Automatic Lung Cancer Detection Using Volumetric CT Imaging Features

A

Research Project Report

Submitted

To

Computer Science Department
Brown University

By

Dronika Solanki
(B01159827)

Abstract

Lung cancer is the most prevailing cancer and the leading cause of cancer related deaths throughout the world [1,2]. Early detection of lung cancer can increase the survival rate of cancer patients. In United States, current statistics shows that about 1 out of 4 cancer deaths are from lung cancer among both men and women than other cancers. The goal of this study is to develop machine-learning models that can detect malignant lung nodules from CT using volumetric imaging features. we used scale invariant feature transform (SIFT) to identify important image features and linear support vector machine (SVM) is our learning algorithm. To understand the shape of healthy and unhealthy data sets or to understand the correlation pattern, we employed tools from topological data analysis (TDA). We have used 80% of the data for training the SVM and remaining 20% of the data for testing.

Introduction

Lung cancer is the most prevailing cancer and the leading cause of cancer related deaths throughout the world. Early detection of lung cancer and treatment can help to reduce the number of deaths every year. Medical imaging helps Radiologists to diagnose diseases however they may miss a small nodule or subtle area, which is difficult to recognize. Scientists have collected large amount of cancer data and is available to research community. In this study image processing procedures and important features extraction have been discussed. We will also discuss machine learning techniques, which helps us to automatically detect malignant lung nodules from CT using volumetric imaging features. Scale invariant feature transform (SIFT) is used to identify important image features. We will train a linear SVM for automatically detecting a malignant lung nodules from CT. To estimate the classifier's performance we have measured accuracy, sensitivity, specificity and area under the curve (AUC). We also employed tools from topological data analysis to understand the topological characteristics of our data set.

Data

CT volumes from the National Lung Screening Trial (NLST) were analyzed for this study. In this section, I am writing instructions for downloading the NLST CT images. To download data from NLST, we need to submit a request for NLST data through the CDAS website. Once request approved, go to (https://biometry.nci.nih.gov/cdas/login/?next=/cdas/projects/nlst/147/data/deliverables/). After successfully logged in we need to follow these steps:

- Click on “My project” on upper right corner and you will see all your approved projects list.
- Click on the specific project
- Click on the “Data Delivery” tab on the right side
- Click on the “Data Deliverables” tab on the top left side

We can either click “Download Data” (to download the trial clinical metadata in CSV format in a zip file), or we can click “Access Images” to use the online TCIA query tool in Figure 1. As we only want to study CT images for our project so first identify people of interest using the metadata files (which we have downloaded from above link), and then download their images using the Query tool. In all metadata files, for our project prsn, sctabn, sctabnc and sctimageinfo

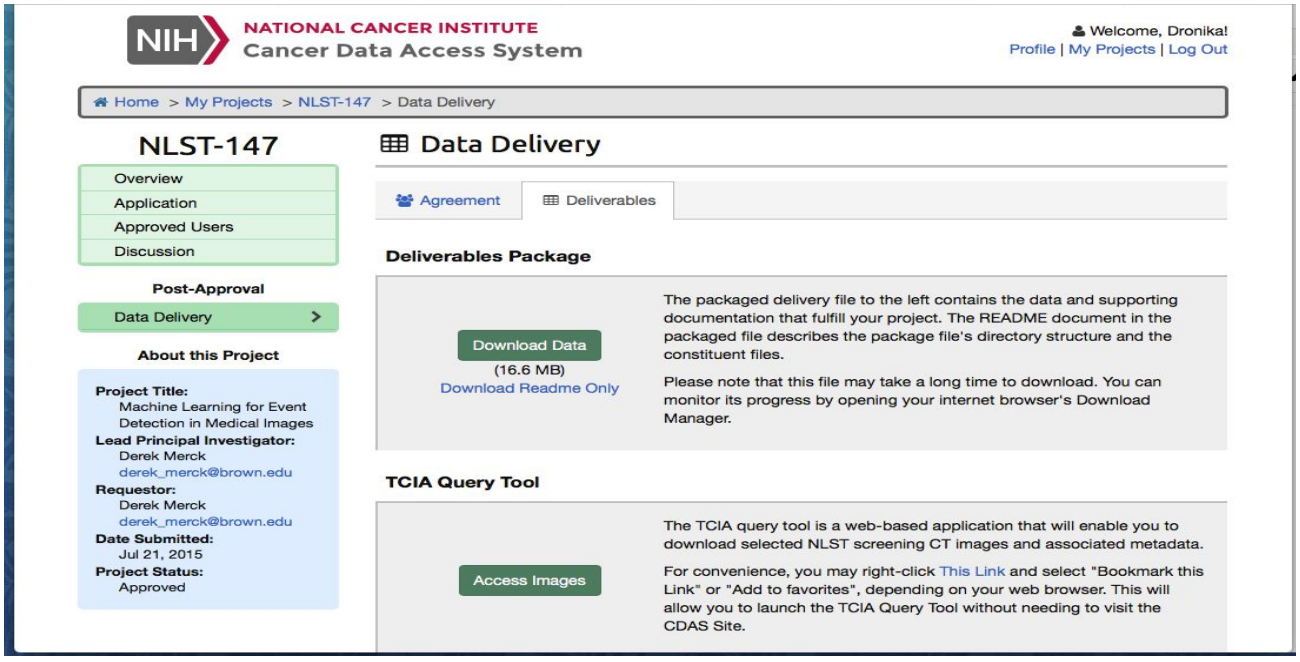

Figure 1 : NLST dashboard for data downloading

files used. The final exam for each subject selected was used, along with the radiologists recorded overall screening results: a positive screening result (suspicious for lung cancer) was assigned if any non-calcified nodules or masses >= 4 mm in diameter, confirmed as malignant with biopsy, were present; and a negative screen was assigned if no significant abnormalities were present. We went through the screen results (scr_res0-scr_res2 in prsn.csv file) to determine the last screen. If a participant was screened, they will have values of 1-6. For this study, the true positives were defined as having conflc = 1 and can_scr = 1 in prsn.csv file. The variable "conflc" represents "status of lung cancer report" and have value in the range of (0-4). Here 0 - "No Report", 1 - "Follow-up collected confirmed Lung cancer", 2 - "Follow-up collected confirmed Not Lung Cancer", 3 - "Medical Records can not be obtained" and 4 - "pending" respectively. The variable "can_scr" represents "whether the cancer followed a positive, negative, or missed screen, or whether it occurred after the screening years" and have values in the range of 0-4. Here 0 - "No Cancer", 1 - "Positive Screen", 2 - "Negative Screen", 3 - "Missed Screen", 4 - "Post Screening" respectively. Finally, anyone with a screen result of 1 and no cancer diagnosis was classified as a true negative. To download the images, we have a text (.txt) file containing a column of PID (participant ID) values for each selected participants. It always recommend to break up the download into several chunks (3 to 5), in case the download is interrupted for some reason. To download images we need to follow these steps :

- Log in to CDAS when prompted.
- The Query Tool should now appear on a separate tab in your web browser.

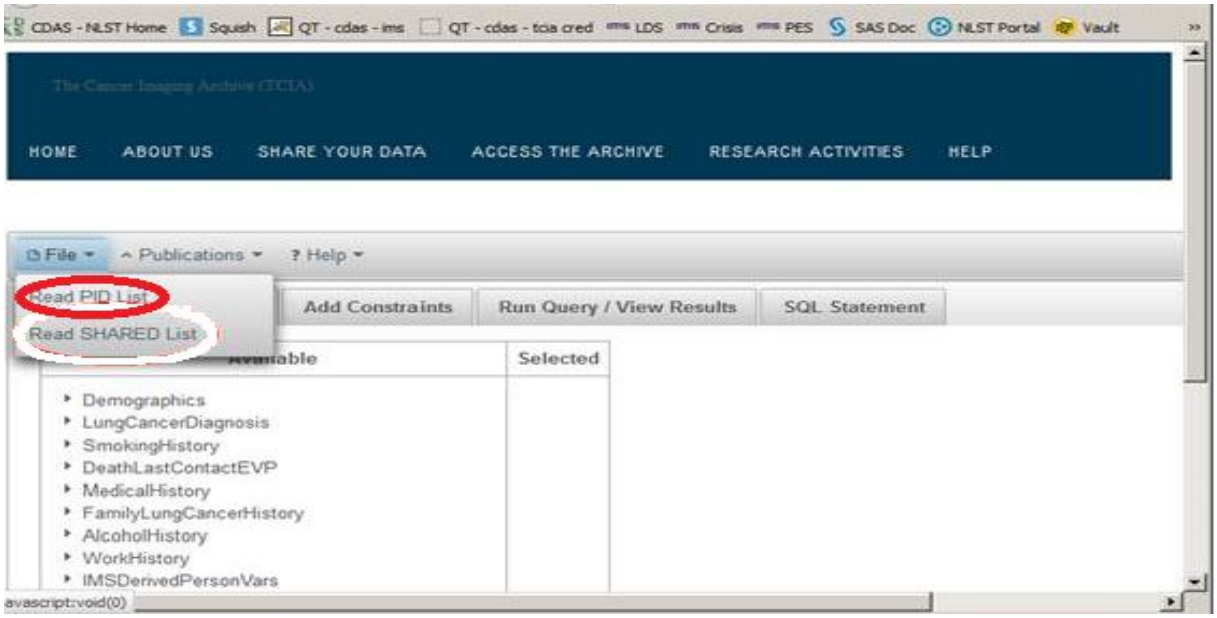

Figure 2 : Display of Query tool screen

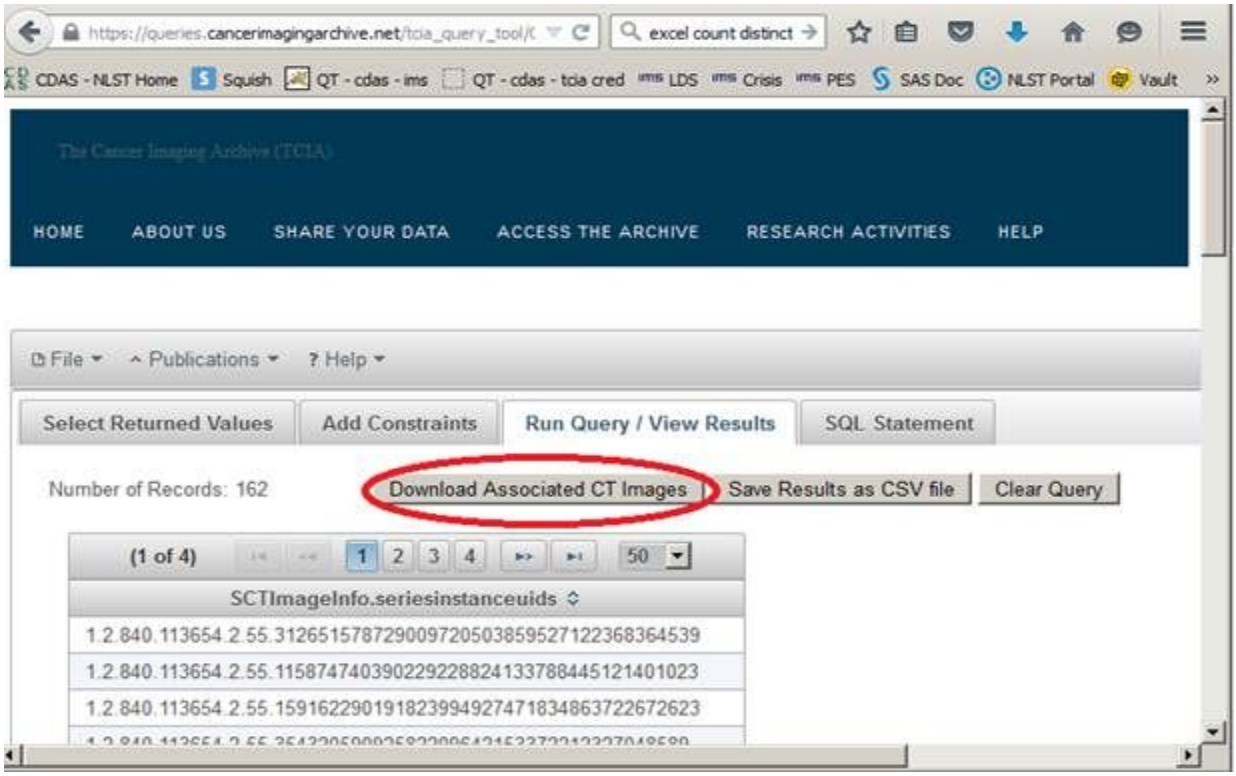

Figure 3 : Download associated CT images

- Click on the File button and then click “Read PID List”. After uploading file we will receive a success message with instructions on how to download the images as shown in Figure 2.
- Click the “Run Query / View Results” tab and then select the “Download Associated CT Images” button as shown in Figure 3.
- We will then be prompted to open the NBIA Download Manager by opening a file named **main.xhtml**, (which is a Java Net Launch Protocol file) with Java Web Start Launcher as shown in Figure 4.

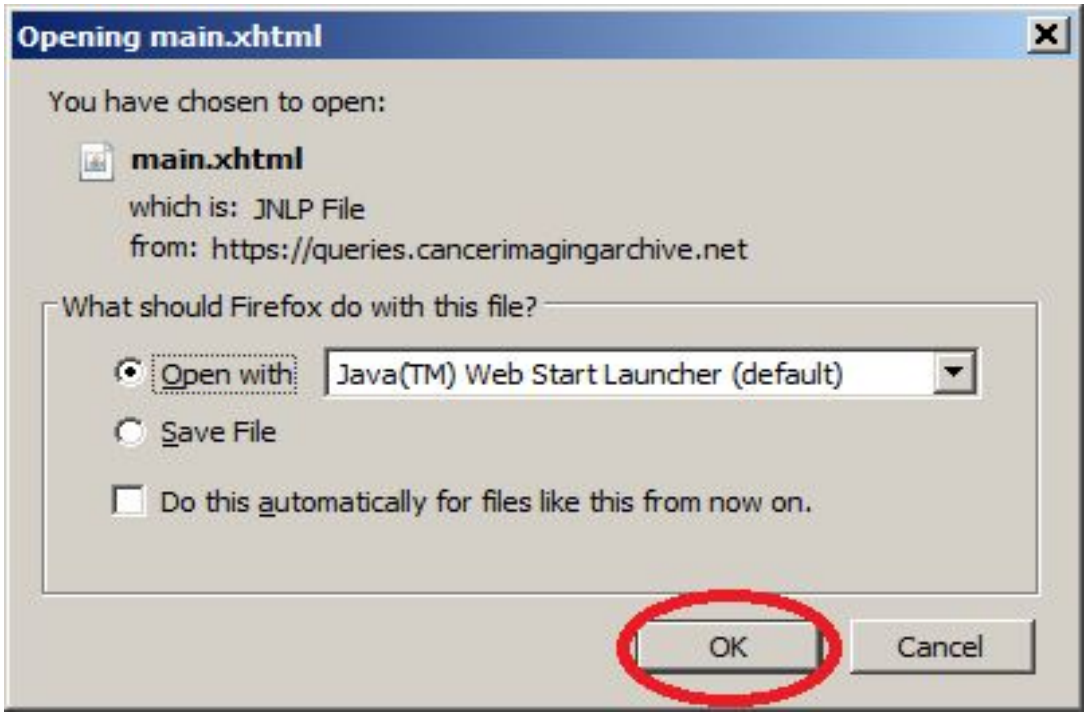

Figure 4 : Java Net Launch Protocol File

- The Download Manager will launch. We will have the option to choose where you want to save the images. Once we have chosen the location, it starts the download. Progress bars indicates the % complete for each image series.

Our dataset comprised 620 lung cancer-positive CT scans and 950 lung cancer-negative CT scans from 1570 unique subjects. After downloading all scans, I have realized that for each patient folder it has two or three more folders which contains dcm file. As discussed above, most patients had three scans, at one-year intervals. Within each of those "study year folders", I found several more folders, one per image series. Within each series folder, I saw either 1 .dcm file or 100+ .dcm files. The folders with 1 .dcm image file are localizer scans of the whole chest. The folders with 100+ image files are the main scans. The csv file called sctimageinfo.csv contains one record per image series. For example :

/nlst/104567/1.22.55.4958795734570974570 5/1.25.76.54.059740957857235034/000001.dcm

/(A) /(B) /(C) /(D) / (E)

Here :

A. Root folder name
B. PID, participant identifier, contained in all CSV metadata files
C. STUDYUID: tells us which of the three study years the scan is from (T0, T1, or T2). Use imginfo.csv file, and find the row with the STUDYUID that matches the folder name, and find the value of STUDY_YR in the same row of the imginfo.csv file.
D. SERIESINSTANCEUIDS: tells us which image series in a given study year. We have 2 kinds of image series: axial reconstructions of the main CT scan (which are the series of

interest) and single localizer images of the chest used by the CT technician to line up the scanner before performing the main CT scan. We are using axial reconstructions for our study which contains 100+ slices spaced out through the chest.

E. Folder containing all 100+ slices.

Some of the final year scans were empty for some patients and we have removed these patients PID from our dataset. Each PID (study) were labeled as positive if it contains the malignant lung nodule identified by radiologist and labeled as negative if its healthy and not containing any nodule.

Methodology

We have divided our methodology into two parts, feature extraction and machine learning algorithms respectively. Our goal was to identity the important image features that can help to distinguish between healthy and unhealthy lung. We observed that nodule detection should be invariant to scale, which means it should be independent of the size of nodule. It should also invariant to rotation and translation, which means it should be independent of the specific orientation or location of the nodule. We extracted volumetric imaging features from each CT scan using a 3D version of the scale-invariant feature transform (SIFT) [3]. The location and scale of distinctive image patches are detected as extrema of a difference-of-Gaussian operator. Once detected, patches are reoriented, rescaled to a fixed size (11^3 voxels) and transformed into a GoH representation over 8 spatial bins and 8 orientation bins, resulting in a

64-element feature descriptor [4]. First we pool all SIFT descriptors from training cases and vector quantize them using K-means clustering algorithm, which results into K clusters centers each of 64 dimensional vector. We used Vector of locally aggregated descriptors (VLAD) to represents each CT slice using SIFT features, which results into 4096-dimensional features for each CT volume. 80% of the data was used for training machine learning models and 20% of the data was used for testing its performance. We choosed linear SVM [5] as our machine learning algorithm. 5-fold cross-validation was used during training for model selection. After obtaining a classification model it is also important to evaluate the classifier's performance. The classifier's performance can be calculated in terms of accuracy, sensitivity, specificity and area under the curve (AUC). Sensitivity determines the amount of true positives which is correctly observed by the classifier and on the other hand Specificity determines the amount of true negatives which is correctly identified.

Results

We have used Python and MATLAB software for all computations. For SIFT feature extraction and SVM training, we have used vlfeat library. The support vector machine (SVM) [5] was the best performing model with an area under the ROC curve of 0.8329 (95% CI: 0.7901, 0.8765) as shown in Figure 5. At a score threshold of 0.5, the SVM had 86% specificity and 59% sensitivity whereas at a score threshold of 0.35, the SVM had 39% specificity and 98% sensitivity, showing that the classifier has a broad range of utility for different threshold settings.

We used Ayasdi software to build the simplicial complex representation of the SIFT VLAD data obtained above from CT slices[6,7].

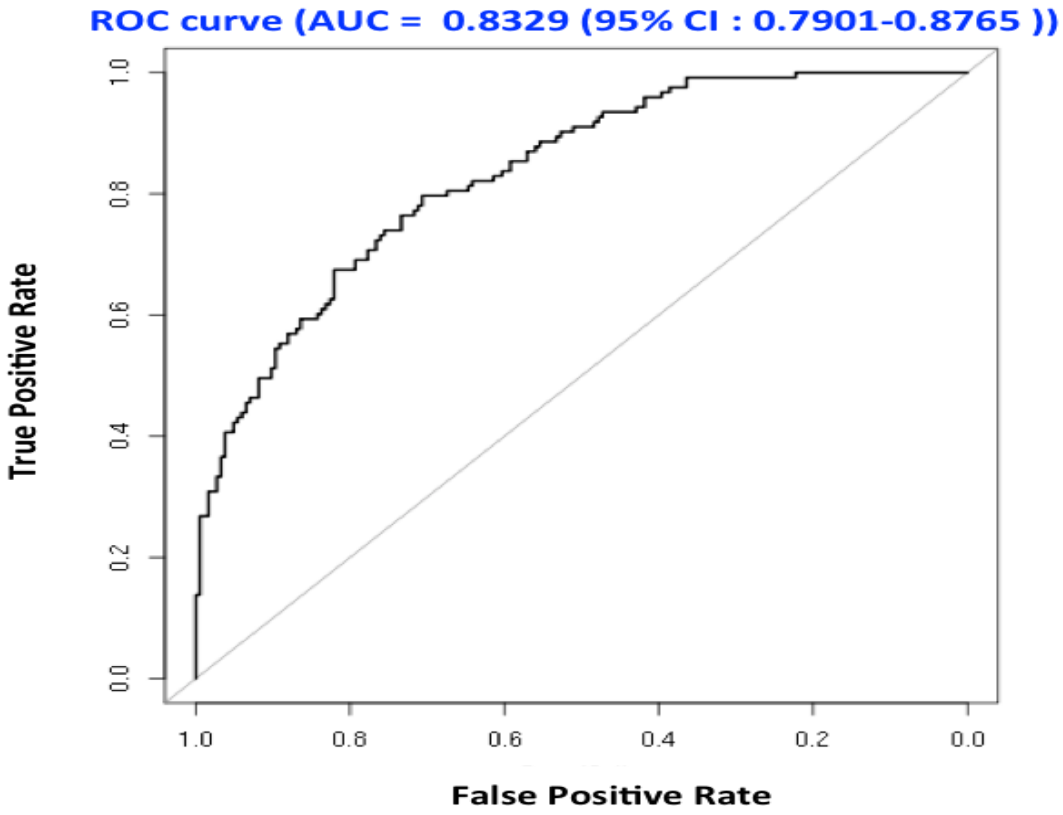

Figure 5 : ROC analysis for our classifier. As we see, the classifier achieves high area under the curve.

In this software, first we need to choose lens function, which helps us to reduce the data to lower dimensional. The lens function is divided into overlapping intervals and further clustered using a chosen distance metric in the data space. Clusters who has common data points are connected through links and visualized by 2D projection. Visualization of the topological structure of the feature space shows how the positive and negative cases form separable clusters

as shown in Figure 6. Here each node represents clusters of cases that are similar in the feature space under normalized Euclidean metric.

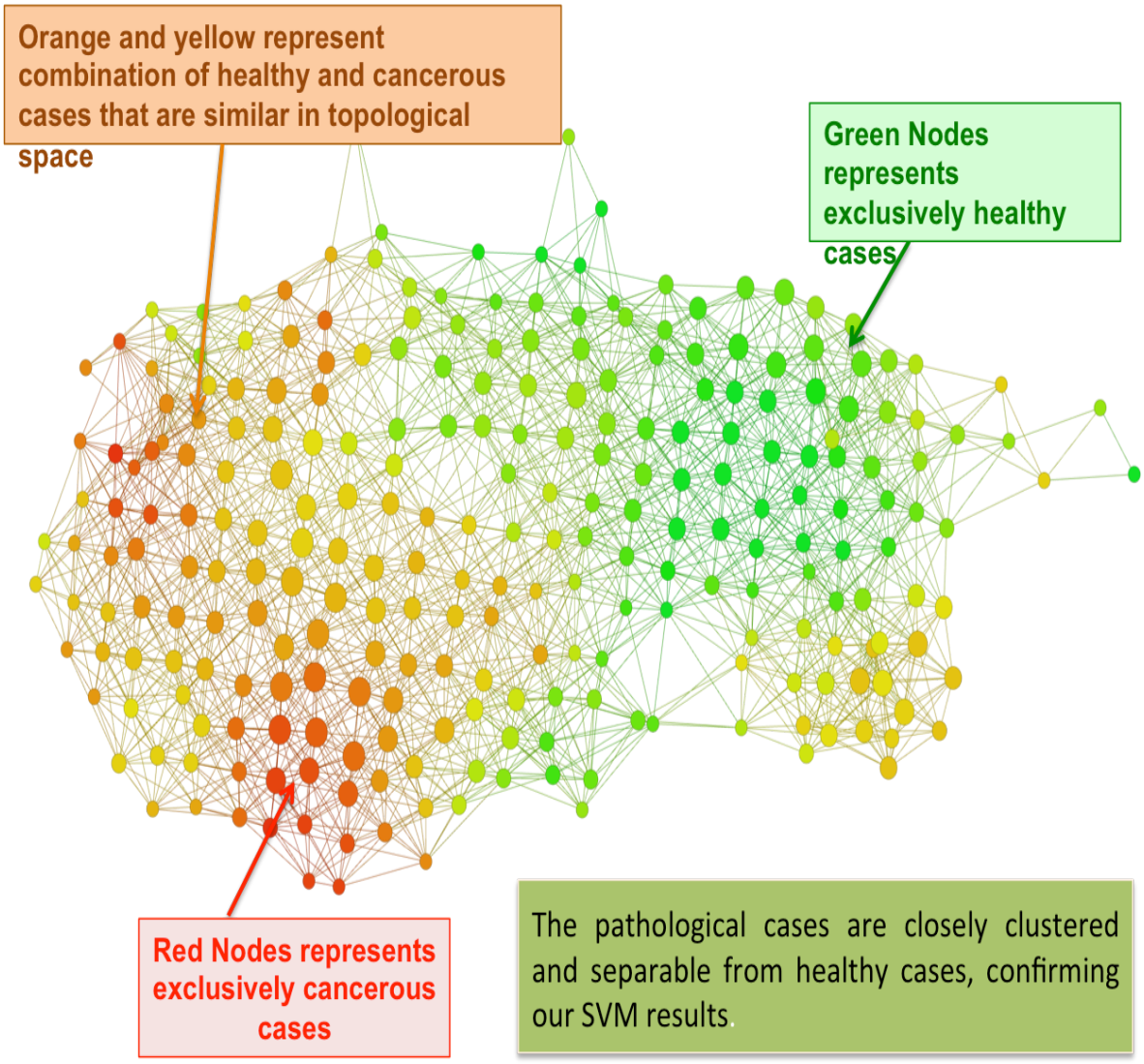

Figure 6 : Topological data analysis. Nodes represents clusters of cases that are similar in the feature space under normalized Euclidean metric.

Conclusion

Machine learning with a dataset of over 1500 volumetric studies achieved respectable discriminatory power in differentiating between CT volumes with and without cancerous lung nodules using volumetric imaging features. Natively volumetric imaging features have only recently been developed and applied in the medical imaging domain. Future work will use incorporate more data from the NLST and investigate clinical implementation.

Clinical Relevance

Machine learning models can help guide radiologists in determining whether a CT scan contains a potentially malignant nodule; and if so, whether to recommend the nodules for biopsy.

Acknowledgement

I wish to thank my advisor Professor Stan Zdonik, without whom this project would not have happened. I would also like to thank Dr. Derek Merck for his excellent guidance and mentorship on this project. My additional thanks go to Krishna Keshavamurthy for all his help throughout this project.

References

1. Jemal, A. et al. Global cancer statistics. CA Cancer J. Clin. 61, 69–90 (2011).

2. Siegel, R., Naishadham, D. & Jemal, A. Cancer statistics, 2013. CA Cancer J. Clin. 63, 11–30 (2013).
3. Lowe, D.G., “Distinctive image features from scale-invariant keypoints,” Int J Computer Vision 60(2), 91-110 (Nov 2004).
4. Matthew Toews, “A Feature-based Approach to Big Data Analysis of Medical Images.” Inf Process Med Imaging. 2015 ; 24: 339-350.
5. Fan, R., Chang, K., Hsieh, C., Wang, X., and Lin, C., “LIBLINEAR: A library for large linear classification,” J Machine Learning Res 9, 1871-1874 (Jun 2007)
6. Lum, P.Y., Singh, G., Lehman, A., Ishkanov, T., Veidemo-Johansson, M., Alagappan, M., Carlsson, J., and Carlsson, G., “Extracting insights from the shape of complex data using topology,” Scientific Reports 3, 1236 (Feb 2013).
7. Carlsson, G., “Topology and data,” Bull Amer. Math. Soc. 46, 255-308 (2009).