### Dataset Description

Write in your own words (1-2 sentences) what this dataset is about.

**a) Does your description specify features and instances?**

The dataset contains 535 pictures of chest x-ray imaging, with 31 different features including 'lymphocyte\_count', 'neutrophil\_count', etc., 25 different labels including different medical diseases such as 'Influenza', 'Klebsiella'.

**b) Do you describe the source and reliability of the data?**

The data is collected from public sources as well as indirect collection through hospitals and physicians. Some of the data is also scraped from websites run by non-profit organisations, which can be contributed by the public.

### Project Title

Hastening the Detection of COVID-19 in Chest X-Ray Imaging of Humans via Machine Learning

### Motivation

Explain why this project is interesting and important.

**a) Does your motivation clearly describe a problem?**

COVID-19 is a prevalent issue which has affected millions of lives and will continue doing so without further intervention.

**b) Does it justify the problem’s significance? What are the benefits of addressing this problem? Who benefits from solving it?**

It is justified as addressing this problem can potentially save millions of lives globally. With a more rapid and accurate diagnosis of a patient’s medical status, doctors can provide quicker assistance to those who have COVID-19. As a result, patients diagnosed with COVID-19 can recover quicker as their diagnosis can be done much faster.

### 

### Problem Statement

A statement of the problem, issue, or task that you’re interested in studying. In particular, try to formulate the key questions (2 to 4 questions is probably a good number) that you will answer in the project.

As this proposal document is self-contained, you should restate your project topic and domain. The information you provided in the submission system is to help with appropriate review assignments.

**a) Does the proposal outline a problem statement, issue, or task that the team is interested in studying?**

Due to the complications that arise from COVID-19, there is a need for quicker diagnosis of patients who potentially have it. This would greatly assist doctors. Current diagnostic methods, such as PCR tests, can be time-consuming, costly, and sometimes limited by supply chain constraints. Chest X-ray imaging presents a potential alternative for COVID-19 detection due to its widespread availability and relatively low cost. However, distinguishing COVID-19 from other respiratory conditions using X-ray images poses significant challenges, requiring the development of robust machine learning models capable of high accuracy and reliability.

**b) Does it formulate a few (2 to 4) questions that the team proposes to address?**

1. How helpful are X-ray images in the differentiation between COVID-19 related cases and other diseases that have symptoms similar to those of COVID-19?
2. From the X-ray scan what key characteristics stand out most as being associated with a COVID-19 infection?
3. Regarding the effectiveness of the model, which of the ML algorithms can provide the highest TPR value and the lowest FNR rate, this is extremely important since it has to capture as many infected persons as possible for containment of the virus. The second optimization criterion is False Negative Rate (FNR), which should be minimised in order not to miss any people that might have diseases.
4. What are the limitations in using chest X-ray imaging for the detection of COVID-19, and are there any ways to enhance the process? Are there any additional solutions that will help to improve the rate of COVID-19 diagnosis?

### 

### General Approach

A high-level description of the general approach you’ll use to address the questions. Sketch out what evidence you are planning to gather (e.g. how you can answer the questions through experiments on data). Survey on the current progress on the problem/task.

**a) Does the proposal contain a high-level draft description of the general approach proposed to address the questions?**

**b) Does it include preliminary plans for evaluation, data gathering? I.e., how the team plans to answer the questions through experiments on data.**

* The current dataset contains 535 entries of data.
* Each entry of data contains

1. An image which is stored as a np.ndarray object which is a greyscale X-ray of the patient’s chest (Varying in angle taken, 2 dimensional)
2. A pd.DataFrame which contains the medical report of the specific patient (Not all aspects of the medical report can be used however, as data is not in a usable state, i.e. NaN values)
3. A pd.DataFrame which contains 25 different lab tests, as well as results (float values, 0.0 for negative and 1.0 for positive)

* Points 1 and 2 may be used separately, or in conjunction in order to predict 3.
* Improvements to the model would be made by obtaining more data to tackle the skewness of the data

**Field of models in consideration:** KNN, SVM and NN/CNN

* To answer **PB.b.1** (Problem statement part b, question 1), we would be trying to predict using different models, using accuracy and loss as a metric to determine the best model, choosing the model with the highest accuracy and lowest loss.
* To answer **PB.b.2**, we would first identify the most accurate model using the metric defined earlier. From there, we would extract the feature that has the highest influence on the predicted value. Influence in this scenario may be the correlation between the feature with the outcome, or weight of the feature, depending on the model
* To answer **PB.b.3**, we would need to assess the recall (sensitivity) for different ML models used for image classification.
* To answer **PB.b.4**. During the data collection/processing phase, we may need to perform some data augmentation in order to gather more normal data points to tackle the skewed dataset. Common augmentation techniques include:

1. Geometric transformations - Translation, Rotation, Scaling, Stretching and Flipping
2. Colour Space transformation - RGB channels, contrast, brightness and saturation
3. Kernel Filters + Random erasing

* Complementary solutions to data gathering: using Generative adversarial networks (GANs) to generate new images. This method does not require existing data to generate synthetic data.

### Evaluation

Include how you will evaluate your project. Propose what your team thinks is a satisfactory project outcome (C grade) and an excellent project outcome (A grade). Remember that performance is secondary to analysis and understanding.

**a) Does the proposal contain a high-level draft description of the general approach proposed to address the questions?** (Yes, no need to write for this)

**b) Does it include preliminary plans for evaluation, data gathering? I.e., how the team plans to answer the questions through experiments on data.**

* A satisfactory project outcome would be:
* Clear explanation of framework behind models (i.e. choosing of hyper parameters)
* Differences in performances of models can be explained
* Clear limitations and benefits of utilising each model
* Potential Strategies to make up/bolster the model’s weakness/strengths
* Acceptable accuracy
* An excellent project outcome would be:
* Analysis on scalability of model with larger sets of data (Theoretical)
* Analysis of retraining/updating of model to maintain its usability in the medical field (Theoretical)
* Draw links as to how these models can help the medical field
* High recall, as predicting falsely when COVID-19 is present is the worst case scenario
* Striking a balance between precision and recall by using F1 score

### Resources

A list of resources you have/need to conduct the project. This includes additional reading, software, datasets, code(github link), etc., beyond your chosen dataset. Are these resources public? How are you planning to get these resources?

**a) Does the proposal give a short list of resources the team plans to use to execute the project (inclusive of readings, software, datasets, etc)?**

**b) Does the team describe any strategy for getting the resources?**

* The primary dataset now contains only 535 entries. To enhance the dataset size, we will be sourcing large amounts of data from publicly available datasets from data sources such as

1. <https://www.kaggle.com/datasets/prashant268/chest-xray-covid19-pneumonia>
2. [NIH ChestX-ray14 Dataset](https://nihcc.app.box.com/v/ChestXray-NIHCC/folder/174256157515) (To train for pneumonia, many negatives labelled as ‘No Finding’ to mitigate the skewed dataset we currently have. Its size is significantly larger than our current dataset 100000 compared to 535 so it can be used to help us train a highly complex model for our task using the negatives)

* Data augmentation library: <https://imgaug.readthedocs.io/en/latest/>
* Existing image detection

Software wise:

* Utilising Python libraries such as Pandas, NumPy for data manipulation and Matplotlib/Seaborn for visualisation.
* Sci-kit learn for baseline models, preprocessing, and performance metrics.
* Pytorch and its deep learning libraries for building and training of more complex/ advance Convolutional Neural Networks (CNN) (Deep learning covered in week 10)

### 

### Scheduling

A schedule of work indicating the dates by which you plan to complete components of the project. Make sure the schedule is plausible.

You may find that a table format with the remaining weeks of the course helpful to describe this goal.

**a) A schedule indicating dates by which the team plans to complete the project components?**

Data Exploration - 22/9/2024 (7 days from now)

Data Preprocessing - 6/10/2024 (14 days)

Model Selection and Training - 20/10/2024 (14 days)

Model Evaluation and Tuning - 3/11/2024 (14 days)

* Evaluation of model performance, increasing/decreasing model complexity

Model Optimization - 17/11/2024 (14 days)

* Final adjustments to improve model through techniques such as bagging and ensembling to prevent overfitting and to reduce variance in model performance

Final Presentation - 24/11/2024 (7 days)

**b) An assignment of the team members to the deliverables (inclusive of peer reviewing duties)?**

At every junction, all (or any) team member(s) can contribute by working on their own branches. At the end of every scheduled date for a component, we will review and compile all of our ideas, and merge them into one. This way, we have more brain power to think through the intricacies of the dataset, more creativity in data exploration, ability to learn about different machine learning models and split the use of them on the dataset, etc. For any team members that have great ideas in any component, then they will be given more credit for that component.

**c) Is the schedule feasible given the timeline, expertise and load of the team members?**

Yes, it is feasible. There are 2 weeks between every major component of the project. Since the project is about digital image processing, having 2 weeks is sufficient time to read and learn more about machine learning methods that deal with digital image processing. The code will be finished by 17th November, and we will have 2 weeks till submission date, so it is best if we complete the final presentation by 24th November, with 1 week left to spare for any finishing touches.

**d) For projects with possibly too large a data source (e.g., Kaggle projects), does the team propose a way to scope the data or problem accordingly to make it feasible?**