2018

Advanced Imaging Analysis for Predicting Tumor Response and Improving Contour Delineation Uncertainty

Rebecca N. Mahon
Virginia Commonwealth University

Advanced Imaging Analysis for Predicting Tumor Response and Improving Contour Delineation Uncertainty

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at Virginia Commonwealth University.

By

Rebecca Nichole Mahon, MS

Master of Science,
Virginia Commonwealth University, May 2015

Bachelor of Science,
University of Virginia, December 2009

Major Director: Dr. Elisabeth Weiss,
Professor,
Department of Radiation Oncology

Virginia Commonwealth University
Richmond, VA
June, 2018

Dedication

To my family, in particular my mom, Caroline Mahon, and dad, John Mahon Jr. for always believing in me despite being a terrible teenager, and my siblings John, Rachael, and Renee Mahon for making life interesting and inspiring me to never give up. To my aunts, uncles, cousins, sister-in-law, niece and friends for encouraging me, always being there, assisting in proof reading my dissertation, and reminding me to take a breaks for fun. To my grandparents Margret “Amah” and John “Pop-Pop” Mahon, Epkje Brouwer (Wilk) Janka and Wladyslaw “John” Wilk, who did not live to see this work but nevertheless made a profound impact on my life. To my cat, Princess, for being at my side for 18 and a half years through thick and thin with unconditional love. To the Riblett family for their support.

And to Matthew Riblett for keeping me sane, challenging me to be a better person, and coercing… I mean encouraging me into become a programmer. I look forward to our many adventures and crazy projects to come. I love you with all my heart.

Acknowledgement

I would like to thank my advisor Dr. Elisabeth Weiss for her mentoring and guidance through the field of radiation oncology. As a physician, she encouraged me to try and see a problem from all sides instead of just looking narrowly at the physics and taught me how to be a member of the radiation oncology team instead of working in the background. In addition to being my advisor, she also contributed countless hours of contouring, and advising to make this project possible. I would also like to thank my unofficial co-advisor Dr. Geoffrey D. Hugo. He taught me how to push beyond my comfort zone and challenged me to explore areas of research in computer science despite my protest of being “not a programmer.” In addition, he has become my role model for an outstanding medical physicist and phenomenal leader that I one-day hope to emulate. I am indebted to both Dr. Weiss and Dr. Hugo for their selfless sharing of knowledge and encouragement and cannot thank them enough for having me as one of their students.

I would also like to thank my committee for their feedback and influence on my project. I am thankful to Dr. Monica Ghita for always being there to answer questions about imaging, chat about the project, and provide guidance into avenues of exploration. I would also like to thank Dr. Milos Manic for taking me under his wing,

teaching me almost everything I know about machine leaning, for answering my questions with questions, and encouraging me to explore the possibilities.

This work would not have been possible without the assistance and support of many other people. A large part of my project would not be possible without the assistance of Dr. Kishor Karki. I cannot thank him enough for his assistance with image acquisition and analysis of contour delineation uncertainty. I would like to thank Dr. Chet Ford for his feedback on my project and assistance with understanding radiomics, and Dr. Nitai Mukhopadhyay for his guidance on the statistical analysis in this work. I thank Ann Lyons for helping to shape my professionalism and giving me the final push I needed to go back to school. With all on my heart, I thank Matthew Riblett for his endless patience with my programming questions, use of his computational resources, hours of debate on project directions, and of course putting up with me in the same office for five years. I would like to thank the VCU medical physics faculty for their wisdom and feedback as well as Deanna Pace, Shana Ryman and Katie Goracke, Judith Larios, Janet Salmon, and Ron Broman for their assistance navigating the administrative side of the program and technical knowledge. I would like to thank the staff of the VCU Center for High Performance Computing (CHiPC) for use of computation resources to complete my project and my classmates for their insight, questions, and friendship. I want to thank my family both actual and “adopted” for their love and support over the years. And finally, everyone who told me I would not be here today for helping to fuel my motivation to keep going no matter what.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AAPM: American Association of Physicists in Medicine

ADC: Apparent Diffusion Coefficient

ANTs: Advanced Normalization Toolkits

AT: Atelectasis

AW: Air Way

BC: Bias Corrected

CAD: Computer Aided Diagnosis

CBCT: Cone Beam Computed Tomography

CCC: Concordance Correlation Coefficient

CNN: Convolutional Neural Network

CRF: Conditional Random Fields

CT: Computed Tomography

CTV: Clinical Target Volume

CW: Chest Wall

DCE: Dynamic Contrast Enhanced

DNA: Deoxyribonucleic Acid

DW: Diffusion Weighted

E-Net: Efficient Neural Network

18F-FDG: Fluorine 18 labeled flurodeoxyglucose

FDR: False Discovery Rate

FWER: Familywise Error Rate

GLCM: Gray Level Co-occurrence Matrix

GLRLM: Gray Level Run Length Matrix

GLSZM: Gray Level Size Zone Matrix

GPU(s): Graphic Processing Unit(s)

GTV: Gross Tumor Volume

Gy: Gray

HU: Hounsfield unit

IGRT: Image Guided Radiation Therapy

ILSVRC: ImageNet Large Scale Visual Recognition Challenge

ITV: Internal Target Volume

Med: Mediastinum

MR(I): Magnetic Resonance (Imaging)

NCI: National Cancer Institute

NGTDM: Neighborhood Gray Tone Difference Matrix

NSCLC: Non-Small Cell Lung Cancer

OAR(s): Organ(s) at Risk

PCA: Principal Components Analysis

PDF: Probability Density Function

PET: Positron Emission Tomography

PTV: Planning Target Volume

RECIST: Response Evaluation Criteria in Solid Tumors

ReLU: Rectified Linear Unit

ResNet(s): Residual Network(s)

ROC: Receiver Operating Characteristic

ROI(s): Region(s) of Interest

RMS: root mean squared

SBRT: Stereotactic Body Radio Therapy

SGD: Stochastic Gradient Decent

SUV: Standard Uptake Value

TRUFISP: True Fast MRI with Steady State Precession

VATS: Video Assisted Thoracic Surgery

VCU: Virginia Commonwealth University

VIBE: Volumetric Interpolation Breath-Hold Examination

Abstract

ADVANCED IMAGING ANALYSIS FOR PREDICTING TUMOR RESPONSE AND IMPROVING CONTOUR DELINEATION UNCERTAINTY

By Rebecca Nichole Mahon, MS

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at Virginia Commonwealth University.

Virginia Commonwealth University, 2018

Major Director: Dr. Elisabeth Weiss,
Professor,
Department of Radiation Oncology

Radiomics, an advanced form of imaging analysis, is a growing field of interest in medicine. Radiomics seeks to extract quantitative information from images through use of computer vision techniques to assist in improving treatment. Early prediction of treatment response is one way of improving overall patient care. This work seeks to explore the feasibility of building predictive models from radiomic texture features extracted from magnetic resonance (MR) and computed tomography (CT) images of lung cancer patients. First, repeatable primary tumor texture features from each imaging modality were identified to ensure a sufficient number of repeatable features existed for model development. Then a workflow was developed to build models to

predict overall survival and local control using single modality and multi-modality radiomics features. The workflow was also applied to normal tissue contours as a control study. Multiple significant models were identified for the single modality MR- and CT-based models, while the multi-modality models were promising indicating exploration with a larger cohort is warranted.

Another way advances in imaging analysis can be leveraged is in improving accuracy of contours. Unfortunately, the tumor can be close in appearance to normal tissue on medical images creating high uncertainty in the tumor boundary. As the entire defined target is treated, providing physicians with additional information when delineating the target volume can improve the accuracy of the contour and potentially reduce the amount of normal tissue incorporated into the contour. Convolution neural networks were developed and trained to identify the tumor interface with normal tissue and for one network to identify the tumor location. A mock tool was presented using the output of the network to provide the physician with the uncertainty in prediction of the interface type and the probability of the contour delineation uncertainty exceeding 5mm for the top three predictions.

1 Introduction

Advances in computer science and technology in medicine have opened new avenues of research to assist in the personalization of medicine. Particularly in the area of cancer treatment, there has been an increased interest in individualized therapy to improve patient outcome. The need for further individualization of treatment stems from the fact that cancer is a multi-faceted disease that is unique to almost every patient. Cancer is caused by one or more mutations that cause a cell to divide continuously by suppressing its ability to "switch off" when in close contact with other cells. Normally, these aberrant cells are detected by the body and the immune system removes them before they can cause harm. However, cancer cells have adapted ways of tricking the immune system into ignoring them and in some cases even helping them to grow. As solid tumors get larger, they coerce the body into providing them with a network of blood vessels through angiogenesis to fuel their growth while simultaneously reducing the available nutrients for the rest of the body. As cancer is a result of mutations to an individual's deoxyribonucleic acid (DNA), the resulting cancer cells are as unique as the DNA from which it mutated.

In the early days of cancer treatment, there was a one size fits all approach. Every patient was treated the same way, however, some patients responded while others did not. As we learned more about cancer physiology and the interactions between radiation

and the body, new techniques were developed to reduce exposure of the normal tissue and organs, and to better target the cancer lesions. Research began to uncover which genetic mutations were present in patients who responded to treatment and those who did not, and which mutations lead to higher risk of developing cancer. Existing treatments began to be individualized, different amounts of radiation were prescribed to individuals with the same cancer location, and alternative treatments such as chemotherapy were developed. Combinations of different treatment techniques were employed for some cancers. More recently, immunotherapy has grown to further target individual mutations to improve results. Researchers have also begun to incorporate research from the field of computer vision to advance our understanding of cancer and to improve the ability to locate and treat cancer lesions.

Medical images are acquired in nearly every cancer treatment diagnosis and during cancer treatment and response verification. These images provide a snap shot of the current tumor environment at a macroscopic level. The ability for the human eye to discern phenotypic features of a lesion is limited by the spatial resolution of the image being viewed and the observer's ability to discern complex patterns, such as the range of characteristics present in a solid tumor, which can take years of training. During tumor growth, angiogenesis typically results in a network of chaotic blood vessels that leads to areas of proliferating cells, hypoxia, and necrosis as the nutrients source grows further away. Biopsies can only sample a small area of the tumor, leaving physicians without a detailed picture of the whole tumor volume and taking enough biopsies to characterize the full tumor volume would be extremely uncomfortable for the patient. The use of medical images has an advantage over biopsies due to their non-invasive nature and

ability to capture the full tumor volume. The underlying biology of a tumor, such as the cell sizes, density, amount of vascularization and areas of necrotic, or dead, tumor and actively proliferating tumor, cannot easily be detected in medical images with high precision at present. One area of active research in radiation oncology is to use computer vision techniques, such as texture analysis, to identify phenotypic signatures of the tumors to inform treatment options. The biological differences in tumors are hypothesized to lead to changes in the visible tumors which may appear through more advanced analysis of the medical images. By identifying these patterns, researchers may be able to link these differences to the underlying pathophysiology and more successful treatment regiments, thereby improving patient outcome and quality of life.

In cancer treatment, there are two competing objectives: to eradicate the tumor cells and to prevent damage to the normal tissue. One of the best ways to reduce damage in healthy tissue is to reduce the amount of radiation delivered to it. During the course of radiation treatment, the physician outlines the boundaries of the tumor or structure to be treated during contour delineation. This contour of the tumor is referred to as the gross tumor volume (GTV). The GTV is expanded upon during the treatment planning process by a margin to account for microscopic disease that is not evident to the human eye on the images. The expanded volume is referred to as the clinical target volume (CTV). Depending on the treatment method, the CTV is expanded again to account for uncertainty in the ability to localize the tumor during treatment, such as set-up inconsistencies, mechanical tolerance of the equipment, etc., into the planning target volume (PTV). For lung cancer and other mobile tumors, the CTV is first expanded into an internal target volume (ITV) due to the large amount of motion that can occur during

breathing combined with cardiac motion. An ITV is constructed from a free breathing computed tomography (CT) scan where the entire path of the tumor from end inhalation to end exhalation is considered part of the ITV. This ITV is again expanded into a PTV to account for the uncertainty in set-up and mechanical tolerances. During treatment, the prescribed radiation dose is delivered to the entire PTV with a desired coverage percentage, usually 95%. The larger the margins used to expand the tumor, the more normal tissue may be included in the treated volume and thereby irradiated to higher doses, increasing the chance of damage to the normal tissue. On the other hand, margins that are not large enough risk missing a portion of the tumor, allowing it to recur and the patient to potentially undergo treatment again. Recent advances in machine learning have shown the ability for a computer to learn how to differentiate between classes of images and locate different objects of interest by learning different features present in the images. These networks are being actively explored to segment an image and could, in the future, assist with defining the target volumes with greater accuracy.

This work will explore two different avenues of improving individualized treatment: using texture analysis to explore correlations between medical images and treatment response, and using machine learning via deep neural networks to improve contour delineation uncertainty. The following section provides further details on the treatment process for non-small cell lung cancer (NSCLC), image types, predictive modeling through radiomics, contour delineation uncertainty, and techniques for image segmentation/classification using machine learning. The section concludes with an overview of the dissertation, specific aims, and innovation.

1.1 Non-Small Cell Lung Cancer

Lung cancer is the second most commonly diagnosed cancer in the United States behind skin cancer, and has the highest mortality rate of any cancer for both men and women in the world.[1, 2] The World Health Organization estimates 19.4% of cancer related deaths are from lung cancer.[2] The overall incidence rates for lung cancer have been declining; however, the overall 5-year survival rate for all stages of lung cancer remains poor at 17.7%.[3] Lung cancer has two main classifications, small cell lung cancer and NSCLC. NSCLC, which is the focus of this work, accounts for approximately 86% of diagnosed lung cancer cases.[4] As with most cancers, the earlier the cancer is diagnosed, the better the chance of survival. Unfortunately, only 16% of cancers are diagnosed at a local stage which has a survival rate of 59.2%.[3] For the majority of NSCLC patients, the cancer is more advanced at the time of diagnosis with 55% of patients being diagnosed with distant tumor spread and 24% being diagnosed with local-regional spread where there is a 31.4% 5-year survival rate.[3]

NSCLC is typically diagnosed by a chest CT image, and the identified nodules are biopsied to determine histology. In addition to the CT scan, a positron emission tomography (PET) image is typically acquired to help identify involvement of lymph nodes, metastases, and active tumor, especially in presence of collapsed lungs, called atelectasis, or other pathologies that make the tumor boundaries difficult to ascertain from CT images alone. Atelectasis typically occurs near the tumor where the lesion growth has obstructed one or more air ways. Depending on the location, this could collapse a small portion of the lobe, the entire lobe, or the entire lung. Lung cancer can metastasize to different areas of the body, but most frequently metastasizes to the lymph nodes,

adrenal glands, and brain. For advanced stage lung cancer, a brain magnetic resonance image (MRI) is also typically obtained to check for metastasis as the PET scan would not be sufficient. The high level of activity in the brain causes a proportionally high uptake of the PET tracer obscuring the location of any metastasis making MR, which has excellent soft tissue contrast, the ideal image for diagnosing metastases to the brain.

As mentioned earlier, a biopsy is often taken in order to determine the histology of the tumor and also to look for the presence of genetic markers. There are three common types of histology for NSCLC: adenocarcinoma, squamous cell carcinoma, and large cell carcinoma. The adenocarcinoma typically begins development in the alveolae of the lung and is slower to grow. Squamous cell carcinoma typically begins in the cells that line the airways of the respiratory track and is faster growing than adenocarcinoma. The large cell carcinoma has large and abnormal cells and is assumed to have neuroendocrine origin.

Several different treatment options for NSCLC exist, and the treatment selected depends on several factors including location, size, and stage of tumor as well as the overall health of the patient. For early stage and smaller tumors where the patient is healthy enough to withstand removal of part or all of a lung, the tumor can be removed by surgery. Surgical options include pneumonectomy, lobectomy, segmental resection or video assisted thoracic surgery (VATS). Pneumonectomy removes the entire affected lung while a lobectomy removes the entire affected lobe. If the tumor is small enough, only a portion of the lobe can be removed through a segmental resection or VATS. Surgery can be combined with radiation and/or chemotherapy to treat any possible microscopic disease remaining following surgery. When surgical resection is not possible,

radiation therapy with or without chemotherapy is the common course of action. Depending on which genetic markers are present, there are several emerging and established immunotherapy treatments that can be used to increase the overall survival of patients, protect normal tissue, or increase effectiveness of the radiation or chemotherapy. During radiation therapy, high energy photos or protons are aimed at the lesion from several different angles. As these photons and protons travel through different mediums, they deposit energy to the surrounding tissue or air. The amount of energy imparted to the medium by charged and uncharged particles interacting within the volume per unit mass is referred to as the absorbed dose an is measured in Gray (Gy). During radiation treatment, typically the tumor is treated to a prescribed dose of between 50-70 Gy. Small tumors are candidates for Stereotactic Body Radiation Therapy (SBRT) where a small number (typically 1-5) of high dose treatment sessions, or fractions, are delivered to the tumor with ablation as the goal. Larger tumors receive smaller dose fractions, approximately 1.8-2 Gy at a time, over a longer period of time, typically 6 weeks.

The goal of the radiation treatment is to induce tumor cell death by destroying the DNA or preventing further cell division. The radiation dose that is delivered to tissue can directly or indirectly damage the cells. Direct damage is when the photon, or charged particles that the photon has excited such as electrons, interact with the DNA strands and break them. More commonly these photons and charged particles cause indirect damage by interacting with the water in the body to create free radicals that damage the DNA. The higher the dose, the more damaging interactions occur. As previously mentioned, the beams are spread out at different angles in order to reduce the dose delivered to the healthy tissue while overlapping the beams within the tumor and delivering the full

prescription dose. In addition, breaking the total dose to be delivered into smaller fractions allows the healthy tissue to repair damage to the DNA between treatment sessions and can reduce the risk of normal tissue complications. Tumor cells are also able to repair during the time between fractions, but they are less efficient and are preferentially damaged during radiation therapy.

Response to radiation treatment is often determined by the response evaluation criteria in solid tumors (RECIST). The RECIST criteria are based on a change in the sum of the longest diameter(s) of the lesion(s) and involved lymph nodes. Treatment response is classified as complete response, partial response, progressive disease, or stable disease depending on the percentage decrease or increase in this longest length.[5, 6] Lung tumors are known to shrink during the course of radiation treatment with various studies finding approximately 1.2% reduction of the tumor volume per day,[7] or 44-51% reduction in volume by the end of treatment.[8, 9] However, the decrease is not consistent from fraction to fraction.[8] Therefore, final tumor response often cannot be determined using RECIST criteria until weeks following the completion of treatment. There is a desire to predict the treatment response earlier in order to provide the best patient care. By predicting response and tumor control earlier, physicians can potentially change the treatment regimen by modifying overall tumor dose, fractionation schedule, chemotherapy regimen, treatment goal, and other options. The additional information before or early into treatment also allows for the patient and physician to make better informed decisions about continuing and follow-up care. Researchers have been exploring various methods of predicating the tumor response with particular emphasis on pretreatment and early treatment features based predictive models.

1.2 Medical Imaging

The models mentioned before are in part based on imaging features, including texture features, extracted from images that are routinely used in the treatment of various tumors and are acquired at various stages of the treatment process according to the individual studies. Diagnostic images are used to determine the type and location of a lesion and to suggest areas for biopsies. Once the cancer is diagnosed, planning or simulation images are taken. The gold standard for radiation therapy is the CT image. The treatment process begins by acquiring CT simulation images of the patient in the treatment body position with any immobilization devices that will be used throughout the treatment process. These images are transferred to the treatment planning system and used to define the treatment target(s), organs at risk (OARs), and accurately construct the dose deposition within the body. The tumor and OARs delineated on the planning CT images are typically used throughout the treatment process. If the patient undergoes significant changes during treatment, such as drastic weight loss, resolution of atelectasis, or significant tumor volume reduction, additional CT images would be acquired to determine the need for re-planning. Follow-up CT images are used to determine the RECIST classification of a patient after completion of treatment. Cone Bean Computed Tomography (CBCT) images are acquired prior to treatment to insure correct patient alignment and to monitor changes in patient anatomy for significant changes.

For lung cancer and other advanced cancers, a PET scan is often taken in addition to the planning CT. Unlike a CT image, PET images show little to no anatomy and are functional images instead. Radioisotope-labeled glucose, typically Fluorine 18 labeled

fluorodeoxyglucose (^{18}F-FDG), is injected prior to the image acquisition. The intensity, or brightness, of a PET image is directly related to the uptake of the labeled glucose in the active tissue. Since actively proliferating tumor cells have a higher metabolic need than the surrounding health tissue, the active tumor appears bright on the PET images. The PET image is useful in determining the lymph node involvement, identifying any distant metastases, and distinguishing tumor from similar appearing tissue surroundings such as near atelectasis. These PET images are frequently registered to the planning CT and used to assist in the delineation of the tumor in the previously mentioned cases.

MRI has superior soft tissue contrast to CT images and is the imaging modality used extensively for brain tumors. The contours are delineated for the brain on the MR image before being registered with the planning CT and transferred. For lung tumors, it is hoped that the improved soft tissue contrast will aid in contour delineation near the mediastinum and in distinguishing between atelectaisis and tumor. Depending on the imaging signal sequence applied to the tissue, MR scanners can generate anatomical images or functional images. One example of functional MR imaging is the diffusion weighted (DW) sequence. DW imaging seeks to capture the motion of water in and around cells. A sequence of diffusion gradients are applied to the tissue to produce different signal strengths related to the amount and direction of water movement. More restricted water, such as in areas of inflammation due to injury, appear brighter and freely moving water appears darker.[10] The DW images are acquired at different b-values, or strength of diffusion weighting as described by equation (1), typically at least a low b-value and a high b-value.

$$b = \gamma^2 G^2 \delta^2 \left(\Delta - \frac{\delta}{3}\right) \quad (1)$$

Where γ is the gyromagnetic ratio of the element of interest, most commonly hydrogen, G is the amplitude of the diffusion gradient pulse, δ is the duration of the diffusion gradient pulse, and Δ is the time between diffusion gradient pulse pairs. At least two different b-value DW images are then used to generate the apparent diffusion coefficient (ADC) map which captures the magnitude of diffusion of water within the tissues. The ADC map is being researched as an alternative functional imaging modality to PET imaging. The ADC map's intensity has the inverted meaning of the DW images where a dark signal indicates areas of restricted water movement. In tumor cells, the movement of water is more limited given the irregular shape and spacing of the cells in the tumor mass and necrotic regions which, similar to the standard uptake value (SUV) in PET images, can better discriminate between tumor and healthy tissue than anatomical appearance alone. Several MRI sequences are designed to show anatomical features, unlike PET, potentially allowing for better specificity in location given the clearer definition of the anatomy without the additional CT scan. These anatomical MR images can be acquired in the same session as functional MR image sequences. Furthermore, the spatial resolution of a PET image is relatively poor when compared with CT and MR images, so a functional MR image could potentially increase the target definition. For these reasons, there has been an increased interest in exploring the potential of MR in lung cancer predictive modeling as will be explored in this work.

1.3 Predictive Modeling

Predictive modeling is using patient data to discern trends that have the ability to give insight into the likely response to therapeutic treatment. These predictive models will

hopefully provide insight into probable treatment response allowing physicians to adapt treatment plans as necessary, determine appropriate follow-up care, and provide more detailed information to patients. Statistical analysis of retrospective patient data is preformed to identify which factors correlate with desired or adverse outcomes. The patient data used in predictive modeling ranges from clinical factors, such as tumor stage, volume, location, and lung function test performance, to more advanced imaging features derived from computer vision techniques and radiomics and, more recently, genomic data. The hypothesis underlying most of these studies is that the tumor microenvironment exhibits observable characteristics that can be used to predict response to different treatment regiments. For lung cancer patients, overall survival, local control, freedom from distant metastasis, and radiation induced lung injury are the major clinical outcomes, or endpoints, that have been explored. Different modeling techniques have been explored including single modality image analysis for pretreatment images and, more recently, time series analysis on images acquired at multiple time points during the course of treatment. CT and PET are the most commonly studied imaging modalities for lung cancer, but with superior soft tissue contrast, MRI is experiencing increased interest.

Pre-treatment images are of particular interest to researchers as they have the potential to provide insight into a treatment before it begins. One model, by Balagurunathan et al.,[11] found texture features related to homogeneity of the tumor were significant predictors in overall survival. In particular, tumor with indicators of high homogeneity were predictive of longer survival.[11] Another later model by Fried et al.,[12] used texture features extracted from pretreatment 4D contrast enhanced CT images and the 50% phase of the 4D CT image to stratify patients into risk groups based on Kaplan-

Meyer curves. They found a significant improvement in the stratification over models that only used clinical prognostic features. The model's classification repeatability was approximately 80% for overall survival, local regional control, and freedom from distant metastasis.[12] Coroller et al.[13] used pretreatment CT scans to predict distant metastases in lung cancer for adenocarcinoma histology. They investigated features from the CT images, such as texture and shape descriptors, and clinical factors, such as the pretreatment tumor volume, for their ability to predict distant metastases. Univariate analysis identified 35 prognostic features and multi-variate analysis was used to create a final model. The final model with the combined clinical factors demonstrated a significant improvement in identifying patients with distant metastasis, p-value= 1.56×10^{-11}.[13]

In developing predictive models, the emphasis has primarily been on pretreatment imaging features and longitudinal studies have not been conducted in many instances. One early longitudinal predictive model for tumor response by Bral et al.,[14] looked at the amount of volume regression as a predictor for metabolic complete remission of NSCLC. They found, by calculating the regression coefficient from fitting the volume change to a negative exponential curve and using a cut off of 0.03, they could predict the non-responders with 80% accuracy while misclassifying only 16.4% of patients who achieve complete remission.[14] George et al.,[15] evaluated PET images acquired at two different time points: pretreatment and follow-up. Texture features were extracted from the regions of interest (ROIs), in the case of this study the tumor volume, and used to create a subspace signature, which defines the collection of features identified by principal component analysis (PCA) to explain the most variation, for each subject at each time point. Using the distance between the subspaces of two time points, the study was able

to predict the RECIST classification with an area under the time dependent receiver operating characteristic (ROC) curve of 0.6676 to 0.6817.[15] This study only included imaging features extracted from PET images. In another study, Jabbour et al.,[16] used the lung tumor volume reduction seen from weekly CBCT images acquired prior to treatment fractions from day 1 to end treatment on day 43. The Cox proportional hazard models showed a 44.3% decrease in death for every 10% decrease in tumor volume between day 1 and day 43.[16] More recently, Fave et al.[17] explored the change in texture features extracted from weekly 4D CT scans' ability to increase the predictive power for local recurrence, survival, and freedom from distant metastasis. They found adding the change in texture features did improve the model for overall survival when compared to clinical factors and pretreatment features alone, but the same was not true of models predicting distant metastasis. They also found local recurrence was significantly predicted by the change in texture features alone.[17]

1.3.1 Radiomics

Recently, the fields of radiomics and radiogenomics have gained popularity. The field of radiogenomics combines the information from clinical data and extracted imaging features with genetic markers to identify correlations between imaging features, genetic markers, and clinical information to predict a variety of clinical endpoints or genetic expression.[18–20] Radiomics, on the other hand, seeks to combine features extracted from imaging and clinically available information to identify potential imaging biomarkers and utilize features in ways to predict response,[11–13, 21–24] segment tissue,[25, 26] classify lesions as benign or malignant,[27–30] and evaluate other characteristics of a tumor.[31–34] One advantage to investigating texture and other imaging features is the non-invasive nature

of the imaging. Images are routinely acquired as part of the radiation therapy workflow, and therefore, by using those images, the patient is afforded no additional dose or time requirements. The main advantage to extracting the imaging features is it quantifies characteristics of the intensity level patterns in the images that may not be readily apparent to a human observer. Imaging, unlike a biopsy, is able to assess the entire tumor volume and quantify the patterns of heterogeneity within the tumor. These patterns are hypothesized to arise from physiological and genomic characteristics of the tumor giving the physician insight into radio resistance and/or sensitivity, and genetic expression which, in turn, can influence treatment decisions.[35, 36]

The general workflow for radiomics is comprised of three basic steps: imaging, feature extraction, and analysis. The process begins with image acquisition in which single or multiple imaging modalities acquire images of the target. After image acquisition, the ROIs are defined specific to the problem being addressed and features are extracted. The extracted features are then analyzed for the end goal of the study, Figure 1.[36] Each step has its unique challenges. Multiple aspects of image acquisition can affect the features extracted from the images including different manufactures, reconstruction, slice thicknesses, imaging protocol, and contour delineation methods.[29, 34, 37–41] Target delineation and generation of predictive modeling are two common uses for the extracted features. Delineation seeks to identify regions that exhibit distinct characteristics from the surrounding medium that helps define the location of a tissue or ROI, where predictive modeling seeks to identify the features within an ROI that are correlated with treatment outcome or other clinical endpoints. There are thousands of possible features that can be extracted from the ROIs. This, coupled with various image preprocessing steps, makes

radiomics a very high dimensional problem. High dimensional problems with limited data introduce another challenge: insufficient data to evaluate all the potential parameters. As a result, different methods for reducing the number of features have been employed and analysis should control for false discovery rates. Feature reduction seeks to employ statistical techniques to reduce the number of correlated features, thereby reducing redundancy, and to evaluate the features that are most clinically relevant. By decreasing the number of features, the dimensionality of the problem is reduced. False discovery rates arise when multiple hypotheses are being tested using same data. If a large number of features are tested for significance, then some of them are bound to be significant due to chance. If 100 features are tested at the 5% error rate, 5 features can be expected to be significant by chance alone. A more detailed description of false error rate control can be found in 2.4 False Error Rate Control. While early research shows there is potential promise in radiomics, experts stress an overall need, as this field of study matures, to create best practices and standardize methods.[36, 42–44]

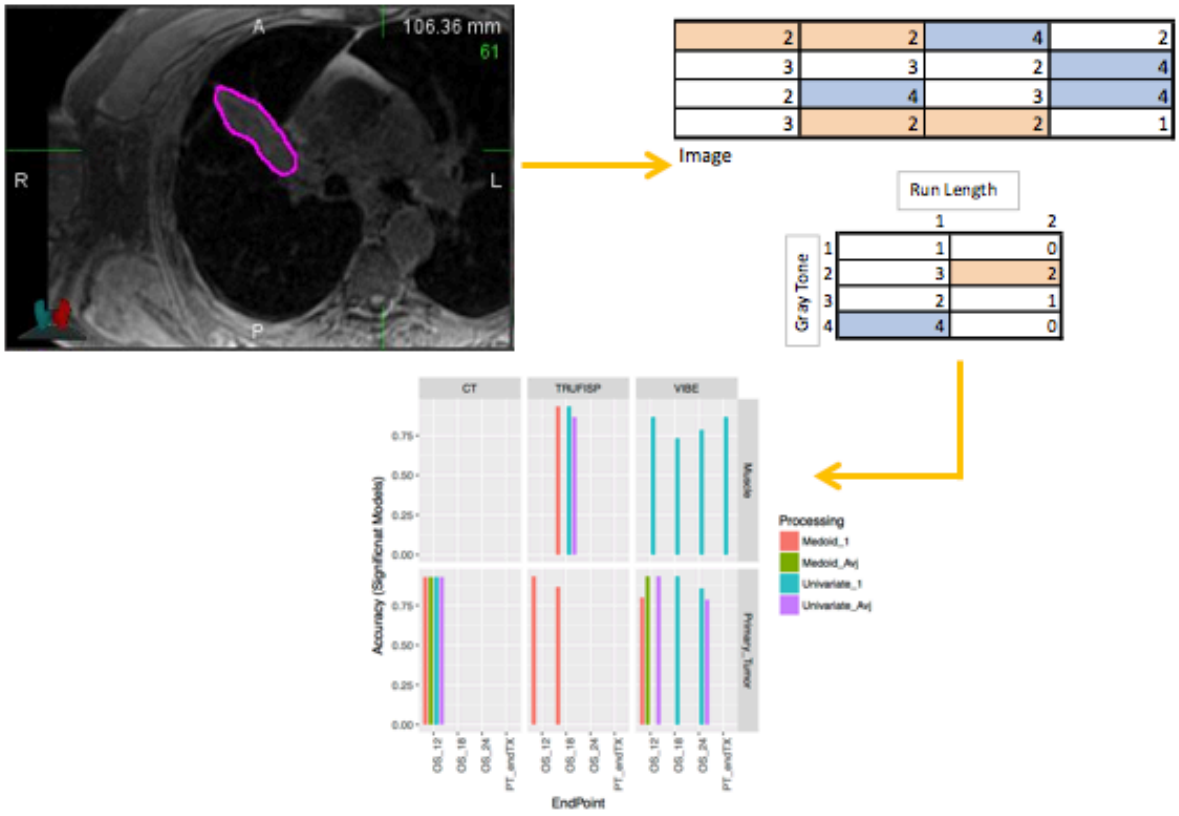


Figure 1: Visual representation of the radiomics workflow. The target is defined (top left), features are extracted (top right) and analyzed for the study end goal (bottom).

In the following section, the texture features and image processing steps employed for this work will be described in further detail. The texture features and imaging processing steps used are not intended to be an exhaustive sampling of all available texture features and image processing techniques but, rather, an application of the frequently reported and promising techniques from lung and other sites applied to NSCLC for MR and CT images.

1.3.1.1 Texture Features

At the core of radiomics are the image features. There are several types of image features that can be extracted from an image, many of them stemming from the field of

computer vision. Information about the tumor volume, surface area, and shape can be determined from the physician delineated contours. Texture features can be calculated to describe the different patterns of intensity within the contour or a neighborhood surrounding a location of interest. Texture analysis, or the use of texture features to describe an image, is used in the field of computer vision to classify images, perform segmentation, enrich details of objects in video games, and determine the shape of an object.[45–47] First order texture features include basic histogram features derived from the intensity values. Higher order texture features include the gray level co-occurrence matrix (GLCM), gray run length matrix (GLRLM), neighborhood gray tone difference matrix (NGTDM), gray level size zone matrix (GLSZM), and others which seek to capture spatially varying texture features. These texture features are frequently investigated in the literature and are regarded as the easiest to compute.[22]

The histogram texture features are the more traditional texture features relating to the intensity distribution within the ROI and do not include spatial information. The ROI is first defined using autosegmentation, semi-automatic segmentation, or manual delineation. Afterwards, a histogram of the intensity levels is created and used to compute statistical descriptions of the intensity distribution such as the mean, minimum, maximum, standard deviation, skewness, etc. Figure 2 shows an example of an intensity histogram derived from an example image slice.

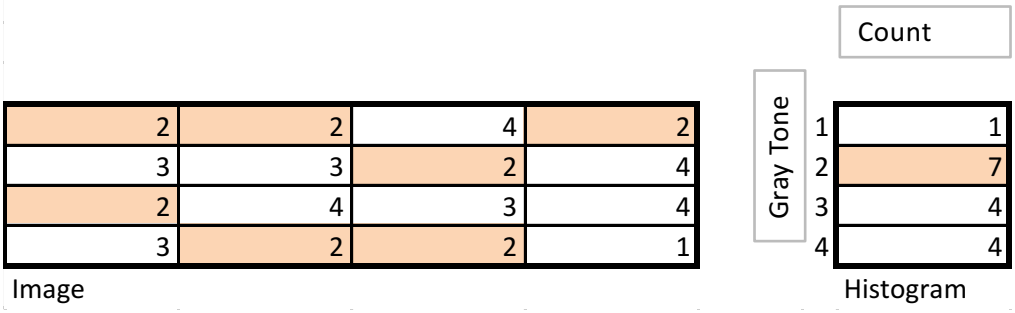


Figure 2: Sample image with corresponding histogram. The highlighted portion of the image corresponds to the highlighted count in the histogram.

The GLCM extracts features based on the probability of pixels (2D) or voxels (3D) at a given distance apart being from two different gray levels.[48–51] The GLCM is an N by N matrix where N is the number of gray levels in the image matrix post any desired processing. There are two hyper parameters that govern the GLCM calculation: the direction and the distance. Each element in the GLCM matrix, p(i,j), represents the number times pixels/voxels of gray level i and j appear in a designated direction such as 0, 90, or 45 degrees and distance, such as 1, 2 or 5 pixels/voxels, away from each other, seen in Figure 3. This matrix is normalized prior to calculation of the texture features making each element represent a probability as opposed to the number of elements. The GLCM as described by Haralick et al.[48] is symmetric and therefore p(1,2) is the same as p(2,1). By exploiting this symmetry, the number of connections needed to fully describe the image is reduced by a factor of two.[48] A different GLCM can be calculated for each combination of the connection directions and distances. In order to make the features more robust, the average of the 4 connected directions for 2D or the 13 directions for 3D, when exploiting symmetry, can also be averaged together thereby removing the directional dependence of the texture features.

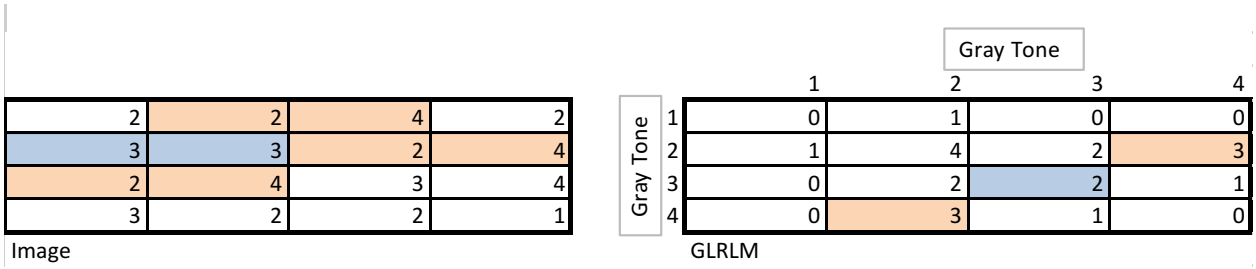
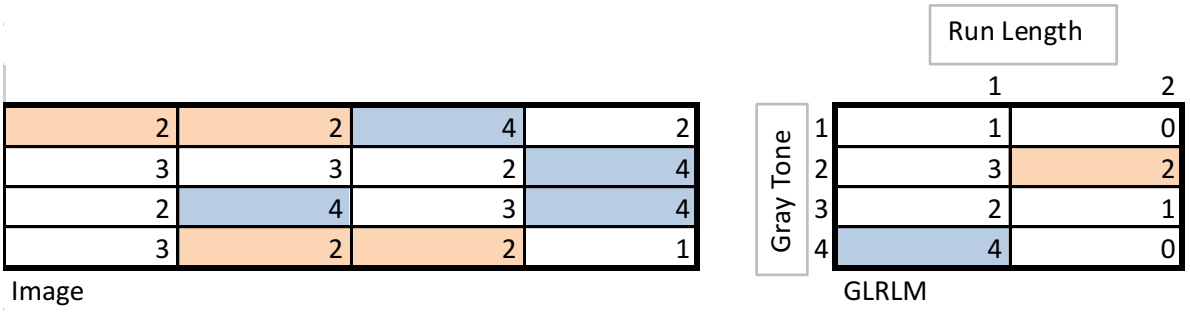


Figure 3: Sample image with corresponding gray level co-occurrence matrix(GLCM). Here the 0 degree and, with symmetry, 180 degree angles along the x-axis at a distance of 1 pixel hyper parameters are being used to calculate the GLCM. The highlighted boxes show corresponding calculations. Notice the blue highlighted boxes in the image correspond to a value of 2 in the GLCM due to symmetry.

The GLRLM describes features extracted from the probability of pixel or voxel with the same gray levels having an unbroken linear connection along a give direction of varying lengths.[52–56] The GLRLM is an N by M matrix where N is the number of gray levels and M is the length of the longest continuously connected run of a single gray level in the specified direction, see Figure 4. Originally, the directions used were the principal directions such as 0 and 90 degrees as opposed to 45 degrees, but the diagonal directions can be calculated as well. This type of analysis is useful in evaluating the linear structure of an image. Each element in the GLRLM, p(i,j), represents the number of runs of gray level i with length j. The GLRLM is normalized before calculating the texture features and each element in the matrix now represents the probability of having a gray level i with length j. The different directions can be averaged together like the GLCM to increase robustness by calculating a directionally independent feature. The original

features proposed by Galloway[52], were inspired by the Haralick features and later papers expanded these features.[53–56]



2	2	4	2
3	3	2	4
2	4	3	4
3	2	2	1

1	1	0
2	3	2
3	2	1
4	4	0

Figure 4: Sample image with corresponding gray level run length matrix(GLRLM). The run length along the x+ direction is calculated with highlighted boxes showing corresponding calculations.

The GLZSM is an extension of the GLRLM in which connections in all directions are considered not just linearly in one direction.[57, 58] The GLSZM is a N by M matrix where N is the number of gray levels desired in the image, and M is the number of pixels/voxels, making up the largest connected patch of a single gray level. Each element of the GLSZM, p(i,j), represents the number of clusters of gray level, i, being comprised of a total j pixels/voxels, see Figure 5. Unlike the two previous texture feature classes, the GLSZM does not have a directional dependence. Cluster sizes are determined by looking for pixels/voxels of the specified gray level that are adjacent to another pixel or voxel of the same desired gray level in any of the 8 (2D) or 26 (3D) connected directions and continuing until all connected pixels or voxels are identified. Where the GLRLM captures linear structural information, the GLSZM, which was originally developed by Thibault et al.[58] to classify cell nuclei, seeks to provide insight into large homogenous areas and intensity changes. Prior to calculating the texture features, the matrix is normalized again transforming the raw counts into probabilities. The texture features calculated from the GLSZM are the same as for the GLRLM.

2	2	4	2
3	3	2	4
2	4	3	4
3	2	2	1

Image

Gray Tone \ Zone Size	1	2	3	4
1	1	0	0	0
2	0	0	1	1
3	1	0	1	0
4	1	0	1	0

GLSZM

Figure 5: Sample image with corresponding gray level size zone matrix (GLSZM). Connections in any direction are considered in determining the zone size. Highlighted boxes show corresponding calculations.

The NGTDM extracts features from the variation of gray levels in a local neighborhood around a selected pixel or voxel.[59] The NGTDM is a 1 by N matrix where N is the desired number of gray levels. Unlike the other texture feature matrices, the elements of the NGTDM are not based on the number of different gray levels in a defined relationship with the gray levels spatial connected to it, but rather the average of the gray levels surrounding pixels/voxels of the specified gray level. The NGTDM element values, p(1,i) represent the sum of average difference between pixels/voxels of intensity i and their surrounding neighbors throughout the entire ROI. First, the average gray level in a neighborhood of defined size, such as the elements directly connected to the central pixel or voxel, of a desired gray level, i, excluding the central pixel or voxel of interest is calculated. Then this average is subtracted from the central gray level, i, of interest thereby calculating the neighborhood gray tone difference. This calculation is repeated for every pixel/voxel of intensity, i, throughout the valid portion of the image. Finally, all the calculated differences per value of i are summed together to create the final element p(1,i) value in the NGTDM. The valid portion of the image includes all pixels/voxels whose surrounding neighborhood is completely within the image boundaries.[59] As with the other

matrices, the values are normalized prior to calculating the texture features. An example of the NGTDM can be seen in Figure 6.

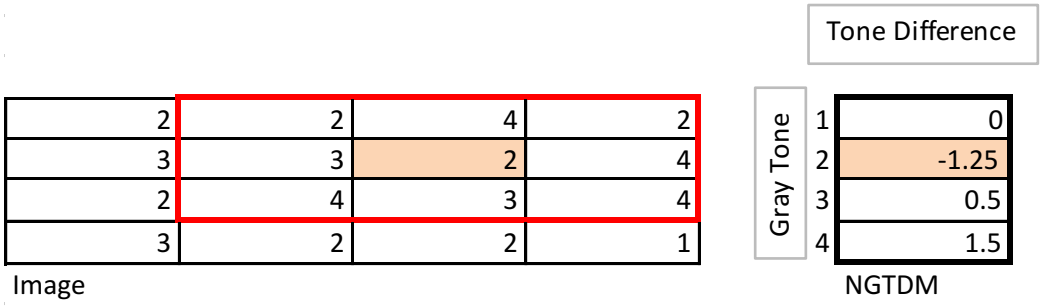


Figure 6: Sample image with corresponding neighborhood gray tone difference matrix (NGTDM). The average of the non-highlighted cells within the red box is subtracted from the highlighted central box in the image to calculate the highlighted value in the NGTDM.

Texture features are dependent on the distribution of gray levels within the image. As such, even subtle changes to the gray level intensities can alter the value of the texture features calculated. Ways to reduce or change this dependence can include quantizing the data into fewer bins than the number of gray levels in the ROI, and applying filters to strengthen texture features either directionally, as with wavelet filters and Gabor filters, or by scale, as with Laplace of Gaussian filters. Wavelet filters act as simultaneous low and high pass filters and decompose an original image into pure edge (high pass), pure contrast (low pass), and combination of directional edges and contrast images.[60] For a 3D-image, the image is decomposed into 8 different images by applying different permutations of either the high pass or low pass portion of the filter on the row, column, and slice directions. The decomposed image can be reconstructed with different weights to emphasize edges or contrast in different directions or throughout the image. The Laplace of Gaussian filter identifies intensity changes within an image via the Laplace gradient operator after Gaussian blurring has been applied to the image. The Gaussian

filter blurs the features finer than the chosen width, allowing emphasis of different coarseness levels of detail.[28] Gabor filter banks are similar to wavelets in that they are spatially and directionally defined; however, they have the form of a Gaussian modulated complex sinusoidal that highlights edges along the given direction and spacing. These filters can be made rotationally invariant by either using a circularly symmetrical filter[61] or by aligning the feature vectors for each image along the highest total energy, or dominant orientation, filtered image.[62]

1.3.1.2 Biomarker Characteristics

In order for the texture features extracted from the image to be clinically useful, they must meet certain requirements. The potential for using radiomics features as biomarkers is still under investigation with varying degree of success having been reported thus far. Some features, such as the SUV in PET, have been studied in greater detail and can be linked to tumor characteristics such as metabolic activity.[63] Other texture features, such as the ones presented earlier, are not directly linked to tumor phenotypes or other biological characteristics definitively and are being actively researched.[11, 12, 18, 20, 22, 26, 40, 64–68] Repeatability, reproducibility, dynamic range, non-redundancy, ease of calculation, and a reasonable link to a biological characteristic, radiographic appearance, or endpoint are all traits that are highly desirable for imaging features to be used in the predictive models for this work.[36, 69, 70] Repeatability refers to the ability of a potential imaging biomarker to obtain the same value given a short time interval and under the same imaging acquisition conditions; while reproducibility is the ability to obtain a similar value for the imaging protocol under changes in equipment or centers.[69] The dynamic range refers to the range of potential values of a feature. If the dynamic range is too small, it

may not have the ability to distinguish between different classifications.[36] With the large range of texture features being investigated, there is a high probability that one or more features are correlated. The set of features used in predictive models should avoid highly correlated or redundant features as this leads to multicollinearity issues with the final model. To this end, researchers have been reducing the number of features by identifying the redundancies and utilizing only those features that are the most repeatable and have a larger dynamic range.[36, 40, 71] In this work, the number of texture features will also need to be reduced due the limited number of patients available for the study. In regression, at least twice as many observations as variables are recommended in order to have a decent fit. Investigating the direct link between the texture features and tumor phenotype or genetic expression is in its infancy and is currently an area of research for different imaging modalities.[20, 72] Despite the lack of a concrete link between the texture features and the underlying biological cause, texture features and their changes over the course of treatment can be useful if they are linked to the appearance of the tumor to an observer. The texture features are being investigated in this work, as said by Hunter et al., "under the hypothesis that they are related to gene expression and phenotype."[73] In other words, this work is assuming that texture features and changes in texture features beyond the threshold of repeatability are related to a biological or gene expression change. This is not completely unfounded as areas of necrotic tissue appear different on images and chaotic vasculature may lead to inhomogeneity of the solid tumor appearance.

Texture features are being actively investigated for CT, MR, and PET imaging modalities. Several studies have successfully been conducted on the repeatability and reproducibility of texture features utilizing CT[11, 12, 24, 26, 40, 41, 66, 71] and PET.[31, 32, 34

However, researchers have had less success in determining the reproducibility of MR texture features in T1-weighted and T2-weighted images due to changes in imaging parameters,[37] machine parameters,[23, 29, 39, 74] imaging protocols,[39, 74] and image characteristics.[38, 75, 76] One recent study published by Gourtsoyianni et al.,[77] on MR texture features for T2-weighted turbo spin echo sequences in liver images showed very low repeatability of higher order texture features (GLRLM, NGTDM, GLSZM) and more repeatable features in the global (histogram) and GLCM features. Other than the Gourtsoyianni et al. study, the repeatability of the T1- and T2-weighted texture features has not been prominently studied in the literature. On the other hand, the repeatability of apparent diffusion coefficients (ADC) as calculated from different b-value diffusion weighted MR (DW-MR) images and dynamic contrast enhanced MR (DCE-MR) images have found success.[64, 78, 79]

This work will seek to identify if there are any T1- and T2-weighted MR texture features that are repeatable under the same imaging conditions as a starting point for finding potential predictive models utilizing MR texture features. DW-MR and ADC texture features will also be evaluated for repeatability and compared to the literature. In addition, this work will also seek to identify CT features from the imaging protocols used at Virginia Commonwealth University (VCU) and compare them to the features found in the literature for agreement.

1.4 Contour Delineation Uncertainty

Accurate target localization is essential to radiation therapy. In the radiation treatment process, as previously mentioned, the physician delineates the GTV on the planning CT. Afterwards the GTV is expanded by adding a margin to account for

microscopic disease that may not be seen with the unaided eye on the clinical images to the CTV often about 5-8mm for lung cancer. The CTV is further expanded to account for different uncertainties which have been reported by Sonke and Belderbos[80] as having standard deviations for interfraction setup errors of 4mm systematic error and 4mm random error, motion of 0-7mm systematic and 0-7mm random error, delineation uncertainty of 2-7mm systematic error, baseline shifts of 4mm systematic error and 3mm random error, and intrafraction target motion of 4mm systematic error and 4mm random error from their literature search. The traditional margin recipe used was proposed by van Herk[81] and consisted of estimating systematic and random errors, such as those reported above, adding the all systematic errors in quadrature to get an overall estimate of the systematic errors and then doing the same for the random errors before using the overall estimates in a population based margin formula to achieve 90% of the population receiving a cumulative dose of 95% of the prescription to the CTV. In some cases, this population based margin could be quite large, such as the 12mm illustrated by van Herk for a prostate case.[81] One particular danger for lung cancer patients, is the risk of toxicities, such as radiation induced pneumonitis, which are shown to correlate with the mean lung dose and the volume of the lung receive more than 20Gy, meaning smaller margins are desirable. The use of 4D gating, breath hold, and image guided radiation therapy (IGRT) can be used to reduce the uncertainty from motion and set up. With 4D gating, the respiratory motion is monitored and the beam is only turned on during a certain phase of motion. Alternatively, with breath hold techniques, motion is limited by arresting breathing for periods of approximately 20s while the beam is on and allowing the patient to breath freely between breath holds. With IGRT, images are taken prior to delivering

radiation to align the planning CT image with the anatomy of the day, reducing the uncertainty in the set up error. These techniques move away from the population based margin to a patient specific margin model aimed at reducing margins where possible.[80] However, these methods do not address uncertainties from contour delineation.

The physician-drawn contour is taken as ground truth throughout the radiation treatment process. However, the contour delineation process can be complicated in areas where there is a lack of a clear boundary, such as low intensity difference between the tumor and the surrounding tissue. In such instances, for example near areas of atelectasis in the lung, the boundary between the tumor and surrounding tissue is difficult to distinguish with the eye and can lead to multiple interpretations by single or multiple observers. Studies have been conducted to quantify the amount of inter- and intra-observer delineation uncertainty. In a study comparing the contour delineation of NSCLC with 14 radiation oncologists and hematologic oncologists, Vorwerk et al.[82] found good agreement (defined as more than 70% overlap) in 23.7% of radiation oncologists (different departments) and 35.9% (same department) on the PTV. Karki et al.[83] in a study with seven physicians found an average delineation uncertainty in the GTV of 2.96mm, 2.06mm, and 2.77mm for CT only, PET/CT, and MRI respectively. Giraud et al.[84] found uncertainty of 3.1 cm laterally, 2.8 cm anterioposteriorly, and 2.1 cm craniocaudally in a study on lung GTV delineation among radiation oncologists and radiologists.

Different approaches have been studied to reduce the delineation uncertainty such as using a prescribed protocol,[85] matched CT-PET instead of just CT images,[86] and auto-contouring;[87] however, the delineation uncertainties remain larger than the mechanical

uncertainties. Steenbakkers et al.[86] noted a reduction in lung cancer delineation uncertainty, as measured by the standard deviation of the difference between the individual contours and the median contour, from 1cm with CT only to 0.4 cm with CT-PET. The greatest gains were made in regions bordering atelectasis from 1.9 cm to 0.5 cm. The auto-contouring and, in addition, autosegmentation based techniques seek to create an automatic or semiautomatic process by which an algorithm defines the boundaries of the tumor and/or thoracic organs with little to no input from the operator. These contours can then be checked and adjusted manually if needed. Several of these algorithms are based off intensity changes. Baardwijk et al.,[87] suggested an auto-contouring technique based on the source-to-background ratio for PET/CT images. The manual contours were compared to the auto-contours, which had been edited by the physicians. Auto-contours showed a significant reduction in the variation of the contoured GTV. Others use a grow region technique where the physician chooses a start seed, or seeds, within the target or at the boundary and the algorithm grows the contour from the identified seeds. Gu et al.[88] developed a single click grow region algorithm that would create an ensemble based final contour from the single internal start seed. This method was able to achieve a similarity index of almost 80% with two different observers with 97% repeatability of the contours with 20 different starting seeds.[88] Other methods include graph-cutting and snakes which seek to minimize energy and mutual information to determine a contour, but these methods are time consuming and computationally expensive in some cases.[88, 89] Lu and Higgins[89] suggested a live wire algorithm that created a suggested contour by connecting successively chosen points along the boundary of the contour. The operator selects a starting point then moves the cursor to

another point along the boundary. The algorithm suggests a contour path from the start point to the cursor location which can be accepted by clicking (thereby creating a new starting location for the next piece of the contour) or modified by moving the mouse until the desired boundary is created. Both the inter- and intra-observer reproducibility were found to have a mean of about 98% with a maximum standard deviation of 0.98% for 2D and 3D contours, with a speed increase of up of 14 times for 2D and 28 times for 3D over manual contouring.[89]

Autosegmentation algorithms, like those described above, are often based on intensity driven mechanisms, such as region growing and mutual information, that may fail in areas that are similar in intensity, such as atelectasis, for centrally located tumor near the mediastinum. Other methods, like the live wire method, involve a lot of user intervention. Machine learning and in particular deep learning based convolutional neural networks (CNNs) approaches have an advantage over purely intensity driven methods as they are able to incorporate information about texture and subtle changes in intensity patterns in addition to intensity level. The incorporation of learned features may be able to differentiate areas of similar intensity and appearance. After a neural network has been trained, results can be produced in seconds with very little user input giving them an advantage over live wire and other semi-automatic contouring processes. Neural networks are also learning algorithms which means by adding corrected predictions into the training set, the network can continually be refined to get better with time allowing the algorithm to potentially perform better on difficult cases in the long run.

The reduction of uncertainty in contour delineation would lead to a better and more consistent definition of the GTV and, by extension, normal tissue as well. An increase in

accuracy of the target contour could also have an effect on the amount of dose escalation possible in the target. Some studies have shown that an escalation in the dose is related to an increase in local control of the tumor and increased overall survival for NSCLS.[90–92] However, this effect is not fully understood as other studies have suggested that dose escalation could be harmful, most notably in the randomized stage III clinical trial by Bradley et al.[93] Reducing the uncertainty in the contour delineation could potentially allow the proposed dose escalations without an increase in normal tissue complications.

1.4.1 Machine Learning

Machine learning is the use of programming algorithms to extract information from a dataset to perform a designated task and improve upon the results without each step being explicitly programmed. In machine learning, the user provides a dataset, framework for the algorithm, and rules for updating results, but does not explicitly program each update step instead letting the algorithm learn the appropriate variables for each step through iteration. For example, machine learning techniques have been used to find the separation between different classes such as species of iris flowers. In this dataset, the algorithm uses the input data: sepal length, sepal width, petal length, and petal width, to predict the species of iris without being explicitly told that iris species “A” has a sepal length “x” and petal width “y.” Instead the network learns the appropriate levels of “x and “y” necessary to distinguish the iris species.

Machine learning and neural networks have been around since the early 20th century.[94] The very first neural networks, often referred to as McCulloch-Pitts neurons, were designed to mimic the function neurons in the brain and can be thought of as a logic gate, or series of logic gates, where each neuron accepts an input, calculates an activation, and if the activation is high enough, produces an output signal, Figure 7.

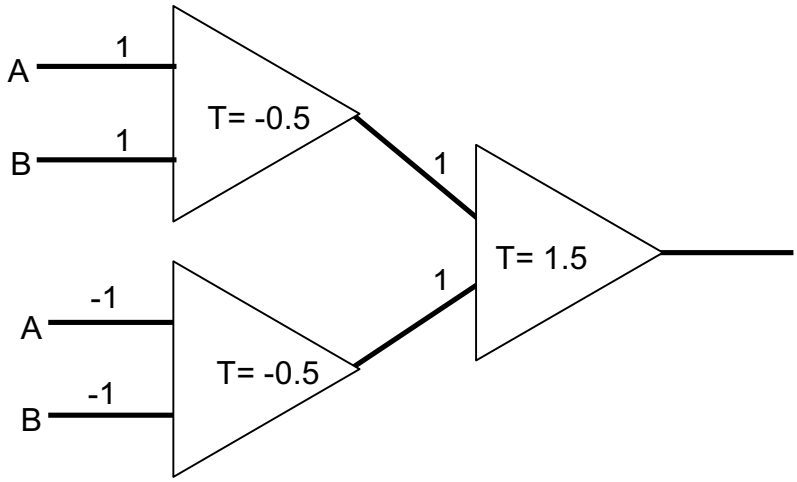


Figure 7: Example of a simple exclusive or (XOR) two-layer neural network. Here the neuron is activated if the sum of the inputs "A" and "B" exceeds the threshold. The numbers along the line indicate the weight to multiply the signal by. All inputs and outputs in this example are 0 or 1.

As seen in Figure 7, each triangle represents a neuron where the input signals are summed and compared to a defined threshold, in this case a hard threshold, where the output is either 1 if larger or equal to the threshold, or 0 if less than the threshold. Each of the inputs is multiplied by a weight indicated by the number above the input and output lines. The input "A" and "B" are binary either 0 or 1 indicating either "on" or "off," or activated (1) and not activated (0). The neurons are arranged into layers as defined by inputs. The first layer in the figure above has two neurons; they both accept the initial inputs. The next layer contains only one neuron and is considered a new layer because

it accepts inputs from the layer before it, regardless of whether it accepts the original input or not, as seen in Figure 8. The final output of the network provides the result of the network analysis on the inputs. Not pictured in either figure is the bias input to each neuron. This bias input can be thought of as similar to the intercept term in the equations of a line. The entire network can alternatively be visualized as a series of linear equations of form $y = ax + b$ that are being solved with the "x" representing the input, "a" the weight, "b" the bias, and "y" the output of each neuron.

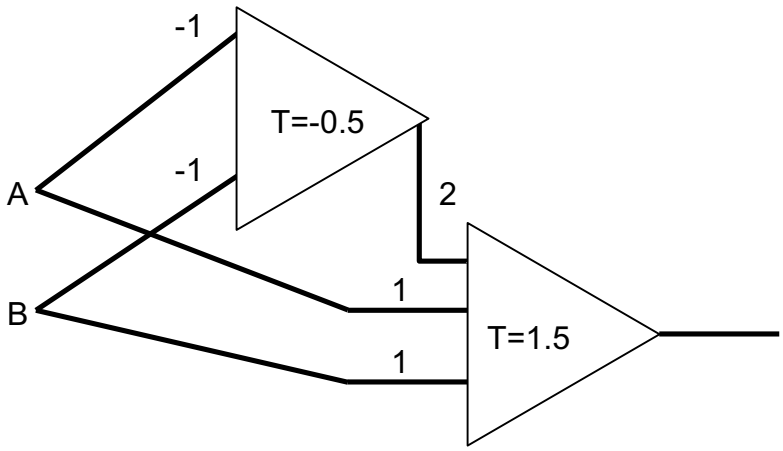


Figure 8: Another example of a two-layer XOR neural network. This time the original inputs "A" and "B" are transferred to the first and second layers of the network.

The learning part of machine learning developed when the networks were able to update their weights according to learning rules. There are two types of learning: supervised and unsupervised. In unsupervised learning, the network is not presented with a "correct" answer to compare the network output to, but rather is given a set of rules and tends to learn patterns that the network thinks are important within the confines of the rules. On the other hand, in supervised learning, the network compares the output with the provided, or human defined, "correct" answer. With either learning method, the

network calculates the error of the network output and uses this error throughout the network to update the weights iteratively until told to stop or it has met some criteria defined by the loss function. The loss function and the optimization method govern how each network updates the weights.

The loss function is chosen for the problem being solved. In the case of a linear regression problem, the mean square error might be chosen; for a binary classification problem, binary cross entropy might be a better choice. In semantic image segmentation, a Dice similarity loss is often employed. The value of the loss function represents the error of the network, and the goal of the optimizer is to update the weights in a manner to minimize the loss.

Early optimizers, or learning rules, were simple such as the Hebb rule[95] for unsupervised learning where the output was multiplied by the learning rate, which dictates how large a step to take at each iteration, and the perceptron rule,[96] where the difference between the output and "correct" answer is multiplied by the learning rate. Today, there are several different optimizers available with one of the most common being stochastic gradient decent (SGD). With the SGD optimizer, the derivative of the activation functions at each layer with respect to the bias and weights are multiplied by the learning rate and error from the loss function before being added to the existing weight. The resulting values then travel backwards to the previous layer where they are again multiplied by the learning rate and derivative of the loss with respect to the weights and biases for the previous layer which are used to update the weights in the current layer and so forth until all weights in the network are updated in a process known as backpropagation.[97] The learning rate dictates how large a change, or step, is taken along the gradient towards

the minimum. Later variations of the SGD algorithm have modifications for regularizing the process to prevent overfitting, such as momentum, which retains a weighted "memory" of the previous weights to help prevent getting caught in local minima.

The activation functions are essential to the learning process and should be chosen for the needs of the problem. The hard activation function mentioned earlier is not the most useful and was replaced early on by soft activation functions. These soft activation functions could output a range of values as opposed to just 0 and 1 for the hard activation functions. Examples of soft activation functions include the sigmoid and hyperbolic tangent functions, which are characterized by low and high plateaus at large positive and negative values respectively, and a gradual slope between them allowing for maximum learning. These soft activation functions allowed the network to reach convergence faster, or at all in some cases. Today, there are a variety of different activation functions such as the very popular rectified linear units (ReLU), which keeps the positive portions of a linear function while setting the negative portion to 0, and the softmax normalized exponential function, which is used frequently in multi-class classification networks.

As technology has advanced, neural networks have become more and more complicated but remained limited in application until the wide spread use of graphic processing units (GPUs). Before GPUs, larger networks remained limited because of the time and computational power required to train a network and the large datasets needed for success. GPUs allowed the time to train a network to be cut down from months to days and today the time has reduced further to minutes, fueling further exploration. With new neural network techniques the input expanded from binary signals to images. In addition, as networks got larger, more and more training images were needed to

adequately train the networks. Labeling of training examples began to be crowd-sourced in order to build larger training datasets. In addition, techniques such as data augmentation, transfer learning, and utilizing image patches were employed to create adequate training sets for networks, especially in the medical field. For medical, and image analysis in general, CNNs have become a particularly powerful tool.

1.4.2 Convolutional Neural Networks

CNNs are networks that use a series of filters of size n in the dimension space of the image, for example n by n for 2D images or n by n by n for 3D images, that are then convolved with the images in a kernel like fashion in order to build a map of hierarchical features. One of the earliest networks, that later became known as a convolutional neural network, was developed by LeCun[98] and was a series of shared weight convolutions. These networks were designed to "read" handwritten numbers and were later expanded to read hand written zip-codes.[99] From this zip-code network, the basic structure of the CNNs used today began to appear: 1) a convolution layer, 2) a down sampling layer, 3) a convolution layer,4) a down sampling layer, and 5) a fully connected output layer. These layers will be described in greater detail in the following sections herein. Today, the CNN is the power house behind almost all image classification, object recognition and localization, and segmentation tasks, such as Alex NET[100], GoogLeNet[101], VGG Nets[102], residual networks (ResNets)[103, 104], efficient neural networks (E-Nets)[105], CIFAR-net[106], and others. The structures of these newer networks have become more complicated than the basic structure outlined above.

As research and network architectures evolved a new concept called deep learning gained popularity. Deep learning refers to using several weight bearing layers in a

network. Though there is no consensus as to how many layers are needed for a network to be considered "deep," the term is generally used in the literature for networks having greater than 5 weight bearing layers. These deeper layered systems, such as ResNets, VGG, GoogLeNet, and others are able to learn more features from images. However, they include a much larger parameter set and need larger datasets to train on.

CNNs have become the go to for any type of neural network trained on raw image inputs. Two areas of research that utilize CNNs, that are relevant to this work, are image classification and image segmentation. The task of image classification is to identify which category an image belongs to. The output is the best guess as to which category, such as cat, dog, bird, etc., the image represents. Image segmentation strives to partition an image into regions having similar characteristics without assigning classes or labels. Semantic image segmentation, on the other hand, returns a map of what class each pixel or voxel is predicted to be, thereby determining the location of the class(es) throughout the image. Semantic image segmentation is similar to another CNN heavy task called object detection. However, the end goal of most object detection algorithms is a bounding box around the object rather than a pixel by pixel map.

1.4.2.1 Convolution layers

The convolution layer can be thought of as the feature detection layer. The user defines the number and shape of the filters employed by this layer, as well as what stride and padding to use. The stride defines how many pixels/voxels the kernel skips before the next calculations, while the padding adds additional pixels/voxels around the image to allow the kernel to get closer to the edges of the image. Padding can be any amount, however, "same" and "valid" are the two most commonly used. The "same' padding

provides only enough padding around the image to allow for the same size after convolution with the filters, while 'valid' adds no padding allowing the output shape to be reduced.

The output size of any layer can be calculated using the following equation:

$$O = \frac{I-F+2P}{S} + 1 \quad (2)$$

where, I is the size of one dimension of the input image, F is the size of the filter, P is the amount of padding, and S is the stride. This calculation is repeated for each image dimension. Square or cube images and filters are often used, so this calculation is usually performed once per layer. There are two other dimensions included in the size of the output layer: the channels and the number of filters. Color images use 3 channels one for each of the red, blue, and green values, while gray scale images only have one channel. Some techniques when working with medical images provide different slices of a 2D medial image to each of the 3 channels.[107, 108] It is important to keep track of the output shape of each layer when designing a network as typical CNNs tend to reduce the size of the image and increase the number of filters as the network gets deeper.

Each filter has its own set of weights and bias terms and seeks to capture information from a local portion of the input image. At lower levels in the network, the convolutions act much like edge detectors and color pattern filters. For example, when trying to classify an image as containing a face, the early filters may look for vertical edges such as the side of a nose or jaw and oval patches of white for the eyes, Figure 9. The subsequent layers look at local patches of the lower level features from the previous layer and begin to construct higher level features of an image like, in continuing the example, eyes, lips, or nose. This continues until the network learns which features need to be present in order to positively identity the image as a face. Convolution layers are typically

followed by a non-linear activation function such as the ReLU activation function described earlier.

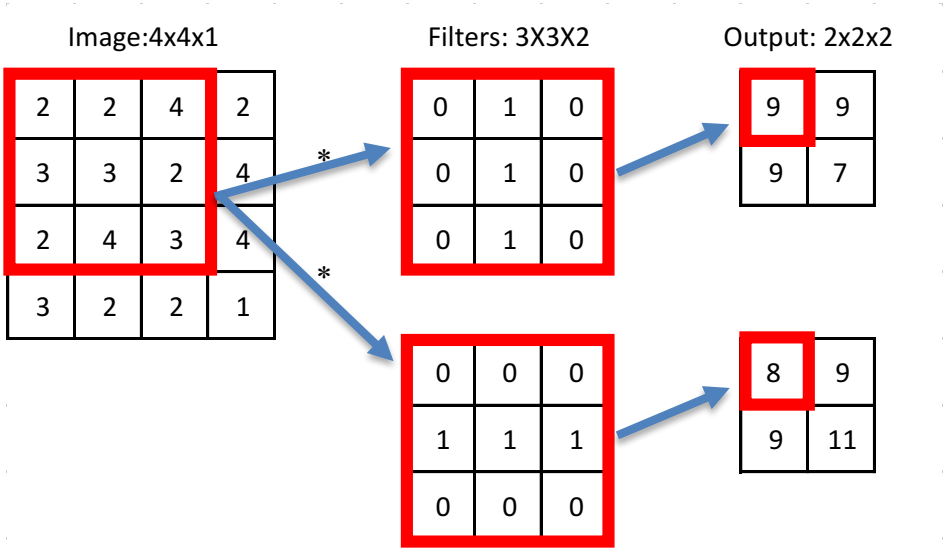


Figure 9: Example of a convolution layer for a 4x4x1 image with two 3x3 filters, valid padding and a stride of 1. The filter weights are seen in the second column with the resulting output in the third column. The portion of the image in the red box is convolved with the weights to produce the values within the red box in the output layer. The entire filters slides over by one horizontally or vertically to fill the rest of the output image.

CNNs can have a large number of weights and biases, also referred to as free parameters. For one layer with 16 filters of size 5 by 5 convolved using a stride of 1 with an original image of size 50 by 50 and same padding, there are about 40,000 free parameters. With CNNs, it is not uncommon to have tens to hundreds of thousands of free parameters. In order to reduce the number of free parameters at a given layer, periodic down sampling or pooling is often employed.

1.4.2.2 Pooling layers

Pooling layers are often used to reduce the size of the layer input and number of parameters. These layers employ a user defined filter size and stride but do not have

weights or biases to tune. These layers are often not included in the count of how many layers a network contains. Frequently, the stride will be chosen to have the same length as the filter size so as to have non-overlapping connections. Others, however, have employed a non-overlapping pooling layer with success.[100] There are different types of pooling that can be applied. The most common are maximum pooling and average pooling, Figure 10. In maximum pooling, the largest value within the filter at each kernel stride is kept while average pooling, as the name suggests, averages the value within the filter and assigns the average to the corresponding pixel/voxel in the output. Pooling layers typically are not followed by an activation function.

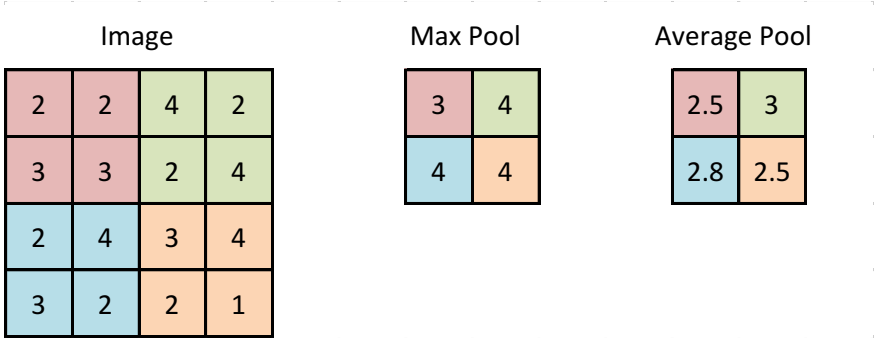


Figure 10: Maximum and Average pooling layers with a 2x2 filter and a stride of 2. Colored boxes in the image correspond to colored boxes in the output following pooling calculations.

1.4.2.3 Fully Connected layers

Fully connected layers, as the name suggests, connects each element, pixel/voxel or neuron depending on the network architecture, of the input layer with each element of the output layer. These layers serve to bring all the information the network has learned together to provide a final output. In more recent years, the traditional dense fully connected layer has been replaced by the use of convolution layers using a filter size of

1 for the given image dimensions, as it is computationally faster and accomplishes the same end goal. The user defines how many output neurons are needed depending on the goals of the network. For example, in the final layer of a classification network, the fully connected layer typically has the same number of neurons as classes being classified. Each neuron in the final layer can be seen as measuring the probability of each class given the activation levels from each of the high level features learned by the network. The fully connected layers are typically followed by an activation function with the very last layer utilizing the activation function necessary to provide the desired output from the network as described previously.

1.4.2.4 Batch Normalization

Batch normalization layers seek to normalize the input values across the batch, or chunk of data with a user specified size, being analyzed by the network to maximize learning. The activation functions that can be applied to a particular layer often have a region where learning is maximized and excessively large or small values saturate the activation function, making learning a very slow process. During the process of training, the weights of the network are updated at the end of a batch. The size of a batch is usually determined by how large the dataset is, where smaller datasets use batch sizes as small as 1, the desired speed of training, where larger batches trained faster, and the memory limits of the system, which provides the upper limit on the possible batch sizes. The batch normalization layer standardizes the batch of inputs to between a range of values such as -1 to 1 or 0 to 1 depending on the nature of the data in the batch to prevent saturation of the activation functions. While many networks employ normalization as a

pre-processing step, the batch normalization has been found to be particularly useful in ResNets.[103]

1.4.3 Image Classification

Image classification was one of the first areas to see significant improvement with machine learning. The annual ImageNet Large Scale Visual Recognition Challenge (ILSVRC) began in 2010 and is a competition where programmers compete to have the algorithm with the highest accuracy on classifying images into one of 1000 classes.[109] In 2012, a machine learning algorithm employing CNNs came in first in the ILSVRC-2012 competition with an error rate over 10% lower than the second place contestant.[100] Since then, the winning algorithms have all been based on machine learning techniques.

In medical imaging, image classification techniques have been employed for a variety of tasks such as classification of lesions as benign or malignant,[110–113] classification of interstitial lung disease,[114–116] type and staging of brain tumors,[117] and other computer aided diagnosis (CAD) tasks. In 2016, the American Association of Physicists in Medicine (AAPM), National Cancer Institute (NCI), and SPIE hosted the LungX challenge, where participants classified lung nodules as benign or malignant. Three machine learning based techniques had an area under the ROC curve (AUC) better than obtained by guessing. All participants had AUCs between 0.5 and 0.68 with the CNN based model having an AUC of 0.59.[110] In 2017, Shen et al.[113] developed a multi-crop CNN that replaced one of the maximum pooling layers with a multi-scale maximum-pooling layer. This network achieved an AUC of 0.93 for detecting the likelihood of metastasis in a lung nodule.

As mentioned previously, to train a network from scratch there needs to be a large amount of data available for training. One method that has been applied in the medical field is transfer learning. In the process of transfer learning, a pre-trained network such as AlexNet, which is trained in the classification of images such as banana, cat, dog, human, etc., is used with all but the last few layers frozen or very minimally modified. Frozen layers do not have their weights updated and subsequently calculate the features they were trained to detect for the application they were originally optimized for. Minimally modified layers are trained on the new data but with a very small learning rate so as to preserve most of the pre-trained information. The last layer, or last few layers, depending on the case, are retrained using the new input images for the desired task to essentially teach the output layers which of the pre-trained features relate to the new task. Transfer learning can be accomplished with a much smaller training set than the original data, however, the new input data must be manipulated into the same format as the original task. For instance, in order to use AlexNet for medical images as Gao et al.[116] did in classifying interstitial lung disease, the input images first have to be sized to 244 by 244 and artificially generate 3 “color” channels by using different Hounsfield unit (HU) windows for low attenuation, normal CT range, and high attenuation. They were able to achieve an overall accuracy of 87.9% on a patch size similar to the literature (31 by 31 pixels) with 6 classes and 68.6% on their “holistic” patch method using large portions of the CT image. By contrast, Microbiana et al.[114] built their own network from scratch to classify the interstitial lung disease into 7 classes. Their network took as input a 32 by 32 patch of the image and had a similar architecture to AlexNet consisting of 5 convolution layers and an average pooling layer before 3 fully connected layers. They were able to obtain an

overall accuracy with 7 classes of 85.6%, which is comparable to the pre-trained network by Gao et al. mentioned previously.

In this work, image classification will be used to classify the type of tumor/normal tissue interface in an image patch using built from scratch CNNs. The hope is that the CNN may recognize patterns within the CT image that would assist physicians in distinguishing the tumor from seemingly similar non-cancerous tissue, such as atelectasis, and reducing the contouring uncertainty in these regions.

1.4.4 Image Segmentation

Semantic image segmentation is the task of labeling each pixel/voxel in an image as belonging to a particular class or background. The object of semantic images segmentation is to provide the location of the class(es) within the image frame. For instance, the machine learning algorithms in new self-driving car software must identify and locate where objects such as traffic lights, signs, other vehicles, people, etc. are in order to navigate and stop appropriately. Image segmenting is also a natural next step from image classification.

Image segmentation is of particular interest to the medical community. Within the radiation oncology workflow, as described earlier, the physician takes the time to draw out the tumor contours as well as OARs. This is a very time consuming process. Autosegmentation, such as the PET/CT method described earlier by Baardwijk et al.,[87] that are not based on machine learning techniques have been shown to reduce contour delineation uncertainty. With the advent of deep learning, there has been an increased interest in using the technology for medical images segmentation.

The initial image segmentation algorithms arose from the patch wise image classification networks. The classification was performed in a sliding window fashion over the entire image creating a map of classes predicted for each central pixel of each window, as described in the neuronal membrane segmentation paper by Ciresan et al.[118] Since then, a few variations of an extended concept have been developed for image segmentation as well, mainly the V-Net,[119] U-Net,[120] and E-Net.[105] All three networks revolve around the same architecture consisting of two paths: the first path, or the encoder path, is the same basic structure as the image classification network usually modeled off the VGG[102] network; the second path, or decoder path, is the reverse of the encoder path with up sampling instead of down sampling between the convolution layers. The information from the encoder path with the same output size is passed to the decoder path following each up sampling to preserve the original spatial information. The result is a pixel by pixel map of the classes. The V-Net, U-Net, and E-Net vary in small details. For speed, the E-Net does not perform as many convolutions on the decoding path as the encoding path. The U-Net appears to perform a set of convolutions on the smallest down sampled image before beginning the decoding path whereas the V-Net is a true mirror. For image segmentation tasks, a Dice coefficient based cost function has been employed by some to improve accuracy.[119, 121] This cost function measures the amount of overlap between the predicted segmentation and the true label segmentation.

Image segmenting has been applied to a wide variety of anatomical locations including but not limited to: the brain[122–124], prostate[119, 121], knee[125], heart[126], and lung.[127] Novikov et al.[127] modified the U-Net architecture to create three new variations by 1) adding drop out layers, which randomly modifies the weights to prevent overfitting, after

each of the convolution layers, 2) inverting the number of filters used so the largest number of filters was at the top layer, and 3) replacing the pooling layer with non-overlapping convolution layers to segment the lungs, heart, and clavicles on chest radiographs. The drop out layers and inverted number of filters had the best Dice similarity, 83.7% - 97.3% depending on the organ. In another paper by Moeskops et al.,[128] they used a classification network passed over the image to segment tissue on brain MRI, breast MRI and chest CT angiography with the same network. The network showed very little confusion between image classes at less than 0.0005% and the Dice score for the brain tissue was between 80-90%, with the exception of the ventricular cerebrospinal fluid, which has a Dice dissimilarity of about 70%, the breast about 70%, and the cardiac ventricles about 60%.

This work aims to utilize the image segmentation techniques to build an uncertainty model for the tumor boundary. This can be accomplished though expanding the classification network for the different interfaces, or training a new network to predict the presence of an interface and the probability associated with the label to provide a picture of uncertainty. This tool could then be used by physicians to reduce contour delineation uncertainty.

1.5 Overview of Dissertation

1.5.1 Problem Statement and Purpose

Despite the advances in cancer treatment, lung cancer, in particular, still has very low five-year survival rates. This work seeks to investigate the feasibility of building early prediction of lung tumor response and survival, and using machine learning techniques to provide additional information to physicians during tumor delineation about expected

uncertainty. Predicting response prior to or early in the treatment course could inform physician decision making, provide better patient care and improve identification of the best treatment plan for the patient. Measuring change in tumor volume is one method of determining treatment response, but the classification of complete response, partial response, stable disease, or progressive disease is often determined after the completion of treatment. One area of growing interest is radiomics, which seeks to utilize imaging and texture features extracted from routine images combined with clinical information to assist in a variety of clinical applications. Critically, image features that cannot be repeated precisely between acquisitions should not necessarily be relied upon in predictive models. Thus, before use in the clinic, repeatable and robust texture features that are capable of describing the changes due to treatment and differences in tissue must be identified.

A factor potentially contributing to the low survival rate in lung cancer is the uncertainty in tumor delineation. Multi-modality imaging has been used to decrease physician target delineation uncertainty. However, inter- and intra-observer variation remains one of the largest uncertainty factors in radiation treatment, possibly leading to excess irradiation of normal tissue or under treatment of tumors. Another field of growing interest is machine learning and in particular deep learning. These techniques are being investigated for automatic segmentation to reduce contouring uncertainty for organs at risk and tumors, and to distinguish between malignant and benign lesions in computer vision because of their success segmenting and classifying parts of images as different materials that appear the same to a human observer. Reducing the contour delineation uncertainty could lead to better patient outcomes through reliable targeting of the tumor

tissue and sparing of normal tissue structures. In addition, knowing the level of uncertainty to expect in an image could help to more accurately define the tumor and spare additional normal tissue.

This thesis will seek to explore multi-modality imaging texture features in predictive and machine learning techniques for uncertainty modeling. The first aim is to investigate texture features from pretreatment and, where available, images acquired during treatment to determine the texture features' repeatability within the tumor, and test their feasibility along with clinical features in a predictive model for tumor response and survival. The second aim is to explore the feasibility of building a probabilistic model of inter-observer contour delineation uncertainty given patches from images to aid physicians in contour delineation.

1.5.2 Specific Aims

<u>Specific Aim 1:</u> Develop and evaluate robust texture features extracted from multiple modality images for potential use in predictive modeling of non-small cell lung cancer tumor response.

SA 1.1: Assess the robustness of texture features extracted from different imaging modalities. Robustness will be evaluated based on repeatability between scans and, where possible, under different scanning conditions. Image pre-processing techniques will be evaluated for their ability to improve robustness.

SA1.2: Investigate the feasibility of building a predictive model for tumor response utilizing the identified texture features extracted from multi-modality images, clinical factors, and observed changes throughout treatment for a limited number of patients. Models for CT and MR will be investigated, as well as a multi-modality model utilizing the

similarities and independent features among the modalities. The success of the models will be validated and the limitations of the models examined.

Specific Aim 2: Build an uncertainty model utilizing imaging features from single and multi-modality images and tumor characteristics to support physician contour delineation.

SA 2.1 Determine the ability to predict the uncertainty in contour delineation from the tissue interface. This sub-aim will first determine the degree of uncertainty in a tissue-tumor interface, and then establish how well the interface type predicts the level of uncertainty.

SA 2.2: Investigate deep machine learning techniques to distinguish between different tumor/normal tissue interfaces given a patch input of the image. This sub aim will determine the extent to which machine learning techniques are able to learn features from input patches to distinguish different interfaces.

SA 2.3: Investigate feasibility of building a tool using machine learned features to predict the level of uncertainty at a point of interest. If the network to determine the interfaces is successful, this network could be extended to produce a probability map of interface location and uncertainty derived from the interface predictions. If the network is not successful, a new network could be developed to predict the uncertainty directly.

1.5.3 Innovation

Predictive models for patient outcome after cancer treatment have utilized a variety of parameters acquired from different imaging modalities, but few have investigated different parameters across different modalities and different time points in treatment. The focus of many predictive models has often included only one imaging modality with

pretreatment imaging features. This work seeks to explore the possibility of expanding the scope of the features utilized in the predictive models for tumor response to include features acquired during treatment for each imaging modality as image data allows and further explore the feasibility of combining multiple modalities into a single predictive model. MR image features are not as well studied due to the effect different imaging protocol parameters have on the resulting texture features, therefore this work will seek to establish the reliability of the MR derived texture features by evaluating the repeatability of the MR derived texture features and identify a set of non-redundant features that can be used, similar to those identified for CT images and PET, in an effort to develop an MR predictive model. As mentioned earlier, there are several factors affecting the reproducibility of the MR texture features, therefore this work will only focus on establishing the repeatable texture features for the imaging protocols used at the VCU hospital and seek to establish the feasibility of producing an institution protocol specific predictive model for MR. The exploration of the multimodality predictive models will seek to determine the features from CT and MR that when combined give the “best” tumor response predictive power. Furthermore, the incorporation of longitudinal features, data permitting, into models predicting response could identify texture features that change as a result of irradiation. These changes in underlying physiological processes caused by irradiation in the tumor could arise from tumor cell death resulting in changes in cell density as necrotic regions develop and/or are cleared, cell vascularity as tumor size and areas of proliferation change, and reoxygenation as previously hypoxic regions gain access to oxygen.

Recent methods of improving contour delineation have generally focused on assisted or autosegmentation. However, this may be a very challenging approach for lung tumors due to subtle boundaries and the highly variable appearance of tumor tissue between patients. Instead, a novel approach is proposed herein to stratify delineation uncertainty by tumor to normal tissue 'interface' type. Then, the focus of this work is to use this uncertainty estimate to assist the physician with manual contouring, rather than the challenging task of automatically delineating the tumor directly. The developed tool seeks to predict the amount of uncertainty in the contour delineation given the tissue interface or around a point to help guide the physician and inform them of the level of uncertainty to expect. Deep convolution neural networks seem particularly well suited to take an image output and classify the different tumor/normal tissue interfaces in the lung, and by expanding the classification task into an image segmentation for the interface, could be a tool used by physicians to reduce contour delineation uncertainty.

2 Specific Aim 1: Repeatability of Magnetic Resonance Image Derived Texture Features and Use in Predictive Models for Non-Small Cell Lung Cancer Outcome

2.1 Introduction

As mentioned previously in the introduction chapter, there has been an increased interest in early prediction of treatment outcome to allow for changes to the treatment regimen with the goal of providing better patient care. In order to accomplish this, the previously mentioned sub aims were devised:

- SA 1.1: Assess the repeatability of texture features extracted from different imaging modalities.
- SA 1.2: Investigate the feasibility of building a predictive model for tumor response.

This aim seeks to first identify texture features appropriate for predictive models based on the repeatability of the texture features and image processing applied. Then the identified texture features are used to investigate predictive models for tumor response at end of treatment, and overall survival at 12, 18, and 24 months. The feasibility of a multi-modality model is also investigated by combining texture features across CT and MR modalities. The predictive capability of the resulting models is evaluated and

compared. In addition, the developed workflow is also applied to normal tissue ROIs as a control and to identify potential spurious results.

The first part of the study investigated the repeatability of a variety of images including: T1-weighted Volumetric Interpolation Breath-Hold Examination (VIBE), T2-weighted True fast MRI with steady state precession (TRUFISP), DW-MRI, ADC maps, and helical 4D CT scans for 15 patients with NSCLC before and during the course of radiotherapy. See Table 1 for summary of patient characteristics.

While the repeatability portion of the study investigates several imaging protocols and incorporates multiple time points, the predictive modeling focuses on pretreatment CT, VIBE, and TRUFISP images only due to the limited availability of during treatment data. The detailed results of the predictive models as well as descriptions of the methods, developed workflow, and analysis can be found in manuscript provided in Appendix I. The key finding in the repeatability study identified several features from the MRI and CT images that were candidates for use in predictive models due to repeatability and stability. Multiple significant models for overall survival were constructed from single modality MR and CT, as well as multi-modality models in the predictive modeling portion of the study. In addition, normal tissue was investigated as a control to help reduce spurious results, which led to identifying the medoid feature selection process as being more robust with the small number of subjects in our study. It is recommended that the reader return to this chapter following reading the manuscript. The following sections will provide additional details regarding methods and analysis not provided in the manuscript. Chapter 3 will discuss preliminary research used to develop the workflow presented in

the manuscript as well as exploratory research on the delta radiomics data from the different time points.

Table 1: Summary of patient characteristics for texture feature repeatability

Sex	
Male	10
Female	5
Mean Age (Range)	
59.1 (50.0-73.4) years	
Histology	
Squamous cell carcinoma	9
Adenocarcinoma	6
Stage	
IIB	3
IIIA	6
IIIB	4
IV	2
Chemotherapy	
Yes	12
No	3
Mean Dose (Range)	
61.6 (59.4-66) Gy	

2.2 Code Modifications

As mentioned in Appendix I, this work used the open source radiomics toolkit develop by Vallieres et al.[22] and was extended in-house. For this work, only the texture feature calculation portion of the toolkit and associated functions were utilized including the volume preparation code. The original GLCM texture feature calculation code included only 8 texture features: energy, contrast, entropy, homogeneity, correlation, variance, sum average, and dissimilarity. This code was expanded to include the additional proposed features by Haralick[48] and in subsequent papers[49, 50] and, in the case of contrast, rewritten in a vectorized form to decreased calculation time. Appendix II provides the mathematical formulation for all texture features used in this work. In addition, the original code provided different methods for quantizing the gray levels before computing the texture features. A new quantization method was written to use the gray

levels as the bin levels in effect allowing for a “no quantization” option which was not present in the original code.

For all images except the DW-MR and ADC maps, the texture feature extraction and image processing was performed in 3D. The image protocol used at VCU to acquire the DW-MR and ADC maps only consisted of 7 slices, which was not sufficient to cover the entire tumor volume in one image set for some patients. Therefore, repeated 7 slice image sets were obtained to cover the whole tumor volume. For all the DW images and ADC maps, a volume weighted average of the texture features extracted from 2D slices was taken and used as a surrogate for the 3D volume. Additional details regarding this procedure will be discussed in the 3D Surrogate for Diffusion Weighted Images and Apparent Diffusion section. The Vallieres code was again modified to perform a 2D wavelet decomposition.

2.3 Extended workflow

A simplified diagram of the workflow developed for this project is depicted in Figure 1 of the manuscript in Appendix I. Presented, in Figure 11 is a more detailed diagram of the workflow developed. The workflow can be thought of as three overarching steps: repeatability, clustering/feature reduction, and modeling. The repeatability step identifies the repeatable and stable features which are then passed to the clustering step. The clustering takes the repeatable and stable features and clusters them, determines the optimum number of clusters, and then selects a representative feature from each cluster. The representative features are then combined with the response data and model selection is performed in the final step. Specific details on the methodologies employed are given in the Methods section of Appendix I.

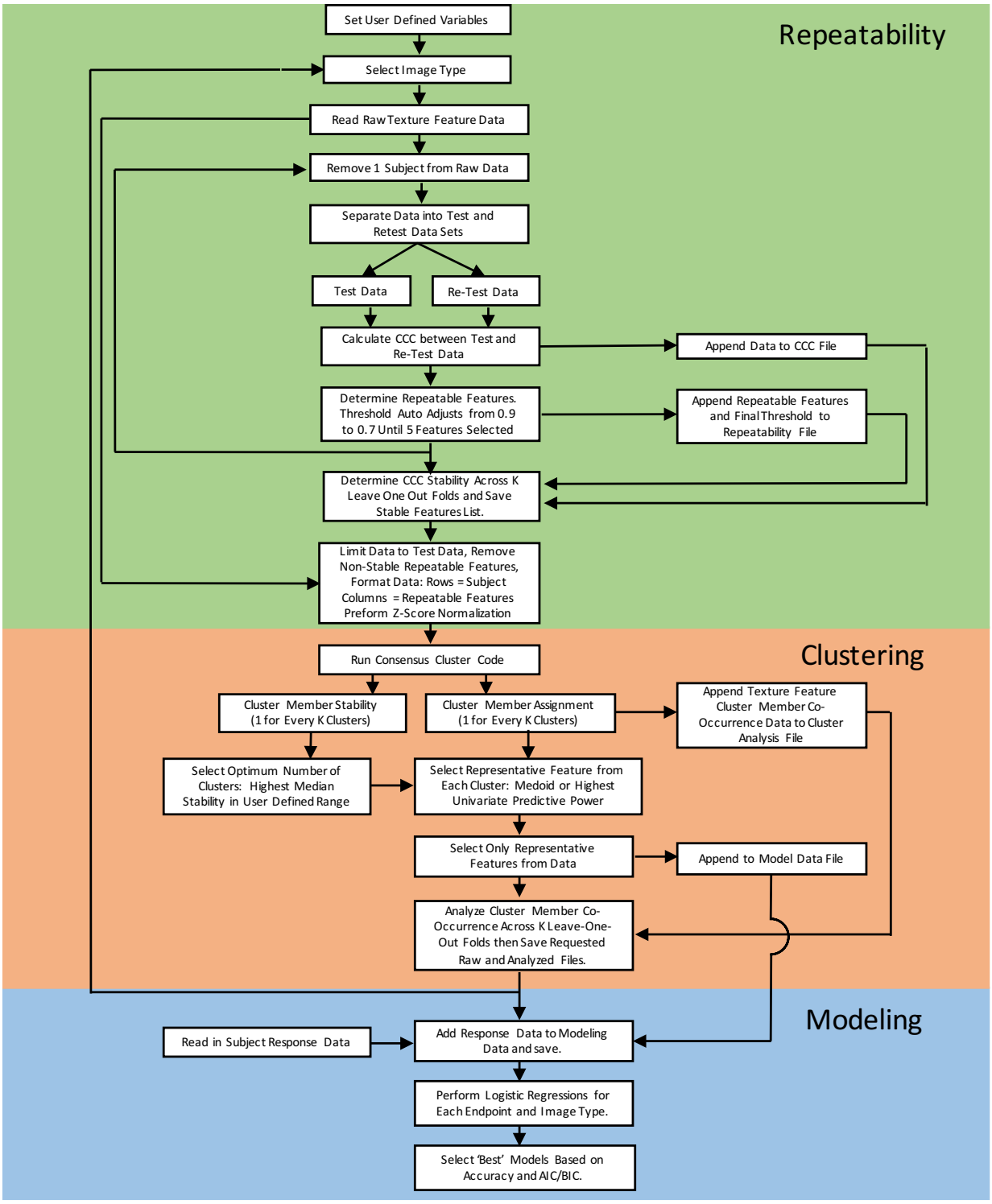


Figure 11: Diagram of workflow developed for radiomic texture feature selection and modeling.

2.4 False Error Rate Control

In radiomics, a large number of texture features and models are often tested as part of the hypothesis,increasing the probability of finding a statistically significant result purely by chance. Often the type 1 error detection rate is set at 0.05 indicating a 5% chance of saying a result is significant when in reality it is not, or more commonly described as rejecting the null hypothesis when it is true. If more than one family of comparisons is being tested for significance, the chance of committing a type 1 error increases. In order to account for the increase in type 1 errors due to multiple comparisons, different familywise error rate control (FWER) and false discovery rate (FDR) procedures have been proposed.

One of the most popular, and most stringent, FWER correction procedures is the Bonferroni correction. The Bonferroni correction seeks to adjust the critical p-value by dividing it by the number of comparisons and using the corrected critical p-value to determine significance.[129] However, the traditional variation of the Bonferroni correction lowers the power to correctly reject one or more false components of the hypotheses. To maintain the power, a sequential Bonferroni correction was proposed by Holm[130] and later popularized by Rice.[131] In the sequential implementation of the Bonferroni correction, the model or univariate p-values from the multiple comparisons are first ordered from smallest to largest and compared to the appropriate level of the corrected critical p-value beginning with the smallest p-value compared to the first level of corrected critical p-value: the desired significance level divided by the number of comparisons. If the smallest p-value is significant when compared to the first level of the corrected critical p-value, the next smallest p-value is compared to the second level of the corrected critical p–value: the

desired significance level divided by the quantity the number of comparisons – 1. If the second smallest p-value is significant, then the third smallest is compared to the third level of the corrected critical p-value: the desired significance level divided by the quantity the number of comparisons – 2 and so forth until the calculated p-value is no longer significant when compared to the appropriate level of the corrected critical p-value.[130, 131]

The Bonferroni correction in the case of a small dataset can be too restrictive as it increases the chance of a type 2 error (accepting the null hypothesis when it is false or saying a variable is not significant when in reality it is significant). In this case, using a FDR controlling procedure may be more appropriate as was used in the manuscript in Appendix I. The FDR procedure adjusts the correction of the p-value using a predetermined acceptable error rate. The Benjamini-Hochberg-Yekutieli (BHY)[132, 133] procedure is similar in nature to the sequential Bonferroni correction procedure except that the significance level is replaced by the acceptable error rate and is multiplied by the factor in equation (3) as opposed to just divided by the number of comparisons.

$$\frac{i}{m * \sum_{j=1}^{m} \frac{1}{j}} \quad (3)$$

where i is the rank of the ordered p-value, and m is the number of comparisons. The corrected p-values resulting from the BHY procedure will always be less than or equal to the Bonferroni corrected p-values when the significance level and acceptable error rate are the same.

2.5 Normal Tissue Determination

The same workflow was applied to unirradiated normal tissue as a control experiment to identify spurious results under the assumption that texture features extracted from unirradiated normal tissue would not have a biological correlation with

treatment outcome. Three different normal tissues were investigated including: air within the main airways, blood within the descending aorta, and contralateral out-of-field muscle in either the erector spinae or the infraspinatus muscles. All three normal tissues where chosen because they would not experience changes due to radiation treatment, and were to a certain degree homogenous and allowed for reproducible ROI definition.

The concordance correlation coefficient (CCC), and other correlation measures, are sensitive to the range of the values present in the population.[134] If the range of values across the sampled population is very narrow, then small deviations from perfect correlation will result in a low CCC score, as seen in Figure 12.[135, 136] For the air and blood samples, the repeatability of texture features was lower than expected for values such as the mean, which numerically only deviated by about 4 HU on a CT scan. This was because the CT value across all patients was very similar as expected but was sensitive to any artifacts induced by the movement of blood in and out of the imaging slice. Similarly, the air contour was sensitive to noise given the small range of values across patients. The muscle contours produced a number of repeatable features in the same range as the tumor and was selected for further investigation as described in Appendix I.

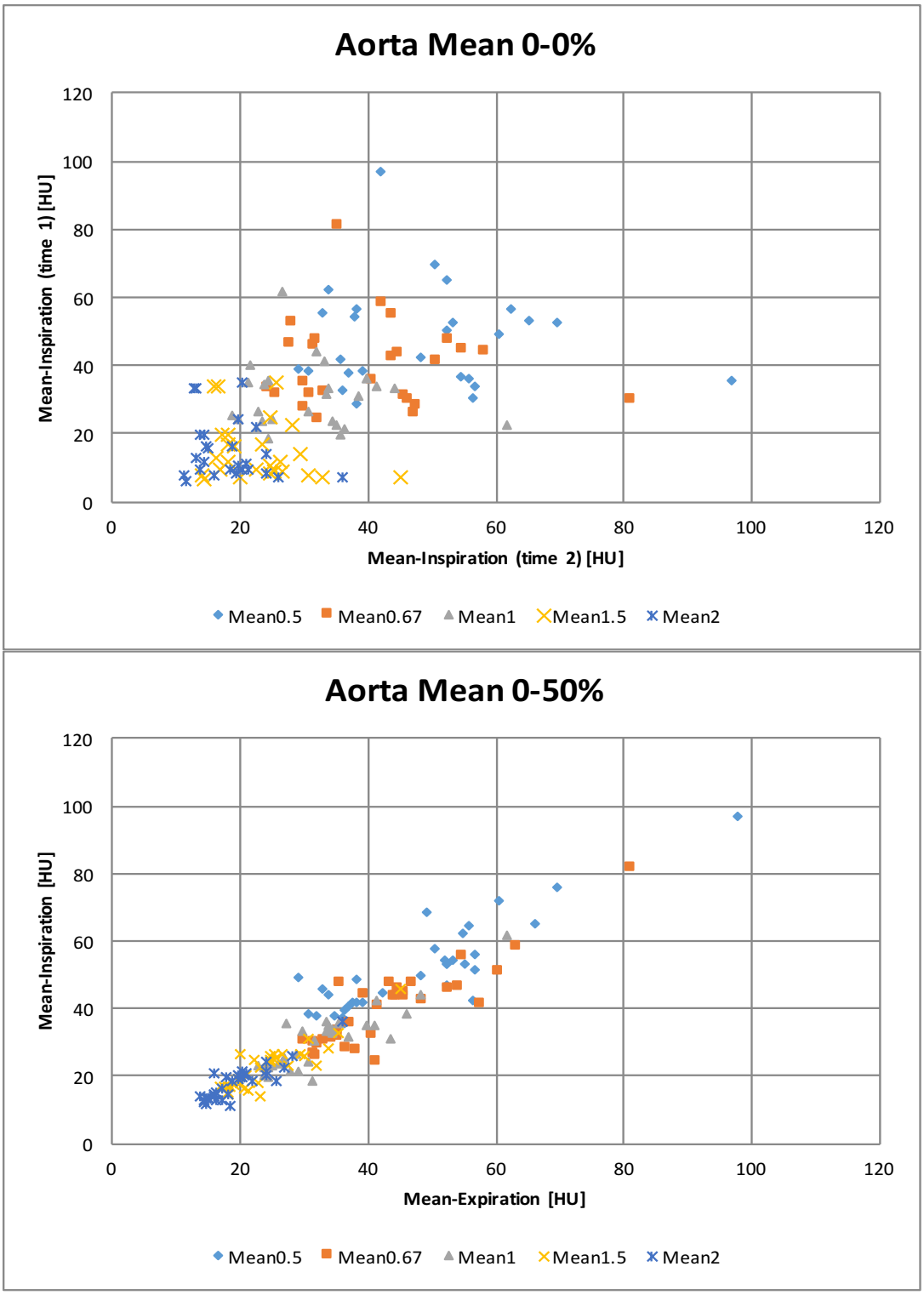


Figure 12: Comparison of the Aorta ROI Mean HU values from the inspiration (0%) phase images from different times (top) and same scan inspiration (0%) and expiration (50%) phase (bottom) CT images demonstrating the difference observed in values for a homogeneous tissue.

2.6 Conclusion

Texture features extracted from MR and CT images were utilized to develop predictive models for overall survival and local control for both single and multi-modality models for a small number of subjects. The workflow developed ensured that repeatable and stable features were used as candidates for clustering and model selection. The medoid representative cluster feature selection method appeared to select more robust texture features than the univariate selection method when compared to the model derived from unirradiated normal tissue. The results are encouraging and further study into MR features for predictive models is warranted.

3 Specific Aim 1: Preliminary and Supplementary Experiments for Radiomics Workflow Development and Predictive Modeling

3.1 Introduction

The work presented in the previous chapter and Appendix I was based on a final workflow developed through testing different methods of image processing and selecting features with desirable characteristics, such as repeatability and stability. The first step in developing the workflow was to determine which preliminary processing steps maximize repeatability of the texture features. The focus of the preliminary workflow development research was on MR images and establishing if any of the texture features noted therein were repeatable under very similar scan conditions. The repeatability of CT and PET features have been studied to a greater extent by others,[11, 20, 34, 40, 68, 71, 73, 137, 138] however, CT repeatability was still evaluated for completeness and comparison for applicable processing steps.

The work presented in this chapter sought to evaluate some of the common image processing techniques and their effect on the repeatability of the texture features extracted from the different image types, such as bias correction for MR images, and wavelet decomposition and quantization of gray level values for both MR and CT. Each

of these steps changes the distribution of the gray level values and the resulting texture feature calculations, thereby affecting the repeatability. By evaluating the different processes' effects on repeatability as measured by the CCC, the image pre-processing steps were determined for the workflow.

Methods of limiting and correcting visual artifacts in the image, such as applying HU thresholds in CT images and bias correction in MR images, aim at improving the qualitative look of an image. Previous work by Hunter et al.,[71] showed that the number of repeatable features decreased with increasing HU thresholds, and for this reason, we did not apply any threshold limits to the CT images. However, research into the repeatability of MR texture features revealed a dependence on the imaging parameters and acquisition methods.[23, 29, 37–39, 74–77] To alleviate some of these dependencies, a single imaging protocol and MR instrument was used to acquire the MR images allowing for an assessment of the repeatability of texture features when image processing techniques, such as bias correction were applied, and will be discussed in the following sections.

Wavelet decomposition and gray level quantization are other processing techniques that have been applied to make texture features more robust or to emphasize different structural contributions within an image. The effect of quantization on texture features has been studied by various groups.[50, 139, 140] Quantization seeks to reduce the sparsity of the GLCM, GLRLM, and GLSZM that can arise when the number of gray level is not quantized. It can be thought of as smoothing the gray levels and reducing the noise in the image. However, by quantizing the images, fine textures can also be lost. Wavelet decomposition can be thought of as a simultaneous high pass and low pass filter decomposing the image into contrast and edge emphasized images.[60] By recombining

the different decomposed images at different levels, edges and other structural information can be strengthened before computing texture features as explored by Vallieres et al.[22]

The remaining sections of this chapter are dedicated to the development of the surrogate 3D texture features derived from 2D slices for the DW-MR and ADC images and exploration of time dependent changes of the texture features through treatment. The imaging protocol for acquiring the DW-MR and ADC maps was limited in range due to recommendation from the manufacturer. For some patients, the entire primary tumor was not covered by the imaging sequence and multiple overlapping sequences were acquired. A method for computing a volume averaged 3D texture feature surrogate was developed from the 2D slices. To investigate the differences in the texture features over the course of treatment, both population and individual changes were explored. A summary of the available images at different time points can be seen in Table 2. The '_Thickness' label denotes a comparison between images with different slice thicknesses, and the '_Order' label denotes a comparison between images with different slice acquisition orders.

Table 2: Summary of available images at different time points.

	Time 1	Time 2	Time 3
CT	15	9	9
TRUFISP	15	10	9
VIBE	16	9	7
DWI_Thickness	4	2	2
DWI_Order	4	6	6
ADC_Thickness	4	2	2
ADC_Order	4	6	6

3.2 Bias Correction

The bias artifact is the result of signal intensity non-uniformity causing a smooth non-uniform change in the signal intensity unrelated to anatomical variation. Bias correction seeks to correct the non-uniformity of the signal in post processing leading to more uniform looking image intensity. An example of the bias artifact and a corrected image can be seen in Figure 13. The VIBE image gradually loses intensity toward the center of the images in the top row. The restored uniform intensity from the bias correction procedure can be seen in the bottom row. There are several potential causes for the signal non-uniformity including: non-uniform B_0 magnetic field, RF coil homogeneity, sensitivity of the surface coils, gradient fields inducing eddy currents, and others.[141]

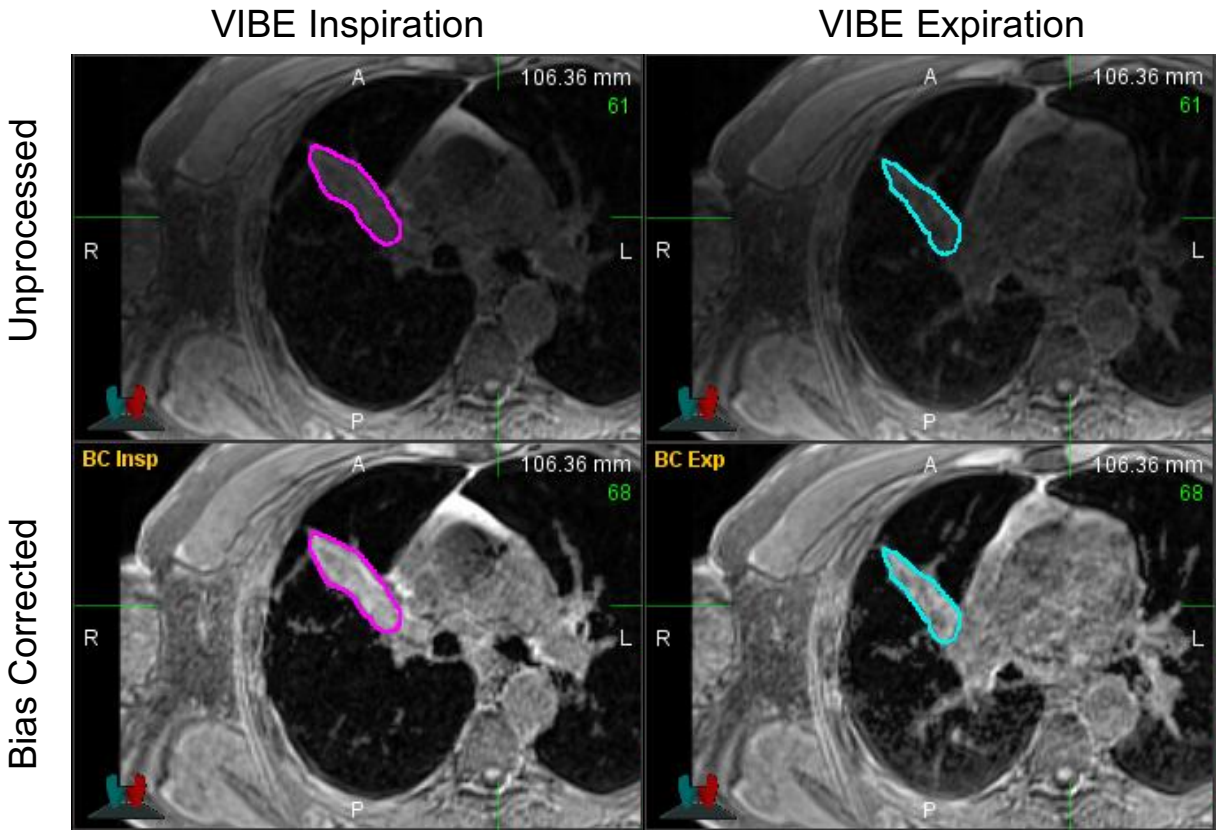


Figure 13: Example of bias artifact (top) and the corresponding bias corrected image (bottom) for one pair of inspiration/expiration VIBE images used in this study.

This study utilized the 32 non-contrast enhanced VIBE inspiration/expiration image pairs and the 34 non-contrast enhanced T2-weighted TRUFISP inspiration/expiration image pairs from all 15 patients and all time points throughout the course of radiation treatment. One experienced radiation oncologist delineated the primary tumor on either the inspiration or expiration image at random per patient to prevent systematic bias. A ridged registration was performed utilizing the MIM version 6.6 (MIM Software Inc., Cleveland, OH) program between the inspiration and expiration images prior to transferring the contour to the corresponding image and adjusting as necessary. Volumes of the final tumors corresponded within ±10% of each other.

Since the MR images exhibited the bias artifact, applying a bias correction would be a reasonable pre-processing step. To evaluate the effect a bias correction pre-processing step has on the repeatability of texture features, a bias correction was applied utilizing the N4 Bias Correction Algorithm in the Advanced Normalization Toolkits (ANTs) developed by Tustison et al.[142, 143] This algorithm uses an iterative approach to estimate the unbiased image, calculate the bias field correction and preform a comparison to the initial image. To ensure a smoothly varying bias field correction, b-splines were used to produce the bias field correction.[142] A mask of the air outside the body and the normal lung tissue, since the lungs are primarily air, was supplied to the algorithm. The procedure for determining the ideal threshold for air within the lung of the TRUFISP and VIBE mask images and bias correction parameters was determined by experimentation on a phantom and approved by a radiation oncologist.

Fifty-nine different texture features, detailed in Table 3, were extracted from the unprocessed and bias corrected images utilizing the modified version of the radiomics code by Vallieres et al.,[22] via the MIM MatLab (MatLab 2016b, MathWorks, Natick, MA) extensions. Details regarding the modifications are described in the 2.2 Code Modifications section of the previous chapter. The mathematical description of each texture feature can be found in Appendix II. Prior to calculating the texture features, the images were isotopically resampled to the axial in-plane pixel width of the image and pixels greater than 3 standard deviations from the mean of the ROI were removed as suggested by Collewet et al.[75] to improve feature stability. Texture features were computed for 5 different wavelet decomposition ratios, $1/2$, $2/3$, 1, $3/2$ and 2, where the ratio of 1 represents the unfiltered image.[22]

The repeatability between the inspiration/expiration images for each image pair was again assessed by the CCC, where texture features with a CCC of greater than 0.9 are considered repeatable.[134, 144] First proposed in 1989, the CCC seeks to characterize the departure of a test and re-test measurement from the 45 degree line through the origin which represents perfect one-to-one correlation.[134] Other methods of agreement exist, such as the Pearson correlation coefficient, paired t-test, and the interclass coefficient, but each has potential draw backs. The Pearson correlation coefficient measures the linear agreement of paired points, but does not characterize deviation from perfect agreement. The paired t-test addresses the mean of all samples but not individual deviations, and the inter class coefficient treats the test and re-test measurements as replicates instead of distinct measurements, though the appropriate ICC has been shown to be equivalent to the CCC.[134, 145, 146]

Table 3: List of the texture features evaluated in this work.

Histogram	GLCM	NGTDM	GLRLM	GLSZM
Variance	Energy	Coarseness	Short Run Emphasis (SRE)	Small Zone Emphasis (SZE)
Skewness	Contrast	Contrast	Long Rum Emphasis (LRE)	Large Zone Emphasis (LZE)
Kurtosis	Entropy	Busyness	Gray Level Non-Uniformity (GLN)	Gray Level Non-Uniformity (GLN)
Standard Deviation (SD)	Homogeneity	Complexity	Run Length Non-Uniformity (RLN)	Zone Size Non-Uniformity (ZSN)
Mean	Correlation	Strength	Run Percentage (RP)	Zone Percentage (ZP)
Minimum	SumAverage		Low Gray Level Run Emphasis (LGRE)	Low Gray Level Zone Emphasis (LGZE)
Maximum	Variance		High Gray Level Run Emphasis (HGRE)	High Gray Level Zone Emphasis (HGZE)
Median	Dissimilarity		Short Run Low Gray Level Emphasis (SRLGE)	Small Zone Low Gray Level Emphasis (SZLGE)
Quartlie1	Mean Pair Sum (MeanPS)		Short Run High Gray Level Emphasis (SRHGE)	Small Zone High Gray Level Emphasis (SZHGE)
Quartile3	Variance Pair Sum (VariancePS)		Long Run Low Gray Level Emphasis (LRLGE)	Large Zone Low Gray Level Emphasis (LZLGE)
	Entropy Pair Sum (EntorpyPS)		Long Run High Gray Level Emphasis (LRHGE)	Large Zone High Gray Level Emphasis (LZHGE)
	Variance Pair Difference (VariancePD)		Gray Level Variance (GLV)	Gray Level Variance (GLV)
	Entropy Pair Difference (EntropyPD)		Run Length variance (RLV)	Zone Size Variance (ZSV)
	Information Correlation Measure 1 (InfoCorr1)			
	Information Correlation Measure 2 (InforCorr2)			
	Auto-Correlation (AutoCorr)			
	Cluster Prominence (ClusterProm)			
	ClusterShade			

Results from the preliminary research demonstrated that there were texture features for the bias corrected (BC) and non-bias corrected images, both VIBE and TRUFISP images, that had a CCC of greater than 0.9 as can be seen in Figure 14 and Figure 15 respectively. The high number of texture features with a CCC greater than 0.9 for the non-bias corrected images was encouraging as it indicated several texture features that could be candidates for further analysis for both MR a multi-modality and MR specific predictive models. The same was not true for the bias corrected images as seen in the VIBE_BC and TRUFISP_BC results in Figure 14 and Figure 15. The number of texture feature that had a CCC of greater than 0.9 was much less than the non-bias corrected images indicating that the bias correction, while improving the visual appearance of the images, did not improve the stability of the texture features. This could be due to the iterative nature of the bias correction algorithms not finding exactly the same solution for both the inspiration and expiration images. Bias correction was not implemented as part of the workflow to increase the number of repeatable features.

VIBE (T1-weighted)					
Wavelet Ratio	0.5	0.67	1	1.5	2
RLN	0.97	0.97	0.97	0.97	0.97
GLNS	0.97	0.96	0.97	0.96	0.97
GLN	0.96	0.96	0.96	0.96	0.97
ZSN	0.97	0.97	0.96	0.96	0.96
LZE	0.96	0.96	0.94	0.95	0.95
Coarseness	0.96	0.96	0.94	0.93	0.93
Entropy	0.93	0.93	0.93	0.93	0.92
Skewness	0.93	0.93	0.93	0.93	0.92
EntropyPD	0.94	0.93	0.93	0.92	0.91
Busyness	0.88	0.91	0.95	0.94	0.94
EntropyPS	0.93	0.92	0.92	0.92	0.92
SumAverage	0.93	0.93	0.93	0.92	0.91
SZE	0.93	0.93	0.92	0.92	0.91
Quartile	0.92	0.92	0.92	0.92	0.92
ContrastN	0.94	0.94	0.93	0.91	0.87
ClusterShade	0.92	0.92	0.92	0.92	0.91
Dissimilarity	0.93	0.93	0.92	0.91	0.90
Energy	0.92	0.92	0.92	0.92	0.91
Median	0.91	0.91	0.92	0.92	0.92
ZP	0.92	0.92	0.92	0.91	0.90
Homogeneity	0.93	0.92	0.92	0.91	0.89
RP	0.92	0.92	0.92	0.91	0.90
LRE	0.92	0.92	0.92	0.91	0.90
SRE	0.92	0.92	0.92	0.91	0.90
Mean	0.91	0.91	0.91	0.91	0.91

VIBE_BC (T1-weighted)					
Wavelet Ratio	0.5	0.67	1	1.5	2
RLN	0.97	0.97	0.97	0.97	0.97
GLNS	0.96	0.96	0.96	0.96	0.96
Coarseness	0.96	0.96	0.95	0.91	0.94
ZSN	0.95	0.94	0.94	0.93	0.93
GLN	0.94	0.94	0.94	0.94	0.94
Skewness	0.93	0.93	0.94	0.94	0.92
Kurtosis	0.87	0.87	0.87	0.87	0.85
Strength	0.87	0.88	0.85	0.83	0.87
ContrastN	0.92	0.90	0.85	0.82	0.81
Busyness	0.87	0.86	0.88	0.86	0.81
InfoCorr	0.84	0.82	0.81	0.83	0.86
ZSV	0.81	0.90	0.84	0.81	0.76
SZLGE	0.89	0.64	0.89	0.86	0.71
SRLGE	0.89	0.64	0.87	0.86	0.71
LGRE	0.89	0.64	0.87	0.86	0.71
LRLGE	0.89	0.64	0.88	0.85	0.69
LGZE	0.89	0.63	0.87	0.84	0.70
LZHGE	0.78	0.80	0.84	0.81	0.68
GLVS	0.68	0.86	0.83	0.70	0.75
SZE	0.79	0.76	0.77	0.74	0.72
ZP	0.78	0.77	0.76	0.74	0.72
SumAverage	0.81	0.81	0.78	0.71	0.63
SRE	0.77	0.76	0.75	0.74	0.73
RP	0.77	0.76	0.75	0.74	0.72
LRE	0.76	0.76	0.75	0.74	0.72

GLCM	GLRLM	GLSZM	Highly Repeatable	Potentially Repeatable
HIST	NGTDM		Repeatable	Not Repeatable

Figure 14: Top 25 repeatable texture features for the VIBE images. Highly repeatable texture features (CCC ≥ 0.95) are green, repeatable features (0.90 ≤ CCC < 0.95) are yellow, potentially repeatable features (0.85 ≤ CCC < 0.90) are orange, and not repeatable feature (CCC < 0.85) are pink.

TRUFISP					
Wavelet Ratio	0.5	0.67	1	1.5	2
Coarseness	0.97	0.97	0.97	0.97	0.97
GLNS	0.96	0.96	0.96	0.97	0.97
RLN	0.96	0.96	0.96	0.96	0.96
GLN	0.96	0.96	0.96	0.96	0.96
ZSN	0.96	0.96	0.96	0.96	0.96
Variance	0.96	0.96	0.96	0.96	0.95
Energy	0.96	0.96	0.96	0.95	0.95
VariancePS	0.96	0.96	0.96	0.96	0.95
SD	0.96	0.96	0.95	0.95	0.94
Complexity	0.96	0.96	0.96	0.94	0.91
ClusterProm	0.95	0.95	0.95	0.94	0.92
Entropy	0.96	0.95	0.95	0.93	0.92
InfoCorr	0.91	0.93	0.94	0.94	0.94
SZHGE	0.94	0.95	0.95	0.94	0.89
EntropyPS	0.95	0.95	0.93	0.92	0.91
SZLGE	0.91	0.92	0.93	0.95	0.95
LGZE	0.91	0.92	0.93	0.94	0.95
Quartile	0.93	0.93	0.93	0.93	0.93
HGZE	0.94	0.94	0.95	0.93	0.89
SZE	0.94	0.94	0.93	0.93	0.91
SRHGE	0.94	0.94	0.94	0.93	0.89
AutoCorr	0.94	0.94	0.94	0.93	0.89
HGRE	0.94	0.94	0.94	0.93	0.89
LZLGE	0.90	0.92	0.93	0.94	0.94
SRLGE	0.90	0.91	0.93	0.94	0.94

TRUFISP_BC					
Wavelet Ratio	0.5	0.67	1	1.5	2
RLN	0.97	0.97	0.97	0.97	0.97
ZSN	0.96	0.96	0.96	0.96	0.96
GLNS	0.95	0.95	0.96	0.96	0.96
Energy	0.94	0.96	0.96	0.95	0.95
GLN	0.95	0.95	0.95	0.95	0.95
Coarseness	0.93	0.95	0.95	0.95	0.95
InfoCorr	0.87	0.91	0.92	0.92	0.93
Entropy	0.92	0.92	0.90	0.89	0.87
GLVS	0.89	0.93	0.87	0.87	0.87
Correlation	0.87	0.88	0.89	0.87	0.86
Busyness	0.84	0.85	0.87	0.88	0.89
EntropyPS	0.89	0.88	0.86	0.85	0.84
SD	0.87	0.87	0.86	0.85	0.83
Minimum	0.85	0.85	0.87	0.85	0.80
SZE	0.83	0.85	0.84	0.84	0.82
VariancePS	0.83	0.83	0.83	0.83	0.81
ZSV	0.71	0.76	0.88	0.88	0.89
ZP	0.83	0.84	0.82	0.81	0.80
Variance	0.83	0.83	0.83	0.82	0.80
Maximum	0.82	0.82	0.81	0.80	0.78
RLV	0.70	0.61	0.89	0.90	0.88
ContrastN	0.82	0.76	0.79	0.79	0.80
GLV	0.62	0.72	0.79	0.88	0.87
Quartile	0.78	0.77	0.75	0.76	0.79
Kurtosis	0.77	0.77	0.76	0.76	0.75

GLCM	GLRLM	GLSZM	Highly Repeatable	Potentially Repeatable
HIST	NGTDM		Repeatable	Not Repeatable

Figure 15: Top 25 repeatable texture features for the TRUFISP images. Highly repeatable texture features (CCC ≥ 0.95) are green, repeatable features (0.90 ≤ CCC < 0.95) are yellow, potentially repeatable features (0.85 ≤ CCC < 0.90) are orange, and not repeatable feature (CCC < 0.85) are pink.

3.3 3D Surrogate for Diffusion Weighted Images and Apparent Diffusion Coefficient Maps

As mentioned earlier in 2.2 Code Modifications, the DW and ADC images were not calculated in 3D. A surrogate 3D texture feature was calculated by taking a volume

weighted average of the texture features extracted from 2D slices and is described in further detail later in this section.

In order to calculate a 3D-like texture feature, a surrogate 3D texture feature was calculated from the 2D slices. Contour delineation was performed in the same manner as the VIBE and TRUFISP contour delineation described in the previous section. Instead of the repeat inspiration/expiration images, the DW images and ADC maps had either: different slice thicknesses, 4 mm or 6 mm, or different slice order acquisition, ascending or interleaved. In the case of the DW images and ADC maps, the portion of the primary tumor present in each image acquisition was delineated and the process was repeated for all image sets necessary to cover the tumor. Overlap between images covering the entire tumor was removed using the MIM software to calculate the intersection of contours in adjacent images and to subtract the intersection from one of the two overlapping contours. The direction of subtracting a superior and inferior portion of the tumor contour was varied randomly to prevent systematic bias. Texture features were then calculated for each 2D slice covering the entire primary tumor contour volume. The final texture feature value for the tumor was calculated by taking the weighted average by volume of the texture features. The different slice thickness images were denoted with "_Thickness" after the image name and the image pairs with different slice acquisition orders were denoted with "_Order" after the image name.

The surrogate 3D texture feature was tested for robustness to missing voxels, which could be introduced during the subtraction of overlapping slices, by evaluating the convergence of the coefficient of variation for the texture feature values extracted from randomly resampled ROIs. Five different resampling levels were tested: 10%, 25%, 50%,

75% and 90%. For each of the resampling levels, a random sampling of the indicated percentage of voxels was used to calculate the texture feature. As the percentage of voxels utilized increased, the coefficient of variation for most of the texture feature value converged to less than 5 percent for both the DW images and ADC maps. An example of the convergence for the mean ROI value can be seen in Figure 16. The low coefficient of variation combined with the convergence of most texture features seemed to indicate the surrogate 3D texture features were appropriate for use in the repeatability study as they are robust to changes in the voxels.

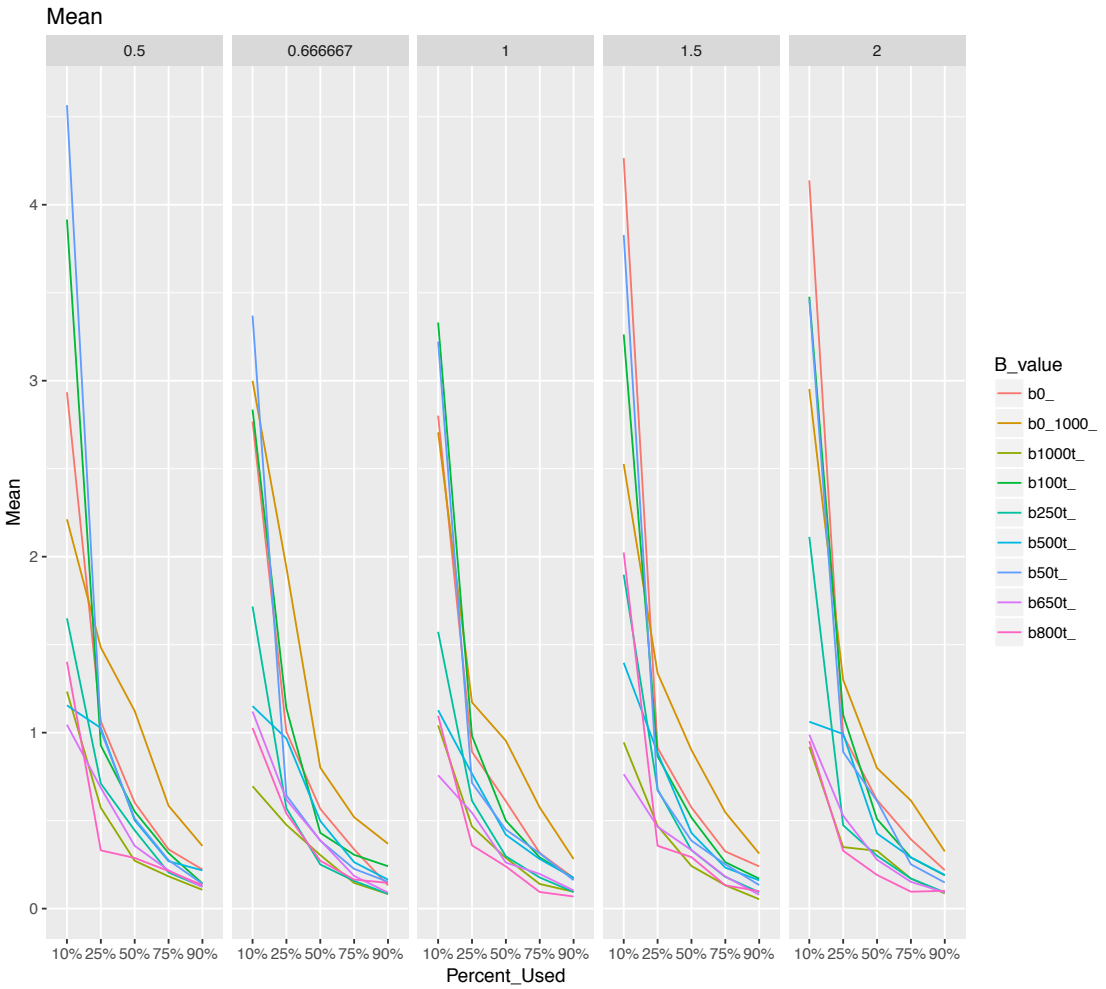


Figure 16: Example of convergence of the coefficient of variation of the 3D surrogate mean texture feature for different b-value DW_Thickness images and the ADC_Thickness map (b0_1000_) as a function of percentage of pixels used for each of the wavelet ratios tested (top row shows wavelet ratios).

The DW image pulse sequence used an echo planar read out where the tissue was first excited and then multiple slices were read as the signal decays causing a slice to slice variation in the signal intensity with lower intensity present in slices acquired later in the sequence. The slice to slice variation can be seen in the alternating bright and dark bands of the interleaved slice acquisition DW image in Figure 17. In order to minimize the impact of the slice variation, the DW and ADC image were normalized prior to texture feature calculation by finding the average minimum and maximum intensity values of the slices and uniformly quantizing the intensity levels between this range.

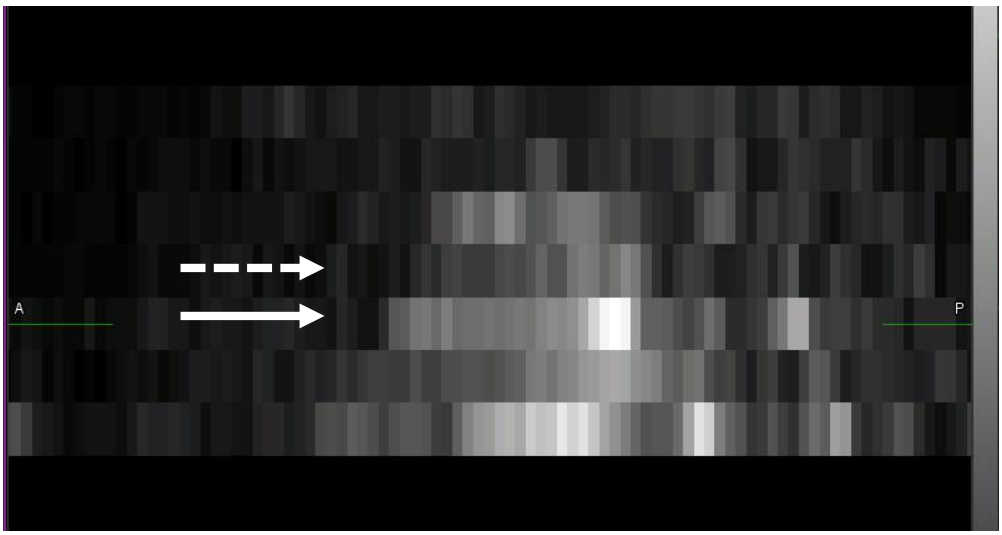


Figure 17: Example of slice intensity variation in sagittal DW 1000 b-value image. This image was acquired with an interleaved slice acquisition pattern creating alternating slices with high (solid arrow) and low (dashed arrow) intensity.

The same repeatability analysis was performed on the DW-MR and ADC 3D surrogate texture features as described in the previous section. Eight different b-value DW images were acquired: 0, 50, 100, 250, 500, 650, 800 and 1000 mm^2/s and all 8 b-value images were used to create the ADC map. The b-value image that had the highest number of repeatable features was the 650 mm^2/s for the DW_Thickness. The DW_Order

image with the most repeatable texture features was a tie between the 50 mm^2/s and $100mm^2/s$, with the 100 mm^2/s having more repeatable features across all wavelet ratios. However, the lower b-value represents perfusion rather than diffusion, so for analysis of the diffusion component was limited to the 650 mm^2/s image which has the third largest number of repeatable features. The results of the repeatability analysis were presented and discussed in the manuscript in Appendix I.

3.4 Quantization and Wavelet Transforms

The Vallieres[22] code used for this work includes support for applying the wavelet transform at different ratios and quantizing the gray levels to a desired number of bins. The wavelet transform ratio selects the weighting of the band-pass sub bands, mixed high and low pass filter on different axes, to the pure high pass and low pass sub-bands, high pass or low pass along all axes. By adjusting the ratio, different structural features or contrast features can be emphasized. The quantization option allows for different binning schema for the gray levels which reduced the sparsity of the texture feature matrices when compared to the gray levels present. While reducing the number of bins may reduce the effect of noise present in the image, it can also smooth over fine texture patterns.

The VIBE, TRUFISP, and CT images were further analyzed for the effect quantization would have on the repeatability of the texture features. The DW and ADC images were not included in this analysis due to the limited amount of data. The texture features were analyzed by quantizing the number of gray level into 8, 16, 32, 64, and 256 bins and comparing them to using the gray levels as the bins, or no quantization. The wavelet ratios for all image types were compared. Aside from the different wavelet ratios

and quantization methods, all aspects of the image processing and texture feature extraction were the same. In September of 2016, Vallieres released an update to the radiomics code that effected the calculation of some of the texture features, predominately the GLRLM and NGTDM. This lead to a reduction in the number of repeatable features when compared to the results in the bias correction section seen in Figure 14 and Figure 15.

Initial experiments with clustering revealed that the different wavelet ratios of a texture feature appeared in the same cluster suggesting they contained redundant information. The wavelet ratios with the most repeatable features for the VIBE, TRUFISP, CT, DW, and ADC images were 0.67 and 1, the unfiltered image. In light of this, the unfiltered image was kept and an average wavelet ratio texture feature was calculated by averaging the five wavelet texture feature values together. The unfiltered and average texture feature values only were considered for clustering.

Six different quantization levels were investigated and compared: 8, 16, 64, 128, 256, and the number of gray levels. For the VIBE, TRUFISP, and CT unfiltered images setting the number of bins equal to the number of gray levels resulted in at least 13 more repeatable texture features than any of the other quantization levels. A similar pattern could be seen in the other wavelet ratios. Since quantifying the gray levels did not improve the repeatability of the texture features, the final workflow used the number of gray levels as the bins.

3.5 Delta Radiomics

Investigation into the changes in texture features over the course of radiation treatment has been a new area of interest in the field of radiomics. The hope with delta

radiomics is to detect features that are changing as a result of radiation treatment and therefore could more reliably predict a response to treatment. Fave et al.[17] recently published results that included delta radiomics features in predictive models for overall survival. While the present work does not have enough data from different time points to create models, the texture features were evaluated for a significant change in the features throughout treatment.

To assess if there was a difference in the population median of texture features between different time points, the Wilcoxon Rank sum test was used. To investigate if a significant change had occurred in an individual patient between the different time points, the confidence interval for the repeatability coefficient as described by Barnhart and Barboriak[69] was calculated and evaluated under the hypothesis that a change occurred. All combinations of time point comparisons, time 1 to time 2, time 2 to time 3, and time 1 to time 3, were analyzed for both the population and individuals.

The repeatability coefficient to detect individual change at the 95% confidence level with 2 repeat images per subject is defined as:

$$\widehat{RC} = 1.96\sqrt{2 * wSD^2} = 2.77 * wSD \quad (4)$$

Where wSD, the within subject standard deviation, is the standard deviation of the difference between the test and retest values of the texture feature per subject divided by $\sqrt{2}$ and summed over all subjects. The confidence interval for the different time points becomes:

$$(L, U) = \left(Y_a - Y_b - \widehat{RC}, Y_a - Y_b + \widehat{RC}\right) \quad (5)$$

where a and b are two different time points, L is the lower bound, U is the upper bound, and Y is the feature value of interest at the identified time point. If the confidence interval includes 0, then no significant change for an individual has occurred.

The Wilcoxon Rank test showed there were only a few texture features that had population differences that were significant at the 0.05 level. The significant features were for the CT images: median from the intensity histogram between time points 1 and 2 for 0.5 wavelet ratio (p=0.04) and homogeneity from the GLCM between time points 1 and 2 for the 1, 1.5, and 2 wavelet ratios (p=0.04, p=0.03, and p=0.04 respectively); and for the VIBE images: correlation from the GLCM between time points 1 and 3 for 1.5 and 2 wavelet ratios (p=0.045 and p=0.03 respectively). There were no significant differences found for the TRUFISP, ADC_Order, ADC_Thickness, DWI_Order, or DWI_Thickness images.

Delta radiomics for the ADC and DWI images sets were not calculated due to the small number of individuals with repeat imaging at different time points. Most of the repeatable texture features showed a statistically significant change in at least 2 subjects at one or more time-point comparisons for the VIBE, TRUFISP, and CT images. For the unfiltered image and wavelet ratio 0.67, there were 7 texture features for the VIBE, 8 texture features for the TRUFISP, and 0 for the CT images that did not exhibit a significant change in at least 2 individuals. Due to the same sample size of subjects containing all three time points, delta radiomics were not included in the creation of predictive models.

3.6 Conclusion

The preliminary work helped to determine the image preprocessing steps that maximized the number of repeatable texture features for the CT and MR images. The preprocessing steps maximized the number of repeatable texture features available for clustering and model selection. For the MR images, bias correction did not increase the repeatability of texture features. In light of this, unprocessed images were used for all the MR and CT images. Additional research provided insight into the effects of quantization and the ratios of the wavelets. The number of gray levels was selected to again maximize the number of repeatable features. The unfiltered wavelet ratio was tied for the most number of repeatable features and, since the features of all wavelets clustered together, the unfiltered wavelets and the average of all the wavelet ratio texture features were used for clustering. The preprocessing steps developed the initial steps of the workflow used in the manuscript in Appendix I.

With the limited amount of repeat time data available for repeatability analysis on the DW, ADC, and the longitudinal data, a preliminary analysis was completed. A surrogate 3D texture feature was developed to test the repeatability of the DW and ADC features. The surrogate texture features were robust to changes in the contour due to the subtraction of overlapping images. The longitudinal data, while not showing a significant change in the population data, revealed that many features appeared to demonstrate change on an individual level. With increased data, a population change may be evident. Further investigation of MR texture feature and delta radiomics is warranted.

4 Specific Aim 2: Build an Uncertainty Model Utilizing Machine Learning Techniques

4.1 Introduction

This second aim seeks to build an uncertainty model utilizing imaging features and tumor characteristics to support physician contour delineation. As part of this second aim three sub aims were proposed:

- SA 2.1 Determine the ability to predict the uncertainty in contour delineation from the tissue interface.
- SA 2.2: Investigate deep machine learning techniques to distinguish between different tumor/normal tissue interfaces given a patch input of the image.
- SA 2.3: Investigate feasibility of building a tool using machine learned features to predict the level of uncertainty at a point of interest.

Machine learning techniques have recently been employed successfully in medical imaging for various tasks ranging from detecting lung nodules in images,[113, 147] classifying radiation induced lung injury,[114–116, 148] to image segmentation,[105, 119, 121, 122, 125, 128, 149–151] making machine learning an ideal tool to attempt classification of different tumor/normal tissue interfaces in lung tumors. The expectation is that the CNN would be able to learn parameters that enable it to identify the boundary of the tumor with pathology and anatomy that are difficult for humans to distinguish from tumor, such as atelectasis. Using

the output of the neural networks, more information could be provided to physicians to aid in contour delineation with increased accuracy.

Our research group has previously investigated the amount of contour delineation uncertainty at the tumor and lymph node interfaces with normal tissues. In the study by Karki et al.,[83] seven observers contoured the primary tumor and affected lymph nodes for ten subjects on three image sets: MR, CT only, and PET/CT. A median contour was calculated from the individual contours, and the bilinear distance was computed from each point on the median contour to each individual contour for each modality. The contour delineation uncertainty was estimated for each contour on each imaging modality by taking the root mean square (RMS) of the standard deviation of the bilinear distances. An experienced physician identified the regions on the median contour corresponding to the interface of the tumor, or affected lymph nodes, with the chest wall, lung parenchyma, hilum, mediastinum, vessels, and atelectasis. The uncertainty for each interface type was determined utilizing bilinear distance as before, but only using the points identified as belonging to each interface. The largest amount of uncertainty among the interfaces for all three modalities was between the primary tumor and atelectasis ($p = 0.0006$), while the interfaces with the least uncertainty were the tumor/lung for CT only and the tumor/mediastinum for PET and MRI. A study by Steenbakkers et al.[86] compared the difference in multi-physician contour agreement and delineation uncertainty, as measured by minimum distance between individual and median contour, using CT only and PET/CT images. They found a reduction in the uncertainty and increase in agreement when using the PET images in addition to the CT images. The anatomical area with the largest

improvement was near atelectasis, though improvement was seen in all interfaces examined: lung, mediastinum, chest wall, and lymph nodes.

The amount of uncertainty in contour delineation appeared to be explained in part by the interface type with the tumor. The first part of this aim endeavors to explore the correlation between interface type and contour delineation uncertainty and investigate the feasibility of using machine learning techniques to identify the interface type without physician input in order to provide additional information about the expected contour delineation uncertainty as an aid in improving contour delineation accuracy.

4.1.1 Interface Uncertainty

Using the bilinear distance gathered by Karki et al. for the primary tumor, further analysis of the uncertainty of each interface was investigated focusing on the CT only and PET/CT images. The characteristics of the patients can be seen in Table 4. The median uncertainty was analyzed using R (3.3.1) in R Studio (1.0.143, RStudios Inc., Boston, MA). Analysis was preformed across all subjects and found the uncertainty was largest for the tumor/atelectasis interface for both the CT only and PET/CT images, and lowest for the tumor/vessel interface as seen in Figure 18 and Figure 19 respectively. However, the trend for each patient varied with atelectasis, when present, generally exhibiting the highest median uncertainty while the other interfaces were more evenly spread as can be seen in Appendix III and Appendix IV.

Table 4: Summary of Characteristics used in the study by Karki et al. and for the first portion of this aim.

Sex	
Male	7
Female	3
Mean Age (Range)	
57.5 (50.0-64.6) years	
Histology	
Squamous cell carcinoma	6
Adenocarcinoma	3
Carcinoma	1
Stage	
IIB	1
IIIA	3
IIIB	6
Chemotherapy	
Yes	7
No	3
Mean Dose (Range)	
63 (45-66) Gy	

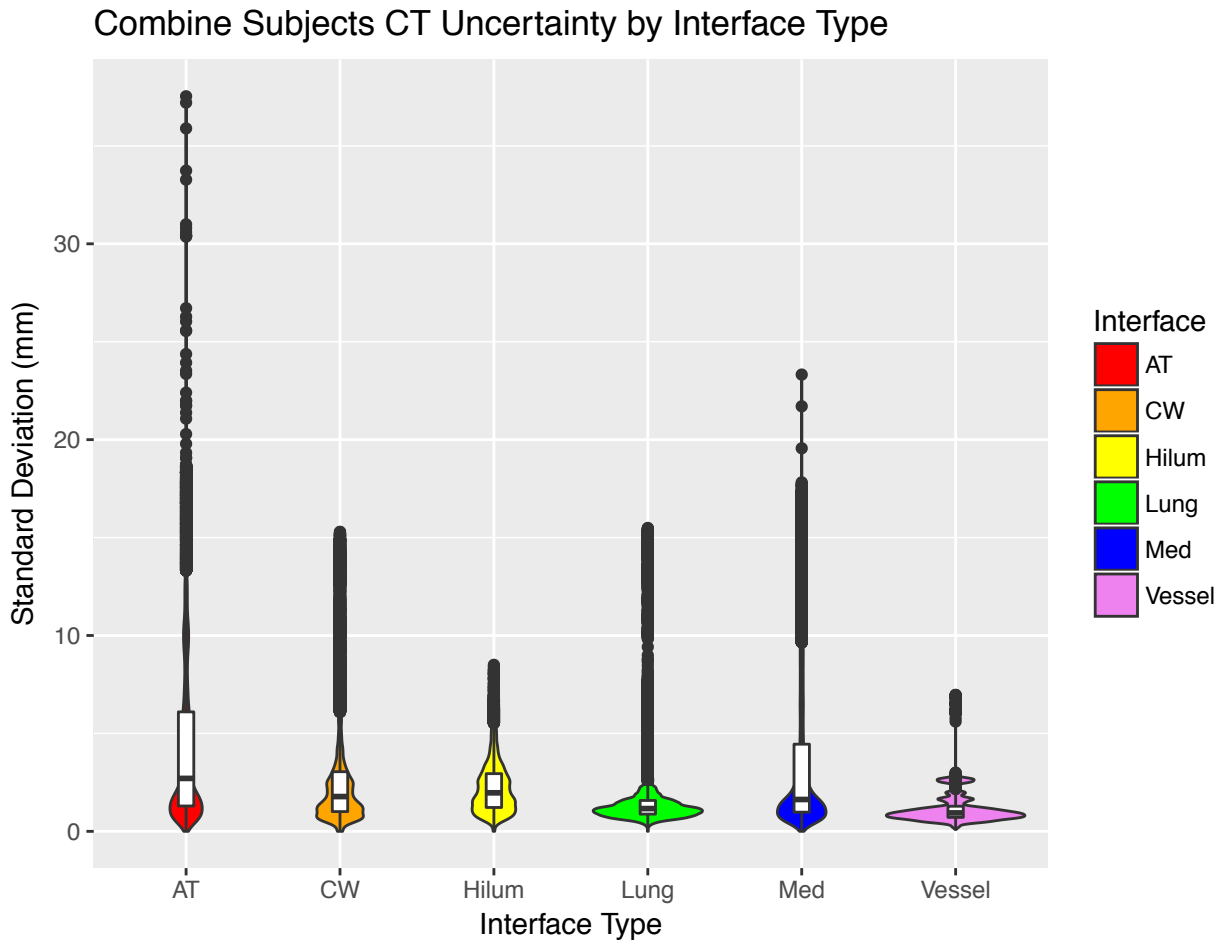


Figure 18: Violin plots of the CT only uncertainty by interface type of all subjects where the width of the plot represents the probability density function (PDF) of values with the indicated standard deviation. Along the center line is a box plot showing the first quartile, median, and third quartile along with the extreme minimum and maximum values within 1.5 times the inner quartile range. Values more extreme are indicated by dots along the whiskers. AT indicates atelectasis interface, CW indicates the chest wall interface, and Med indicates the mediastinum interface.

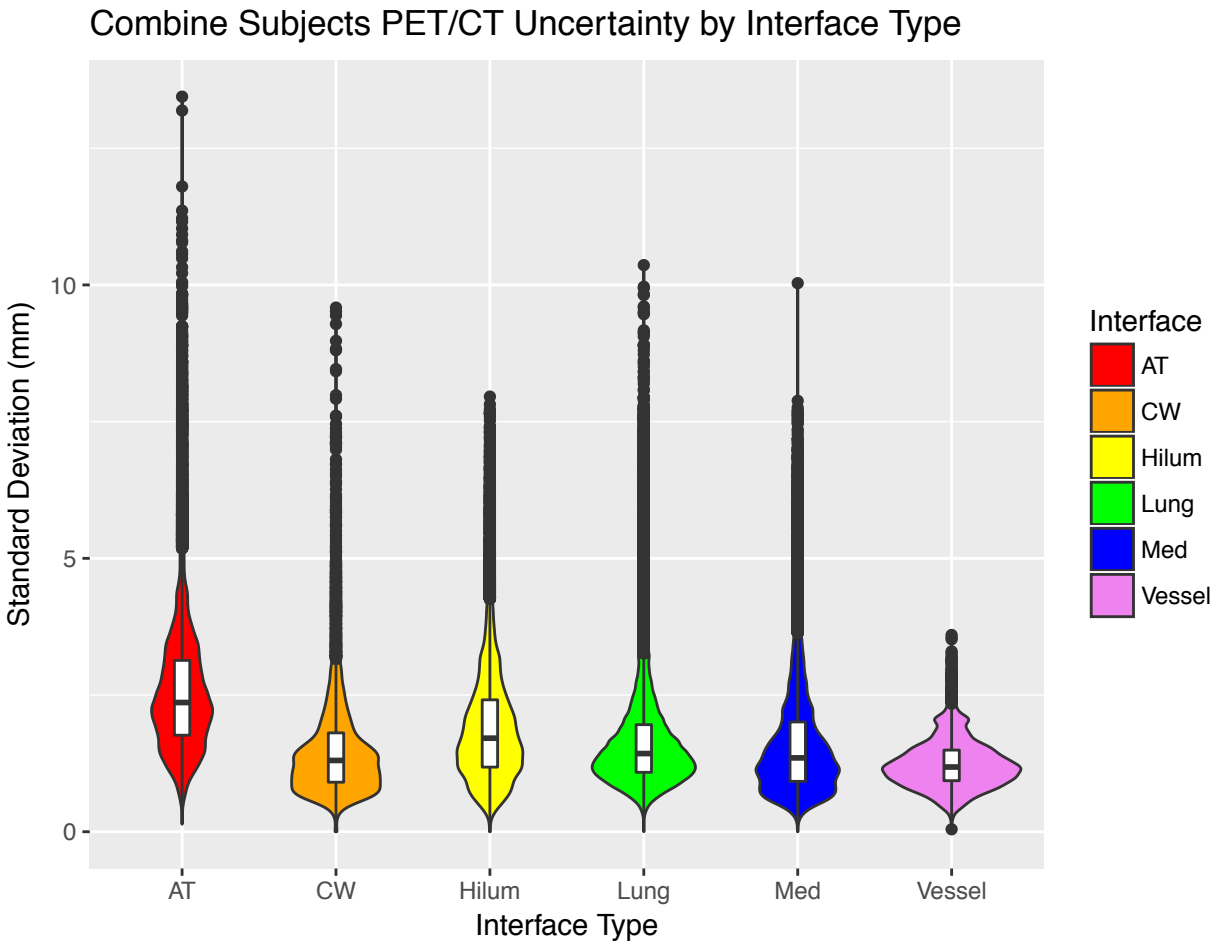


Figure 19: Violin plots of the PET/CT uncertainty by interface type of all subjects where the width of the plot represents the probability density function (PDF) of values with the indicated standard deviation. Along the center line is a box plot showing the first quartile, median, and third quartile along with the extreme minimum and maximum values within 1.5 times the inner quartile range. Values more extreme are indicated by dots along the whiskers. AT indicates atelectasis interface, CW indicates the chest wall interface, and Med indicates the mediastinum.

Statistical analysis of the relationship between uncertainty and interface type revealed, for the PET/CT images, no significance for any of the interfaces. The atelectasis/tumor interface was trending towards significance with a p-value of 0.0752, but all other interfaces had p-values between 0.1073 and 0.7256. K-means clustering of the uncertainty and spatial information of each point on the median contour for a given subject was attempted in Python (2.7) using the Sklearn package from scikit[152] to try and

recover the interface boundaries. The results of the k-means clustering were dominated by spatial information and failed to recover the interface boundaries. When the spatial information was removed, the uncertainty also did not cluster by interface. A comparison of k-means clustering can be seen in Figure 20. Qualitative inspection of the uncertainty showed it was largest in protruding regions of the tumor but within an interface the uncertainty did not have observable trends of homogeneity or increased uncertainty near boundaries with other interfaces.

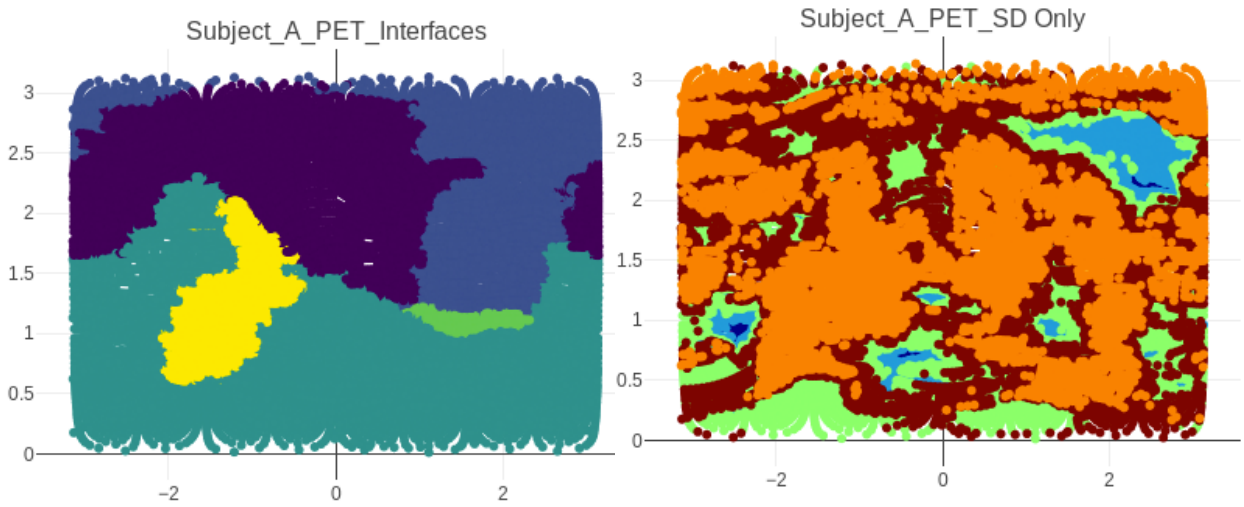


Figure 20: K-means clustering for an example subject with the PET/CT 3D contour surface unrolled to be displayed in 2D. The figure on the left shows the ground truth interface while the panel on the right shows the k-means clustering with only the uncertainty. Color was used to differentiate different clusters

While the interface type alone does not explain the amount of contour delineation uncertainty observed, knowing the amount of expected uncertainty in a given region of the image can provide additional information to the physician while contouring, allowing them to take the uncertainty into account while defining the region to be treated. A comparison of the RMS uncertainty for the PET/CT delineated tumor overall and by interface can be seen in Figure 21. The relationship between the interface type and the

uncertainty level was explored and a convolutional neural network (CNN) was investigated for its ability to determine the interface type without aid from a physician as part of this work. A tool to provide physicians with additional information about the uncertainty was proposed as the final product.

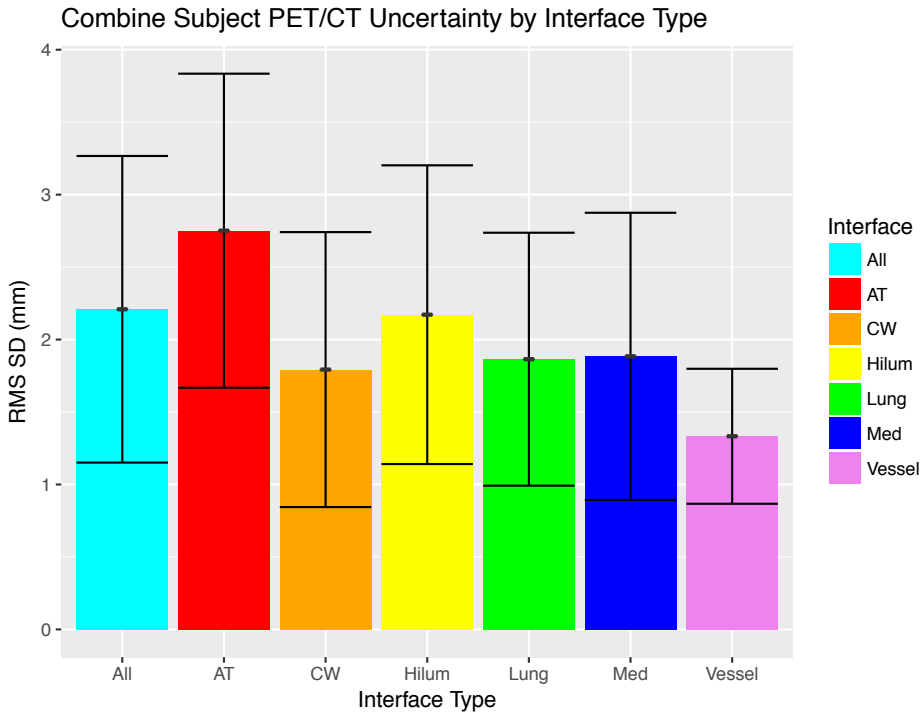


Figure 21: Overall uncertainty RMS of the standard deviation of the bilinear distance for the primary tumor (PT) interfaces PET/CT imaging modality.

4.2 Interface Type Identification Convolutional Neural Network

The next phase of this specific aim endeavored to build a neural network to predict the interface type without user input. Following the results of the initial network, additional network designs, techniques, and subjects were evaluated. The following sections begin by describing the initial network design and dataset curation, and are followed by a description and results of the additional network designs implemented.

4.2.1 Methods

The initial dataset for the first CNN used the same patients characterized in Table 4. The training dataset was created using the interface contours derived from the median contours used in the study by Karki et al.[83] All contours were drawn with MIM (MIM Maestro v6.X, Cleveland, OH) and extracted using the MatLab extensions (2016b, MathWorks, Natick, MA) before being processed in Python (2.7) for use in the developed CNNs. An example of the interface contours can be seen in Figure 22. First, each image and corresponding contours were resampled to have an isotropic voxel size of 0.5mm^3. Then the bounding box containing all the interface contours for a subject was identified and a 4mm margin in each direction was added. From the bounding box, patches of 12.5 mm^3, or 25 voxels^3, were extracted along with the interface type label or a label indicating not an interface for the central voxel. This resulted in several thousand labeled patches across all subjects.

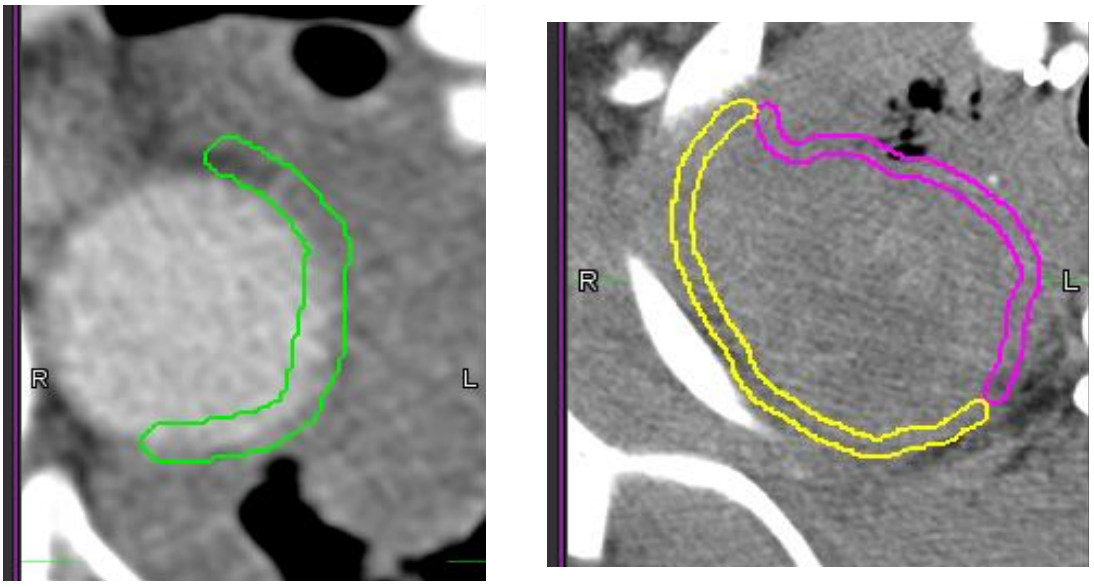


Figure 22: Examples of the aorta/tumor interface in green, the chest wall/tumor interface in yellow and atelectasis/tumor interface in pink.

As the number of patches extracted from each patient and interface type depended on the size and presence of an interface, the dataset was highly unbalanced. To account for this unbalance in the data, a two stage training inspired by Havaei et al.[123] was employed whereby the network was first trained on a balanced number of examples from each class randomly sampled across all subjects. The second phase of training initialized the network weights using the weights learned in the first phase of training and continued to train the network on an expanded training set representing a “natural probability” where the number of examples from each training class is proportional to the frequency observed in the subjects. The images are normalized across all training examples before training to have a mean of 0 and standard deviation of 1. Examples of the center slice of the normalized patches can be seen in Figure 23. In addition, 20% of the training examples were withheld for an unseen testing validation set resulting in approximately 13,000 examples for phase one training, 72,000 for phase two training, and 38,000 for testing.

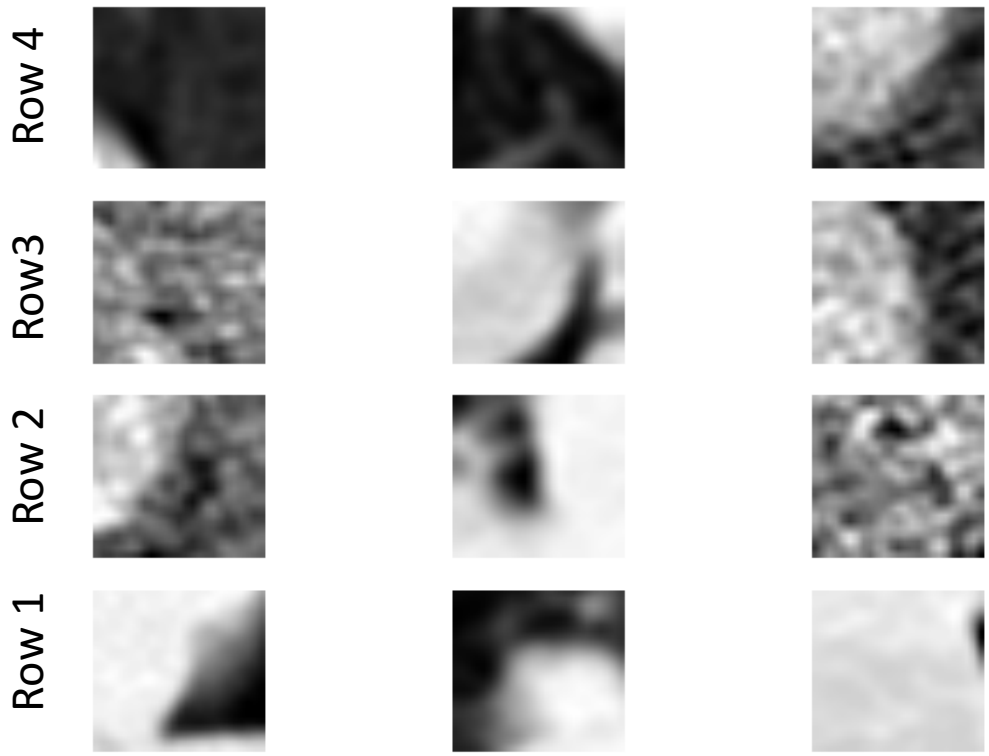


Figure 23: Example of normalized center slice of the input patch. The label is for the central pixel in each input example are as follows: Row 1: aorta, lung, not interface; Row 2: atelectasis, hilum, not interface; Row 3: not interface, mediastinum, lung; Row 4: chest wall, hilum, atelectasis.

The network employed for both phases of training was a 3D CNN with a structure similar to the VGG net.[102] This small network consisted of two blocks of two convolutional layers followed by a maximum pooling layer, followed by two "fully connected" layers implemented as convolutional layers with a drop out layer in between. A detailed explanation of each layer type can be found in section 1.4.2 Convolutional Neural Networks. Unlike the VGG net, the filter size employed in this work decreased in size as the layers got deeper similar to the InvertedNet used by Novikov et al.[127] without skip connections. A diagram of the network architecture can be seen in Figure 24. Valid padding was used in all layers. Training was conducted using the Keras package[153] (version 2.1.6) with a Tensorflow (version 1.6) backend an M2050 Maxwell Nvidia GPU

(Nvidia, Santa Clara, CA) with an SGD optimizer with momentum for regularization using a categorical cross-entropy loss function. The final class prediction was performed by the Softmax activation.

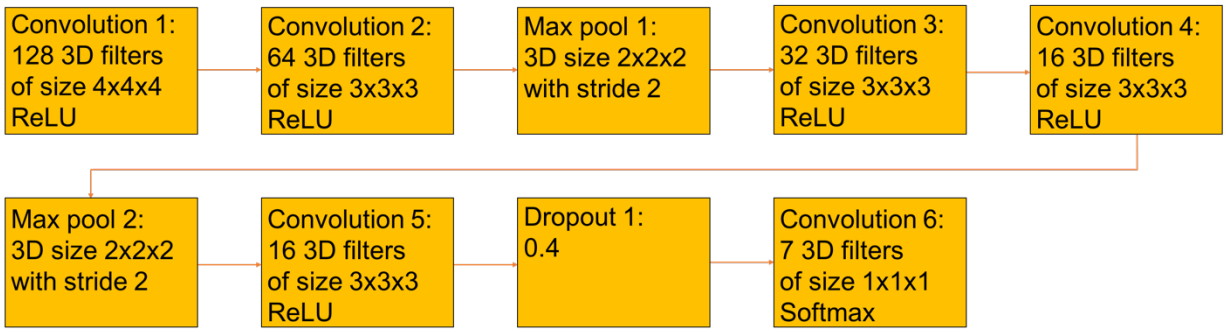


Figure 24: Network architecture for the initial CNN network consisting of two blocks of two convolution layers followed by maximum pooling followed by two fully connected layers implemented as convolutional layers and dropout.

The network was evaluated for accuracy by computing the raw accuracy, number of correctly labeled test examples divided by the total number of test examples, and by analyzing the confusion matrix across all classes. The accuracy for each class was assessed by calculating:

$$A_c = \frac{TP}{TP+FP_c+FN_c} \quad (6)$$

where c is the class being evaluated, TP is the true positive, sum of all correctly identified class, FP_c is the false positives, sum of all examples incorrectly labeled as class c, and FN_c is the false negatives, the sum of all examples of class c labeled as the incorrect class. The error rates, $\varepsilon = 1 - A_c$, are reported in the results section below.

4.2.2 Results

Phase one training achieved accuracy on the unseen test set of 65% while the second phase of training improved the network accuracy to 70%. The top two accuracy of the network following second phase training was 85%. The confusion matrix results

following the second phase of training can be seen in Table 5. The class with the highest error was the atelectasis (AT)/tumor with 18.7% accuracy followed closely by the not interface label with 16.8%. The remaining interfaces had error rates of 13.4% hilum/tumor, 10.9% lung/tumor, 8.7% mediastinum (Med)/tumor, 3.8% chest wall (CW)/tumor, and 0.6% vessel(Aorta)/tumor.

Table 5: Confusion matrix following second phase of training for the initial CNN.

		Predicted Label							
		Not Interface	Lung	Hilum	AT	Med	CW	Aorta	Total Labels
True Label	Not Interface	2721	1561	808	1971	518	210	55	7844
	Lung	180	7231	329	135	72	0	3	7950
	Hilum	28	357	4847	1334	379	0	11	6956
	AT	62	486	515	4613	658	272	7	6613
	Med	50	165	382	287	2998	18	0	3900
	CW	7	1	22	505	26	3687	0	4248
	Aorta	9	21	41	10	10	0	973	1064
	Total Prediction	3057	9822	6944	8855	4661	4187	1049	

4.2.3 Discussion

The result of this initial CNN showed it may be feasible to train a neural network to identify the interface type without physician input for a local patch. The lower overall accuracy of the network suggested that improvements could be made by creating a deeper network and adding additional patients to the training dataset. The need for additional training examples can be seen in the overfitting of the vessel(aorta)/tumor class. In the dataset, only one subject had the vessel/tumor interface, meaning the unseen test examples were all from the same subject seen in the training, but different views, leading to the low error rate. The testing set was comprised of a randomized selection of patches from each class from all subjects exposing the testing set to same type of bias seen by the vessel/tumor interface but to a lesser degree given the variety of

subjects and views to be randomly selected from. To reduce this bias and give a better real world accuracy, the test set should be comprised of completely unseen subject(s).

In addition, the low class accuracy for the atelectasis and not interface suggested the network was struggling to identify these patches. Part of the confusion for the atelectasis could come from the similarity of the atelectasis to the normal tissue on CT scans, which makes visually distinguishing the tumor difficult. The mediastinum and hilum are both similar hybrid structures in the center of the thoracic cavity that contain multiple types of tissue, such as airways and/or blood vessels. To reduce the confusion between these two structures, they could be combined into one class, as attempting to differentiate between mediastinum and hilum structures during contouring resulted in an anatomically often ambiguous separation. The errors from the not interface class could be induced by the wide variety of possible patches included in the not interface class including normal tissue and tumor only tissue alike.

Different pathways were investigated to implement the larger dataset and network design improvements to see if a deeper network is able to improve the accuracy by designing a network that could be used to predict the presence of an interface in general, and/or by designing a network to predict the presence of the tumor.

4.3 Further Exploration of Network Predictions

Following the success of the initial network, the number of patients was expanded to include a total of 39 patients. Since the expanded patient set did not have median contours like the previous study, the tumor contour from the planning CT was used to create the interface contours for all patients. Like the dataset before, the planning contours were created using the PET information as well as the CT scan. The specificity

of the interface labels was also changed by expanding to include a label for airways/tumor and bone/tumor in addition to lung/tumor, atelectasis/tumor, mediastinum/tumor, chest wall/tumor, and vessel/tumor. Also, the previous hilum/tumor was combined with the mediastinum, and bone was removed from the chest wall contours. The updated patient characteristics can be seen in Table 6. These patients were utilized for the reminder of the experiments in this aim. In addition, training was conducted on either the K80 Kepler or P100 Pascal Nvidia GPUs.

Table 6: Summary for patient characteristic for machine learning study.

Sex	
Male	24
Female	15
Mean Age (Range)	
60.6 (50.0-74.6) years	
Stage	
I	1
IIA	1
IIB	4
IIIA	16
IIIB	10
IV	7
Chemotherapy	
Yes	32
No	7
Mean Dose (Range)	
62.2 (40.0-70.2) Gy	

4.3.1 Interface Prediction Networks

4.3.1.1 Expanded Traditional CNN

Building upon the architecture used previously, the network was made deeper by adjusting the padding and adding an additional block of two convolutional layers followed by a maximum pooling layer. This expanded network architecture can be seen in Figure

25. The two phase training was again employed, at first to balance the dataset, but early testing showed no improvement in the accuracy of the network with the second phase of training, so only phase one with balanced interface types across all subjects was used to train the revised CNN architectures. The same method was employed to create the labeled dataset by extracting patches from each subject. During training, one subject was withheld to be the unseen test set while a class balanced dataset for training was randomly selected from the remaining subjects. This network was again trained using SGD with a momentum term and categorical cross-entropy as the loss function.

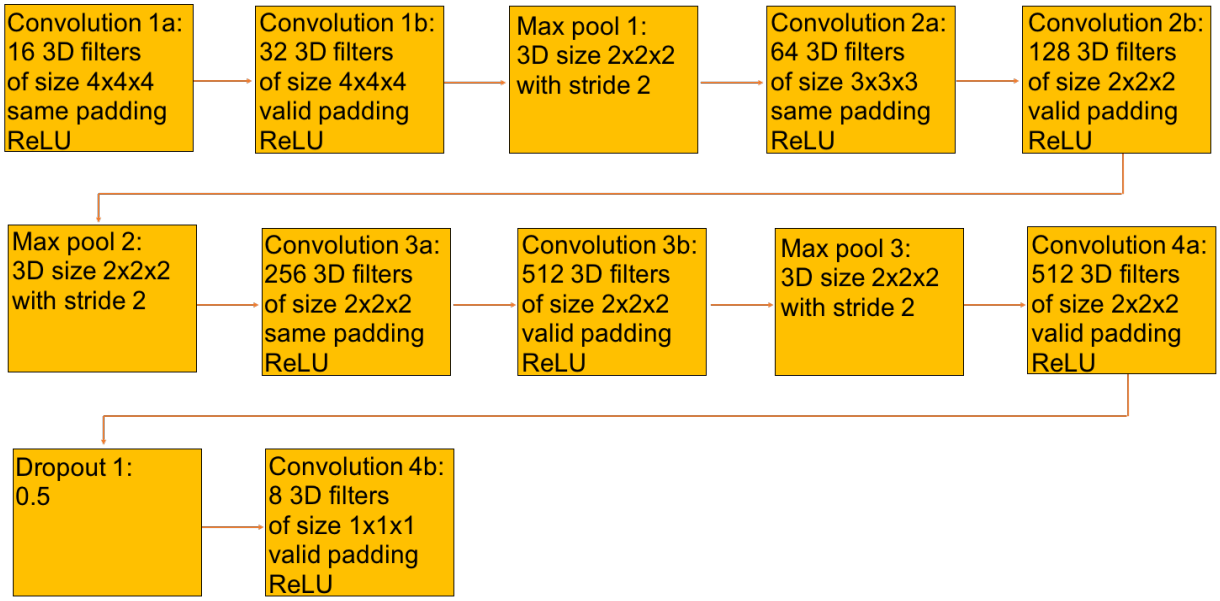


Figure 25: Diagram of the revised CNN architecture to include an addition block of two convolutional layers followed by a maximum pooling layer.

The expanded network was first tested using an unseen test set following training that consisted of a reserved subset of the training set derived from all subjects to assess the improvement in network accuracy due to the additional subjects. The overall accuracy of the network rose to 78% for best guess. The network was then retrained and optimized this time by withholding one subject entirely for the test set and training the network on the remaining subjects. The overall accuracy of the unseen subject for the optimized

parameters was 40%. The largest source of error was the atelectasis interface where a large number of image patches were classified as atelectasis/tumor despite the subject not exhibiting atelectasis. Following these results, the network was adapted again to employ the skip connection techniques of the ResNet architecture to see if more accuracy was possible with an even deeper network.

4.3.1.2 ResNet CNNs

The ResNet like network, referred to as the ResNet CNN, was also developed incorporating the batch normalization scheme suggested by He et al. in their follow-up paper[103] to the original paper[104] introducing the ResNet architecture. The ResNet architecture differs from the extended tradition CNN described in the previous section by employing skip connections to pass information learned in previous layers to later layers and the use of identity blocks to prevent overfitting. These skip connections improve the ability of the SGD optimization to learn by removing the exploding and vanishing gradient problem often seen as networks get deeper allowing the potential for additional benefits from deeper architectures.[104]

The ResNet architecture begins like the traditional CNN where the input undergoes a series of convolution layers and a maximum pooling layer. However, following the initial maximum pooling layer, a series of residual and identity blocks are added in place of the additional traditional convolutional layers before the final, fully connected layers. The identity blocks and single residual blocks implemented in this work are descried in the 2016 paper by He et al.[103] using the “full pre-activation” set up. The double residual block added an additional convolution layer with desired filters followed by batch normalization with ReLU activation before the final weight layer seen in the single residual block. To

create the double ResNet CNN, convolution 2a and convolution 2b as a group and convolution 3a and convolution 3b as a group in Figure 25 were replaced with the double convolution block, one for each group, with varying numbers of identity blocks following before the max pool layers. This network was still trained using SGD, with a momentum term and categorical cross-entropy as the loss function.

The overall accuracy of the unseen test subject for the double ResNet CNN varied with the number of identity layers used from 49% without any identity layers to 52%, 43% and 30% for 1, 2, and 3 identity layers respectively. For comparison, the test subject was kept the same as used to test the extended CNN in the previous section. For the best preforming network and hyper parameters, the individual class confusion matrix can be seen in Table 7. The error for each individual class in order from highest to lowest was as follows: 31.9% mediastinum (Med)/tumor, 26.8% vessel/tumor, 24.0% not interface, 21.0% atelectasis (AT)/tumor, 17.2% air way (AW)/tumor, 14.5% lung/tumor, 1% bone/tumor, and 0% chest wall (CW)/tumor. There were three interfaces not present in the test subject: the chest wall, atelectasis, and bone, as the subject has a centrally located tumor that did not obstruct an airway. The mediastinum/tumor was most often confused with the vessel/tumor and air way/tumor. The not interface patches also had a low number of correctly predicted labels most often being mistaken for atelectasis and mediastinum.

Table 7: Confusion matrix for the double ResNet CNN with one identity layer.

		Not Interface	Lung	AT	Med	CW	Vessel	AW	Bone	Total Labels
True Label	Not Interface	169	98	133	194	0	72	85	2	753
	Lung	24	598	12	35	0	0	84	0	753
	AT	0	0	0	0	0	0	0	0	0
	Med	12	18	118	414	0	141	50	0	753
	CW	0	0	0	0	0	0	0	0	0
	Vessel	4	9	261	233	0	240	6	0	753
	AW	2	55	3	127	0	0	560	0	747
	Bone	0	0	0	0	0	0	0	0	0
	Total Prediction	211	778	527	1003	0	453	785	2	

4.3.1.3 Encoder Network

Following analysis of the results of the ResNet CNN, one more modification to the network architecture was attempted. In order to improve the accuracy of the not interface labels, the labels were modified to change the not interface label from a class identification to an encoder, indicating the presence of any interface by the value of 1 and no interface with 0. The labels of the remaining interface were unchanged. This change to the labels resulted in changes to the network architecture, the loss function utilized, and the balancing method for the data given to the network.

To better train the network using the encoder label and interface labels, the dataset provided to the network had to not only have a balanced number of patches between the different interface types excluding the not interface patches but also between the total number of patches containing interfaces and number of patches without any interfaces. This resulted in more examples of the not interface patches being added to the training set. The extended CNN was again modified by replacing convolution 2a, convolution 2b, convolution 3a, and convolution 3b in Figure 25 this time individually with the single residual blocks described in the previous section. The encoder portion of the label was evaluated by the binary cross-entropy loss, while the remaining interface labels

were evaluated with the categorical cross-entropy. The total loss for the network was the summation of the binary cross-entropy and the categorical cross-entropy. Ideally, this allows the network to learn features that indicate an interface, and then for those that are an interface, distinguish which interface is present.

The accuracy of the Encoder Net showed a compromise in the accuracy of identifying the interface type with the accuracy of identifying the presence of an interface. The highest accuracy achieved for identifying if an interface was present was 77% while the accuracy of identifying the interface type for the same network was only 37%. With different hyper parameters, the interface type accuracy was able to reach 54% while the accuracy of identifying if any interface was present was only 66%. This trade off suggested that creating two separate networks, one for each task, may be beneficial. It further highlighted that the network may not be able to distinguish the tumor tissue from the surrounding healthy tissue, thereby identifying where interfaces between the tumor and normal tissue exist. This was illustrated by creating a prediction map by preforming a sliding window prediction over the entire area of the image surrounding the tumor contour and plotting the prediction for the interface identification portion of the network only. An example slice can be seen in Figure 26.

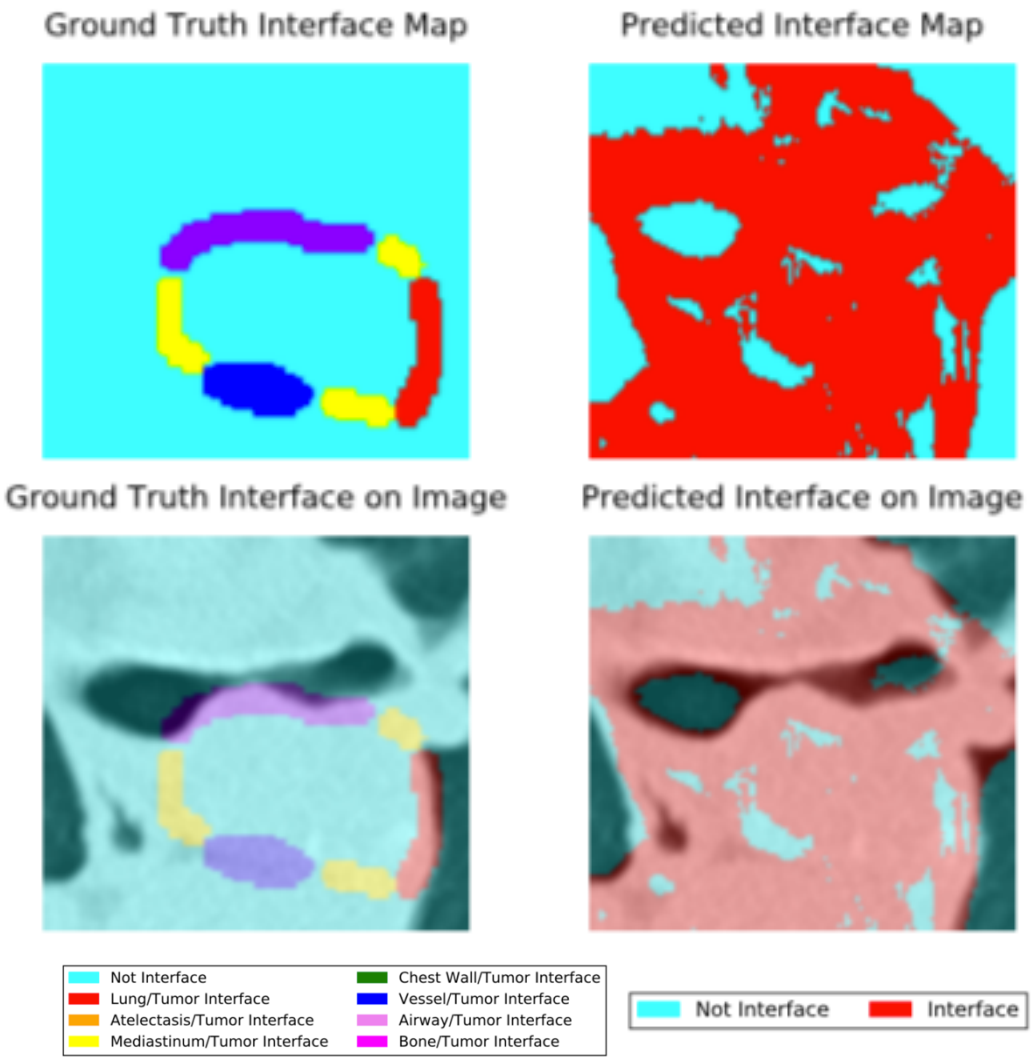


Figure 26: Illustrative slice from the prediction map for the unseen test subject from the encoder net. The top and bottom images on the left side detail the ground truth interface type labels while the top and bottom images on the right demonstrate the predicted output of the encoder network for the presence of an interface of any type.

4.3.2 Single Task Networks

Following analysis of the results of the previous networks, it appeared the task of determining where the tumor was located in order to identify the interfaces, and then identifying which interface was present was too ambitious for a single pass of the networks designed so far. Considering this, it seemed appropriate to take a step back and evaluate simpler criteria for the network to identify. Three different networks were

developed: the first and second were to break the Encoder Net into two separately trained networks, one to identify the location of any interfaces and the other to identify the type of interface given an input that was known to be an interface; the third network was developed to identify the location of the tumor. The network architecture used for all three networks was the same: the Encoder Net architecture described above without the modifications to the loss function. The data fed into each of the networks was different depending on the desired outcome and the filter size of the final layer was adjusted accordingly. Instead of using only one test subject, the network was retrained three to five times with a different subject left out each time.

4.3.2.1 BinaryRes_IF Network

The existing dataset was used to create the datasets for the interface location prediction network, referred to as BinaryRes_IF network, by modifying the labels. The labels first disregarded the interface type information and instead collapsed them into one label indicating if any interface was present and leaving the not interface label unmodified. This created a binary set of labels indicating if the patch belonged the interface or not interface class. The binary cross-entropy was used to train the BinaryRes_IF network. For this network, training was completed by alternating one of four test subjects.

The accuracy of the new work for the first subject was 78%, the second subject was 63%, the third test subject was 80%, and fourth subject was 66%. However, when the prediction maps were created, the results, while more specific, seemed to identify the region of the tumor and some interfaces instead of just the interface, Figure 27. This observation was the motivation for the BinaryRes_Tumor network.

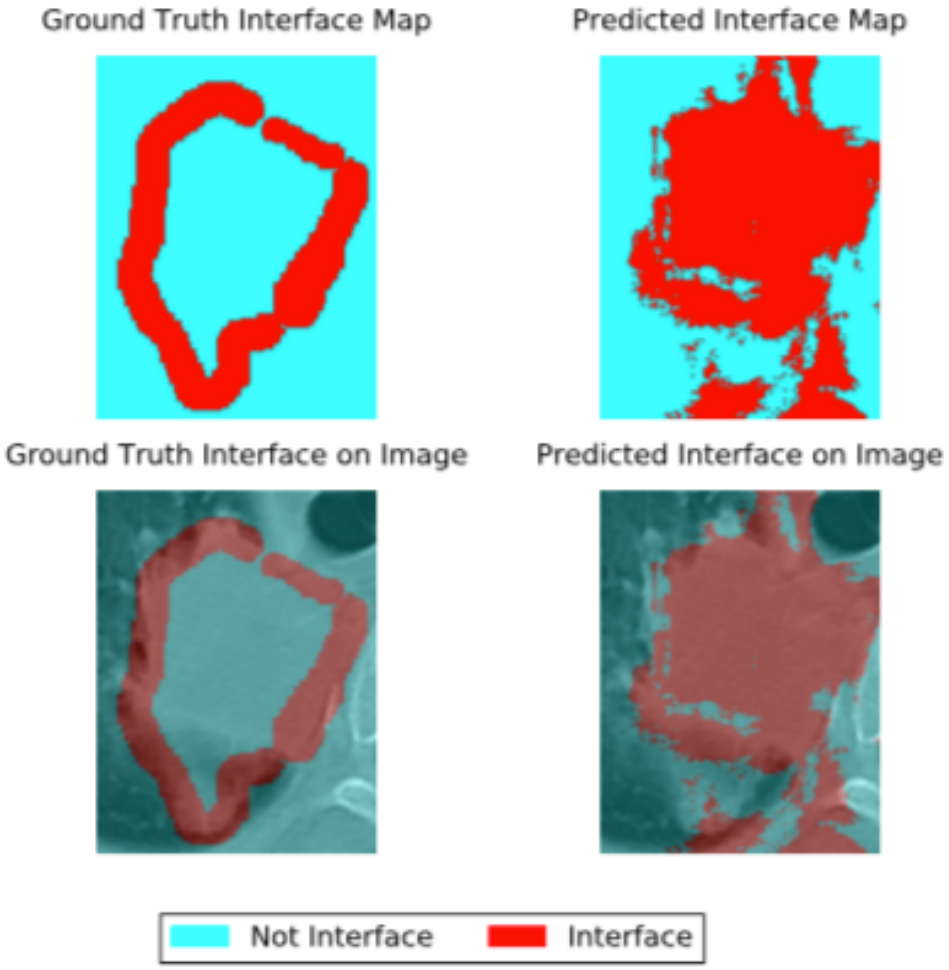


Figure 27: Illustrative slice of the comparison of the ground truth interface location with the output of the BinaryRes_IF network. The top row represents the ground truth and output maps while the bottom rows shows the maps overlaid on the corresponding images slice.

4.3.2.2 BinaryRes_Tumor Network

In an attempt to determine a network's ability to detect the location of the tumor, the existing network structure was repurposed. A new dataset was created for the BinaryRes_Tumor network in the same manner as described in section 4.2.1 with the planning tumor contour being supplied instead of the interface contours. This resulted in a series of patches with binary labels identifying the center pixel of the patch as being within the tumor contour or not. Like the contour used in the initial and subsequent

networks, the tumor contours were created by physicians using information from the PET scan as well as the CT scan. The network was again trained with SGD optimization with a binary cross-entropy loss function five times with a different subject withheld for the testing each time.

The network was able to achieve a relatively high accuracy for all 5 test subjects on the reserved set of patches with the accuracies ranging from 84% to 87%. Prediction maps were again created for all subjects to visualize how well the network would perform on the entire image. Analyzing these images indicated the network seemed to predict false positives in the muscle and atelectasis, Figure 28 , but overall seemed to capture most of the tumor. Illustrative slices from all test subjects can be seen in Appendix V. The Dice score was calculated as a measure of how well the predicted tumor and original tumor contour were in agreement. Conditional random fields (CRF) post processing was also applied using the pydensecrf package[154] implementing the fast CRF by Krahenbuhl and Koltun[155] to remove small areas of predicted classes that were not near similar labels. The Dice for the network output and CRF post processing can be seen in Table 8.

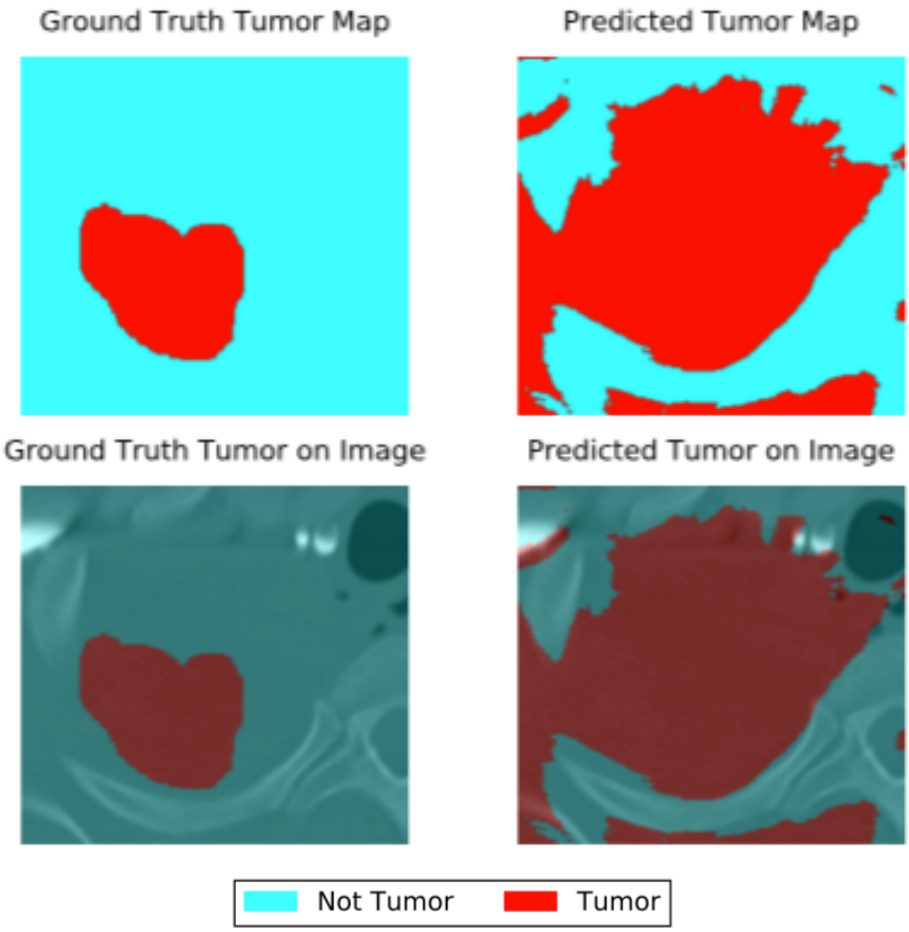


Figure 28: Illustrative slice of the comparison of the ground truth tumor location with the output of the BinaryRes_Tumor network. The top row represents the ground truth and output maps while the bottom rows show the maps overlaid on the corresponding images slice.

Table 8:Dice similarity coefficient for the BinaryRes_Tumor network predictions before and after CRF post-processing.

	Tumor	
Subject	Pre-CRF	Post-CRF
1	0.68	0.77
2	0.68	0.75
3	0.60	0.69
4	0.60	0.68
5	0.68	0.78

While the BinaryRes_Tumor network is able to identify the tumor with some precision, the number of false positives from regions such as atelectasis, at this time, is not conducive to building an uncertainty model. Since the test subject seen above also had median contours drawn using the CT only, as well as the PET/CT from the study by Karki et al.,[83] a Dice similarity coefficient between the network output and the median contours was calculated. The Dice similarity between the network output and the CT only median contour was 0.65, while the Dice similarity between the network output and the PET/CT median contour was 0.69. However, when the post-processing was applied, the Dice for the PET/CT dropped to 0.66, while the CT only Dice similarity rose to 0.70. This suggests the network predictions for the tumor location may be closer to those made by physicians only looking at a CT image rather than discerning any information from the CT image alone that would indicate areas of high metabolism as indicated on PET image.

4.3.2.3 IF_Only Network

The last network explored was the IF_Only network which endeavors to predict the interface type of a patch known to be from an interface. The existing labeled interface patch dataset was again used to create the datasets for the IF_Only network by modifying the labels. This time the existing dataset was first limited to only the patches known to contain an interface before the not interface class was removed from the labels of the remaining patches before training and testing. The network was trained using SGD with a momentum term, categorical cross-entropy for the loss function and was trained five times using a different test subject each time.

The IF_Only network resulted in an accuracy of between 38% and 65% across the five test subjects on the randomly selected test patches. Prediction maps for all the

patches within each subject image were created like with the previous networks to visualize the network output. For these prediction maps, only the patches identified in the physician defined interface contour, used as the ground truth, were extracted and fed to the trained network for predictions. From the prediction maps, there seemed to be greater accuracy in prediction of the interface than the test accuracy demonstrated as can be seen in Figure 29 and the illustrative slices from all test subjects in Appendix VI. The areas of the image not identified by the physicians were labeled as not interface for the purpose of illustration and were not part of the network prediction.

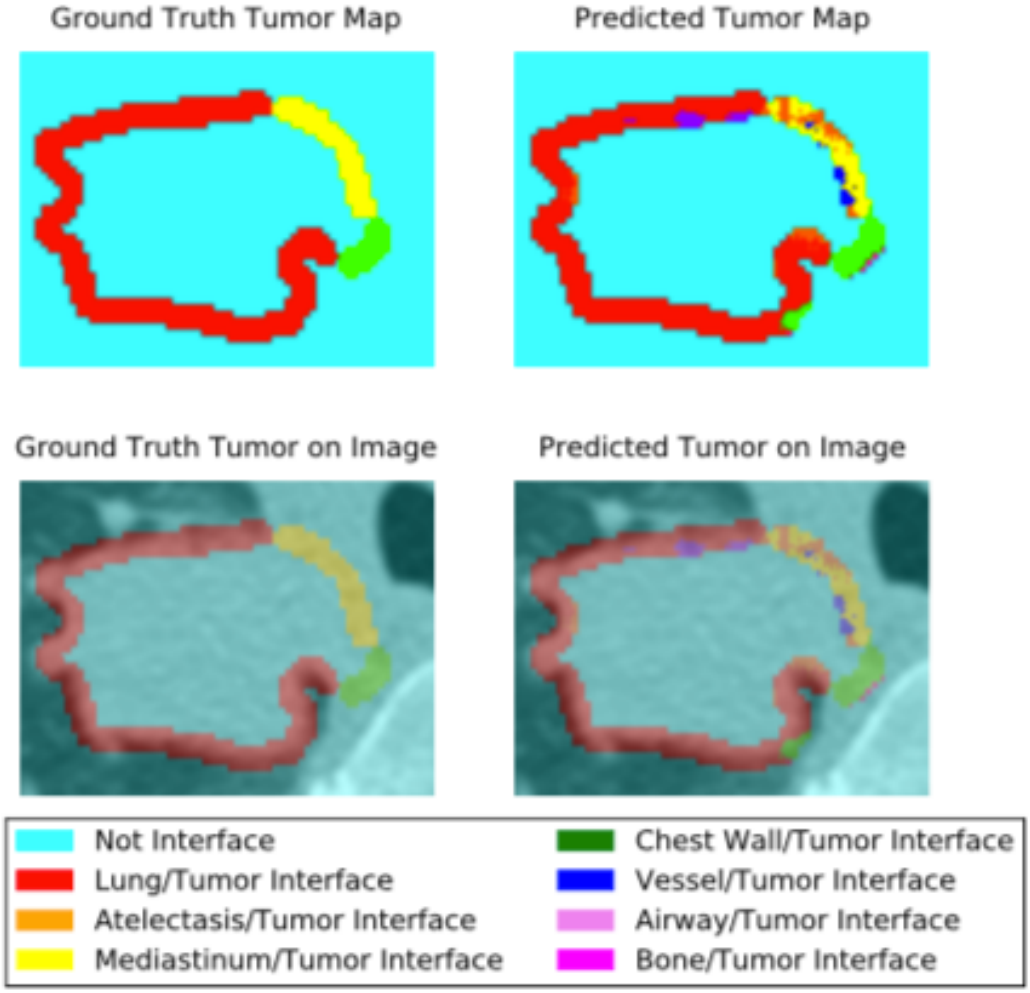


Figure 29: Illustrative slice comparing the ground truth physician interface labels with the IF_Only network output without post-processing. The top row indicates the maps of predicted and true interface types while the bottom row shows the maps overlaid on the corresponding image slice.

The Dice overlap between the network prediction and the ground truth was calculated for each interface to assess the agreement between the network output and the physician identified interface locations across the whole image. CRF post-processing was applied to the interface prediction labels to try and improve the Dice overlap. The results can be seen in Table 9.

Table 9: Dice similarity coefficient comparison for IF_Only network predictions before and after CRF post-processing.

	Lung		AT		Med		CW		Vessel		AW	
Subject	Pre-CRF	Post-CRF	Pre-CRF	Post-CRF	Pre-CRF	Post-CRF	Pre-CRF	Post-CRF	Pre-CRF	Post-CRF	Pre-CRF	Post-CRF
1	0.70	0.83	--	--	0.12	0.22	0.73	0.84	0.46	0.64	0.31	0.42
2	0.64	0.75	0.04	0.04	0.27	0.25	--	--	--	--	--	--
3	0.85	0.96	--	--	0.42	0.68	0.56	0.67	--	--	--	--
4	0.65	0.80	0.51	0.79	0.11	0.12	0.54	0.68	0.22	0.44	0.42	0.51
5	0.69	0.84	0.34	0.54	0.18	0.19	--	--	0.41	0.57	0.38	0.57

-- indicates interface was not present for test subject, AT is the atelectasis/tumor interface, Med is the mediastinum/tumor interface, CW is the chest wall/tumor interface, and AW is the air way/tumor interface.

The interface that the network had the most difficulty identifying was the mediastinum/tumor interface. This interface was often confused with atelectasis, vessel, chest wall, and airways which all share a similar intensity on CT scans, and in the case of vessels and air ways, may be found throughout the mediastinum. One interesting observation, however, was the network's ability to pick up on nearby structures, such as bone, that were not part of the physician ground truth label, Figure 30. These incorrect labels may in some cases still be informative and show the ground truth may benefit from review or consensus and could affect the accuracy of labels such as the mediastinum. In the subjects with large regions of atelectasis, the network seemed able to identify the interface with higher accuracy. The network also seemed to be able to predict the lung, airways, bone, and chest wall interfaces with relatively high accuracy. The average class accuracy for the interfaces are: lung/tumor 82%, atelectasis/tumor 74%,

mediastinum/tumor 67%, chest wall/tumor 89%, vessel/tumor 82%, air way/tumor 91%, and bone/tumor 95%. Similarly, the physician contour delineation uncertainty was also low for these interfaces as seen in Figure 21, while the atelectasis region had the largest contour delineation uncertainty.

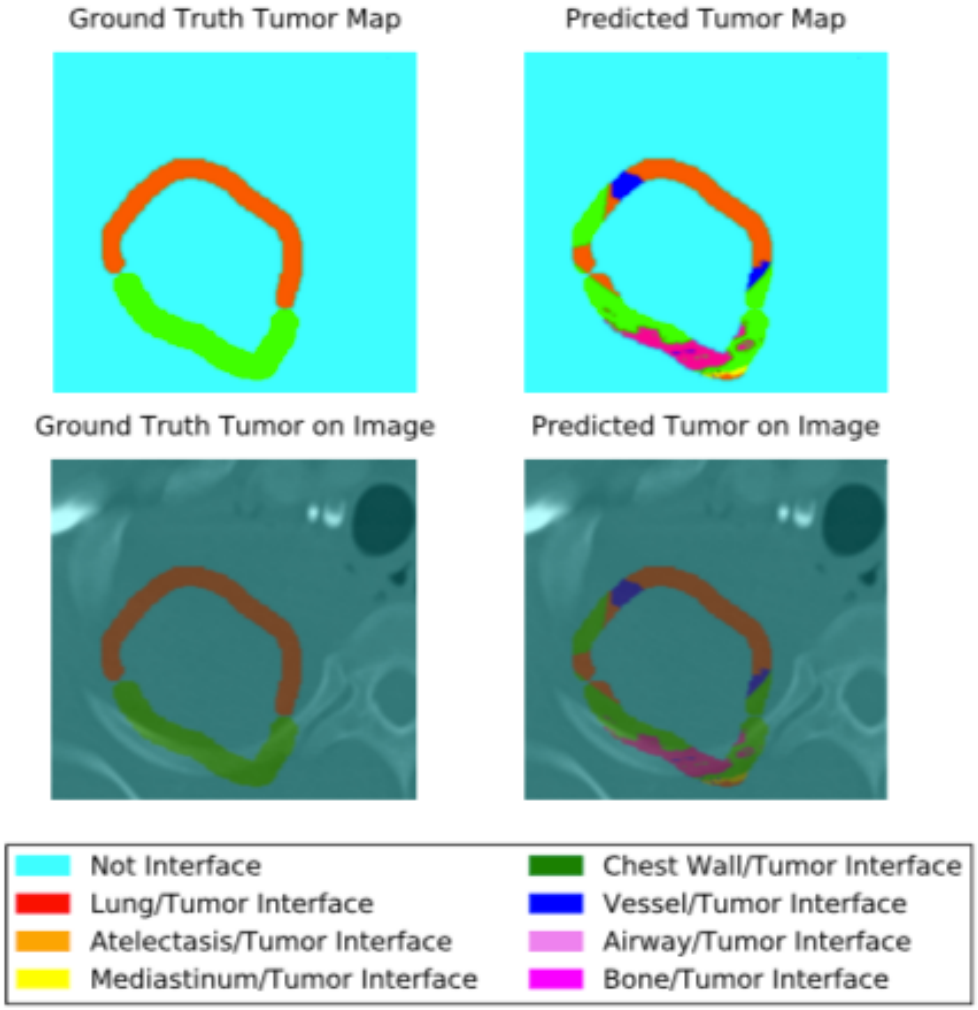


Figure 30: Illustrative slice of comparison of the ground truth and predicted maps (top row) with the maps overlaid on the corresponding image. Of interest is the area predicted as bone and its proximity to the rib in the corresponding image.

A summary of all the networks investigated as part of specific aim 2 can be seen in Table 10.

Table 10: Summary of networks investigated and findings.

Network Name	Task	Description	Results
Initial CNN	Identify interface type including 'not interface' class	Traditional CNN structure with 6 weight bearing layers and small subject set.	Interface types able to be identified with moderate accuracy
Extended CNN	Identify interface type including 'not interface' class	Traditional CNN structure with 8 weight bearing layers and expanded subject set.	Increased accuracy of predictions, but false positives for 'not interface' class
ResNet CNN	Identify interface type including 'not interface' class	Modified extended CNN with skip connections across 2 convolution layers and followed by varying number of identity layers	Reduced number of false positives from extended ResNet CNN, and achieved moderate accuracy
Encoder Net	Identify presence of interface and which type	Modified extended CNN with skip connections across 1 convolution layer and followed by varying number of identity layers. Preforms 2 tasks simultaneously combined loss function and two predictions per patch.	Trade off in accuracy regarding if a patch is an interface or which interface type. Hinted the network may not identify tumor location
BinaryRes_IF	Identify presence of interface	Modified encoder net for only interface presence. Input became binary labels with only one loss function and output per patch	Confirmed suspicion that network was not identifying tumor location before identifing interface
BinaryRes_Tumor	Identify presence of tumor	Same network as BinaryRes_IF, input data altered to contain labeled image patches indicating presence of tumor	Network achieved good accuracy for most patients. False positives in regions of involved lymph nodes, muscle, and atelectasis
IF_Only	Identify interface type without 'not interface' class	Modified encoder net for only interface presence. Input became binary labels with only one loss function and output per patch	Network achieved good accuracy for most interfaces particularly with top 2 and 3 results. Basis of information provided to physician

4.4 Uncertainty tool

The uncertainty tool proposed by this aim seeks to combine the predictive power of the IF_Only network with a measure of uncertainty in the interface. As discussed in 1.4 Contour Delineation Uncertainty, a margin is added to the GTV to account for uncertainty such as inconsistencies in set-up, motion, mechanical uncertainties, and others. The contour delineation uncertainty is only one of the uncertainties counted as part of this margin which for lung cancer is approximately 5mm. The proposed uncertainty tool seeks to provide the probability that the contour delineation uncertainty exceeds the 5mm threshold for the interface at the physician selected point. The confidence the network has in its interface type prediction is displayed along with the probabilities for the top three predictions to provide the physician with additional information.

The probability of the contour delineation uncertainty exceeding a threshold was calculated utilizing the bilinear distances from the median contour to the individual contours from the study by Karki el al.[83] For this work, only the data for the contours drawn using both the PET and CT images for the primary tumor were considered. The probability of exceeding a threshold was calculated by dividing the total number of points exceeding the desired threshold by the total number of points evaluated. The results of various thresholds by interface type can be seen in Table 11.

Table 11: Probability of contour delineation uncertainty exceeding various thresholds by interface type.

Interface/ Threshold(mm)	AT	CW	Hilum	Lung	Med	Vessel
1	0.9677	0.6871	0.8297	0.8106	0.7062	0.6851
2	0.6615	0.2033	0.3871	0.2375	0.2519	0.0797
3	0.2856	0.0683	0.1375	0.0735	0.0894	0.0036
4	0.0840	0.0300	0.0431	0.0212	0.0285	0.0000
5	0.0264	0.0085	0.0144	0.0081	0.0131	0.0000
6	0.0091	0.0049	0.0047	0.0044	0.0051	0.0000
7	0.0047	0.0031	0.0009	0.0019	0.0009	0.0000
8	0.0018	0.0011	0.0000	0.0004	0.0000	0.0000
9	0.0007	0.0005	0.0000	0.0001	0.0000	0.0000
10	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000

The probability of exceeding the 5mm threshold is highlighted for all interface types

The IF_Only network was able to predict the type of interface with suitable accuracy for the best guess. Looking at the top two predictions for the network, the accuracy increased to 55%-92% where 14% is guessing. If the top three predictions are considered, the accuracy improves to 75%-97%. The output of the IF_Only network provides the level of activation from the network for all classes providing a measure of certainty for the final class assignment. The activation level of the top three interface predictions provides the physician with a measure of certainty from the network that can be used with the associated probability of exceeding the 5mm threshold.

The proposed usage for this tool would be for the physician to click on a point of interest along the interface of the tumor. The uncertainty tool extracts the image patch around the point of interest and runs the IF_Only prediction. The uncertainty tool returns the top three interface predictions with the network's activation levels and associated

probability of exceeding the 5mm threshold. The physician can then use this information to adjust their contour accordingly. A mock visualization can be seen in Figure 31.

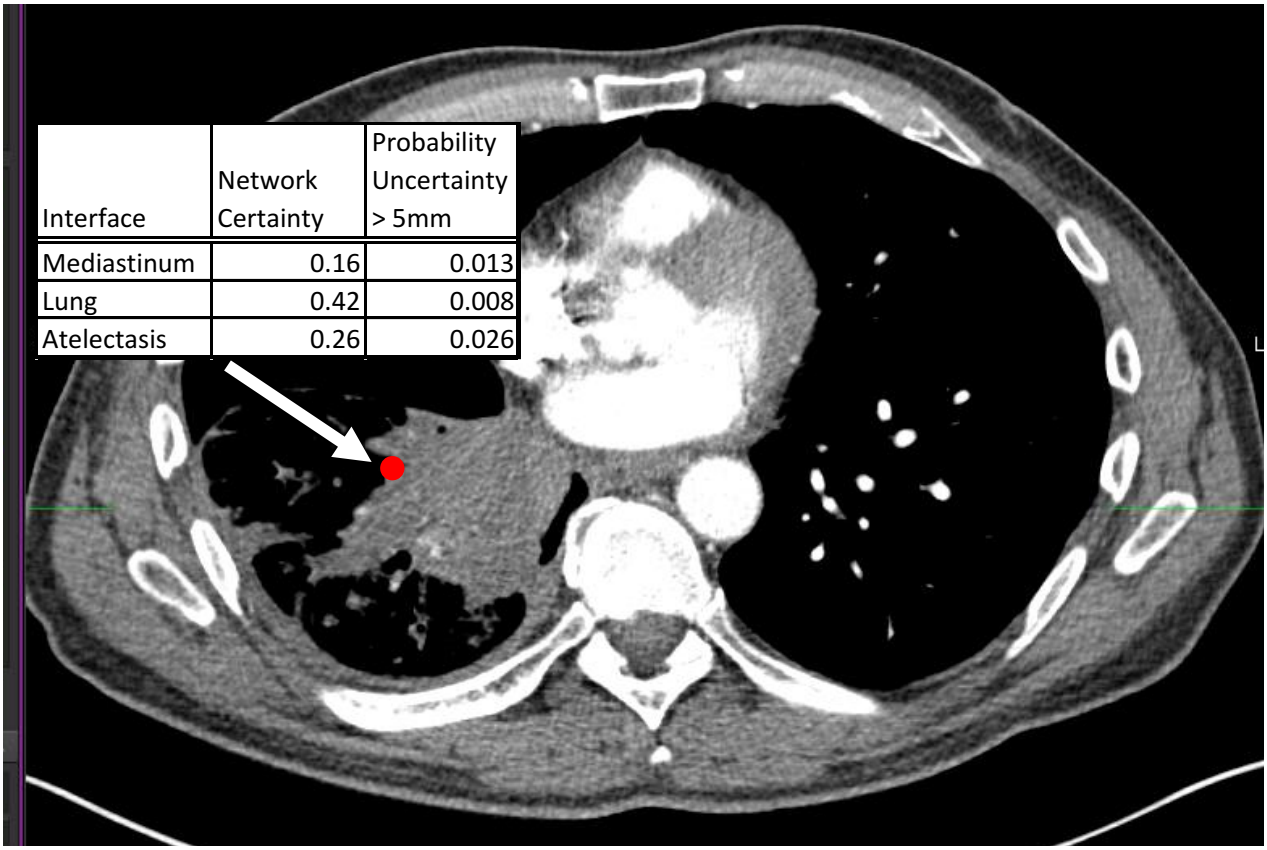


Figure 31: Mock contour assistance tool. The network prediction with certainty and probability uncertainty exceeds 5mm is presented for the selected point (exaggerated red dot with arrow).

4.5 Conclusion and Future Work

This aim developed a potential tool to provide additional information to physicians during contouring. During the course of network exploration, two networks were developed that succeeded at performing tumor location and interface type identification for known interface patches. The earlier network designs produced promising results at first glance; however, they seemed to struggle with identifying the location of the interfaces without guidance. The initial task appeared too complex to complete in one

pass for the simple CNN network structures investigated in this work. The final tool gives additional information to the physicians, but does not propose a tumor contour as has been suggested to reduce contour delineation by other research.[87] The information provided by the developed tool seeks to improve the contour accuracy by making physicians aware of how much uncertainty to expect allowing them to use their best judgement in defining the treatment target.

The networks in their current form have good accuracy, but not high enough for clinical use. Additional techniques could be implemented to refine and increase the accuracy of the network predictions to provide more accurate information to physicians. One such technique is a cascade network where the results of one network are fed to the next in order to increase the accuracy of the second network or to refine the output.[123,156] Christ et al.[156] first trained a network to identify the liver then the second network identified the liver lesions. For the interface prediction, this technique could increase the accuracy by first training the network to identify if an interface belongs to an easier to identify class such as lung/tumor interface and the second network could be trained to distinguish between similar tissues such as atelectasis and the mediastinum.

In future work, the BinaryRes_Tumor network could be modified to include the PET data as an additional channel during network training and prediction which may improve the accuracy of the tumor identification. This type of network could be used in an alternative manner to reduce tumor uncertainty by first suggesting a tumor contour and allowing the physician to refine the contour as necessary. An alternative potential extension with cascade networks could be to first train a network to identify the lung, including pathology within a candidate region, and then identify the tumor. This could

reduce the false positives in muscle tissue outside the lung. The percentage certainty of the network for each voxel could then be used to highlight were the network was uncertain in the tumor location and a physician could expect larger uncertainty.

5 Conclusion

The goal of this dissertation was to investigate the feasibility of building MR and multi-modality predictive models for lung cancer and to investigate the feasibility of using machine learning techniques to build an uncertainty model to aid in contour delineation.

The first aim of this work was to investigate the feasibility of building predictive models using single CT and MR modalities and multi-modality models. Repeatability for the MR texture features was first determined and compared with the repeatability of CT features. Several texture features were identified as repeatable for MR and CT images suggesting MR could be used as a basis for predictive modeling provided standardized imaging techniques are used. Next, a workflow for creating predictive models was developed and used to identify predictive models for single modality MR and CT features, as well as multi-modality models for local control at the end of treatment, and overall survival at 12, 18, and 24 months. Two different feature selection techniques were investigated as part of the work flow. After controlling for false discovery rates, multiple significant models were identified for overall survival while the local control at the end of treatment models were not significant. A control experiment was conducted on normal tissue to further aid in identifying spurious results. The normal tissue models identified the medoid feature selection method as a more robust method for the small subject group used in this study

as it produced a smaller number of significant models for tumor outcome. The accuracy of the significant MR models was comparable to the significant CT models and MR features appeared in the top multi-modality models as well, suggesting that further study of MR features in a larger patient cohort is warranted.

The second aim of this work sought to investigate the feasibility of building an uncertainty model to aid in physician contour delineation. This work built on previous work by Karki et al.[83] to investigate the link between the tumor interface type and the amount of uncertainty. While the interface type alone was not enough to explain the level of uncertainty, there appeared to be a relationship between uncertainty and interface type. Several convolution neural networks were tested to predict the interface type and tumor location from image patches. The accuracy of the interface type prediction network developed achieved moderate accuracy on the best guess but was significantly improved by scoring the top 2 or 3 interface class accuracy. The tumor location network had an acceptable accuracy for most patients; however, still struggled with similar looking physiology, such as near atelectasis. Further refinement would be needed for either network to be clinically acceptable, but the results show neural networks warrant further research in lung cancer.

Improved patient treatment and outcome is a constant goal in radiation oncology. Being able to predict treatment outcome prior to, or early during treatment, and improving accuracy of treatment targets are two ways to improve on current practices. The results of this work investigated the potential uses of radiomics in MR imaging for lung cancer as well as convolutional neural networks to assist in contour delineation by providing

physicians with addition information. These techniques show promise and should be investigated further.

6 References

[1] American Cancer Society, “Cancer facts & figures 2016,” (2016).

[2] B.W. Stewart and C.P. Wild (eds.), *World Cancer Report 2014* (International Agency for Research on Cancer, Lyon, FRA, 2014).

[3] C.K. Howlader N, Noone AM, Krapcho M, Miller D, Bishop K, Altekruse SF, Kosary CL, Yu M, Ruhl J, Tatalovich Z, Mariotto A, Lewis DR, Chen HS, Feuer EJ, *National Cancer Institute Seer cancer statistics review* (Bethesda, MD, 2016).

[4] American Lung Association Epidemiology and Statistics Unit Research and Program Services Division, “Trends in Lung Cancer Morbidity and Mortality,” (November), 1–34 (2014).

[5] P. Therasse, S.G. Arbuck, E. a Eisenhauer, *et al.*, “New Guidelines to Evaluate the Response to Treatment in Solid Tumors,” JNCI J. Natl. Cancer Inst. **92**(3), 205–216 (2000).

[6] E.A. Eisenhauer, P. Therasse, J. Bogaerts, *et al.*, “New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1),” Eur. J. Cancer **45**(2), 228–247 (2009).

[7] P.A. Kupelian, C. Ramsey, S.L. Meeks, *et al.*, “Serial megavoltage CT imaging during external beam radiotherapy for non-small-cell lung cancer: Observations on tumor regression during treatment,” Int. J. Radiat. Oncol. Biol. Phys. **63**(4), 1024–1028 (2005).

[8] G. Lim, A. Bezjak, J. Higgins, *et al.*, “Tumor regression and positional changes in non-small cell lung cancer during radical radiotherapy,” J. Thorac. Oncol. **6**(3), 531–536 (2011).

[9] J. Fox, E. Ford, K. Redmond, J. Zhou, J. Wong, and D.Y. Song, “Quantification of Tumor Volume Changes During Radiotherapy for Non-Small-Cell Lung Cancer,” Int. J. Radiat. Oncol. Biol. Phys. **74**(2), 341–348 (2009).

[10] J.T. Bushburg, J.A. Seibert, E.M. Leidholt Jr., and J.M. Bonne, *The Essential Physics of Medical Imaging*, 3rd ed. (Lippincott Wlliams & Wilkins, Philadelphia, 2012).

[11] Y. Balagurunathan, Y. Gu, H. Wang, *et al.*, “Reproducibility and Prognosis of Quantitative Features Extracted from CT Images,” Transl. Oncol. **7**(1), 72–87 (2014).

[12] D. V Fried, S.L. Tucker, S. Zhou, *et al.*, “Prognostic value and reproducibility of pretreatment CT texture features in stage III non-small cell lung cancer.,” Int. J. Radiat. Oncol. Biol. Phys. **90**(4), 834–842 (2014).

[13] T.P. Coroller, P. Grossmann, Y. Hou, *et al.*, “CT-based radiomic signature predicts distant metastasis in lung adenocarcinoma.,” Radiother. Oncol. **114**(3), 345–350 (2015).

[14] S. Bral, M. Duchateau, M. De Ridder, *et al.*, “Volumetric response analysis during chemoradiation as predictive tool for optimizing treatment strategy in locally advanced unresectable NSCLC,” Radiother. Oncol. **91**(3), 438–442 (2009).

[15] J. George, P. Claes, K. Vunckx, *et al.*, “A textural feature based tumor therapy response prediction model for longitudinal evaluation with PET imaging,” in *2012 9th IEEE Int. Symp. Biomed. Imaging*(IEEE, 2012), pp. 1048–1051.

[16] S.K. Jabbour, S. Kim, S.A. Haider, *et al.*, “Reduction in tumor volume by cone beam computed tomography predicts overall survival in non-small cell lung cancer treated with chemoradiation therapy,” Int. J. Radiat. Oncol. Biol. Phys. **92**(3), 627–633 (2015).

[17] X. Fave, L. Zhang, J. Yang, *et al.*, “Delta-radiomics features for the prediction of patient outcomes in non–small cell lung cancer,” Sci. Rep. **7**(1), 588 (2017).

[18] A.M. Rutman and M.D. Kuo, “Radiogenomics: Creating a link between molecular diagnostics and diagnostic imaging,” Eur. J. Radiol. **70**(2), 232–241 (2009).

[19] A.K. Das, M.H. Bell, C.S. Nirodi, M.D. Story, and J.D. Minna, “Radiogenomics predicting tumor responses to radiotherapy in lung cancer,” Semin. Radiat. Oncol. **20**(3), 149–155 (2010).

[20] H.J.W.L. Aerts, E.R. Velazquez, R.T.H. Leijenaar, *et al.*, “Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach,” Nat. Commun. **5**, (2014).

[21] E. Scalco, C. Fiorino, G.M. Cattaneo, G. Sanguineti, and G. Rizzo, “Texture analysis for the assessment of structural changes in parotid glands induced by radiotherapy.,” Radiother. Oncol. **109**(3), 384–7 (2013).

[22] M. Vallières, C.R. Freeman, S.R. Skamene, and I. El Naqa, “A radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities.,” Phys. Med. Biol. **60**(14), 5471–96 (2015).

[23] Y. Peng, Y. Jiang, T. Antic, M.L. Giger, S. Eggener, and A. Oto, “A study of T 2 -weighted MR image texture features and diffusion-weighted MR image features for computer-aided diagnosis of prostate cancer,” **8670**(773), 86701H (2013).

[24] H. Mi, C. Petitjean, P. Vera, and S. Ruan, “Robust Feature Selection to Predict Lung Tumor Recurrence,” Comput. Methods Mol. Imaging **22**, 103–112 (2015).

[25] S. AlZubi, N. Islam, and M. Abbod, “Multiresolution Analysis Using Wavelet, Ridgelet, and Curvelet Transforms for Medical Image Segmentation,” Int. J. Biomed. Imaging **2011**, 1–18 (2011).

[26] C. Parmar, E. Rios Velazquez, R. Leijenaar, *et al.*, “Robust Radiomics Feature Quantification Using Semiautomatic Volumetric Segmentation,” PLoS One **9**(7), e102107 (2014).

[27] C.E. McLaren, W.-P. Chen, K. Nie, and M.-Y. Su, “Prediction of Malignant Breast Lesions from MRI Features,” Acad. Radiol. **16**(7), 842–851 (2009).

[28] B. Ganeshan, K. a. Miles, R.C.D. Young, and C.R. Chatwin, “Texture analysis in non-contrast enhanced CT: Impact of malignancy on texture in apparently disease-free areas of the liver,” Eur. J. Radiol. **70**(1), 101–110 (2009).

[29] J. Fruehwald-Pallamar, J. Hesselink, M. Mafee, L. Holzer-Fruehwald, C. Czerny, and M. Mayerhoefer, “Texture-Based Analysis of 100 MR Examinations of Head and Neck Tumors – Is It Possible to Discriminate Between Benign and Malignant Masses in a Multicenter Trial?,” RöFo - Fortschritte auf dem Gebiet der Röntgenstrahlen und der Bildgeb. Verfahren (2015).

[30] A. Wibmer, H. Hricak, T. Gondo, *et al.*, “Haralick texture analysis of prostate MRI: utility for differentiating non-cancerous prostate from prostate cancer and differentiating prostate cancers with different Gleason scores,” Eur. Radiol. **25**(10), 2840–2850 (2015).

[31] M. Grkovski, A. Apte, J. Schwartz, *et al.*, “Reproducibility of 18F-FMISO intratumor distribution and

texture features in NSCLC," J. Nucl. Med. **56**(supplement_3), 126- (2015).

[32] S. Chicklore, V. Goh, M. Siddique, A. Roy, P.K. Marsden, and G.J.R. Cook, "Quantifying tumour heterogeneity in 18F-FDG PET/CT imaging by texture analysis," Eur. J. Nucl. Med. Mol. Imaging **40**(1), 133–140 (2013).

[33] F. Yang, M.A. Thomas, F. Dehdashti, and P.W. Grigsby, "Temporal analysis of intratumoral metabolic heterogeneity characterized by textural features in cervical cancer.," Eur. J. Nucl. Med. Mol. Imaging **40**(5), 716–27 (2013).

[34] F. Tixier, M. Hatt, C.C. Le Rest, A. Le Pogam, L. Corcos, and D. Visvikis, "Reproducibility of Tumor Uptake Heterogeneity Characterization Through Textural Feature Analysis in 18F-FDG PET," J. Nucl. Med. **53**(5), 693–700 (2012).

[35] P. Lambin, E. Rios-Velazquez, R. Leijenaar, *et al.*, "Radiomics: Extracting more information from medical images using advanced feature analysis," Eur. J. Cancer **48**(4), 441–446 (2012).

[36] V. Kumar, Y. Gu, S. Basu, *et al.*, "Radiomics: the process and the challenges," Magn. Reson. Imaging **30**(9), 1234–1248 (2012).

[37] M.E. Mayerhoefer, P. Szomolanyi, D. Jirak, A. Materka, and S. Trattnig, "Effects of MRI acquisition parameter variations and protocol heterogeneity on the results of texture analysis and pattern discrimination: An application-oriented study," Med. Phys. **36**(4), 1236 (2009).

[38] S.J. Savio, L.C. V Harrison, T. Luukkaala, *et al.*, "Effect of slice thickness on brain magnetic resonance image texture analysis.," Biomed. Eng. Online **9**(1), 60 (2010).

[39] D. Jirák, M. Dezortová, and M. Hájek, "Phantoms for texture analysis of MR images. Long-term and multi-center study.," Med. Phys. **31**(3), 616–22 (2004).

[40] Y. Balagurunathan, V. Kumar, Y. Gu, *et al.*, "Test–Retest Reproducibility Analysis of Lung CT Image Features," J. Digit. Imaging **27**(6), 805–823 (2014).

[41] D. Mackin, X. Fave, L. Zhang, *et al.*, "Measuring Computed Tomography Scanner Variability of Radiomics Features.," Invest. Radiol. **50**(8), 1–9 (2015).

[42] S.S.F. Yip and H.J.W.L. Aerts, "Applications and limitations of radiomics," Phys. Med. Biol. **61**(13), (2016).

[43] G. Lee, H.Y. Lee, H. Park, *et al.*, "Radiomics and its emerging role in lung cancer research, imaging biomarkers and clinical management: State of the art," Eur. J. Radiol. **86**, 297–307 (2017).

[44] M. Scrivener, E.E.C. de Jong, J.E. van Timmeren, T. Pieters, B. Ghaye, and X. Geets, "Radiomics applied to lung cancer: a review," Transl. Cancer Res. **5**(4), 398–409 (2016).

[45] W.H. Nailon, "Texture Analysis Methods for Medical Image Characterisation," Biomed. Imaging 75–100 (2010).

[46] F. Albregtsen, "Digital Image Analysis - Texture," 15 (2011).

[47] M. Tuceryan, M. Tuceryan, A.K. Jain, and A.K. Jain, "The Handbook of Pattern Recognition and Computer Vision (2nd Edition), Texture Analysis," Pattern Recognit. 207–248 (1998).

[48] R.M. Haralick, K. Shanmugam, and I. Dinstein, "Textural Features for Image Classification," IEEE Trans. Syst. Man. Cybern. **3**(6), 610–621 (1973).

[49] M. Bevk and I. Kononenko, "A statistical approach to texture description of medical images: A

preliminary study," Proc. IEEE Symp. Comput. Med. Syst. 239–244 (2002).

[50] L.-K. Soh and C. Tsatsoulis, "Texture analysis of SAR sea ice imagery using gray level co-occurrence matrices," IEEE Trans. Geosci. Remote Sens. **37**(2), 780–795 (1999).

[51] D. Assefa, H. Keller, C. Ḿnard, N. Laperriere, R.J. Ferrari, and I. Yeung, "Robust texture features for response monitoring of glioblastoma multiforme on T1 -weighted and T2 -FLAIR MR images: A preliminary investigation in terms of identification and segmentation," Med. Phys. **37**(4), 1722–1736 (2010).

[52] M. Galloway, "Texture Analysis Using Gray Level Run Lengths," Comput. Graph. Image Process. **4**(2), 172–179 (1975).

[53] A. Chu, C.M. Sehgal, and J.F. Greenleaf, "Use of gray value distribution of run lengths for texture analysis," Pattern Recognit. Lett. **11**(6), 415–419 (1990).

[54] B. V. Dasarathy and E.B. Holder, "Image characterizations based on joint gray level—run length distributions," Pattern Recognit. Lett. **12**(8), 497–502 (1991).

[55] X. Tang, "Texture information in run-length matrices," IEEE Trans. Image Process. **7**(11), 1602–1609 (1998).

[56] D.-H. Xu, A.S. Kurani, J.D. Furst, and D.S. Raicu, "Run-Length Encoding for Volumetric Texture," Int. Conf. Vis. Imaging Image Process. 452–458 (2004).

[57] G. Thibault, J. Angulo, and F. Meyer, "Advanced statistical matrices for texture characterization: application to cell classification.," IEEE Trans. Biomed. Eng. **61**(3), 630–7 (2014).

[58] G. Thibault, B. Fertil, C. Navarro, *et al.*, "Texture Indexes and Gray Level Size Zone Matrix Application to Cell Nuclei Classification," Pattern Recognit. Inf. Process. 140–145 (2009).

[59] M. Amadasun and R. King, "Textural features corresponding to textural properties," IEEE Trans. Syst. Man Cybern. **19**(5), 1264–1273 (1989).

[60] A. Meyer-Baese and V. Schmid, "The Wavelet Transform in Medical Imaging," in *Pattern Recognit. Signal Anal. Med. Imaging*(Elsevier, 2014), pp. 113–134.

[61] R. Porter and C. Canagarajah, "Robust rotation-invariant texture classification: wavelet, Gabor filter and GMRF based schemes," Vision, Image Signal Process. IEE Proc. - **144**(3), 180–188 (2009).

[62] S. Arivazhagan, L. Ganesan, and S.P. Priyal, "Texture classification using Gabor wavelets based rotation invariant features," Pattern Recognit. Lett. **27**(16), 1976–1982 (2006).

[63] W. Huang, T. Zhou, L. Ma, *et al.*, "Standard uptake value and metabolic tumor volume of 18F-FDG PET/CT predict short-term outcome early in the course of chemoradiotherapy in advanced non-small cell lung cancer," Eur. J. Nucl. Med. Mol. Imaging **38**(9), 1628–1635 (2011).

[64] L. Bernardin, N.H.M. Douglas, D.J. Collins, *et al.*, "Diffusion-weighted magnetic resonance imaging for assessment of lung lesions: Repeatability of the apparent diffusion coefficient measurement," Eur. Radiol. **24**(2), 502–511 (2014).

[65] A. Kassner and R.E. Thornhill, "Texture Analysis: A Review of Neurologic MR Imaging Applications," Am. J. Neuroradiol. **31**(5), 809–816 (2010).

[66] L. Hunter, "Radiomics of NSCLC: Quantitative CT Image Feature Characterization and Tumor Shrinkage Prediction," UT GSBS Diss. Theses (Open Access) (2013).

[67] L.P. Clarke, B.S. Croft, R. Nordstrom, H. Zhang, G. Kelloff, and J. Tatum, “Quantitative imaging for evaluation of response to cancer therapy.,” Transl. Oncol. **2**(4), 195–7 (2009).

[68] G. Doumou, M. Siddique, C. Tsoumpas, V. Goh, and G.J. Cook, “The precision of textural analysis in 18F-FDG-PET scans of oesophageal cancer,” Eur. Radiol. 2805–2812 (2015).

[69] H.X. Barnhart and D.P. Barboriak, “Applications of the repeatability of quantitative imaging biomarkers: a review of statistical analysis of repeat data sets,” Transl. Oncol. **2**(4), 231–235 (2009).

[70] D.P. Schuster, “The opportunities and challenges of developing imaging biomarkers to study lung function and disease,” Am. J. Respir. Crit. Care Med. **176**(3), 224–230 (2007).

[71] L.A. Hunter, S. Krafft, F. Stingo, *et al.*, “High quality machine-robust image features: identification in nonsmall cell lung cancer computed tomography images.,” Med. Phys. **40**(12), 121916 (2013).

[72] K.M. Panth, R.T.H. Leijenaar, S. Carvalho, *et al.*, “Is there a causal relationship between genetic changes and radiomics-based image features? An in vivo preclinical experiment with doxycycline inducible GADD34 tumor cells,” Radiother. Oncol. **116**(3), 462–466 (2015).

[73] L.A. Hunter, Y.P. Chen, L. Zhang, *et al.*, “NSCLC tumor shrinkage prediction using quantitative image features,” Comput. Med. Imaging Graph. **49**, 29–36 (2015).

[74] S. Herlidou-Même, J.M. Constans, B. Carsin, *et al.*, “MRI texture analysis on texture test objects, normal brain and intracranial tumors,” Magn. Reson. Imaging **21**(9), 989–993 (2003).

[75] G. Collewet, M. Strzelecki, and F. Mariette, “Influence of MRI acquisition protocols and image intensity normalization methods on texture classification,” Magn. Reson. Imaging **22**(1), 81–91 (2004).

[76] L.C. V Harrison, M. Raunio, K.K. Holli, *et al.*, “MRI texture analysis in multiple sclerosis: toward a clinical analysis protocol.,” Acad. Radiol. **17**(6), 696–707 (2010).

[77] S. Gourtsoyianni, G. Doumou, D. Prezzi, *et al.*, “Primary Rectal Cancer : Repeatability of Global and Local- Regional MR Imaging Texture,” Radiology **000**(2), 1–10 (2017).

[78] D. Malyarenko, C.J. Galbán, F.J. Londy, *et al.*, “Multi-system repeatability and reproducibility of apparent diffusion coefficient measurement using an ice-water phantom,” J. Magn. Reson. Imaging **37**(5), 1238–1246 (2013).

[79] S. Doblas, G.S. Almeida, F.-X. Blé, *et al.*, “Apparent diffusion coefficient is highly reproducible on preclinical imaging systems: Evidence from a seven-center multivendor study,” J. Magn. Reson. Imaging n/a-n/a (2015).

[80] J.-J. Sonke and J. Belderbos, “Adaptive Radiotherapy for Lung Cancer,” Semin. Radiat. Oncol. **20**(2), 94–106 (2010).

[81] M. Van Herk, “Errors and Margins in Radiotherapy,” Semin. Radiat. Oncol. **14**(1), 52–64 (2004).

[82] H. Vorwerk, G. Beckmann, M. Bremer, *et al.*, “The delineation of target volumes for radiotherapy of lung cancer patients,” Radiother. Oncol. **91**(3), 455–460 (2009).

[83] K. Karki, S. Saraiya, G.D. Hugo, *et al.*, “Variabilities of Magnetic Resonance Imaging–, Computed Tomography–, and Positron Emission Tomography–Computed Tomography–Based Tumor and Lymph Node Delineations for Lung Cancer Radiation Therapy Planning,” Int. J. Radiat. Oncol. **99**(1), 80–89 (2017).

[84] P. Giraud, S. Elles, S. Helfre, *et al.*, “Conformal radiotherapy for lung cancer: Different delineation

of the gross tumor volume (GTV) by radiologists and radiation oncologists," Radiother. Oncol. **62**(1), 27–36 (2002).

[85] P. Bowden, R. Fisher, M. Mac Manus, *et al.*, "Measurement of lung tumor volumes using three-dimensional computer planning software," Int. J. Radiat. Oncol. Biol. Phys. **53**(3), 566–573 (2002).

[86] R.J.H.M. Steenbakkers, J.C. Duppen, I. Fitton, *et al.*, "Reduction of observer variation using matched CT-PET for lung cancer delineation: A three-dimensional analysis," Int. J. Radiat. Oncol. Biol. Phys. **64**(2), 435–448 (2006).

[87] A. van Baardwijk, G. Bosmans, L. Boersma, *et al.*, "PET-CT-Based Auto-Contouring in Non-Small-Cell Lung Cancer Correlates With Pathology and Reduces Interobserver Variability in the Delineation of the Primary Tumor and Involved Nodal Volumes," Int. J. Radiat. Oncol. Biol. Phys. **68**(3), 771–778 (2007).

[88] Y. Gu, V. Kumar, L.O. Hall, *et al.*, "Automated delineation of lung tumors from CT images using a single click ensemble segmentation approach," Pattern Recognit. **46**(3), 692–702 (2013).

[89] K. Lu and W.E. Higgins, "Interactive segmentation based on the live wire for 3D CT chest image analysis," Int. J. Comput. Assist. Radiol. Surg. **2**(3–4), 151–167 (2007).

[90] J. Willner, K. Baier, E. Caragiani, A. Tschammler, and M. Flentje, "Dose, volume, and tumor control prediction in primary radiotherapy of non-small-cell lung cancer," Int. J. Radiat. Oncol. Biol. Phys. **52**(2), 382–389 (2002).

[91] J.D. Bradley, N. Ieumwananonthachai, J.A. Purdy, *et al.*, "Gross tumor volume, critical prognostic factor in patients treated with three-dimensional conformal radiation therapy for non-small-cell lung carcinoma," Int. J. Radiat. Oncol. **52**(1), 49–57 (2002).

[92] R. Rengan, K.E. Rosenzweig, E. Venkatraman, *et al.*, "Improved local control with higher doses of radiation in large-volume stage III non-small-cell lung cancer," Int. J. Radiat. Oncol. Biol. Phys. **60**(3), 741–747 (2004).

[93] J.D. Bradley, R. Paulus, R. Komaki, *et al.*, "Standard-dose versus high-dose conformal radiotherapy with concurrent and consolidation carboplatin plus paclitaxel with or without cetuximab for patients with stage IIIA or IIIB non-small-cell lung cancer (RTOG 0617): A randomised, two-by-two factorial p," Lancet Oncol. **16**(2), 187–199 (2015).

[94] W.S. Mcculloch and W. Pitts, "a Logical Calculus of the Ideas Immanent in Nervous Activity," Bull. Math. Biophys. **5**(1), 99–115 (1943).

[95] D. Hebb, *The organization of Behaviour* (Weily, New York, 1949).

[96] F. Rosenblatt, "The perceptron: A probabilistic model for information storage and organization in …," Psychol. Rev. **65**(6), 386–408 (1958).

[97] D.E. Rumelhart, G.E. Hinton, and R.J. Williams, "Learning representations by back-propagating errors," Nature **323**(6088), 533–536 (1986).

[98] Y. LeCun and others, "Generalization and network design strategies," Connect. Perspect. 143–155 (1989).

[99] Y. LeCun, B. Boser, J.S. Denker, *et al.*, *Backpropagation Applied to Handwritten Zip Code Recognition*, Neural Comput. **1**(4), 541–551 (1989).

[100] A. Krizhevsky, I. Sutskever, and G.E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," Adv. Neural Inf. Process. Syst. 1–9 (2012).

[101] C. Szegedy, Wei Liu, Yangqing Jia, *et al.*, “Going deeper with convolutions,” in *2015 IEEE Conf. Comput. Vis. Pattern Recognit.*(IEEE, 2015), pp. 1–9.

[102] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” Int. Conf. Learn. Represent. 1–14 (2015).

[103] K. He, X. Zhang, S. Ren, and J. Sun, “Identity mappings in deep residual networks,” Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics) **9908 LNCS**, 630–645 (2016).

[104] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” arXiv 1–12 (2015).

[105] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, “ENet: A Deep Neural Network Architecture for Real-Time Semantic Segmentation,” arXiv 1–10 (2016).

[106] A. Krizhevsky, Learning Multiple Layers of Features from Tiny Images (2009).

[107] P.-P. Ypsilantis, M. Siddique, H.-M. Sohn, *et al.*, “Predicting Response to Neoadjuvant Chemotherapy with PET Imaging Using Convolutional Neural Networks.,” PLoS One **10**(9), e0137036 (2015).

[108] F. Ciompi, K. Chung, S.J. van Riel, *et al.*, “Towards automatic pulmonary nodule management in lung cancer screening with deep learning,” arXiv cs.CV **10**(March), 09157 (2016).

[109] Stanford Vision Labs, *ImageNet*, (2016).

[110] S.G. Armato, K. Drukker, F. Li, *et al.*, “LUNGx Challenge for computerized lung nodule classification,” J. Med. Imaging **3**(4), 044506 (2016).

[111] K. Sirinukunwattana, S.E.A. Raza, Y.W. Tsang, D.R.J. Snead, I.A. Cree, and N.M. Rajpoot, “Locality Sensitive Deep Learning for Detection and Classification of Nuclei in Routine Colon Cancer Histology Images,” IEEE Trans. Med. Imaging **35**(5), 1196–1206 (2016).

[112] L. Wei, Y. Yang, R.M. Nishikawa, and Y. Jiang, “A study on several machine-learning methods for classification of malignant and benign clustered microcalcifications,” IEEE Trans. Med. Imaging **24**(2), 371–380 (2005).

[113] W. Shen, M. Zhou, F. Yang, *et al.*, “Multi-crop Convolutional Neural Networks for lung nodule malignancy suspiciousness classification,” Pattern Recognit. **61**, 663–673 (2017).

[114] B. Microbiana, D. Hidalgo, M. Anthimopoilos, *et al.*, “Lung Pattern Classification for Interstitial Lung Diseases Using a Deep Convolutional Neural Network,” IEEE Trans. Med. Imaging **35**(5), 1207–1216 (2016).

[115] Q. Li, W. Cai, X. Wang, Y. Zhou, D.D. Feng, and M. Chen, “Medical image classification with convolutional neural network,” 2014 13th Int. Conf. Control Autom. Robot. Vis. **2014**(December), 844–848 (2014).

[116] M. Gao, U. Bagci, L. Lu, *et al.*, “Holistic classification of CT attenuation patterns for interstitial lung diseases via deep convolutional neural networks,” Comput. Methods Biomech. Biomed. Eng. Imaging Vis. (Cidi), 1–6 (2016).

[117] E.I. Zacharaki, S. Wang, S. Chawla, *et al.*, “Classification of brain tumor type and grade using MRI texture and shape in a machine learning scheme,” Magn. Reson. Med. **62**(6), 1609–1618 (2009).

[118] D.C. Ciresan, A. Giusti, L.M. Gambardella, and J. Schmidhuber, “Deep Neural Networks Segment Neuronal Membranes in Electron Microscopy Images,” Nips 1–9 (2012).

[119] F. Milletari, N. Navab, and S.A. Ahmadi, “V-Net: Fully convolutional neural networks for volumetric medical image segmentation,” Proc. - 2016 4th Int. Conf. 3D Vision, 3DV 2016 565–571 (2016).

[120] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation,” in *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, edited by N. Navab, J. Hornegger, W.M. Wells and A.F. Frangi (Springer International Publishing, Cham, 2015), pp. 234–241.

[121] T. Clark, A. Wong, M.A. Haider, and F. Khalvati, “Fully Deep Convolutional Neural Networks for Segmentation of the Prostate Gland in Diffusion-Weighted MR Images,” in *Image Anal. Recognition.*, edited by F. Karray, A. Campilho and F. Cheriet (Springer International Publishing, Cham, 2017), pp. 97–104.

[122] P. Moeskops, M.A. Viergever, A.M. Mendrik, L.S. De Vries, M.J.N.L. Benders, and I. Isgum, “Automatic Segmentation of MR Brain Images with a Convolutional Neural Network,” IEEE Trans. Med. Imaging **35**(5), 1252–1261 (2016).

[123] M. Havaei, A. Davy, D. Warde-Farley, *et al.*, “Brain tumor segmentation with Deep Neural Networks,” Med. Image Anal. **35**, 18–31 (2017).

[124] S. Pereira, A. Pinto, V. Alves, and C.A. Silva, “Brain Tumor Segmentation Using Convolutional Neural Networks in MRI Images,” IEEE Trans. Med. Imaging **35**(5), 1240–1251 (2016).

[125] A. Prasoon and E. Al., “Deep Feature Learning for Knee Cartilage Segmentation Using a Triplanar Convolution Neural Network,” Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics) **8150**(PART 2), 599–606 (2013).

[126] J. Lieman-Sifry, M. Le, F. Lau, S. Sall, and D. Golden, “Fastventricle: Cardiac segmentation with ENet,” Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics) **10263 LNCS**, 127–138 (2017).

[127] A.A. Novikov, D. Lenis, D. Major, J. Hladůvka, M. Wimmer, and K. Bühler, “Fully Convolutional Architectures for Multi-Class Segmentation in Chest Radiographs,” arXiv 1–9 (2017).

[128] P. Moeskops, J.M. Wolterink, B.H.M. van der Velden, *et al.*, “Deep Learning for Multi-task Medical Image Segmentation in Multiple Modalities,” in *MICCAI 2011 14th Int. Conf. (Vision, Pattern Recognition, Graph.*, edited by S. Ourselin, L. Joskowicz, M.R. Sabuncu, G. Unal and W. Wells (Springer International Publishing, Cham, 2016), pp. 478–486.

[129] O.J. Dunn, “Multiple Comparisons Among Means,” J. Am. Stat. Assoc. **56**(293), 52–64 (1961).

[130] S. Holm, “A Simple Sequentially Rejective Multiple Test Procedure,” Scand. J. Stat. **6**(2), 65–70 (1979).

[131] W.R. Rice, “Analyzing Tables of Statistical Tests,” Evolution (N. Y). **43**(1), 223–225 (1989).

[132] Y. Benjamini and D. Yekutieli, “The control of the false discovery rate in multiple testing under dependency,” Ann. Stat. **29**(4), 1165–1188 (2001).

[133] Y. Benjamini and Y. Hochberg, *Controlling the false discovery rate: a practical and powerful approach to multiple testing*, J. R. Stat. Soc. **57**(1), 289–300 (1995).

[134] L.I. Lin, “A Concordance Correlation-Coefficient to Evaluate Reproducibility,” Biometrics **45**(1), 255–268 (1989).

[135] G. Atkinson and A. Nevill, “Comment on the Use of Concordance Correlation to Assess the Agreement Between Two Variables,” Biometrics **53**(2), 775–777 (1997).

[136] L. I-Kuei and V. Chinchilli, “Rejoinder to the Letter to the Editor from Atkinson and Nevill,” Biometrics **53**(2), 777–778 (1997).

[137] X. Fave, D. Mackin, J. Yang, *et al.*, “Can radiomics features be reproducibly measured from CBCT images for patients with non-small cell lung cancer?,” Med. Phys. **42**(12), 6784–6797 (2015).

[138] O. Grove, A.E. Berglund, M.B. Schabath, *et al.*, “Quantitative Computed Tomographic Descriptors Associate Tumor Shape Complexity and Intratumor Heterogeneity with Prognosis in Lung Adenocarcinoma,” PLoS One **10**(3), e0118261 (2015).

[139] M.-C. DESSEROIT, F. TIXIER, W.A. Weber, *et al.*, “Reliability of PET/CT shape and heterogeneity features in functional and morphological components of Non-Small Cell Lung Cancer tumors: a repeatability analysis in a prospective multi-center cohort,” J. Nucl. Med. (2016).

[140] Y. Chen and F.Y. Yang, “Research on Characteristic Properties of Gray Level Co-Occurrence Matrix,” Appl. Mech. Mater. **204–208**, 4755–4759 (2012).

[141] B. Belaroussi, J. Milles, S. Carme, Y.M. Zhu, and H. Benoit-Cattin, “Intensity non-uniformity correction in MRI: Existing methods and their validation,” Med. Image Anal. **10**(2), 234–246 (2006).

[142] N.J. Tustison and J.C. Gee, “N4ITK: Nick’s N3 ITK Implementation For MRI Bias Field Correction,” InsightJournal 1–8 (2009).

[143] N.J. Tustison, B.B. Avants, P.A. Cook, *et al.*, “N4ITK: Improved N3 Bias Correction,” IEEE Trans. Med. Imaging **29**(6), 1310–1320 (2010).

[144] G. McBride, “A proposal for strength-of-agreement criteria for Lin’s Concordance Correlation Coefficient,” NIWA Client Rep. **HAM2005-06**, 14 (2005).

[145] J.L. Carrasco, T.S. King, and V.M. Chinchilli, “The concordance correlation coefficient for repeated measures estimated by variance components.,” J. Biopharm. Stat. **19**(1), 90–105 (2009).

[146] C.-C. Chen and H.X. Barnhart, “Assessing agreement with intraclass correlation coefficient and concordance correlation coefficient for data with repeated measures,” Comput. Stat. Data Anal. **60**, 132–145 (2013).

[147] Y.-J. Yu-Jen Chen, K.-L. Hua, C.-H. Hsu, W.-H. Cheng, and S.C. Hidayati, “Computer-aided classification of lung nodules on computed tomography images via deep learning technique,” Onco. Targets. Ther. 2015 (2015).

[148] L. Lu, H. Shin, H.R. Roth, *et al.*, “Deep Convolutional Neural Networks for Computer-Aided Detection : CNN Architectures , Dataset Characteristics and Transfer Learning Deep Convolutional Neural Networks for Computer-Aided Detection : CNN Architectures , Dataset Characteristics and Transfer,” IEEE Trans. Med. Imaging **35**(5), 1285–1298 (2016).

[149] A. Cruz-Roa, A. Basavanhally, F. González, *et al.*, “Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks,” in *SPIE*, edited by M.N. Gurcan and A. Madabhushi (2014), p. 904103.

[150] H. Chen, Q. Dou, L. Yu, J. Qin, and P.A. Heng, “VoxResNet: Deep voxelwise residual networks for brain segmentation from 3D MR images,” Neuroimage (April), 1–10 (2017).

[151] C. Wachinger, M. Reuter, and T. Klein, “DeepNAT: Deep convolutional neural network for segmenting neuroanatomy,” Neuroimage **170**(February 2017), 434–445 (2017).

[152] F. Pedregosa, G. Varoquaux, A. Gramfort, *et al.*, “Scikit-learn: Machine Learning in Python,” J. Mach. Learn. Res. **12**, 2825–2830 (2012).

[153] F. Chollet and et.al., *Keras*, (2015). https://keras.io

[154] L. Beyer, Y. Hold, A. Nikkou, and Swehrwein, *pydensecrf*, (2016). https://github.com/lucasb-eyer/pydensecrf

[155] P. Krähenbühl and V. Koltun, “Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials,” in *Adv. Neural Inf. Process. Syst. 24*, edited by J. Shawe-Taylor, R.S. Zemel, P.L. Bartlett, F. Pereira and K.Q. Weinberger (Curran Associates, Inc., 2011), pp. 109–117.

[156] P.F. Christ, M.E.A. Elshaer, F. Ettlinger, *et al.*, “Automatic liver and lesion segmentation in CT using cascaded fully convolutional neural networks and 3D conditional random fields,” Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics) **9901 LNCS**, 415–423 (2016).

Appendix I

Repeatability of Magnetic Resonance Image Derived Texture Features and Use in Predictive Models for Non-Small Cell Lung Cancer Outcome

Title

Repeatability of Magnetic Resonance Image Derived Texture Features and Use in Predictive Models for Non-Small Cell Lung Cancer Outcome

Authors

Mahon, R.N.[1], Hugo G.D.[1,2], Weiss E.[1]

Affiliations

1. Radiation Oncology, Virginia Commonwealth University, Richmond, VA 23298, USA
2. Radiation Oncology, Washington University, St. Louis, MO, 63105. USA

Abstract

Purpose: To evaluate the repeatability of MRI and CT derived texture features and investigate feasibility of use in predictive single and multi-modality models for radiotherapy of non-small cell lung cancer.

Methods: Fifty-nine texture features were extracted from unfiltered and wavelet filtered images. Repeatability of test-retest features from helical 4D CT scans, true fast MRI with steady state precession (TRUFISP), volumetric interpolation breath-hold examination (VIBE), and diffusion weighted MRI (both the acquired images and processed apparent diffusion coefficient (ADC) images) was determined by the concordance correlation coefficient (CCC). A workflow was developed to predict overall survival at 12, 18, and 24 months and tumor response at end of treatment for tumor features, and normal tissue features as a control. Texture features were reduced to repeatable and stable features before clustering. Cluster representative feature

selection was performed by univariate or medoid analysis before model selection. P-values were corrected for false discovery rate.

Results: Repeatable (CCC ≥ 0.9) features were found for both the tumor and normal tissue: CT: 54.4% for tumor and 78.5% for normal tissue, TRUFISP: 64.4% for tumor and 67.8% for normal tissue, VIBE: 52.6% for tumor and 72.9% for normal tissue, DWI: 10.2% for tumor, ADC: 18.6% for tumor. Normal tissue control analysis found 7 significant models with 6 of 7 models utilizing the univariate representative feature selection technique. Tumor analysis revealed 12 significant models for overall survival and 0 for tumor response at end of treatment. The accuracy of significant single modality was about the same for MR and CT. Multi-modality tumor models had comparable performance to single modality models.

Conclusion: MR derived texture features may add value to predictive models and should be investigated in a larger patient cohort. Control analysis demonstrated the medoid representative feature selection method may result in more robust models.

Keywords

Radiomics; magnetic resonance imaging; computed tomography; texture features; repeatability; predictive modeling; lung cancer

Introduction

In recent years, predictive models built on imaging features have been investigated as biomarkers for various clinical endpoints and anatomical sites. This research area, termed

'radiomics', seeks to extract a large number of pre-determined features from an image and use these features to phenotypically characterize the tissue in question. Imaging features that quantitatively describe tissue characteristics such as texture and shape are used in various models to predict treatment response. In addition, a sub area of radiomics, delta radiomics, seeks to extract information from the changes in the extracted texture features as a result of treatment-induced changes.[3–5] Both single time point, such as pre-treatment, and delta radiomics features have been used to explore correlations between the extracted features and histology[6], gene mutation[5, 7], local control[8–11], distant metastasis, and overall survival[3, 12–15].[3, 16–18] Utilizing the information contained in medical images is attractive for two reasons: it is non-invasive and gives information of the whole tumor volume unlike a biopsy, and can often be extracted from images obtained for routine patient management purposes. Radiomics for lung cancer has primarily focused on computed tomography (CT) and positron emission tomography (PET) images.[4, 5, 9, 19–22] Magnetic Resonance (MR) images, on the other hand, have been less prominently studied for use in lung cancer, partly due to the high variability in the extracted features due to variations in imaging protocol and acquisition signal.[23–31] However, the superior soft tissue contrast and lack of additional dose to a patient treatment make MR a potentially useful imaging modality for lung cancer radiomics. In addition, with the introduction of MR simulation and use of MR for image guidance, MR images may become more common in the radiation treatment workflow which makes the exploration of their potential desirable.

This work aims to first characterize the repeatability of MR texture features of lung cancer extracted from a single machine and imaging protocol and compare them to CT texture features extracted with the same workflow. Then, for pretreatment CT and MR images, utilize the repeatable features to assess the feasibility of predicting local control and overall survival for each modality individually and for a set of multi-modality models. For validation of the selected models in lung cancer and to serve as a control, predictive models were also constructed for

normal tissue features which were expected not to change in response to radiation treatment or predict treatment outcome for the same endpoints.

Methods

Patient characteristics

This study utilized images of 15 subjects with non-small cell lung cancer enrolled on an IRB approved study. Diffusion weighted and morphological MRI and CT images before and during the course of radiotherapy were acquired in order to investigate the potential use of MRI in lung cancer treatment planning and response evaluation. All subjects underwent radiation therapy with or without concurrent chemotherapy for stage IIB to IV non-small cell lung cancer per department protocol. See Table 1 for summary of patient characteristics.

Table 1: Summary of patient characteristics for texture feature repeatability study.

Sex	
Male	10
Female	5
Mean Age (Range)	
59.1 (50.0-73.4) years	
Histology	
Squamous cell carcinoma	9
Adenocarcinoma	6
Stage	
IIB	3
IIIA	6
IIIB	4
IV	2
Chemotherapy	
Yes	12
No	3
Mean Dose (Range)	
61.6 (59.4-66) Gy	

Images

Five different imaging types were evaluated in this study: T1-weighted Volumetric Interpolation Breath-Hold Examination (VIBE), True fast MRI with steady state precession (TRUFISP), diffusion weighted MRI (both the acquired images and processed apparent diffusion coefficient (ADC) images), and helical 4D CT scans. The ADC maps were created utilizing 8 b-value DWI images (b-values between 0 and 1000 s/mm^2). MR images were all acquired on a 1.5 T scanner (Avanto, Siemens, Munich, Germany), CT images were acquired as 4D images (Brilliance Big Bore, Philips, Amsterdam, Netherlands). Table 2 details the different imaging parameters used for each modality.

Sets of CT and MR images were acquired on the same day before treatment ("Time 1"),

Table 2: Summary of imaging parameters for each modality.

	CT	TRUFISP	VIBE	DWI
Machine	Phillips Brilliance Big Bore	Siemens Avanto 1.5 T	Siemens Avanto 1.5 T	Siemens Avanto 1.5 T
Contrast	No	No	No	No
Pixel Spacing	0.98mm to 1.37mm	0.74mm	1.57mm	2mm
Slice Thickness	3mm	5mm	1.6mm or 2mm	4mm or 6mm
Gap	3mm	6.4mm	0	4.8mm to 9.6mm
In-plane Matrix Size	512 x 512	512 x 512	208 x 256	144 x 192
Breathing Regulation	Free Breathing	Breath Hold	Breath Hold	Respiration Triggered
kVp	120 or 140	NA	NA	NA
Coil	NA	Body	Body	Body
Flip angel	NA	57 or 66 degrees	12 degrees	90 degrees
TR	NA	3.65ms	3.56ms	Various
TE	NA	1.82ms to 1.89ms	1.28ms	74
Number of Echos	NA	1	1	1
Echo Train length	NA	1	1	1
Number of Averages	NA	1	1	2
b-values	NA	NA	NA	0, 50, 100, 250, 500, 650, 800, and 1000 mm/s^2

approximately three weeks into treatment ("Time 2") and/or at the conclusion of treatment ("Time 3"). To evaluate the repeatability of texture features extracted from the various imaging modalities, test-retest image sets were acquired utilizing the same protocol with a short break ("coffee break") in between scans without repositioning the patients. Images were acquired utilizing 4D acquisition for the CT images, breath hold for the TRUFISP and VIBE images and respiration triggering for the DWI images. The short time interval between the two image sets allows for evaluation of random changes in a texture feature as it can be assumed that there is no true physiological change in the tumor between the two images. The 4D CT images were comprised of 33 pairs of inspiration and expiration phase images. While the TRUFISP and VIBE images totaled 34 and 32 same session inspiration-expiration breath-hold image pairs respectively. For the diffusion weighted and corresponding ADC maps, there were 8 imaging pairs with slice thickness difference of 4mm or 6mm and 16 ascending or interleaved acquisition pattern image pairs for the repeatability analysis referred to as "Thickness" or "Order" respectively. Table 3 outlines the test – retest image pairs available for all subjects in the study.

Region of Interest

The primary gross tumor volume was delineated as the region of interest (ROI) by one experienced radiation oncologist on one of the images in the image pair using clinical segmentation and registration software (MIM Maestro v6.X, Cleveland, OH). The test and retest images were rigidly registered together, and the contour was transferred to the corresponding image and manually adjusted as necessary to ensure the volumes had a no greater than 10% difference and visually defined the tumor. The normal tissue contours used to build a control model consisted of a cylindrical volume of 5.75 mL, sampled from an erector spinae or infraspinatus muscle outside the radiotherapy fields depending on primary tumor location. In addition to the muscle contours, blood contoured inside the descending aorta and air contoured inside the trachea were also tested.

Table 3: Summary of test-retest image pairs available for all subjects in the study.

	Time 1	Time 2	Time 3
CT	15	9	9
TRUFISP	15	10	9
VIBE	16	9	7
DWI_Thickness	4	2	2
DWI_Order	4	6	6
ADC_Thickness	4	2	2
ADC_Order	4	6	6

"Thickness" refers to different slice thicknesses between the test and re-test images and "Order" refers to different slice acquisition orders.

Texture Features

Fifty-nine texture features were extracted, Table 4, at five different wavelet ratios, 0.5, 0.67, 1, 1.5, and 2, utilizing the Radiomics Matlab code (2016b, MathWorks, Natick, MA) by Vallières et al. which was extended in-house to include additional texture features, quantization by the gray level number, and to increase speed of code.[16] Prior to feature extraction, all images were

Table 4: List of all 59 texture features extracted from images.

Histogram	GLCM	NGTDM	GLRLM	GLSZM
Variance	Energy	Coarseness	Short Run Emphasis (SRE)	Small Zone Emphasis (SZE)
Skewness	Contrast	Contrast (ContrastN)	Long Rum Emphasis (LRE)	Large Zone Emphasis (LZE)
Kurtosis	Entropy	Busyness	Gray Level Non-Uniformity (GLN)	Gray Level Non-Uniformity (GLNS)
Standard Deviation (SD)	Homogeneity	Complexity	Run Length Non-Uniformity (RLN)	Zone Size Non-Uniformity (ZSN)
Mean	Correlation	Strength	Run Percentage (RP)	Zone Percentage (ZP)
Minimum	Sum Average		Low Gray Level Run Emphasis (LGRE)	Low Gray Level Zone Emphasis (LGZE)
Maximum	Variance (VarianceG)		High Gray Level Run Emphasis (HGRE)	High Gray Level Zone Emphasis (HGZE)
Median	Dissimilarity		Short Run Low Gray Level Emphasis (SRLGE)	Small Zone Low Gray Level Emphasis (SZLGE)
Quartile1	Mean Pair Sum (MeanPS)		Short Run High Gray Level Emphasis (SRHGE)	Small Zone High Gray Level Emphasis (SZHGE)
Quartile2	Variance Pair Sum (VariancePS)		Long Run Low Gray Level Emphasis (LRLGE)	Large Zone Low Gray Level Emphasis (LZLGE)
	Entropy Pair Sum (EntropyPS)		Long Run High Gray Level Emphasis (LRHGE)	Large Zone High Gray Level Emphasis (LZHGE)
	Variance Pair Difference (VariancePS)		Gray Level Variance (GLV)	Gray Level Variance (GLVS)
	Entropy Pair Difference (EntropyPD)		Run Length variance (RLV)	Zone Size Variance (ZSV)
	Information Correlation Measure 1 (InfoCorr1)			
	Information Correlation Measure 2 (InfoCorr2)			
	Auto-Correlation (AutoCorr)			
	Cluster Prominence (ClusterProm)			
	ClusterShade			

Texture catagories are: Grey Level Co-Ocurrence Matrix (GLCM), the Neighborhood Gray Tone Difference Matirx (NGTDM), Gray Level Run Length Maritx (GLRLM) and Gray level Size Zone Matrix (GLSZM)

resampled to the in-plane voxel size, and for the MRI images, any voxels with intensity greater than +/- 3σ from the mean were removed as suggested by Collewet et al. for greater feature stability.[24] The gray level values were used as the bins for quantization resulting in a bin width of 1HU.

The extracted texture features were from five different texture feature categories: the intensity histogram, the gray level occurrence matrix (GLCM) with features calculated as described by Haralick et al.[32] and extended by Conners et al.[33], the neighborhood gray tone difference matrix (NGTDM) with features calculated as described by Amadasun and King[34], the gray level run length matrix (GLRLM) with feature described by Galloway[35], and the gray level size zone difference matrix (GLSZM) with feature calculated as described by Thibault et al.[36] The different

texture categories capture first order texture features through the intensity histogram as well as second order features which encode information about the spatial distribution of the voxels through the other categories. The higher order features were derived from the GLCM, which measures the co-occurrence of different voxel intensities in a defined distance and direction away, the GLRLM, which measures characteristics of the distribution of connected isointense voxels in a given direction, the GLSZM, which extends the GLRLM feature to connected isointense voxels in all direction, and the NGTDM, which characterizes the difference between a voxel and its neighbors. The GLCM and GLRLM feature were averaged across all directions to make features rotationally robust. For all image sets except the DWI/ADC images, 3D texture features were calculated. For several patients, the large tumor volumes were not completely covered by the DWI protocol. Therefore, multiple overlapping acquisitions were used to cover the entire tumor volume as recommended in the imaging protocol. In this situation, because of slice to slice intensity variations, a surrogate 3D texture feature value was calculated for the DWI/ADC images by using a volumetrically weighted average of the texture features calculated from non-overlapping 2D slices.

Repeatability Analysis

An analysis of the repeatability of texture features extracted from each image type was performed with R (3.3.1) in R Studio (1.0.143, RStudios Inc., Boston, MA). First, repeatability of texture features was assessed for all image types at all 3 time points by using the concordance correlation coefficient (CCC) as described by Lin et. al. utilizing all available time points for each image type.[37] The CCC measures the correlation between two paired measurements by calculating the deviance from perfect one-to-one correlation.

Outcomes Modeling

Texture features were evaluated for predictive power in logistic regression models for local control and overall survival. Prior to model selection, the features were first reduced to repeatable and stable features before clustering. After clustering, a representative feature from each cluster was selected as a candidate for modeling. The "best" models were selected based on accuracy and fit according to each clinical endpoint. The workflow process can be seen in Figure 1 and is described in greater detail below.

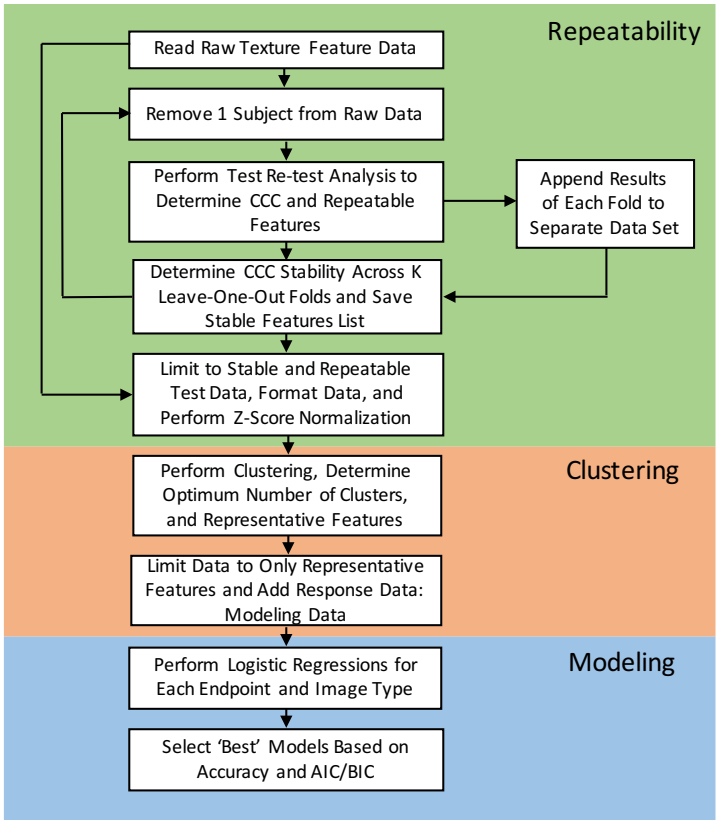


Figure 1: Workflow for model selection.

Feature Reduction

Repeatability was again computed as part of the model selection workflow using a leave-one-out cross validation of only the pretreatment images in order to select stable features and increase robustness of the models. During model selection, a feature was deemed repeatable if the CCC value exceeded a cut off of 0.9 as recommended by McBride.[38] If the 0.9 cutoff failed to produce more than 5 repeatable texture features, the cutoff was lowered in 0.05 steps until at least 5 features were found with 0.7 being the minimum allowed cutoff to compensate for having less than the 25 recommended samples. The stability of the CCC calculation was determined by the coefficient of variation (COV) across all folds. Texture features with a COV less than 5% were considered stable repeatable features and included in the feature reduction and model selection process.

Representative Feature Selection

A z-score normalization was then applied to the subset of repeatable and stable features before undergoing hierarchical clustering based on the absolute value of the Spearman distance using the Cluster Consensus Plus package (1.38.0, open source, Bioconductor.org) in R with an 85% subject resample and 1000 iterations.[39] The optimum number of clusters was determined by identifying the highest median cluster consensus from the range of cluster numbers k=[5,7]. The upper bound was selected in order to have twice as many pretreatment subjects images as texture features in the model as is generally recommended in regression, and the lower bound selected to be the minimum number of clusters where the relative change in the area under the cumulative density function curve, which is the 'Delta Area' plot returned as part of the consensus cluster plus package output, appeared to become stable.[39]

A representative feature was selected from each of the k clusters of the optimum number previously established by one of two methods: medoid or univariate prognostic power. With the

medoid selection process, the Spearman correlation among cluster members was calculated and the feature with the highest average correlation was selected.[20, 40] For the univariate prognostic power selection, a log likelihood ratio test was performed comparing the model with each individual feature to the model containing only the intercept, and the feature with the most significant log likelihood p-value was selected as the representative feature.[3]

Model Selection

Following feature reduction and representative feature selection, an exhaustive search of all possible models was performed for the CT, VIBE and TRUFISP pre-treatment images only. The sample size of the remaining image types and time points was too small for consideration. The logistf package[41] in R was used to regress the selected variables to four different endpoints: Overall survival at 12 (OS_12), 18 (OS_18) and 24 (OS_24) months and tumor response at end of treatment (PT_endTx). For the OS_12, OS_18 and PT_endTx there were 14 available subjects. One subject was lost to follow up after 20 months and was not included in the OS_24 dataset resulting in a total of 13 subjects' data being available for modeling at this time point. The maximum likelihood equation was penalized using the Firth method[42] to reduce small sample bias, and the "best" models were selected based on a combination of high leave-one-out cross validation accuracy, low Akaike information criterion (AIC), and low Bayesian information criterion (BIC). One model for each outcome, image type, and representative feature selection method was selected and compared resulting in a total of 48 models. The selection methods and image processing methods were compared on the basis of corrected significance, and accuracy to determine the "best" 4 models for each image type. For significance testing, the log likelihood ratio test was used to compare the selected model to the model containing only the intercept. The log likelihood ratio test p-values were corrected using the Benjamini-Hochberg-Yekutieli (BHY) procedure[43] to control for multiple dependent comparisons with an acceptable false discovery rate of 0.05.

The single modality and muscle studies utilized all available pretreatment images, while the multi-modality study utilized the subset of 9 patients with all three pre-treatment images. The multi-modality data set was created by concatenating all the repeatable and stable features from each of the three modalities into one data set then proceeding with clustering, representative feature determination, and model selection. For comparison, the single modality clustering, representative feature determination and model selection process was repeated for the same subset of 9 patients. In the case of the single modality utilizing the full image set and muscle studies, only the best model was selected, while for the multi-modality and comparison single modality image study, the top 3 models were compared for trends due to the sample size.

Results

Repeatability

Three different tissue types were investigated for the normal tissue control, blood within the descending aorta, air within the trachea and muscle. The blood and air were not used as repeatability was difficult to establish using the CCC calculation as small difference between test and re-test value caused a large drop in CCC value with the narrow range of values in the population. Analysis for the normal tissue control modeling was completed on texture features derived from the muscle contours and revealed a number of repeatable and stable features. For the wavelet with the highest number of repeatable features, each of the modalities achieved the following results: CT: 55.9% of features were considered highly repeatable (CCC ≥ 0.95) and an additional 18.6% were considered repeatable (CCC ≥ 0.9), TRUFISP: 59.3% highly repeatable features and an addition 8.5% repeatable features, VIBE: 47.5% highly repeatable

features and an additional 25.4% repeatable features. The DWI and ADC images were not tested for the normal tissue.

The CCC score for the top 25 repeatable primary tumor features from all time points for each image type can been seen in Table 5. The b-value of 650 s/mm^2 for the diffusion weighted images was selected because it had the most repeatable features for the images without perfusion contamination for both the DWI_Order and DWI_Thickness image sets. For the wavelet with the highest number of repeatable features, each of the modalities achieved the following results: CT: 39% of features were considered highly repeatable (CCC ≥ 0.95) and an additional 24.4% were considered repeatable (CCC ≥ 0.9), TRUFISP: 16.9% highly repeatable features and an addition 47.5% repeatable features, VIBE: 3.4% highly repeatable features and an additional 49.2% repeatable features, DWI_Order: 10.2% repeatable features, ADC_Order: 18.6% repeatable features, DWI_ Thickness and ADC_Thickness: 0% repeatable features. There were several features that were repeatable across multiple modalities as can be seen in Table 5. In addition, texture features were found to be repeatable across the majority of the different wavelet filtered images in 97.5% for CT images, 76.1% for TRUFISP images, 84.8% for the VIBE images, 28.6% of the DWI_Order image, and 100% in the ADC_Order image. The percentage of repeatable texture features from each category (histogram, GLCM,…) was approximately the same within each imaging modality. For the CT image all texture categories had a percentage of repeatable features between 60% and 80%, for TRUFISP the majority of texture feature categories were between 83%-100% with the exception of histogram features where only 40% were repeatable at any wavelet ratio, and for VIBE all texture categories were between 46% and 62% repeatable. For the DWI_Order, the GLRLM and GLSZM were both at 8% and all other categories at 0% and for the ADC_Order, the GLCM, GLRLM and GLSZM were between 15% and 23% while the NGTDM had the most at 40% and the histogram feature the least at 0%. There were no repeatable features for the DWI_Thickness or ADC_Thickness.

Model Selection

Preliminary analysis of feature clusters revealed that individual texture features from different wavelets were in the same clusters regardless of image type. Therefore, only the texture feature values from the wavelet ratio 1, the unfiltered image, denoted “image_1” or “feature_1”, and texture feature values averaged across all wavelet ratios, denoted “image_avj” or “feature_avj”, were considered. During feature reduction, the minimum repeatability cutoffs used for any leave one out fold for the primary tumors were: 0.9 for CT_avj and CT_1, 0.85 for VIBE_avj, VIBE_1, and TRUFISP_avj, and 0.8 for TRUFISP_1. The minimum threshold for repeatability in any leave one out fold for the muscle was 0.9 for all image types.

Regarding the normal tissue contours, for the VIBE_avj , VIBE_1 and CT_avj there were 32 features available for clustering, TRUFISP_avj had 35, TRUFISP_1 had 29, and CT_1 had 25. The optimum number of clusters as determined by the highest median cluster consensus was 7 for CT_1 and VIBE_avj, 6 for the CT_avj and TRUFISP_avj, and 5 for the VIBE_1, TRUFISP_1 images. For CT, no models were found to be significant predictors of tumor outcome. For TRUFISP 6 of 16 models and 14/16 models for VIBE were found to be significant predictors of tumor outcome. The primary tumor had a similar number of repeatable and stable features for clustering and optimum cluster number as the normal tissue. The number of repeatable and stable features remaining for clustering for the VIBE_avj was 41, VIBE_1 was 34, TRUFISP_avj was 25, TRUFISP_1 was 30, CT_avj was 32, and CT_1 was 35. The optimum number of clusters as determined by the highest median cluster consensus was 7 for the VIBE_avj, TRUFISP_avj, and TRUFISP_1, 6 for the CT_avj and CT_1, and 5 for the VIBE_avj images.

For the CT images, each filtering and feature selection method produced the same significant model. The selected single modality model found to be significant by the BHY procedure was

for OS_12, with an accuracy of 0.93 ± 0.13The only significant models based on the TRUFISP images were derived from the unfiltered image with medoid feature selection. The significant models from the BHY procedure were: OS_12 with accuracy 0.93 ± 0.13, and OS_18 with accuracy 0.87 ± 0.18. OS_24 and PT_endTx were not determined to be significant. The VIBE significant models were derived from all four combinations of filtering and representative feature selection processes with the univariate selection method producing more significant models than the medoid method PT_endTx was not predicted significantly by any of the tumor models. The highest accuracy achieved for the OS_12 was 0.93± 0.13, the OS_18 was 0.93 ± 0.13, the OS_24 was 0.86 ± 0.22. The accuracy of the significant models for the muscle and primary tumor can be seen in Figure 4.

Figure 3 depicts the frequency of texture features occurring within the significant models. It can be seen that while the muscle texture features did produce significant models very few overlapped with texture features selected in the best primary tumor models.

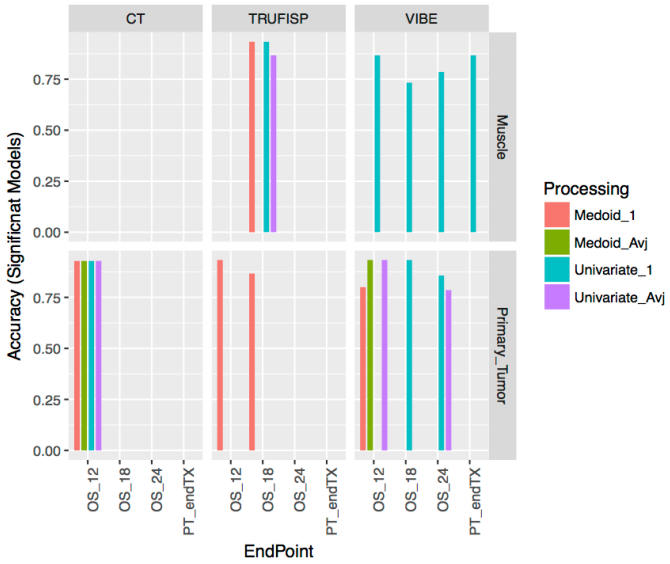


Figure 4: Accuracy of significant models for primary tumor and muscle by modality and image filtering/representative feature selection technique.

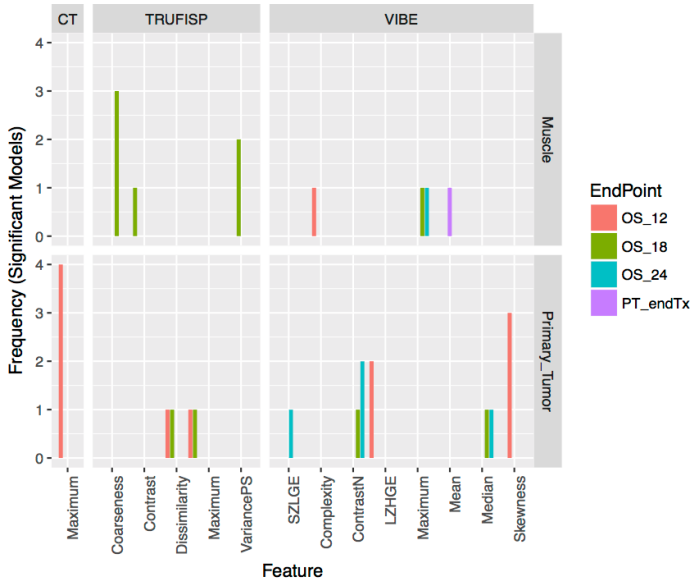


Figure 3: Comparison of texture features utilized within significant models derived from both the primary tumor and muscle tissue by modality and endpoint.

Following the analysis of the normal tissue and single modalities models, a multi-modality model selection and reduced subject single modality, referred to as reduced_SM, model selection was performed. Multi-modal and reduced_SM model selection was performed for the PT_endTx, OS_12 and OS_24 only as the responses for the reduced patient set for OS_24 and OS_18 end points were identical. In addition, the number of clusters was set to 5 due to the number of patient remaining. None of the models were significant under the BHY procedure for any of the reduced_SM or multi-modality models.

The top 3 multi-modality models had features from all three modalities represented. The TRUFISP derived features appeared in 26 of the 36 models more often than the VIBE (11/36) and CT (12/36) features. The average accuracy of the top 3 models was comparable or more accurate than the reduced_SM models. For the OS_12 and OS_24, the multi-modality average accuracy by wavelet filtering and representative feature selection ranged from 67% to 85%

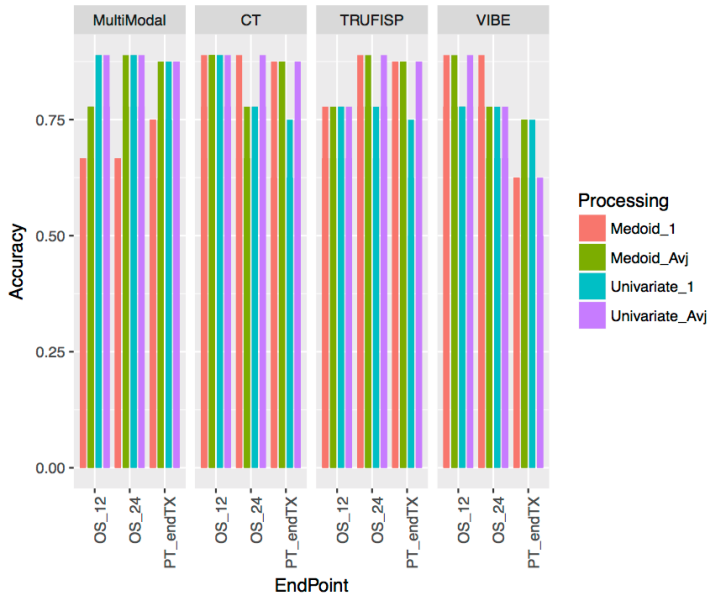


Figure 5: Highest model accuracy by endpoint and modality for the primary tumor.

while the corresponding single modalities ranged in accuracy from 72% to 81%. For the PT_endTx, the multi-modality accuracy had a range of 71%-88% and the corresponding single modalities ranged from 54%-88%. Compared to CT only the OS_12 and OS_24 predictive ranges were 78%-81% and the PT_endTx was 67%-88% suggesting the multi-modal models may add some small accuracy benefits. A comparison of the highest accuracy for top 3 models for all modalities can be seen in Figure 5.

Discussion

Our work evaluated the repeatability of MR derived texture features and their usefulness in predictive models as compared to and in combination with CT derived features and models. We were able to identify repeatable texture features and predictive models for the primary tumor in both the VIBE and TRUFISP image sequences that were significant under the BHY procedure and had promising accuracy. In addition, we demonstrated the feasibility of constructing multi-modality based predictive models that were comparable to the single modality predictive models for the primary tumor. However, since the multi-modality models did not outperform the single modality models, this particular combination of multi-modality imaging does not appear warranted for the purpose of outcome prediction using above described radiomics approach.

Finally, we evaluated the ability of radiomics in out-of-field un-irradiated normal tissue to predict tumor outcome. Because this tissue is minimally affected by radiation and tumor response, we expected to find no significant relationship between texture features in any modality and outcome. The results of the normal tissue portion of this work emphasizes the need for caution as a number of MR based models were found to be significant predictors of outcome, but also demonstrates how such a technique can be used to assist in model development. For example, the use of univariate feature selection from a cluster gave a higher number of spurious results

compared to medoid based feature selection. From these results, one can select the medoid-based method as potentially being more robust, presuming the muscle region has no true ability to predict patient response to therapy. Furthermore, the features from the significant models derived from primary tumor and muscle had no overlap.

By using only one scanner and imaging protocol with manual contours from one physician, we were able to reduce many potentially confounding factors that affect texture feature repeatability. The features we found to be repeatable for the CT images are comparable to findings by Larue et. al.[12]for their 4D CT images. The repeatability of MR features has not been as highly investigated in the literature. One recent study published by Gourtsoyianni et. al.[31] on MR texture feature repeatability showed very low repeatability of higher order texture features (GLRLM, NGTDM, LGSZM) and more repeatable features in the global (histogram) features and GLCM. Some of the features our work found to be repeatable were in agreement with the study by Gourtsoyianni et al.; however, it should be noted that the Gourtsoyianni et al. study focused on T2-weighted turbo spin echo image technique for rectal cancer whereas this work utilized VIBE and TRUFISP image techniques and focused on non-small cell lung cancer.

This study was the first, to our knowledge, to investigate the predictive power of un-irradiated, out-of-field normal tissue for clinical endpoints. The discovery of significant predictive models on MRI derived from muscle tissue throws some doubt on the connection between the radiomics texture features extracted from the tumor and underlying biological processes as it suggests the correlation may be spurious. The lack of significant muscle-based models for the CT images and the reduced number of significant muscle-based models for the VIBE and TRUFISP image derived features, on the other hand, suggests that not all the results may be spurious and that refinement and further exploration of the general approach described here in a larger sample size is needed. It is of interest that the texture features present in the best

models were different for the primary tumor and muscle tissue. In addition, the method of representative feature selection seemed to have a large impact on the significance. Of the 7 significant normal tissue models, 6 were formulated utilizing the univariate representative feature selection process. Several authors[3, 6, 12] have used univariate based feature selection methods to maximize the chance of predictive power, however, this may also increase the risk of finding spurious results. The medoid method, on the other hand, may select more robust features by selecting the most similar feature within a cluster, therefore the authors suggest the use of the medoid representative feature selection method with the workflow presented in this work.

Whereas CT and VIBE based primary tumor models did not seem to favor either the unfiltered or filtered images as the models have about equal accuracy and significance with the texture features averaged across wavelet ratios or unfiltered, TRUFISP tumor models prefer the unfiltered features. Overall, the VIBE images resulted in more repeatable and stable features than the TRUFISP with a minimum threshold and number of features closer to those derived from CT images. In addition, the texture features selected can be related to tumor in-homogeneity such as dissimilarity and contrast, from the NGTDM, that have been identified as correlating with outcome.[45] In addition, when considering the significant normal tissue models as spurious, the pre-treatment images seemed to be a better predictor of overall survival at earlier time points as the significant medoid models were for the OS_12.

Evaluating the primary tumor clusters across the different methods and image modalities, there appeared to be several general themes that could be identified. All of the imaging modalities had a single cluster of features related to tumor homogeneity and another containing coarseness. Another theme present in all but two images were derived from the histogram features. Variance and energy related clusters appeared in CT and TRUFISP images. The final

cluster theme consisted of image specific features that were unique clusters to each image type. These cluster themes suggest the repeatable and stable features seem to capture some similar underlying features of the tumor phenotypes regardless of modality such as homogeneity as well as identify where MR images may capture different subtleties suggesting that MR features may add additional value to predictive models over just using CT features alone. These results were further explored in a limited fashion through the multi-modality portion of this work.

For the multi-modality study, our results showed that the accuracy of the top 3 multi-modality features were comparable to the top 3 reduced_SM derived models since there were no significant models for either reduced_SM or multimodality models with 9 subjects. Multi-modality models may have a small increased accuracy. The TRUFISP image derived features were present in 72% of the top multi-modality models. With comparable performance and predominately TRUFISP based features, the results seem to suggest that the MR models could stand alone for predictive power. However, with the small sample size further testing will be required. Future work should also include information from the routinely acquired pre-treatment PET scans which have been shown to have some added benefit to CT[3, 46] and MR.[16]

Our work did not seek to evaluate robustness of texture features acquired on different imaging machines and locations which would be necessary to establish imaging texture features as biomarkers. One of the main challenges facing radiomics as a whole is establishing a standardized protocol for image acquisition and texture feature extraction.[44] There is a great variability of techniques currently used in radiomics and, so far, there has not been a firm conclusion on recommended procedure or best practices as far as image processing. As a result, the goal of this study was to assess feasibility and potential value for MR texture features while comparing them to CT models derived with the same texture feature extraction procedure.

The main limitation of our study is the small sample size. Due to the challenging nature and burden on patients of collection of multi-modality imaging at multiple time points along with repeatability studies, only a small pilot cohort was available. However, techniques to minimize bias due to the small sample size were employed throughout. The Firth small sample bias penalization to the maximum likelihood values was used during models selection when preforming the log likelihood ratio test against a model containing only the intercept.[41] This procedure adds a small penalty to the maximum likelihood inversely proportional to the sample size. When analyzing the CCC, all time points were considered together to provide a larger picture of the range of the texture features present in the subject population which seemed justified by the overall lack of a significant difference in the texture features when comparing the various time points of the populations under the Wilcoxon Rank test. In addition, clinical factors were not included as model selection variables to allow maximum exploration of the texture features. Studies have shown that adding radiomics features to models with clinical factors have increased the predictive ability of models[3] and future work should explore this possibility with MR features in a larger dataset.

Conclusion

In this study, we measured repeatability of MR and CT texture features and then used these to build models for estimating outcome after radiation therapy for non-small cell lung cancer. The results show that MR images may hold valuable information in addition to the features from CT images and should be investigated further in a larger patient cohort.

Acknowledgements

We would like to thank Dr. Julian Rosenman for valuable discussions related to the use of normal tissue controls. We would also like to thank Dr. Kishor Karki for assistance with image acquisition and valuable discussion related to the imaging sequences used.

This work was supported in part by a research grant from the National Cancer Institute of the National Institutes of Health under award number P30CA016059. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Disclosure of Conflicts of Interest

We disclose the following potential conflicts of interest in the manuscript: Virginia Commonwealth University has a research agreement with Varian Medical Systems. EW and GD receive funding from NIH. EW receives royalties from UpToDate.

Table 5: Table of top 25 repeatable texture features for all image types.

	CT		TRUFISP		VIBE (T1-weighted)		DWI Order (b-value 650)		ADC Order		DWI Thickness (b-value 650)		ADC Thickness	
Wavelet Ratio	0.67	1	0.67	1	0.67	1	0.67	1	0.67	1	0.67	1	0.67	1
EntropyPS	0.951	0.945	0.946	0.934	0.924	0.922	0.855	0.866	0.926	0.927	0.729	0.728		
InfoCorr1	0.960	0.954	0.932	0.940			0.816	0.808	0.827	0.867	0.772	0.773	0.834	0.827
Coarseness	0.966	0.966	0.968	0.967	0.957	0.941	0.846	0.802			0.731	0.722		
Entropy			0.955	0.946	0.932	0.929	0.879	0.898	0.931	0.931			0.794	0.796
EntropyPD	0.956	0.956			0.934	0.928	0.714	0.735	0.869	0.879			0.765	0.744
Median	0.959	0.960			0.914	0.916			0.852	0.852	0.806	0.806	0.713	0.713
Quartile3			0.930	0.929	0.922	0.921			0.869	0.869	0.809	0.809	0.738	0.738
VariancePS	0.952	0.950	0.962	0.961					0.847	0.847	0.781	0.781	0.740	0.740
AutoCorr			0.941	0.941					0.840	0.840	0.741	0.738	0.785	0.785
Complexity	0.954	0.946	0.956	0.955					0.912	0.914			0.819	0.820
Dissimilarity	0.962	0.959			0.926	0.921			0.837	0.837			0.801	0.801
Energy			0.962	0.958	0.918	0.921	0.864	0.876	0.917	0.921				
GLN			0.956	0.938	0.934	0.935	0.889	0.890	0.921	0.925				
GLNS			0.951	0.943	0.937	0.936	0.902	0.905	0.921	0.925				
Homogeneity	0.943	0.947			0.923	0.918			0.870	0.864			0.776	0.772
LGRE	0.970	0.969	0.913	0.931			0.904	0.895	0.900	0.902				
LGZE	0.970	0.969	0.920	0.936			0.902	0.874	0.926	0.927				
Mean	0.959	0.959			0.911	0.912			0.847	0.846	0.800	0.802		
SD	0.947	0.946	0.956	0.953							0.758	0.755	0.782	0.782
Skewness					0.929	0.932	0.870	0.869			0.789	0.794	0.769	0.769
SRHGE			0.942	0.942					0.834	0.834	0.740	0.737	0.787	0.787
SRLGE	0.970	0.969	0.913	0.931			0.906	0.899	0.912	0.914				
SZHGE			0.948	0.950					0.834	0.834	0.733	0.726	0.789	0.789
SZLGE	0.970	0.969	0.920	0.936			0.905	0.884	0.904	0.901				
Variance	0.951	0.949	0.962	0.961							0.773	0.769	0.764	0.764
Busyness					0.909	0.949	0.768	0.782	0.905	0.914				
ClusterProm			0.951	0.947					0.834	0.834	0.823	0.824		
ClusterShade	0.952	0.948			0.919	0.919					0.835	0.834		
HGRE			0.941	0.942							0.745	0.741	0.787	0.787
HGZE			0.944	0.945							0.747	0.742	0.788	0.788
LRHGE			0.939	0.939							0.755	0.759	0.785	0.785
Maximum									0.850	0.850	0.807	0.807	0.750	0.750
SumAverage					0.933	0.932	0.837	0.840					0.741	0.741
SZE	0.941	0.948	0.941	0.937	0.930	0.921								
VarianceG	0.963	0.961					0.828	0.834	0.881	0.882				
ZSN	0.945	0.952	0.941	0.938	0.931	0.924								
Contrast	0.954	0.952											0.809	0.809
ContrastN	0.955	0.960			0.939	0.926								
Correlation							0.887	0.890			0.765	0.765		
Kurtosis							0.852	0.839					0.729	0.729
LRLGE	0.970	0.969					0.893	0.877						
LZHGE											0.684	0.771	0.780	0.781
MeanPS											0.715	0.714	0.801	0.801
Quartile1	0.959	0.961									0.789	0.789		
RLV							0.912	0.895			0.626	0.688		
VariancePD	0.950	0.951											0.826	0.826
GLV							0.771	0.732						
GLVS							0.801	0.796						
LRE					0.912	0.917								
LZE					0.932	0.953								
LZLGE							0.743	0.795						
Minimum							0.692	0.692						
RLN					0.914	0.917								
RP					0.913	0.917								
SRE					0.914	0.917								
Strength											0.827	0.817		
ZP					0.920	0.920								
ZSV							0.844	0.632						

Highly Repeatable
Repeatable
Potentially Repeatable
Not Repeatable

GLCM GLRLM
HIST NGTDM
GLSZM

Highly Repeatable: CCC ≥ 0.95 Repeatable: 0.90 ≤ CCC < 0.95
Potentially Repeatable: 0.85 ≤ CCC < 0.90 Not Repeatable: CCC < 0.
Texture feature abbreviations are listed in table 4. Texture features are arranged in decreasing order of frequency across all modalities. Gray Level Co-Occurrence Matrix (GLCM), Gray Level Run Length Matrix (GLRLM), Gray Level Size Zone Matrix (GLSZM) and Neighborhood Gray Tone Difference Matrix (NGTDM).

References

1 American Cancer Society, “Cancer facts & figures 2016,” (2016).

2 *National Cancer Institute Seer cancer statistics review* (2005).

3 X. Fave, L. Zhang, J. Yang, *et al.*, “Delta-radiomics features for the prediction of patient outcomes in non–small cell lung cancer,” Sci. Rep. **7**(1), 588 (2017).

4 S. Carvalho, R.T.H. Leijenaar, E.G.C. Troost, *et al.*, “Early variation of FDG-PET radiomics features in NSCLC is related to overall survival - the ‘delta radiomics’ concept,” Radiother. Oncol. **118**(Icc), S20–S21 (2016).

5 H.J.W.L. Aerts, P. Grossmann, Y. Tan, *et al.*, “Defining a Radiomic Response Phenotype: A Pilot Study using targeted therapy in NSCLC,” Sci. Rep. **6**(September), 33860 (2016).

6 W. Wu, C. Parmar, P. Grossmann, *et al.*, “Exploratory Study to Identify Radiomics Classifiers for Lung Cancer Histology,” Front. Oncol. **6**(March), 1–11 (2016).

7 K.M. Panth, R.T.H. Leijenaar, S. Carvalho, *et al.*, “Is there a causal relationship between genetic changes and radiomics-based image features? An in vivo preclinical experiment with doxycycline inducible GADD34 tumor cells,” Radiother. Oncol. **116**(3), 462–466 (2015).

8 A.J. Wong, A. Kanwar, A.S. Mohamed, and C.D. Fuller, “Radiomics in head and neck cancer: from exploration to application,” Transl. Cancer Res. **5**(4), 371–382 (2016).

9 T.P. Coroller, V. Agrawal, V. Narayan, *et al.*, “Radiomic phenotype features predict pathological response in non-small cell lung cancer,” Radiother. Oncol. **119**(3), 480–486 (2016).

10 H. Peulen, F. Mantel, M. Guckenberger, *et al.*, “Validation of high-risk CT features for detection of local recurrence after stereotactic body radiotherapy for early stage non-small cell lung cancer,” Int. J. Radiat. Oncol. **96**(1), 134–141 (2016).

11 G.J. Anthony, A. Cunliffe, R. Castillo, *et al.*, “Incorporation of pre-therapy 18 F-FDG

uptake data with CT texture features into a radiomics model for radiation pneumonitis diagnosis," Med. Phys. (2017).

12 R.T.H.M. Larue, L. Van De Voorde, J.E. van Timmeren, *et al.*, "4DCT imaging to assess radiomics feature stability: An investigation for thoracic cancers," Radiother. Oncol. **125**(1), 147–153 (2017).

13 L.A. Hunter, Y.P. Chen, L. Zhang, *et al.*, "NSCLC tumor shrinkage prediction using quantitative image features," Comput. Med. Imaging Graph. **49**, 29–36 (2015).

14 L. Hunter, "Radiomics of NSCLC: Quantitative CT Image Feature Characterization and Tumor Shrinkage Prediction," UT GSBS Diss. Theses (Open Access) (2013).

15 C. Lian, S. Ruan, T. Denœux, F. Jardin, and P. Vera, "Selecting radiomic features from FDG-PET images for cancer treatment outcome prediction," Med. Image Anal. **32**, 257–268 (2016).

16 M. Vallières, C.R. Freeman, S.R. Skamene, and I. El Naqa, "A radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities.," Phys. Med. Biol. **60**(14), 5471–96 (2015).

17 T.P. Coroller, P. Grossmann, Y. Hou, *et al.*, "CT-based radiomic signature predicts distant metastasis in lung adenocarcinoma.," Radiother. Oncol. **114**(3), 345–350 (2015).

18 J. Wu, M.F. Gensheimer, X. Dong, *et al.*, "Robust Intratumor Partitioning to Identify High-Risk Subregions in Lung Cancer: A Pilot Study," Int. J. Radiat. Oncol. **95**(5), 1504–1512 (2016).

19 X. Fave, D. Mackin, J. Yang, *et al.*, "Can radiomics features be reproducibly measured from CBCT images for patients with non-small cell lung cancer?," Med. Phys. **42**(12), 6784–6797 (2015).

20 C. Parmar, R.T.H. Leijenaar, P. Grossmann, *et al.*, "Radiomic feature clusters and Prognostic Signatures specific for Lung and Head & Neck cancer," Sci. Rep. **5**(1), 11044 (2015).

21 M. Scrivener, E.E.C. de Jong, J.E. van Timmeren, T. Pieters, B. Ghaye, and X. Geets, "Radiomics applied to lung cancer: a review," Transl. Cancer Res. **5**(4), 398–409 (2016).

22 F.H.P. van Velden, G.M. Kramer, V. Frings, *et al.*, "Repeatability of Radiomic Features in Non-Small-Cell Lung Cancer [18F]FDG-PET/CT Studies: Impact of Reconstruction and Delineation," Mol. Imaging Biol. **18**(5), 788–795 (2016).

23 M.E. Mayerhoefer, P. Szomolanyi, D. Jirak, A. Materka, and S. Trattnig, "Effects of MRI acquisition parameter variations and protocol heterogeneity on the results of texture analysis and pattern discrimination: An application-oriented study," Med. Phys. **36**(4), 1236 (2009).

24 G. Collewet, M. Strzelecki, and F. Mariette, "Influence of MRI acquisition protocols and image intensity normalization methods on texture classification," Magn. Reson. Imaging **22**(1), 81–91 (2004).

25 J. Fruehwald-Pallamar, J. Hesselink, M. Mafee, L. Holzer-Fruehwald, C. Czerny, and M. Mayerhoefer, "Texture-Based Analysis of 100 MR Examinations of Head and Neck Tumors – Is It Possible to Discriminate Between Benign and Malignant Masses in a Multicenter Trial?," RöFo - Fortschritte auf dem Gebiet der Röntgenstrahlen und der Bildgeb. Verfahren (2015).

26 L.C. V Harrison, M. Raunio, K.K. Holli, *et al.*, "MRI texture analysis in multiple sclerosis: toward a clinical analysis protocol.," Acad. Radiol. **17**(6), 696–707 (2010).

27 S. Herlidou-Même, J.M. Constans, B. Carsin, *et al.*, "MRI texture analysis on texture test objects, normal brain and intracranial tumors," Magn. Reson. Imaging **21**(9), 989–993 (2003).

28 S.J. Savio, L.C. V Harrison, T. Luukkaala, *et al.*, "Effect of slice thickness on brain magnetic resonance image texture analysis.," Biomed. Eng. Online **9**(1), 60 (2010).

29 D. Jirák, M. Dezortová, and M. Hájek, "Phantoms for texture analysis of MR images. Long-term and multi-center study.," Med. Phys. **31**(3), 616–22 (2004).

30 Y. Peng, Y. Jiang, T. Antic, M.L. Giger, S. Eggener, and A. Oto, "A study of T 2 -weighted MR image texture features and diffusion-weighted MR image features for computer-aided diagnosis of prostate cancer," **8670**(773), 86701H (2013).

31 S. Gourtsoyianni, G. Doumou, D. Prezzi, *et al.*, "Primary Rectal Cancer : Repeatability of Global and Local- Regional MR Imaging Texture," Radiology **0**(2), 1–10 (2017).

32 R.M. Haralick, K. Shanmugam, and I. Dinstein, "Textural Features for Image Classification," IEEE Trans. Syst. Man. Cybern. **3**(6), 610–621 (1973).

33 R.W. Conners, M.M. Trivedi, and C.A. Harlow, "Segmentation of a high-resolution urban scene using texture operators," Comput. Vision, Graph. Image Process. **25**(3), 273–310 (1984).

34 M. Amadasun and R. King, "Textural features corresponding to textural properties," IEEE Trans. Syst. Man Cybern. **19**(5), 1264–1273 (1989).

35 M. Galloway, "Texture Analysis Using Gray Level Run Lengths," Comput. Graph. Image Process. **4**(2), 172–179 (1975).

36 G. Thibault, B. Fertil, C. Navarro, *et al.*, "Texture Indexes and Gray Level Size Zone Matrix Application to Cell Nuclei Classification," Pattern Recognit. Inf. Process. 140–145 (2009).

37 L.I. Lin, "A Concordance Correlation-Coefficient to Evaluate Reproducibility," Biometrics **45**(1), 255–268 (1989).

38 G. McBride, "A proposal for strength-of-agreement criteria for Lin's Concordance Correlation Coefficient," NIWA Client Rep. **HAM2005-06**, 14 (2005).

39 M.D. Wilkerson and D.N. Hayes, "ConsensusClusterPlus: A class discovery tool with confidence assessments and item tracking," Bioinformatics **26**(12), 1572–1573 (2010).

40 Z.-C. Li, Q.-H. Li, B.-L. Song, *et al.*, "Clustering of MRI Radiomics Features for Glioblastoma Multiforme: An Initial Study," in edited by T. Dohi, I. Sakuma and H. Liao (Springer Berlin Heidelberg, Berlin, Heidelberg, 2016), pp. 311–319.

41 G. Heinze, P. Meinhard, D. Dunkler, and H. Southworth, *logistf: Firth's Bias-Reduced Logistic Regression*, 1–33 (2016).

42 D. Firth, "Bias Reduction of Maximum Likelihood Estimates," Biometrika **80**(1), 27–38 (1993).

43 Y. Benjamini and D. Yekutieli, "The control of the false discovery rate in multiple testing under dependency," Ann. Stat. **29**(4), 1165–1188 (2001).

44 S.S.F. Yip and H.J.W.L. Aerts, "Applications and limitations of radiomics," Phys. Med. Biol. **61**(13), (2016).

45 H.J.W.L. Aerts, E.R. Velazquez, R.T.H. Leijenaar, *et al.*, "Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach," Nat. Commun. **5**, (2014).

46 D. V Fried, S.L. Tucker, S. Zhou, *et al.*, "Prognostic value and reproducibility of pretreatment CT texture features in stage III non-small cell lung cancer.," Int. J. Radiat. Oncol. Biol. Phys. **90**(4), 834–842 (2014).

Appendix II

Formula for the fifty-nine texture features used in specific aim 1.

Histogram Texture Features

Texture features are calculated from the intensity values in a region of interest.

Feature	Formula	Description
Mean	$\mu = \frac{1}{N}\sum_{i=1}^{N} p_i$	$p_i$is element i in the region of interest and N is the total number of elements in the region of interest
Standard Deviation	$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(p_i - \mu)^2}$	$p_i$is element i in the region of interest, N is the total number of elements in the region of interest and μ is the mean
Variance	$V = \frac{1}{N-1}\sum_{i=1}^{N}\lvert p_i - \mu\rvert$	$p_i$is element i in the region of interest, N is the total number of elements in the region of interest and μ is the mean
Skewness	$s = \frac{\frac{1}{N}\sum_{i=1}^{N}(p_i - \mu)^3}{\sigma^3}$	$p_i$is element i in the region of interest, N is the total number of elements in the region of interest, μ is the mean and σ is the standard deviation.
Kurtosis	$k = \frac{\frac{1}{N}\sum_{i=1}^{N}(p_i - \mu)^4}{\sigma^4}$	$p_i$is element i in the region of interest, N is the total number of elements in the region of interest, μ is the mean and σ is the standard deviation.
Minimum	Smallest p_i in region of interest	$p_i$is element i in the region of interest
Median	Middle p_i in ordered list of all elements if odd or average of two middle values if even	$p_i$is element i in the region of interest
Maximum	Largest $p_i$in region of interest	$p_i$is element i in the region of interest
Quartile 1	The p_i separating the lowest 25% of values from the upper 75% in ordered list of all elements	$p_i$is element i in the region of interest
Quartile 3	The p_i separating the lowest 75% of values from the upper 25% in ordered list of all elements	$p_i$is element i in the region of interest

Gray Level Co-occurrence Matrix Texture Features

Texture features are calculated from the gray level co-occurrence matrix

Feature	Formula	Description
Energy	$f_1 = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p(i,j)^2$	$p(i,j)$ is the row i and column j element of the GLCM, N_g is the number of gray levels
Contrast	$f_2 = \sum_{n=0}^{N_{g-1}} n^2 \left\{\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p(i,j)\right\}$	$p(i,j)$ is the row i and column j element of the GLCM, N_g is the number of gray levels. Sum over i,j for all i,j where $\lvert i-j\rvert = n$ only
Entropy	$f_3 = -\sum_{i=1}^{X}\sum_{j=1}^{Y} p(i,j)\log\left(p(i,j)\right)$	$p(i,j)$ is the row i and column j element of the GLCM, N_g is the number of gray levels
Homogeneity	$f_4 = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} \frac{p(i,j)}{1+(i-j)^2}$	$p(i,j)$ is the row i and column j element of the GLCM, N_g is the number of gray levels
Correlation	$f_5 = \frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_g}(ij)p(i,j) - \mu_x\mu_y}{\sigma_x\sigma_y}$	$p(i,j)$ is the row i and column j element of the GLCM, $\mu_x, \mu_y, \sigma_x,$ and σ_y are the mean μ and standard deviation σ of the marginal distributions (sum along the columns and rows respectively)
Sum Average	$f_6 = \frac{1}{2}\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} ip(i,j) + jp(i,j)$	$p(i,j)$ is the row i and column j element of the GLCM, and N_g is the number of gray levels
Variance	$f_7 = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} (i-\mu_x)^2 p(i,j)$	$p(i,j)$ is the row i and column j element of the GLCM, N_g is the number of gray levels, and μ_x is the mean of the row marginal distributions (sum along the columns)

Dissimilarity	$f_8 = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g}\lvert i-j\rvert p(i,j)$	(i,j) is the row i and column j element of the GLCM, and N_g is the number of gray levels
Mean Pair Sum	$f_9 = \sum_{x=2}^{2N_g} xS(x)$	$p(i,j)$ is the row i and column j element of the GLCM, N_g is the number of gray levels and $S(x) = \sum_{i+j=x} p(i,j)$
Variance Pair Sum	$f_{10} = \sum_{x=2}^{2N_g}(x - f_9)^2 S(x)$	$p(i,j)$ is the row i and column j element of the GLCM, N_g is the number of gray levels, and $S(x) = \sum_{i+j=x} p(i,j)$
Entropy Pair Sum	$f_{11} = -\sum_{x=2}^{2N_g}\{S(x)\log(\mathrm{S(x)})\}$	$p(i,j)$ is the row i and column j element of the GLCM, N_g is the number of gray levels, and $S(x) = \sum_{i+j=x} p(i,j)$
Variance Pair Difference	$f_{12} = \sum_{x=0}^{N_g-1}(x - xD(x))^2 D(x)$	$p(i,j)$ is the row i and column j element of the GLCM, and N_g is the number of gray levels and $D(x) = \sum_{\lvert i-j\rvert=x} p(i,j)$
Entropy Pair Difference	$f_{13} = -\sum_{x=0}^{N_g-1} D(x)\log(D(x))$	$p(i,j)$ is the row i and column j element of the GLCM, and N_g is the number of gray levels and $D(x) = \sum_{\lvert i-j\rvert=x} p(i,j)$
Information Correlation Measure 1	$f_{14} = \frac{f_3 + \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p(i,j)log(p_{x,i}p_{y,j})}{-\sum_{i=1}^{N_g} p_{x,i}\log(p_{x,i})}$	$p(i,j)$ is the row i and column j element of the GLCM, and N_g is the number of gray levels, p_x and $p_y$are the distribution of sums along the column and rows respectively
Information Correlation Measure 2	$f_{15} = \sqrt{\lvert 1 - e^{-2\left(-\sum_{i=1}^{Ng}\sum_{j=1}^{Ng} p_x p_y \log(p_x p_y) + f_3\right)}\rvert}$	p_x and $p_y$are the distribution of sums along the column and rows respectively

Auto-Correlation	$f_{16} = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g}(ij)\,p(i,j)$	$p(i,j)$ is the row i and column j element of the GLCM
Cluster Shade	$f_{17} = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g}\left(i+j-\mu_x-\mu_y\right)^3(i,j)$	$p(i,j)$ is the row i and column j element of the GLCM and μ_x, μ_y is the mean marginal distributions (sum along the columns and rows respectively)
Cluster Prominence	$f_{17} = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g}\left(i+j-\mu_x-\mu_y\right)^4 p(i,j)$	$p(i,j)$ is the row i and column j element of the GLCM and μ_x, μ_y is the mean marginal distributions (sum along the columns and rows respectively)

Gray Level Run Length Matrix Texture Features

Texture features are calculated from the gray level run length matrix

Feature	Formula	Description
Short Run Emphasis	$SRE = \frac{1}{N}\sum_{i=1}^{N_g}\sum_{j=1}^{N_r}\frac{p(i,j)}{j^2}$	$p(i,j)$ is the row (gray level) i and column (run length) j element of the GLRLM, N_g is the number of gray levels, N_r is the maximum run length, and N is the total number of elements.
Long Rum Emphasis	$LRE = \frac{1}{N}\sum_{i=1}^{N_g}\sum_{j=1}^{N_r}p(i,j)j^2$	$p(i,j)$ is the row (gray level) i and column (run length) j element of the GLRLM, N_g is the number of gray levels, N_r is the maximum run length, and N is the total number of elements.
Gray Level Non-Uniformity	$GLN = \frac{1}{N}\sum_{i=1}^{N_g}\left(\sum_{j=1}^{N_r}p(i,j)\right)^2$	$p(i,j)$ is the row (gray level) i and column (run length) j element of the GLRLM, N_g is the number of gray levels, N_r is the maximum run length, and N is the total number of elements.
Run Length Non-Uniformity	$RLN = \frac{1}{N}\sum_{j=1}^{N_r}\left(\sum_{i=1}^{N_g}p(i,j)\right)^2$	$p(i,j)$ is the row (gray level) i and column (run length) j element of the GLRLM, N_g is the number of gray levels, N_r is the maximum run length, and N is the total number of elements.
Run Percentage	$RP = \frac{N}{p(i,j)j}$	$p(i,j)$ is the row (gray level) i and column (run length) j element of the GLRLM, and N is the total number of elements.
Low Gray Level Run Emphasis	$LGRE = \frac{1}{N}\sum_{i=1}^{N_g}\sum_{j=1}^{N_r}\frac{p(i,j)}{i^2}$	$p(i,j)$ is the row (gray level) i and column (run length) j element of the GLRLM, N_g is the number of gray levels, N_r is the maximum run length, and N is the total number of elements.

High Gray Level Run Emphasis	$HGRE = \frac{1}{N}\sum_{i=1}^{N_g}\sum_{j=1}^{N_r} p(i,j)i^2$	$p(i,j)$ is the row (gray level) i and column (run length) j element of the GLRLM, N_g is the number of gray levels, N_r is the maximum run length, and N is the total number of elements.
Short Run Low Gray Level Emphasis	$SRLGE = \frac{1}{N}\sum_{i=1}^{N_g}\sum_{j=1}^{N_r} \frac{p(i,j)}{i^2 j^2}$	$p(i,j)$ is the row (gray level) i and column (run length) j element of the GLRLM, N_g is the number of gray levels, N_r is the maximum run length, and N is the total number of elements.
Short Run High Gray Level Emphasis	$SRHGE = \frac{1}{N}\sum_{i=1}^{N_g}\sum_{j=1}^{N_r} \frac{p(i,j)i^2}{j^2}$	$p(i,j)$ is the row (gray level) i and column (run length) j element of the GLRLM, N_g is the number of gray levels, N_r is the maximum run length, and N is the total number of elements.
Long Run Low Gray Level Emphasis	$LRLGE = \frac{1}{N}\sum_{i=1}^{N_g}\sum_{j=1}^{N_r} \frac{p(i,j)j^2}{i^2}$	$p(i,j)$ is the row (gray level) i and column (run length) j element of the GLRLM, N_g is the number of gray levels, N_r is the maximum run length, and N is the total number of elements.
Long Run High Gray Level Emphasis	$LRHGE = \frac{1}{N}\sum_{i=1}^{N_g}\sum_{j=1}^{N_r} p(i,j)i^2 j^2$	$p(i,j)$ is the row (gray level) i and column (run length) j element of the GLRLM, N_g is the number of gray levels, N_r is the maximum run length, and N is the total number of elements.
Gray Level Variance	$GLV = \sqrt{\frac{1}{N}\sum_{i=1}^{N_g}\sum_{j=1}^{N_r} ip(i,j) - \mu_g}$	$p(i,j)$ is the row (gray level) i and column (run length) j element of the GLRLM, N_g is the number of gray levels, N_r is the maximum run length, N is the total number of elements, and $\mu_g = \frac{1}{N}\sum_{i=1}^{N_g}\sum_{j=1}^{N_r} ip(i,j)$

Run Length Variance	$RLV = \sqrt{\frac{1}{N}\sum_{i=1}^{N_g}\sum_{j=1}^{N_r} jp(i,j) - \mu_r}$	$p(i,j)$ is the row (gray level) i and column (run length) j element of the GLRLM, N_g is the number of gray levels, N_r is the maximum run length, N is the total number of elements, and $\mu_r = \frac{1}{N}\sum_{i=1}^{N_g}\sum_{j=1}^{N_r} jp(i,j)$

Gray Level Size Zone Matrix Texture Features

Texture features are calculated from the gray level size zone matrix

Feature	Formula	Description
Small Zone Emphasis	$SZE = \frac{1}{N}\sum_{i=1}^{N_g}\sum_{j=1}^{N_z}\frac{p(i,j)}{j^2}$	$p(i,j)$ is the row (gray level) i and column (zone size) j element of the GLSZM, N_g is the number of gray levels, N_z is the maximum run length, and N is the total number of elements.
Large Zone Emphasis	$LZE = \frac{1}{N}\sum_{i=1}^{N_g}\sum_{j=1}^{N_z}p(i,j)j^2$	$p(i,j)$ is the row (gray level) i and column (zone size) j element of the GLSZM, N_g is the number of gray levels, N_z is the maximum run length, and N is the total number of elements.
Gray Level Non-Uniformity	$GLNS = \frac{1}{N}\sum_{i=1}^{N_g}\left(\sum_{j=1}^{N_r}p(i,j)\right)^2$	$p(i,j)$ is the row (gray level) i and column (zone size) j element of the GLSZM, N_g is the number of gray levels, N_z is the maximum run length, and N is the total number of elements.
Zone Size Non-Uniformity	$ZSN = \frac{1}{N}\sum_{j=1}^{N_z}\left(\sum_{i=1}^{N_g}p(i,j)\right)^2$	$p(i,j)$ is the row (gray level) i and column (zone size) j element of the GLSZM, N_g is the number of gray levels, N_z is the maximum run length, and N is the total number of elements.
Zone Percentage	$ZP = \frac{N}{p(i,j)j}$	$p(i,j)$ is the row (gray level) i and column (zone size) j element of the GLSZM, and N is the total number of elements.
Low Gray Level Zone Emphasis	$LGZE = \frac{1}{N}\sum_{i=1}^{N_g}\sum_{j=1}^{N_z}\frac{p(i,j)}{i^2}$	$p(i,j)$ is the row (gray level) i and column (zone size) j element of the GLSZM, N_g is the number of gray levels, N_z is the maximum run length, and N is the total number of elements.

High Gray Level Zone Emphasis	$HGZE = \frac{1}{N}\sum_{i=1}^{N_g}\sum_{j=1}^{N_z} p(i,j)i^2$	$p(i,j)$ is the row (gray level) i and column (zone size) j element of the GLSZM, N_g is the number of gray levels, N_z is the maximum run length, and N is the total number of elements.
Small Zone Low Gray Level Emphasis	$SZLGE = \frac{1}{N}\sum_{i=1}^{N_g}\sum_{j=1}^{N_z} \frac{p(i,j)}{i^2 j^2}$	$p(i,j)$ is the row (gray level) i and column (zone size) j element of the GLSZM, N_g is the number of gray levels, N_z is the maximum run length, and N is the total number of elements.
Small Zone High Gray Level Emphasis	$SZHGE = \frac{1}{N}\sum_{i=1}^{N_g}\sum_{j=1}^{N_z} \frac{p(i,j)i^2}{j^2}$	$p(i,j)$ is the row (gray level) i and column (zone size) j element of the GLSZM, N_g is the number of gray levels, N_z is the maximum run length, and N is the total number of elements.
Large Zone Low Gray Level Emphasis	$LZLGE = \frac{1}{N}\sum_{i=1}^{N_g}\sum_{j=1}^{N_z} \frac{p(i,j)j^2}{i^2}$	$p(i,j)$ is the row (gray level) i and column (zone size) j element of the GLSZM, N_g is the number of gray levels, N_z is the maximum run length, and N is the total number of elements.
Large Zone High Gray Level Emphasis	$LZHGE = \frac{1}{N}\sum_{i=1}^{N_g}\sum_{j=1}^{N_z} p(i,j)i^2 j^2$	$p(i,j)$ is the row (gray level) i and column (zone size) j element of the GLSZM, N_g is the number of gray levels, N_z is the maximum run length, and N is the total number of elements.
Gray Level Variance	$GLVS = \sqrt{\frac{1}{N}\sum_{i=1}^{N_g}\sum_{j=1}^{N_z} ip(i,j) - \mu_g}$	$p(i,j)$ is the row (gray level) i and column (zone size) j element of the GLSZM, N_g is the number of gray levels, N_z is the maximum run length, N is the total number of elements, and $\mu_g = \frac{1}{N}\sum_{i=1}^{N_g}\sum_{j=1}^{N_z} ip(i,j)$

Zone Size Variance	$ZSV = \sqrt{\frac{1}{N}\sum_{i=1}^{N_g}\sum_{j=1}^{N_z} jp(i,j) - \mu_z}$	$p(i,j)$ is the row (gray level) i and column (zone size) j element of the GLSZM, N_g is the number of gray levels, N_z is the maximum run length, N is the total number of elements, and $\mu_z = \frac{1}{N}\sum_{i=1}^{N_g}\sum_{j=1}^{N_z} jp(i,j)$

Neighborhood Gray Tone Difference Matrix Texture Features

Texture features are calculated from the neighborhood gray tone difference matrix.

Feature	Formula	Description
Coarseness	$f_1 = \frac{1}{\sum_{i=0}^{G} \frac{N_i}{(N-2d)^2} s(i)}$	$s(i)$ is the ith element of the NGTDM, G is the maximum gray level, N_i is the number of elements of with gray tone i in the image, N is the total number of elements in the image, and d is the size of the neighborhood used to calculate the NGTDM
Contrast	$f_2 = \left[\frac{1}{N_g(N_g-1)} \sum_{i=0}^{G}\sum_{j=0}^{G} p_i p_j (i - j)^2\right]\left[\frac{1}{N^2}\sum_{i=0}^{G} s(i)\right]$	$s(i)$ is the ith element of the NGTDM, G is the maximum gray level, $p_i, p_j = \frac{N_{i(j)}}{(N-2d)^2}$, $N_{i(j)}$ is the number of elements of with gray tone $i(j)$ in the image, N is the total number of elements in the image, d is the size of the neighborhood used to calculate the NGTDM, and N_g is the number of distinct gray levels.
Busyness	$f_3 = \frac{\sum_{i=0}^{G} p_i s(i)}{\sum_{i=0}^{G}\sum_{j=0}^{G} ip_i - jp_j}$	$s(i)$ is the ith element of the NGTDM, G is the maximum gray level, $p_i, p_j = \frac{N_{i(j)}}{(N-2d)^2}$, $N_{i(j)}$ is the number of elements of with gray tone $i(j)$ in the image, N is the total number of elements in the image, and d is the size of the neighborhood used to calculate the NGTDM; $p_i \neq 0, p_j \neq 0$
Complexity	$f_4 = \sum_{i=0}^{G}\sum_{j=0}^{G} \frac{\lvert i-j\rvert}{N^2(p_i+p_j)}\{p_i s(i) + p_j s(j)\}$	$s(i)$ is the ith element of the NGTDM, G is the maximum gray level, $p_i, p_j = \frac{N_{i(j)}}{(N-2d)^2}$, $N_{i(j)}$ is the number of elements of with gray tone $i(j)$ in the image, N is the

		total number of elements in the image, and d is the size of the neighborhood used to calculate the NGTDM; $p_i \neq 0, p_j \neq 0$
Strength	$f_5 = \frac{\sum_{i=0}^{G}\sum_{j=0}^{G}(p_i + p_j)(i-j)^2}{\sum_{i=0}^{G} s(i)}$	$s(i)$ is the ith element of the NGTDM, G is the maximum gray level, $p_i, p_j = \frac{N_{i(j)}}{(N-2d)^2}$, $N_{i(j)}$ is the number of elements of with gray tone $i(j)$ in the image, N is the total number of elements in the image, and d is the size of the neighborhood used to calculate the NGTDM; $p_i \neq 0, p_j \neq 0$

Appendix III

Violin plots of the CT only interface uncertainty by individual subject.

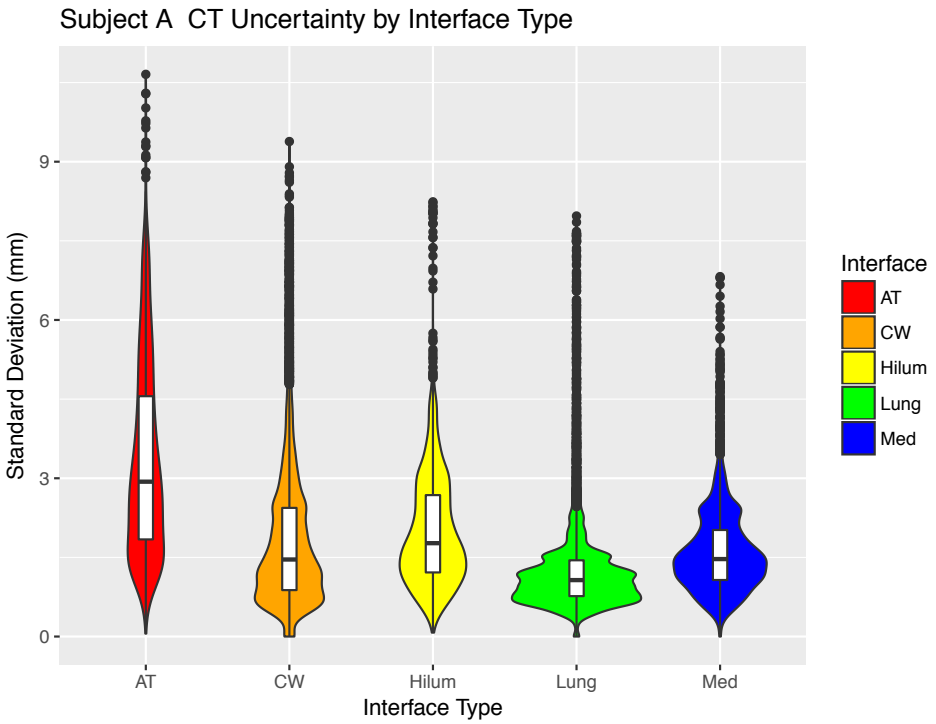


Figure 1: Violin plots of the CT only uncertainty by interface type of subject A where the width of the plot represents the probability density function (PDF) of values with the indicated standard deviation. Along the center line is a box plot showing the first quartile, median, and third quartile along with the extreme minimum and maximum values within 1.5 times the inner quartile range. Values more extreme are indicated by dots along the whiskers. AT indicates atelectasis interface, CW indicates the chest wall interface, and Med indicates the mediastinum interface.

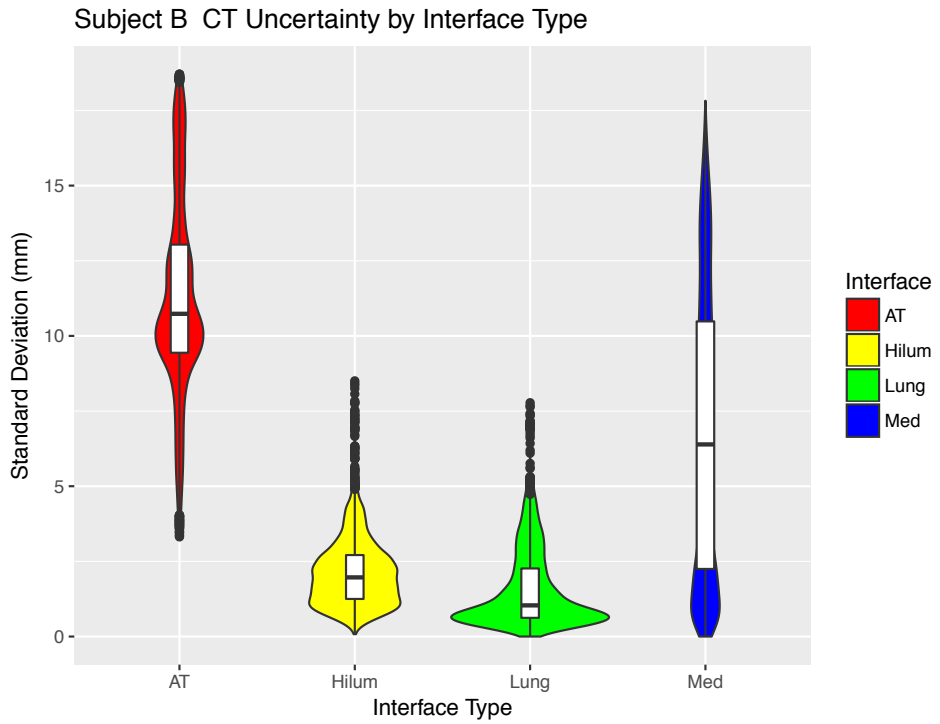


Figure 2: Violin plots of the CT only uncertainty by interface type of subject B where the width of the plot represents the probability density function (PDF) of values with the indicated standard deviation. Along the center line is a box plot showing the first quartile, median, and third quartile along with the extreme minimum and maximum values within 1.5 times the inner quartile range. Values more extreme are indicated by dots along the whiskers. AT indicates atelectasis interface, CW indicates the chest wall interface, and Med indicates the mediastinum interface.

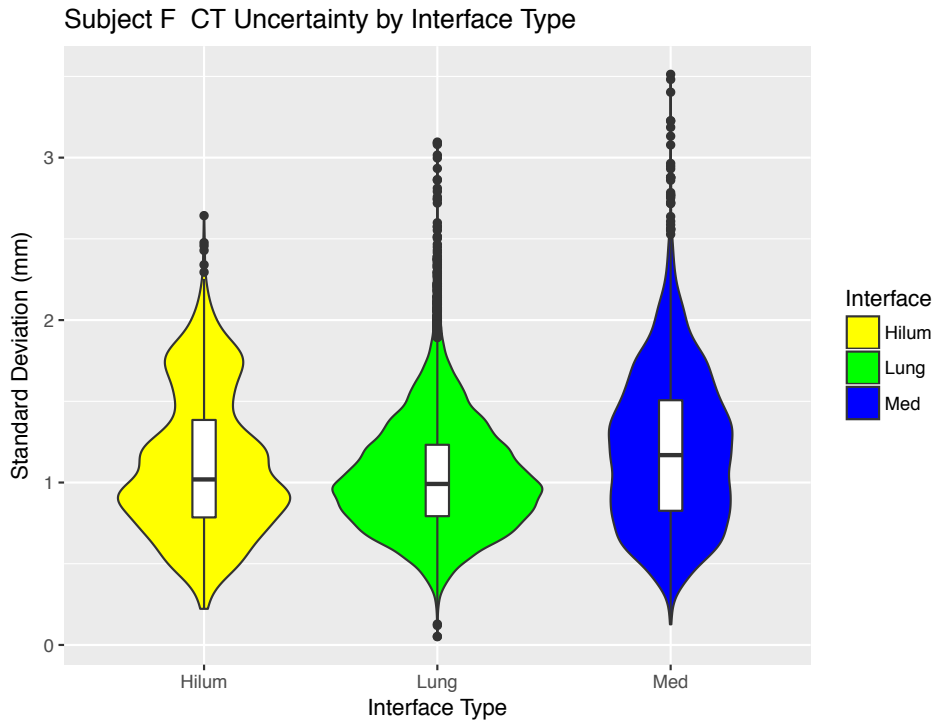


Figure 3: Violin plots of the CT only uncertainty by interface type of subject F where the width of the plot represents the probability density function (PDF) of values with the indicated standard deviation. Along the center line is a box plot showing the first quartile, median, and third quartile along with the extreme minimum and maximum values within 1.5 times the inner quartile range. Values more extreme are indicated by dots along the whiskers. AT indicates atelectasis interface, CW indicates the chest wall interface, and Med indicates the mediastinum interface.

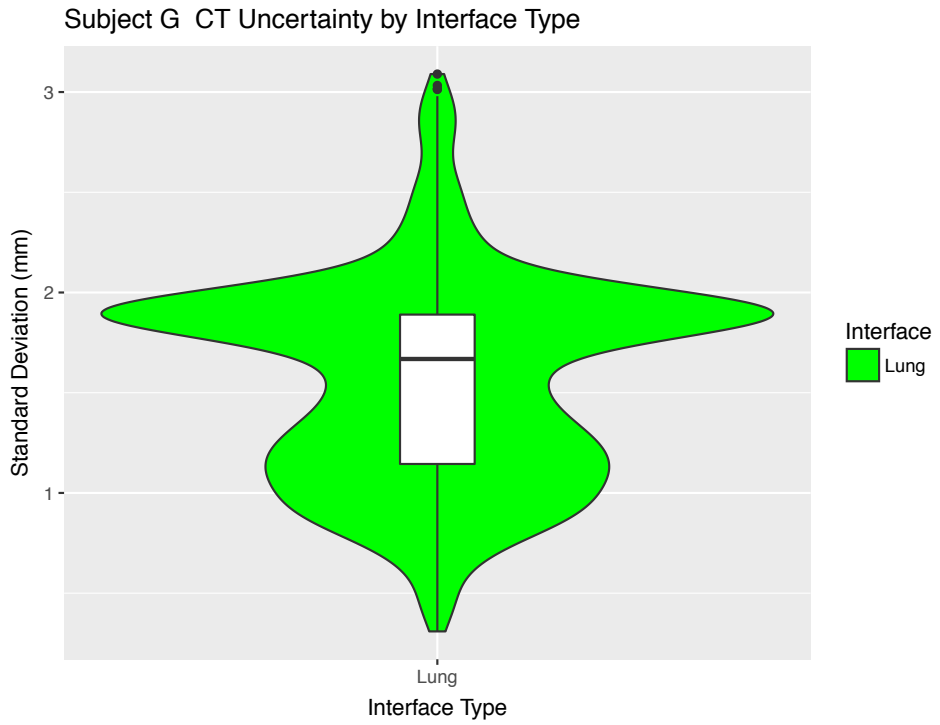


Figure 4: Violin plots of the CT only uncertainty by interface type of subject G where the width of the plot represents the probability density function (PDF) of values with the indicated standard deviation. Along the center line is a box plot showing the first quartile, median, and third quartile along with the extreme minimum and maximum values within 1.5 times the inner quartile range. Values more extreme are indicated by dots along the whiskers. AT indicates atelectasis interface, CW indicates the chest wall interface, and Med indicates the mediastinum interface.

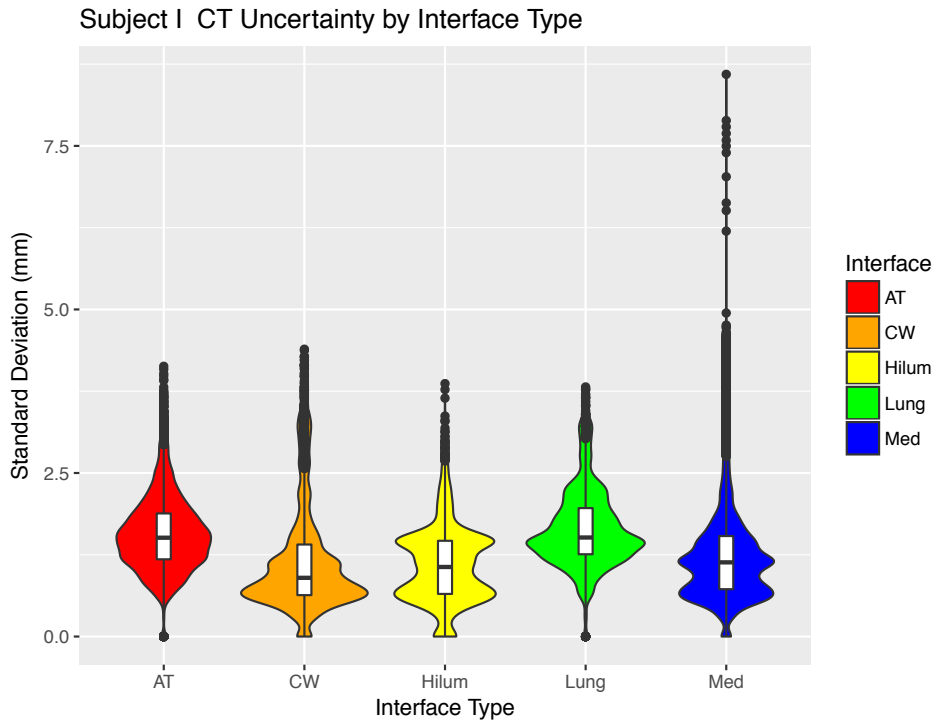


Figure 5: Violin plots of the CT only uncertainty by interface type of subject I where the width of the plot represents the probability density function (PDF) of values with the indicated standard deviation. Along the center line is a box plot showing the first quartile, median, and third quartile along with the extreme minimum and maximum values within 1.5 times the inner quartile range. Values more extreme are indicated by dots along the whiskers. AT indicates atelectasis interface, CW indicates the chest wall interface, and Med indicates the mediastinum interface.

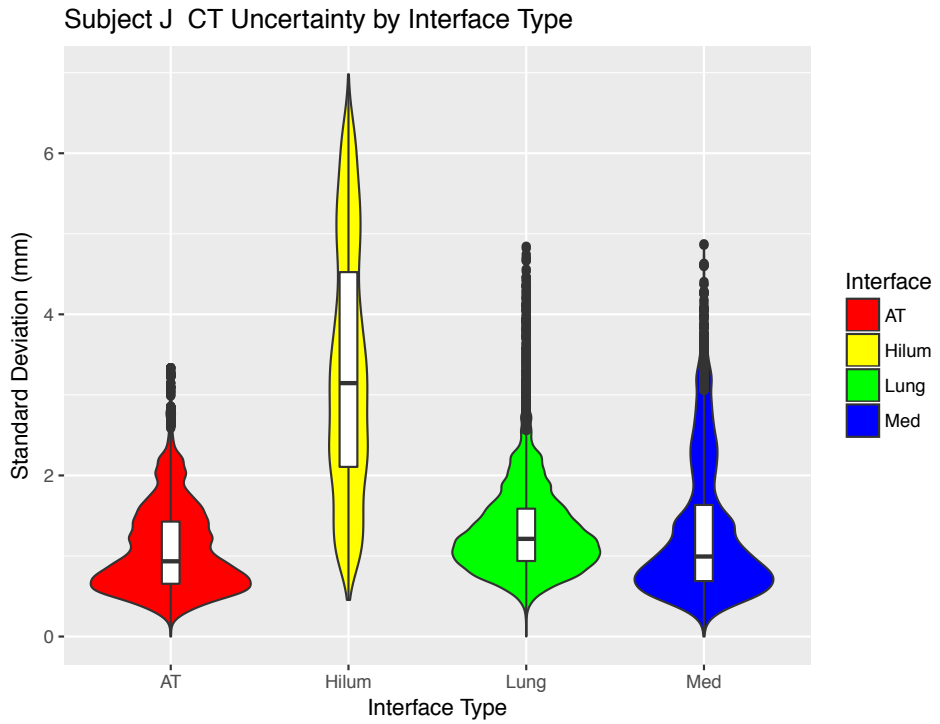


Figure 6: Violin plots of the CT only uncertainty by interface type of subject J where the width of the plot represents the probability density function (PDF) of values with the indicated standard deviation. Along the center line is a box plot showing the first quartile, median, and third quartile along with the extreme minimum and maximum values within 1.5 times the inner quartile range. Values more extreme are indicated by dots along the whiskers. AT indicates atelectasis interface, CW indicates the chest wall interface, and Med indicates the mediastinum interface.

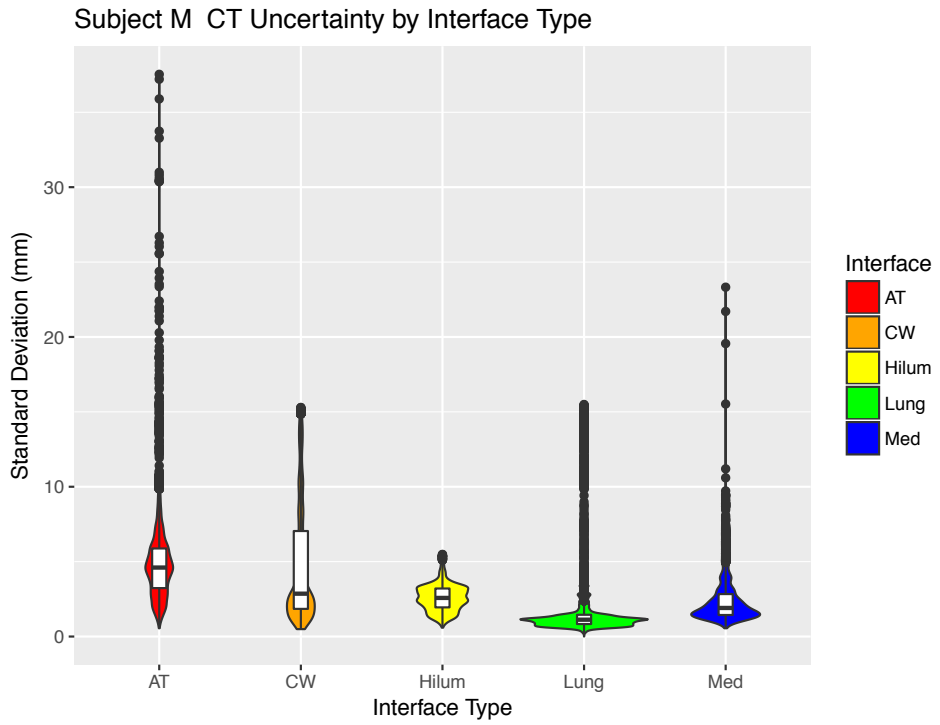


Figure 7: Violin plots of the CT only uncertainty by interface type of subject M where the width of the plot represents the probability density function (PDF) of values with the indicated standard deviation. Along the center line is a box plot showing the first quartile, median, and third quartile along with the extreme minimum and maximum values within 1.5 times the inner quartile range. Values more extreme are indicated by dots along the whiskers. AT indicates atelectasis interface, CW indicates the chest wall interface, and Med indicates the mediastinum interface.

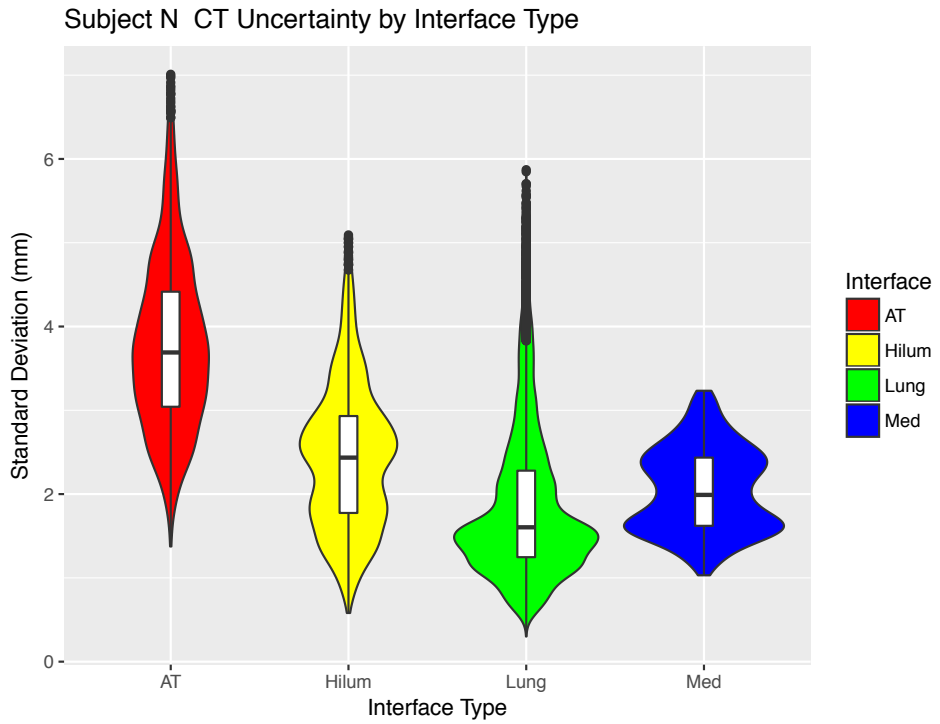


Figure 8: Violin plots of the CT only uncertainty by interface type of subject N where the width of the plot represents the probability density function (PDF) of values with the indicated standard deviation. Along the center line is a box plot showing the first quartile, median, and third quartile along with the extreme minimum and maximum values within 1.5 times the inner quartile range. Values more extreme are indicated by dots along the whiskers. AT indicates atelectasis interface, CW indicates the chest wall interface, and Med indicates the mediastinum interface.

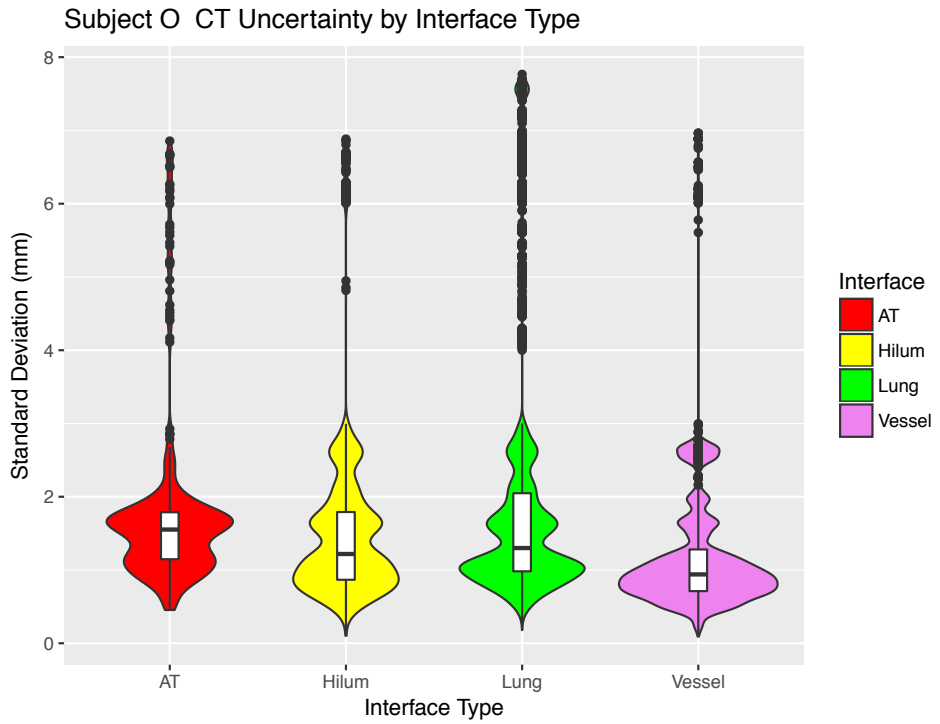


Figure 9: Violin plots of the CT only uncertainty by interface type of subject O where the width of the plot represents the probability density function (PDF) of values with the indicated standard deviation. Along the center line is a box plot showing the first quartile, median, and third quartile along with the extreme minimum and maximum values within 1.5 times the inner quartile range. Values more extreme are indicated by dots along the whiskers. AT indicates atelectasis interface, CW indicates the chest wall interface, and Med indicates the mediastinum interface.

Appendix IV

Violin plots of the PET/CT only interface uncertainty by individual subject.

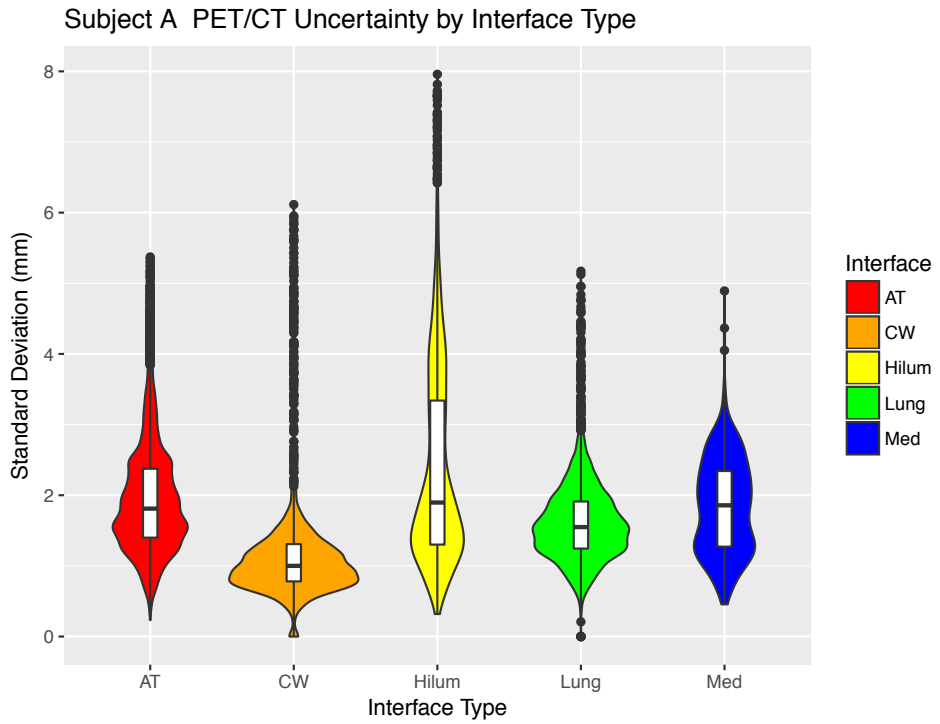


Figure 1: Violin plots of the PET/CT uncertainty by interface type of subject A where the width of the plot represents the probability density function (PDF) of values with the indicated standard deviation. Along the center line is a box plot showing the first quartile, median, and third quartile along with the extreme minimum and maximum values within 1.5 times the inner quartile range. Values more extreme are indicated by dots along the whiskers. AT indicates atelectasis interface, CW indicates the chest wall interface, and Med indicates the mediastinum.

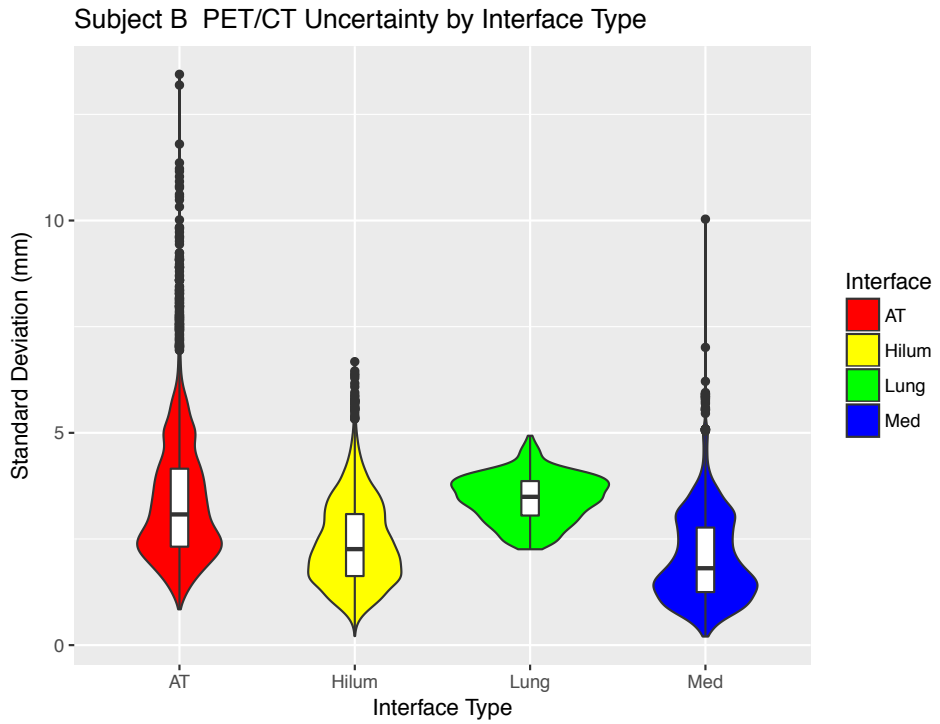


Figure 2: Violin plots of the PET/CT uncertainty by interface type of subject B where the width of the plot represents the probability density function (PDF) of values with the indicated standard deviation. Along the center line is a box plot showing the first quartile, median, and third quartile along with the extreme minimum and maximum values within 1.5 times the inner quartile range. Values more extreme are indicated by dots along the whiskers. AT indicates atelectasis interface, CW indicates the chest wall interface, and Med indicates the mediastinum.

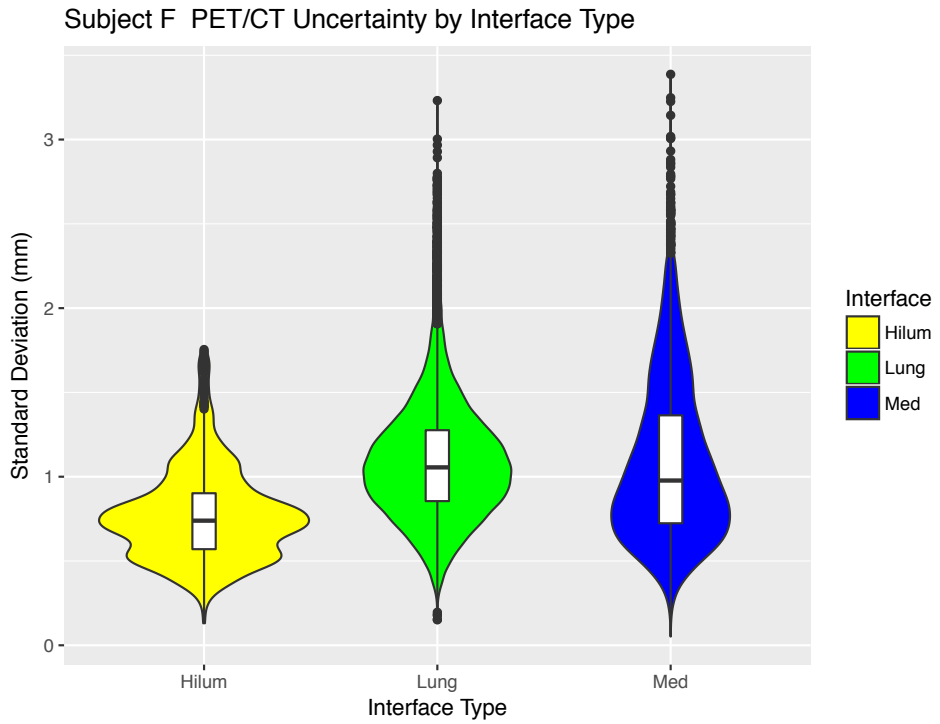


Figure 3: Violin plots of the PET/CT uncertainty by interface type of subject F where the width of the plot represents the probability density function (PDF) of values with the indicated standard deviation. Along the center line is a box plot showing the first quartile, median, and third quartile along with the extreme minimum and maximum values within 1.5 times the inner quartile range. Values more extreme are indicated by dots along the whiskers. AT indicates atelectasis interface, CW indicates the chest wall interface, and Med indicates the mediastinum.

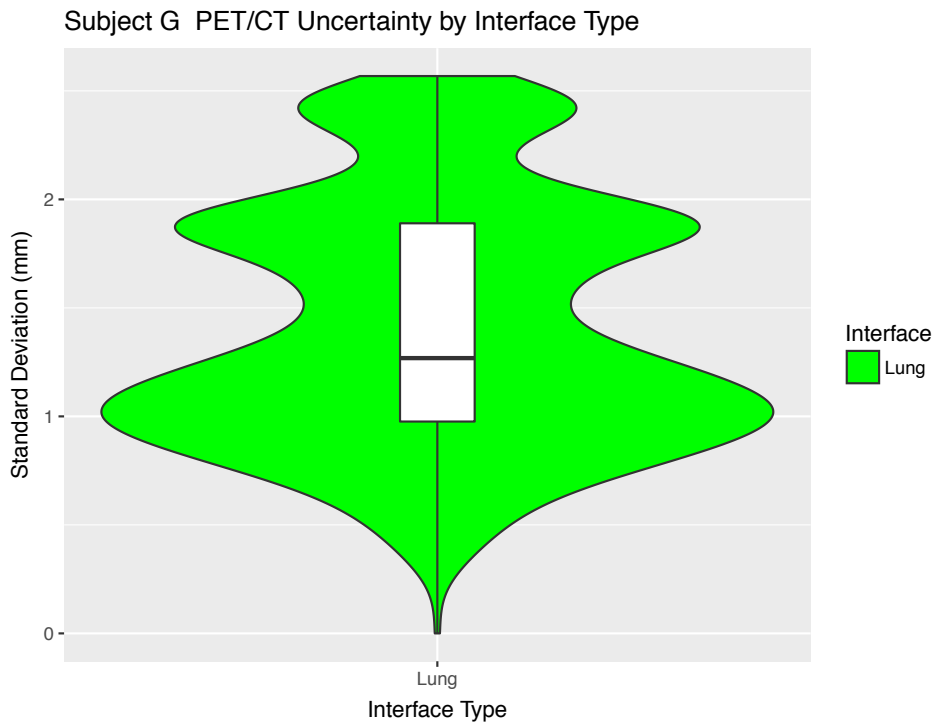


Figure 4: Violin plots of the PET/CT uncertainty by interface type of subject G where the width of the plot represents the probability density function (PDF) of values with the indicated standard deviation. Along the center line is a box plot showing the first quartile, median, and third quartile along with the extreme minimum and maximum values within 1.5 times the inner quartile range. Values more extreme are indicated by dots along the whiskers. AT indicates atelectasis interface, CW indicates the chest wall interface, and Med indicates the mediastinum.

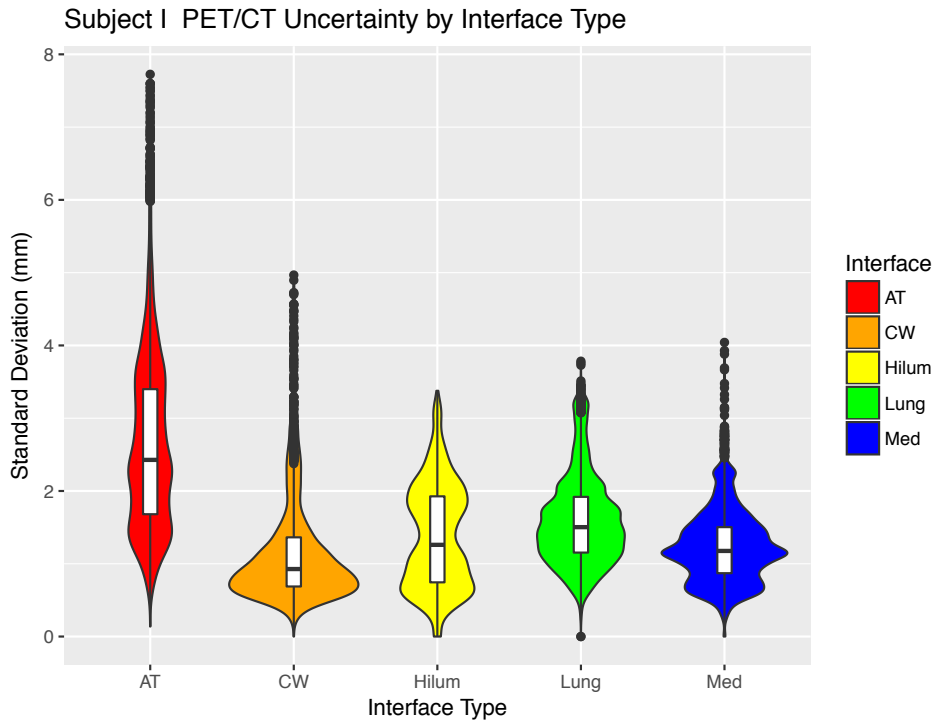


Figure 5: Violin plots of the PET/CT uncertainty by interface type of subject I where the width of the plot represents the probability density function (PDF) of values with the indicated standard deviation. Along the center line is a box plot showing the first quartile, median, and third quartile along with the extreme minimum and maximum values within 1.5 times the inner quartile range. Values more extreme are indicated by dots along the whiskers. AT indicates atelectasis interface, CW indicates the chest wall interface, and Med indicates the mediastinum.

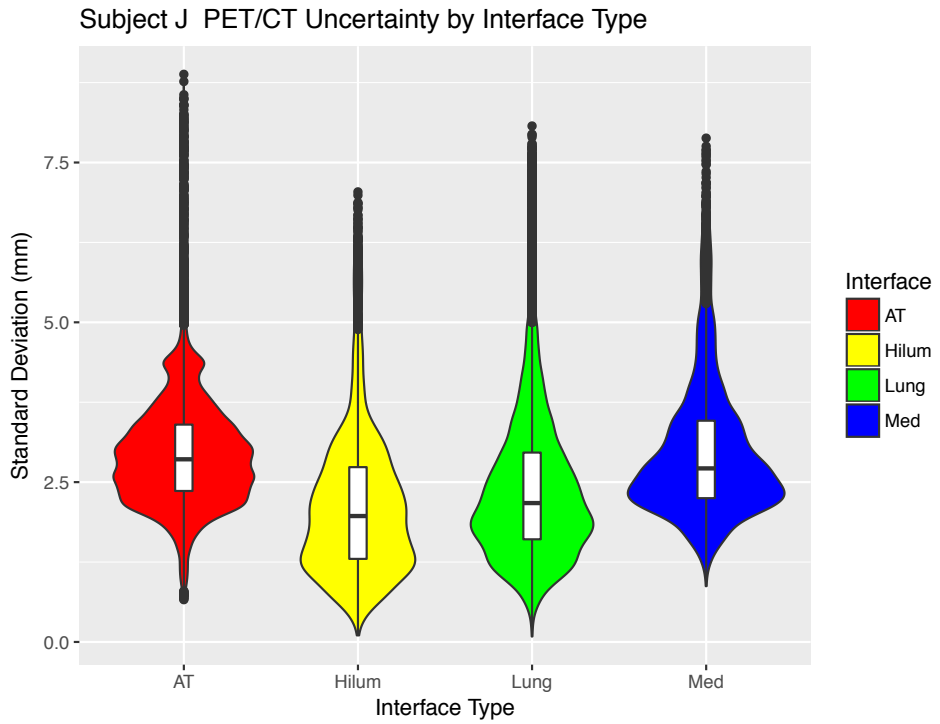


Figure 6: Violin plots of the PET/CT uncertainty by interface type of subject J where the width of the plot represents the probability density function (PDF) of values with the indicated standard deviation. Along the center line is a box plot showing the first quartile, median, and third quartile along with the extreme minimum and maximum values within 1.5 times the inner quartile range. Values more extreme are indicated by dots along the whiskers. AT indicates atelectasis interface, CW indicates the chest wall interface, and Med indicates the mediastinum.

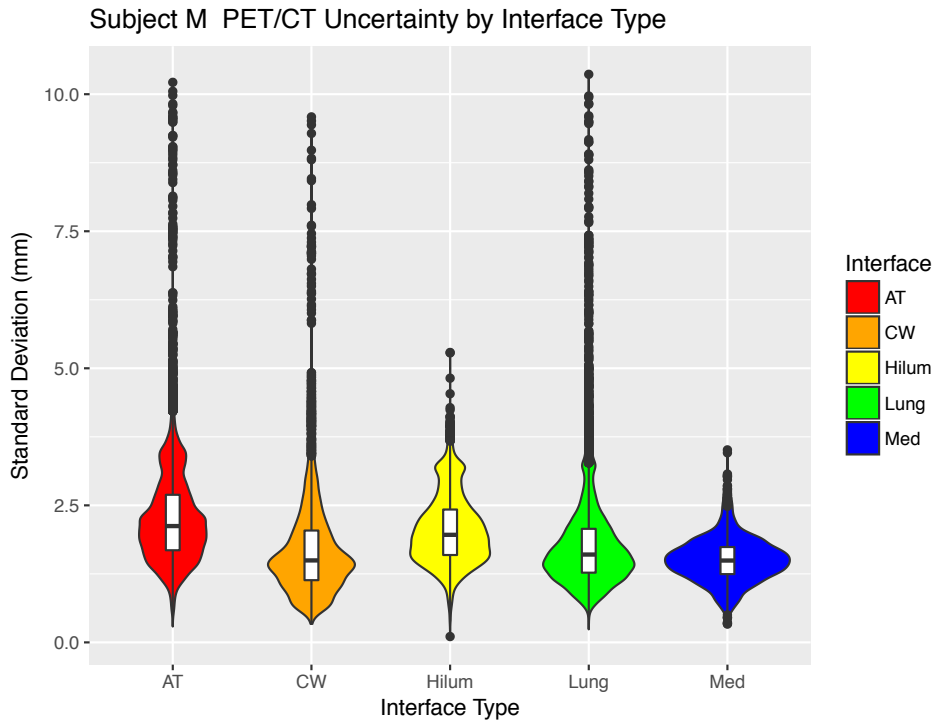


Figure 7: Violin plots of the PET/CT uncertainty by interface type of subject M where the width of the plot represents the probability density function (PDF) of values with the indicated standard deviation. Along the center line is a box plot showing the first quartile, median, and third quartile along with the extreme minimum and maximum values within 1.5 times the inner quartile range. Values more extreme are indicated by dots along the whiskers. AT indicates atelectasis interface, CW indicates the chest wall interface, and Med indicates the mediastinum.

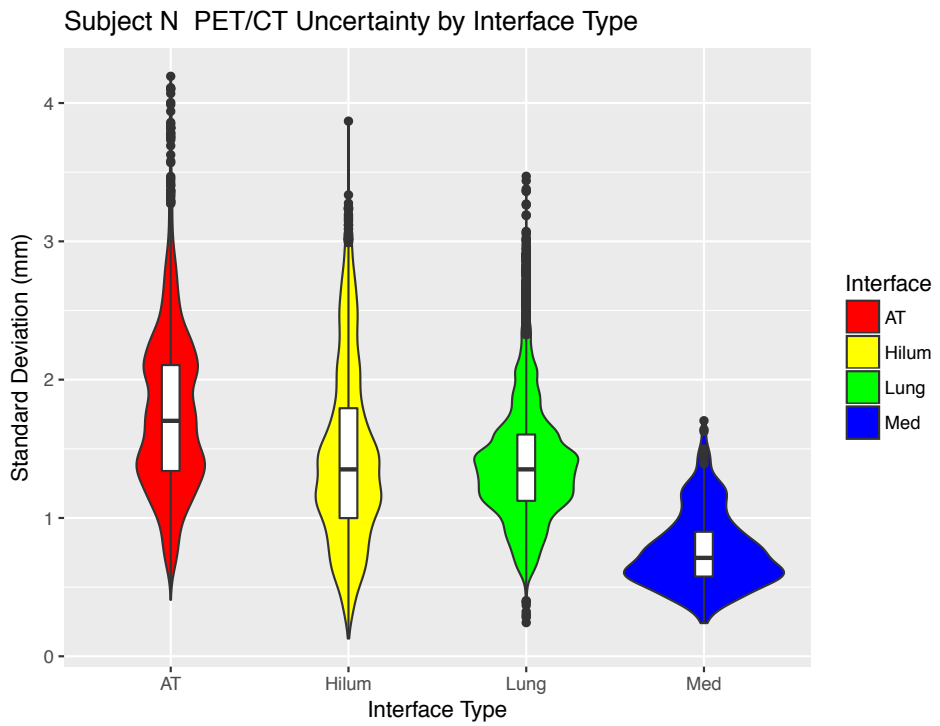


Figure 8: Violin plots of the PET/CT uncertainty by interface type of subject N where the width of the plot represents the probability density function (PDF) of values with the indicated standard deviation. Along the center line is a box plot showing the first quartile, median, and third quartile along with the extreme minimum and maximum values within 1.5 times the inner quartile range. Values more extreme are indicated by dots along the whiskers. AT indicates atelectasis interface, CW indicates the chest wall interface, and Med indicates the mediastinum.

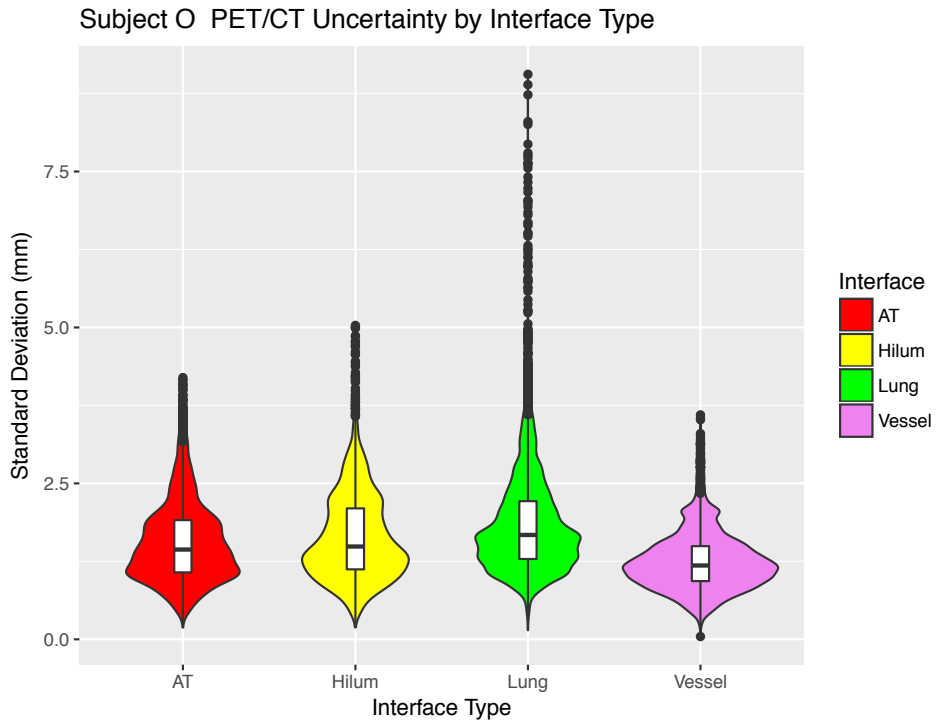


Figure 9: Violin plots of the PET/CT uncertainty by interface type of subject O where the width of the plot represents the probability density function (PDF) of values with the indicated standard deviation. Along the center line is a box plot showing the first quartile, median, and third quartile along with the extreme minimum and maximum values within 1.5 times the inner quartile range. Values more extreme are indicated by dots along the whiskers. AT indicates atelectasis interface, CW indicates the chest wall interface, and Med indicates the mediastinum.

Appendix V

Examples of prediction map crated from the BinaryRes_Tumor network output. Illustrative slice from all test subjects spread evenly throughout the tumor contour.

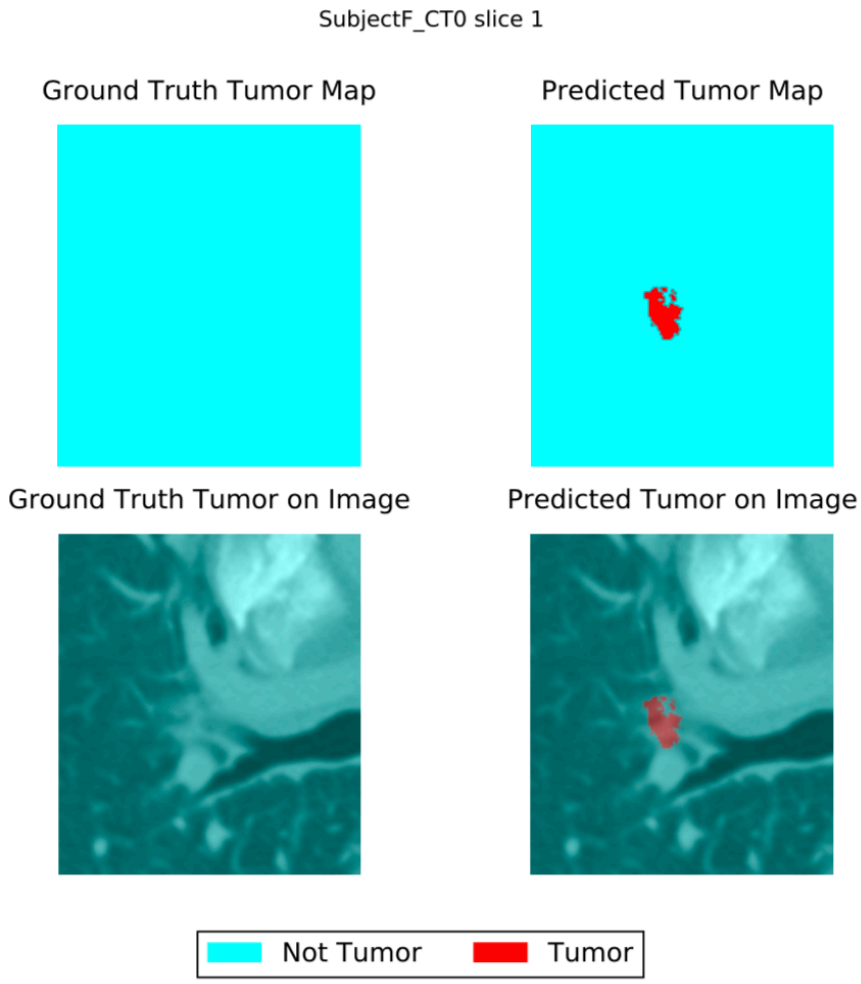


Figure 1: Comparison of the ground truth tumor location map with the predicted location map from the BinaryRes_Tumor network following CRF post processing. The top row represents the map of true and predicted labels while the bottom row shows the maps overlaid on the corresponding image slice.

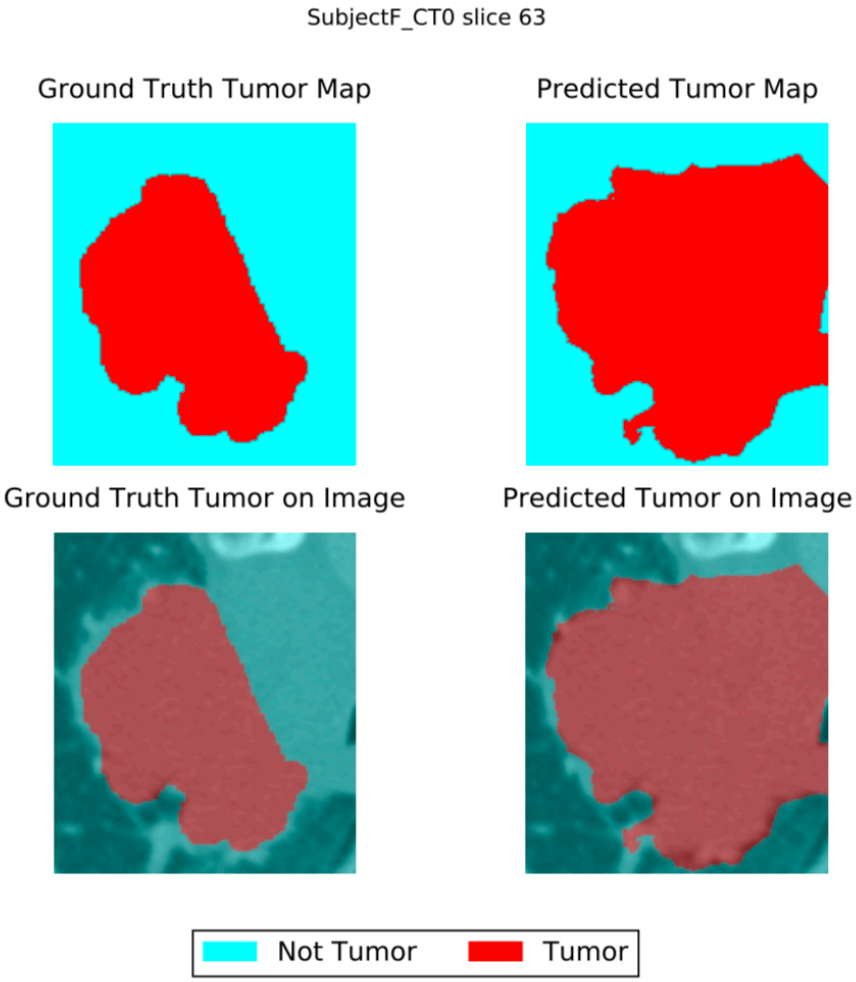


Figure 2: Comparison of the ground truth tumor location map with the predicted location map from the BinaryRes_Tumor network following CRF post processing. The top row represents the map of true and predicted labels while the bottom row shows the maps overlaid on the corresponding image slice.

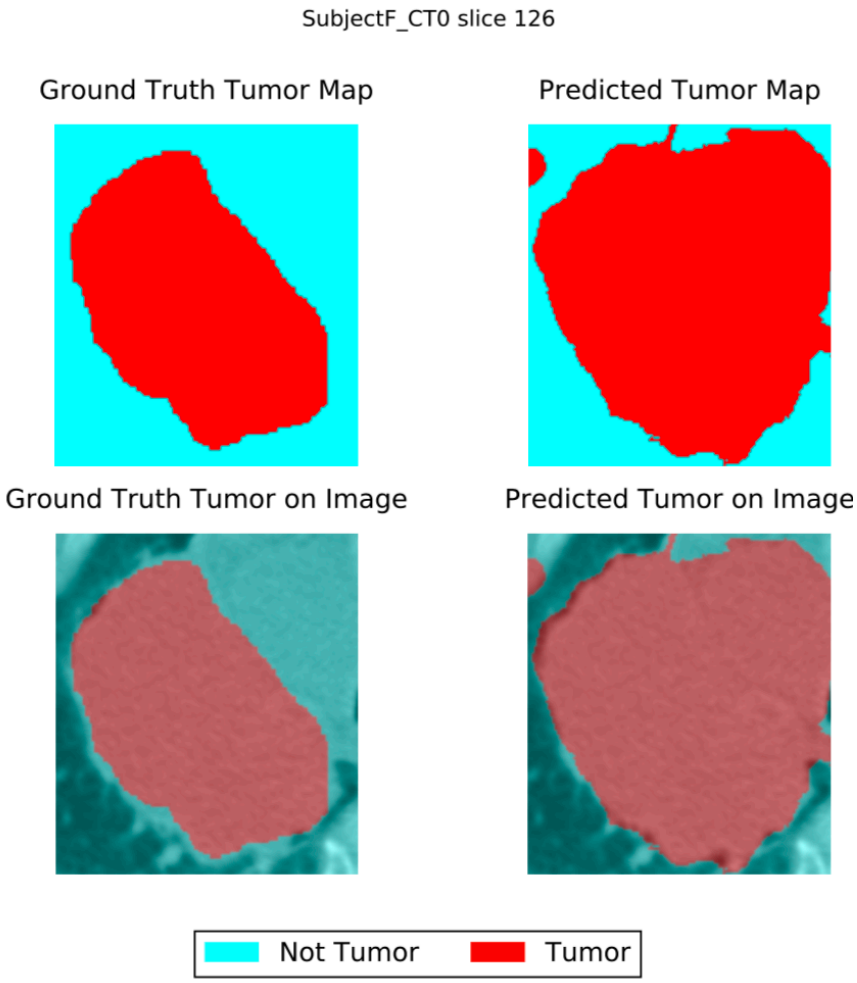


Figure 3: Comparison of the ground truth tumor location map with the predicted location map from the BinaryRes_Tumor network following CRF post processing. The top row represents the map of true and predicted labels while the bottom row shows the maps overlaid on the corresponding image slice.

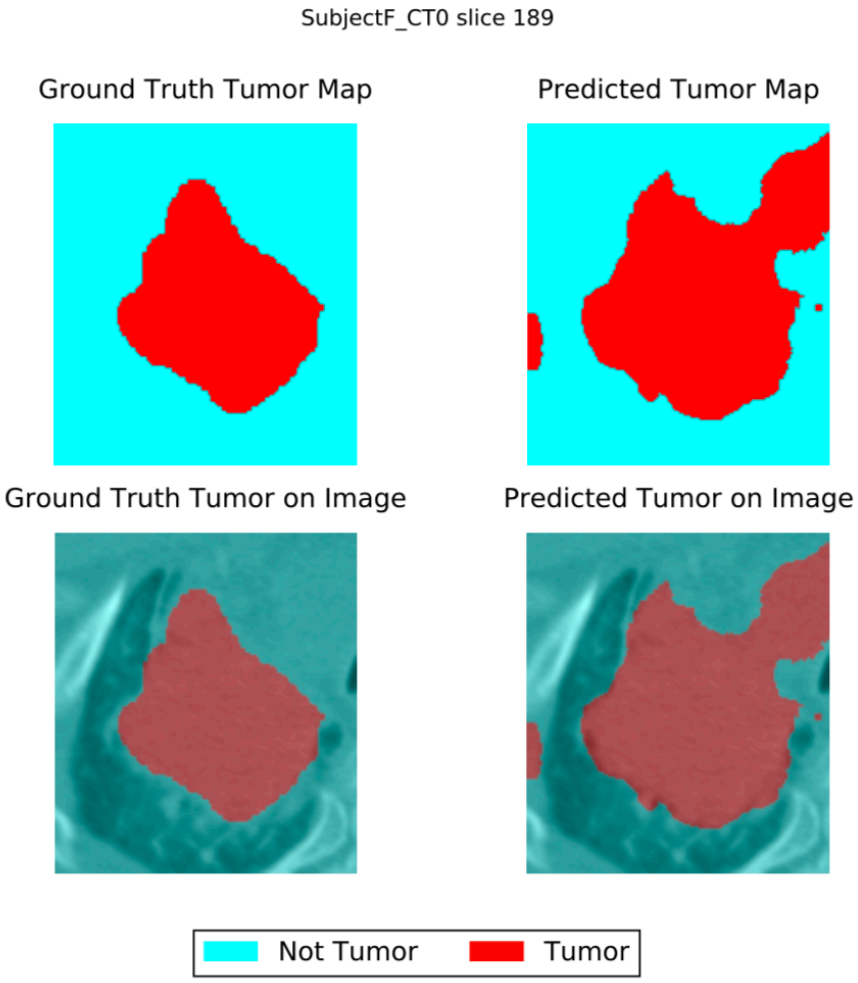


Figure 4: Comparison of the ground truth tumor location map with the predicted location map from the BinaryRes_Tumor network following CRF post processing. The top row represents the map of true and predicted labels while the bottom row shows the maps overlaid on the corresponding image slice.

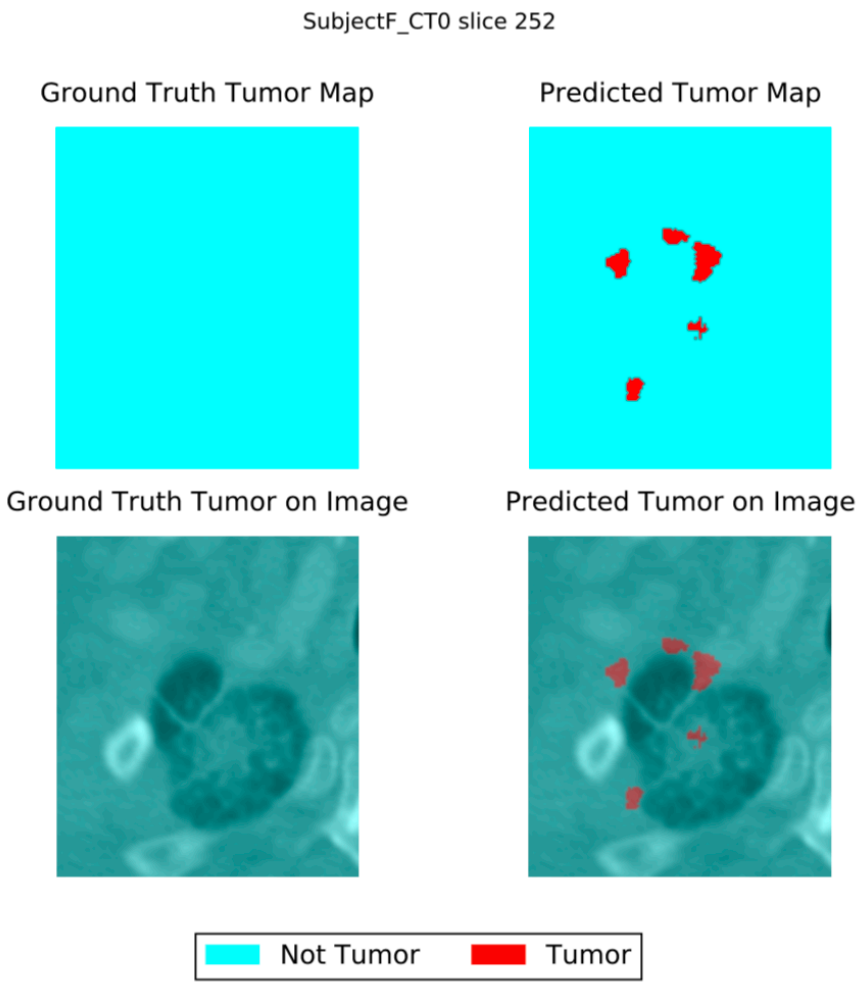


Figure 5: Comparison of the ground truth tumor location map with the predicted location map from the BinaryRes_Tumor network following CRF post processing. The top row represents the map of true and predicted labels while the bottom row shows the maps overlaid on the corresponding image slice.

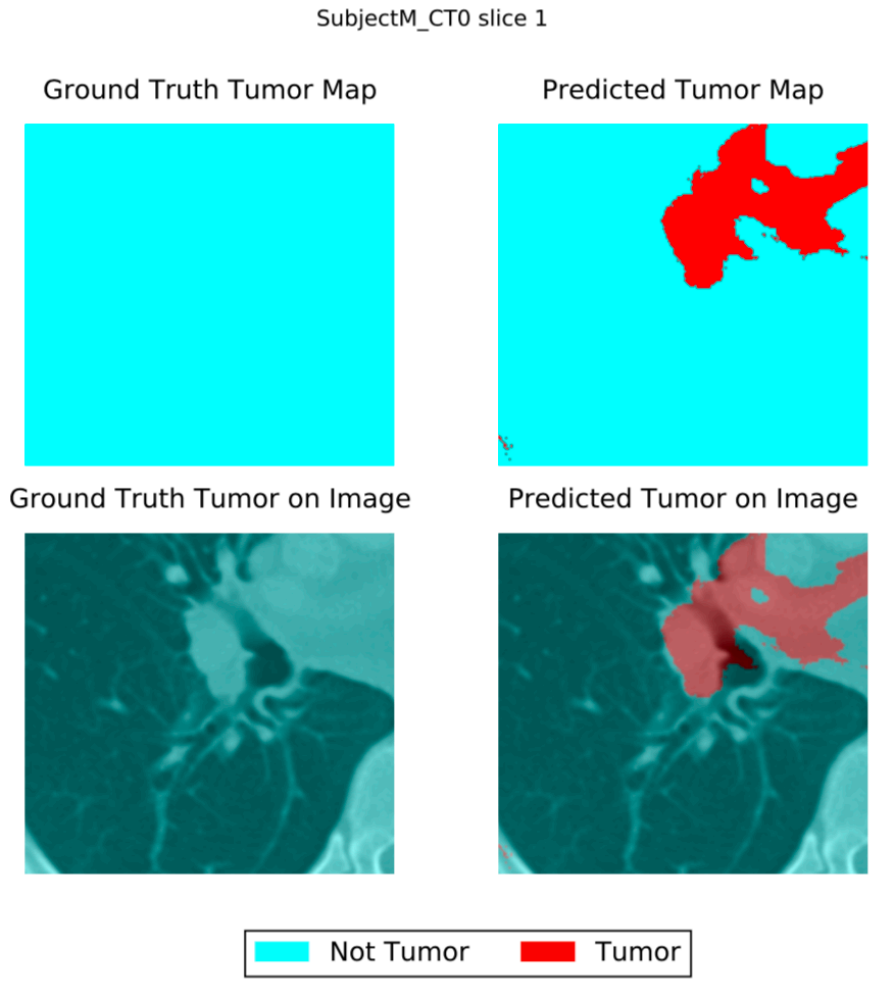


Figure 6: Comparison of the ground truth tumor location map with the predicted location map from the BinaryRes_Tumor network following CRF post processing. The top row represents the map of true and predicted labels while the bottom row shows the maps overlaid on the corresponding image slice.

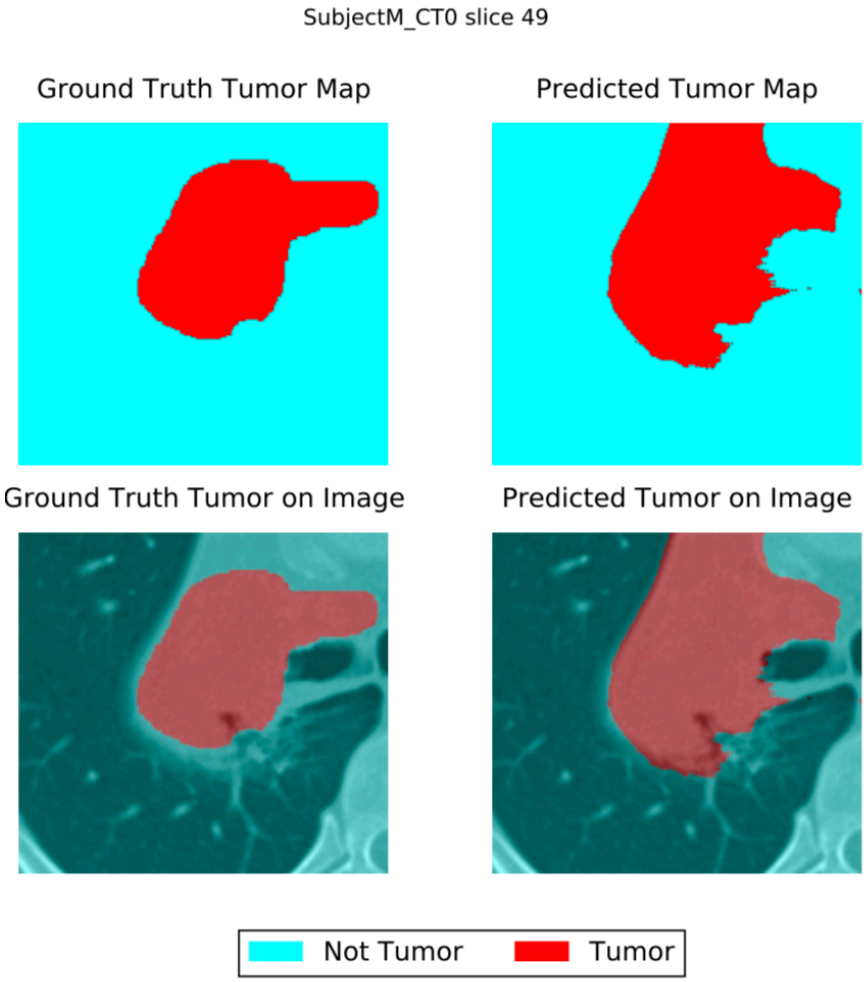


Figure 7: Comparison of the ground truth tumor location map with the predicted location map from the BinaryRes_Tumor network following CRF post processing. The top row represents the map of true and predicted labels while the bottom row shows the maps overlaid on the corresponding image slice.

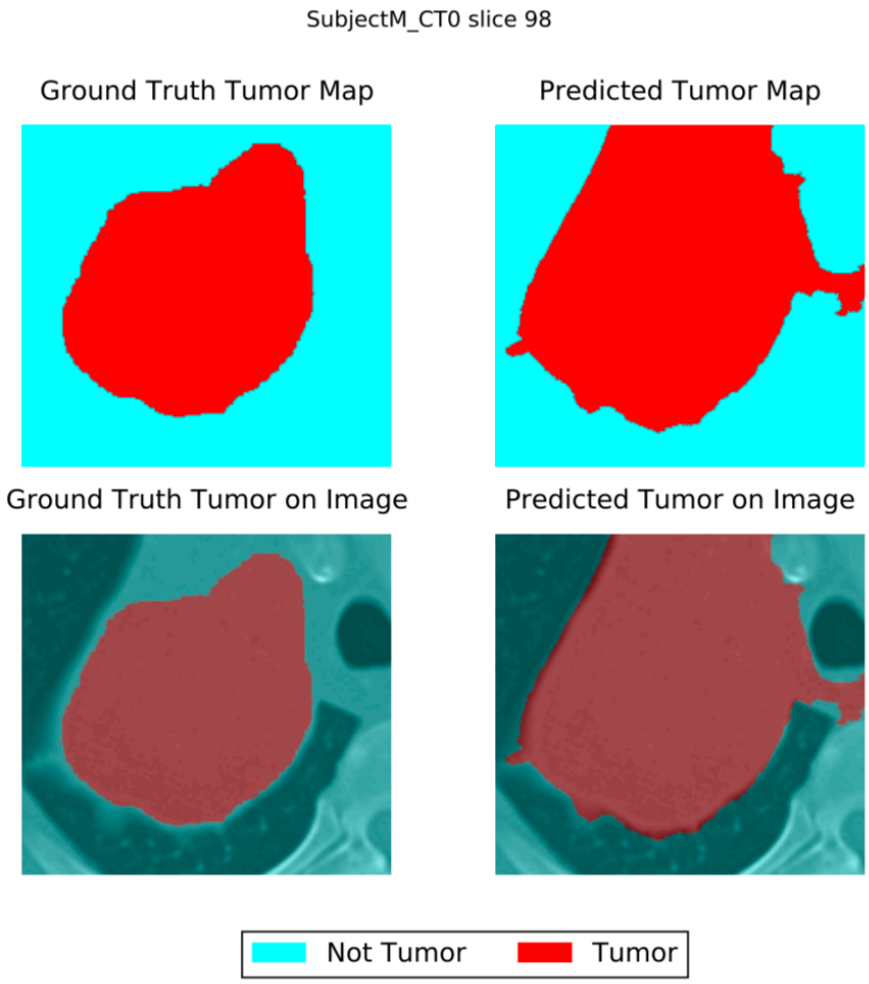


Figure 8: Comparison of the ground truth tumor location map with the predicted location map from the BinaryRes_Tumor network following CRF post processing. The top row represents the map of true and predicted labels while the bottom row shows the maps overlaid on the corresponding image slice.

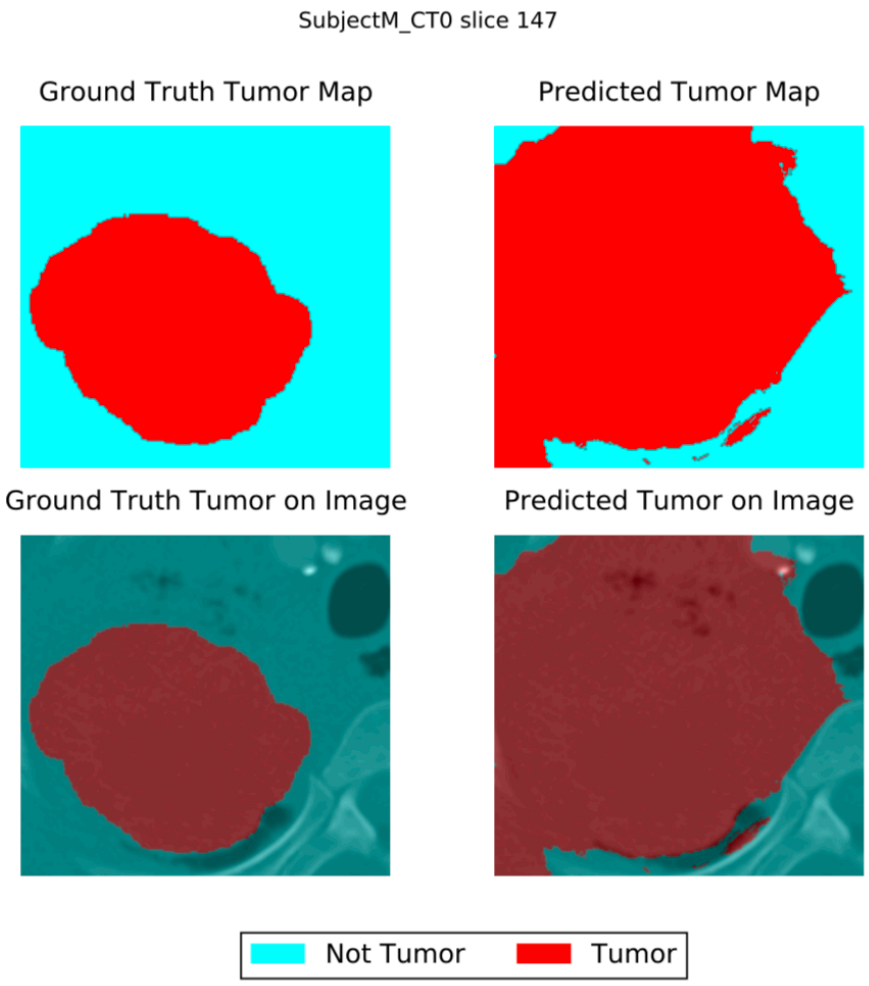


Figure 9: Comparison of the ground truth tumor location map with the predicted location map from the BinaryRes_Tumor network following CRF post processing. The top row represents the map of true and predicted labels while the bottom row shows the maps overlaid on the corresponding image slice.

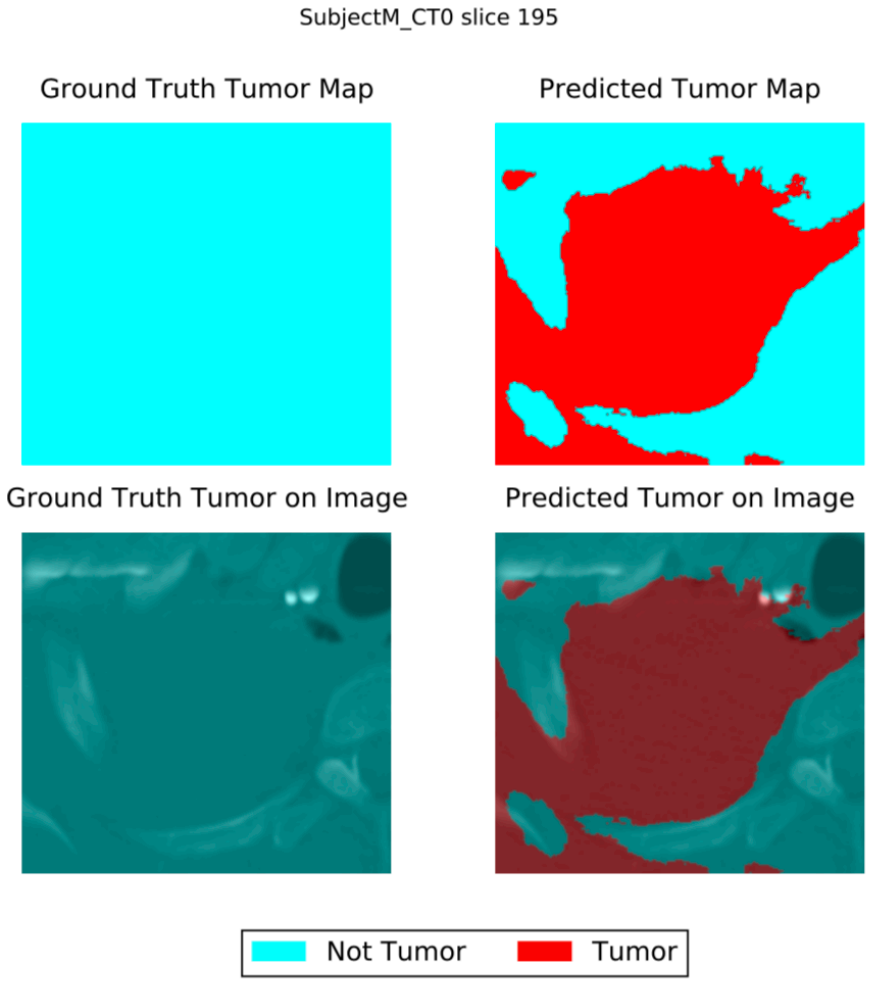


Figure 10: Comparison of the ground truth tumor location map with the predicted location map from the BinaryRes_Tumor network following CRF post processing. The top row represents the map of true and predicted labels while the bottom row shows the maps overlaid on the corresponding image slice.

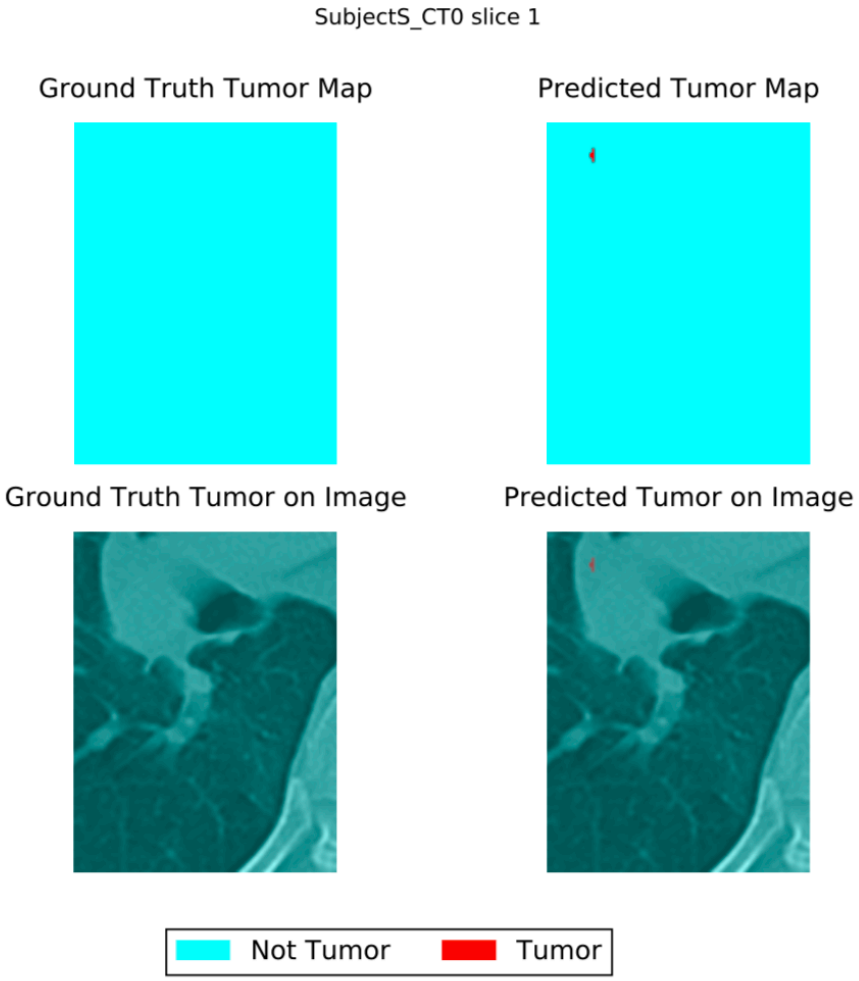


Figure 11: Comparison of the ground truth tumor location map with the predicted location map from the BinaryRes_Tumor network following CRF post processing. The top row represents the map of true and predicted labels while the bottom row shows the maps overlaid on the corresponding image slice.

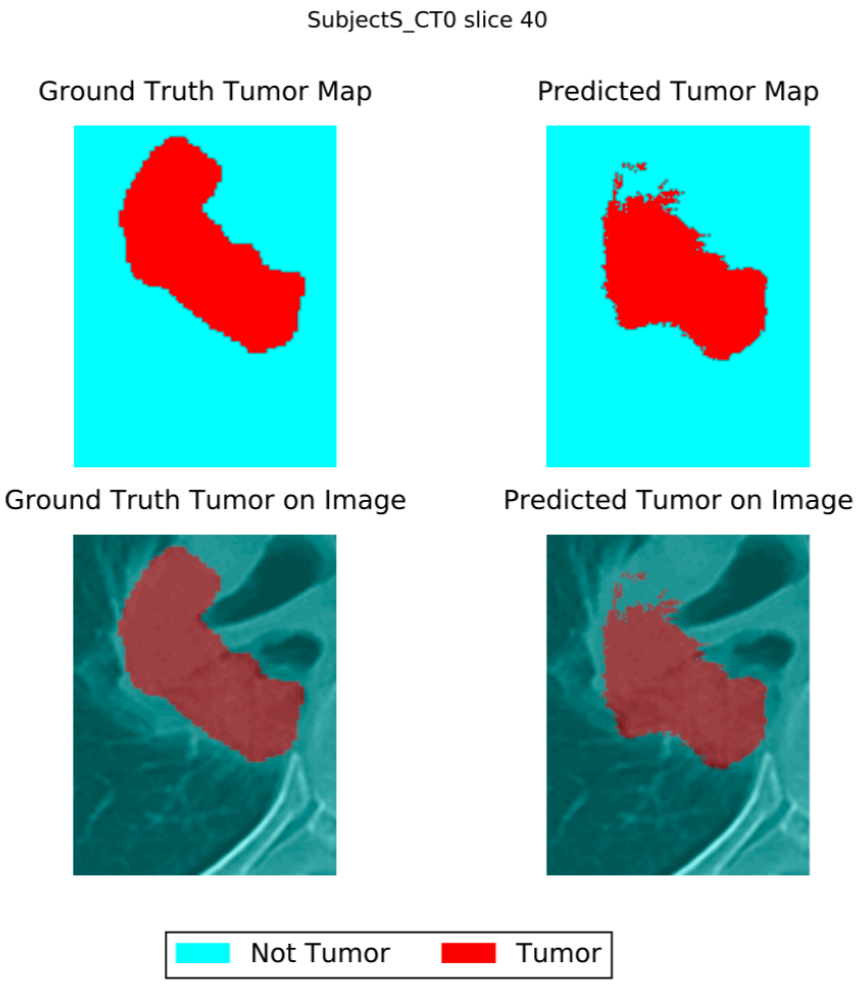


Figure 12: Comparison of the ground truth tumor location map with the predicted location map from the BinaryRes_Tumor network following CRF post processing. The top row represents the map of true and predicted labels while the bottom row shows the maps overlaid on the corresponding image slice.

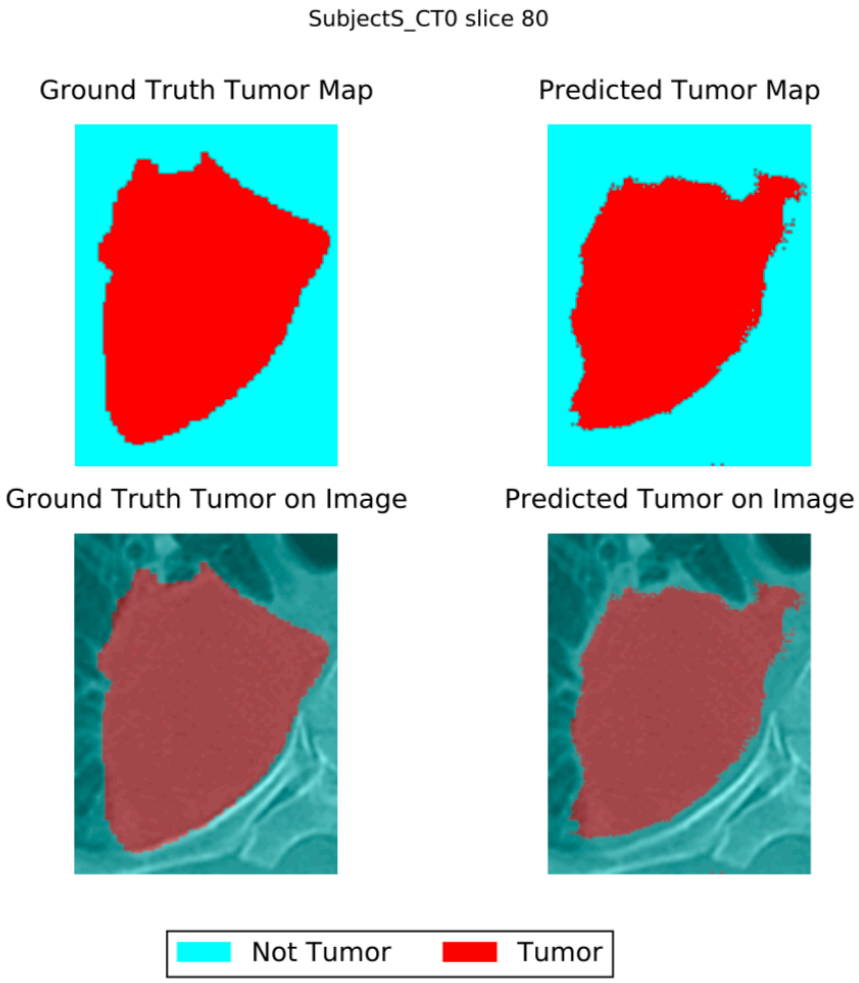


Figure 13: Comparison of the ground truth tumor location map with the predicted location map from the BinaryRes_Tumor network following CRF post processing. The top row represents the map of true and predicted labels while the bottom row shows the maps overlaid on the corresponding image slice.

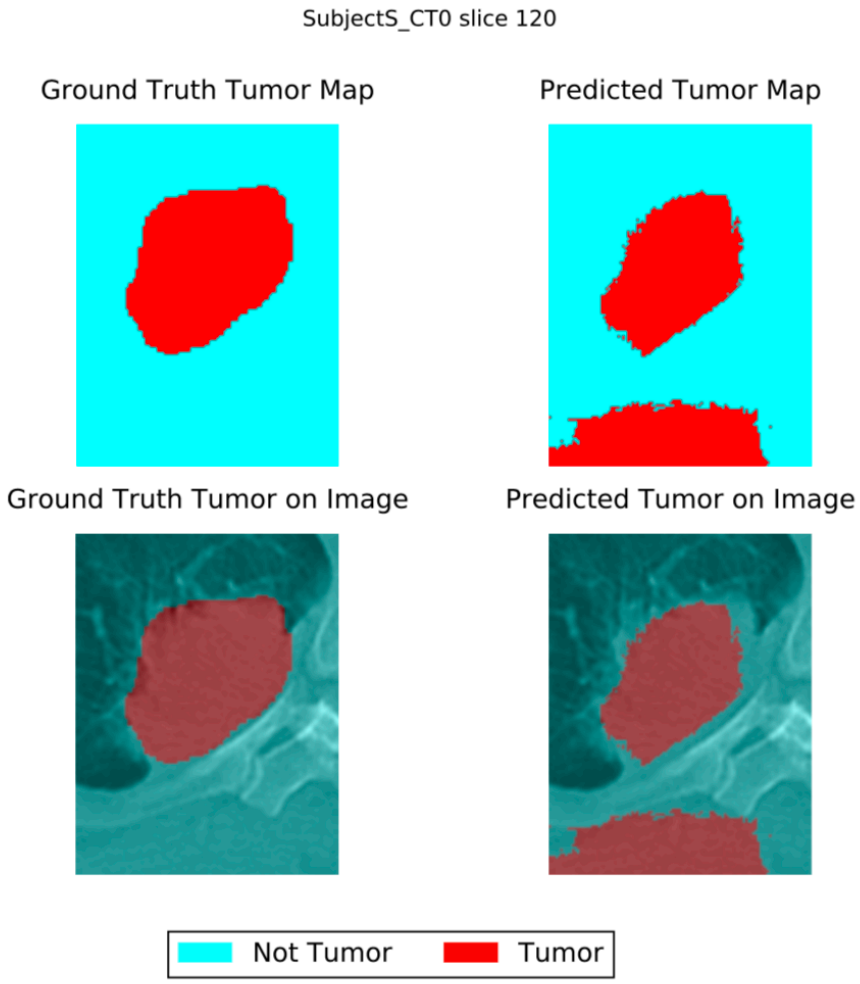


Figure 14: Comparison of the ground truth tumor location map with the predicted location map from the BinaryRes_Tumor network following CRF post processing. The top row represents the map of true and predicted labels while the bottom row shows the maps overlaid on the corresponding image slice.

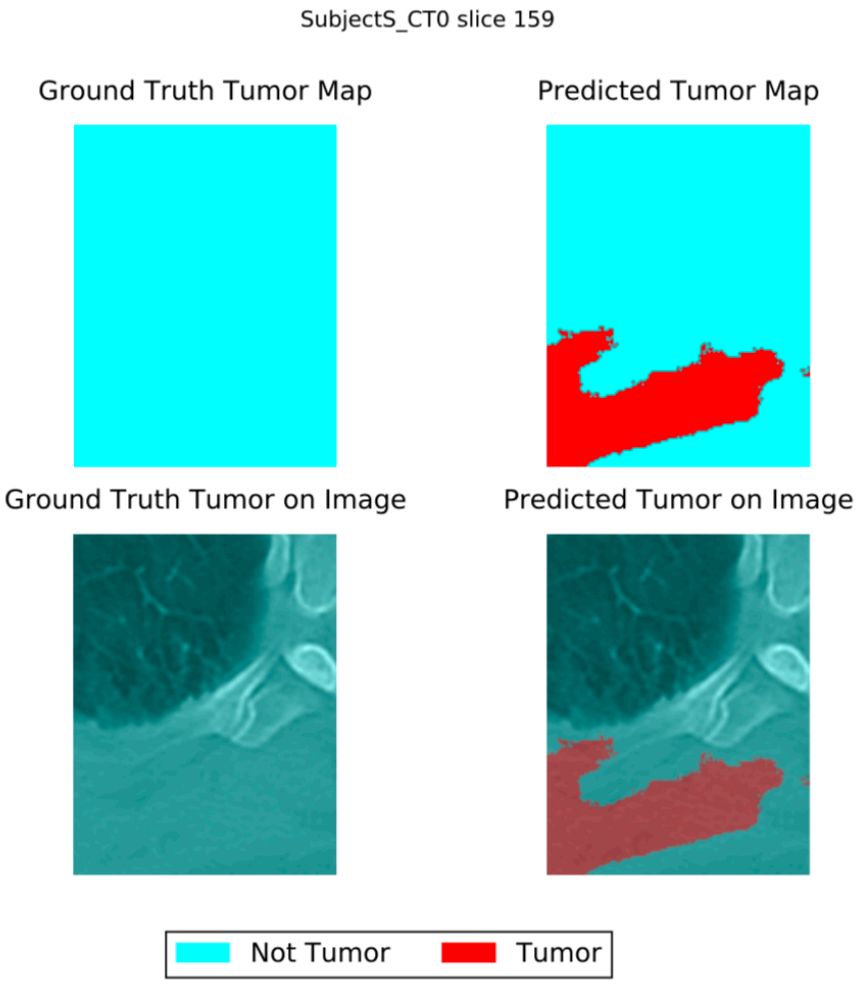


Figure 15: Comparison of the ground truth tumor location map with the predicted location map from the BinaryRes_Tumor network following CRF post processing. The top row represents the map of true and predicted labels while the bottom row shows the maps overlaid on the corresponding image slice.

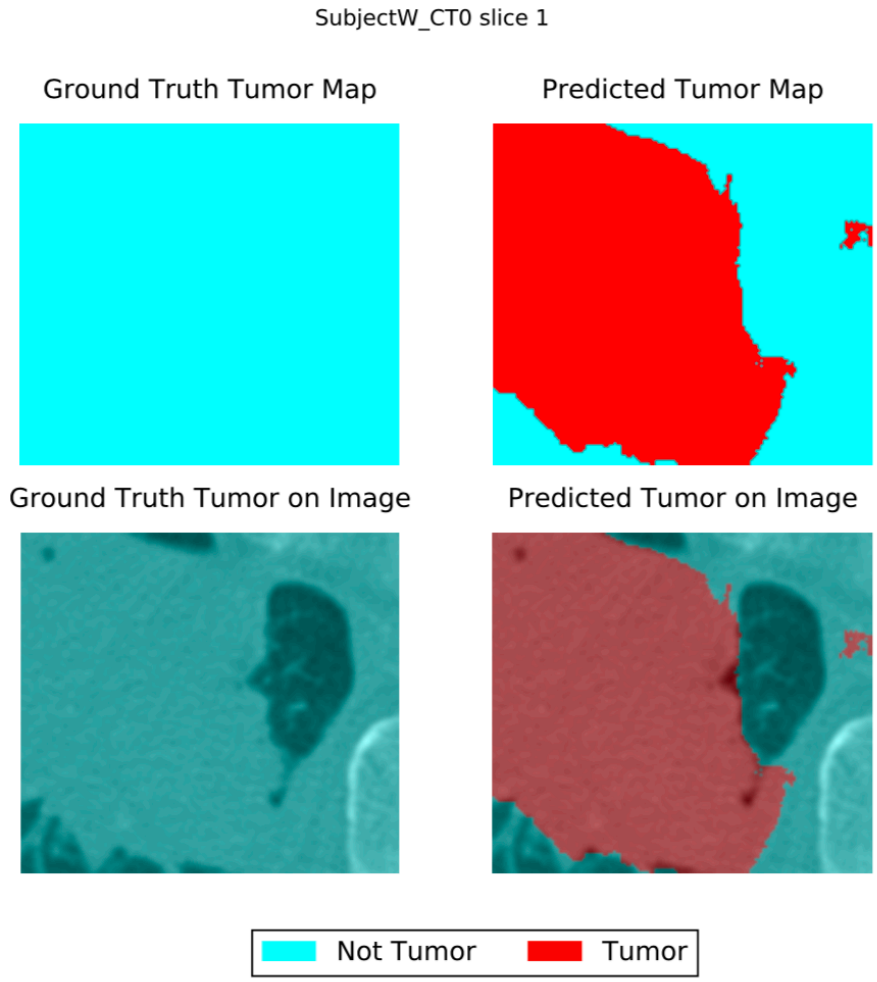


Figure 16: Comparison of the ground truth tumor location map with the predicted location map from the BinaryRes_Tumor network following CRF post processing. The top row represents the map of true and predicted labels while the bottom row shows the maps overlaid on the corresponding image slice.

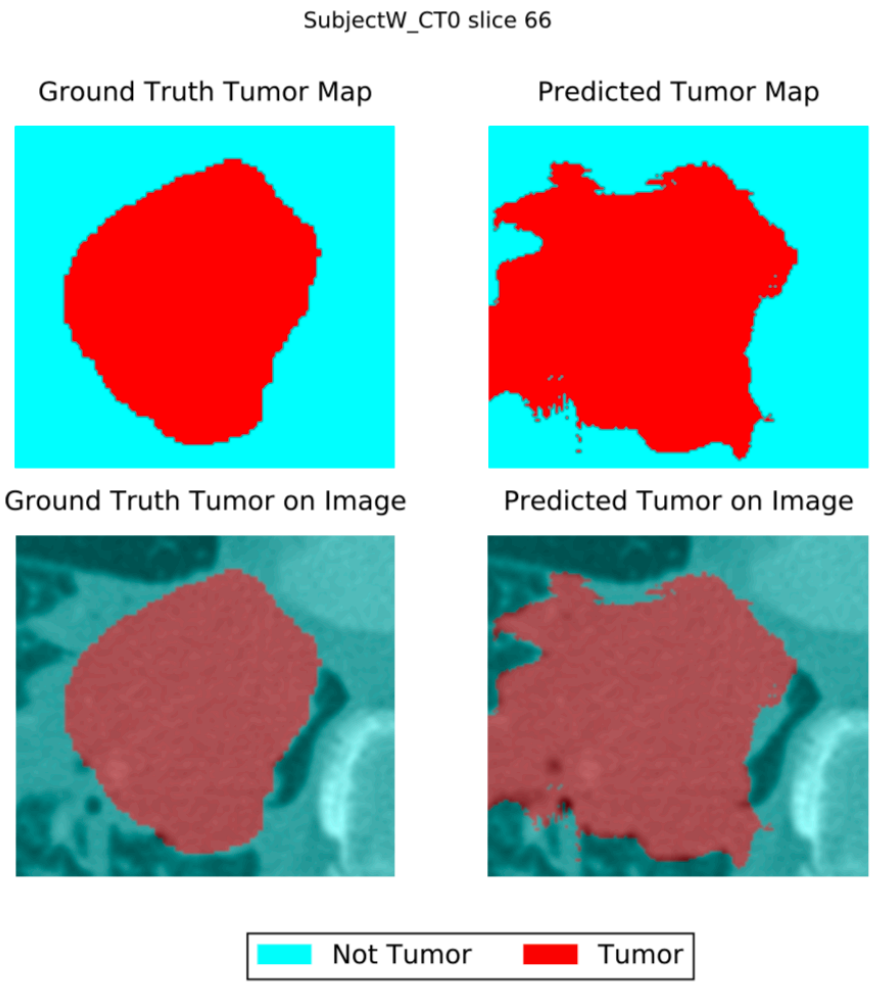


Figure 17: Comparison of the ground truth tumor location map with the predicted location map from the BinaryRes_Tumor network following CRF post processing. The top row represents the map of true and predicted labels while the bottom row shows the maps overlaid on the corresponding image slice.

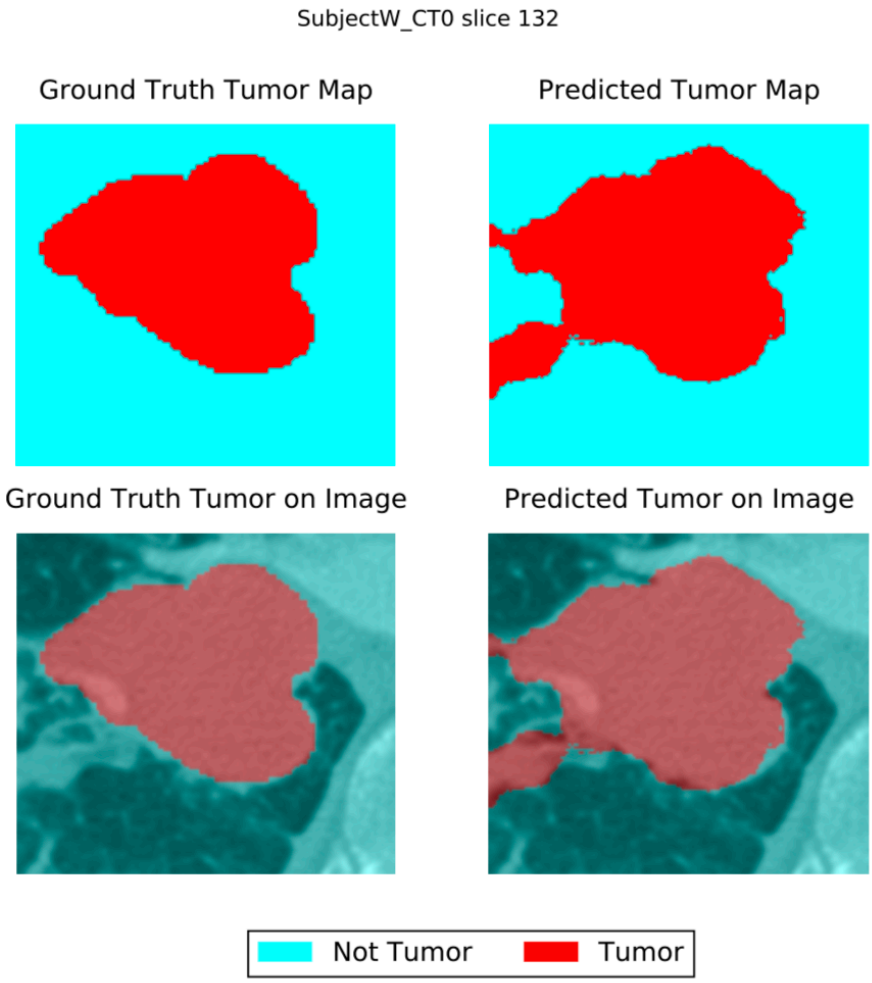


Figure 18: Comparison of the ground truth tumor location map with the predicted location map from the BinaryRes_Tumor network following CRF post processing. The top row represents the map of true and predicted labels while the bottom row shows the maps overlaid on the corresponding image slice.

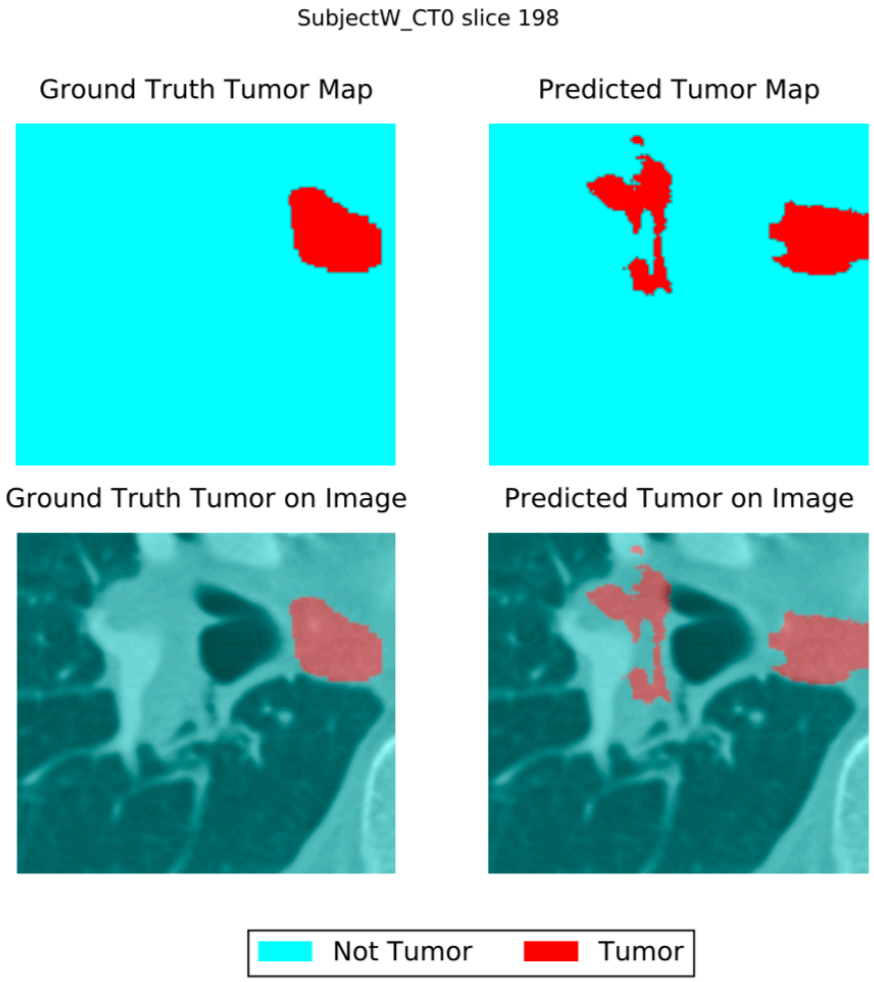


Figure 19: Comparison of the ground truth tumor location map with the predicted location map from the BinaryRes_Tumor network following CRF post processing. The top row represents the map of true and predicted labels while the bottom row shows the maps overlaid on the corresponding image slice.

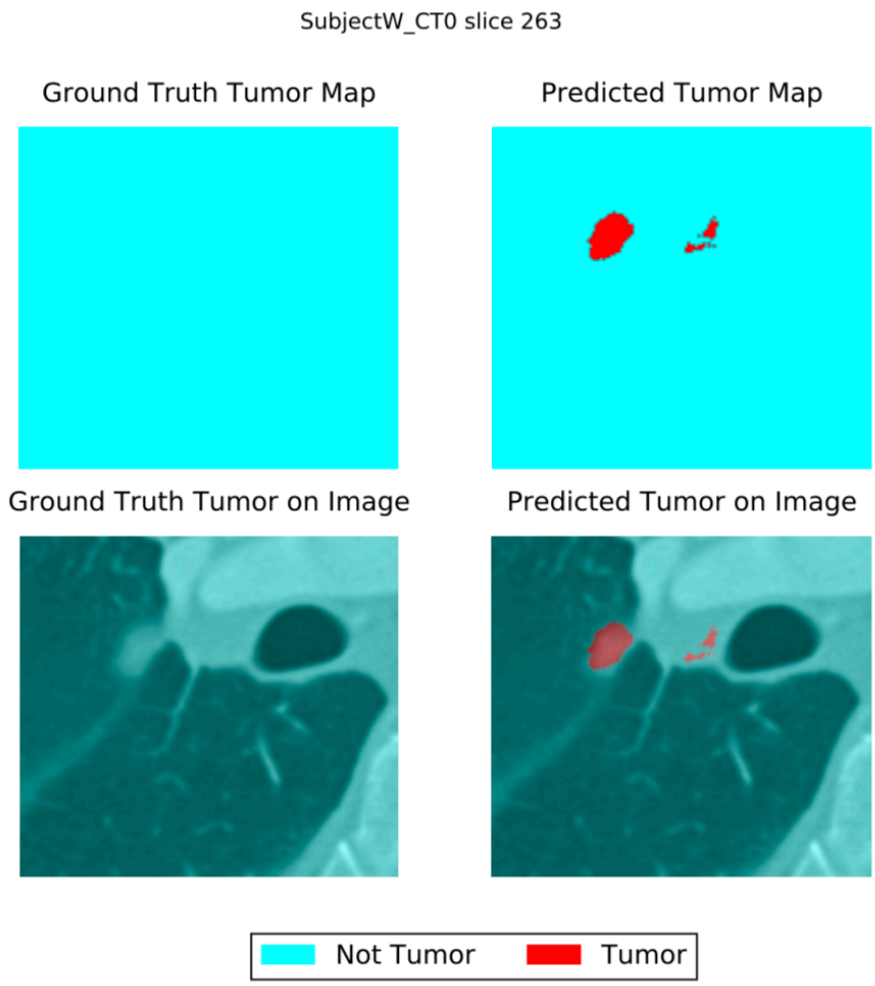


Figure 20: Comparison of the ground truth tumor location map with the predicted location map from the BinaryRes_Tumor network following CRF post processing. The top row represents the map of true and predicted labels while the bottom row shows the maps overlaid on the corresponding image slice.

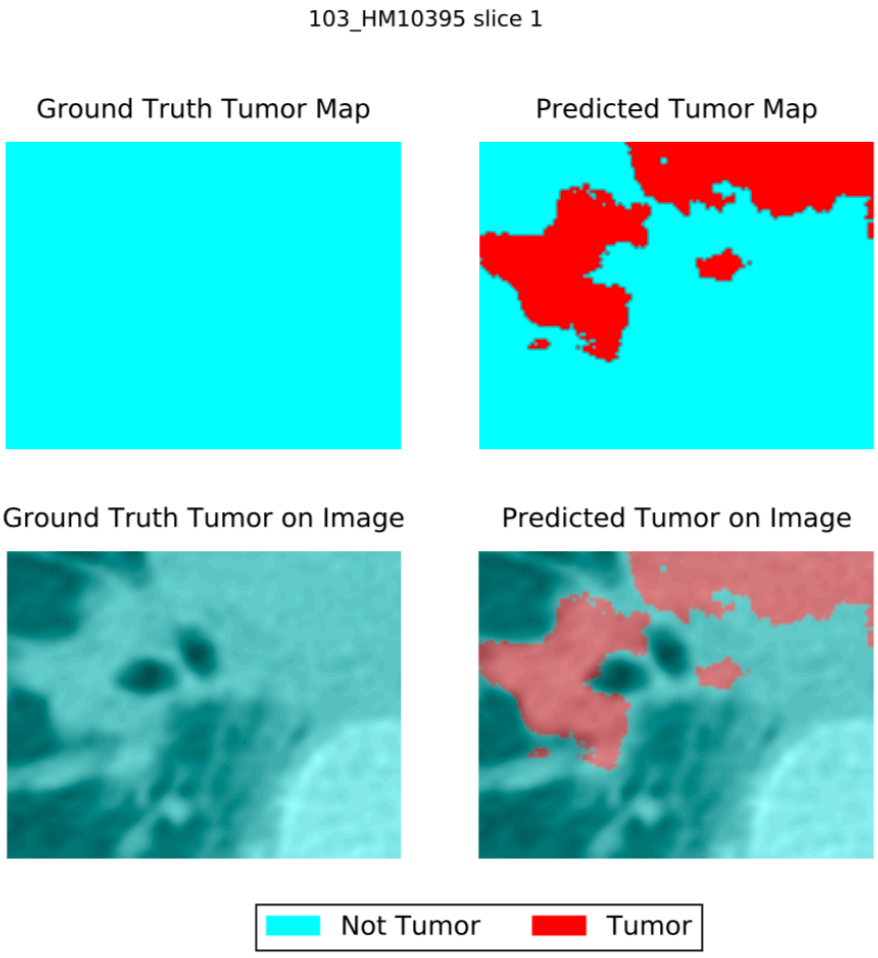


Figure 21: Comparison of the ground truth tumor location map with the predicted location map from the BinaryRes_Tumor network following CRF post processing. The top row represents the map of true and predicted labels while the bottom row shows the maps overlaid on the corresponding image slice.

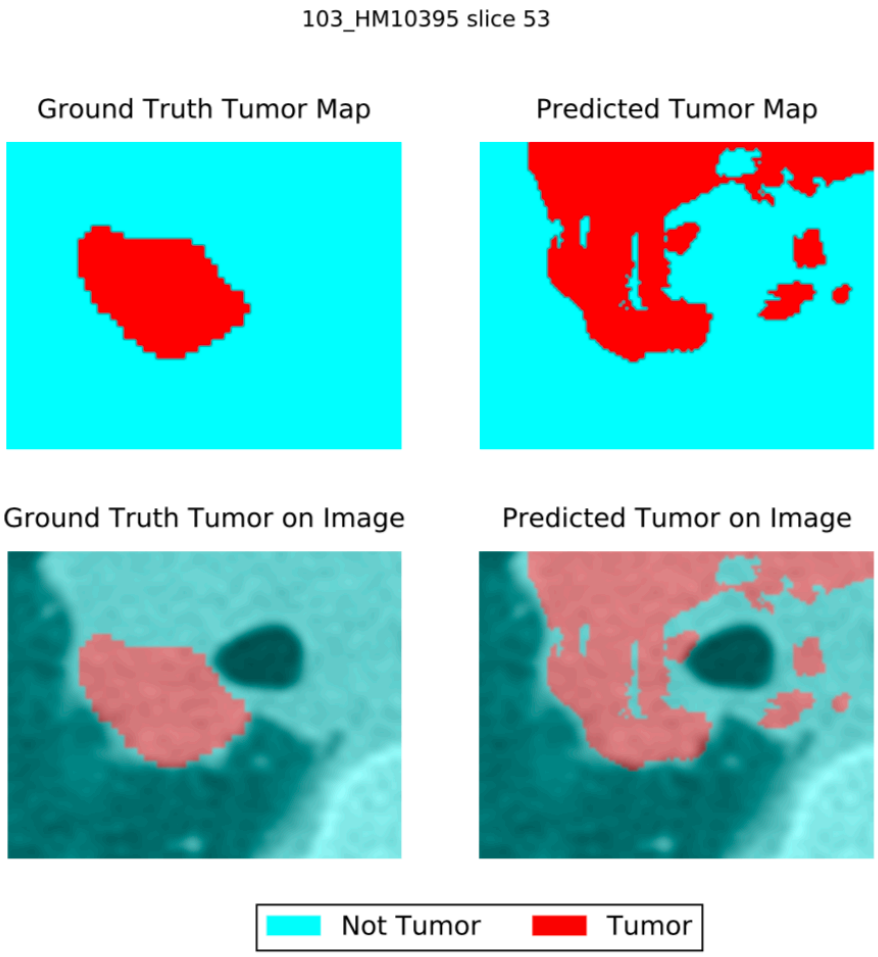


Figure 22: Comparison of the ground truth tumor location map with the predicted location map from the BinaryRes_Tumor network following CRF post processing. The top row represents the map of true and predicted labels while the bottom row shows the maps overlaid on the corresponding image slice.

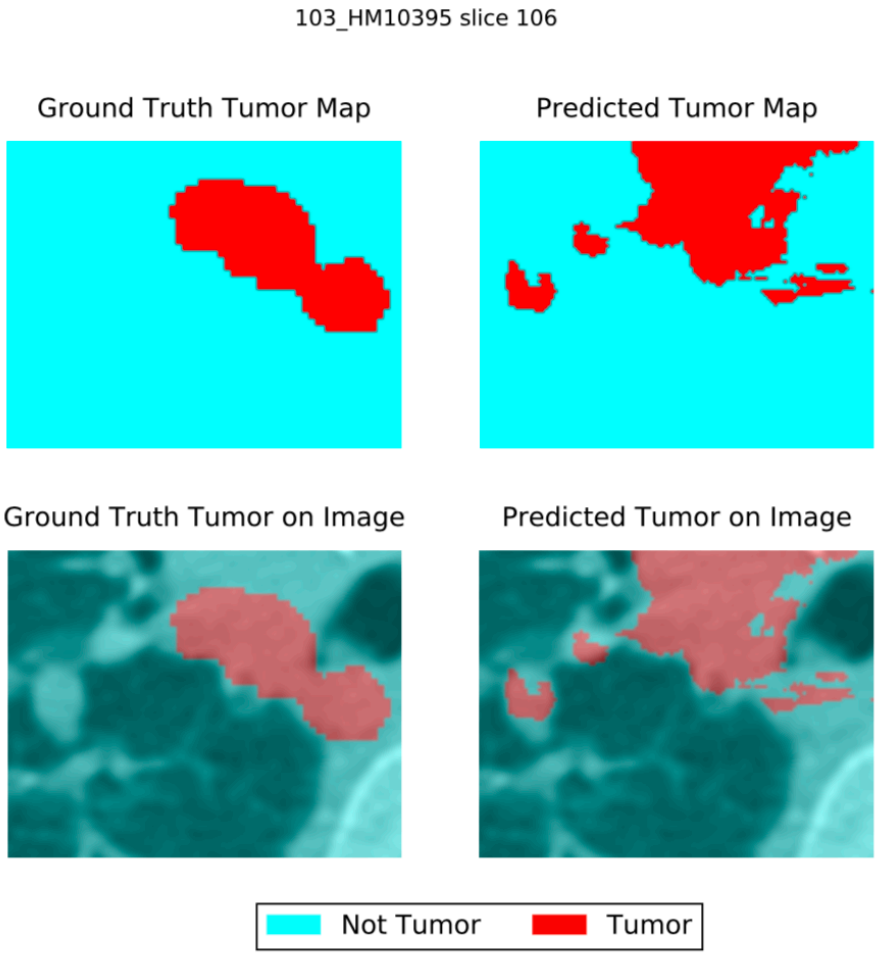


Figure 23: Comparison of the ground truth tumor location map with the predicted location map from the BinaryRes_Tumor network following CRF post processing. The top row represents the map of true and predicted labels while the bottom row shows the maps overlaid on the corresponding image slice.

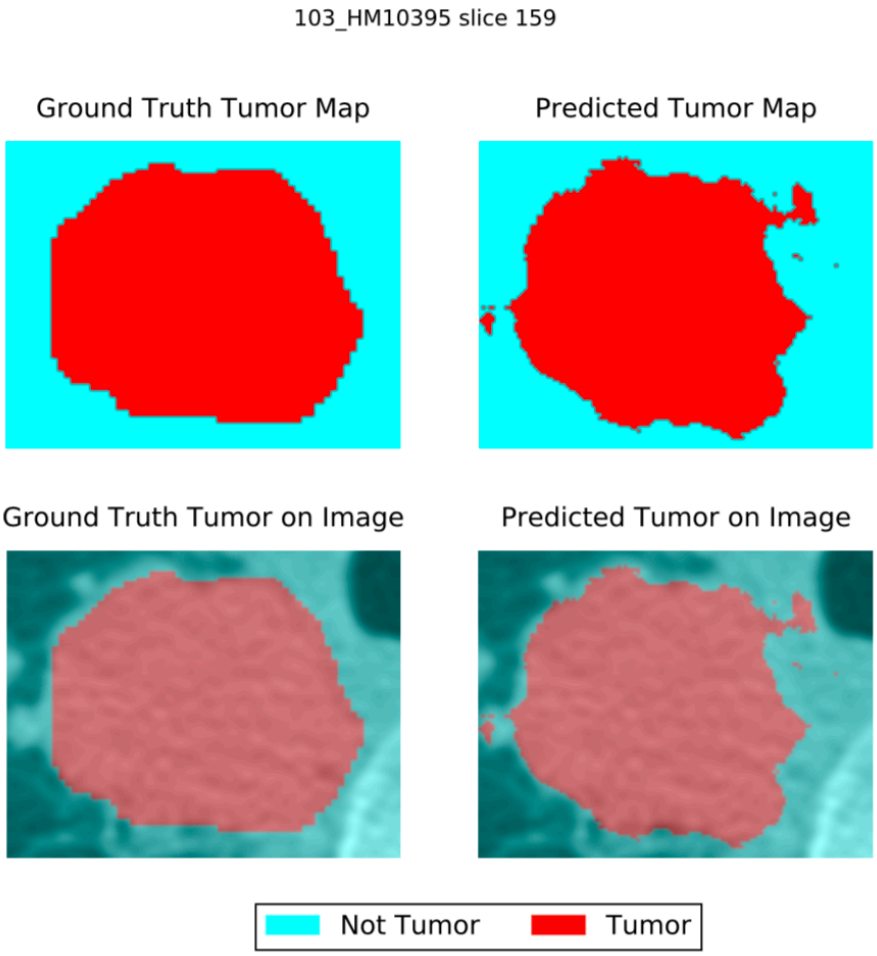


Figure 24: Comparison of the ground truth tumor location map with the predicted location map from the BinaryRes_Tumor network following CRF post processing. The top row represents the map of true and predicted labels while the bottom row shows the maps overlaid on the corresponding image slice.

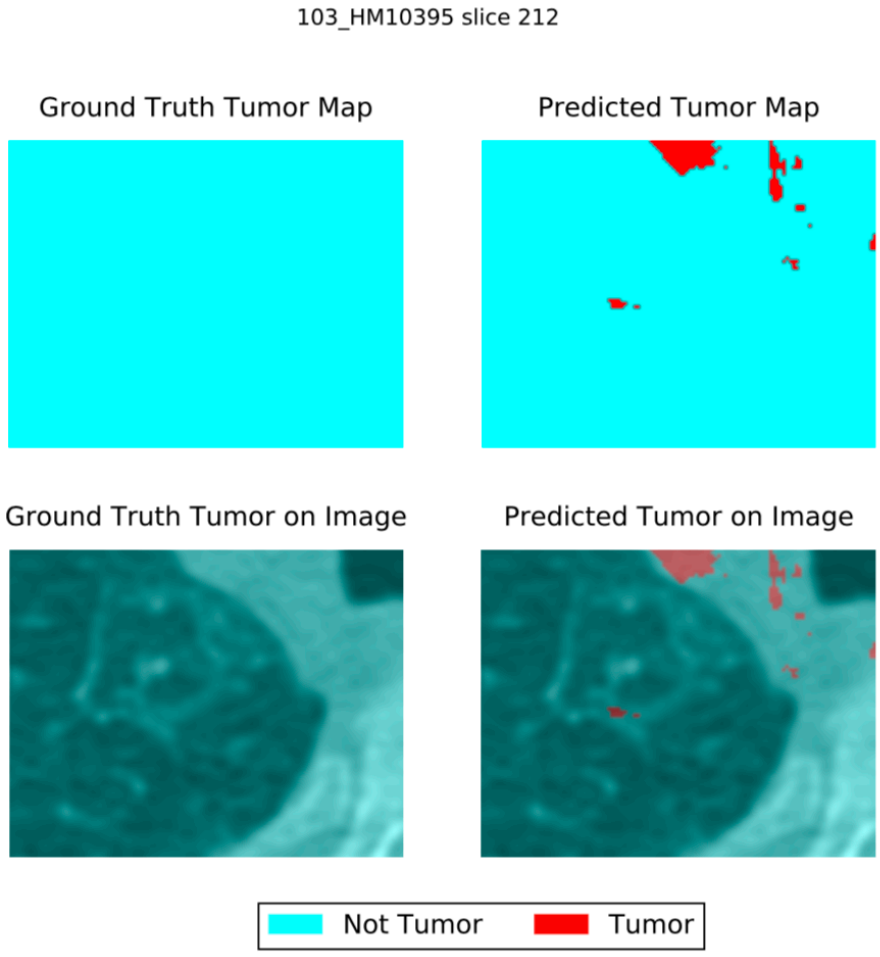


Figure 25: Comparison of the ground truth tumor location map with the predicted location map from the BinaryRes_Tumor network following CRF post processing. The top row represents the map of true and predicted labels while the bottom row shows the maps overlaid on the corresponding image slice.

Appendix VI

Examples of prediction map crated from the IF_Only network output. Illustrative slice from all test subjects spread evenly throughout the interface contours.

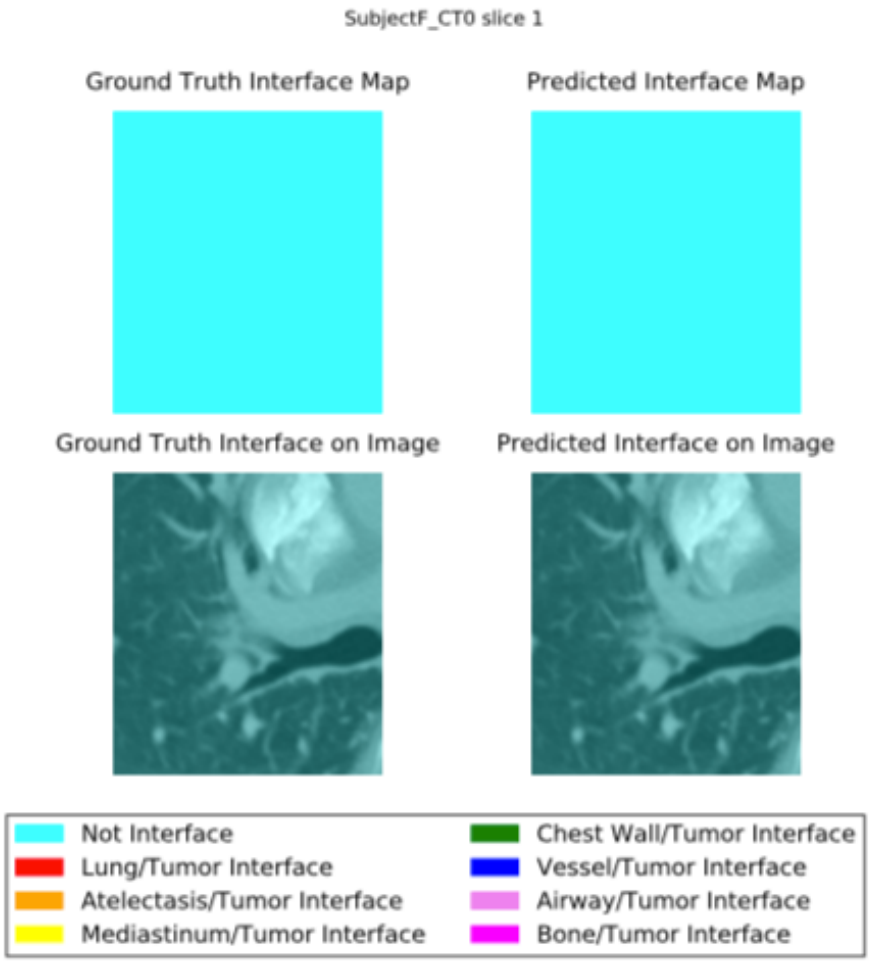


Figure 1: Comparison of the ground truth interface labels map with the predicated labels map from the IF_Only network no post processing. The top row represents the map of true and predicted labels while the bottom row shows the maps overlaid on the corresponding image slice.

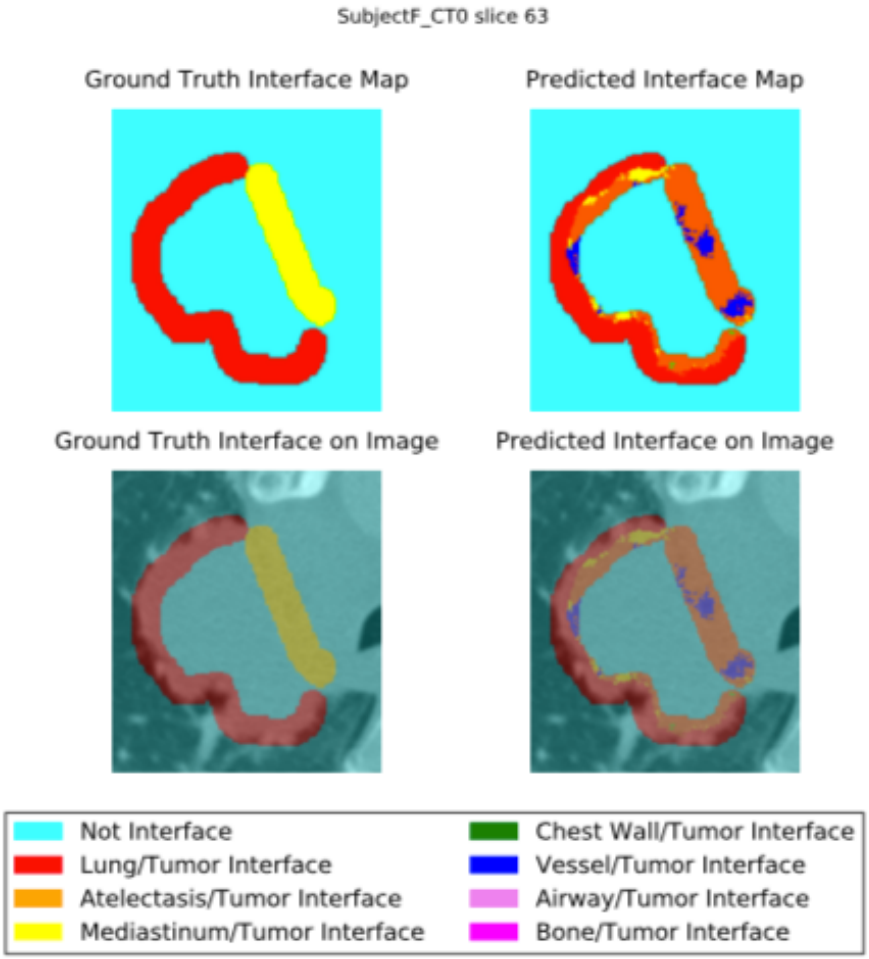


Figure 2: Comparison of the ground truth interface labels map with the predicated labels map from the IF_Only network no post processing. The top row represents the map of true and predicted labels while the bottom row shows the maps overlaid on the corresponding image slice.

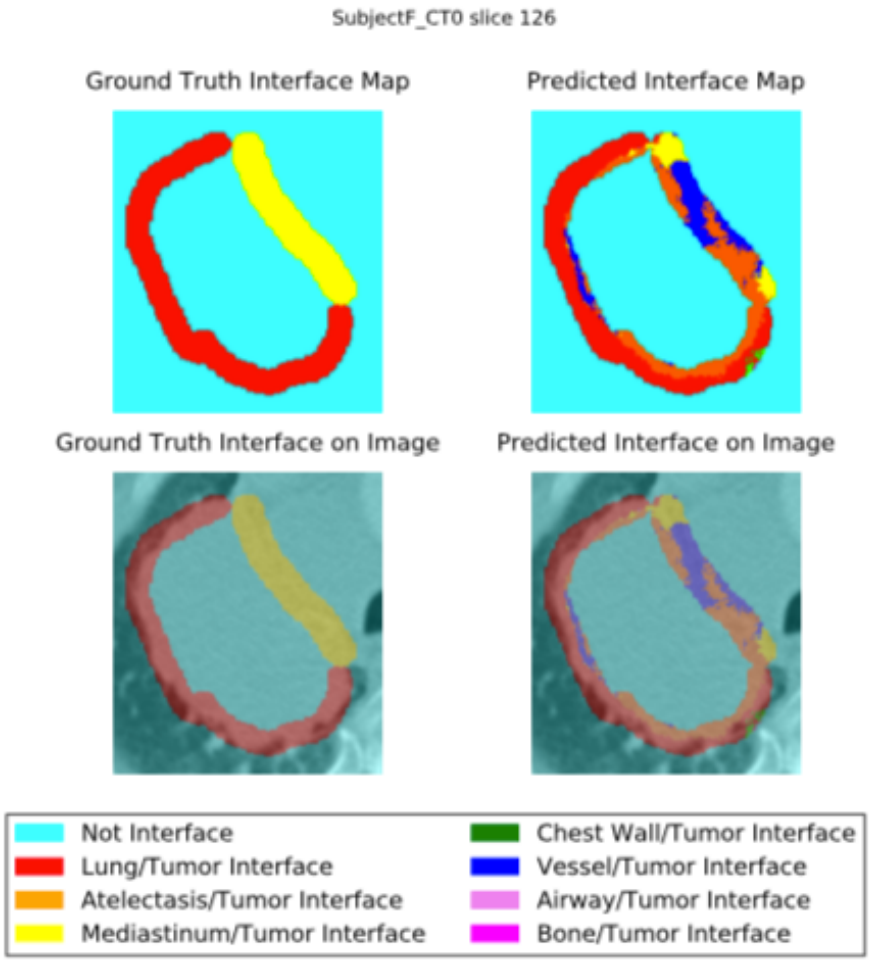


Figure 3: Comparison of the ground truth interface labels map with the predicated labels map from the IF_Only network no post processing. The top row represents the map of true and predicted labels while the bottom row shows the maps overlaid on the corresponding image slice.

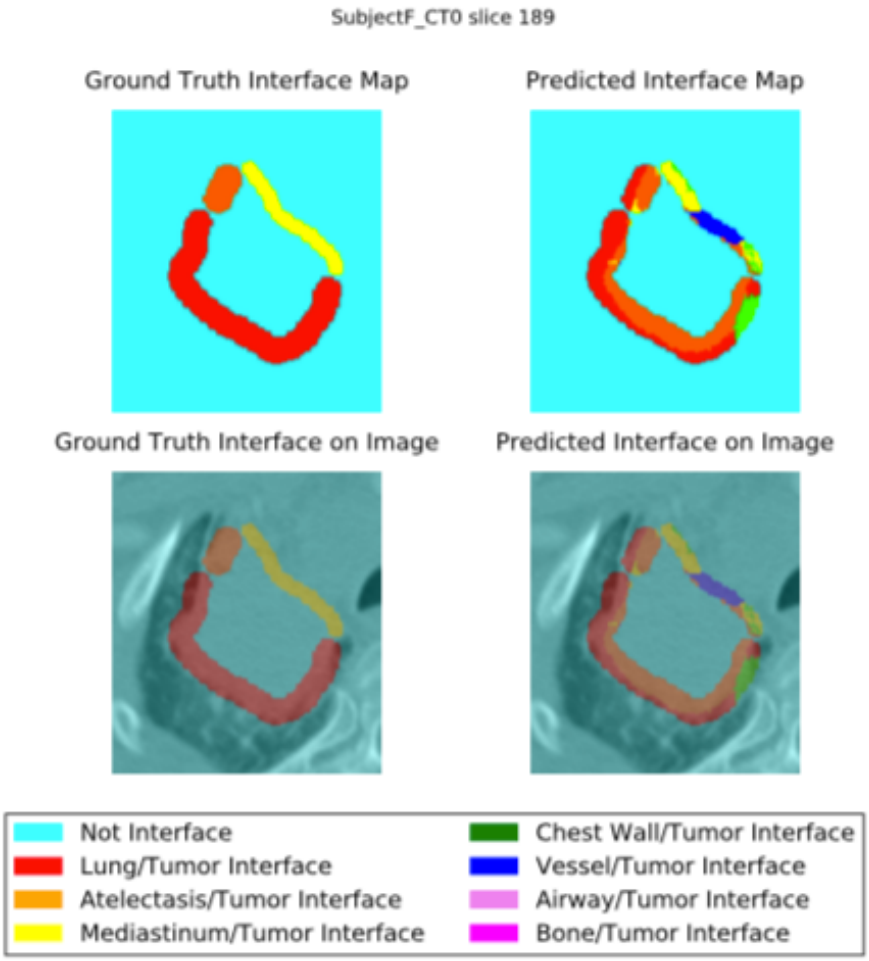


Figure 5: Comparison of the ground truth interface labels map with the predicated labels map from the IF_Only network no post processing. The top row represents the map of true and predicted labels while the bottom row shows the maps overlaid on the corresponding image slice.

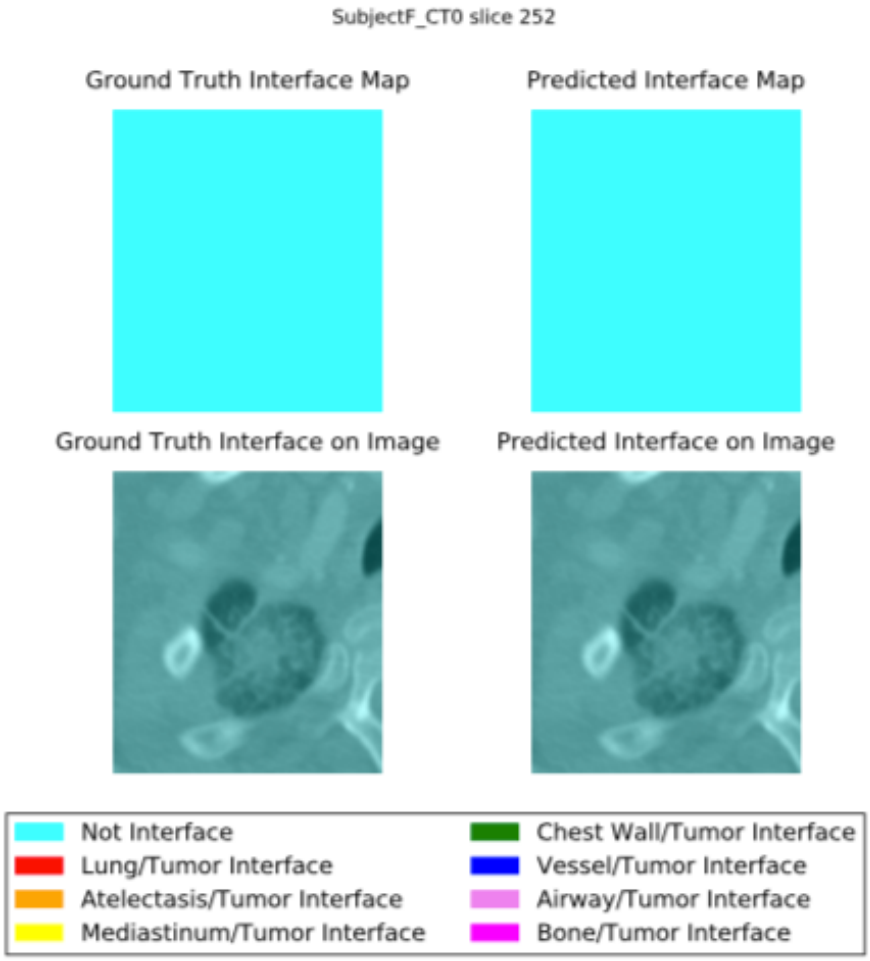


Figure 4: Comparison of the ground truth interface labels map with the predicated labels map from the IF_Only network no post processing. The top row represents the map of true and predicted labels while the bottom row shows the maps overlaid on the corresponding image slice.

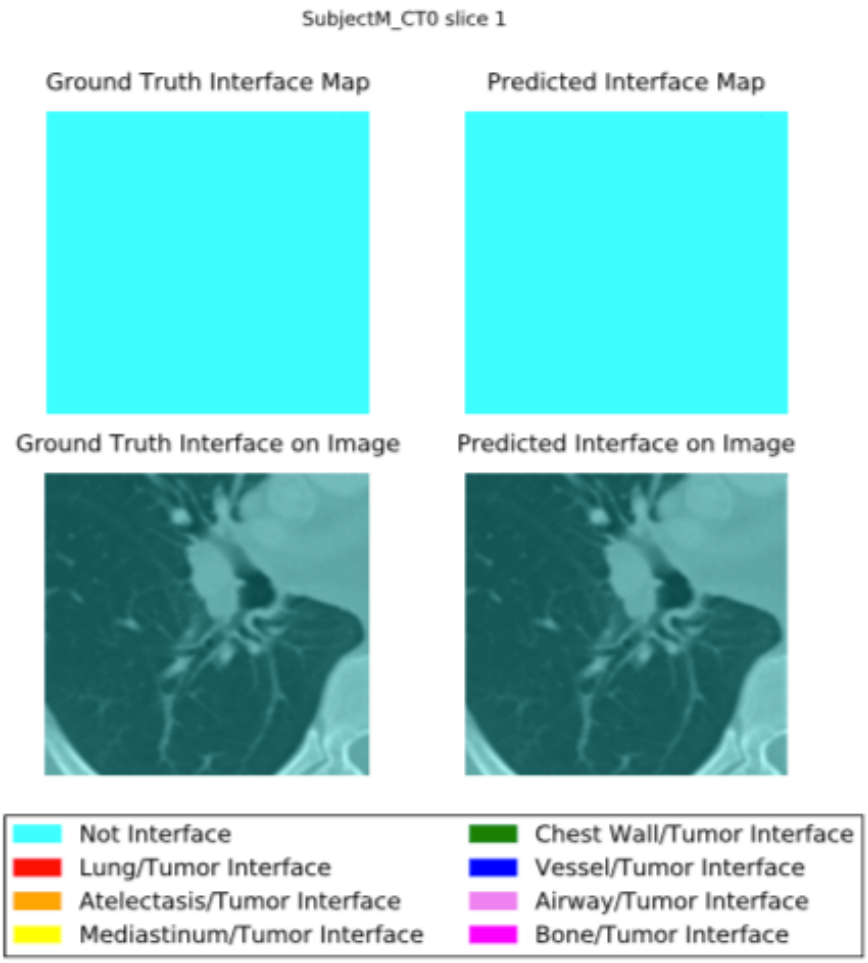


Figure 6: Comparison of the ground truth interface labels map with the predicated labels map from the IF_Only network no post processing. The top row represents the map of true and predicted labels while the bottom row shows the maps overlaid on the corresponding image slice.

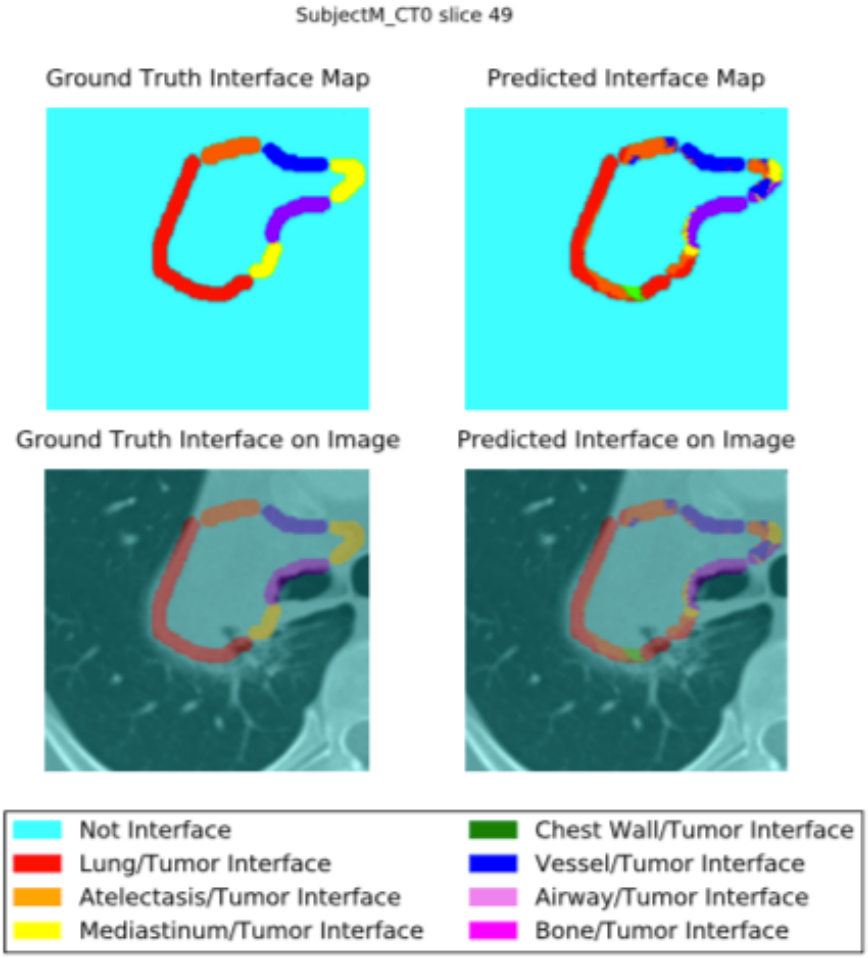


Figure 7: Comparison of the ground truth interface labels map with the predicated labels map from the IF_Only network no post processing. The top row represents the map of true and predicted labels while the bottom row shows the maps overlaid on the corresponding image slice.

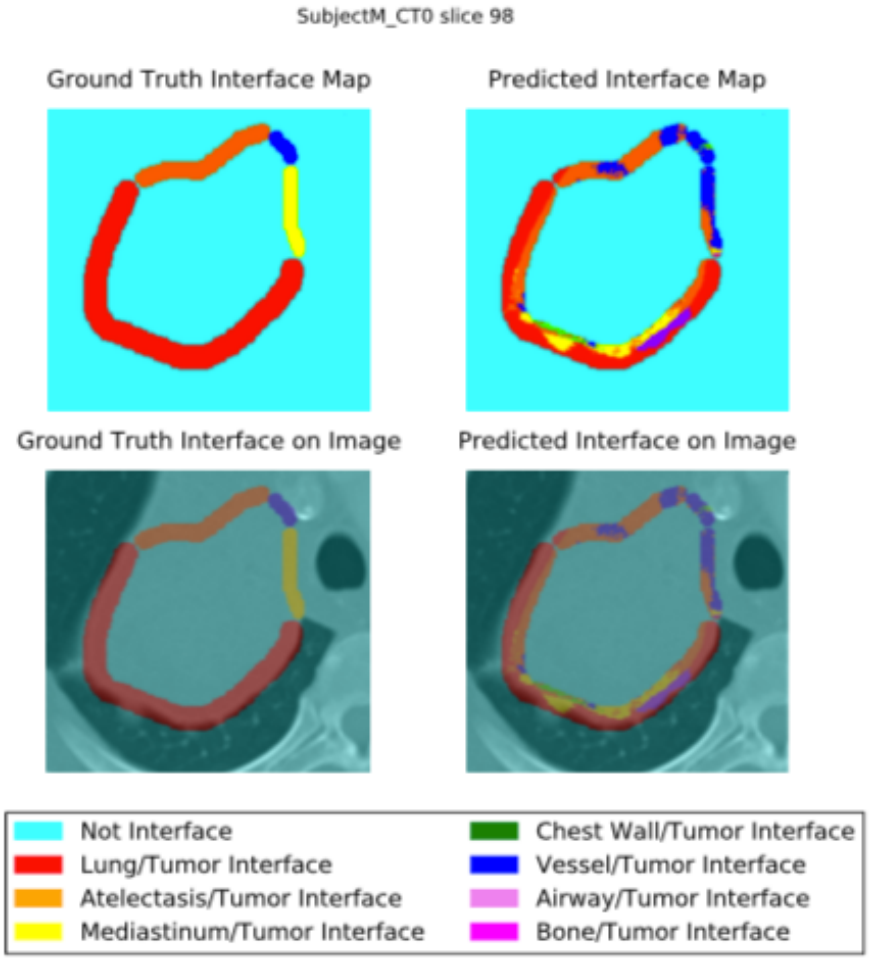


Figure 8: Comparison of the ground truth interface labels map with the predicated labels map from the IF_Only network no post processing. The top row represents the map of true and predicted labels while the bottom row shows the maps overlaid on the corresponding image slice.

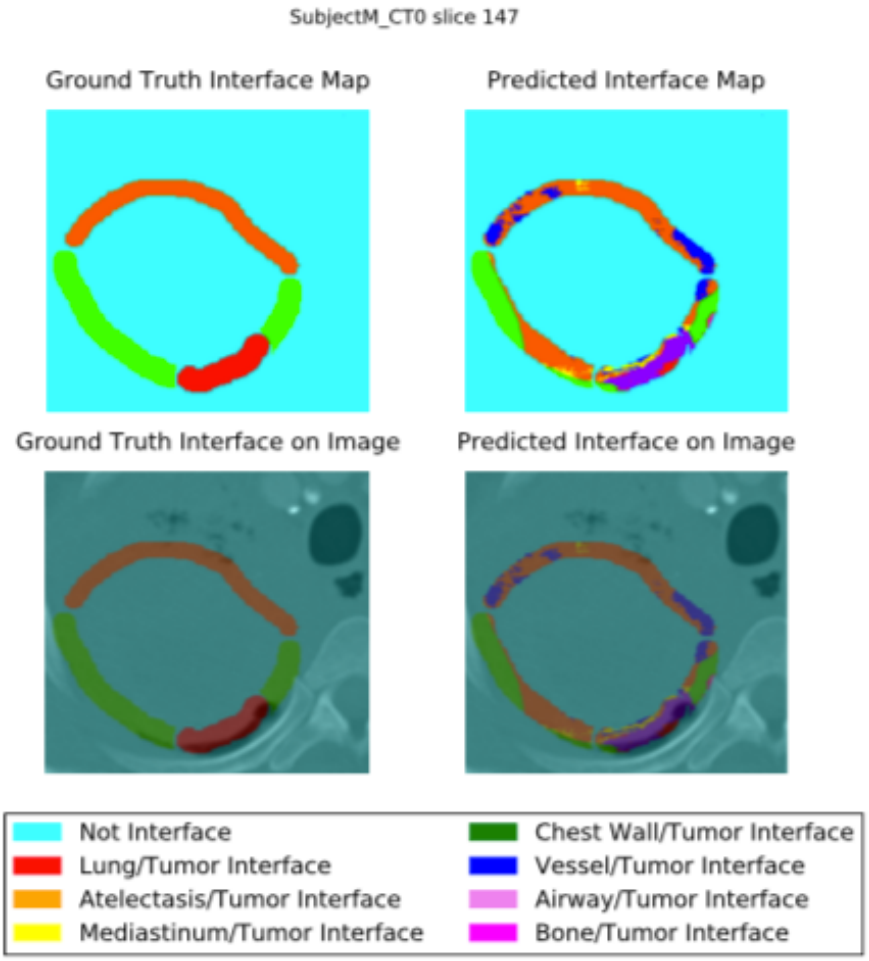


Figure 9: Comparison of the ground truth interface labels map with the predicated labels map from the IF_Only network no post processing. The top row represents the map of true and predicted labels while the bottom row shows the maps overlaid on the corresponding image slice.

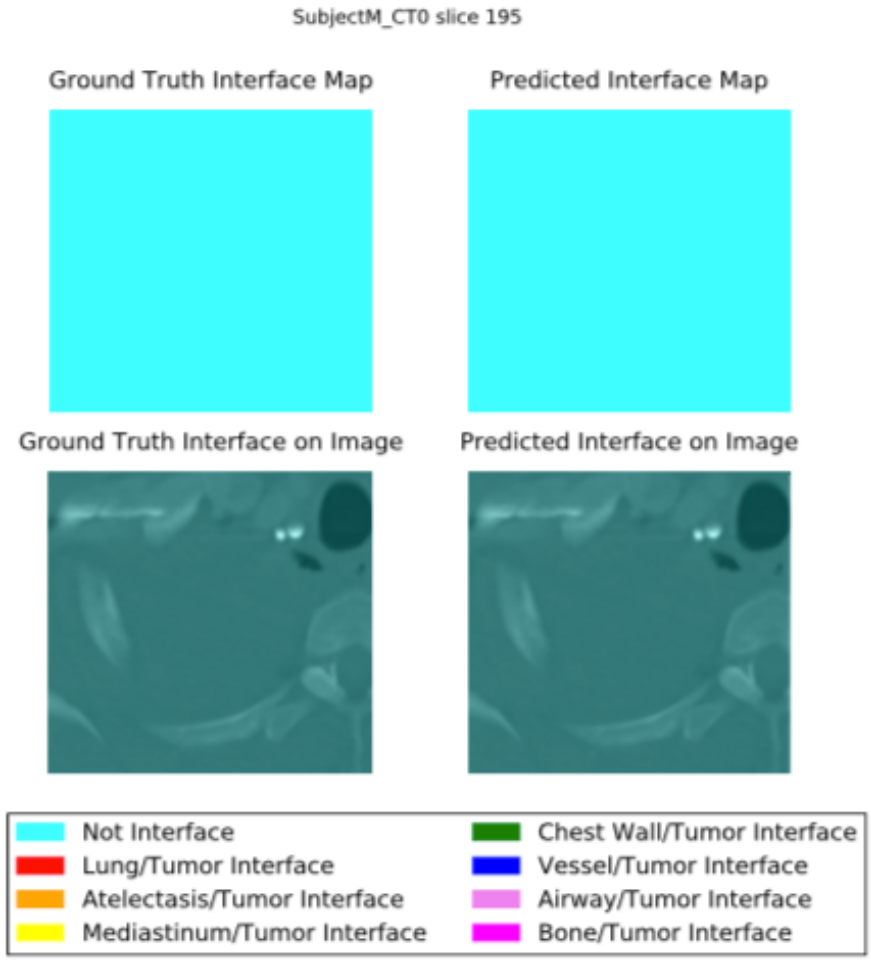


Figure 10: Comparison of the ground truth interface labels map with the predicated labels map from the IF_Only network no post processing. The top row represents the map of true and predicted labels while the bottom row shows the maps overlaid on the corresponding image slice.

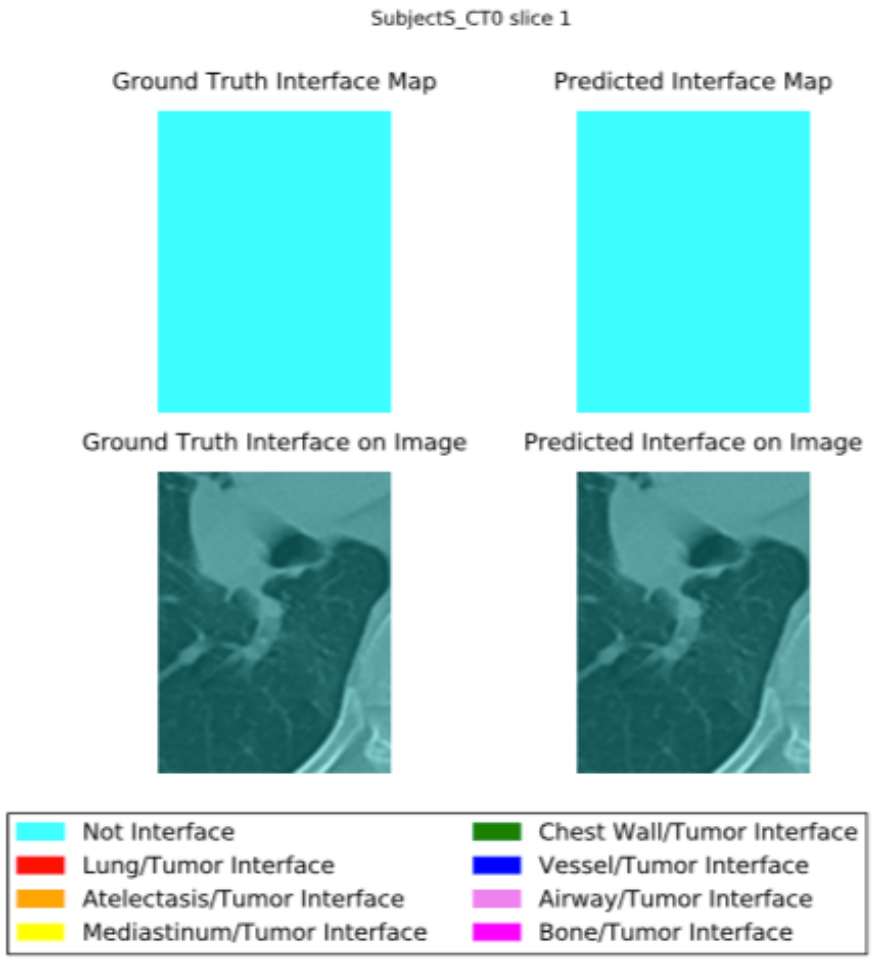


Figure 11: Comparison of the ground truth interface labels map with the predicated labels map from the IF_Only network no post processing. The top row represents the map of true and predicted labels while the bottom row shows the maps overlaid on the corresponding image slice.

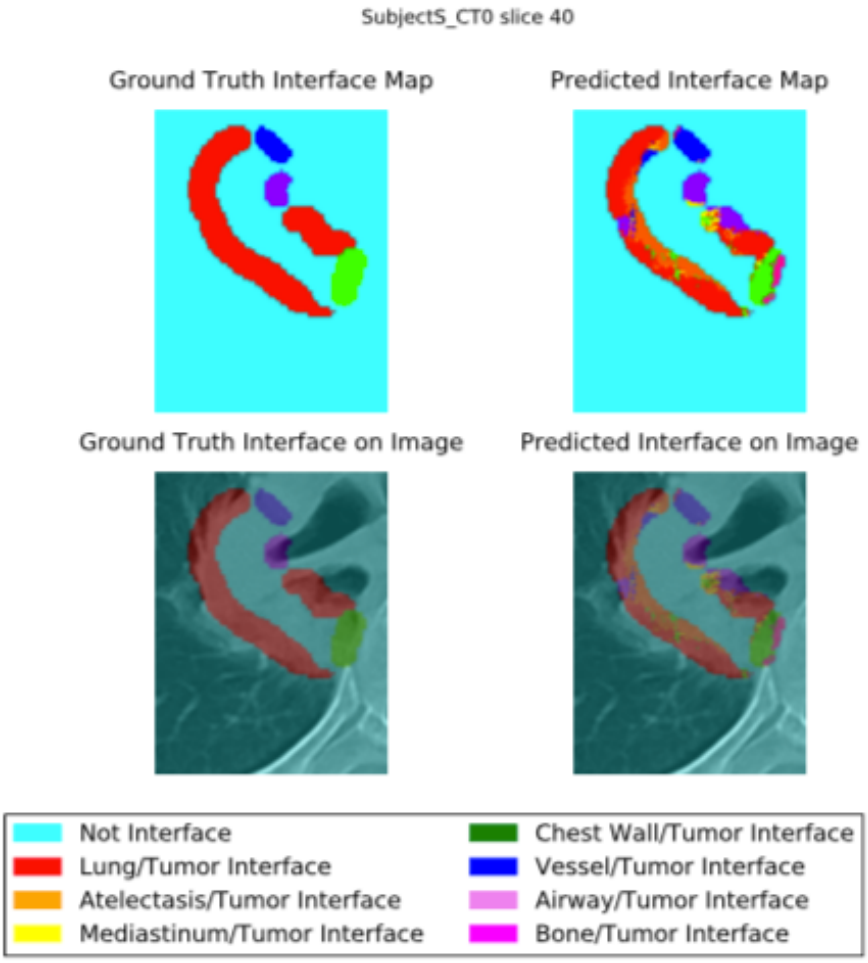


Figure 12: Comparison of the ground truth interface labels map with the predicated labels map from the IF_Only network no post processing. The top row represents the map of true and predicted labels while the bottom row shows the maps overlaid on the corresponding image slice.

c

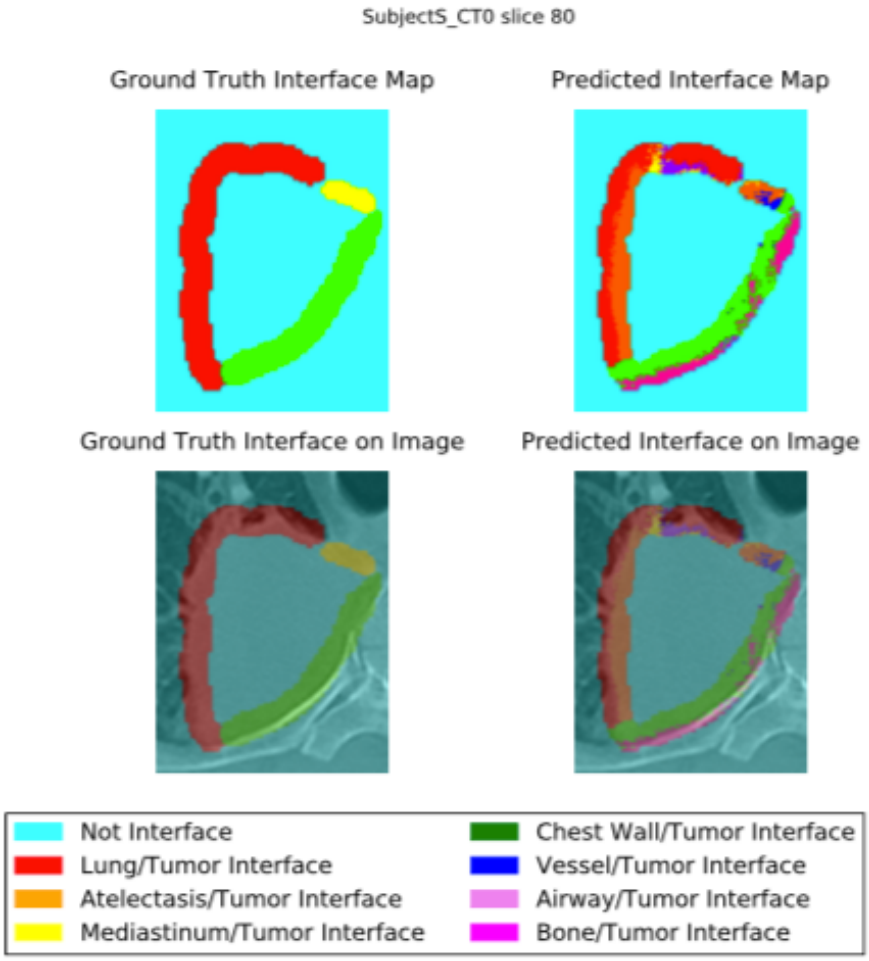


Figure 13: Comparison of the ground truth interface labels map with the predicated labels map from the IF_Only network no post processing. The top row represents the map of true and predicted labels while the bottom row shows the maps overlaid on the corresponding image slice.

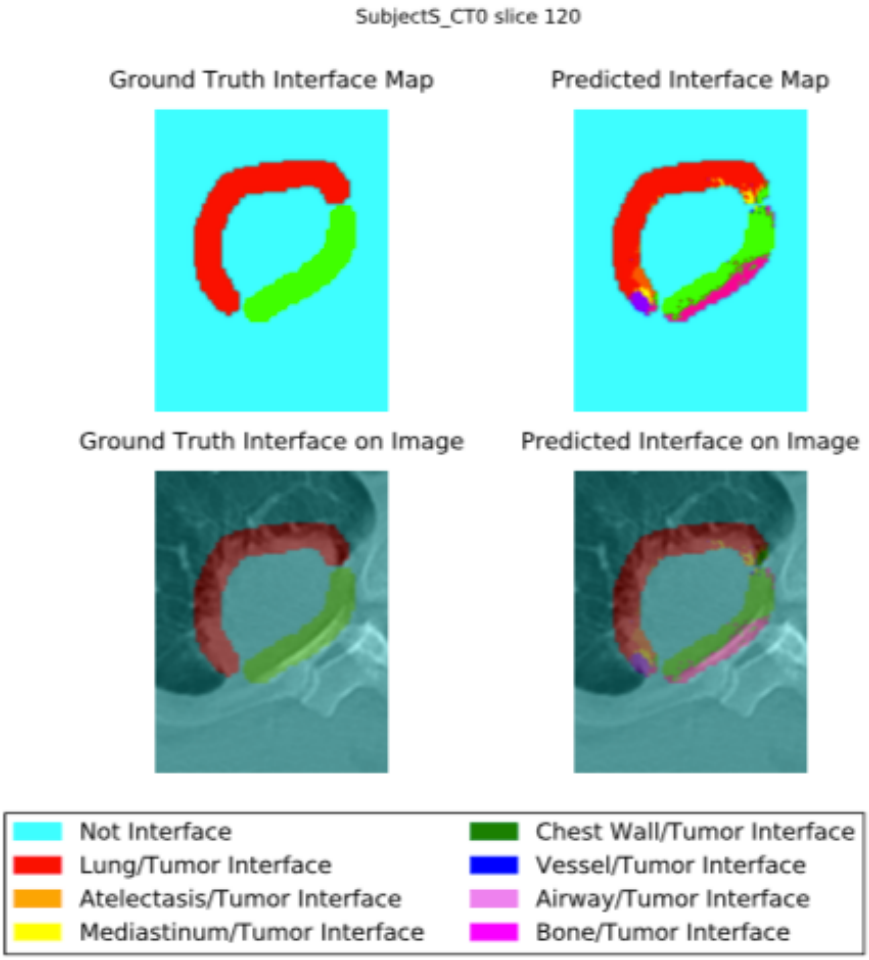


Figure 14: Comparison of the ground truth interface labels map with the predicated labels map from the IF_Only network no post processing. The top row represents the map of true and predicted labels while the bottom row shows the maps overlaid on the corresponding image slice.

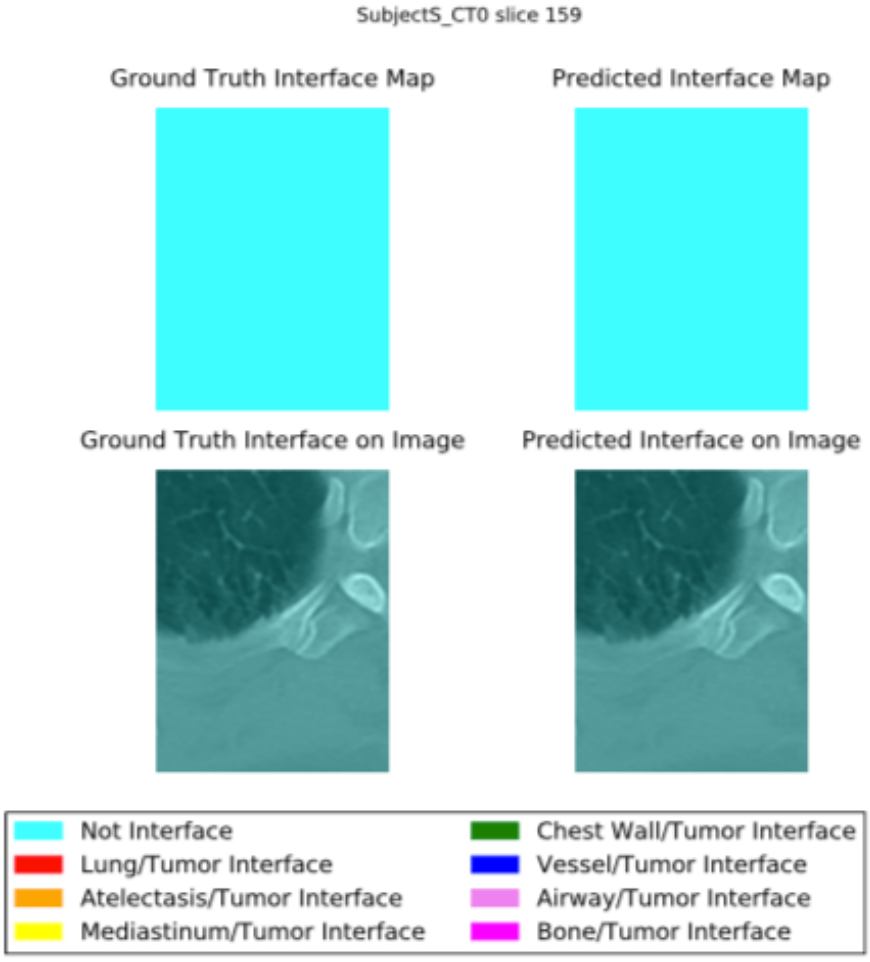


Figure 15: Comparison of the ground truth interface labels map with the predicated labels map from the IF_Only network no post processing. The top row represents the map of true and predicted labels while the bottom row shows the maps overlaid on the corresponding image slice.

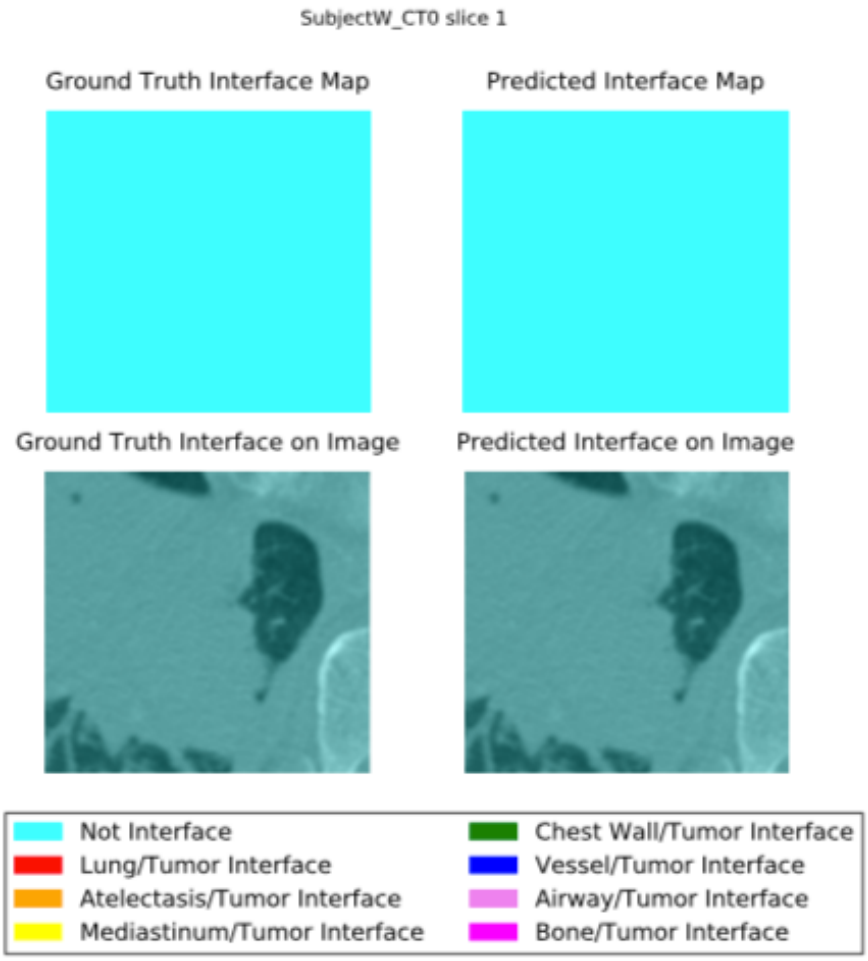


Figure 16: Comparison of the ground truth interface labels map with the predicated labels map from the IF_Only network no post processing. The top row represents the map of true and predicted labels while the bottom row shows the maps overlaid on the corresponding image slice.

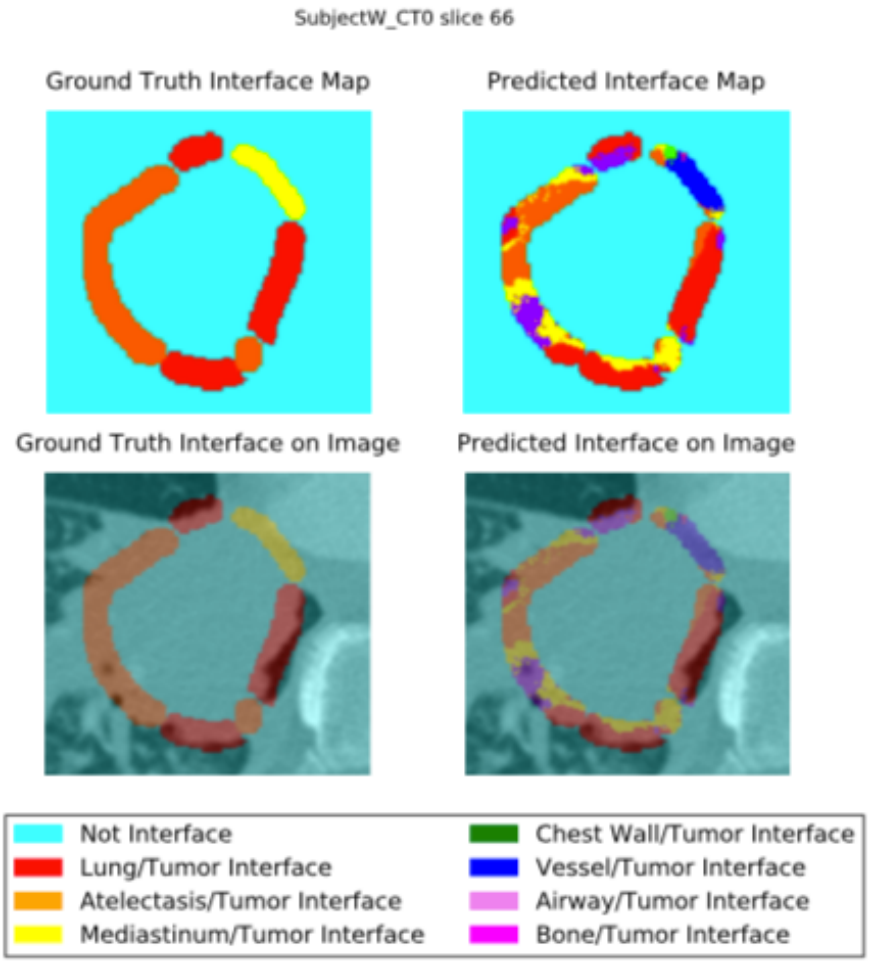


Figure 17: Comparison of the ground truth interface labels map with the predicated labels map from the IF_Only network no post processing. The top row represents the map of true and predicted labels while the bottom row shows the maps overlaid on the corresponding image slice.

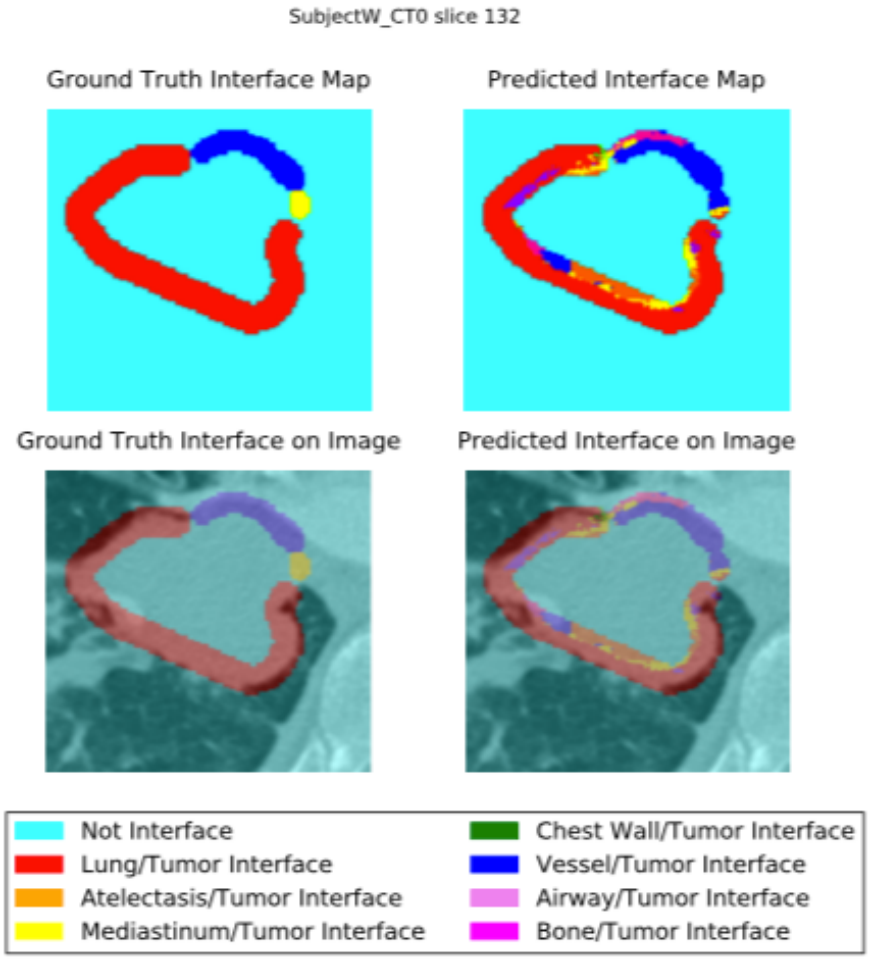


Figure 18: Comparison of the ground truth interface labels map with the predicated labels map from the IF_Only network no post processing. The top row represents the map of true and predicted labels while the bottom row shows the maps overlaid on the corresponding image slice.

e

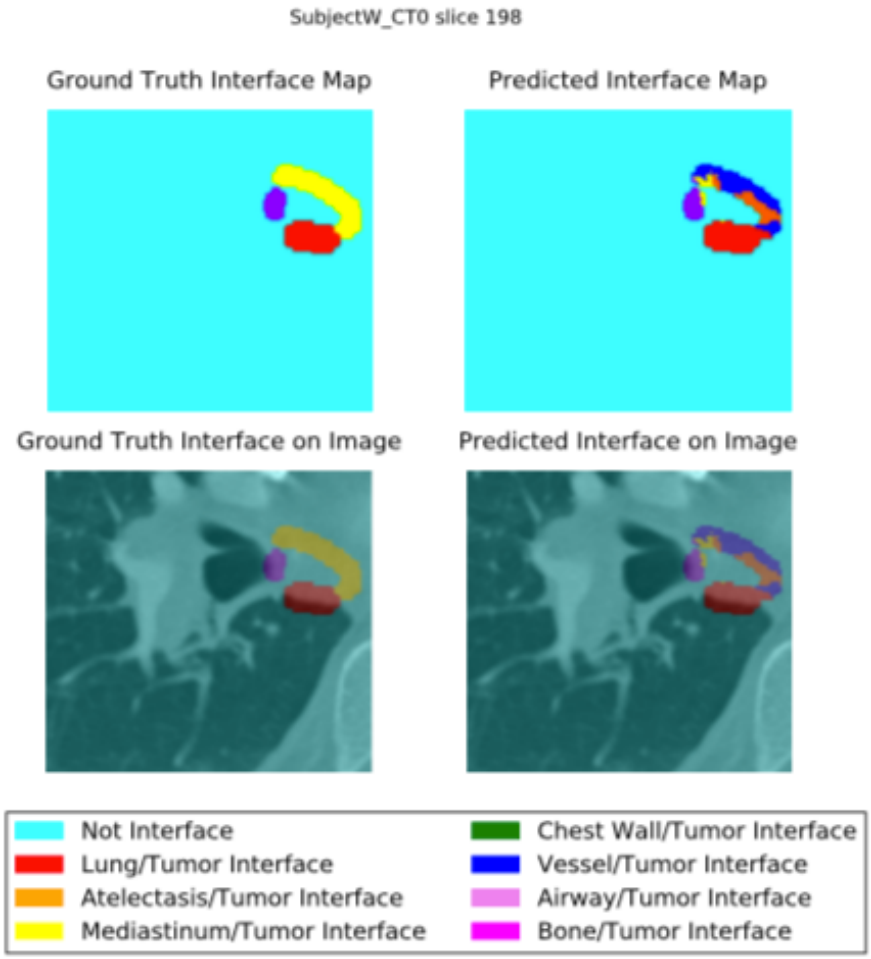


Figure 19: Comparison of the ground truth interface labels map with the predicated labels map from the IF_Only network no post processing. The top row represents the map of true and predicted labels while the bottom row shows the maps overlaid on the corresponding image slice.

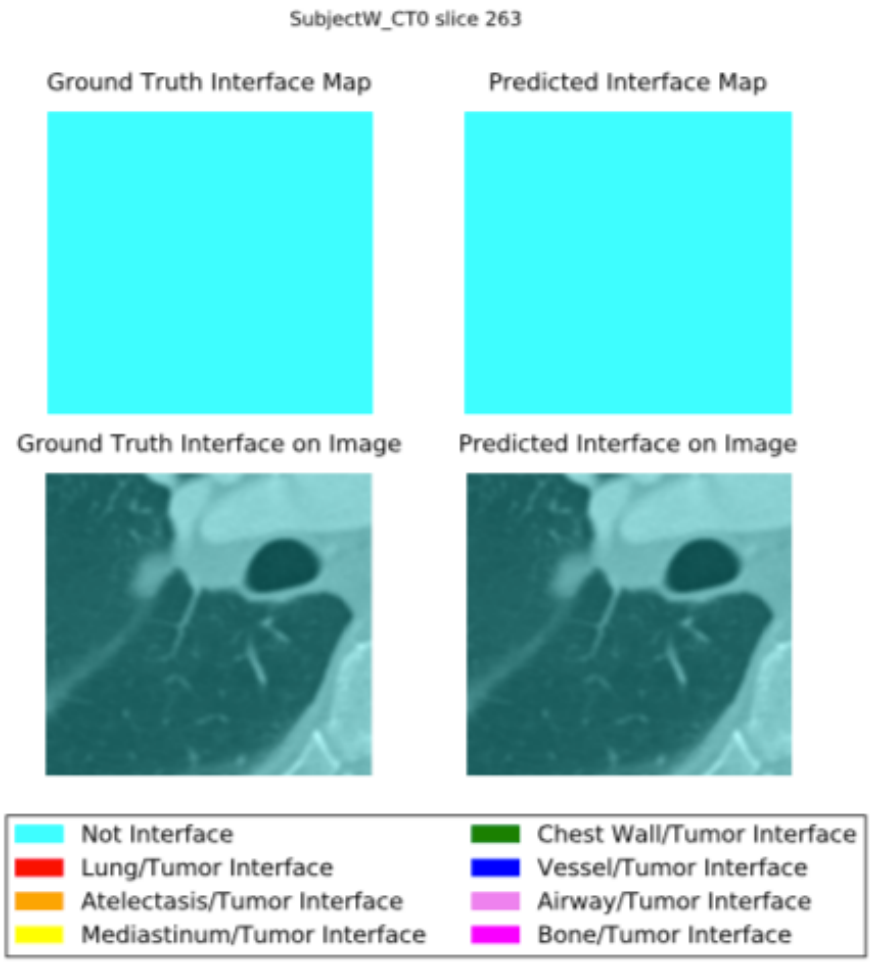


Figure 20: Comparison of the ground truth interface labels map with the predicated labels map from the IF_Only network no post processing. The top row represents the map of true and predicted labels while the bottom row shows the maps overlaid on the corresponding image slice.

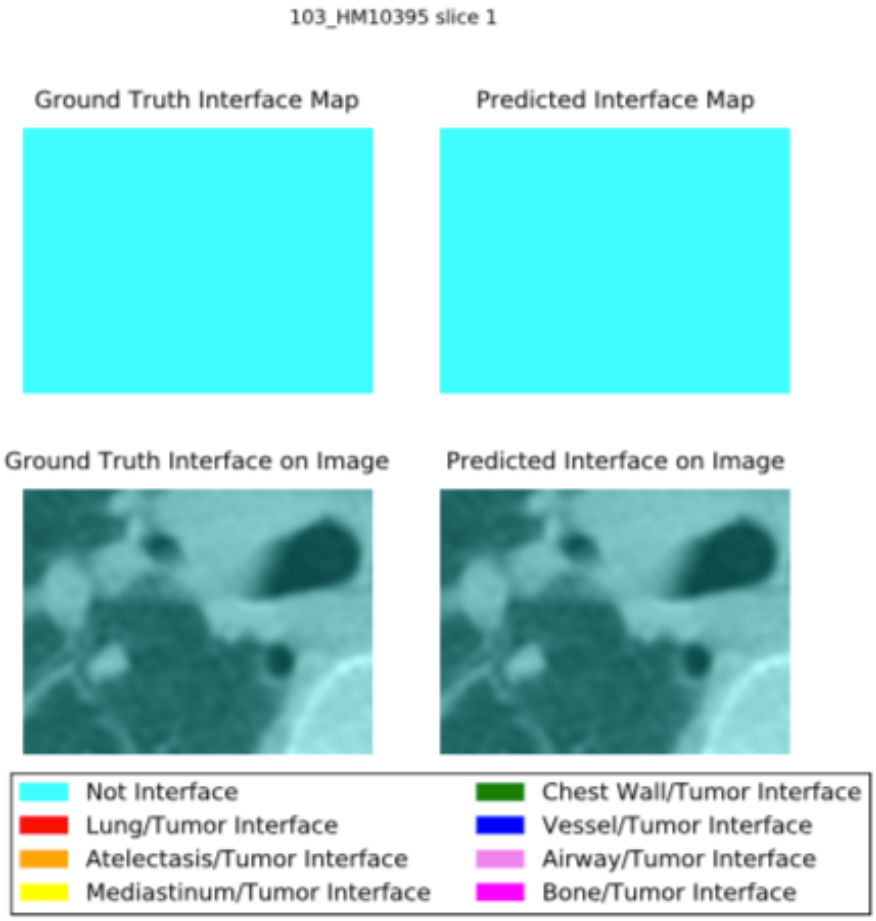


Figure 21: Comparison of the ground truth interface labels map with the predicated labels map from the IF_Only network no post processing. The top row represents the map of true and predicted labels while the bottom row shows the maps overlaid on the corresponding image slice.

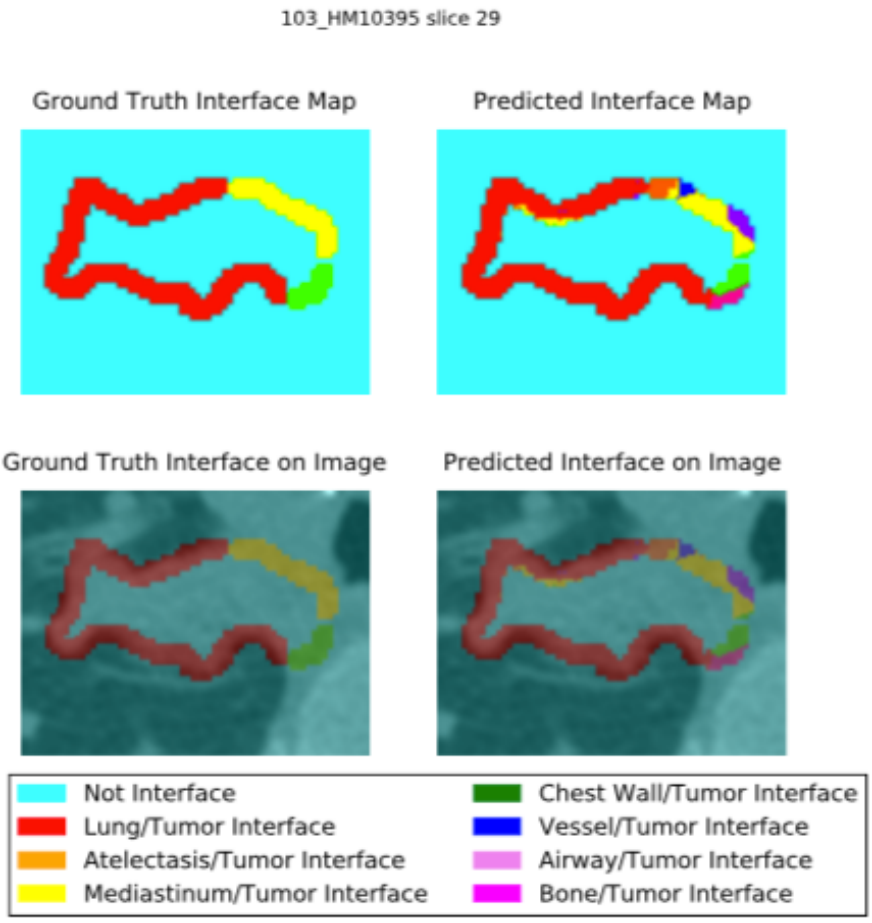


Figure 22: Comparison of the ground truth interface labels map with the predicated labels map from the IF_Only network no post processing. The top row represents the map of true and predicted labels while the bottom row shows the maps overlaid on the corresponding image slice.

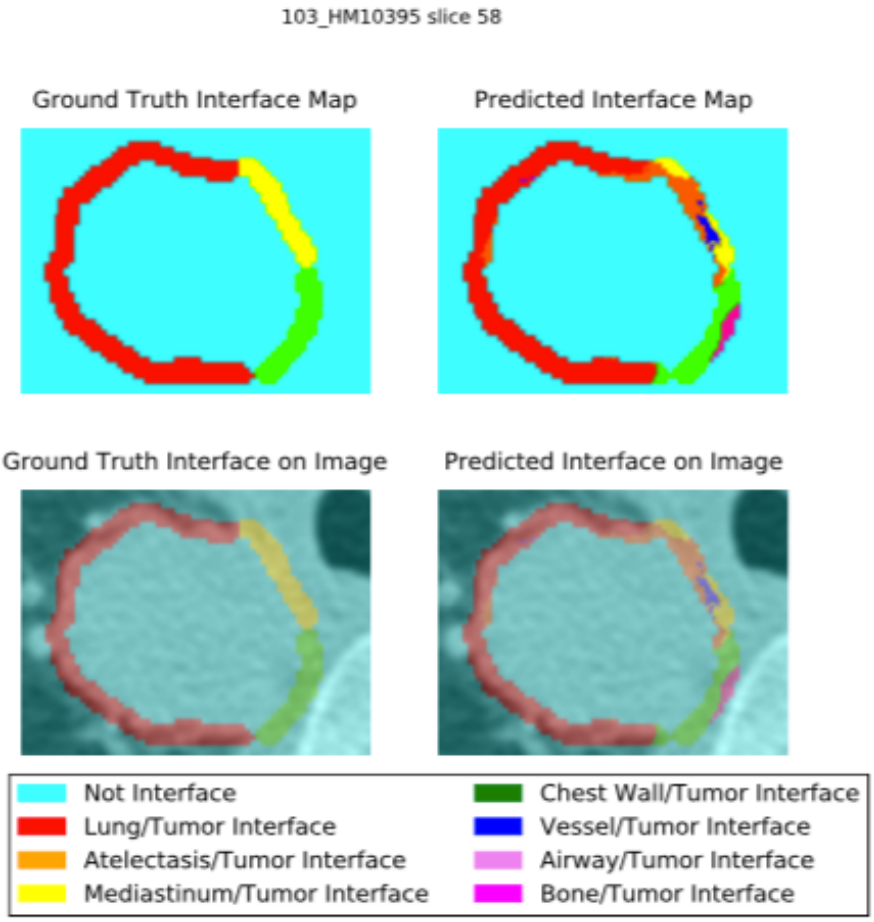


Figure 23: Comparison of the ground truth interface labels map with the predicated labels map from the IF_Only network no post processing. The top row represents the map of true and predicted labels while the bottom row shows the maps overlaid on the corresponding image slice.

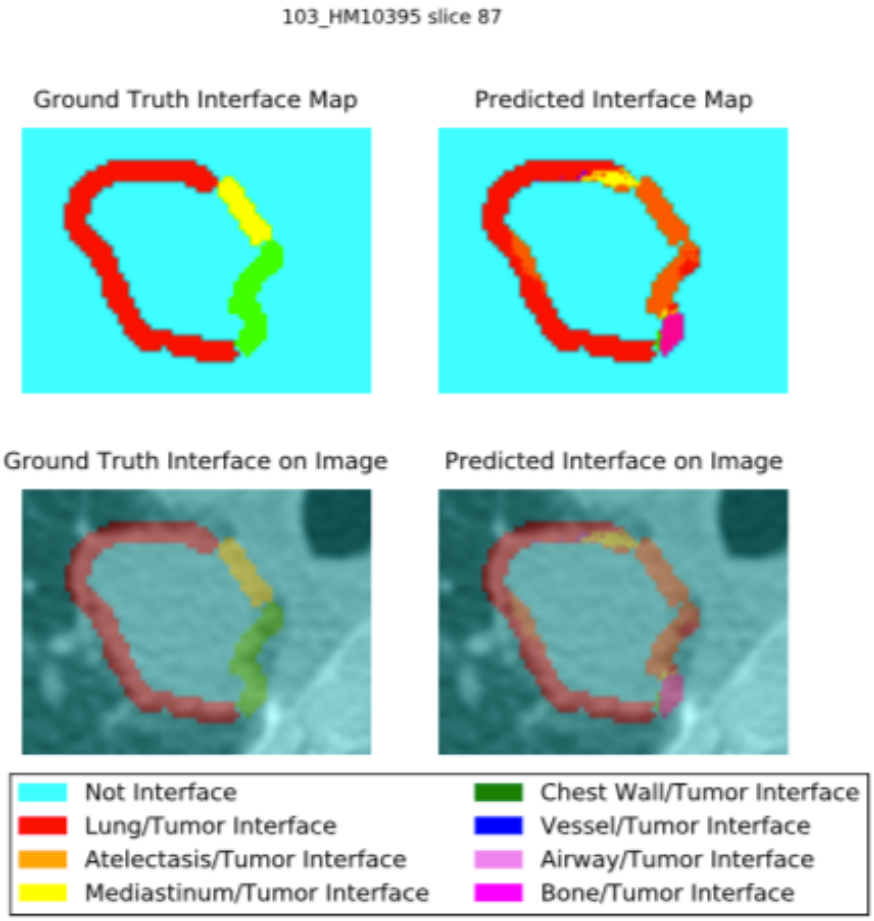


Figure 24: Comparison of the ground truth interface labels map with the predicated labels map from the IF_Only network no post processing. The top row represents the map of true and predicted labels while the bottom row shows the maps overlaid on the corresponding image slice.

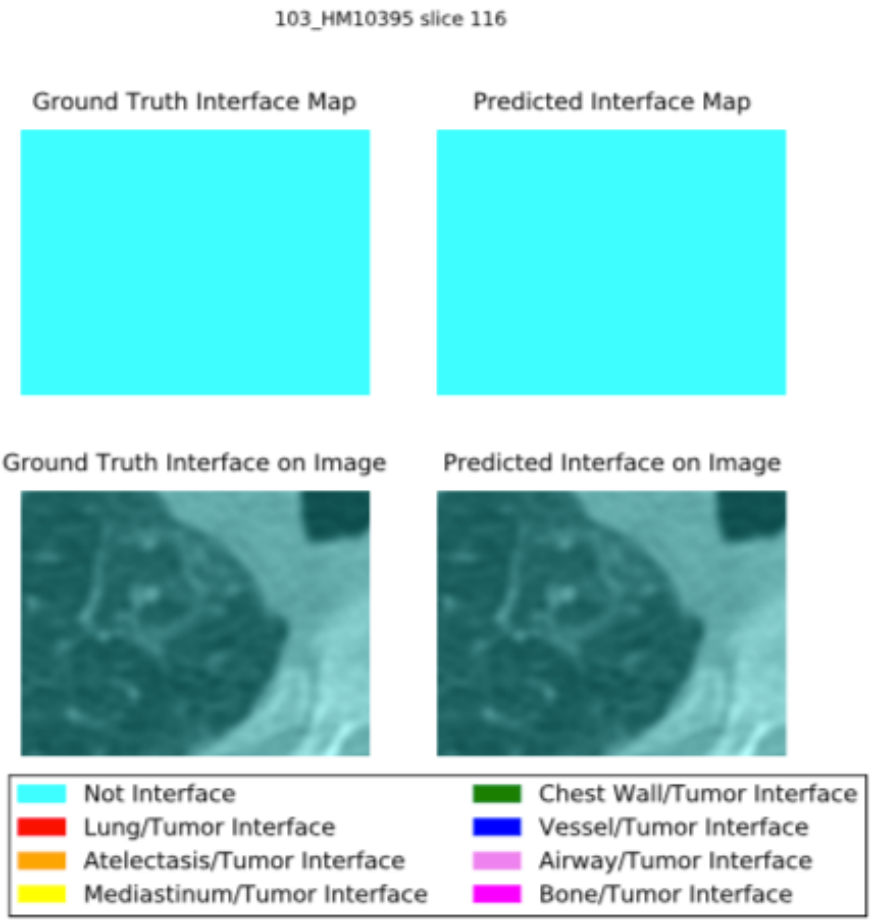


Figure 25: Comparison of the ground truth interface labels map with the predicated labels map from the IF_Only network no post processing. The top row represents the map of true and predicted labels while the bottom row shows the maps overlaid on the corresponding image slice.

Vita

Rebecca Nichole Mahon, “Nicky”, was born October 17, 1987 in Newport, Rhode Island to Lieutenant Commander John Littleton Mahon Jr., US Navy, and Caroline Theresa Wilk Mahon and is a US citizen. Nicky graduated from North Stafford High School in 2005. She received her Bachelors of Science with a distinguished major in physics and a minor in theater arts from the University of Virginia in December of 2009. Following graduation, she attended the Eugene O’Neil Theater Institute for acting. Between August, 2011 and November, 2013, she served as the inaugural Math Instructional Assistant for Germanna Community College department of Tutoring Services, where she established the Math Center, assisted with implementing the Virginia Community College System (VCCS) developmental math redesign at Germanna, and supervised the Supplemental Instruction program. She received two Outstanding Contribution awards and was part of the team winning the VCCS Excellence in Education: Innovative use of Technology award in 2013. She returned to graduate school in August, 2013 and completed her Master of Science degree with a major in medical physics in 2015 at Virginia Commonwealth University (VCU) before continuing on to her Doctorate of Philosophy. During her time at VCU, she served as the Vice President of Social Affairs, Vice President of Academic Affairs, and President of the Medical Physics Scholastic Society, was inducted into the Phi Kappa Phi society,

was the recipient of the Phi Kappa Phi Susan E. Kennedy scholarship for promoting women in higher education and Phi Kappa Phi academic achievement award, placed first in the Mid Atlantic Chapter of the American Association of Physicists in Medicine (AAPM) young investigators symposium, and was appointed to the Student and Trainee Subcommittee of the AAPM. In addition, she was the first author, or co-first author, on six abstracts presented at national conferences, and co-author on two published articles and four abstracts submitted to regional and national conferences. In addition to her academic achievements, Nicky was nominated for best lighting design for “Les Miserables” by the Maryland Theater Guild and has achieved level I certification for the Chesterfield County Community Emergency Response Team (CERT) and is in the process of completing level II certification.