Case Study: Detecting Pneumoconiosis

Analytics for Engineers

*Developed by: Ramasubramanian (Ramsu) Sundararajan[[1]](#footnote-1)[[2]](#footnote-2), Horace Xu, Pavan Annangi, Xiaodong Tao*

# Background

A leading hospital wishes to develop a screening program for coal miners, in order to facilitate early detection of Pneumoconiosis. The standard detection procedure involves taking a chest x-ray and examining it for abnormalities that indicate the onset of Pneumoconiosis. A typical doctor’s report divides each lung into three zones (upper, middle and lower) and labels them as normal/abnormal. However, due to the lack of trained doctors with expertise and the large number of patients to be screened, they have requested GE to develop a computer-aided detection system.

A team of image analysts have already developed algorithms to segment the lung and divide it into three zones. They have done this for a set of images where the doctor’s labeling for the lung zones is known, and characterized each lung zone in terms of a set of features. Each patient is identified by a unique patient number. The feature data for the various lung zones for each patient, along with the zone label (0=Normal, 1=Abnormal) is given in the attached Excel workbook.

# Assignment

You are required to build an automated screening program that will identify whether a particular patient is abnormal or not. You will need to ensure that you identify as many of the abnormal cases as possible, while keeping the false positives low. Use leave-one-patient-out cross-validation to determine the correct model and report the results in terms of average performance across cross-validation samples.

1. Let be the patients for whom data is provided, where are the patients who have been diagnosed with Pneumoconiosis and are those haven’t been diagnosed with the disease.
2. For each patient , let the vectors represent the feature vectors extracted for the 6 lung zones (Right Upper, Right Middle, Right Lower, Left Upper, Left Middle, Left Lower), and let represent the zone-level labels (1=Pneumoconiosis, 0=healthy). Note that, even if one of the zones show evidence of Pneumoconiosis, the patient is diagnosed as having the disease. In other words, the patient label .
3. For each patient :
   1. Build a model (or a set of models) on the data for . This can be a single model that uses all the data across all 6 zones, or individual zone-level models whose predictions are then combined.
   2. Predict the label for . Part of the challenge will be to figure out whether to predict individual labels for each   and then combine them in some fashion, or to use some method that predicts a single patient-level label on the basis of all the feature vectors in one go. Let the predicted label be .
   3. Steps a and b are to be repeated for each patient; therefore, the predicted labels   will each be a prediction from a different model (with respect to which that patient is an unseen sample), and any aggregate performance metric that compares them with can be seen as a measure of generalization ability (i.e., performance on unseen examples).
4. Now, compare the actual labels with predicted labels for these patients as obtained above. You will need to report 3 measures:

# Feature description

Once a region of interest (lung zone) is segmented, it is characterized in terms of a set of features. We extract two types of features in order to describe each region of interest. These are described below:

1. **Intensity based** We extract a set of 6 features based on the histogram of intensity values – mean, standard deviation, skewness, kurtosis, energy and entropy. Apart from calculating these on the original ROI, we also extract these features after applying a difference filter on the image for the purpose of local enhancement. If denotes the image gray value at , the first and second order filters are defined as:

where is the difference scale and is the orientation at which the difference is computed. and represent the first order difference while represent the second order difference. We use the first and second order difference filter bank with given orientations and given scale . We can calculate 6 intensity-based features (mean, variance, skewness, kurtosis, energy, entropy) for each filtered image, along with the same features for the raw image without filtering, amounting to a total of 222 features. A subset of 34 features from this set has been provided in the attached data sheet. These features are labeled with the prefix .

1. **Co-occurrence matrix based**: We also extract a set of 5 features based on the gray level co-occurrence matrix computed for the ROI, namely energy, entropy, local homogeneity, correlation and inertia. The co-occurrence matrix allows us to capture the level of similarity and dissimilarity among adjacent pixels in an ROI. Thus, an ROI with an opacity will contain adjacent pixels with similarly high intensities, whereas a normal ROI will not contain such adjacent pixels. Computing these features for various orientations captures this information for various types of adjacency. A subset of 5 of out of 25 such features has been provided in the attached data sheet. These features are labeled with the prefix .

Thus, a total of 39 features for each lung zone has been provided in the attached Excel spreadsheet. The first column in each worksheet (one sheet per zone) gives the patient number, while the last column gives the label.

1. Email: ramasubramanian.sundararajan@ge.com [↑](#footnote-ref-1)
2. Acknowledgements: Data based on study conducted with Shanghai Pulmonary Hospital. [↑](#footnote-ref-2)