# Classification of Chest X-Ray images to distinguish between COVID-19 and Pneumonia
# COMP 6721 Applied Artificial Intelligence - Proposal - By Group D

## A. Problem Statement and Application

COVID-19 is an infectious illness brought on by the coronavirus strain SARS-CoV-2 responsible for the severe acute respiratory syndrome. Patients suffering from COVID-19 can also present with abnormalities on chest X-ray images that are characteristic of infection [1]. According to WHO, the most common diagnosis in severe COVID-19 patients is severe pneumonia [2].

With people all round the world suffering from COVID, it is important to diagnose it correctly by identifying it from numerous similar cases of pneumonia, Tuberculosis, etc. [2]. Hence, it is our model topic to distinguish between the chest X-ray images of COVID and other diseases related to chest as it can be helpful in the early diagnosis of COVID-19.

With our proposed ML and CNN based model, we aim to differentiate the X-ray images into Normal, COVID-19, Pneumonia and few other classes.

The challenges that might arise with our problem is training the model with similar chest X-ray images where pneumonia can be mistaken with COVID. As the dataset consists of images of different format and hence the pre-processing of data would be a challenge.

The goal throughout developing the application is to study the impact of different training models in our application by interchanging the datasets, CNN architectures, hyper parameters etc.

## B. Image Dataset Selection

For the implementation of the model, we have selected 3 datasets from the Kaggle.

| Datasets | Classes | Images | Average |
|---|---|---|---|
| Dataset1 | 4 | 21164 | 3616 |
| Dataset2 | 3 | 6939 | 2313 |
| Dataset3 | 22 | 2482 | 49 |

**Dataset1 :** [5] "COVID-19 Radiography Database" consists of 21,164 images which are categorised into 4 classes namely, COVID Positive, Normal, Non-COVID lung infection, Pneumonia. There are 3616 COVID-19 positive cases along with 10,192 Normal, 6012 Lung Opacity (Non-COVID lung infection) and 1345 Viral Pneumonia images. All the images are in Portable Network Graphics (PNG) file format and resolution are 299299 pixels. Data is collected from Germany medical schools, GitHub, Kaggle, Radiological Society of North America(RSNA) and various other public datasets. [3] [4]

**Dataset2 :** [6] "COVID19, Pneumonia, Normal Chest X-ray Dataset" has 3 classes containing 6939 multi-type images. A total of 1401 samples of COVID-19 were collected using GitHub, the Radiopaedia, Italian Society of Radiology (SIRM) , Figshare data repository websites. Then, 912 augmented images were collected from Mendeley. Finally, 2313 samples of normal and pneumonia cases were obtained from Kaggle. Every class has an average of 2313 images and the images are in (jpg/jpeg) file format. [8] [9] [10] [11]

**Dataset3 :** [7] "XrayBodyParts-fastai" contains 2,482 images with 1,738 train images of 22 categories. With 724 images of the Chest, other parts are represented by 1,014 images of 21 classes, less than 49 images per class in average, with some rare classes having just few images, like Clavicles with 9 images only, and several multi-labels classes with a single image. These are all PNG images. [12]

## C. Possible Methodology

**Gathering Data :** The data is collected from different sources as mentioned in section[Image Dataset Selection]

**Data Pre-Processing :** The data-images are in different forms, so it will require proper pre-processing methods like image enhancement, image rescaling and image normalisation, image augmentatiion etc to remove noise.

**Choosing a model :** CNN architectures as LeNet, ResNet and VGG models will be used to implement our project.

**Model Training :** A Deep Learning model needs to be trained for several epochs and the loss is back propagated to update the weights in the model. The models chosen are explored further using the learning parameters and optimization techniques such as SGD.

**Model Evaluation :** In order to assess our classification model, metrics like Accuracy, precision, recall, and F1 score will be used. Area under of ROC Curve can be plotted to learn the performance of the model at all thresholds. Confusion matrix can also be plotted to determine the performance of the model.

**Parameter Tuning :** Learning parameters will affect the deep learning models. Some of them are to be fine tuned to converge the loss when encountered in the model. Optimization of our CNN model can be analysed by using the parameters like learning rate, loss function and batch size.

**Transfer Learning :** After training our 9 models, we will train 2 models with transfer learning using the "Fine-Tuning" method in PyTorch

**Prediction :** Data visualisation will be done using the activation methods with filters to understand input patterns and T-SNE tool to visualize high-dimensional data. .

CVPR #*****

## D. Gantt Chart

| TASK TITLE | START TIME | DUE TIME | WEEK 1 | WEEK 2 | WEEK 3 | WEEK 4 | WEEK 5 | WEEK 6 | WEEK 7 | WEEK 8 | WEEK 9 | WEEK 10 | WEEK 11 | WEEK 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Project Announcement | Week 1 | - | | | | | | | | | | | | |
| Project subject Finalization | Week 1 | Week 5 | | | | | | | | | | | | |
| Appropriate Dataset selection | Week 3 | Week 5 | | | | | | | | | | | | |
| Proposal Submission | Week 4 | Week 5 | | | | | | | | | | | | |
| Proposal Revision & Re-submission | Week 5 | Week 6 | | | | | | | | | | | | |
| Model building | Week 6 | Week 7 | | | | | | | | | | | | |
| Data pre-prosessing | Week 6 | Week 8 | | | | | | | | | | | | |
| Project Progress Reporting | Week 8 | Week 9 | | | | | | | | | | | | |
| Model training | Week 8 | Week 10 | | | | | | | | | | | | |
| Model Optimization | Week 9 | Week 10 | | | | | | | | | | | | |
| Project Revision | Week 10 | Week 12 | | | | | | | | | | | | |
| Evaluation Metrics | Week 11 | Week 11 | | | | | | | | | | | | |
| Final Reporting | Week 11 | Week 12 | | | | | | | | | | | | |
| Final Project Submission | Week 12 | Week 12 | | | | | | | | | | | | |

Figure 1. Gantt Chart

**Deliverable Contents :**

Part 1: Project Proposal
Part 2: Progress Report
Part 3: Working Model
Part 4: Trained and Optimized Models
Part 5: Analysis
Part 6: Final Report and Presentation

CVPR #*****

## E. Bibliography

## References

[1] Ng Ming-Yen, et al. *Imaging profile of the COVID-19 infection: radiologic findings and literature review.* Radiology: Cardiothoracic Imaging 2.1 (2020). 1

[2] *https://www.who.int/docs/default-source/coronaviruse/clinical-management-of-novel-cov.pdf*. 1

[3] M.E.H. Chowdhury, T. Rahman, A. Khandakar, R. Mazhar, M.A. Kadir, Z.B. Mahbub, K.R. Islam, M.S. Khan, A. Iqbal, N. Al-Emadi, M.B.I. Reaz, M. T. Islam *"Can AI help in screening Viral and COVID-19 pneumonia?"* IEEE Access, Vol. 8, 2020, pp. 132665 - 132676. 1

[4] Rahman, T., Khandakar, A., Qiblawey, Y., Tahir, A., Kiranyaz, S., Kashem, S.B.A., Islam, M.T., Maadeed, S.A., Zughaier, S.M., Khan, M.S. and Chowdhury *Exploring the Effect of Image Enhancement Techniques on COVID-19 Detection using Chest X-ray Images.* arXiv preprint arXiv:2012.02238, M.E., 2020. 1

[5] Dataset 1: *https://www.kaggle.com/datasets/tawsifurrahman/covid19-radiography-database*. 1

[6] Dataset 2: *https://www.kaggle.com/datasets/amanullahasraf/covid19-pneumonia-normal-chest-xray-pa-dataset*. 1

[7] Dataset 3: *https://www.kaggle.com/code/alexanderyyy/xraybodyparts-fastai/notebook*. 1

[8] *https://github.com/agchung* 1

[9] *https://radiopaedia.org/?lang=gb* 1

[10] *https://www.sirm.org/en/category/articles/covid-19-database/* 1

[11] *https://figshare.com/articles/dataset/COVID-19ChestX-RayImageRepository/12580328/2* 1

[12] *https://www.kaggle.com/datasets/ibombonato/xray-body-images-in-png-unifesp-competion* 1