UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

PREDICTION OF EARLY STAGE LUNG CANCER PROGNOSIS AFTER SURGERY USING A NEW CAD-GENERATED IMAGING MARKER

A THESIS

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

MASTER OF SCIENCE

By

VENKATA SAI CHANDRA BOYAPALLY
Norman, Oklahoma
2018

PREDICTION OF EARLY STAGE LUNG CANCER PROGNOSIS AFTER SURGERY USING A NEW CAD-GENERATED IMAGING MARKER

A THESIS APPROVED FOR THE
SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING

BY

_____
Dr. Bin Zheng, Chair

_____
Dr. Yuchen Qiu

_____
Dr. Samuel Cheng

Acknowledgements

This work is supported by Peggy and Charles Stephenson Cancer Center, University of Oklahoma. The author would like to gratefully acknowledge his thesis advisor Dr. Bin Zheng who gave the author this opportunity to perform his study, and mentored him to accomplish his master degree thesis.

The author would also like to thank all his committee members: Dr. Yuchen Qiu and Dr. Samuel Cheng, for their efforts of serving in the author's M.S. program committee.

Finally, author expresses his deep gratitude to his family and friends for their endless support and encouragement.

Table of Contents

Acknowledgements iv

List of Tables vi

List of Figures vii

Abstract viii

Chapter 1: Introduction 1

Lung Cancer 1

Computed Tomography (CT) Image Modality 2

Lung Cancer Staging 3

Lung Cancer Recurrence and Its's Prognostic Factors 5

Chapter 2: Materials 8

Dataset 8

Chapter 3: Methods 11

Segmentation of Lung 11

Segmentation of Lung Tumors 13

Density Mask 15

Quantitative Tumor and Emphysema related Features 16

Feature Selection 21

Machine Learning Model 22

Performance Assessment and Data Analysis 23

Chapter 4: Results 25

Chapter 5: Discussion 32

Chapter 6: Conclusion 35

References 36

List of Tables

Table 1: Demographic information of the patients …… 7

Table 2: Distribution of two groups of cases in tumor and cancer cell characteristics . 8

Table 3: Emphysema features …….. 19

Table 4: comparing AUC values of individual tumor related features ……... 22

Table 5: comparing AUC values of individual Emphysema1 related features…… 22

Table 6: comparing AUC values of individual Emphysema2 related features …… 23

Table 7: comparing correlation coefficients between tumor and emphysema1 feature..23

Table 8: comparing correlation coefficients between tumor and emphysema2 features24

Table 9: Comparing Performance of different Image Markers ……... 24

Table 10: Confusion matrix generated using the feature fusion based image marker……...…….....26

Table 11: Confusion matrix generated using the tumor feature based image marker …. 26

Table 12: Confusion matrix generated using the emphysema1 feature based image marker……....26

Table 13: Confusion matrix generated using the emphysema2 feature based image marker…….27

List of Figures

Figure 1: Segmentation of the Lung region ……..……... 11

Figure 2: Flowchart to segment Lung tissue …… 11

Figure 3: Segmentation of tumor region …….……..……. 13

Figure 4: ROC curves comparing tumor feature based classifier, emphysema feature based classifier and combine feature based classifier …….……. 25

Abstract

Due to cancer heterogeneity, identifying new clinical markers to more effectively predict prognosis of cancer patients plays an important role to improve efficacy of cancer treatment. Objective of this study is to develop and test a new quantitative imaging (QI) marker to predict prognosis of early stage non-small cell lung cancer (NSCLC) patients after surgery. For this purpose, this study includes following research tasks or steps. First, a new computer-aided detection (CAD) scheme was developed to automatically segment lung and tumor regions from the chest computed tomography (CT) images of all the slices simultaneously using an adaptive pixel value thresholding and/or region growing method. Next, CAD scheme was used to compute a large number of image features related to tumor shape, size, circularity, density heterogeneity, and lung background tissue patterns. Then, a machine learning approach was applied to build a multi-feature fusion based prediction model, which enables to produce a CAD-generated quantitative image (QI) marker for predicting diseases-free survival (DFS) of the NSCLC patients within 3 years after surgery. In order to achieve more robust result of training and testing the machine learning model, a leave-one-case-out (LOCO) cross-validation method was used. A feature selection process using a correlation-based feature subset evaluator and a synthetic minority oversampling technique (SMOTE) were embedded in LOCO based training process. Finally, prediction performance of the QI marker or prediction model was evaluated using the receiver operating characteristic (ROC) and other statistical data analysis. In summary, the goal of this study is to select more effective image features computed from both segmented lung tumors and emphysema related background regions for producing a new CAD-generated QI marker and demonstrate the feasibility of

applying this new QI marker to yield higher performance in predicting prognosis of early stage NSCLC patients

Chapter 1: Introduction

Lung Cancer

Lung cancer is one of the leading cancer that has the highest mortality rate, An estimated 1,685,210 Americans are expected to die from lung cancer in 2016, accounting for approximately 27 percent of all cancer deaths, which is more than a quarter of cancer deaths in U.S. Currently, the most important cause of lung cancer is smoking (resulting in nearly 85% of lung cancer cases in U.S). Lung cancer has poor prognosis; nearly 90% of lung cancer patients die of this disease. Lung is a relatively large human organ that may involve many other chronic lung diseases, and the lung tumors can often grow for a long time before they are found. Even when some abnormal symptoms, such as coughing and fatigue, do occur many people are likely to consider these symptoms due to other chronic lung diseases. For this reason in the normal clinical practice, it is difficult to detect early-stage lung cancer (stage 1) that has a better prognosis. Hence, a large number of patents with the lung cancer are currently detected and diagnosed at the advanced stage with the low survival rate. Therefore, great effort has been made to import lung cancer screening program using low-dose computed tomography (CT) image modality for the last decade[1]. In July 29, 2103 U.S. Preventive Services Task Force issued a draft recommendation in favor of lung cancer screening for long-term smokes using low-dose CT tests. Meanwhile, the trend of using CT for screening and/or detecting other types of lung diseases (e.g., chronic obstructive pulmonary diseases) is also on the rise. As a result, more early stage lung cancer are detected during the regular lung cancer screening or other incident finding. In lung cancer, there are two major categories based on the size of the cancer cell when seen under microscope: small cell lung cancer and non-small cell

lung cancer which are treated differently. Over 75% of lung cancer cases are non-small-cell lung cancer (NSCLC); hence this study was focused on prognosis of NSCLC patients[2]. NSCLC has three sub types of adenocarcinoma, squamous cell carcinoma, and large cell carcinoma which differ in size, shape and chemical make-up, but have similar prognosis. Research shows that COPD patients who are smokers have a higher risk of getting lung cancer. However, there is increasing evidence that even those non-smokers with COPD have a greater risk of developing lung cancer. The link between the two could be that smoking is an acknowledged cause of COPD and a cause of lung cancer. But, recent evidence suggests that COPD itself is an independent risk factor for developing lung cancer, separate from any smoking history.

Computed Tomography (CT) Image Modality

X-ray Computed Tomography (CT), formerly known as Computerized Axial Tomography (CAT), was introduced into clinical practice in 1972 by G.N.Housefield. Currently, high-resolution CT is an important diagnosis tool in clinical radiology specifically in lung cancer imaging. Due to CT's higher accuracy, wide accessibility and cost-effectiveness, it remains the most popular imaging modality whereas many other advanced imaging modalities including, Positron Emission Tomography (PET), PET-CT and magnetic resonance imaging (MRI) has been investigated and applied in lung cancer imaging[3].

On the contrary to the conventional X-ray imaging techniques, CT is a slice-imaging (or cross-sectional imaging) modality that is capable of acquisition of images from whole body in spiral and axial scanning modes, which provide three-dimensional information from internal organs and tissues of interest. A large volume of body can be

scanned through this modality towards imaging of multiple slices simultaneously. Cross-sectional image of the anatomy is obtained through reconstruction of X-ray projected data by a computer. There are two types of reconstruction techniques: analytical reconstruction and iterative reconstruction. Filtered back-projection is an analytical projection method that has been widely used for CT scanners.

Throughout the time, development in X-ray, detector and scanner technology have led to a renaissance of CT. Up to now, CT scanner have 4generation. Frist generation CT scanner use single X-ray source and single detector that translate and rotate to scan anatomy with a small fan beam. Second generation CT scanner have multiple detector. Source and detector translate while scanning via fan beam over a larger rotating array of detector using wide fan beam, while the fourth generation scanner have rotating X-ray source and a larger fan beam[4]. The two earlier generation are considered a translation/rotation systems, and the two latest generation are considered as continuous rotation system which have improved scanning speed, acquisition of image data, as well as image quality (i.e. 3D spatial resolution).

Lung Cancer Staging

What is staging?

Cancer staging determine how much the disease has grown or spread within patient's body. Staging provide information on the extent or the severity of the cancer. Staging can be determined through tests such as laboratory tests, imaging test, physical exams, and pathology reports[5]. Cancer patients will be staged when they are

diagnosed with the cancer, and they will be referred to this stage even if the cancer has progressed or has got worse.

Cancer staging is important as:

- Knowing the cancer stage helps physicians in making decisions for the treatment.
- It is used in providing cancer prognosis for patients.
- Aids is figuring out beneficial clinical trials for treating patients.
- It also enables exchanging information about patients, evaluating and comparing data gathered from different clinical trials.

Common Staging Factors

Based on National Cancer Institute fact sheet the most common factors used in determination of cancer stage are:

- Location of the primary tumor and the cell type (e.g., adenocarcinoma, squamous cell carcinoma)
- Tumor size and/or extent(reach)
- Involvement of regional lymph node (the spread of cancer to nearby lymph nodes).
- Number of tumors (the primary tumor and the presence of metastatic tumors, or metastases)
- Tumor grade (how closely the cancer cell and tissue resemble normal cells and tissue)

What Is Stage I Non-Small Cell Lung Cancer?

The American Joint Committee on Cancer (AJCC) TNM staging for lung cancer is included in appendix A. Here NSCLC staging is focused specifically. The information gathered from three T, N and M factors will be combined to determine an overall stage

(0, I, II, III or IV) for NSCLC cases. Lower stage indicates a better outlook for patient. Stage I NSCLC cases are divided into two groups[5].

Stage IA: The tumor is not large than 3cm, has not reached membranes surrounding the lungs, does not affects main branches of the bronchi, and has not spread to lymph nodes or distant sites.

Stage IB: The cancer has not spread to lymph nodes or distant sites and has one or more of the following properties:

- The tumor is larger than 3cm across but not larger than 5cm.
- The tumor has grown into a main bronchus, but is not within 2cm of the carina (and it is not larger than 5cm).
- The tumor is partially clogging the airways (and is not larger than 5cm).

Lung Cancer Recurrence and Its's Prognostic Factors

As it was mentioned earlier patients diagnosed with lung cancer at early stages (such as stage I) will have better prognosis and early cancer detection and treatment can improve the survival rate of lung cancer patient. However, lung cancer patient may still suffer from cancer recurrence after surgical resection of the malignant tumor. Based on different studies 30% to 60% of stage I NSCLC patients have cancer recurrent.

This indicates that the mortality rate among the stage I NSCLC patient is much higher that many other types of cancer (e.g., breast cancer) detected at early stage.

According to the date from the National Cancer Institute's Surveillance, Epidemiology, and End Results database, current 5-Years survival rates are 49% and 45% for Stage IA and Stage IB NSCLC patients, respectively[1]. Identifying or developing effective

clinical markers or prediction models will not only lead to an accurate and reliable cancer prognosis for patients who went through a surgical resection, but also aid in effective treatment management. As through a reliable marker, patients with high risk of cancer recurrence will be identified, specific adjuvant chemotherapy will be applied after surgery to prevent or minimize the risk of cancer recurrence for such patient. Currently there is no clinical standards for assessing the risk of post-surgery recurrence of date. Great number of prognostic factors has been examined due to the growing number of literature in identifying factors with predictive capability of patient survival.

The overall prognosis for patients with COPD and lung cancer is worse than that of patients with lung cancer without COPD. Certainly those patients denied surgery, or offered only limited resection because of impaired pulmonary function, may not have the option of surgical cure. In addition, nonsurgical treatment options (limited by scant available supporting data and often reserved for poor surgical candidates) such as radiation therapy, radiofrequency ablation, stereotactic body radiotherapy, and cryotherapy have resulted in poorer survival and increased rates of local recurrence compared with surgical treatment. The impact of COPD on survival after resection of lung cancer is uncertain. One series demonstrated that for patients with stage I disease and low predicted postoperative FEV_1 values (less than 40%), 5-year survival post resection is significantly lower, compared with patients with better lung function (35 vs. 65%)[6]. Given that the immediate postoperative mortality and rates of tumor recurrence were similar in the two groups, the increased 5-year mortality in the high-risk group was presumed to be due to non-oncological factors. This lower survival rate for patients with severely limited respiratory reserve is consistent with reports by other groups.

Current international standard, Response Evaluation Criteria in Solid Tumors (RECIST) guideline suggests that the response of the tumor to the targeted treatment is evaluated based on the tumor size (measured by the longest diameter of the tumor) and their suitability for accurate repeated measurements. According to RECIST, the sum of the longest diameter (LD) for all identified lesions is considered as a reference to evaluate the tumor response to the treatment. However to subjectively measure tumor size in one-dimension and evaluate the size change during the multiple (sequential) CT image examination is not reliable, as there will be larger inter-reader variability, and that this method has low correlation to the clinical outcome of the patients. As a result, identifying new quantitative image markers computed from CT images has received increasing interest recently.

Chapter 2: Materials

Dataset

The test dataset for this study was retrospectively acquired under an institutional review board approved data collection protocol from the first affiliated Hospital of Guangzhou Medical University, Guangzhou China. The dataset includes image of thoracic CT examination of 107 patients who underwent lung cancer diagnosis and treatment in the hospital. All of the patients in this dataset were diagnosed with the verified stage I NSCLC. Based on the current clinical guideline, a lung surgery was performed on each patient to reset verified malignant lung tumor. After lung surgery, the tumor specimens were extracted. Table shows the demographic information of 107 patients along with corresponding subjectively assigned scores in the cases of this testing dataset.

Table 1: Demographic information of the patients

Cancer recurrence	Age	Yes	No	total
Male				
	<= 60	6	23	
	>60	9	24	
Total		15	47	62
Female	<= 60	4	19	
	>60	7	15	
Total		11	34	45

Among these 107 stage I NSCLC patients, 62 are male and 45 are female. The age of these patients ranged from 39 to 85 years old. The average age 61 years old with a standard deviation of 9.2. Among these patients, 52 are younger than 60 years old and 55

are older than 60 year old. These 107 patients were divided in to two groups with and without cancer recurrence within the 3 years after the lung cancer surgery. Specifically, 26 patients were assigned to cancer recurrence group and 81 were in progression-free survival (without cancer recurrence) group. The table summarizes the distribution of these two groups of cases in tumor and cancer cell characteristics.

Table 2: Distribution of two groups of cases in tumor and cancer cell characteristics

	Cancer recurrence	Yes	No	Total
Tumor Density				
	uniform	7	16	23
	necrosis	12	37	49
	vacuoles	7	25	32
	inanition	0	3	3
Cell type				
	Squamous Cell Carcinoma	7	13	20
	Adenocarcinoma	16	63	79
	other	3	5	8
Tumor Size				
	<= 3cm	13	61	74
	>3cm & <=5cm	13	20	33
Tumor Boundary				
	smooth	1	4	5
	lobulation	4	22	26
	spiculation	21	55	76
Relation to pleura				
	Normal	4	18	22
	Last card pleura	6	28	34
	pleura indentation	16	35	51

All CT image of these patients were acquired using a 16-detector based Toshiba Aquilion CT machine. In the CT image scanning protocol, X-ray tube voltage ranged

from 120 to 140KVP and the current ranged from 140 to 340mAs depending on patent body size. The CT image slices were reconstructed with an image size of 512x512 pixels with pixel size ranged from 0.51 to 0.74mm also depending on the patient body size. The image slice thickness of all CT images in this dataset was 2mm.

Chapter 3: Methods

The CAD model is developed to automatically segment the lung region from all the slices of the CT scan. Once the lung is segmented from the original image tumor region is segmented using a semi-automated method. After the tumor is segmented we apply the density mask to get the emphysema in the lung tissue. Using the tumor and emphysema region total of 51 features are computed.

Segmentation of Lung

The CAD model operates on all the slices simultaneously and performs a series of operation to segment lung region. It reads in the first slice and gets the slice thickness information present in the Dicom header. If the slice thickness is less than 5mm, the image is passed through a low pass filter or else it will directly proceed to the next step. Next, we try to separate background noise from the lung tissue, for that we set the background pixel values to -200Hu. Then we apply multilevel thresholding to remove the background pixels. Once the background pixels are removed using rescale intercept we get back the original pixel values for the lung tissue. Then we perform region labelling to remove unwanted artifacts and also to fill in small cavities in the lung tissue formed due to thresholding. Even the improperly identified airways are removed by region labelling. Once this is done we get the segmented lung tissue[7]. The same procedure is repeated for all the slices iteratively to get the complete segmented lung tissue.

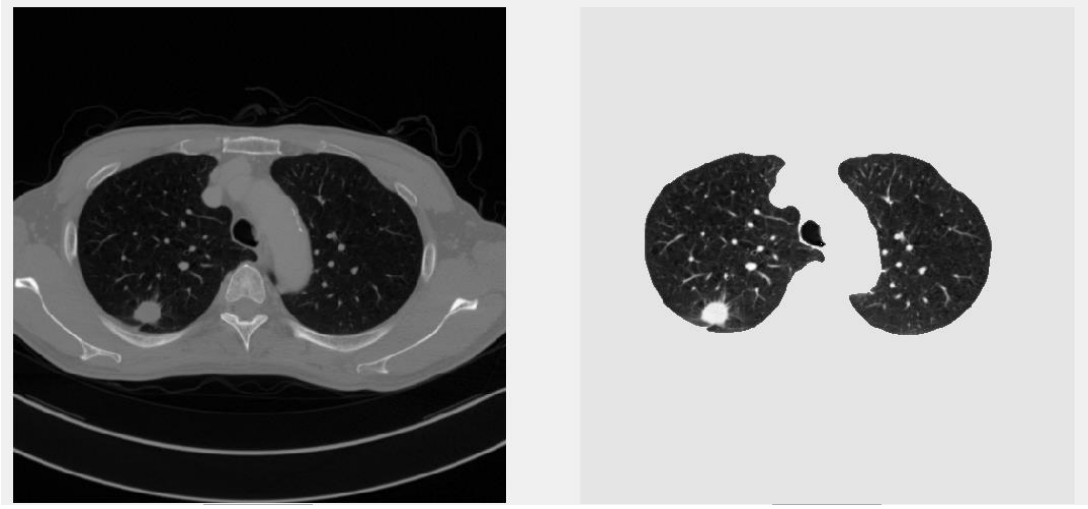

Figure 2: Segmentation of the Lung region

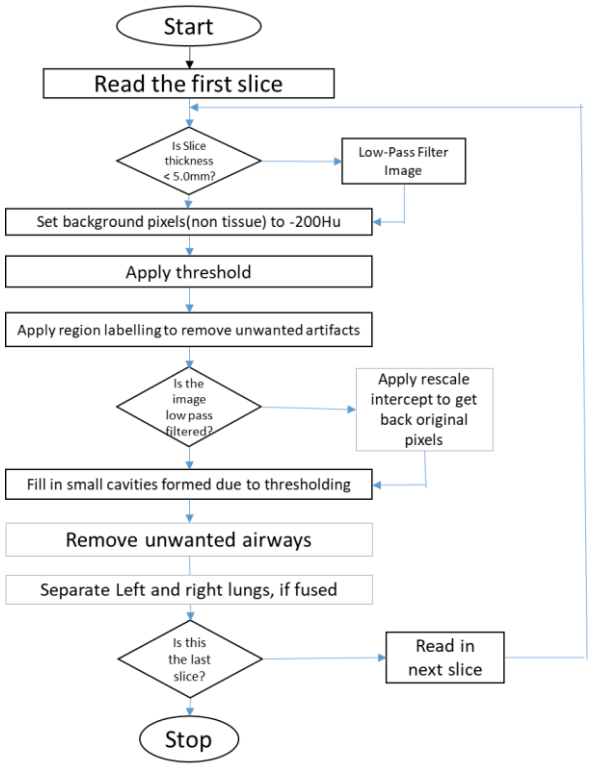

Figure2: Flowchart to segment Lung tissue

Segmentation of Lung Tumors

In this collected image dataset, each tumor center and diameter on the targeted CT image slice have been previously identified and marked by the radiology during the original CT image reading and interpretation. Using the slice number and the location the seed point is given. Based on this seed point, region growing algorithm is applied to compute the tumor. The marked tumor seed in the center slice was mapped in to the next adjunct CT image slices to segment the tumor area depicting on the next slice. This mapping process was iteratively performed until the scheme either reached the slice without remaining tumor area being detected and segmented by the CAD scheme or reached a slice with a distance larger than tumor diameter from center slice. Specifically, in each involved image slice, first a conventional region growing algorithm was applied with an empirically selected threshold of CT number (-450 HU)[8]. This step works well in segmentation of well-circumscribed lung nodules. However, this dataset contained other types of nodules such as Juxtapleural and Vascularized nodules are nodules connected to vessels or other structures. In order to provide a more accurate segmentation for such nodules two other image processing steps were applied.

In order to remove the attachment of Juxtapleural tumors from chest wall, modified convex hull function based algorithm was applied (initially introduced by Kuhgnik et al de[9]. According to the anatomical fact that lung is mostly convex, the convex hull function could efficiently and adaptively remove thoracic lesions (Juxtapleural tumors) from the chest wall. However, applying the convex hull function algorithm is also likely to generation a few minor (isolated) regions due to image noise.

The scheme then applied a region labeling algorithm to remove small regions while maintaining the segmented tumor region.

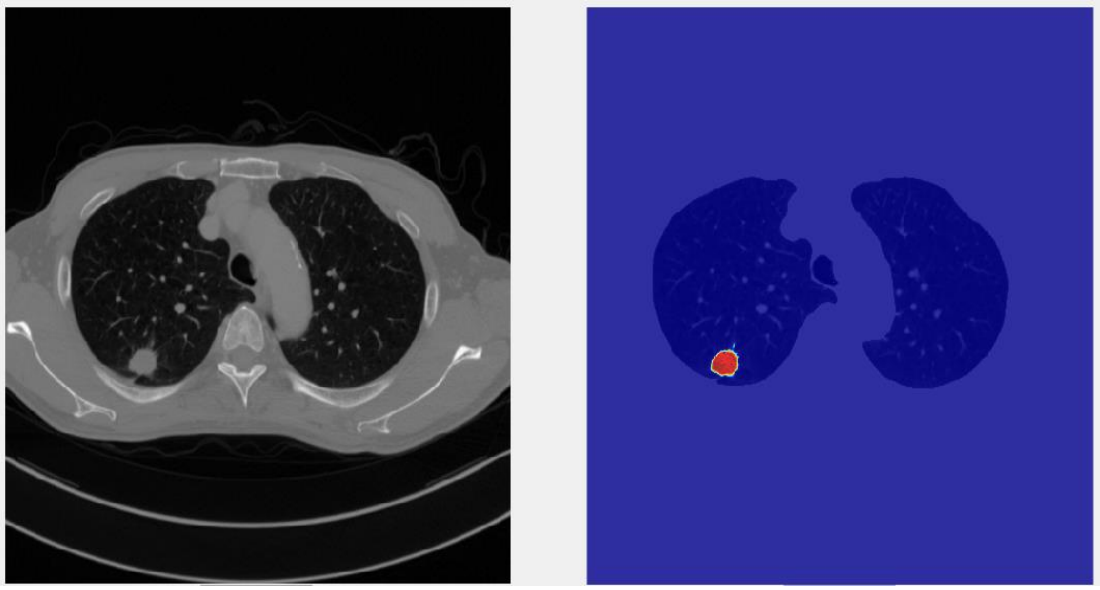

Figure 3: Segmentation of tumor region

Next, to remove vessels and structures connected to tumor regions, distance based morphological operation as proposed by Kuhgnik et al were applied. After tumor were segmented using the initial region growing algorithm and convex hull function the scheme fitted a rectangular window to the initial tumor boundary that was also centered on tumor. Then a Euclidean distance transform of the initial window was taken which converted it into distance map E that contains the minimum distance of each pixel of the tumor region to the tumor boundary pixels. Afterwards, a seed optimization was done by searching for the pixel C with the longest distance in the neighborhood of the initial given seed, and a new radius for the tumor was calculated based on this new seed:

$$r = E(c)$$

Using this process, a normalized distance map E was obtained

$$E=E/r$$

A small window was placed to cover the targeted tumor. If a region growing starts from the center C, it reaches the boundary of the image through the vasculature and the result would not be smoothly detected tumor. Distance based morphological operation was performed on the distance map to remove vascular connections. The erosion and dilation were based on the shortest distance of each pixel to the tumor boundary. To perform erosion and dilation, and adaptive threshold was calculated from the normalized distance map E with the initial value of 1. The threshold was lowered from 1 to a value with which the boundaries of the window were reached.

Condition and the region growing algorithm was reapplied on this slice to segment the tumor region. The similar semi-automated tumor segmentation and performance evaluation method has been reported in the previous studies to segment lung nodules and breast masses to reduce or minimize the erroneous measurement results of the features values computed.

Density Mask

Previous studies have indicated that lung cancer and chronic obstructive pulmonary disease (COPD) are closed associated and both of them are primarily caused by cigarette smoking [10], [11]. Around 50%-90% of lung cancer patients suffer from COPD. Among the variety of COPD symptoms, emphysema is an important one with higher association to lung cancer [11]. However, whether and how the emphysema of the NSCLC patients affect their prognosis or risk of cancer recurrence has not been well investigated before.

This study aims to investigate this issue by integrating emphysema related features to the QI marker or prediction model to predict prognosis of NSCLC patients. For this purpose, a density mask is applied on the lung tissue to analyze the emphysema present in the patients. For this we apply a threshold based on the lung attenuation for all the slices to get the total emphysema. Previous studies have reported two different threshold values -910HU and -950HU for the density mask. So both the thresholds are tested in this study[12][13]. First -910HU threshold value is applied to get the emphysema region in the lung and all the emphysema related features are computed then -950HU is applied and the features are calculated.

Quantitative Tumor and Emphysema related Features

The CAD scheme extracted a total of 56 features in which 35 are tumor-related morphological, CT number distribution and texture features and 21 are emphysema related features from the CT image. The initial 35 features extracted from the tumor region were:

1. **Tumor volume**: This is the total volume of all the tumor voxels. It is calculated by number of pixels inside tumor region X $(\text{pixel size})^3$, Where $(\text{pixel size})^3$ is the voxel volume.
2. **Density or Mean pixel values within the tumor:** It is related to the degree of tumor density and heterogeneity within the tumor.
3. **Standard deviation of density:** It computes the standard deviation of all the pixel values within the tumor.
4. **2D Volume:** This feature calculates the volume in the central region of the tumor

5. **Tumor diameter marked by the radiologist:** This feature is assessed and measured using RECIST.
6. **Convexity:** It describes the smoothness of edges of the tumor, which can be calculated as follows:

$$\textbf{Convexity} = \frac{\boldsymbol{Tu}mor\ region\ area}{convex\ region\ area}$$

7. **Max radius:** All the possible radii between the center and all the tumor surface pixels are computed, among them the maximum radius is considered as the max radius.
8. **Contrast:** It is the difference between the mean of inner ring tumor pixels and the mean of surrounding outer ring boundary pixels.

$$\textbf{Contrast} = \overline{Iinner} - \overline{Ioutter}$$

9. **Skewness:** Skewness measures the asymmetry of the probability distribution curve of all the tumor pixel values about its mean.

$$\textbf{Skewness} = \sqrt{\boldsymbol{N}} * \frac{\sum_{i=1}^{N}(\mathrm{Ii} - \bar{I})^3}{\sum_{i=1}^{N}(\mathrm{Ii} - \bar{I})^{\frac{3}{2}}}$$

10. **Kurtosis**: Kurtosis measures the "tailed-ness" of the tumor density distribution when comparing to the standard normal distribution:

$$\textbf{Kurtosis} = \sqrt{\boldsymbol{N}} * \frac{\sum_{i=1}^{N}(\mathrm{Ii} - \bar{I})^4}{\sum_{i=1}^{N}((\mathrm{Ii} - \bar{I})^2)^2}$$

11. **STD ratio:** It is defined as the ratio of STD of tumor intensity to the tumor boundary intensity.

12. **STD RL:** It is the ratio of the standard deviation and average of all the radii between center and tumor surface pixels.

$$\textbf{STD RL} = \frac{\text{STD of radii}}{\text{mean of radii}}$$

13. **Energy**: It is a sum of the squared tumor pixel values.

$$\text{Energy} = \sum_{i=1}^{N} I_i^2$$

14. **Entropy:** It describes the randomness/ uncertainty in an image.

$$\textbf{Entropy} = \sum_{i=1}^{N_l} \boldsymbol{P_i \, log_2 P_i}$$

Where P is the first order histogram of tumor pixels with N_l discrete intensity levels.

15. **Maximum CT number within tumor pixels:** It is the maximum value of all the tumor pixels.
16. **Mean absolute deviation:** It is defined as the mean absolute deviation between the tumor pixel value and the average tumor intensity:

$$\textbf{Mean absolute deviation} = \frac{1}{N} \sum_{i=1}^{N} |\boldsymbol{I_i - \bar{I}}|$$

17. **Median:** It is the median value of all the tumor pixels.
18. **Minimum CT number within tumor pixels:** It is the minimum value of all the tumor pixels.
19. **Range of pixel values within tumor:** It measures difference between the maximum and minimum values of all the tumor pixels.
20. **RMS:** It is the root mean square value of tumor pixels.
21. **Uniformity:** It is the measure of histogram randomness and can be computed as follows:

$$\textbf{Uniformity} = \sum_{i=1}^{N_l} \boldsymbol{P_i^2}$$

22. **Auto Correlation of tumor pixels:** It computes the autocorrelation between the tumor pixels.
23. **Tumor Cluster Shade of central tumor region:** It measures the skewness of central region of the tumor.
24. **Tumor Cluster Prominence:** It is the measure of asymmetry of the tumor region.
25. **Tumor pixels dissimilarity:** It measures the dissimilarity between tumor pixels. It is calculated as follows;

$$\textbf{Dissimilarity} = \sum_{i,j=0}^{N-1} \boldsymbol{P_{i,j}}\,|\boldsymbol{i-j}|$$

26. **Homogeneity:** It measure the homogeneity in the tumor pixels. It is calculated as following:

$$\text{Homogeneity} = \sum_{i,j=0}^{N-1} \frac{P_{i,j}}{1+(i-j)^2}$$

27. **Maximum Probability:** It is calculated as $\max P_{i,j}$.
28. **Variance:** It is calculated as following:

$$\text{Variance} = \sum_{i,j=0}^{N-1} P_{i,j}(i-\mu_i)^2$$

The GLRL was computed through a package developed by Wei using a zigzag method in four steps to: 1) determine direction, 2) perform zigzag scan, 3) obtain new sequences, and 4) calculate run-length matrix. From each GLRL matrix, the scheme computed 7 texture features. Since each texture feature had different values in each of 4 directions, the final value was represented as the mean of 4 values calculated in 4 directions. Run-

length matrix of p (i, j) is the number of runs with pixels of gray level I and run length j in an image. For such run length matrix M is the number of gray levels and N is the maximum run length, n_r is the total number of runs and n_p is the number of pixels in the image[8].

Following are the features extracted:

29. **Short Run Emphasis (SRE):** Calculated using

    SRE = $\frac{1}{n_r}\sum_{i=1}^{M}\sum_{j=1}^{N}\frac{P(i,j)}{j^2}$ and tends to emphasis on short runs.

30. **Long Run Emphasis (LRE):** Calculated using

    LRE = $\frac{1}{n_r}\sum_{i=1}^{M}\sum_{j=1}^{N}P(i,j).j^2$ and tends to emphasis on long runs.

31. **Gray-Level Non-Uniformity(GLN):** Calculated using

    GLN =$\frac{1}{n_r}\sum_{i=1}^{M}(\sum_{j=1}^{N}P(i,j))^2$, and increases as the gray-level outlier dominates the histogram.

32. **Run-Length Non-Uniformity (RLN):** Calculated using

    RLN =$\frac{1}{n_r}\sum_{i=1}^{N}(\sum_{j=1}^{M}P(i,j))^2$, measures the non-uniformity of the run lengths. This feature will have low values if the runs are equally distributed throughout the lengths.

33. **Run Percentage (RP):** calculate using $\frac{n_r}{n_p}$ will have its lowest value in the images with most linear structures.

34. **Low Gray-Level Run Emphasis (LGRE):** It extracts gray level information in the run-length matrix, and is calculated using $\frac{1}{n_r}\sum_{i=1}^{M}\sum_{j=1}^{N}\frac{P(i,j)}{i^2}$

35. **High Gray-Level Run Emphasis (HGRE):** It extracts gray level information in the run-length matrix, and is calculated using $\frac{1}{n_r}\sum_{i=1}^{M}\sum_{j=1}^{N} P(i,j).i^2$

Similarly for the emphysema 21 features are computed. Two set of similar features for two different density masks applied to compute the emphysema are calculated.

The following are the features computed for the emphysema

Table 3: Emphysema features

Feature class	Feature description
Shape	emphysema volume, emphysema percentage, convexity
Density	Energy, Entropy, Minimum, Maximum, mean, median, range, Uniformity, density STD, skewness, kurtosis
Texture	7 gray level texture based features were computed

Feature Selection

The initial feature pool includes total 56 features in which 35 are tumor related features and 21 are emphysema related features. In order to reduce dimensionality of the feature space and increase robustness of the multi-feature based risk prediction model, a CFS Subset Evaluation attribute selection in WEKA data mining software package with a Best-first heuristic feature [14] was applied to select a subset of effective and non-redundant feature from the initial pool of 51 image feature based on the important sorting of feature in predicating the risk of cancer recurrence. The feature selection method used

in this study, evaluated the worth of a subset of features with respect to discriminative power of each individual feature along with degree of the redundancy between features. As a result, a small and optimal set of 8 image features was built. Among them, 5 are computed from the segmented tumors (which are Stdratio, Max, dissimilarity, HGRE, 2D volume and max tumor diameter as discussed in previous subsection of Quantitative Tumor and Emphysema related Features) and 3 are computed from emphysema regions.

Machine Learning Model

Next, a machine learning method was selected and applied to combine the selected optimal features and build the prediction model or QI marker to predict prognosis of NSCLC patients. Although many different machine learning classifiers can be used for this purpose, in this study, due to unbalanced test dataset, random forest tree based classifier was selected and built [15]. Random Forest (RF) has become an attractive ensemble method in data mining. As a classifier integration method, RF have the features of classifying fast and training simple and is suitable for feature selection according to variable importance[16]. The primary advantage of a random forest is its unexcelled accuracy among current algorithms and it has methods for balancing error in class population in unbalanced datasets[19].

Specifically, three random forest classifiers or models were built in this study. The first model was built to combine 6 selected tumor features and predict the likelihood or risk of a stage I NSCLC patient having cancer recurrence after cancer surgical treatment. In this study, WEKA data mining software package was used to build the classifier. Due to

unbalanced data (26 positive and 81 negative cases), in order to achieve a more balanced optimization results in predicting the cases in two classes, a synthetic minority oversampling technique (SMOTE) [17] was applied to add synthetic data to double the "positive" test cases from 26 to 52. As a result, total 133 cases were used to build and optimize the random forest tree based classifier. The second model was built using 3 selected emphysema related QI features and the third model was built using 9 QI features by combining 6 tumor related features and 3 emphysema related features to predict the cancer recurrence risk, respectively. The same training and testing method applied in the first random forest model using 6 tumor related QI features only was applied to build the second and the third models [18].

Performance Assessment and Data Analysis

First, all the selected 6 tumor features and 3 emphysema features were processed individually by using receiver operating characteristic (ROC) fitting program (ROCKIT) to compute the area under the ROC (AUC). AUC was used as an assessment index to analyze their performance to predict or classify the test cases of the dataset into two classes namely, positive for the cases with cancer recurrence after surgery within 3 years and negative for cases without cancer recurrence within 3 years. The correlation coefficient of AUC values between these features were also computed and compared.

Next, in order to evaluate performance of a random forest model, a leave-one-case-out (LOCO) cross-validation method was applied. Each of 107 cases in the dataset was selected as an independent testing case in LOCO process to yield a CAD or QI marker generated likelihood score to predict the risk of cancer recurrence. AUC value was calculated using the likelihood or prediction scores of these 107 original test cases and

the ROCKIT program. Then, the prediction performance levels or AUC values of three random forest classifier based QI markers or models built using 5 tumor related QI features, 3 emphysema related features, and total 8 QI features, were compared.

In addition, besides the AUC values, an operation threshold was applied to divide the cases into two groups with high and low risk of having cancer recurrence. From the classification result, a confusion matrix was generated. Then, from the confusion matrix, the prediction accuracy, and positive prediction and negative prediction values (PPV and NPV), sensitivity, specificity were also calculated and compared by using the three different random forest models.

Chapter 4: Results

The table 4 shows the AUC values for the selected 6 tumor related features individually in predicting the cancer recurrence. Next table show the AUC values for the selected 3 features each related to emphysema computed using the threshold -910HU(Emphysema1) and emphysema computed using the threshold -950HU(Emphysema2). The result indicate that tumor features 1 and 5 had the maximum and comparable AUC values. In the Emphysema related features, feature 3 of Emphysema1 and feature 2 of Emphysema2 had highest AUC values. Tables summarize the correlation coefficients calculated between each tumor feature and emphysema features. 90% of the absolute values of the correlation coefficients is much smaller than $r < 0.5$, which indicates that the features used in the predictor are not highly correlated or redundant. This result demonstrates that combination of these features has the potential to add supplementary information or has discriminatory power to significantly increase performance in predicting the risk of lung cancer recurrence.

Table 4: comparing AUC values of individual tumor related features

Feature	AUC	STD	95% CI
1	0.78	0.06	[0.64, 0.88]
2	0.64	0.05	[0.53, 0.74]
3	0.59	0.06	[0.46, 0.72]
4	0.67	0.05	[0.55, 0.77]
5	0.67	0.05	[0.56, 0.77]
6	0.76	0.05	[0.64, 0.86]

1 Stdratio, 2 Max, 3 dissimilarity, 4 HGRE, 5 2D volume, 6 max tumor diameter

Table 5: comparing AUC values of individual Emphysema1 related features

Feature	AUC	STD	95% CI
1	0.632	0.07	[0.48, 0.76]
2	0.57	0.06	[0.44, 0.69]
3	0.66	0.06	[0.52, 0.78]

1 entropy of emphysema1, 2 auto correlation of emphysema1, 3 uniformity of emphysema1

Table 6: comparing AUC values of individual Emphysema2 related features

Feature	AUC	STD	95% CI
1	0.56	0.07	[0.41, 0.70]
2	0.55	0.07	[0.41, 0.69]
3	0.59	0.06	[0.46, 0.71]

1 Entropy of emphysema2, 2 Uniformity of emphysema2, 3 homogeneity of emphysema2

As mentioned in previous chapter, the performance of the classifier were evaluated based on their AUC values. AUC value for the tumor feature based classifier involving 5 selected tumor features was 0.794 with a standard error of 0.05 and a 95% confidence interval (CI) of [0.68, 0.87], Whereas the AUC values for the Emphysema related features were 0.700 with a standard error of 0.07 and 95% confidence interval of [0.54, 0.82] and

0.71 with a standard error of 0.05 and 95% confidence interval of [0.59, 0.81] for Emphasema1 and Emphysema2 respectively. However, this difference is not statistically significant difference (p value = 0.4). The results shows that the emphysema features or tumor features alone to build a machine learning classifier did not have a significant AUC value, using combined feature based classifier enabled to yield a significantly higher AUC value.

Table 7: comparing correlation coefficients between tumor features and emphysema1 features

Feature	1	2	3	4	5	6	7	8	9
2	-0.1504								
3	-0.16111	0.363701							
4	-0.11218	0.419922	0.378207						
5	-0.21653	0.304951	0.69608	0.47185					
6	-0.31725	0.121846	0.016811	-0.06618	-0.0099				
7	0.017228	-0.02379	-0.1381	0.070104	-0.04891	-0.07496			
8	-0.07462	-0.00328	0.175647	-0.08576	0.063511	0.051291	-0.93565		
9	-0.12426	0.411489	0.446326	0.990371	0.506635	-0.06431	0.05221	-0.05909	

Standard deviation ratio, Maximum CT number in tumor region, Dissimilarity in tumor region, HGRE, 2D volume, max tumor diameter, Entropy of Emphysema1, Autocorrelation of Emphysema, Uniformity of Emphysema.

Table 8: comparing correlation coefficients between tumor features and emphysema2 features

Feature	1	2	3	4	5	6	7	8	9
1	-0.14635								
2	-0.16666	0.364177							
3	-0.12274	0.421638	0.378202						
4	-0.21429	0.304107	0.696751	0.473599					
5	-0.31162	0.119946	0.017282	-0.06418	-0.01211				
6	-0.37869	-0.05219	0.090315	0.072765	0.119675	0.108774			
7	0.154233	0.08307	-0.03264	-0.05396	-0.04879	-0.01945	-0.93568		
8	-0.1398	0.416407	0.445853	0.990999	0.510468	-0.06134	0.078142	-0.05254	

Standard deviation ratio, Maximum CT number in tumor region, Dissimilarity in tumor region, HGRE, 2D volume, max tumor diameter, Entropy of Emphysema1, Autocorrelation of Emphysema, Uniformity of Emphysema.

Table 9 shows the AUC values of obtained by tumor, emphysema, combining tumor with emphysema1 and combining tumor with emphysema2. The results showed that the maximum AUC = 0.86 ± 0.3, which is significantly higher than the AUC values generated using either tumor features or emphysema features as computed by the ROCKIT program. Figure shows and compares the ROC curves generated using tumor features emphysema features and the optimal fusion of both. The figure below compares the ROC curves for combined features ROC, Tumor features based ROC and emphysema features based ROC.

Table 9: Comparing Performance of different Image Markers

Classifier	**Performance**
Tumor with Emphysema1	0.865 ± 0.03
Tumor with Emphysema2	0.823 ± 0.05
Tumor	0.79 ± 0.04
Emphysema1	0.70 ± 0.06
Emphysema2	0.71 ± 0.05

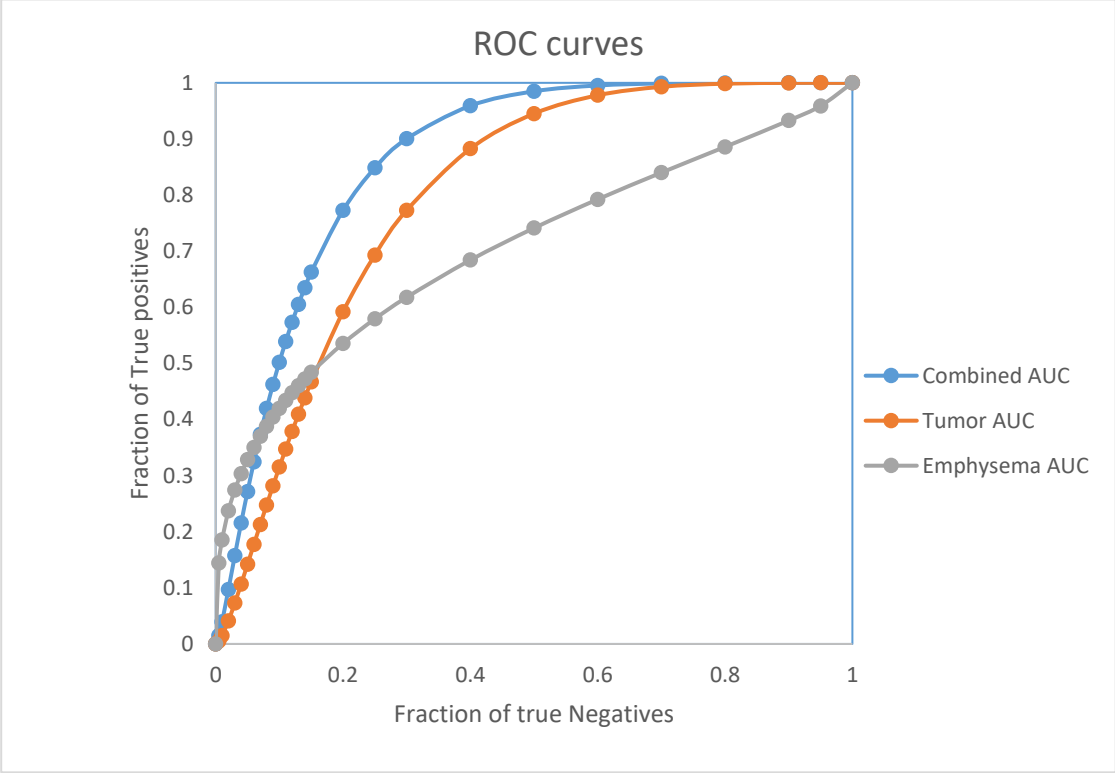

Figure 4: ROC curves comparing tumor feature based classifier, emphysema feature based classifier and combine feature based classifier

Table 9, Table 10, Table 11, Table 12 shows the confusion matrices obtained using tumor and emphysema feature fusion, and their individual feature based classifiers. The feature fusion marker predicts 84 cases as no cancer recurrence cases, among them 70 cases are DFS. Similarly this marker predicts 23 cases as cancer recurrences cases and 15 of them are cancer recurrence cases. This marker yields an overall prediction accuracy of 79% with prediction sensitivity of 91% and corresponding specificity of 64%. For the tumor feature based marker 84 and 23 are no cancer recurrence and cancer recurrence cases are predicted, among which 63 cases have no cancer recurrence and 16 cases have cancer recurrence. Therefore the PPV and NPV are 75% and 69% respectively with an overall prediction accuracy of 73%. Similarly for Emphysema 1, the overall prediction accuracy is 72% with prediction sensitivity of 82% and corresponding specificity of 44%.For the Emphesema2 the overall prediction accuracy is 76% with prediction sensitivity of 87% and corresponding specificity of 51%.

Table 10: Confusion matrix generated using the feature fusion based image marker

Actual / Prediction	DFS-Yes	Cancer Recurrence	
DFS-Yes	70	14	PPV = 0.83
Cancer Recurrence	8	15	NPV = 0.65
	Sensitivity = 89%	Specificity = 51%	

Table 11: Confusion matrix generated using the tumor feature based image marker

Actual / Prediction	DFS-Yes	Cancer Recurrence	
DFS-Yes	63	21	PPV = 0.75
Cancer Recurrence	7	16	NPV = 0.69
	Sensitivity = 90%	Specificity = 43%	

Table 12: Confusion matrix generated using the emphysema1 feature based image marker

Actual / Prediction	DFS-Yes	Cancer Recurrence	
DFS-Yes	66	15	PPV = 0.81
Cancer Recurrence	14	12	NPV = 0.46
	Sensitivity = 82%	Specificity = 44%	

Table 13: Confusion matrix generated using the emphysema2 feature based image marker

Actual / Prediction	DFS-Yes	Cancer Recurrence	
DFS-Yes	65	16	PPV = 0.80
Cancer Recurrence	9	17	NPV = 0.65
	Sensitivity = 87%	Specificity = 51%	

Chapter 5: Discussion

This study demonstrated feasibility of developing a new quantitative imaging (QI) marker to predict prognosis of early stage non-small cell lung cancer (NSCLC) patients after surgery among the stage I NSCLC patients. The study yielded higher prediction performance and also has a number of unique characteristics, which are discussed as follows in this section.

Prediction of lung cancer recurrence risk after initial surgery among the stage I NSCLC patients has high clinical impact on overall efficiency of treatment management and lung cancer screening program. Specifically, due to the recent promotion of lung cancer screening programs using low-dose CT examinations, identifying lung cancer prognostic factors will have a more important role in reaching the ultimate goal of reducing cancer mortality rate. Although developing CAD schemes of lung nodules using chest CT images has been well investigated previously by many research groups, but in this study a new CAD scheme is developed to predict the cancer risk by generating new quantitative image feature.

A new set of tumor related features along with lung emphysema features were computed from the CT images and a machine learning classifier was built to predict lung cancer prognosis. Using best-first heuristic feature search algorithm, 8 non-redundant features were selected from the pool of 51 features. The study also showed that optimally combining multiple image features using a machine learning classifier can significantly increase the prediction performance.

Quantitative image features analysis method can be more efficient and reliable to integrate these features by eliminating inter and intra observer variations. This study applies CAD scheme on predicting lung cancer recurrence risk after surgery for stage I NSCLC patients by generating new quantitative imaging marker. In this study we combined tumor features to emphysema based features to generate new quantitative imaging marker.

Apart from the data analysis mentioned before, several other experiments were performed that gave interesting features. First the optimal feature selection is important in developing a CAD- based quantitative image feature analysis scheme. Training random forest network with all the 51 features yielded a lower performance that the classifier training using 9 selected features

Second, when using original dataset of 107 cases to train two classifiers, the tumor feature based classifier yielded and the emphysema based classifier yielded a maximum value of which are significant advantage of using a SMOTE method to generate synthetic data and balance the number of training cases in two risk classes in training a machine learning classifier.

Despite the promising results, this preliminary study had a few limitations. First the size of the dataset is small, which cannot represent the general population of stage I NSCLC patients. Hence, the robustness of the reported results needs to be tested in future studies with a new large and more diverse datasets acquired from different CT machines and other image scanning protocols.

Second, only 35 tumor related features and 21 emphysema related features were computed. Other lung background image features i.e. features related to COPD which may also provide some supplementary prediction information, have not been incorporated into the quantitative image feature analysis scheme.

Third, in this study a semi-automatic scheme was performed to segment lung tumors. The segmentation results were visually examined. As a results, a small fraction of tumor segmentation boundary contours were manually corrected when the substantial errors were visually detected. Due to the lack of "ground-truth" and the potential inter-observer variability, this is not an optimal tumor segmentation method, which may create errors in computing tumor related features. However, it is believe that this semi-automated segmentation is an efficient approach to perform this proof-of-concept study.

Fourth, we compared two different threshold levels to apply the density mask for computing emphysema. We analyzed their performance individually and also by combining them to the tumor features. We used AUC to analyze the performance. Individually the emphysema computed by applying -910HU had an AUC value of whereas for the emphysema computed with -950 had an AUC value of. But this difference is not statistically significant. So by using emphysema computed using -910HU had a better performance than the other.

Chapter 6: Conclusion

Although early stage non-small cell lung cancer (NSCLC) patients have relatively higher survival rates (i.e., five-year survival rates of 49% and 45% for stage IA and IB stage NSCLC patients, respectively), cancer recurrence rates after surgery of resecting malignant tumors may vary from 30% to 60%. In order to more effectively treat and manage stage I NSCLC patients, it is important to develop an effective clinical marker or predictive model to more accurately predict cancer prognosis (i.e., risk of cancer recurrence or likelihood of disease-free survival (DFS)) after cancer surgery. As a result, patients with a higher risk of cancer recurrence should receive specific or targeted chemotherapy after surgery to minimize the risk of cancer recurrence. Thus, it is important to identify or develop more effective clinical markers to stratify early stage NSCLC patients into high and low risk of cancer recurrence.

Aiming to better address and/or help solve this clinical issue, we in this study investigated a new computer-aided quantitative image analysis method to predict the risk of lung cancer recurrence or DFS of the stage I NSCLC patients after lung cancer surgery. For this purpose, we developed a new computer-aided detection (CAD) scheme to automatically segment lung regions and the malignant tumors depicting on CT images acquired before surgery and compute image features related to tumor shape, size, circularity, density heterogeneity, and lung background tissue patterns. After selecting optimal features, we train and build a machine learning based classifier to generate a new quantitative imaging marker to predict DFS of lung cancer patients. The goal of this study is to test feasibility of applying this new quantitative imaging (QI) marker to predict DFS of the early stage NSCLC patients.

References

[1] S. J. Swensen *et al.*, “Radiology CT Screening for Lung Cancer : Five-year Prospective,” *Cancer*, pp. 259–265, 2005.

[2] M. Jamal-Hanjani *et al.*, “Tracking the Evolution of Non–Small-Cell Lung Cancer,” *N. Engl. J. Med.*, vol. 376, no. 22, pp. 2109–2121, 2017.

[3] D. C. P. Cobben *et al.*, “Emerging Role of MRI for Radiation Treatment Planning in Lung Cancer,” *Technol. Cancer Res. Treat.*, vol. 15, no. 6, p. NP47-NP60, 2016.

[4] W. A. Kalender, “X-ray computed tomography,” *Phys. Med. Biol.*, vol. 51, no. 13, 2006.

[5] “7.” .American cancer society website: http://www.cancer.org/cancer/non-mall-cell-lung-cancer/detection-diagnosis-staging/staging.html.

[6] C. Poleri *et al.*, “Risk of recurrence in patients with surgically resected stage I non-small cell lung carcinoma: Histopathologic and immunohistochemical analysis,” *Chest*, vol. 123, no. 6, pp. 1858–1867, 2003.

[7] J. K. Leader *et al.*, “Automated Lung Segmentation in X-Ray Computed Tomography: Development and Evaluation of a Heuristic Threshold-Based Scheme,” *Acad. Radiol.*, vol. 10, no. 11, pp. 1224–1236, 2003.

[8] N. Emaminejad *et al.*, “Fusion of Quantitative Image and Genomic Biomarkers to Improve Prognosis Assessment of Early Stage Lung Cancer Patients,” *IEEE Trans. Biomed. Eng.*, vol. 63, no. 5, pp. 1034–1043, 2016.

[9] J. M. Kuhnigk *et al.*, “Morphological segmentation and partial volume analysis for volumetry of solid pulmonary lesions in thoracic CT scans,” *IEEE Trans. Med. Imaging*, vol. 25, no. 4, pp. 417–434, 2006.

[10] S. D. Shapiro, “Merging personalized medicine and biology of aging in chronic obstructive pulmonary disease,” *Am. J. Respir. Crit. Care Med.*, vol. 184, no. 8, pp. 864–866, 2011.

[11] D. O. Wilson *et al.*, “Association of radiographic emphysema and airflow obstruction with lung cancer,” *Am. J. Respir. Crit. Care Med.*, vol. 178, no. 7, pp. 738–744, 2008.

[12] J. G. Goldin, “Quantitative CT of emphysema and the airways,” *J. Thorac. Imaging*, vol. 19, no. 4, pp. 235–240, 2004.

[13] Z. A. Aziz *et al.*, “Functional impairment in emphysema: Contribution of airway abnormalities and distribution of parenchymal disease,” *Am. J. Roentgenol.*, vol. 185, no. 6, pp. 1509–1515, 2005.

[14] E. Burns, S. Lemons, R. Zhou, and W. Ruml, “Best-First Heuristic Search for Multi-Core Machines,” pp. 449–455, 1995.

[15] M. Khalilia, S. Chakraborty, and M. Popescu, “Predicting disease risks from highly imbalanced data using random forest.,” *BMC Med. Inform. Decis. Mak.*, vol. 11, no. 1, p. 51, 2011.

[16] P. G. Student, “Predicting Breast Cancer using Apache Spark Machine Learning Logistic Regression,” vol. 4, no. 2, pp. 6–9, 2017.

[17] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic minority over-sampling technique,” *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.

[18] W. G. Touw *et al.*, “Data mining in the life science swith random forest: A walk in the park or lost in the jungle?,” *Brief. Bioinform.*, vol. 14, no. 3, pp. 315–326, 2013.