Project IV: Lung Cancer Detection
PH 244 Big Data: A Public Health Perspective

Thomas Carpenito & Rachael Phillips

May 7, 2018

Summary

Lung cancer is one of the most common and deadly malignant cancers. Like other cancers, the best solution is early diagnosis and timely treatment. In this work, we aim to address this issue by developing an algorithm that is capable of early detection of lung cancer. The novelty of our algorithm is that it constructs several quantitative summary measures from low dose CT images. We test our algorithm and perform variable importance analysis with training data from the *Data Science Bowl 2017* Kaggle competition [1]. Despite low predictive performance, we highlight the scientific relevance and the future potential of our novel quantitative summary measures for CT image analysis.

1 Questions

The *Data Science Bowl 2017* [1] asks competitors to develop an algorithm to improve early detection of lung cancer using about 1,400 low-dose CT images from high-risk patients. Specifically, the competition asked competitors to develop an algorithm capable of predicting lung cancer within one year of the patient's supplied CT scan. Using summary measures derived from a quantitative conversion of these CT images (Hounsfield Unit frequency distributions), we proceed to answer this question and we propose an additional question. From this collection of summary measures derived from the CT images, which are the most predictive of lung cancer?

2 Data Overview

Provided by the National Cancer Institute, the dataset contains low-dose CT images from 1,595 high-risk patients in Digital Imaging and Communications in Medicine (DICOM) format. Of the 1,595 patient files, 198 and 1,397 comprised the test set and training set, respectively. The DICOM header of each patient file was stripped to permit for patient identification. [*Note:* We only use the training data to test our algorithm because the cancer status is unknown for the test data.]

The volumetric thoracic Computed Tomography (CT) is a commonly used imaging tool for lung cancer diagnosis that visualizes all tissues according to their X-ray absorption. Lesions in the lung are called pulmonary nodules and the existence of nodule does not definitely indicate lung cancer. However, cancerous lung nodules contain a complex intermixing of cellular tissue types [2] that differs from non-cancerous nodules. This cellular tissue type differentiation is the basis of our algorithms design.

3 Data Processing

After extracting the raw patient data, we converted it to Hounsfield Units (HU) using the R package `oro.dicom` [3]. HU is a standard quantitative scale for describing radiodensity that ranges from -1024 to 3044. Interestingly, every tissue has its own specific HU range and this range is the same for different people! Thus, we expect HU range differences between cancerous and non-cancerous patients based on the knowledge provided in Section 2.

With `coreCT::coreHist$histData$finalFreq`, we obtained a long-format data frame of the frequency of each HU for each patient [4]. Each row corresponded to each HU range (4,068 total HU ranging from -1,024 to 3,044), each column corresponded to each patient identifier (1,397 columns) and the data table was populated with frequency counts for individual patients and HU. Next, we divided each frequency by its corresponding column sum to scale the data and make every patient's frequency distribution (i.e., every column) sum to one. Thus, our long-format data table was now populated with the *percent* frequency of each HU for each patient. With `reshape`, we transposed our long-format, 4,068 $\times$ 1,397, data frame into into a wide-format, 1,397 $\times$ 4,068, data frame where each row comprised the HU percentage frequencies for one patient. Next, we simply merged our wide-format data set with the data set containing the lung cancer outcomes (using the patient identifier); obtaining a 1,397 $\times$ 4,069 data frame where the last column corresponded to the cancer outcome.

3.1 Quantitative Summary Measures

Because specific tissues comprise specific *ranges* of HU, the individual HU were grouped into clusters (See Table 1). We used `dplyr::group_by` to group the individual HU percentage frequency data into the clusters and then we took the sum within each cluster with `dplyr::summarize_all`. The typical cluster size is 100 HU but the extreme values have larger HU cluster sizes. Slight discrepancies exist on the correspondence HU ranges and substances (i.e., the HU range corresponding to blood, lung, muscle, etc. varies depending on the source). So, the labels of these ranges are agnostic to substance type. After this HU clustering, our data frame was reduced to 1,595 rows comprising the patients and 23 columns (22 HU cluster columns `HU_A` - `HU_V` and one cancer status column).

label	`HU_A`	`HU_B`	`HU_C`	`HU_D`,..., `HU_T`	`HU_U`	`HU_V`
Hounsfield unit range	$\leq$ -1,000	-1,000 to -900	-900 to -800	-800 to -700,..., 800 to 900	900 to 1,000	$>$1,000

Table 1: Classification of clusters of Hounsfield Units

One of the 1,397 individuals with several "NA" values was discarded from training data set. From the 1,396 individuals in the training data, 1,035 individuals were diagnosed with lung cancer and 361 individuals were not diagnosed with lung cancer. We

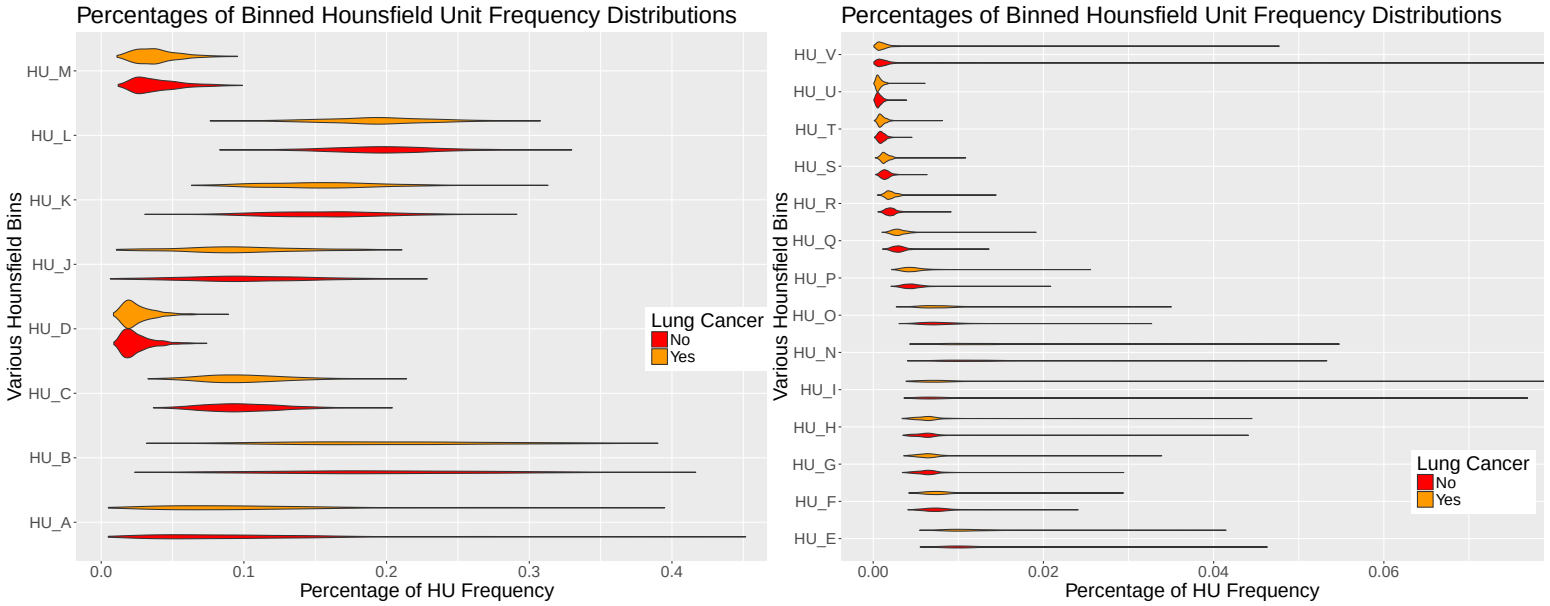

Figure 1: Violin plots of the percentage frequency makeup of Hounsfield Units for patients with and without lung cancer.

envisioned the existence of underlying distributional differences within each HU cluster between individuals with lung cancer and those without. We do not notice any major differences between the two groups of individuals (See Figure 1). We do notice that some of the HU clusters are much more skewed than others. For example, the `HU_P` through `HU_U` cluster distributions are more skewed for those with lung cancer and the `HU_V` distribution is more skewed for those with no lung cancer.

Based on the evidence provided by these violin plots, we considered two additional summary measures: skewness and kurtosis. Skewness is a measure of the symmetry in a distribution and a symmetrical dataset will have a skewness of 0. Essentially, skewness measures the relative size of the two tails. On the other hand, kurtosis is a measure of the combined sizes of the two tails. Kurtosis measures the amount of probability in the tails. As before, we used `dplyr::group_by` to group the individual HU percentage frequency data into the clusters and then we calculated skew and kurtosis within each cluster with `dplyr::summarize_all`. To visualize differences between cancerous and non-cancerous groups, we calculated the cancer specific mean within each HU cluster for both kurtosis and skewness (See Figure 2). We see more variation between the cancer and no cancer groups when we consider the average skewness. There is little variation (aside from the extreme bins) between the two groups average kurtosis. In downstream analysis, we consider the following quantitative summary measures: (1) the bins themselves consisting of the percentage of HU frequencies (`HU_A` - `HU_V`) (2) the skewness of each of these bins (`HU_A_S` - `HU_V_S`) and the (3) kurtosis of the bins (`HU_A_K` - `HU_V_K`). Thus, our final data table has 1,396 rows corresponding to the patients in the training data and 67 columns (22 columns for the raw bins, 22 columns for the skew of these bins, 22 columns for the kurtosis of the bins, and 1 column for the cancer status).

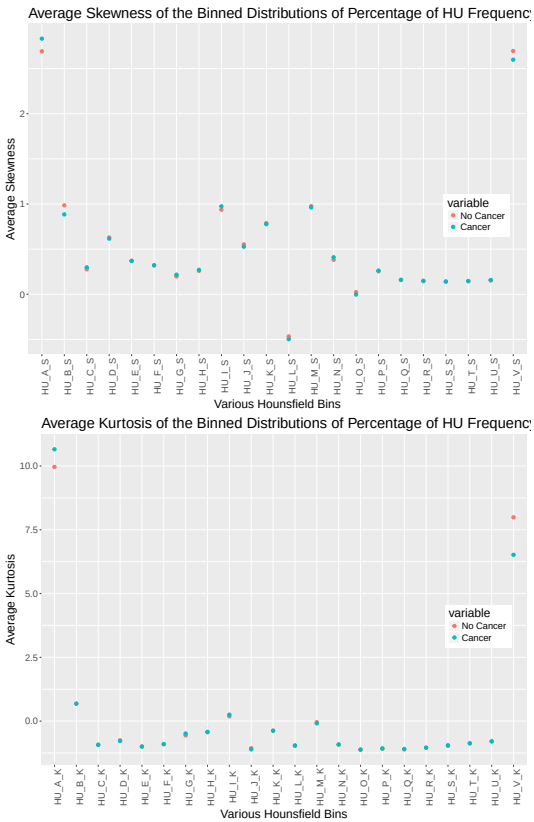

Figure 2: Scatter plots comparing the average skew (top) and the average kurtosis (bottom) across binary cancer groups.

4 Data Analysis

4.1 Winning Methods

The winning methods have been published in [5]. The winning team, *grt123*, implemented a 3D deep neural network model consisting of two modules. The first module is a 3D CNN region proposal network for suspicious nodule detection. The second module selects the top five nodules based on the detection confidence, evaluates their cancer probabilities and obtains the probability of lung cancer for the subject. The algorithm performed very well on the lung cancer classification task.

4.2 Novel Methods

We used 10-fold cross-validation with `CV.SuperLearner` [6] to predict the binary lung cancer status with the training data summary measures. SuperLearner is an ensemble machine learning algorithm that uses cross-validation to estimate the performance of multiple machine learning models. It creates an optimal weighted average of those models, using the test data performance (created internally by the folds) and a specified loss function. This approach has been proven to be asymptotically as accurate as the best possible prediction algorithm that is tested. The `CV.SuperLearner` is identical to the regular SuperLearner aside from a

layer of "nested" cross validation. This involves (1) hiding a subset of the training data so that it is not used to fit SuperLearner and then (2) estimating the performance of the SuperLearner on this unseen data. We utilize the `CV.SuperLearner` so that we can obtain predictive performance of our SuperLearner. Our `CV.SuperLearner` library of base learners contained a few robust algorithms that can handle binomial outcomes: `SL.ranger` – a faster implementation of random forest, `SL.glmnet` – generalized linear model with lasso or elastic net regularization, `SL.gam` – simple generalized additive model, `SL.knn` – K-nearest neighbor classification, and `SL.mean` – the sample mean in the training data.

4.2.1 Predictive Performance

We evaluated the predictive performance of our `CV.SuperLearner` classifier with the area under the ROC curve (AUC) – a summary measure of the predictive accuracy of a binary classification model where a value closer to 1 indicates high predictive accuracy and a score closer to .5 is akin to flipping a coin (See Figure 3). The `CV.SuperLearner` AUC of .5047 indicates little to no predictive capability. We also compare the performance of the `CV.SuperLearner` to the performance of the of the base learners (see Table 2). This table indicates essentially the same AUC using the sample mean and the final super learner model, suggesting little predictive capability in our covariates.

Algorithm	Average AUC
Super Learner	0.505
Discrete SL	0.491
SL.glmnet	0.4757
SL.glm	0.5135
SL.knn	0.5045
SL.gam	0.5216
SL.mean	0.5000
SL.ranger	0.4853

Table 2: Average AUC for each algorithm in the super learner library.

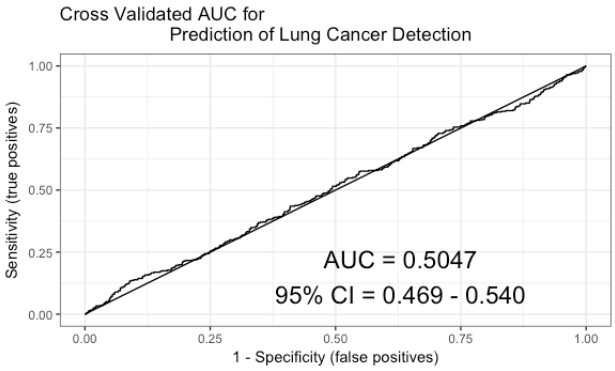

Figure 3: Plot of the AUC for CV.SuperLearner.

4.2.2 Variable Importance Measures

We used the `randomForest` function and R package and the `varImpPlot` function to see which variables are most important in predicting lung cancer. Because `randomForest` is sensitive to class imbalance, we first took a random sample (with `dplyr::sample_n`) from the no cancer group such that we would have equal class sizes ($n_{NoCancer} = 361$ and $n_{Cancer} = 361$). The Mean Decrease in Accuracy (MDA) quantifies the importance of a variable by measuring the change in prediction accuracy, when the values of the variable are randomly permuted compared to the original observations and a high score means the variable was important. According to Figure 4, `HU_P_S` or the skewness of `HU_P` (ranging from 400 to 500 HU) seems to be the most important variable in predicting lung cancer and `HU_A` (ranging from -1,000 or less HU) follows closely behind.

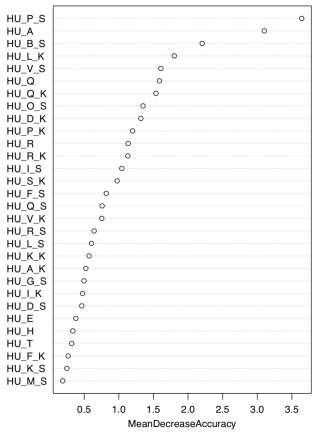

Figure 4: Variable importance plot based on MDA.

5 Discussion

In this work, we explore methods to improve early detection of lung cancer. We constructed several quantitative summary measures from low dose CT images. These summary measures account for distributional shapes of HU bins and the raw distributions of HU bins. We hoped to identify lung tissue type differences between the cancer and non-cancer patients as was done in [2]. However, our AUC of .5047 indicates that there is little success in using these clustered HU summary measures as a predictor for lung cancer detection in at risk patients.

It would be interesting to explore alternative preprocessing procedures. There might be a more appropriate method for scaling the data. Also, we might consider isolating the lung cancer nodules from the images and then performing the analysis with the quantitative summary measures. It would also be great to test the performance of our algorithm on a different data set with less class imbalance. A coherent and consistent conversion between HU range and substance would be very helpful for interpreting our variable importance results. Accompanying this report is complete repository of the R code, tables, and results [7].

References

[1] Data Science Bowl 2017. `https://www.kaggle.com/c/data-science-bowl-2017`, 2017.

[2] JC Sieren, AR Smith, J Thiesse, E Namati, EA Hoffman, JN Kline, and G McLennan. Exploration of the volumetric composition of human lung cancer nodules in correlated histopathology and computed tomography. *Lung Cancer*, 74(1):61–68, 2011.

[3] Brandon Whitcher, Volker J. Schmid, and Andrew Thornton. Working with the DICOM and NIfTI Data Standards in R. *Journal of Statistical Software*, 44(6):1–28, 2011.

[4] Troy D. Hill, Earl Davey. *Programmatic analysis of sediment cores using computed tomography imaging.* Narragansett, RI, 2017.

[5] Fangzhou Liao, Ming Liang, Zhe Li, Xiaolin Hu, and Sen Song. Evaluate the Malignancy of Pulmonary Nodules Using the 3D Deep Leaky Noisy-or Network. *arXiv preprint arXiv:1711.08324*, 2017.

[6] Eric Polley, Erin LeDell, Chris Kennedy, and Mark van der Laan. *SuperLearner: Super Learner Prediction*, 2018. R package version 2.0-23.

[7] Rachael Phillips. Big Data Project IV: Lung Cancer Detection. `https://github.com/rachaelvphillips/ph244-big_data`, 2018.