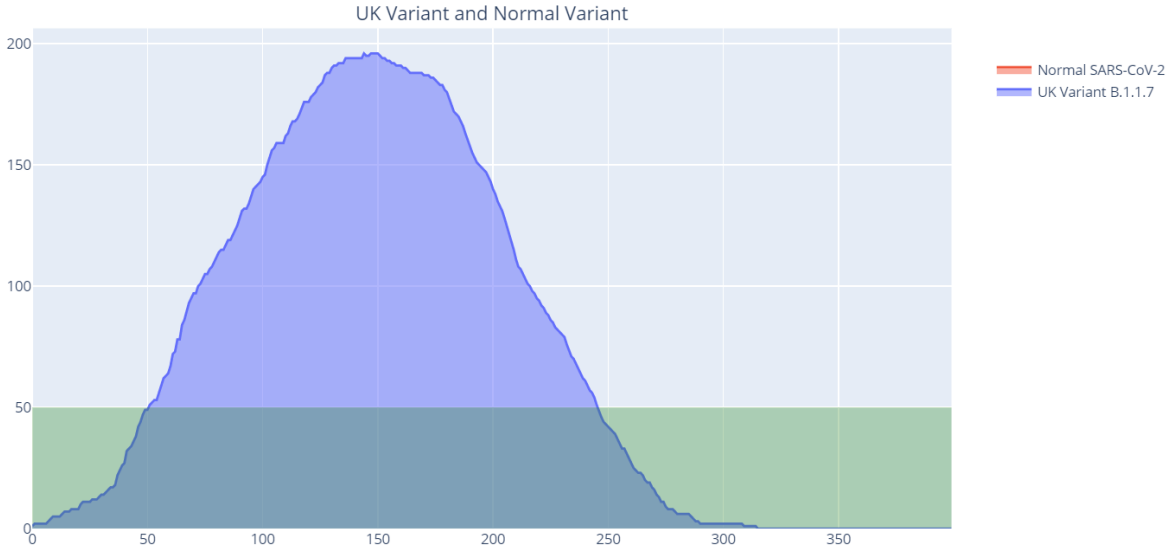

Figure 4: COVID-19 B1.1.7 variant infection number, the green zone indicates the 25% and within medical capacity area.

E COVID-19 and COVID-19 B1.1.7 variant together simulation

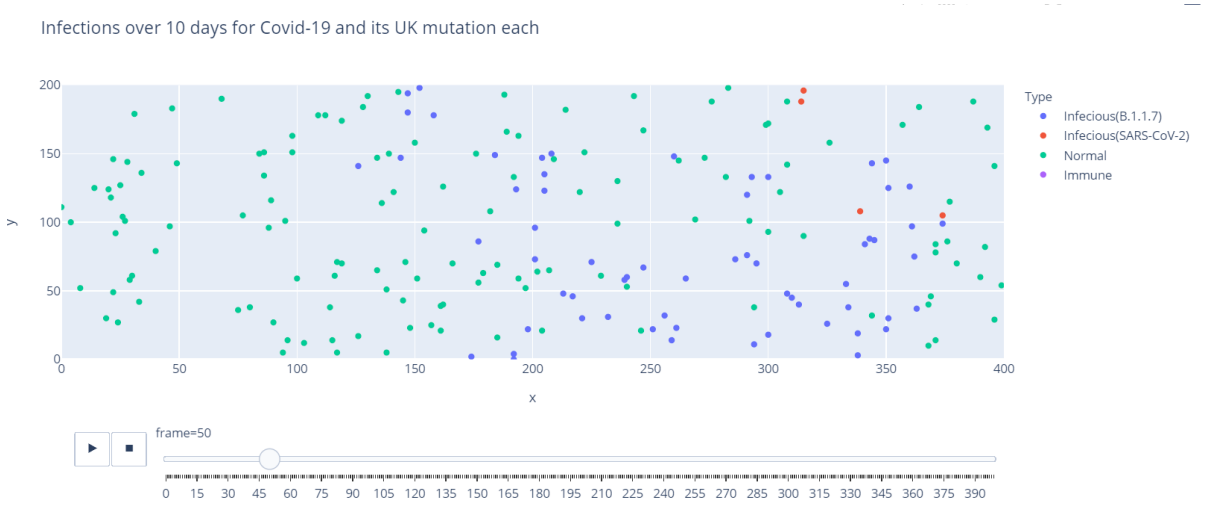

Figure 5: Normal SARS-COV-2 and B1.1.7 variant infection simulation on hypothetical population of 200

F 2xCOVID-19 simulation

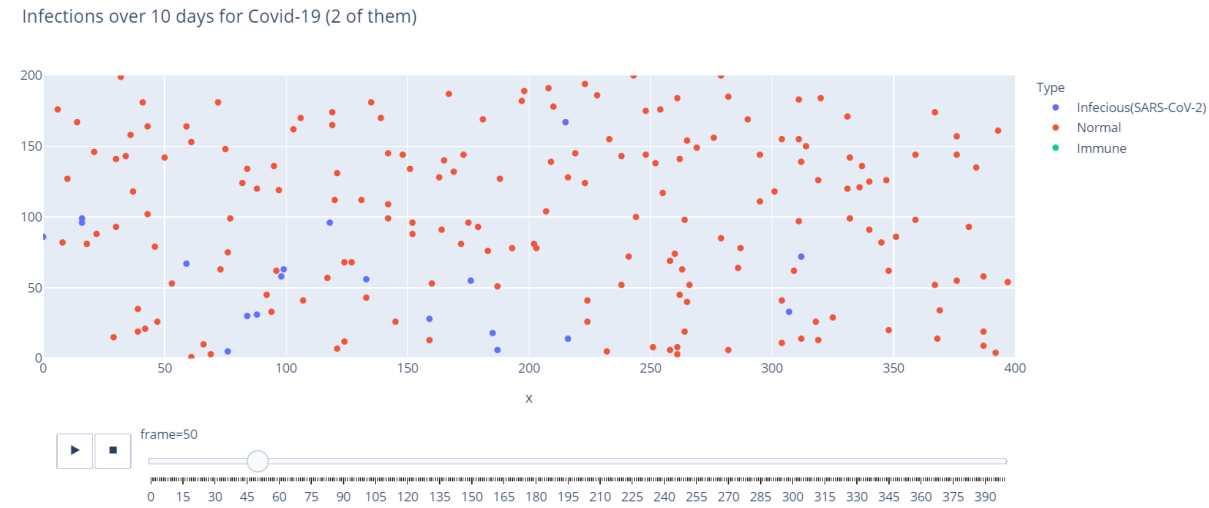

Figure 6: 2x SARS-COV-2 infection simulation on hypothetical population of 200

G Covid-19 B1.1.7 and Normal COVID-19 infection count

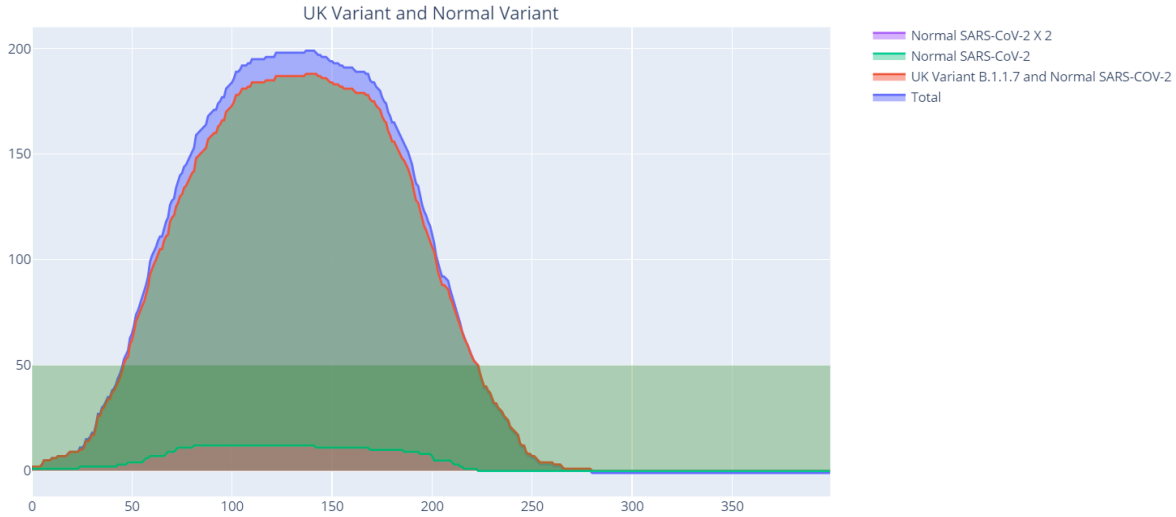

Figure 7: COVID-19 B1.1.7 variant and Normal SARS-COV-2 infection number, the green zone indicates the 25% and within medical capacity area.

H 2x Normal COVID-19 infection count

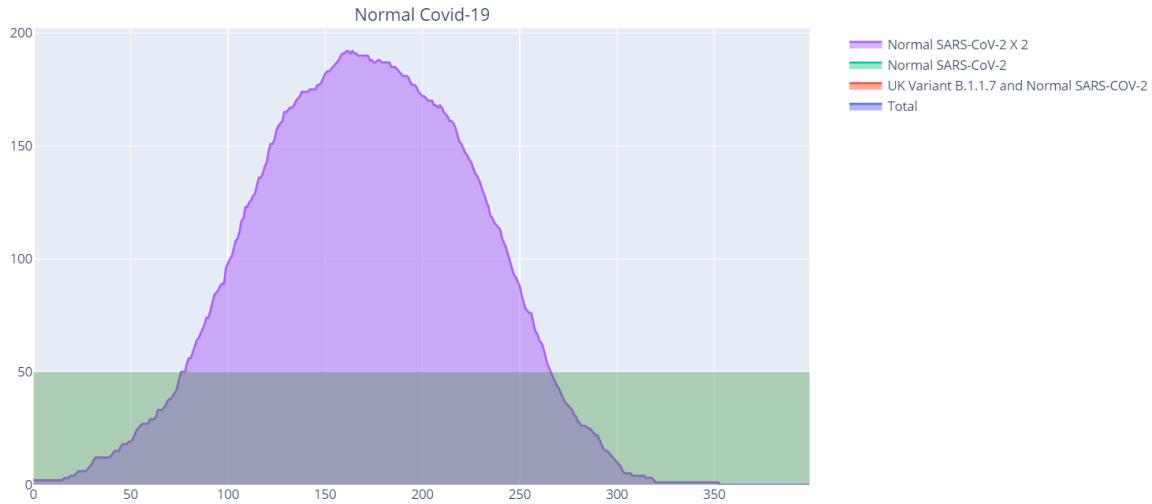

Figure 8: 2x Normal SARS-COV-2 infection number, the green zone indicates the 25% and within medical capacity area.

I Covid-19 Simulation with Curfew (30% movement reduction)

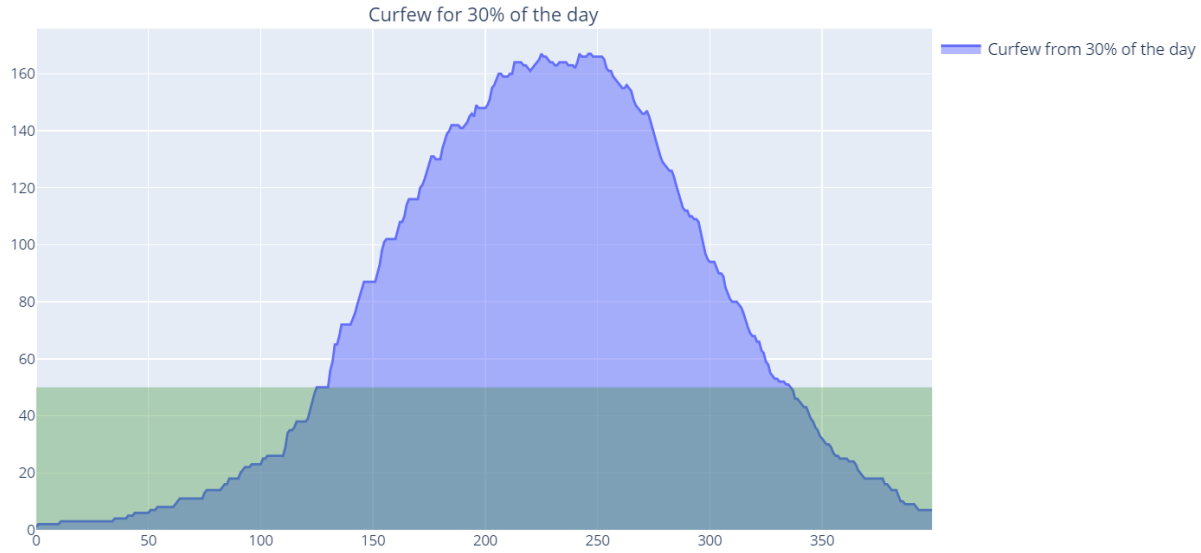

Figure 9: Normal SARS-COV-2 infection number for 30% movement reduction/curfew, the green zone indicates the 25% and within medical capacity area.

J Covid-19 Simulation with Lockdown (Essential Worker movement Only)

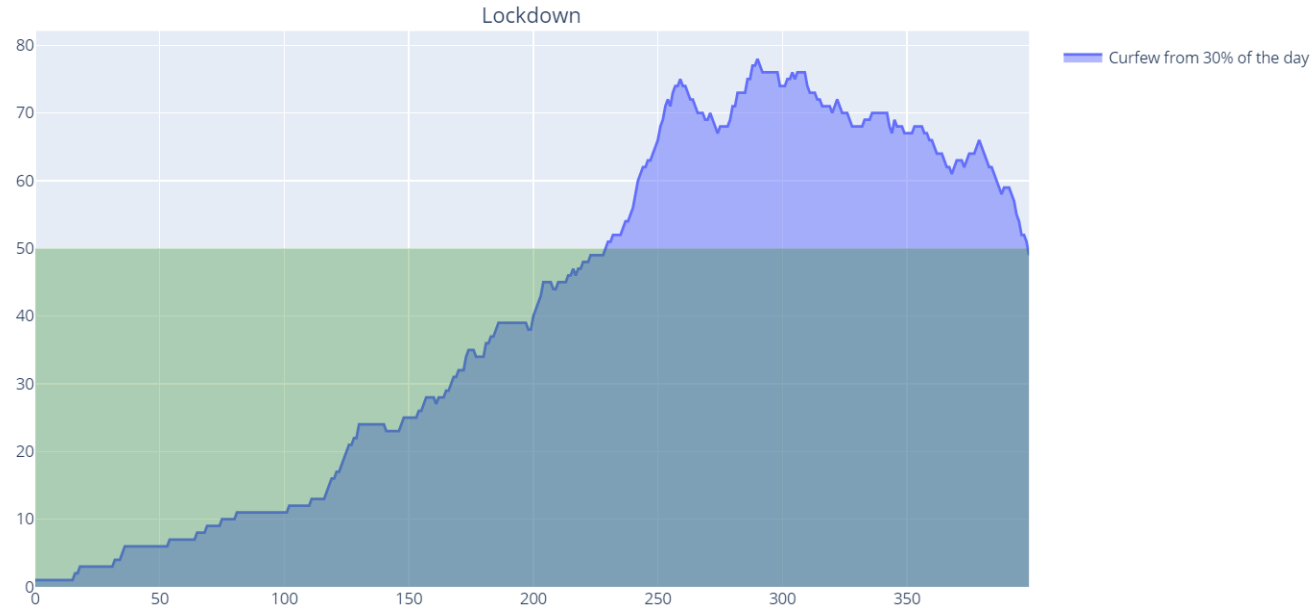

Figure 10: Normal SARS-COV-2 infection number for Lock down population with movement of only essential workers, the green zone indicates the 25% and within medical capacity area.

K Training and Testing loss for COVID-19 cases in Japan

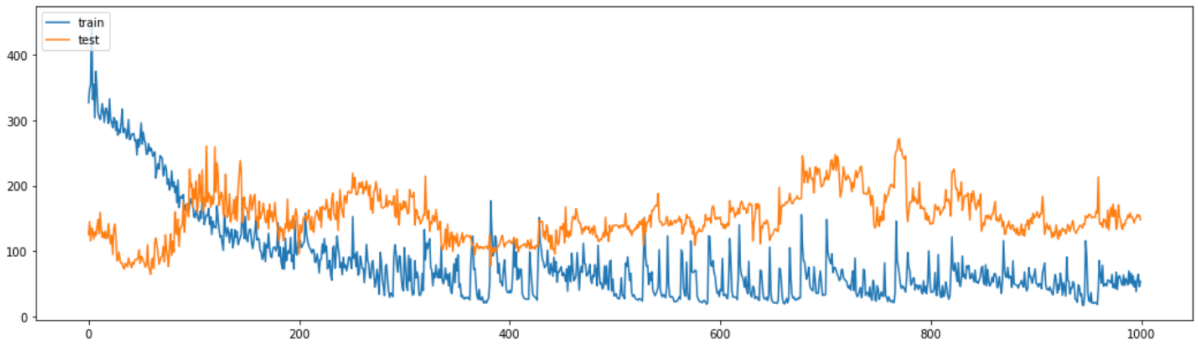

Figure 11: Shows training and testing loss for training to predict covid-19 cases in Japan

L Sample Prediction for Country Japan COVID-19 cases

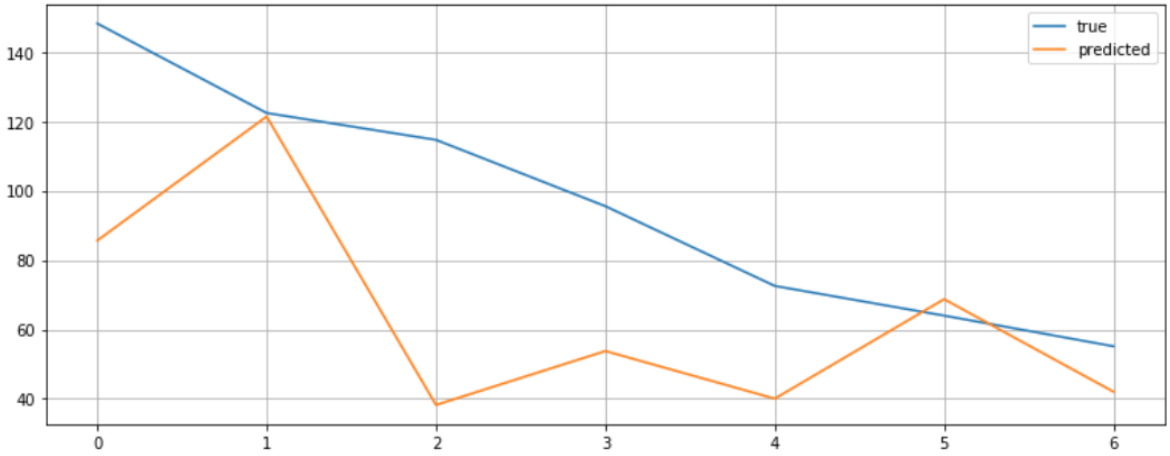

Figure 12: Shows the true covid-19 case numbers and predicted case numbers for Japan Model

M Data division for COVID-19prediction Data

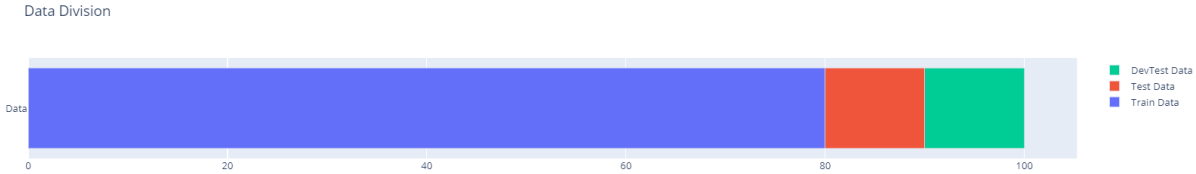

Figure 13: Graphs show how our training and testing data is divided.

N Model Evaluation for Model Japan

```
[21]: model = tf.keras.models.load_model('Model/firstModel.h5')
      model.evaluate(dataGenerator_DevTest(BatchSize, SequenceLength, features, population), steps = 100)

      100/100 [==============================] - 1s 8ms/step - loss: 2104.2078 - mae: 38.3150
[21]: [2104.2077783203126, 38.314995]
```

Figure 14: Evaluation of Model for Japan.

O Mean Accuracy Error

$$mae = \frac{\sum_{i=1}^{n} abs\left(y_i - \lambda(x_i)\right)}{n}$$

Figure 15: Formula for Mean Accuracy Error

P Training and Testing loss for COVID-19 cases in United Kingdom

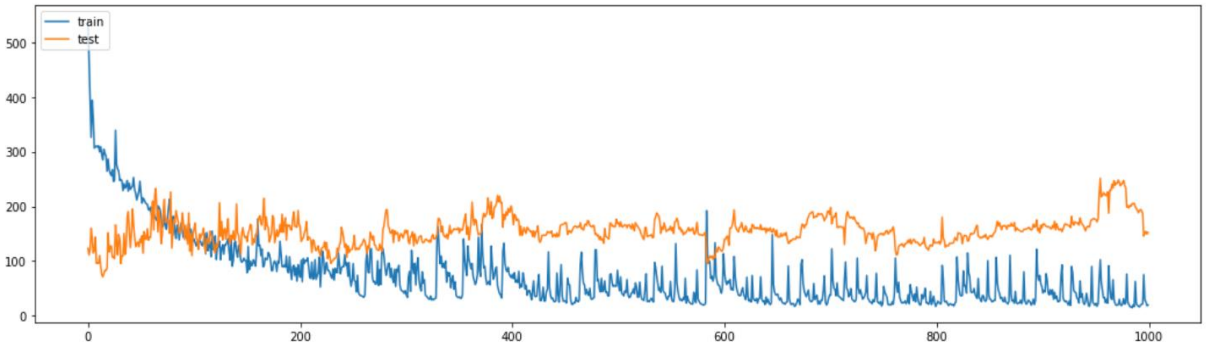

Figure 16: Shows training and testing loss for training to predict covid-19 cases in UK

Q Sample Prediction for Country United Kingdom COVID-19 cases

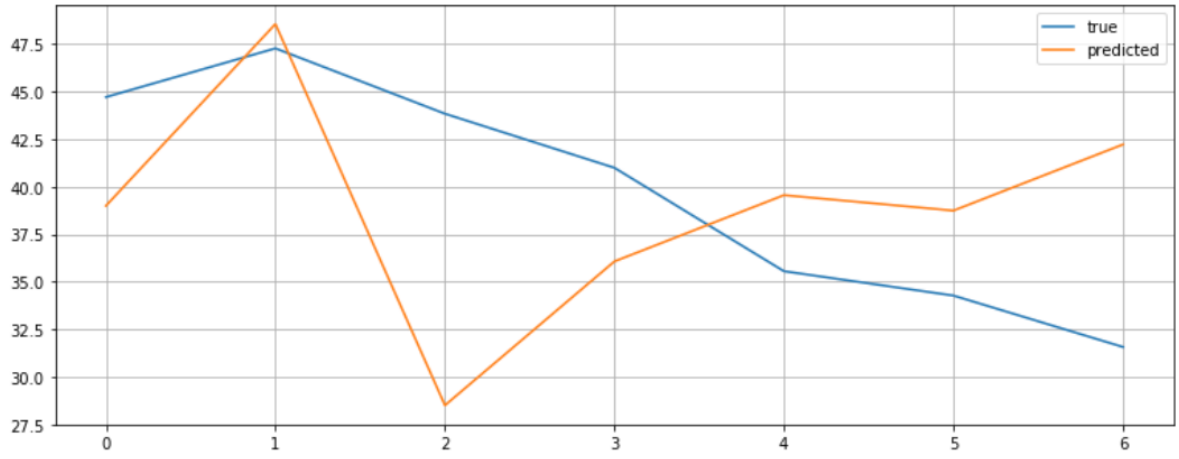

Figure 17: Shows the true covid-19 case numbers and predicted case numbers for UK model

R Model Evaluation for Model UK

```
[32]: ukModel.evaluate(datagen(ukxDevTest, ukyDevTest, 16, 7, 12), steps = 100)

      100/100 [==============================] - 1s 6ms/step - loss: 3068.2620 - mae: 42.0461
[32]: [3068.261951904297, 42.046093]
```

Figure 18: Evaluation of Model for UK.

S COVID-19 vaccinations and new COVID-19 case trend

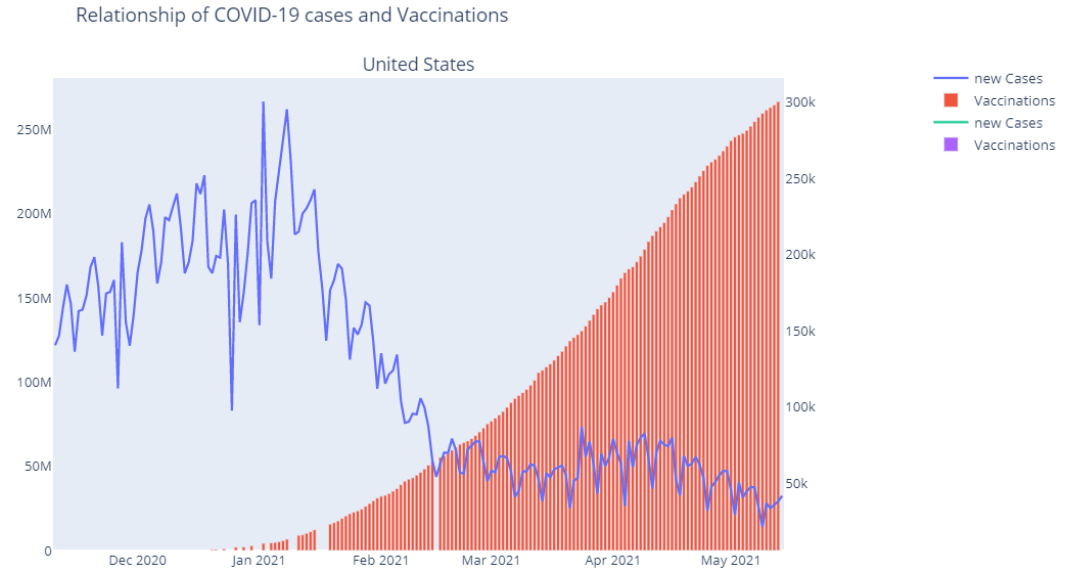

Figure 19: New cases and Vaccination graph for United States

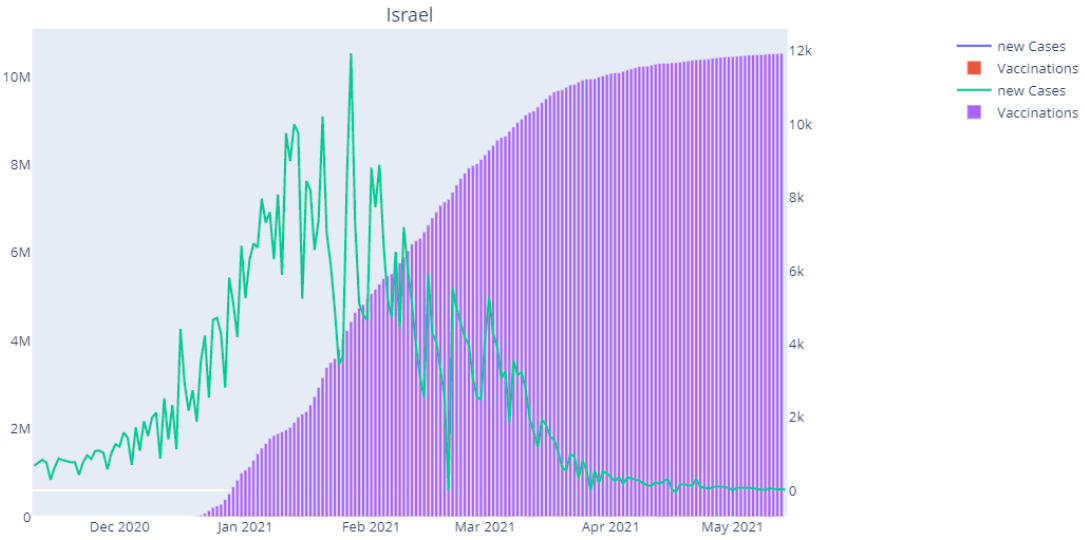

Figure 20: New cases and Vaccination graph for Israel