# Problem Approach

* Labelled set: 12,000 CXR Images
  + 10,000 training images
  + 2,000 test images
* Prediction Set: 3,000 CXR Images
* CXR images > feature detection > nodules & masses

### Task deliverables

1. Adapted code w/ comments explaining changes.
2. Make recommendations on how best to use the unlabelled dataset and has told you there is budget to label a further 300 CXRs
   * Which images can help the model learn best? Where are the problems?
   * Low confidence predictions?
   * Provide 300 UUIDs for labelling.
3. PowerPoint presentation detailing: model training techniques used, performance analysis and explanations for tagging recommendation. This should be included as a pre-read but also prepared as a short verbal presentation for the next stage of the interview. The presentation should be no more than 6 slides long and take no more than 10 minutes to deliver.

## work process recommendation

**Part 1 – Model Training & Optimisation:**

1. Using the original python notebook and datasets provided, adapt the code to train a model which identifies the ‘Nodule/Mass’ class provided.
2. Validate the initial model you have built and record this as a baseline performance
3. Optimise the performance of your model, you may change any aspect of the code and use any data provided but **cannot change the model selected or edit the test dataset.** (This should not be an exhaustive process, **1 or 2 well-selected techniques** would be fine).
4. Record the results from your best performing model, as well as an explanation of what you have implemented.

**Part 2 – Data Selection:**

1. Using your trained model and any performance analysis you think might be useful, identify the 300 CXR uuids that you would recommend for tagging.
2. Provide uuids in a json file.
3. Give appropriate rationale for your choice of data (within presentation).

## assessment

* Code clarity and readability.
* Quality of documentation.
* Selection and implementation of techniques to improve performance is essential.
  + Final model performance is less important than the appropriateness of techniques selected and the quality of your implementation.
* For your presentation, you will be assessed on your ability to communicate complex concepts to a broad audience and the ability to deliver a succinct and effective conclusion.

## improved model

### Data Augmentation

#### HOrizontal flipping

I applied the horizontal flipping torchvision class with the standard probability of p=0.5.

#### brightness variation

Chart, histogram

Description automatically generated

Here I analysed the brightness variation in the training dataset. It had a mean of 0.5407 and a standard deviation of 0.1044. I used the transform.ColorJitter class to vary the brightness by a random amount limited to ± 2 standard deviations.

### imbalanced classes

The major issue with the dataset is the imbalanced classes. To aid prediction here I will investigate, oversampling and changing the loss function to give equal weight to the Nodule/Mass present and not present classes.

#### loss function

#### optimiser

Let's say you have L more times of the abundant class than rare class. for stochastic gradient descent, take int(a\*L) separate steps each time you encounter training data from the rare class. (a will need to be cross-validated).

### tuning prediction threshold

I shouldn’t just use 0.5 as the decision boundary, I should evaluate different thresholds and see how they affect the evaluation metrics (not using 0.5)

Business relevance: What is worse, missing a node/module? Or falsely suggesting one.

False Positive better than False Negative

Recall should be prioritised. Precision should be deprioritised.

## model analysis

* **Confusion Matrix**: A breakdown of predictions into a table showing correct predictions (the diagonal) and the types of incorrect predictions made (what classes incorrect predictions were assigned).
* **Precision**: A measure of a classifiers exactness.
* **Recall**: A measure of a classifiers completeness
* **F1 Score (or F-score)**: A weighted average of precision and recall.
* **ROC curve (AUC**?)

## Q&A

* What other changes would you have made given extra time?
  + Investigated different models.
  + Used pretrained networks.
  + Investigated the effect of batch size.
  + Focused training on hard examples.
  + Investigated ensemble approaches
  + Investigated more types of data augmentation. (tuned hyperparameters for brightness)
  + Exponentially decreasing learning rate