COVID-19 CHEST X-RAY IMAGE CLASSIFICATION USING DEEP LEARNING

Team\_1066\_Covientist

soumav2, gunther6, mb33, rk14

**ABSTRACT**

Due to the rapid growth of COVID-19 cases around the world, it has become imminent to harness artificial intelligence in order to improve the existing medical diagnosis process related to this disease. Conflicting nature of Covid-19 and Pneumonia symptoms pose difficulty in identifying the actual presence of disease in Flu season. It was found in early studies that patients present abnormalities in chest radiography images that are characteristic of those infected with COVID-19. Therefore, chest X-RAY image classification has emerged as an alternative technique to aid medical diagnosis during pandemic time. However, it is cumbersome to manually detect COVID-19 cases from a set of images thus deep learning techniques has been proposed to build a state-of-the-art tool to enhance the current diagnosis process. We aim to implement a set of deep learning models on online available resources of 6432 images and further strengthen by utilizing data augmentation techniques to provide better generalization of the model during testing phase. Furthermore, we would select the model exhibiting best performance metrics during validation phase in order to make an optimistic outcome in detection of COVID-19 from chest X-RAY images.

**Keywords**

Covid-19, Coronavirus detection, Deep learning, X-ray, Pneumonia, Classification.

# INTRODUCTION

The COVID-19, a viral infection causes severe respiratory illness ranging from common cold to life threating disease like Severe Acute Respiratory Syndrome (SARS) and Middle East Respiratory Syndrome (MERS). People suffering from COVID-19 have moderate respiratory illness that can be cured without any special treatment of antibiotics. However, people facing from medical complications like diabetes, chronic respiratory diseases, and cardiovascular diseases are more likely to suffer from this virus. According to the reports of WHO, common symptoms of COVID-19 are same as that of common flu, which include fever, tiredness, dry cough, and shortness of breath, aches, pains and sore throat. Due to the similar nature of flu symptoms, it is difficult to detect the virus at early stage and this can’t be treated by normal anti-biotics material.

# MOTIVATION

The WHO approved method of testing corona virus are the reverse transmission polymerase chain reaction (RT-PCR) method where the short sequences of DNA or RNA are analyzed and reproduced or amplified. However, there have been unforeseen challenges occurred while following this testing procedure:

1) It has been observed that negative results do not rule out the possibility of a person infected with COVID-19.

2) Limited availability of testing kits and screening workstation created roadblocks to carry out mass COVID-19 screening in every part of the world.

The above scenarios create massive implication to medical professionals and staffs to handle the exponential growth of cases with accurate testing measure and making subsequent medical decision accordingly. Therefore, it has become imminent to come up with an alternative approach to support the existing medical diagnosis process. X-ray imaging is frequently used modality by medical practitioners to assert or to deny the possibility of any bacterial pneumonia infection. The same process can be applied to detect COVID-19 by finding out the anomaly between X-ray images of a Covid-19 infectant and a person suffering from bacteria\viral pneumonia. Easy availability of X-ray machines make it need of the time to use for detection COVID-19 cases in the absence of screening workbenches and kits. However, the biggest challenge is to manually examine each X-ray images and extract the findings lead to enormous time and presence of medical professionals. Thus, it is obvious that a computer-aided method driven by state-of-the-art deep learning models will improve the current COVID-19 diagnosis process with more accurate prediction and less time-consuming simulated tasks.

# METHODOLOGY

## Dataset and pre-processing

For this particular project, we will use chest x-ray images extracted from [Chest X-ray Kaggle dataset](https://www.kaggle.com/prashant268/chest-xray-covid19-pneumonia), a publicly available COVID-19 data repository collected data from various public sources as well as through indirect collection from hospitals and physicians.

Here is a simple breakdown of number of images segregated in train and test folders of the original dataset:

|  |  |  |  |
| --- | --- | --- | --- |
| **Classifiers** | **Chest X-Ray dataset(No. of images)** | | **Percentage of total image count** |
| **Train** | **Test** |
| Normal\Healthy | 1266 | 317 | 25% |
| Pneumonia | 3418 | 855 | 66% |
| COVID-19 | 460 | 116 | 9% |

Fig. 1 - 3 show images of all the three cases, that are considered for this project, such as normal\healthy, pneumonia and COVID-19. As a part of image pre-processing, we will apply i) image augmentation train images in order to bypass any overfitting issues, ii) image normalization to convert into a standard size (mostly 224 x 224) and iii) apply gaussian filtering technique to remove noise before we feed the train image data to neural network input layer. We can apply additional image transformation techniques if that will be deemed necessary in order to improve accuracy and reduce loss.



We may apply further resampling method to split the dataset into three different categories by introducing validation set which will be used during hyperparameter tuning step. In that scenario, we may consider splitting the main dataset into 70% train data, 15% validation data and 15% test data.

Note: Although the number of COVID-19 images are relatively smaller than the other group of images, our initial retrospective research showed that we still have higher volume of images for this classification project compared to some previously conducted research on COVID-19 detection process.

## Model Selection

Selection of the deep learning models for this project will be a two-fold approach. At first, we will test different neural network models known to provide good image classification performance such as MLP and CNN, capturing the accuracy of each model separately and select the best one. Second, we can follow transfer learning approach of training the dataset using pre-tested models for computer vision classification such as VGGNet, ResNet, Inception and EfficientNet, and test different model architecture independently to find out the best precision using such pre-existing model. We may leverage some [pre-trained models](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7476608/table/tbl0005/?report=objectonly) built during previous research work related to COVID-19.

At the final stage, the model that produces best accuracy metrics among all others will be considered as the final Neural Network for this project.

**Note:** The number of epochs is the number of times that the entire training dataset is being processed to the network during training. Some networks are sensitive to the batch size, such as LSTM recurrent neural networks and Convolutional Neural Networks. We will consider tuning hyperparameters such as learning rate and epochs, as well as updating the last neuron network layers to produce better model evaluation outcome for our project.

## Analytical Infrastructure

We will use open-source machine learning library PyTorch to build the predictive model for this image classification project. In order to achieve high accuracy and maximum throughput, we may use cloud platforms such as Azure, AWS or Google Colab in order to harness better computing power to train and re-train the model with sufficient number of epochs, helping to scale and achieve higher accuracy within a short span of time. We may also consider the cloud infrastructure as an alternative to deploy our solution publicly.

## Model Evaluation

Footnotes Performance of the two-fold COVID-19 model algorithms will be evaluated using the common statistical measure confusion matrix and area under the curve. From the confusion matrix we will obtain different metrics such as true positive (TP), true negative (TN), false positive (FP), and false negative (FN) which will be calculated as below:

1) Accuracy = (T N + T P)/(T N + T P + F N + F P)

2) Precision = T P/(T P + F P)

3) Recall = T P/(T P + F N)

4) F1 score = 2(Precision×Recall)/(Precision + Recall)

5) Specificity = T N/(T N + F P)

As far as other Area Under the Curve (AUC) matrix concerned, the final model evaluation we will project the ROC curves and compare the different AUC values to define the model with the best performance.

# PROJECT TIMELINE

The We aim to follow a detailed roadmap for this project as mentioned below:

|  |  |  |  |
| --- | --- | --- | --- |
| **Task** | **Start** | **End** | **Days** |
| Data processing/  transformation/embedding | 4/1/2021 | 4/6/2021 | 6 |
| Model build | 4/7/2021 | 4/10/2021 | 4 |
| Model evaluation | 4/11/2021 | 4/13/2021 | 3 |
| Group Discussion on the individual model results | 4/14/2021 | 4/14/2021 | 1 |
| Finalize hyperparameter tuning methods | 4/15/2021 | 4/16/2021 | 2 |
| Result consolidation and submit project draft | 4/17/2021 | 4/18/2021 | 2 |
| Total # of days taken | 4/1/2021 | 4/18/2021 | 18 |

# CONCLUSION

COVID-19 X-ray image classification is considered as a very recent state-of the-art medical diagnosis process which is why we will encounter an imminent data scarcity issue. Hence, we will perform additional pre-processing and leverage various deep learning methods to overcome the shortage of data and build a sustainable model with reasonable higher accuracy.

Additional note: Some of the approaches documented in this project proposal are hypothetical which may differ from the final implementation.

# REFERENCES

1. Zoabi, Y., Deri-Rozov, S. & Shomron, N. Machine learning-based prediction of COVID-19 diagnosis based on symptoms. *npj Digit. Med.* **4,**3 (2021).
2. Jain, G., Mittal, D., Thakur D. & Mittal M. A deep learning approach to detect COVID-19 coronavirus with X-Ray images. *Biocybern. Biomed. Eng.* **40,** 4 (2020).
3. https://www.kaggle.com/prashant268/chest-xray-covid19-pneumonia
4. https://www.connectedpapers.com/main/fedd928db5fbfa07819caacf01debbccca353a19/Deep-Learning-COVID19-Features-on-CXR-Using-Limited-Training-Data-Sets/graph