Improving malignancy prediction through feature selection informed by nodule size ranges in NLST

Computed tomography (CT) is widely used during diagnosis and treatment of Non-Small Cell Lung Cancer (NSCLC). Current computer-aided diagnosis (CAD) models, designed for the classification of malignant and benign nodules, use image features, selected by feature selectors, for making a decision. In this paper, we investigate automated selection of different image features informed by different nodule size ranges to increase the overall accuracy of the classification. The NLST dataset is one of the largest available datasets on CT screening for NSCLC. We used 261 cases as a training dataset and 237 cases as a test dataset. The nodule size, which may indicate biological variability, can vary substantially. For example, in the training set, there are nodules with a diameter of a couple millimeters up to a couple dozen millimeters. The premise is that benign and malignant nodules have different radiomic quantitative descriptors related to size. After splitting training and testing datasets into three subsets based on the longest nodule diameter (LD) parameter accuracy was improved from 74.68\% to 81.01\% and the AUC improved from 0.69 to 0.79. We show that if AUC is the main factor in choosing parameters then accuracy improved from 72.57\% to 77.5\% and AUC improved from 0.78 to 0.82. Additionally, we show the impact of an oversampling technique for the minority cancer class. In some particular cases from 0.82 to 0.87.