**Capstone Three - Project Proposal**

**Problem statement formation**

How can researchers develop algorithms using CT scans to effectively identify if a person is infected by SARS-COV-2?

**Context**

In December 2019, a severe outbreak of pneumonia that originated from China spread rapidly to other countries. The virus was identified and named as Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) and the outbreak was named coronavirus disease (COVID-19).

COVID-19 displayed symptoms such as fever, cold, headache and fatigue, and many countries used Reverse Transcription Polymerase Chain Reaction (RT-PCR) as a confirmation of the infection. However, further analysis showed that the sensitivity of chest CT was greater than that of RT-PCR (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7233365/>) and the scan can be used effectively when patients display symptoms compatible with COVID-19 but the RT-PCR is negative. This amplifies the need for the scan to be as effective as possible to distinguish patients with COVID-19 vs patients without COVID-19.

**Criteria for success**

The metric of success for researchers will be to achieve an accuracy more than 95% on the detection.

**Scope of solution space**

The focus of the project will be to identify a model that can best differentiate patients with COVID-19 vs patients without COVID-19.

**Constraints**

The CT scan doesn’t take into concern if the patient already had pre-existing habits such as smoking or pre-existing health conditions.

**Stakeholders**

Government agencies and researchers.

**Data sources**

The dataset that will be used for this project will be the [SARS-COV-2 Ct-Scan Dataset](https://www.kaggle.com/datasets/plameneduardo/sarscov2-ctscan-dataset) from Kaggle. The data was collected from real patients in hospitals from Sao Paulo, Brazil. Among the 2482 CT scans available, 1252 scans are positive for SARS-CoV-2 infection (COVID-19) and 1230 CT scans are not positive for SARS-CoV-2 infection.

**Problem approach**

1. Perform Data Wrangling
   1. Import images and convert them to arrays
   2. Create labels for COVID-19 images vs non COVID-19 images.
2. Exploratory Data Analysis
   1. Visualize the CT scans
   2. Differentiate CT scans that have COVID-19 infection present and those that do not.
3. Pre-processing
   1. Split the training and test data.
4. Modelling
   1. Perform classification using model architecture such as CNN
   2. Use accuracy as the target metric when predicting on test set.
   3. Assess using ROC curve and AUC

**Deliverables**

* A GitHub repo containing the work completed for each step of the project.
  1. A slide deck
  2. A project report