**Title: Covid-19 Forecasting & X-Ray Image Detection**

**ABSTRACT:**

Coronavirus (COVID-19) pandemic was growing exponentially and has had a significant impact on the whole world. Researchers, technologists, doctors, and other healthcare workers are working day and night on the development of vaccine and medicines to control and treat this virus. With exponentially increasing infections, proportionate fatalities are being reported both from developed and underdeveloped countries. According to WHO coronavirus dashboard, there are total number of 7,62,791,152 confirmed cases till date with 6,897,025 deaths globally. Hence, there is an urgent requirement for conducting academic research on several aspects of this highly contagious disease, to find effective means of containment and treatment of the disease, for now, and in future. We have identified some opportunities for academic research related to COVID-19 and have also provided suggestions to contain, prevent and treat this viral infection. The report concludes that machine learning techniques show great promise in addressing the challenges posed by Covid-19, including accurate forecasting and early detection, which can help to mitigate the virus's spread and impact on society.

**INTRODUCTION:**

The COVID-19 pandemic, also known as the coronavirus pandemic, is an ongoing global pandemic of coronavirus disease 2019 (COVID-19) caused by a severe acute respiratory syndrome. The first ever case was confirmed in the end of year 2019 from Wuhan – a city in China. After a month, an out-break was declared by the World Health Organisation when Public Health Emergency of International Concern on January 30th, 2020, and later a pandemic on March 11th, 2020. As of today, there have been a total of around 762M cases and 6.68M deaths worldwide making it the deadliest pandemic in history. The symptoms vary from undetectable to life-threatening. Already a lot of people were infected with the virus before the out-break was announced, since someone who never develops symptoms can still pass the virus to others. People with underlying health conditions and elderly people were the ones with severe illness. The disease being air borne, it is too easy for the virus to transmit from one person to another, even while breathing in contaminated air. The risk of getting the virus is highest when people are in proximity. but they can be inhaled over longer distances, particularly indoors. A lot of Covid19 vaccines were approved and distributed widely in various countries starting December 2020. The vaccines prevented millions of deaths. However, it was quite difficult since there were many strains of covid-19 developed due to the mutations. Apart from the vaccine, recommended preventive measures were to quarantine the one who’s infected with the virus, wearing of masks in public places, social distancing and improving the ventilation as well as air filtration. A severe social and economic disruption was caused due to the pandemic, leading to the largest global recession. On the bright side, due to less human activities, reduced pollution and green house emission was observed. There were a lot of issues raised of racial and geographic discrimination, health equity, and the balance between public health imperatives and individual rights. This pandemic is the defining global health crisis of our time and the most significant challenge we have faced since World War Two

**DATA:**

In our project, we have used two different datasets which will help in forecasting and diagnosing Covid 19. The forecasting data and the X-Ray Image dataset. We were able to locate both of these datasets through Kaggle.

Forecasting dataset -

The forecasting data, which is an open-access dataset was consolidated by the University of Oxford which updates daily, but we are only using the data from January 2020 to March 2022 for our analysis. Although there are many Covid-19 datasets available online, we decided to go with this specific data because this data is sourced from official data published by governments and ministries of health from countries around the world. Using this dataset, we were able to understand the trajectory of the disease which further can be helpful in anticipating its impact on public health.

The Covid-19 data set is made available to the public by the group "Our World in Data", which in partnership with The University of Oxford has created a reliable repository of datasets about dozens of topics with a focus on major global issues. This dataset uses the most recent official numbers from governments and health ministries worldwide. Population estimates for per-capita metrics are based on the United Nations World Population Prospects. The dataset is updated regularly and includes data on the total number of vaccinations administered, first and second doses administered, daily vaccination rates and population-adjusted coverage for all countries for which data are available (169 countries as of 7 April 2021). It will be maintained as the global vaccination campaign continues to progress. The variables represent all of our main data related to confirmed cases, deaths, hospitalizations, vaccinations and testing, as well as other variables of potential interest from various countries around the world.

X-Ray Image Detection dataset -

The X-Ray Image data is a collection of 5,586 image copies which were collected from various sources including hospitals, medical centres, and public repositories. The images were collected from two different sources: the first set of images is from the National Institutes of Health Clinical Centre, and the second set is from a paediatric hospital in Guangzhou, China. Out of the 5,856 images, 3,883 images are classified as "normal" and 1,273 images are classified as "COVID."

The dataset has been created to aid researchers and machine learning practitioners in developing and evaluating models for the automated detection of COVID-19 in chest X-ray images. Each image in the dataset has been labelled as either COVID-19 positive or negative by a team of expert radiologists. It contains a diverse range of images, including those from different age groups, genders, and ethnicities. The dataset is suitable for training and evaluating machine learning models for automated COVID-19 detection. It can also be used for research purposes, such as developing new algorithms for image segmentation and classification, exploring the impact of different imaging techniques on COVID-19 diagnosis, and investigating the efficacy of different treatment protocols. Overall, the Chest X-Ray Covid Classification Dataset is a valuable resource for the development of automated tools for COVID-19 diagnosis and treatment, as well as for research into the disease and its impact on the human body.

METHOD OF ANALYSIS:

Whenever there is a mention of time series models for forecasting future values with the help of historical data, the most commonly used models come to mind such as ARIMA, Holt-winters, and Prophet. We decided to go with these models and we’ll be comparing them to see which one gives better accuracy. These models take into account historical data, seasonal trends, and other factors to predict future trends in cases, hospitalizations, and deaths.

ARIMA

Prophet

Holt-Winters

Testing set

Pre-processing

Covid-19 Forecasting Data

Training set

Evaluation using RMSE

Trained model

*Fig. 1: Flowchart for analysis methods*

While working with these models, we had faced several issues like there was a lot of noise in the visualizations for most of our models. We fixed it by the Smoothing technique for all our visualizations. While performing Random Forest, the model was not giving accurate measures for RMSE, MSE and accuracy since the data splitting was incorrect. At first, we let the data split randomly which had caused the issue, but after splitting the data date wise, we were able to achieve decent accuracy.

For the X-Ray Image Detection, we have worked on CNN model, Random Forest and VGG-16. We selected these models as our primary goal is to achieve high accuracy in predicting outcomes or classifying data. Therefore, the model that provides the highest accuracy on detecting COVID-19 from X-ray images, has been selected. These models are capable of handling complex data and extracting useful features from the images. They are not complex, can be easily understood and provide more transparency and interpretability. The selection of a model in a machine learning project using X-ray image dataset required careful consideration of several factors, including accuracy, data complexity, model explainability, and the specific task at hand. The utilization of the COVID X-ray image collection for machine learning applications will help in the quick, accurate, and affordable diagnosis of COVID-19, which is critically required during the pandemic. Also, it will ease the burden on testing kits and offer a non-invasive method of disease diagnosis.

During the analysis, we came across challenges like dealing with the quality of data, interpreting the images, class imbalance and it took us a lot of time in execution for VGG-16 model since we had a lot of images. The quality of data or image resolution was affecting the results of our analysis as it was difficult to extract meaningful information. Image interpretation was a complex task because it was challenging to differentiate between Covid-19 changes and other lung conditions which caused similar changes in X-Ray images. The number of Covid-19 cases was much lower compared to the number of non-covid cases which lead to class imbalance. To deal with these challenges, we had to reshape the images.

**REFERENCE:**

[https://www.kaggle.com/Forecasting-dataset](https://www.kaggle.com/datasets/georgesaavedra/covid19-dataset)

[https://www.kaggle.com/X-Ray-Image-dataset](https://www.kaggle.com/code/vivek468/chest-x-ray-pneumonia-classification-keras)

<https://ourworldindata.org/covid-vaccinations>