COVID-19 Predictions using deep learning techniques

Background and Problem

Coronavirus disease 2019 is known as COVID-19 or 2019-nCoV which is an infectious disease. This virus leads to acute respiratory syndrome in many patients. The first alarm of COVID-19 disease was in December 2019 in Wuhan (capital of Hubei province) in China and has since spread globally were ongoing today as coronavirus pandemic. COVID-19 is currently one the most life-threatening problems around the world. The fast and accurate detection of the COVID-19 infection is essential to identify, take better decisions and ensure treatment for the patients which will help save their lives. In this project, we study a fast and efficient way to identify COVID-19 patients with multitask deep learning (DL) methods.

Dataset

In this study, we used github repository for the 2019 Novel Coronavirus Visual Dashboard operated by the Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE https://github.com/CSSEGISandData/COVID-19) The dataset is available in the time series format with date, month and year so that the temporal components are not neglected.

Data Cleaning and Preprocessing:

For this project, dataset from John Hopkins university center for system science and engineering have been used for analysis and visualizations. We have used dataset including 2019 coronavirus dataset (January-July, 2020), which tracks the spread of 2019-nCoV, COVID-19 (nCOV-19) coronavirus spread dataset, which consists of number of confirmed, death, and recovered reported, and 2019-nCoV dataset, which handles the day level information on 2019-nCoV affected cases. Columns latitude longitude and provinces were dropped from our dataset. Time was converted to python format using panda's module. Panda library was used to change the data from wide to long

Exploratory data analysis

We analyzed our datasets with timeseries EDA methods and visualize those data to see the pattern of the outbreak of COVID-19 in selected countries. Here, we present an effort to visualize and analyze data between 22 January 2020 and July 2020 with timeseries plots of selected countries. At the beginning massive number of cases were concentrated in China, later on cases were seen in all over the globe. Now a massive number of cases are reported in US compared to

the rest of the world. As shown in fig 1, fig 2 fig 3 United states top the list on total cases, total deaths and total recoveries.

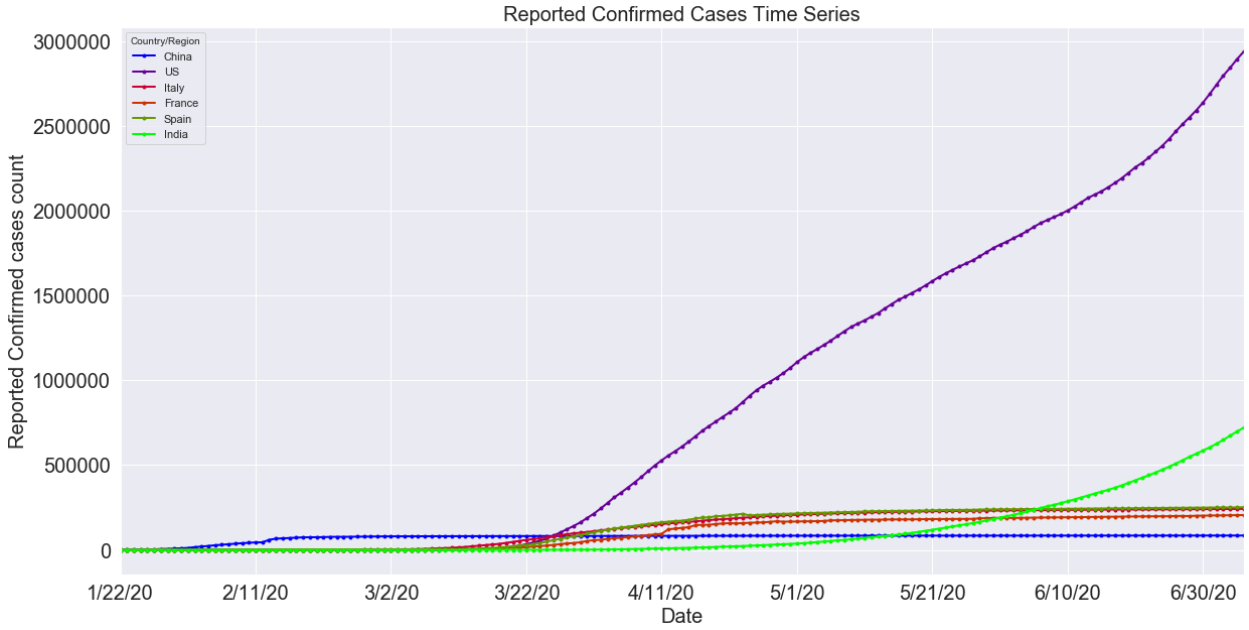

Figure 1 Timeseries plot of Confirmed cases in Selected Countries

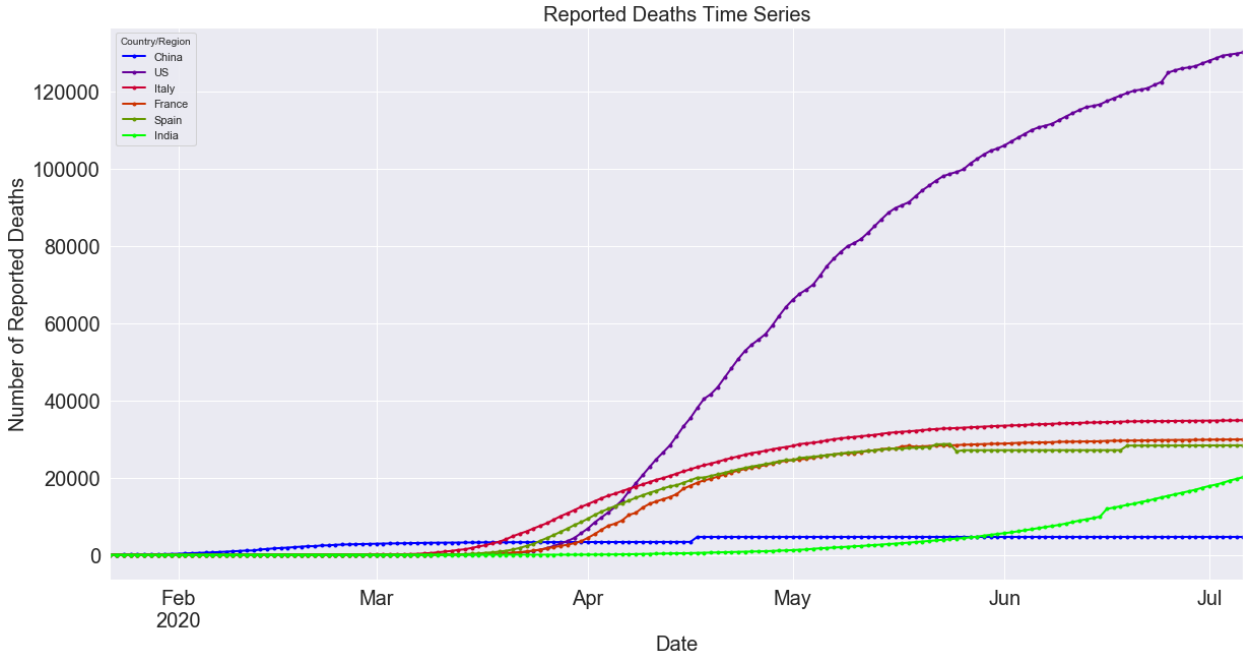

Figure 2 Timeseries plot of reported deaths in Selected Countries

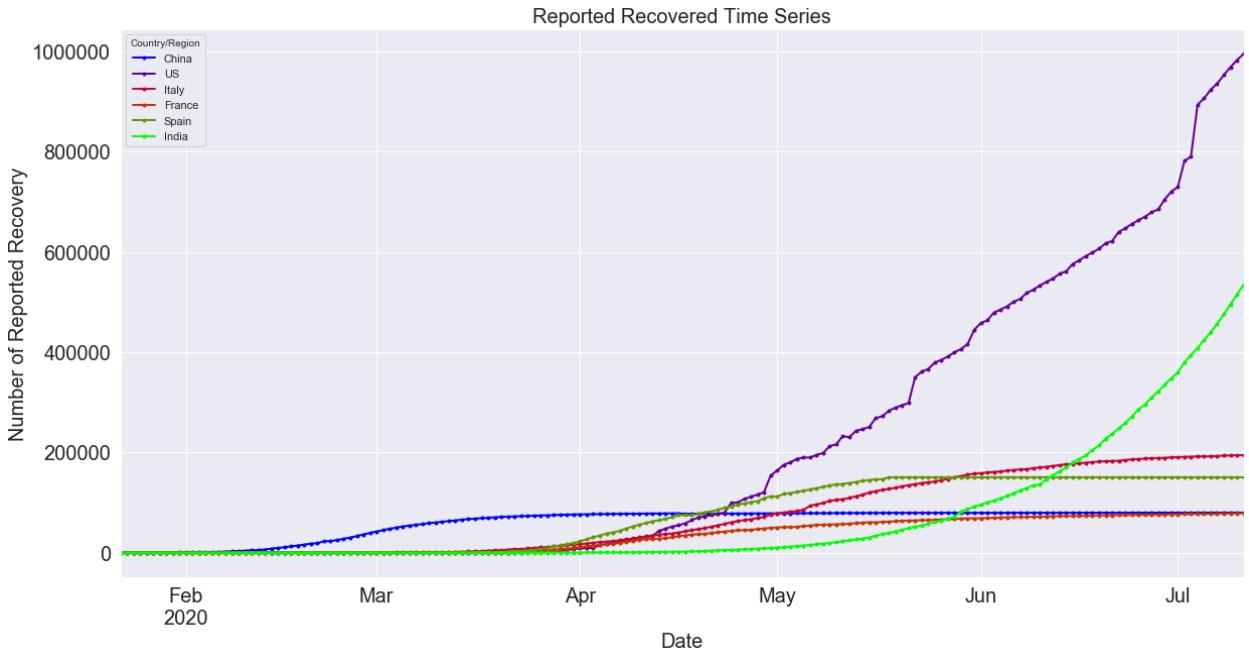

Figure 3 Timeseries plot of Reported Recovery in Selected Countries

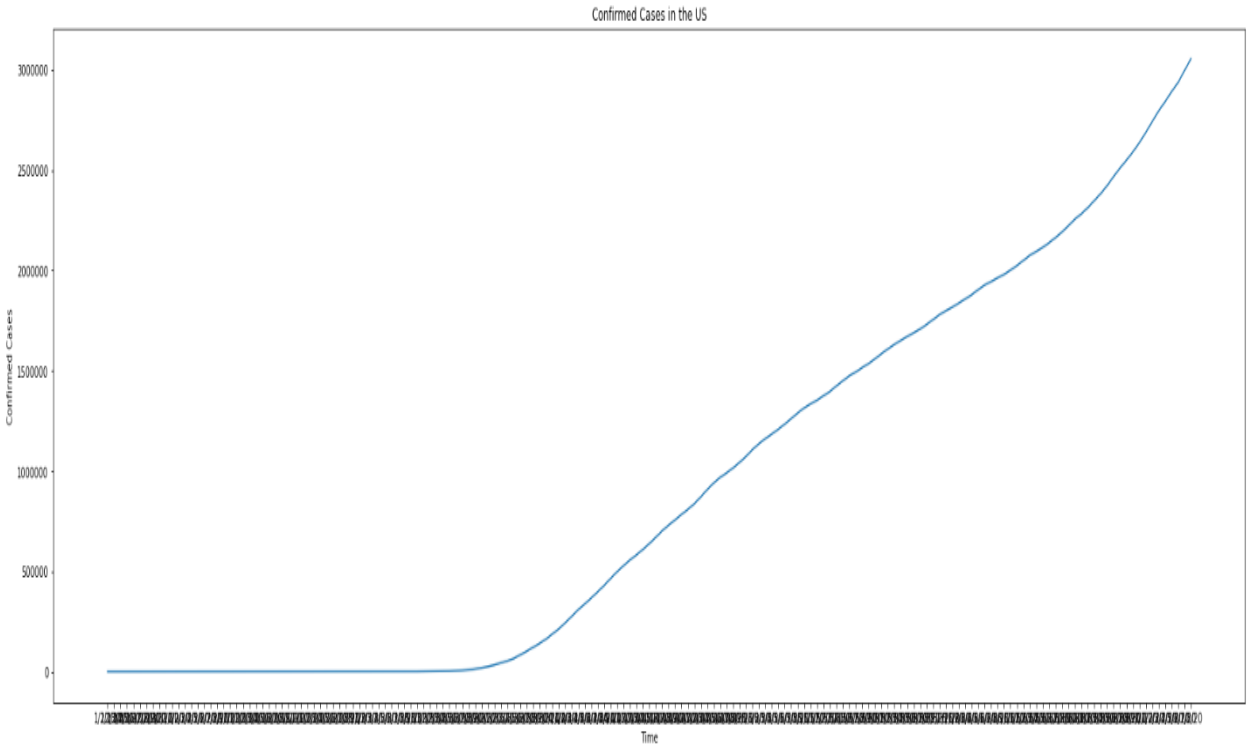

Figure 4 Timeseries Plot for number of cases in United States

Statistical Tests

Augmented Dickey-Fuller test

The augmented dickey-fuller test is a type of statistical test called unit root test. Unit root test helps to find out the trend on time series dataset.
The null hypothesis of the test is that the time series can be represented by a unit root, that it is not stationary (has some time-dependent structure). The alternate hypothesis (rejecting the null hypothesis) is that the time series is stationary.

- **Null Hypothesis (H0)**: If failed to be rejected, it suggests the time series has a unit root, meaning it is non-stationary. It has some time dependent structure.

- **Alternate Hypothesis (H1)**: The null hypothesis is rejected; it suggests the time series does not have a unit root, meaning it is stationary. It does not have time-dependent structure

- **p-value > 0.05**: Fail to reject the null hypothesis (H0), the data has a unit root and is non-stationary.
- **p-value <= 0.05**: Reject the null hypothesis (H0), the data does not have a unit root and is stationary.

Comparing the result of test using statsmodel module of python, test statistic to the critical values, it looks like we would have to fail to reject the null hypothesis that the time series is non-stationary and does have time-dependent structure.

Deep Learning

Deep learning based NN methods has been used for time series prediction problem and different prediction approach has been used for prediction. Among the different deep-learning-based methods, LSTM NN and stacked autoencoders (SAEs) have been reported to have better performance than some commonly used traditional prediction models. In corona virus prediction LSTM NN and SAEs are found to be used to predict the number prevalence of virus and found better performance for LSTM NNs than SAEs. However, which kind of deep neural networks is the most appropriate model to predict corona virus cases remains unsolved. In this project, we apply LSTM NN model for time series prediction of corona virus. The LSTM algorithm is trained on the training set and the model is used to make predictions on the test set. The prediction is compared with the actual values in the test set to evaluate the performance of the trained model.

We performed min/max scaling on the dataset which normalized the data within a certain range of minimum and maximum values. MinMaxScaler class from sklear.preprocessing module is used to scale our data. Next dataset was converted into tensors since Pytorch models are training using

tensors. To convert dataset into tensors our dataset was passed to the constructor of the FloatTensor object. Next, we defined a function called sliding window sequence the function will accept the raw input data and will return a list of tuples. Class LSTM is defined using nn.Module calls of the pytorch library. On the next step an object of the LSTM( ) was created by defining a loss function and the optimizer.

The predictions made by our LSTM are depicted by the orange line depicted in figure 5. Though our algorithm is not too accurate but still been able to capture upward trend for the total number of cases in the coming days.

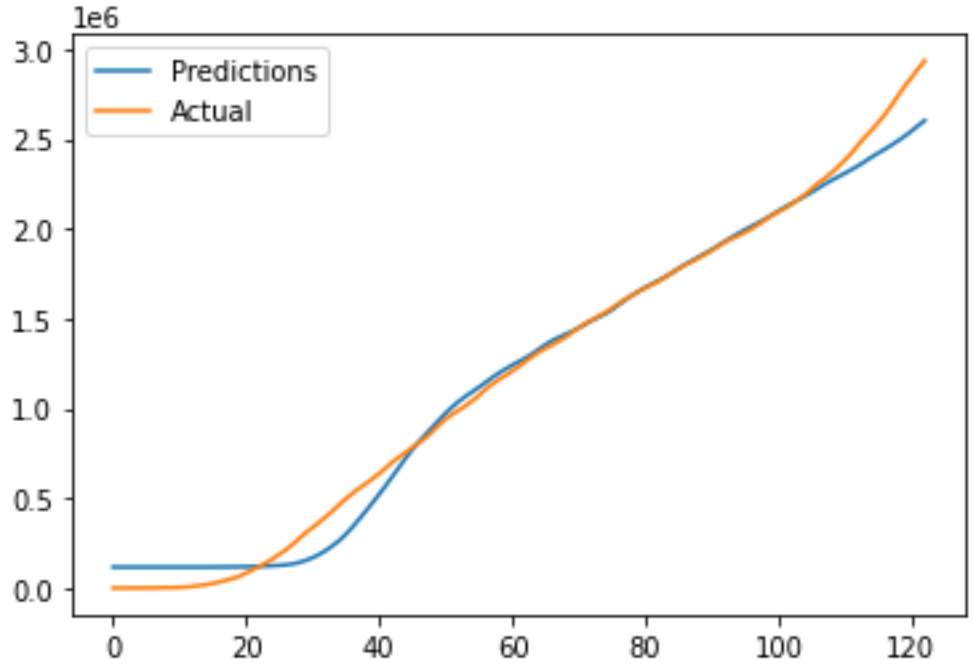

Figure 5 Prediction vs Actual by LSTM model

Conclusions:

In conclusion, the dataset we have used for our experiment 2019 coronavirus dataset (January-July 2020), COVID-19 (nCOV-19) coronavirus spread dataset, and 2019-nCoV dataset can be useful to monitor and predict the emerging outbreaks, such as 2019-nCoV. Such activities can help us to generate and disseminate detailed information to the scientific community, especially in the early stages of an outbreak, when there is a little else available, allowing for independent assessments of key parameters that influence interventions. We observe that deep learning techniques can be used to predict the future time series prediction of COVID-19 allowing health care sectors and health authorities take necessary steps at right time to mitigate the adverse impact of virus.