Input Data Files Cover Page

1. **Untrimmed Dataset**

The results from the LDHCT screening study arm were utilized for this feature selection investigation. The LDHCT data was cleaned up to only include the information of tumors that were between 4mm to 20mm in size, resulting in a dataset with 3,702 patients and 273 features.

Each row represents a different patient, and each column after the first represents a different feature. The column “ca” represents whether or not the nodule was cancerous.

1. **Trimmed Dataset**

As above, with some more processing:

The data was then cleaned by dropping out the feature columns with over 90% missing data as well as removing the patients with over 50% of their data fields missing. Next, any patients missing a data entry in the diagnosis target attribute were also dropped from the dataset. Patients with 50% or less missing information had their field entries filled with the mean of the attribute from this cleaned patient dataset. This processing resulted in a transformation from 273 features to 268 features and 3,702 patients to 3,534 patients.

We used a feature selection technique called Recursive Feature Elimination to rank features and filter out uncorrelated features based on their direct contribution to the overall prediction score. Recursive Feature Elimination is based on the basic idea of repeatedly creating a predictive model (e.g. SVM or random forest) to evaluate a given set of features in the dataset. We start with all of the features in the dataset. In every iteration, we take out a feature from the dataset and re-evaluate the model to check the impact of that given feature on the prediction accuracy. We repeat that process until all features in the dataset are tested.