Pre-treatment Radiomics Models for Clinical Outcomes in Early-stage Non-Small Cell Lung Cancer (NSCLC)

by

Nathan Allen Shaffer

Department of Medical Physics
Duke University

Date:_____

Approved:

_____
Fang-Fang Yin, Advisor

_____
Kyle Lafata

_____
Chunhao Wang

_____
Betty Tong

Thesis submitted in partial fulfillment of
the requirements for the degree of
Master of Science in the Medical Physics
Program in the Graduate School
of Duke University

2021

ABSTRACT

Pre-treatment Radiomics Models for Clinical Outcomes in Early-stage Non-Small Cell Lung Cancer (NSCLC)

by

Nathan Allen Shaffer

Department of Medical Physics
Duke University

Date:_____

Approved:

_____
Fang-Fang Yin, Advisor

_____
Kyle Lafata

_____
Chunhao Wang

_____
Betty Tong

An abstract of a thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in the Medical Physics Program in the Graduate School of Duke University

2021

Abstract

Lung cancer in accounts for 13% of all new cancer diagnoses and is the leading cause of cancer mortality in the United States (Bogart, 2017, Howlader, 2020). Non-small cell lung cancer (NSCLC) in particular, accounts for 80-85% of lung cancer diagnoses and is estimated to cause more than 130,000 deaths this year (American Cancer Society, 2021). Currently, the standard of care for early-stage NSCLC is surgery, with stereotactic body radiation therapy (SBRT) is becoming more accepted as the primary treatment option for patients who are medically inoperable. It remains controversial as to which method is optimal for marginal surgical patients, but it has been shown that SBRT and sublobar resection provide similar local tumor control rates and clinical outcomes in stage I NSCLC (Ackerson, 2018).

The goal of this research work was to develop pre-treatment radiomic models for surgical NSCLC patients to predict cancer recurrence. This was done by investigating two specific aims. The first was to **(1.) build radiomic models based on pre-treatment CT images from surgical patients and evaluate their performance in predicting cancer recurrence** and the second was to **(2.) build radiomic models based on pre-treatment CT images from surgical patients and evaluate their performance in predicting cancer recurrence**. Radiomic features were extracted from the contoured GTV's from pre-treatment CT scans of surgical and SBRT patients. To investigate the first aim, multivariate models were trained and tested on only surgical patients to find associations between the

extracted features and each clinical outcome. To investigate the second aim, these models were first trained on surgical data and tested on SBRT data to investigate the generalizability of each model across treatment modalities. Next, models were trained and tested on a pooled dataset to investigate potential associations of radiomic features with cancer recurrence independent of treatment. Models were evaluated by creating ROC curves and calculating the area under these curves (AUC's).

Models trained and tested on surgical patients showed a stronger association between radiomic features and non-local failure (maximum AUC of 0.82 ± 0.04) and a poor association with local failure (maximum AUC of 0.57 ± 0.04). This may suggest that radiomic features have limited value in predicting local recurrence since the GTV which is used in calculating these features is no longer in the body post-treatment. Despite this, it is difficult to draw strong conclusions based on the variability in the image parameters of surgical patients, such as slice thickness and x-ray tube current, which have been shown to affect feature values (Midya, 2018, Kim, 2019). This is supported by the degraded performance in these models when SBRT data was introduced, further increasing image variability.

Dedication

I would like to dedicate this thesis work to my dad who passed from lung cancer a few months ago as well as my mom, sister, and brother for supporting me throughout my graduate career.

Contents

List of Tables

List of Figures

Acknowledgements

I would like to acknowledge the following groups of people:

- My committee members, Dr. Fang-Fang Yin, Dr. Kyle Lafata, Dr. Chunhao Wang, and Dr. Betty Tong for their valuable time and input.
- The Duke Medical Physics Program for being incredibly supportive through this pandemic and shift to and online environment.
- My fellow classmates for their collaboration, insight, and support throughout the past 2 years

1. Introduction

1.1 Non-Small Cell Lung Cancer (NSCLC)

Lung cancer is currently the second most common cancer and the leading cause of death from cancer for both men and women and is estimated to result in over 130,000 deaths this year (Bogart, 2017, American Cancer Society, 2019). The most common type of lung cancer, resulting in 80-85% of all lung cancer diagnoses, are classified as non-small cell lung cancer (NSCLC) (American Cancer Society, 2019). These statistics on their own warrant increased effort in lung cancer research to help diagnose, manage, and treat new cases.

There are three main types of NSCLC including adenocarcinoma, squamous cell carcinoma, and large cell carcinoma. These types of cancer are grouped together because they have similar treatments and prognosis. Adenocarcinoma is a type of cancer that begins in mucus-producing glandular cells in the body. In the lung, these cells are found in the surface epithelium lining the airways as well as the connective tissue layer beneath the mucosal epithelium. This type of lung cancer commonly occurs in smokers but is also the most common lung cancer in non-smokers. Occasionally, this disease is found before spreading due to its location near the outer parts of the lung, thus resulting in a more favorable prognosis. Squamous cell carcinoma is a type of cancer that originates in thin, flat cells that make up the outer layers of structures in the body. In the lungs these cells line the inside of airways, and many cases start in the center of the lungs near the main

airways or bronchus. Like adenocarcinomas in the lung, smoking is a large contributing factor in squamous cell carcinomas (Harvard Health Online, 2020). The third form of NSCLC's are large cell lung carcinomas, which often develop in the outer regions of the lung but can originate from any part. Generally, large cell lung carcinomas can grow and spread quickly making it harder to treat.

The current standard of treatment for NSCLC is surgery, especially in early stages when the tumor is localized to the lung without significant spreading and metastasis. For stage 0 and I, this may be the only needed treatment for those that are healthy enough for the procedure. If the patient is not healthy enough for surgery, radiation therapy is also commonly used as a local treatment.

1.1.1 Treatment of NSCLC: Surgery

The standard treatment for early-stage NSCLC is surgery is surgical resection with mediastinal and hilar lymphadenectomy according to the National Comprehensive Cancer Network (NCCN) clinical practice guidelines (NCCN Clinical Practice Guidelines, 2021). The goal of surgery is to remove the tumor in its entirety, without leaving behind any residual disease. Currently, the standard operation is lobectomy (removal of one lobe of the lung), when feasible, for early-stage disease (Ginsberg, 1995). For stage 0 NSCLC, the cancer is well localized so a segmentectomy (removal of a small part of the lung) can often be enough. Depending on tumor location, complete surgical resection may require a bilobectomy (removal of two lobes of the right lung) or even

pneumonectomy (removal of an entire lung). For stage I NSCLC, it is common to undergo a lobectomy or segmentectomy for smaller stage I cancers. However, it has been shown that a lobectomy is still the preferred method due to its higher survival rate and lower locoregional recurrence (Raman, 2018, Bogart, 2017). Nearby lymph nodes are also removed to check for cancer spread. Treatment for stage II NSCLC generally includes a lobectomy or pneumonectomy if it is determined that the patient can recover. Regional lymph nodes are also removed and examined for cancer which are important in determining future steps for treatment which usually includes adjuvant chemotherapy.

Risks from surgery should also be considered when evaluating its viability as a treatment modality. These risk factors include bleeding, infection, myocardial infarction (heart attack), stroke, respiratory failure, and cardiac arrhythmia (Paul, 2010). There are many contributing factors that can increase the chance of these risks such as patient health and other competing comorbidities. These risk factors contribute to the hesitation to perform invasive procedures and have led to an interest in other less invasive techniques such as video-assisted thoracoscopy (VATS) in which a scope is inserted into the chest through a small incision through which a tumor can be removed. Patients who have access to this technique may recover more rapidly but may not be offered to all patients, depending on surgeon preference.

1.1.2 Treatment of NSCLC: Radiation Therapy

Radiation therapy has had a changing role as a treatment modality for early-stage NSCLC since the introduction of newer technology in image guidance, tumor tracking, and dose control. Traditionally radiation has been used for patients that are not candidates for surgery for various reasons according to the National Comprehensive Cancer Network (NCCN) clinical practice guidelines. In particular, stereotactic body radiation therapy (SBRT) has shown improving local control rates and survival rates for early-stage NSCLC (Parashar, 2013).

SBRT allows for the delivery of high-dose radiation to a tumor while simultaneously sparing healthy tissue. This technique has been made possible due to advancements in image guidance, allowing for precise tumor tracking and margin control. SBRT allows radiation delivery in high doses per fraction up to 20 Gy as opposed to 2 Gy in other radiation therapy treatments. In this technique, multiple beams shaped by multileaf collimators (MLC's) are used that converge to an isocenter to deliver high doses while sparing nearby tissues and structures such as the spinal cord or heart. The dose profile can be further shaped using inverse planning optimization techniques and dynamic MLC movement commonly referred to as intensity-modulated radiation therapy (IMRT).

1.1.3 Comparing Surgery and SBRT

Surgery has kept its place as the standard of care for early-stage lung cancer since it is well tested, reliable, and readily available. Despite this, there are notable problems with deferring to surgery as a default treatment modality. Surgery exposes the patient to potential risks with anesthesia, infection, difficult recovery that can last weeks or more, and other complications. Radiation therapy tends to have fewer initial risks and a less volatile recovery period, with the most common side effects consisting of nausea, itchiness, redness of the skin, and tiredness.

Due to the promising results of SBRT in early-stage NSCLC for inoperable patients, there has been an increased interest in utilizing SBRT for operable early-stage lung cancer (Song, 2009). There have been multiple studies that have shown successful utilization of SBRT for operable patients such, one of which showing 1- and 3-year survival rates of 94.7% and 84.7% respectively (Lagerwaard, 2012). In a similar fashion there have also been studies that have shown that SBRT and sublobar resection provide similar rates of local tumor control and clinical outcomes in stage I NSCLC (Ackerson, 2018). These studies suggest the need for further investigation into methods that can help determine the best treatment modality for a given patient with early-stage NSCLC (McDonald, 2017).

1.2 Radiomics

The term radiomics refers to the study of biological information or data derived specifically from radiological images. Radiomics is a method that extracts a significant number of features from radiographic images using specific data-characterization algorithms that can be used to quantify certain disease characteristics quantitatively (Lambin, 2012). This approach aims to enhance or provide additional data to that which is already available to clinicians to aid in clinical decision making.

The radiomic analysis method has been proposed due to the significant improvement in medical imaging and computing in the past several decades. This has allowed for better quality images to be taken, stored at a higher capacity, and analyzed quicker. Consequently, the extraction of multiple features from a set of images based on the spatial relationship of intensity levels, descriptors of shape, and texture heterogeneity have made it an appealing candidate for analysis using different techniques used in artificial intelligence. These techniques have been used in multiple studies to show an association between radiomic features and clinical outcomes.

Radiomics has demonstrated value in predicting clinical outcomes for lung SBRT patients in areas such as such as patient survival and recurrence rates (Huynh, 2016, Lafata, 2019a). However, fewer studies have focused on their application to surgery patients. This demonstrates a need for investigation of the viability of radiomic features in predicting clinical outcomes for surgical patients.

1.3 Purpose and Specific Aims

The purpose of this thesis work is to develop pre-treatment radiomic models for surgical NSCLC patients to predict cancer recurrence. To achieve this purpose, this work focused on two specific aims.

1. **Specific Aim 1:** To build radiomic models based on pre-treatment CT images from surgical patients and evaluate their performance in predicting cancer recurrence.
2. **Specific Aim 2:** To generalize our surgical models by incorporating features from pre-treatment SBRT images in model training and testing.

2. Methods

The overall workflow for this study can be described in six main stages: acquisition, segmentation, feature extraction, data splitting, modeling, and analysis.

2.1 Acquisition

2.1.1 Patient Scans

For each patient, pre-treatment CT scans were acquired under free-breathing conditions. For SBRT patients, images were acquired on either a GE scanner with 2.5mm slice thickness and 120 kVp, or a Siemens scanner with 2 mm slice thickness 120 kVp. For surgical patients, images were acquired on one of 6 GE scanners, 1 Siemens scanner, or 1 Phillips scanner with slice thicknesses ranging from 0.6-5.0 mm at 120 kVp. Surgical scans also had varying x-ray tube current ranging from 48-1550 mA with most scans ranging between 100-600 mA.

2.1.2 Clinical Outcomes

Patient specific outcomes for surgical patients (n = 87) and SBRT patients (n = 63) were scored based on the following categories based on follow-up CT, PET/CT, or other pathological information. A score of 1 was given to patients who experienced the specific type of recurrence and a score of 0 was given to those who did not experience this recurrence. Specifically, the treatment outcomes fell into these specific categories:

- **Failure** ($F \in \{0,1\}$): Cancer recurrence following treatment (n =21)
    - **Local Failure** ($LF \in \{0,1\}$): Recurrence within a few cm of the surgical line (n = 7)

- **Non-Local Failure** (nLF ∈ {0,1}): Regional or distant failure (n = 16)
    - **Regional Failure** (RF ∈ {0,1}): Recurrence in regional lymph nodes (n = 8)
    - **Distant Failure** (DF ∈ {0,1}): Metastatic disease development (n = 10)

2.2 Segmentation

Lesions were identified on pre-treatment computed tomography scans under free-breathing conditions. For each image, the gross tumor volume (GTV) was manually segmented using commercially available contouring software (ARIA Eclipse, Varian). A single physician with experience in interpretation and manual delineation of lung lesions performed all contouring. No pre-defined directions regarding display settings (e.g., window/level, thresholding, pixel representation) were used or specified. Lesion size and appearance were cross-checked against prior radiological interpretation to ensure appropriate delineation from lymphovascular structures.

2.3 Radiomic Feature Extraction

After the GTV's were segmented, the resulting CT volume and segmentation object was exported into MATLAB for analysis. For each GTV, 45 radiomic features were extracted as potential biomarkers for recurrence of NSCLC following treatment. This resulted in the given feature spaces each defined by a matrix of the form,

$$\boldsymbol{\mathcal{F}} = \left(f_{i,j}\right) \in \mathbb{R}^{m \times n} \tag{1}$$

Where n can be taken as the number of patients in the subset, and m the number of extracted features from each subset. That is, the $(i,j)^{th}$ coordinate represents the i^{th} feature from the j^{th} patient. This implies that the j^{th} patients feature vector will be defined as

$$\mathbf{f}_j = \left(f_{i,j}\right)_{i=1}^{m} \in \mathbb{R}^m, \qquad j = 1,2,\dots,n. \tag{2}$$

A complete list of features used in this study is shown in Figure 1. These features are separated into the following categories:

Intensity

#	Feature Name
1	Energy
2	Kurtosis
3	Skewness

Fine Texture

#	Feature Name
4	Auto Correlation
5	Cluster Prominence
6	Cluster Shade
7	Cluster Tendency
8	Contrast
9	Correlation
10	Differential Entropy
11	Dissimilarity
12	Second Angular Moment
13	Joint Entrpy
14	Inverse Difference
15	Inverse Difference Moment
16	Information Correlation 1
17	Information Correlation 2
18	Normalized Inverse Difference Moment
19	Normalized Inverse Difference
20	Inverse Variance
21	Joint Maximum
22	Sum Average
23	Sum Entropy

Coarse Texture

#	Feature Name
24	Short Run Emphasis
25	Long Run Emphasis
26	Gray Level Non-uniformity
27	Run Length Non-uniformity
28	Run Percentage
29	Low Gray Level Run Emphasis
30	High Gray Level Run Emphasis
31	Short Run Low Gray Level Emphasis
32	Short Run High Gray Level Emphasis
33	Long Run Low Gray Level Emphasis
34	Long Run High Gray Level Emphasis
35	Small Zone Emphasis
36	Large Zone Emphasis
37	Gray Level Non-uniformity
38	Zone Size Non-Uniformity
39	Zone Percentage
40	Low Gray Level Zone Emphasis
41	High Gray Level Zone Emphasis
42	Small Zone Low Gray Level Emphasis
43	Small Zone High Gray Level Emphasis
44	Large Zone Low Gray Level Emphasis
45	Large Zone High Gray Level Emphasis

Figure 1: List of radiomic features used in this study, grouped by categrory. Feauture names are consistent with IBSI standard (Zwanenburg, 2020).

- **Intensity based features:** Intensity based features are defined based on the gray level histogram of each image. This can be considered a probability density function

that describes the frequency of different gray levels in the image. They are understood to measure the density characteristics of a tumor (Aerts, 2014).

- **Fine Texture Features:** Fine texture features focus on the small-scale heterogeneity of a tumor within the high-resolution structure. They are calculated from the Gray-Level Co-occurrence Matrix of the image (Haralick, Shanmugam, & Dinstein, 1973).
- **Coarse Texture Features:** Coarse texture features capture the larger scale heterogeneity of a tumor within the low-resolution structure. They are calculated from the Gray-Level Run Length Matrix (Tang, 1998) and Gray-Level Size Zone Matrix (Thibault, 2009). The first measures the frequency of run-lengths, or the size of a set of consecutive pixels with the same intensity. The second provides information on the size of homogeneous zones for each gray level.

Feature extraction was calculated using in-house radiomics software developed in MATLAB.

2.4 Data Splitting

In each model, generalization was evaluated using stratified Monte-Carlo cross-validation with 20 iterations (Burman, 1989). Each iteration the data was split into training (80%) and validation (20%) cohorts with equal event ratios with each of the following models trained after each split. The data included in each training and testing split can be described in three separate experiments corresponding to the two specific aims of this work. To satisfy specific aim 1, a single experiment was performed where

training and testing data was taken only from surgical patients. To satisfy specific aim 2, two experiments were performed. In the first, models trained on surgical patients and then tested on SBRT patient data. This is commonly referred to as cross-training. In the second, models were trained and tested on a combined dataset including both treatment methods.

2.4 Multivariate Modeling

The multivariate association between the radiomic features and clinical endpoints described in the previous sections were investigated using three distinct models. When making clinical decisions, an estimated probability of recurrence is much more useful than a binary prediction, especially when comparing the unique situations of patients. For this reason, regression models were favored over classification models since they output a probability of recurrence instead of binary response.

2.4.1 Logistic Regression

The multivariate association between the calculated radiomic features and cancer recurrence was first investigated using a logistic regression model. This model is named after the well-known logistic equation commonly used to model population growth shown in equation 3.

$$f(x) = \frac{1}{1 + e^{-x}} \tag{3}$$

This function takes any real value for $\boldsymbol{x}$ and maps it to a value between 0 and 1. This equation is also known as the sigmoid function due to it taking an S shape. When fitting

to a logistic regression model, the predicted outputs can be interpreted as the probability of event success, with its sigmoid shape lending itself to be especially useful in binary datasets. Since the clinical recurrence data is represented as a logical 1 for recurrence and logical 0 for no recurrence, this model follows as a natural choice for the data.

Generally, a regression model can be fitted by minimizing a cost or loss function that penalizes predictions that are far from the true outcome. For example, basic linear regression uses least squared error as a loss function which results in a convex graph where the vertex is taken to be its global minimum. In logistic regression, this same approach cannot be used since the sigmoid shape of the logistic function suggests that the least squared error would result in multiple local minima. Instead, we choose to use logistic loss which assigns a higher penalty when predicting an event success when the true outcome is an event failure and vice versa. The estimator for this model is expressed in equation 4.

$$\widehat{\beta_n} = \text{argmin}_\beta \, P_n\{Y \, \log(X\beta) + (1 - Y)log(1 - X\beta)\} \tag{4}$$

In this equation, Y represents our response vector of clinical outcomes, X is our design matrix including the features for each patient, β is a vector of fitted model coefficients, and P_n is the empirical measure or average denoted in equation 5 below.

$$P_n f(X) = \frac{1}{n}\sum_{i=1}^{n} f(X_i) \tag{5}$$

Despite being a preferred method for binary datasets, there are notable disadvantages to a logistic regression model to consider. The first obstacle to consider is the tendency for a

logistic regression model to overfit as the dimension of the dataset increases. This can result in our trained model to generalize poorly to different datasets, decreasing the predictive power considerably. Secondly, logistic regression assumes little to no multicollinearity, which is not necessarily the case in this study. To combat this, regularization techniques were implemented to add bias to the model and constrain the fitting coefficients for each feature. The regularization techniques described in the future sections should reduce the tendency to overfit and reduce the dependence of our model on possible sources of multicollinearity.

2.4.2 Random Forest Regression

Random forest models are composed of multiple decision trees that separate data into groups based on their similarities and differences. Generally, a decision tree can be thought of as a series of if-else statements that categorize a given set of training data into decision nodes which are used as indicators to describe that data. In a random forest regression model multiple decision trees are made, and the average of the predictions is produced as the output to the model.

In a random forest model, one important parameter to consider is the number of decision trees in the forest. As you increase the number of trees in your model, the training data becomes better represented, and increases the performance of the model. This comes at the cost of computation time.

A random forest model has many advantages, first of which being its versatility. Random forest models can be used for classification tasks and can even handle non-linear relationships between observations and outcomes. Random forests also tend to reduce overfitting with enough trees in the model which is and important concern in any machine learning model. One significant disadvantage to this model is the lack of descriptive power in comparison to predictive power. While a random forest tends to excel in accurate predictions, looking for a quantitative description of the relationships in the data can be difficult.

2.4.3 Probit Regression

Probit regression is used to model binary or dichotomous variables where the inverse standard normal distribution of the probability is taken as a linear combination of the predictors. This link function is represented in equation 6 below.

$$Y^* = \Phi^{-1}(X) \tag{6}$$

Here, Y^* is a predicted output and Φ is the cumulative distribution function (CDF) of the standard normal distribution. Like logistic regression, probit regression lends itself to a similar sigmoid shape and can also map any real value to a value between 0 and 1. This means both models are effective in giving a predicted probability of a given outcome. Aside from their link functions the main difference between logistic and probit models is the assumption that each one makes about the distribution of the errors. Logistic regression assumes a standard logistic distribution of errors while probit regression

assumes a normal distribution. A significant difference in these two models could give insight as to how we interpret the error in our data.

Probit regression also suffers from many of the same problems as logistic regression including multicollinearity and the tendency to overfit. For this reason, regularization will be used in conjunction with this model as well, described in the following sections.

2.4.4 L2 Regularization: Ridge Regression

An often first choice in regularization is to add an L2 penalty term to the loss function which encourages the regression coefficients to shrink toward zero. This adds bias to the model while reducing the variance, thus possibly reducing MSE and limiting the possibility of overfitting. This technique is commonly referred to as ridge regression, especially when paired with a least squares estimator as shown in equation 7.

$$\widehat{\beta_n^{\lambda}} = \text{argmin}_{\beta} P_n(Y - X\beta)^2 + \lambda||\beta||_2^2 \quad (7)$$

Here, $\lambda > 0$ is a tuning parameter and is chosen during optimization during cross validation while $||\beta||_2$ is the standard L2 Euclidean norm. We can see from this equation that when λ increases the regression coefficients are forced toward zero. This creates a stronger emphasis on features that tend to impact the outcome by increasing their regression coefficients in comparison to those that have a less significant effect. Though ridge regression tends to shrink regression coefficients toward zero, they will never equate to zero. This means that using ridge regression or a L2 regularization cannot

perform any kind of feature selection without imposed thresholding. Graphically this idea is shown in Figure 2 below using an ordinary least squares (OLS) estimator in two dimensions.

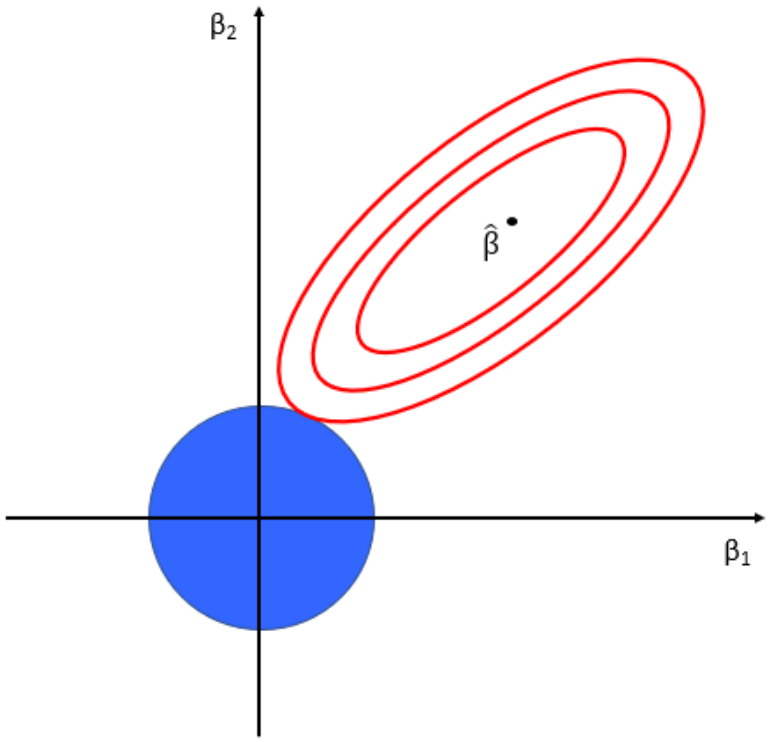

Figure 2: A contour of the error (red) and constraint (blue) functions for ridge regression. The intersection represents the optimized solution where both regions are as small as possible, known as the ridge estimate.

From this figure we can see the tradeoff between constraint for our regression coefficients $\beta_1^2 + \beta_2^2 \leq c$ and the residual sum of squares of the estimator $\hat{\beta}$, while showing that the solution will not set one of the coefficients to zero.

2.4.5 L1 Regularization: Lasso Regression

Another common technique for regularization is to use the least absolute shrinkage and selection operator (lasso) constraining the L1 norm instead of the L2 norm. While similar to ridge regression, this method is unique in that it can produce sparse solutions, setting some coefficients to 0. The estimator for lasso regression is shown in equation 8 below.

$$\widehat{\beta_{\mathrm{n}}^{\lambda}} = \mathrm{argmin}_{\beta}\, \mathrm{P_n}(\mathrm{Y} - \mathrm{X}\beta)^2 + \lambda\left|\left|\beta\right|\right|_1 \tag{8}$$

In this case, λ serves the same purpose as in ridge regression. Here as λ increases, coefficients tend toward 0, but allowing for sparse solutions unlike before. This also means that lasso regularization can be used as a feature selection method. Graphically, this is shown in Figure 3.

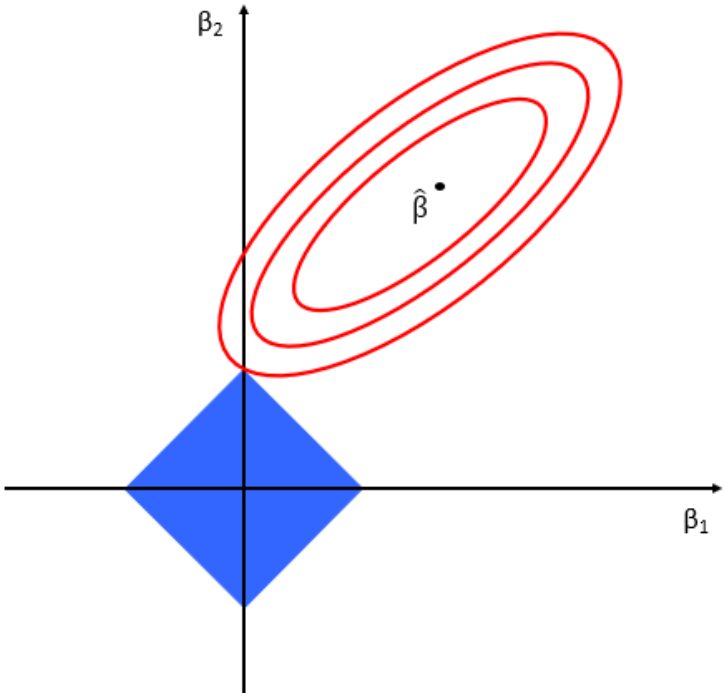

Figure 3: A contour of the error (red) and constraint (blue) functions for lasso regression. The intersection represents the optimized solution where both regions are as small as possible

From this figure we can view a similar tradeoff to ridge between the constraint for our regression coefficients $|\beta_1| + |\beta_2| \leq c$, and the residual sum of squares of the estimator $\hat{\beta}$ but with the possibility of the solution occurring with coefficients set to 0.

2.4.6 Elastic-Net

Elastic-net regression is a method that combines the penalty terms of both the ridge and lasso techniques to gain the benefits from using both. The elastic-net estimator shown in equation 9 below uses a parameter α instead of λ, which balances the weight of the L1 and L2 penalties.

$$\widehat{\beta_n^{\alpha}} = \text{argmin}_{\beta}\, P_n(Y - X\beta)^2 + \alpha||\beta||_2^2 + (1-\alpha)\left|\left|\beta\right|\right|_1 \quad (9)$$

We can see from the above that when α=1, we arrive at the ridge solution and when α=0, we arrive at the lasso solution. With the combination of these penalties, we can still have singularities at vertices which allow for sparse solutions, yet keep the convex edges removing the limitation on number of variables, encourages grouping, and stabilizing the L1 regularization path. This is shown in Figure 4 below in a 2-dimensional illustration with α=0.5.

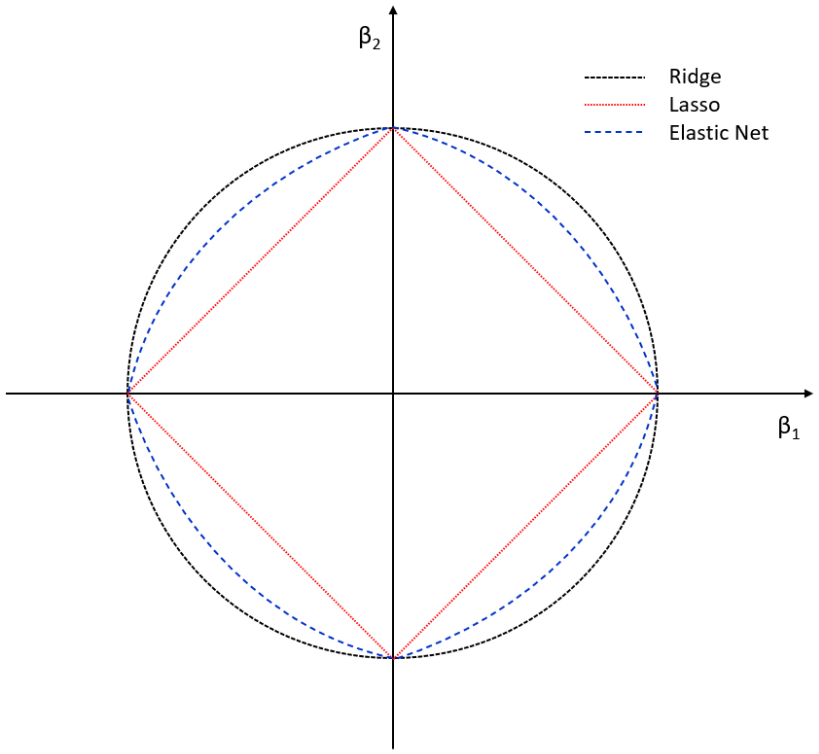

Figure 4: A depiction of the constraint regions for ridge, lasso, and elastic net regression in 2-dimensional space. The elastic net constraining region lies between the ridge constraint, a circle in 2 dimensions, and the lasso constraint, a diamond in 2 dimensions.

In this project, elastic net was used as the final regularization method for both regression techniques. In this work, multiple values of alpha were initially tested by training models on the surgical cohort of patients. In each case, the alpha value of 0.5 produced models with higher AUC's, so this value was kept consistent in all following models.

2.5 Model Analysis

For each model, results were combined to construct receiver operating characteristic (ROC) curves and the area under the curve (AUC) was calculated. The AUC was used as the metric to compare final model performance. Finally, a permutation test was performed to determine the statistical significance between models.

3. Results

3.1 Aim 1: Surgical Models

The first aim in this work was to build radiomic models based on pre-treatment CT images from surgical patients and evaluate their performance in predicting cancer recurrence. These models were created using the methods described above and evaluated by their ROC curves and resulting AUC values. The resulting AUC values are shown in Table 1 below.

Table 1: AUC values for models trained on surgical patients for each type of failure.

	F	**LF**	**nLF**
Logistic Regression	0.70 ± 0.03	0.57 ± 0.04	0.82 ± 0.04
Probit Regression	0.70 ± 0.04	0.57 ± 0.04	0.80 ± 0.04
Random Forest	0.64 ± 0.04	0.53 ± 0.04	0.70 ± 0.04

From this table we can see that the logistic and probit regression models gave similar AUC values that were higher on average than the random forest method. We can also see that that the nLF models gave the highest average AUC values with the LF models showing the lowest. The ROC curves with the highest AUC's for each type of failure are shown in Figure 5.

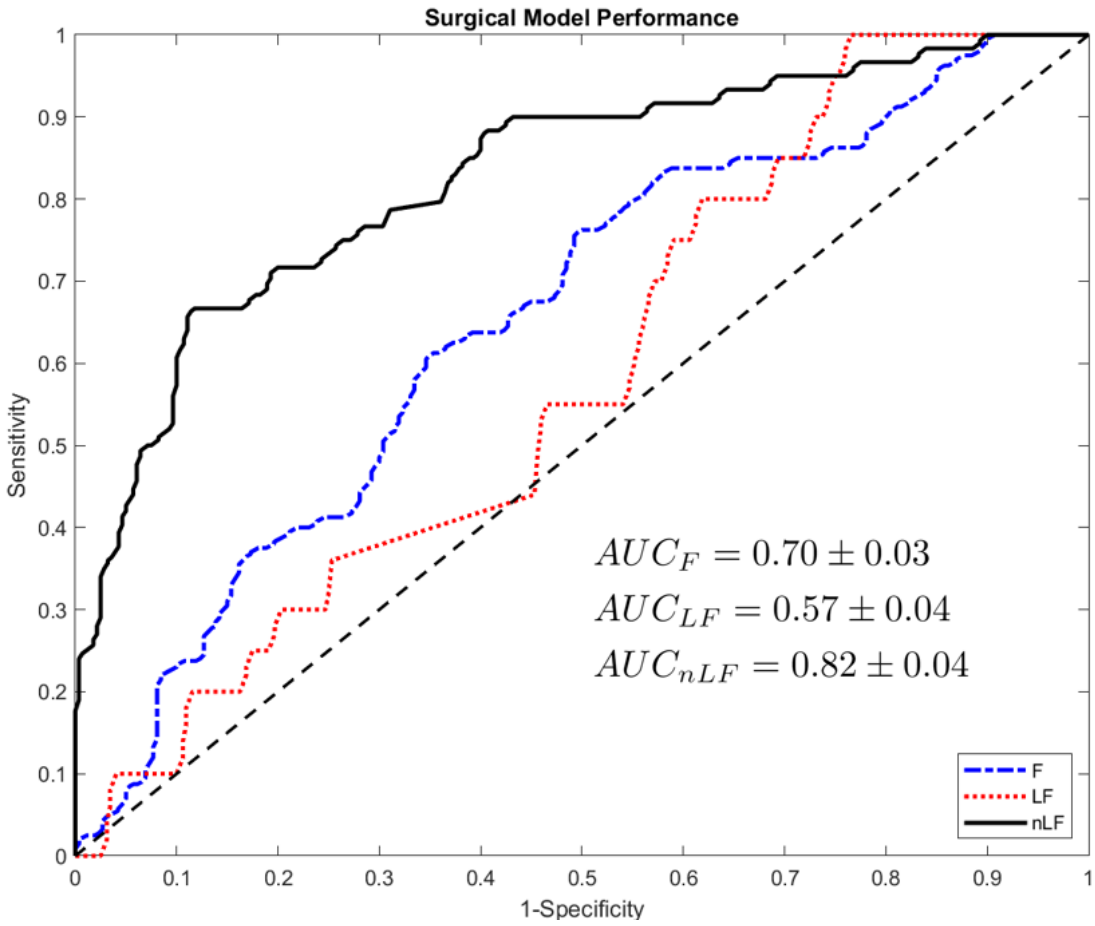

Figure 5: ROC curves for the best perfoming surgical models for each failure type. The largest pairwise p-value was < 0.01.

3.2 Aim 2: Incorporating SBRT Data

The second aim in this work was to generalize our surgical models by incorporating features from pre-treatment SBRT images in model training and testing. This was done in two ways. First, a cross-training approach was taken, training all models on surgical patients features and then tested on SBRT patients. Second, models were trained and tested on a combined set of features from both surgery and SBRT patients.

3.2.1 Cross-Training

The AUC values for the cross-trained models are shown in Table 2 below.

Table 2: AUC values for cross-trained models for each type of failure.

	F	LF	nLF
Logistic Regression	0.54 ± 0.02	0.57 ± 0.03	0.53 ± 0.03
Probit Regression	0.54 ± 0.02	0.56 ± 0.03	0.53 ± 0.03
Random Forest	0.52 ± 0.02	0.59 ± 0.04	0.58 ± 0.04

Here we can that the logistic and probit regression models performed similarly again, which varied slightly from the random forest models. No model in particular demonstrated any predictive power for recurrence with the highest AUC value for any model and recurrence type being 0.59 ± 0.04. The ROC curves with the highest AUC's for each type of failure are shown in Figure 6.

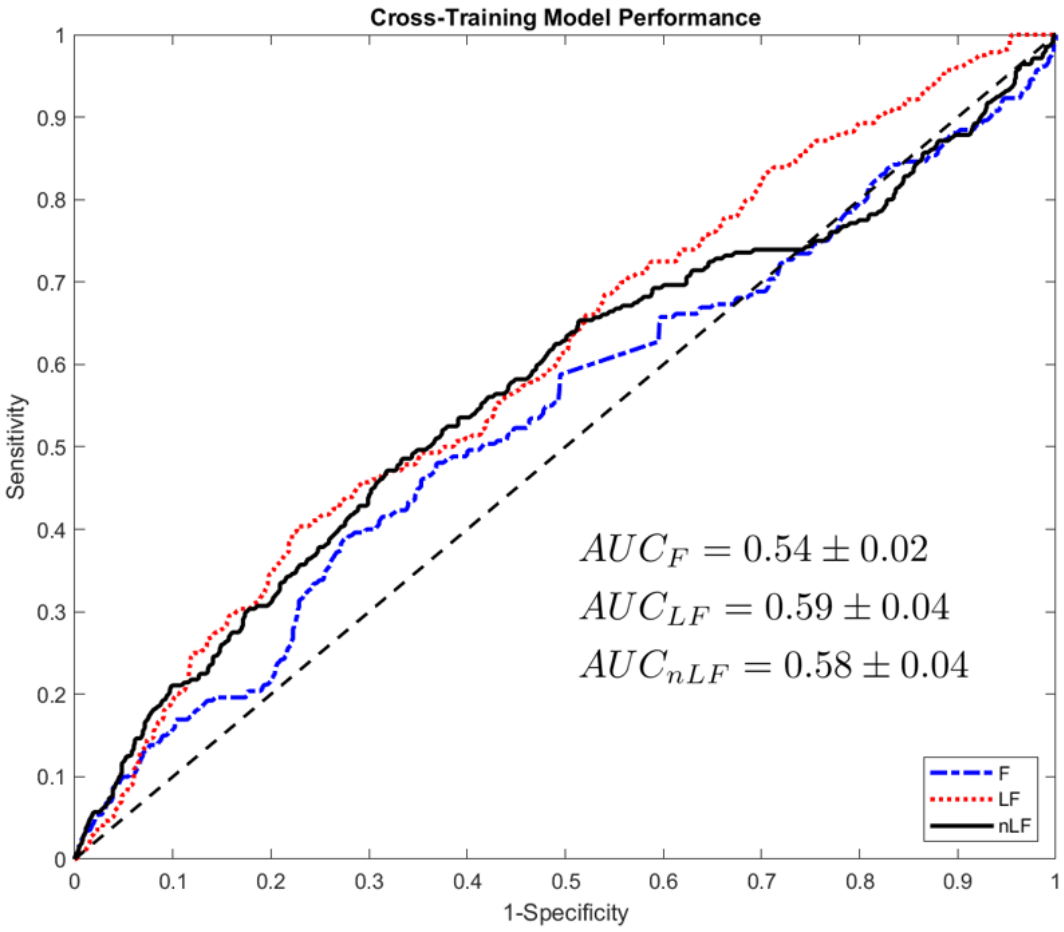

Figure 6: ROC curves for the best perfoming cross-trained models for each failure type. The largest pairwise p-value was < 0.02.

3.2.2 Combined Data Split

The AUC values for the models trained on the combined dataset are shown in Table 3.

Table 3: AUC values for models trained on the combined dataset for each type of failure.

	F	LF	nLF
Logistic Regression	0.54 ± 0.02	0.57 ± 0.02	0.64 ± 0.03
Probit Regression	0.54 ± 0.02	0.57 ± 0.02	0.64 ± 0.03
Random Forest	0.55 ± 0.02	0.59 ± 0.04	0.60 ± 0.03

Like the cross-trained set, the logistic and probit regression models performed similarly, which varied slightly from the random forest models. No model demonstrated significant predictive power for recurrence with the highest AUC value for any model and recurrence type being 0.64 ± 0.03. The ROC curves with the highest AUC's for each type of failure are shown in Figure 7.

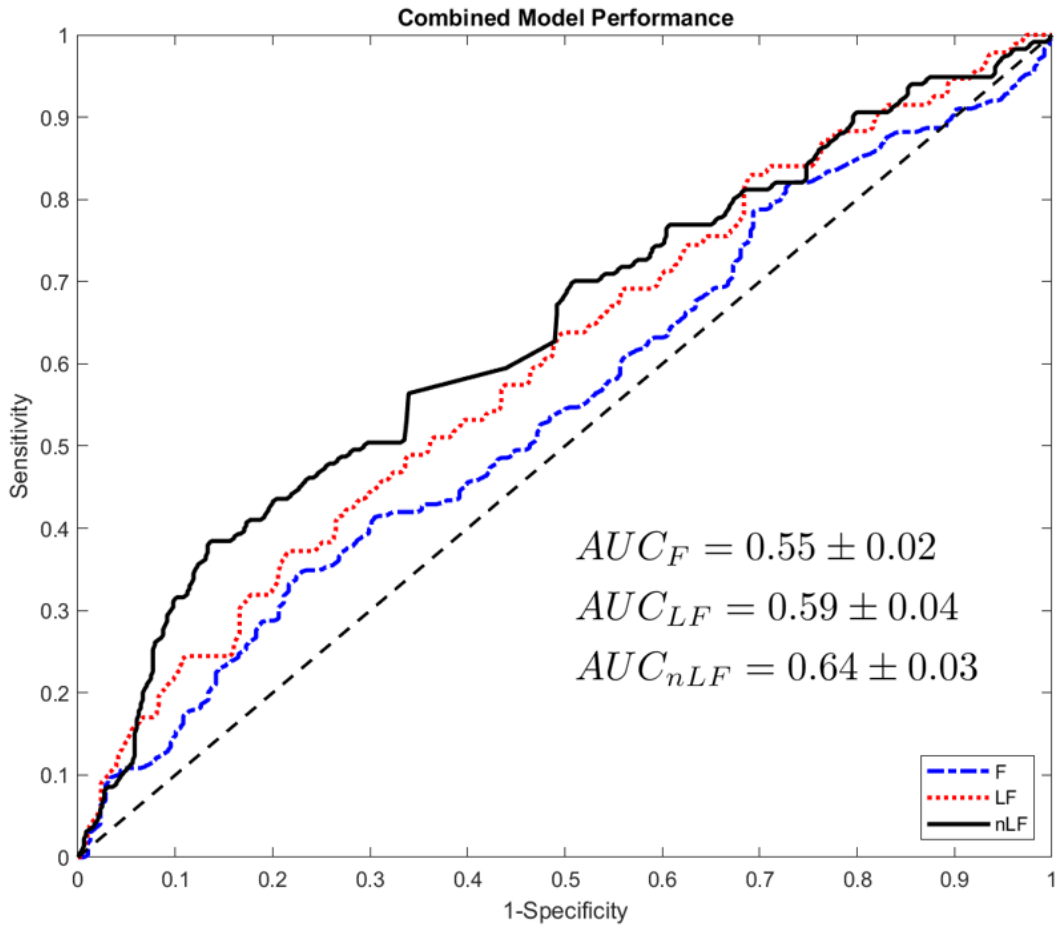

Figure 7: ROC curves for the best perfoming cross-trained models for each failure type. The largest pairwise p-value was < 0.02.

4. Discussion

4.1 Aim 1

From Table 1 and Figure 5, we can see that the models trained on the surgical cohort only performed best when predicting nLF and the worst when predicting LF with the highest AUC values for each recurrence type being 0.82 ± 0.04 and 0.57 ± 0.04 respectively. Lafata (2019a) instead showed the strongest association between radiomic features and LF instead of nLF when training models on SBRT patients. This may demonstrate a key difference in the mechanisms of local recurrence between SBRT and surgery. Radiomic features are calculated directly from the image of the GTV, which is completely removed during surgery. This suggests that that radiomic features may have limited value in predicting local recurrence, considering the part of the image that is used in calculating these features is no longer in the body post-treatment. SBRT on the other hand, leaves all tissue in the body which makes properties of the GTV or calculated features a more attractive choice for predicting recurrence.

It should also be noted that within the surgical cohort of scans, there is a significant amount of variability of image parameters between scans. The variability in scan parameters such as slice thickness, mAs, and reconstruction algorithm have been shown to strongly impact image features (Midya, 2018, Kim, 2019). This variability makes it difficult to draw strong conclusions about the association between radiomic features and

recurrence, even for the better performing surgical models for nLF. This phenomenon is further demonstrated in the results for aim 2.

4.2 Aim 2

From Tables 2 and 3 and Figures 6 and 7, we can see that when we added SBRT data into the training and testing process the performance of the trained models decreased with only a few models being statistically significant when compared to a random guess (AUC = 0.5). This could be a further demonstration of effect of varying image parameters on radiomic features. As stated previously, the surgical patients were taken with multiple scanners under different scanner parameters. Adding SBRT data to this set increases the variation in imaging parameters which could further affect radiomic feature values. This may be a likely contributing factor to the poor performance of models that incorporated both sets of data.

4.3 Model Comparison

In each case the logistic and probit regression models performed similarly, which is to be expected since both methods had the same regularization methods and have a sigmoidal relationship between outcome probability and predictors. The differences between the two methods, mainly their link functions and their assumptions on error distribution did not show in the outcomes of the models. In the surgical cohort models described in Table 1, these models both performed better than the random forest model. This may be due to the added regularization that the other two methods had, whereas the

random forest model did not have a specific feature selection step beforehand. The random forest model performance may improve with an additional feature selection step or an increased amount of decision trees, however this will increase computation time.

4.4 Radiomic Features

The set of coefficients for each regression model was saved in each case. By examining the coefficients that frequently have the largest magnitudes, you can determine which features were important in aiding the predictive power of the model. The model that performed the best throughout this study was logistic regression trained and tested on surgical patients, specifically when predicting non-local failure. In Figure 8, a histogram shows the probability of a feature's regression coefficient appearing among the 10 largest in this model. According to these criteria, the features that drove this model the most were Second Angular Moment and Cluster Tendency, which are both Fine Texture Features. Second Angular Moment is based on the Gray Level Co-occurrence Matrix and is a measure of homogeneous patterns in the image. A greater Second Angular Moment suggests that there are more instances of intensity value pairs that neighbor each other at higher frequencies in an image. Cluster Tendency is also determined from the Gray Level Co-occurrence Matrix and will be greater if there are more groupings of voxels with similar gray levels. The dependence of this model on features that rely on the level of homogeneity in the grayscale of the image could imply that physical homogeneity may contribute to the chance of recurrence. For example, a heterogeneous tumor could result

in a biopsy that provides misleading information or increase the chance of relapse due cell diversity and adaptability to treatment options.

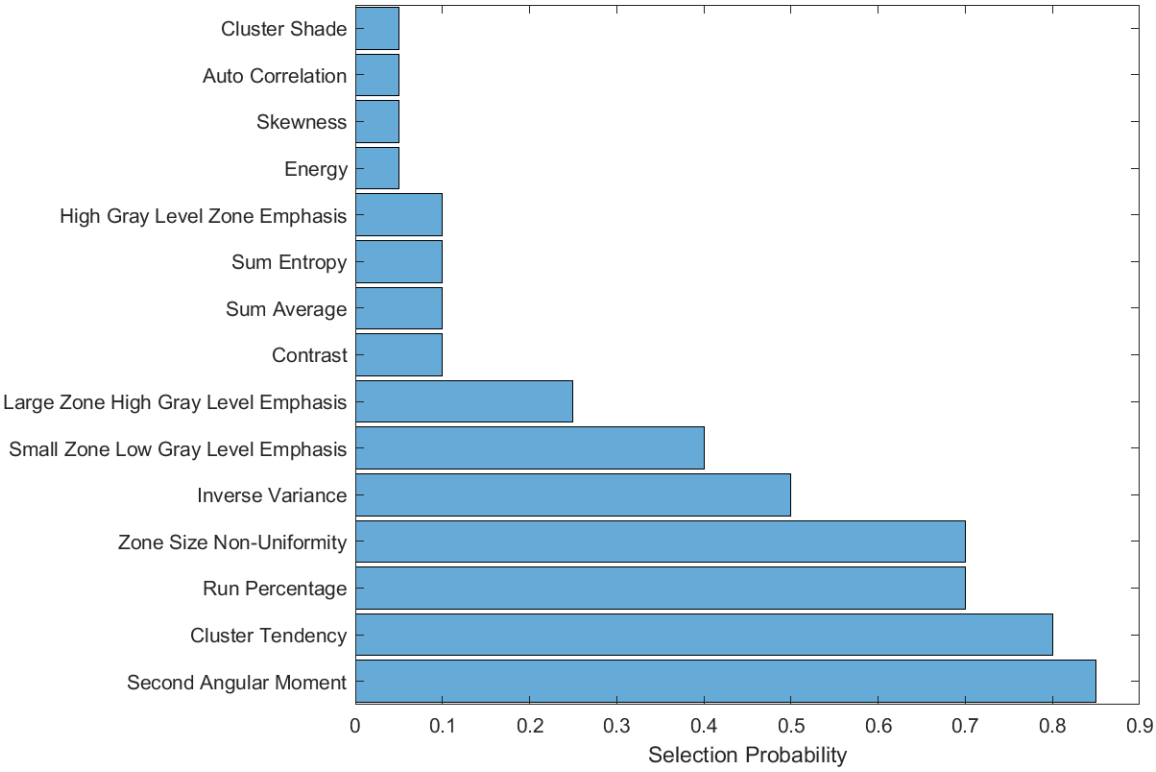

Figure 8: Histogram of the most frequently selected features in the logistic regression nLF models.

5. Limitations and Future Work

The first limitation in this study was the variability in the images of the surgical patients as described by their imaging parameters such as slice thickness and tube current which have been shown to affect imaging features (Midya, 2018, Kim, 2019, Lafata, 2018). Histograms of the varying slice thicknesses and tube currents from each of the surgical images are shown below in Figure 9 to demonstrate this variability.

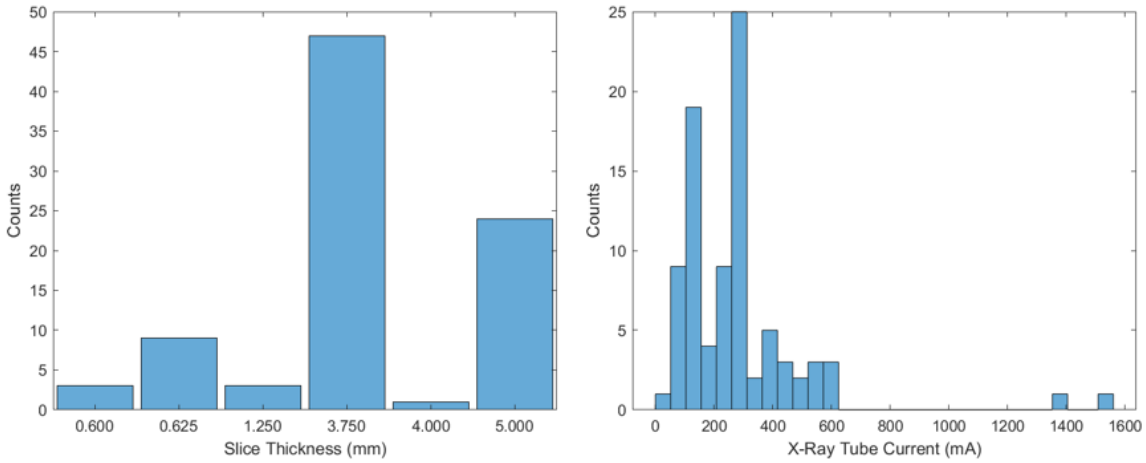

Figure 9: Histograms of slice thickness and tube current for pre-treatment surgical images.

For future work, controlling for these parameters before modeling may be done to minimize the error in feature values from scanner differences. For example, a more uniform dataset could be constructed only from patients with 3.75 mm slice thicknesses and tube currents ranging from 100-300 mA. For the data used in this project, doing this would severely limit our data set in terms of size and recurrence rate making trained models more unstable and less reliable. With more available patient data, imaging parameters like this should be accounted for. Similarly, feature extraction software implementations should also be standardized to maintain consistency when analyzing features (Zwanenburg, 2020, Chang, 2020).

Another limitation in this study is the lack of an independent test set for model validation. This means that the models developed in this study are not completely independent of the testing set which could result in overperformance. With more data,

such a test set can be created, such as designating the most current year scans as a test set while training only on based on scans from previous years.

In this work supervised machine learning models, trained on a labeled dataset so the model can evaluate its accuracy, however future work should consider unsupervised models as well. Specifically, unsupervised clustering techniques have shown to be particularly useful for a variety of radiomic applications (Aerts, 2014, Lu, 2019, Lafata, 2019b, Lafata, 2019c). In addition, radiomics can be integrated with other "-omics" to potentially improve prognostic value with a multi-omic model instead of just imaging. In lung cancer, liquid biopsies to detect cell free DNA and circulating tumor DNA are becoming more prevalent, so radiomics combined with cfDNA/ctDNA markers could be used to develop radiogenomic biomarkers (Corradetti, 2019, Lafata, 2021).

6. Conclusions

Models trained and tested on surgical patients showed a stronger association between radiomic features and non-local failure and a poor association with local failure. This may demonstrate that radiomic features have limited value in predicting local recurrence, considering the part of the image that is used in calculating these features is no longer in the body post-treatment. Despite this, it is difficult to draw strong conclusions based on the variability in the images of surgical patients which has been shown to affect feature values (Midya, 2018, Kim, 2019, Lafata 2018). This is supported by the degraded

performance in these models when SBRT data was introduced, further increasing image variability.

References

Aerts, Velazquez, & Leijenaar. (2014). Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat Commun*.

American Cancer Society. (2019, October). *About Lung Cancer*. Retrieved from https://www.cancer.org/cancer/lung-cancer/about/what-is.html

American Cancer Society. (2021, February). *Treatment Choices for Non-Small Cell Lung Cancer, by Stage*. Retrieved from https://www.cancer.org/cancer/lung-cancer/treating-non-small-cell/by-stage.html

American Society of Clinical Oncology (ASCO). (2020, May). *Lung Cancer - Non-Small Cell: Statistics*. (Cancer.Net) Retrieved from https://www.cancer.net/cancer-types/lung-cancer-non-small-cell/statistics

Bhupesh Parashar, M. S. (2013). Radiation Therapy for Early Stage Lung Cancer. *Seminars in Interventional Radiology, 30*(2), 185-190. doi:10.1055/s-0033-1342960

Bogart, J., & Wallen, J. (2017). Management of Patients With Stage I Lung Cancer. *Journal of Oncology Practice, 13*(2), 69-76.

Bradley G Ackerson, B. C. (2018). Stereotactic body radiation therapy versus sublobar resection for stage I NSCLC. *Lung Cancer, 125*, 185-191. doi:10.1016/j.lungcan.2018.09.020

Burman, P. (1989). A Comparative Study of Ordinary Cross-Validation, v-Fold Cross-Validation and the Repeated Learning-Testing Methods. *Biometrika, 76*(3), 503-514. doi:doi.org/10.2307/2336116

Chang, Lafata, Wang, Duan, Geng, Yang, & Yin. (2020). Digital phantoms for characterizing inconsistencies among radiomics extraction toolboxes. *Biomedical Physics and Engineering Express, 6*(2).

Corradetti, Torok, Hatch, Xanthopoulos, Lafata, Jacobs, . . . Nixon. (2019). Dynamic changes in circulating tumor DNA during chemoradiation for locally advanced lung cancer. *Advances in Radiation Oncology*, 748-752.

Elizabeth Huynh, T. P. (2016). CT-based radiomic analysis of stereotactic body radiation therapy patients with lung cancer. *Radiother Oncol., 120*(2), 258-266. doi:10.1016/j.radonc.2016.05.024

Frank J Lagerwaard, N. E. (2012, May). Outcomes of stereotactic ablative radiotherapy in patients with potentially operable stage I non-small cell lung cancer. *International Journal of Radiation Oncology, Biology, Physics, 83*(1), 348-353. doi:10.1016/j.ijrobp.2011.06.2003

G. Thibault, B. F. (2009). Texture indexes and gray level size zone matrix. Application to cell nuclei classification.

Ginsberg, R. J., & Rubinstein, L. V. (1995). Randomized trial of lobectomy versus limited resection for T1 N0 non-small cell lung cancer. Lung Cancer Study Group. *The Annals of Thoracic Surgery, 60*(3), 615-622. doi:10.1016/0003-4975(95)00537-u

Haralick, R. M., Shanmugam, K., & Dinstein, I. (1973). Textural Features for Image Classification. *IEEE Transactions on Systems, Man, and Cybernetics, 3*(6), 610-621. doi:10.1109/TSMC.1973.4309314

Harvard Health Online. (2020, September). *Squamous Cell Carcinoma of the Lung*. (Harvard Health Publishing) Retrieved from https://www.health.harvard.edu/cancer/squamous-cell-carcinoma-of-the-lung

Howlader N, N. A. (2020, April 15). *SEER Cancer Statistics Review, 1975-2017*. Retrieved from National Cancer Institute: https://seer.cancer.gov/csr/1975_2017/

Hugo Aerts, E. R.-K. (2014). Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature Communications, 5*, 1-8.

Kim, Y. J., & Lee, H.-J. (2019). The Effect of CT Scan Parameters on the Measurement of CT Radiomic Features: A Lung Nodule Phantom Study. *Comput Math Methods Med.* doi:10.1155/2019/8790694

Lafata, Cai, Wang, Hong, Kelsey, & Yin. (2018, November). Spatial-temporal variability of radiomic features and its effect on the classification of lung cancer histology. *Physics in Medicine and Biology, 63*(22).

Lafata, Corradetti, Gao, Jacobs, Weng, Chang, . . . Yin. (2021). Radiogenomic analysis of locally advanced lung cancer based on CT imaging and intra-treatment changes in cell free tumor DNA. *Radiology: Imaging Cancer*.

Lafata, Hong, Geng, Ackerson, Liu, Zhou, . . . Yin. (2019). Association of pre-treatment radiomic features with lung cancer recurrence following stereotactic body radiation therapy. *Physics in Medicine and Biology, 64*(2). doi:10.1088/1361-6560/aaf5a5

Lafata, Zhou, Liu, & Yin. (2019, September). Data clustering based on Langevin annealing with a self-consistent potential. *Q Appl Math*, 591-613.

Lafata, Zhou, Liu, Hong, Kelsey, & Yin. (2019, August). An exploratory radiomics approach to quantifying pulmonary function in CT images. *Sci Rep, 9*(1). doi:10.1038/s41598-019-48023-5

Lu, Arshad, Thornton, Avesani, & Cunnea. (2019, February). A mathematical-descriptor of tumor-mesoscopic-structure from computed-tomography images annotates prognostic- and molecular-phenotypes of epithelial ovarian cancer. *Nat Commun, 10*(1). doi:10.1038/s41467-019-08718-9

McDonald, F. (2017). Management of stage I and II nonsmall cell lung cancer. *European Respiratory Journal, 49*(1). doi:10.1183/13993003.00764-2016

Midya, A., Chakraborty, J., & Gönen, M. (2018). Influence of CT acquisition and reconstruction parameters on radiomic feature reproducibility. *Journal of Medical Imaging, 5*(1). doi:10.1117/1.JMI.5.1.011020

(2021). *NCCN Clinical Practice Guidelines in Oncology (NCCN Guidelines®) Non-Small Cell Lung Cancer Version 4.2021.* National Comprehensive Cancer Network. Retrieved from www.nccn.org/patients

Paul, S. (2010). Thoracoscopic lobectomy is associated with lower morbidity than open lobectomy: A propensity-matched analysis from the STS database. *J Thorac Cardiovasc Surg, 139*(2), 366-378. doi:10.1016/j.jtcvs.2009.08.026

Philippe Lambin, E. R.-V. (2012). Radiomics: extracting more information from medical images using advanced feature analysis. *European Journal of Cancer, 48*(4), 441-446. doi:10.1016/j.ejca.2011.11.036

Si Yeol Song, W. C.-W. (2009, October). Fractionated stereotactic body radiation therapy for medically inoperable stage I lung cancer adjacent to central large bronchus. *Lung Cancer, 66*(1), 89-93. doi:10.1016/j.lungcan.2008.12.016

Tang, X. (1998). Texture information in run-length matrices. *IEEE Transactions on Image Processing, 7*(11), 1602 - 1609. doi:10.1109/83.725367

Timmeren, J. E. (2020). Radiomics in medical imaging—"how-to" guide and critical reflection. *Insights into Imaging, 11*(91). doi:https://doi.org/10.1186/s13244-020-00887-2

Vignesh Raman, C.-F. J. (2018). Surgical treatment for early stage non-small cell lung cancer. *Journal of Thoracic Disease, 10*(7), 898-904. doi:10.21037/jtd.2018.01.172

Wen Yu, C. T. (2018). Development and Validation of a Predictive Radiomics Model for Clinical Outcomes in Stage I Non-small Cell Lung Cancer. *International Journal of Radiation Oncology, Biology, Physics, 102*(4), 1090-1097. doi:10.1016/j.ijrobp.2017.10.046

Zwanenburg A, V. M. (2020, May). The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping. *Radiology, 295*(2), 328-338. doi:10.1148/radiol.2020191145