Pulmonary Lung Diseases Classification Using Haar-Like Features and Multi-Layer Perceptron

Problems Identified:

1. Dataset contains high resolution images over 1 lac among different folders. Reading images with correct labels was a problem especially if you are working on a non-gpu based environment. It can be highly time consuming.

2. Data preparation was a hectic task due to the size of the dataset.

3. The images containing *No Finding* class were more than other classes together, which leads to biasness in the learning process of model.

4. The dataset provided is a *multilabel classification* problem. It can be solved easily using Deep Learning approach but in basic Machine Learning It was my first time exploring such scenario. We have seen such datasets like the one provided in Bioinformatics for Gene Ontology Classification.

5. The quality of X-ray images is not consistent.

6. *Hernia* Diseases had the lowest labelled images in this dataset. This makes it hard for classifier to learn and predict such labels.

Proposed Methodology:

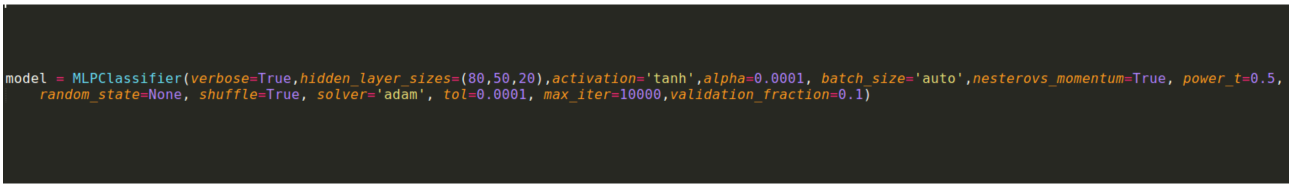
1. The classifier I have used is *Multi Layer Perceptron* (MLP) Network using Sklearn library.

a) It works better on large datasets.

b) It can avoid overfitting, and produces better results in case of unbalanced labeled dataset, provided you have the knowledge of tuning hyperparameters.

c) Multi Layer Perceptron works better for multi-label problem.

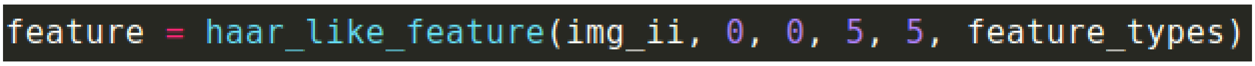
d) It is a similar implementation to Basic Neural network.



```
model = MLPClassifier(verbose=True,hidden_layer_sizes=(80,50,20),activation='tanh',alpha=0.0001, batch_size='auto',nesterovs_momentum=True, power_t=0.5,
    random_state=None, shuffle=True, solver='adam', tol=0.0001, max_iter=10000,validation_fraction=0.1)
```

2. I implemented Multi Label Classification using MultiLabelBinarizer, LabelPowerset approach. *MultiLabelBinarizer* converts labels into a single matrix. Number of columns shows the number of classes and 0,1 value will show the presence or absence of that label in a particular datapoint row. *Label Powerset* on the other hand is a problem transformation approach for multi-label classification that transforms a multi-label problem to a multi-class problem with assigning 1 multi-class classifier trained on all unique label combinations found in the training data.

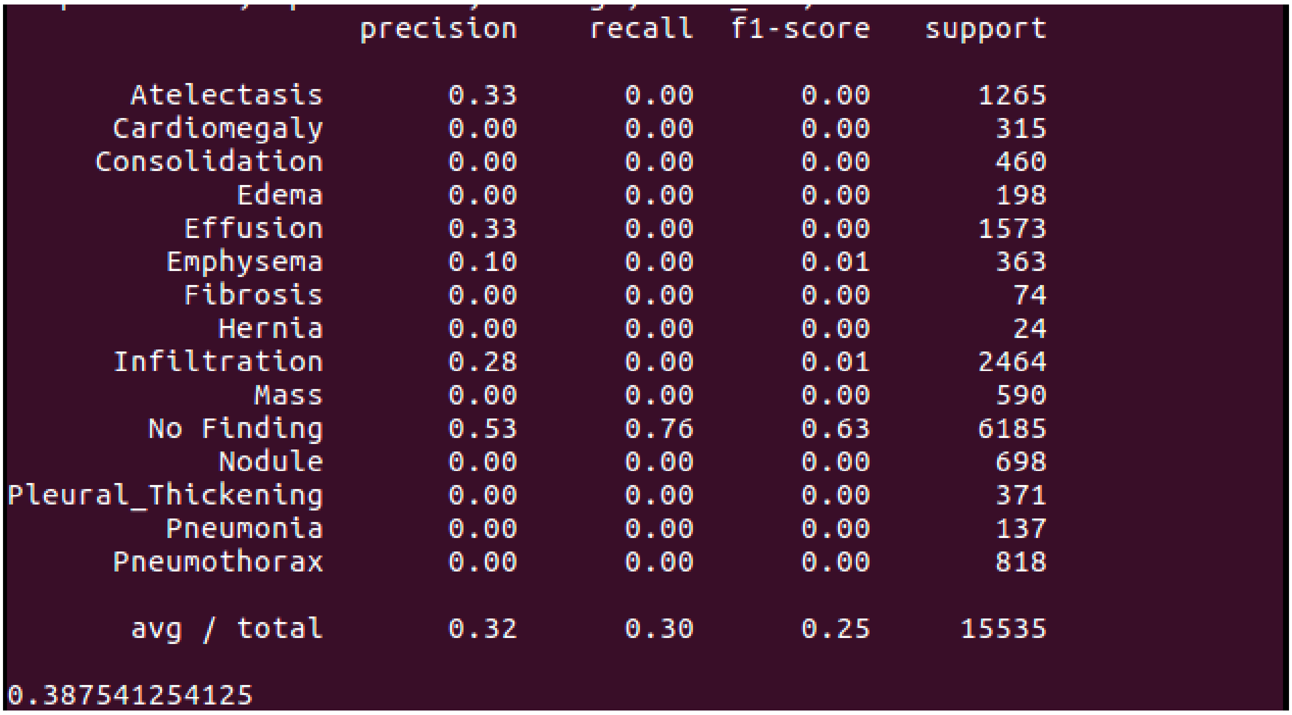
3. Haar-Like Feature is used for the feature vector per image. The feature vector size is approx 200 and the extraction of features on small window is very fast compared to other known feature extraction methods.



```
feature = haar_like_feature(img_ii, 0, 0, 5, 5, feature_types)
```

4. The dataset was splitted into 100,000 training and 15,535 testing images in final evaluation.

5. The accuracy achieved was 38%. *No Finding* label is still affecting the classification process but the classifier is now able to predict more than one class.



```
                  precision    recall  f1-score   support

      Atelectasis       0.33      0.00      0.00      1265
     Cardiomegaly       0.00      0.00      0.00       315
    Consolidation       0.00      0.00      0.00       460
            Edema       0.00      0.00      0.00       198
         Effusion       0.33      0.00      0.00      1573
        Emphysema       0.10      0.00      0.01       363
         Fibrosis       0.00      0.00      0.00        74
           Hernia       0.00      0.00      0.00        24
     Infiltration       0.28      0.00      0.01      2464
             Mass       0.00      0.00      0.00       590
       No Finding       0.53      0.76      0.63      6185
           Nodule       0.00      0.00      0.00       698
Pleural_Thickening       0.00      0.00      0.00       371
        Pneumonia       0.00      0.00      0.00       137
     Pneumothorax       0.00      0.00      0.00       818

      avg / total       0.32      0.30      0.25     15535

0.387541254125
```