The Machine Learning task I chose for this assignment is Classfication. I checked the class distribution and discovered that the data imbalance was very severe for this dataset, where most of the pixels have the label 3 which menas background. Therefore, in the data preprocess, I first cropped all the images to crop out the large amount of background on the edges of the lung scan, then I filtered out images with equal or more than 95% of label 3 in the pixels. After this, my dummy classifier accuracy drop from 81% to 57%. and in total, I deleted 200 images from the training dataset.

For data preparation, to make sure the images and labels dataset work for further prediction, decoding the labels from one-hot encoding to normal labels and also reshaping both images and labels is necessary. There are two functions in utils.py to help simplify this process.

For model selection, after doing train-test split, I tested the accuracy for prediction.

Dummy Classifier Accuracy: 0.57

Baseline Logistic Regression Accuracy: 0.93

Random Forest Accuracy: 0.95

I also tried to use the cropped + filtered images for training and use the original(non-cropped) images for testing the accuracy,

Dummy Classifier Accuracy: 0.84

Baseline LogisticRegression Accuracy: 0.65

Random Forest Accuracy: 0.69

Thus, I chose Random Forest to be my final model, and also limit its maximum depth to increase speed and avoid overfitting. RandomForest is a choice which generates better prediction than the baseline, and at the same time, it is faster compares to other models such as KNN or SVM, which may take much longer time to process and make the prediction.

Although I cropped out some background information, there are still limitations during this process:

1.In my quality report, after I cropped out the background information, some of the other labels may be cropped out in small amount.

2.The imbalance was not completely resolved because this is a multiclass classification, it is aparaent that there are still imbalance between other classes and the amount of background is still high.

3.Because I deleted all images with background more than 95%, this means that some important information from other classes may be missed, and also the training data size drop from 810 to 650 approximately.

4.I also try to make prediction and measure the accuracy on non-cropped original images, however, the accuracy for random forest is not very high, about 70%. Therefore, I think there should be some other models that may generate better result than random forest.

Also, to think about the distribution of this multiclass label dataset, classification task may not be the best machine learning task the resolve this.