2022

Multi-Modality Automatic Lung Tumor Segmentation Method Using Deep Learning and Radiomics

Siqiu Wang
Virginia Commonwealth University

Multi-Modality Automatic Lung Tumor Segmentation Method Using Deep Learning and Radiomics

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

By

Siqiu Wang, M.S.

M.S., The University of Texas at Austin, Dec 2016

B.S., The University of Texas at Austin, May 2014

Committee Members:

Dr. Lulin Yuan (Director),

Department of Radiation Oncology, Virginia Commonwealth University

Dr. Elisabeth Weiss,

Department of Radiation Oncology, Virginia Commonwealth University

Dr. Laura Padilla,

Department of Radiation Medicine and Applied Sciences,

University of California San Diego

Dr. Milos Manic,

Department of Computer Science, Virginia Commonwealth University

Dedication

To my family, especially my mom, Jianping Qiu, and dad, Feng Wang, whom I have not been able to see in person for the last three years, for enduring the separation of thousands of miles from your only child, just so I could have this opportunity to pursue a different life. I cannot wait to translate this for you and tell you what great parents you are and how immensely grateful I am to be your daughter.

Also, mom, I can finally tell you to stop overanalyzing all the smallest “mistakes” you might have made when raising me. I’d say you did a pretty good job.

Acknowledgement

Foremost, I would like to express my deepest gratitude to the three members of our research group: Dr. Lulin Yuan, Dr. Elisabeth Weiss, and Dr. Rebecca Mahon. As my primary advisor, Dr. Yuan provided me with unwavering support from the very beginning. Not only did his clinical expertise and scientific insights in medical physics form a solid foundation for this project, but also did his dedicated mentorship guide me through some of the most difficult times in a pandemic. I would also like to thank my unofficial advisor Dr. Elisabeth Weiss. As the physician in the group, she asked the important clinical questions and encouraged me to view things in different perspectives. The numerous hours she spent reviewing the contours were what made this project possible. Her awe-inspiring efficiency also motivated me to improve myself and become a better researcher. Last but not least, I would like to thank Dr. Mahon for starting me down the road of machine learning and for always being there when I needed her knowledge and wisdom despite the major transitions in her professional and personal life.

I would also like to thank my committee members: Dr. Laura Padilla for contributing her clinical insights to this project and for always being a strong advocate for students during her time at VCU, and Dr. Milos Manic for his excellent teaching that initiated me into the world of machine learning and for encouraging me to explore different possibilities.

I am thankful to the four radiation oncologists, Dr. Nuzhat Jan, Dr. Ross James Taylor, Dr. Philip Reed McDonagh, and Dr. Bridget Quinn, for taking time out of their busy schedules to provide me with invaluable feedback that made the clinical evaluation portion of this project possible. I am also thankful to the High Performance Research

Computing (HPRC) Core Facility at VCU for the high performance computing resources used in this work.

Finally, I would like to thank the past and present members of our program who offered help, guidance, and support throughout my graduate career: Sarah Holler, Ben Lewis, Areej Aljabal, Samantha Conrad, Matthew Riblett, Katie Goracke, Ron Broman, Judith Larios, Janet Salmon, and the faculty of the Medical Physics program.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AIP: Average Intensity Projection

ANN: Artificial Neural Network

BLD: Bi-directional Local Distance

CBCT: Cone-Beam Computed Tomography

CNN: Convolutional Neural Network

CRF: Conditional Random Field

CRN: Convolutional Residual Network

CT: Computed Tomography

CTV: Clinical Target Volume

DSC: Dice Similarity Coefficient

FCM: Fuzzy-c-means Clustering Method

FCN: Fully-Convolutional Network

FDG: Fluorodeoxyglucose

GLDM: Gray-Level Dependence Matrix

GLCM: Gray-Level Cooccurrence Matrix

GLRLM: Gray-Level Run Length Matrix

GLSZM: Gray-Level Size Zone Matrix

GTV: Gross Tumor Volume

HD: Hausdorff Distance

HPRC: High Performance Research Computing

IAEA: International Atomic Energy Agency

IBSI: Image Biomarker Standardization Initiative

ICRU: International Commission on Radiation Units and Measurements

KNN: K-Nearest Neighbors

LoG: Laplacian of Gaussian

MLC: Multi-Leaf Collimator

NGTDM: Neighboring Gray Tone Difference Matrix

NSCLC: Non-Small Cell Lung Cancer

OAR: Orgain-At-Risk

OSEM: Ordered-Subset Expectation Maximization

PACS: Picture Archiving and Communication System

PET: Positron Emission Tomography

PTV: Planning Target Volume

QA: Quality Assurance

RFN: Recurrent Fusion Network

RMS: Root Mean Squared

ROI: Region of Interest

RPM: Real-time Position Management

RW: Random Walk

SGD: Stochastic Gradient Descent

SUV: Standardized Uptake Value

SVM: Support Vector Machine

TCIA: The Cancer Imaging Archive

TPS: Treatment Planning System

VOI: Volume of Interest

Abstract

MULTI-MODALITY AUTOMATIC LUNG TUMOR SEGMENTATION METHOD USING DEEP LEARNING AND RADIOMICS

By Siqiu Wang

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

Virginia Commonwealth University, 2022.

Director: Lulin Yuan,

Assistant Professor, Department of Radiation Oncology

Delineation of the tumor volume is the initial and fundamental step in the radiotherapy planning process. The current clinical practice of manual delineation is time-consuming and suffers from observer variability. This work seeks to develop an effective automatic framework to produce clinically usable lung tumor segmentations.

First, to facilitate the development and validation of our methodology, an expansive database of planning CTs, diagnostic PETs, and manual tumor segmentations was curated, and an image registration and preprocessing pipeline was established. Then a deep learning neural network was constructed and optimized to utilize dual-modality PET and CT images for lung tumor segmentation. The feasibility of incorporating radiomics and other mechanisms such as a tumor volume-based stratification scheme for training/validation/testing were investigated to improve the segmentation performance. The proposed methodology was evaluated both quantitatively with

similarity metrics and clinically with physician reviews. In addition, external validation with an independent database was also conducted.

Our work addressed some of the major limitations that restricted clinical applicability of the existing approaches and produced automatic segmentations that were consistent with the manually contoured ground truth and were highly clinically-acceptable according to both the quantitative and clinical evaluations. Both novel approaches of implementing a tumor volume-based training/validation/ testing stratification strategy as well as incorporating voxel-wise radiomics feature images were shown to improve the segmentation performance. The results showed that the proposed method was effective and robust, producing automatic lung tumor segmentations that could potentially improve both the quality and consistency of manual tumor delineation.

A. Introduction

Within the last decade, applications of machine learning, deep learning in particular, has gained enormous momentum in cancer research due to a combination of the increasing availability of life sciences data and the soaring demand for more precise and individualized medicine. Cancer genome and pathology research, as outstanding examples, benefited tremendously from automated microscopy and tissue labelling as well as the establishment of public databases such as the Cancer Genome Atlas that currently houses over 2.5 petabytes of high-quality data. In comparison, applications of automated image segmentation for the purpose of radiotherapy planning have not experienced the same degree of acceleration. While several vendors have developed software products for normal organ segmentation as well as adaptations of the existing tumor segmentations based on deformable registration, accurate automatic generation of the initial tumor segmentation has proven to be challenging and left many avenues to be explored. Despite many methods have been proposed, the uniqueness of each cancer case combined with the limited access to well-labeled imaging data, due to patient privacy and HIPAA concerns, is a major hurdle to adequate validation and subsequently clinical deployment.

This work aims to combine novel techniques with thorough validation to construct an automated tumor segmentation framework that produce clinically meaningful results.

A.1 Non-Small Cell Lung Cancer

Lung cancer is the second most common cancer in both men and women, and by far the leading cause of cancer mortality, accounting for 23% of all cancer deaths in the United States. About 80-85% of the lung cancers diagnosed are Non-Small Cell Lung Cancer (NSCLC) which is the main focus of this study [1]. Unfortunately, the overall 5-year survival

rate for all stages of lung cancer is only 22.9%. While only about 16% of lung cancer cases are diagnosed at a localized stage, the five-year survival rate for these patients is at a much higher 61.2% [2]. As the primary definitive treatment for patients with locally advanced lung cancer and inoperable patients with early stage lung cancer [3], external beam radiation therapy is the treatment modality-of-choice for tumors of a broad range of sizes, locations, and complexities. Improving the effectiveness of radiotherapy is an important aspect of enhancing lung cancer care.

A.2 Lung Cancer Radiotherapy Workflow and Its Uncertainties

The fundamental principle of radiation therapy is the delivery of a cell-killing high dose to the tumor and the sparing of the surrounding normal tissues according to their specific dose tolerances. Achieving this goal with great precision requires the minimization of uncertainties throughout the radiotherapy process. A standard-of-care radiation treatment for lung cancer involves two phases: planning and delivery. The general workflow can be summarized as the following steps:

1. Diagnosis and clinical workups
2. Consultation and assessment for radiation treatment
3. Simulation
4. Target volume and organs at risk delineation
5. Dose prescription and treatment planning
6. Treatment setup and delivery

The cancer treatment planning phase (step 1 to 5) begins when radiation therapy is chosen as the treatment modality based on diagnosis, patient consultation, and assessment for suitability. A set of CT images specifically needed for planning purposes

is then acquired during simulation where the patient is immobilized and positioned under the treatment conditions. These planning CTs, along with the PET/CTs or diagnostic CTs taken during the clinical workups, are used for volume delineation where the volumes of interest, i.e., the target for dose delivery and the organs-at-risk for avoidance, are contoured. With these images and delineated volumes as basis, a radiation treatment plan is constructed and optimized in the treatment planning system (TPS) that outlines the desired dose distribution according to the prescribed dose and the instructions for the treatment machine to deliver such distribution.

According to the International Atomic Energy Agency (IAEA) report on accuracy and uncertainties in radiotherapy [4], planning uncertainties mostly arise from the volume delineation step which has been recognized as “the weakest link in the search for accuracy in radiotherapy” [5], as will be discussed in detail in the next section. To a lesser extent, uncertainty in the treatment planning system due to relative dose calibration, dose calculation algorithms, and plan-to-deliverable optimization also affect the accuracy.

A treatment plan is normally delivered in multiple irradiation fractions. At the beginning of each fraction, the patient is positioned and immobilized at the treatment machine using the same setup as in simulation with the guidance of the on-board imaging systems. The fractional dose distribution is then delivered according to the TPS-generated plan. This process is repeated until the entire treatment course is completed.

The International Commission on Radiation Units and Measurements (ICRU) [6] recognizes three main sources of uncertainty in the delivery phase that may impact the precise execution of a plan: patient setup, organ motion and/or deformation, and machine errors. Patient setup variations from fraction to fraction are unavoidable, even though immobilization devices and on-board imaging guidance are employed to improve positioning reproducibility. Depending on the specific tumor site, organ motion can have a

varying degree of influence on the consistency of plan delivery. Bladder or rectum fillings can cause inter-fractional variations, while bowel movement, cardiac, or respiratory motion leads to intra-fractional variations. Motion management measures are often adopted to mitigate these effects. Treatment machine-related uncertainties are often not considered significant, with the advent of modern engineering enabling high mechanical accuracy (e.g., mechanical center vs. radiation center, multi-leaf collimator (MLC) positioning, etc.) as well as dosimetric accuracy (e.g., accumulated dose, dose delivery rate, dose gradient, etc.).

Furthermore, imaging is utilized throughout the entire treatment process: PET/CT acquired during clinical workups and CT during simulation are used for planning; planar x-ray as well as cone-beam CT (CBCT) taken with the on-board imagers are essential for patient setup. The limitations of these systems also introduce uncertainty to the process.

Ideally, efforts should be made to keep all uncertainties as low as possible. In practice, the IAEA report provided estimates of the levels of accuracy that are clinically achievable (Table 1).

Table 1. Estimates of external beam radiation therapy related uncertainties. The contents are adapted from the IAEA report to focus on lung cancer treatment only.

Quantity	Dosimetric Uncertainty*	Geometric Uncertainty
Planning		
Dose calibration	1.6–2.6%	
Imaging related to treatment planning		
CT		
Image geometry & resolution		< 2 mm
PET		
Image geometry & resolution		4–7 mm
Volume delineation		
Target definition (site dependent)		5–50 mm
Normal tissue definition		5–20 mm
Treatment planning system	Several % *	2–4 mm
Delivery		

Treatment machine related uncertainties		
Mechanical		< 2 mm
Dosimetric	< 2%	
Patient positioning		
Initial setup		< 1–15 mm
Re-setup at each fraction		2–5 mm
Imaging related to delivery		
EPIDs		
Image geometry & resolution		1–2 mm
MV CT (helical)		
Image geometry & resolution		1–2 mm
kV CBCT		
Image geometry & resolution		1 mm
MV CBCT		
Image geometry & resolution		2 mm
Overall		
EBRT end to end in phantom **	3–10%	2 mm
EBRT end to end in patient ** #	5–10%	5 mm

** TPS uncertainties can reach up to 20% in regions of high density heterogeneity, especially with older systems. However, this is not typically a concern in lung cancer treatment. ** Does not include volume delineation. # Expert consensus.*

A.3 Lung Tumor Delineation and Observer Variability

A.3.1 Lung Tumor Delineation Workflow

Without an invasive and thorough pathological investigation, it is difficult to determine the “true” extent of the disease. Instead, manual delineation by radiation oncologists based on medical images is considered the “ground truth” in current clinical practice. Manual delineation is one of the most laborious and time-consuming tasks for physicians in the treatment planning process [7].

Compared with the other geometric uncertainties, volume delineation uncertainty is evidently much more significant (Table 1). Target delineation, specifically, is one of the most error-prone steps in the radiotherapy chain. Figure 1 briefly summarizes the

workflow and the types of medical images commonly employed for lung tumor delineation.

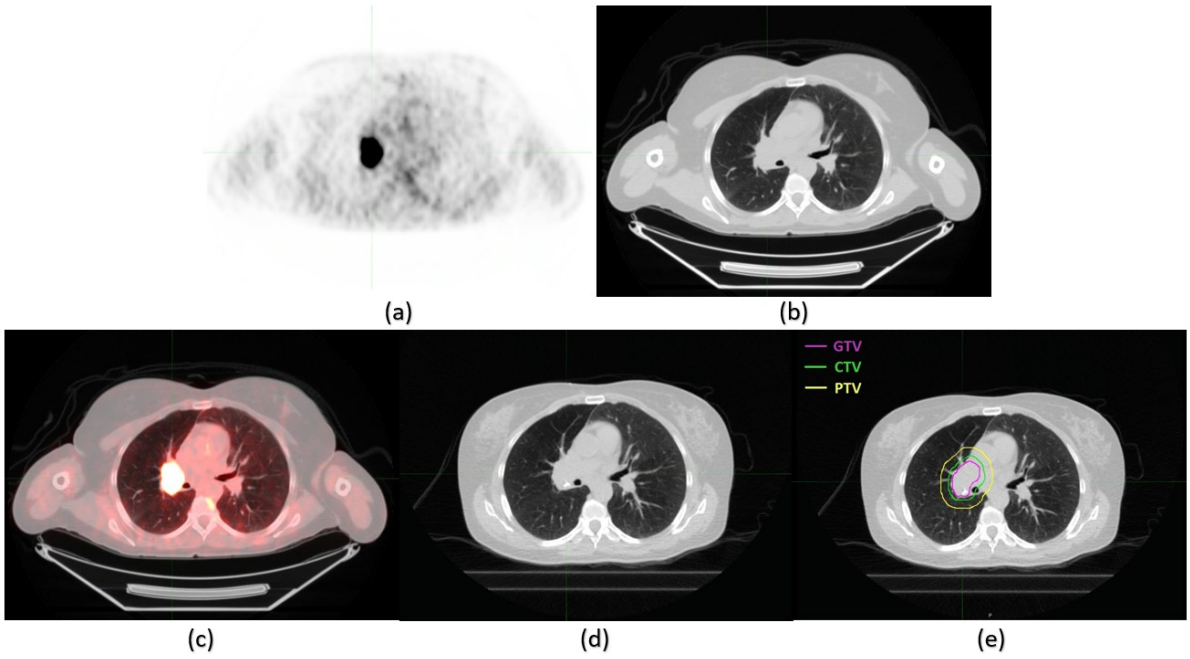


Figure 1. An example of the image modalities involved in lung radiotherapy. A diagnostic set of CT (a) and PET (b) are first acquired on a PET/CT scanner; Information from the fused PET/CT image (c) and the planning CT (d), taken during the simulation, are then combined (through either visual inspection or PET/CT and CT fusion) and assessed by the radiation oncologist who delineates the target volumes (GTV, CTV, and PTV) on the planning CT (e).

Since the defined volumes are the basis on which the treatment plan is constructed, any inaccuracy will lead to a systematic error downstream, potentially resulting in a geographical miss of the target and/or overexposure of the organs-at-risk. To provide a uniform framework for delineation, ICRU introduced the concepts of gross tumor volume (GTV), clinical target volume (CTV), and planning target volume (PTV) [8] (Figure 1.e). In summary, GTV envelopes the gross palpable or visible extent of the malignant growth which includes the primary tumor, any metastatic lymph nodes, and other metastasis. CTV expands upon the GTV to contain the microscopic disease that is not visible to the naked eyes, whereas PTV includes the CTV and a margin to account for organ motion and setup

variation. While the contouring and margin selection for CTV and PTV is a topic that warrants extensive investigation, our project will focus mainly on the delineation of GTV.

There are two interplaying sources of variations in target volume delineation: 1) uncertainty due to the inherent limitations of the imaging modalities, especially in capturing the metastatic disease and 2) the observer variability in the determination of the exact GTV boundary from the images [9].

A.3.2 Imaging-Related Uncertainty

The delineation process is subject to both the strengths and pitfalls of the imaging modalities. For lung tumor delineation, radiation oncologists mainly rely on two types of images: CT (Figure 1.a) and PET (Figure 1.b). The image acquisition mechanisms of these two modalities are fundamentally different: A CT image is the volumetric reconstruction of the transmission of x-rays from the generator through the patient to the detectors, whereas PET depends on the detection and localization of the annihilation photons produced by the positron-emitting radiotracer injected into the patient's body. As a result, CT depicts the anatomical structures according to their electron density, while the signal in PET reflects the radiotracer uptake in the local tissue. In the case of FDG-PET, the radiotracer ^{18}F-FDG congregates in the regions of high glucose consumption, i.e. high metabolic activity. Since malignant lesions tend to be metabolically active, ^{18}FDG-PET in lung cancer imaging is especially useful in identifying metastatic lymph nodes or the primary tumor boundary especially at ambiguous tumor/normal tissue interfaces [10].

The anatomical and functional information from CT and PET, respectively, are complimentary to each other. During diagnosis/staging, a PET image is usually acquired alongside a CT image on a PET/CT scanner with the patient in the same position for anatomical reference and attenuation correction. The images are registered upon

acquisition, forming a set of PET/CT image. Incorporating PET/CT along with the simulation CT in target volume delineation was shown to alter the shape and size of the GTVs in a majority of patients undergoing radiotherapy with curative intent [11]. A multitude of studies have demonstrated the advantage of combining PET and CT over the conventional CT-only planning in improving GTV coverage and changing the size of GTVs, thus enabling target dose escalation while complying with the constraints for organs-at-risk [12]. A systematic study found that the percentage of cases where the integration of PET/CT in radiotherapy planning led to significant changes in the delineated target volumes ranged from 21% to 100% among all reviewed studies [13]. Furthermore, PET is invaluable in delimiting a rough boundary between the tumor and its surrounding tissues when they appear indistinguishable on CT due to their similar electron densities [14]. In the area of atelectasis, especially, incorporating PET/CT was shown to successfully reduce the lung and esophageal doses while maintaining target coverage [15].

However, while the adaptation of multi-modality treatment planning helps accurately define the target volumes, it also introduces additional uncertainties due to a few clinical and technical issues: how to reliably define the tumor boundary in PET and how to combine PET/CT and planning CT for composite delineation.

The tumor/normal tissue boundary on PET is often difficult to pinpoint because of the low spatial resolution and the variations in signal intensity thresholds used for determining active tumor disease. Modern PET/CT scanners in the clinics can achieve spatial resolution of less than 2 mm for CT and up to 7 mm for PET (Table 1). In addition to the inherent low spatial resolution due to its acquisition mechanism, PET also suffers from motion artifacts since the images are taken under free-breathing conditions and take several minutes. Furthermore, since the radiotracer uptake in tissue is governed by a variety of biological and physical processes that vary from patient to patient. there does

not exist a physically meaningful PET signal threshold between malignant lesions and normal tissue [16]. The simplest way to utilize PET information is visual interpretation. However, the apparent tumor volume is significantly affected by the window level and therefore highly subjective. Alternatively, the standardized uptake value (SUV) was coined in an attempt to standardize the signal intensity and provide a semi-quantitative basis for evaluation. SUV is a measure of radiotracer uptake activity in a region of interest (ROI) normalized to a distribution volume. An arbitrary SUV value (2.0 to 2.5) or a percentage of SUV_{max} (40% to 50%) is often used to distinguish between normal and abnormal tissues [12], [17]. However, the clinical validity of SUV thresholds has been repeatedly challenged. Biehl et al. argued that the optimal threshold for delineation depends on the tumor size and can range from $15 \pm 6\%$ for large tumors to $42 \pm 2\%$ for tumors smaller than 3 cm [18]. Biological factors such as heterogeneous uptake, inflammation, blood glucose level, motion artifacts, etc. and technical factors such as scanner variability, reconstruction parameter, residual activity in the system, etc. all may affect SUV calculations by 5% to 50% [19]. PET/CT for lung cancer faces additional challenge, as respiration-induced misalignments between PET and CT within a single PET/CT scan can lead to a significant underestimation of the SUV as the CT is the basis of attenuation correction in PET [20]. In summary, tumor edge definition in PET is challenging and leaves a wide gap for subjective user interpretation. Minimizing this subjectivity is one of the driving forces behind automated computer vision methods where the full range of voxel intensity can be utilized quantitatively instead of a single threshold.

Once the PET/CT and planning CT are acquired, how to combine the information from these two distinct modalities for volume delineation introduces new variables due to the lack of strict clinical guidelines [21]. In the most direct approach, as in our institution, the diagnostic PET/CT is displayed in fused mode (Figure 1.c) next to the planning CT (Figure

1.d). The two images are then visually correlated by the radiation oncologist who delineates the GTV on the planning CT while identifying areas of active malignant lesions on PET/CT for inclusion into the GTV (Figure 1.e). The association of information in this method relies on the radiation oncologist's knowledge and experience and is therefore prone to observer variability. Instead of visual referencing, image fusion is sometimes applied [22]. However, differences in image acquisition conditions between the diagnostic PET/CT and the simulation CT, along with organ movement and respiratory motion, can make registration difficult, potentially leading to interpretation issues and increased observer variability [23]. Deformable registration has been a candidate in addressing differences in acquisition conditions in other applications such as treatment response assessment. However, it was to be of marginal value for treatment planning [24]. There is not yet a unified strategy to level PET and CT images for manual delineation of lung tumor, which contributes to the uncertainty in volume delineation.

A.3.3 Observer-Related Uncertainty

Some of these limitations can be addressed through following delineation protocols and utilizing technological advancements of the imaging modalities, e.g. more sophisticated detection system to achieve higher PET resolution, dedicated PET/CT simulator to reduce registration mismatch, etc. However, even with a consistent imaging and delineation protocol as well as a matching set of PET/CT and planning CT, observer variability persists as the main source of geometric uncertainty compared with setup error and organ motion [25]. As a visual example, 6 physicians from our institution delineated the GTV (Figure 2) for the same patient using the same set of PET/CT and planning CT. Although the metrics used in literature to quantify inter- and intra-observer variability in lung tumor delineation are not uniform, the general findings indicate a significant level of

observer-related uncertainty. The V_{max}/V_{min} ratio among GTVs delineated by different radiation oncologists for the same patient can range from 1.8 to 2.3 for the primary GTV and from 5.2 to > 7 for the nodal GTV [26]. Fox et al. reported a median interobserver percentage of concordance of 70% and a median intraobserver percentage of concordance of 71% using registered PET/CT and planning CT (an improvement from 61% and 58%, respectively, with unregistered images) [27]. Louie et al. found that 4D-CT may help reduce both inter- and intra-observer variabilities. However, disagreement among observers tends to increase with case complexity with the reported primary tumor volume overlap index ranging from 0.556 to 0.915 [28]. The location of the tumor and the type of interface between the tumor and the surrounding tissue were shown to be important factors associated with the degree of observer variability [29]. Karki et al. found the variability at tumor-atelectasis interface to be significantly larger than that at any other interfaces. In general, a thorough review of interventions to reduce inter-observer variability [30] attributed the wide-spread variability to the difference in knowledge (both oncological and anatomical) and its application, interpretation of anatomy on images, as well as understanding of target volume definitions. In addition, each physician's risk/benefit assessment, especially in regions of anatomical ambiguity, also impact the decision making process [9]. Methodological differences such as drawing precision and hand-eye coordination may also play a role in executing the intended target volume [26]. Addressing the observer variability stemming from the differences in subjective interpretation is essential to improving the quality and consistency of lung cancer treatment.

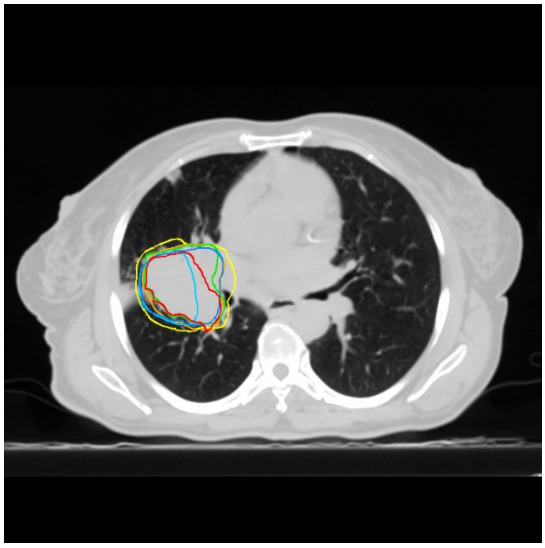


Figure 2. An example of inter-observer variability in manual segmentation. The figure shows 6 different GTVs delineated by 6 different radiation oncologists from our institution.

A.3.4 Metrics for Tumor Delineation Evaluation

A wide variety of metrics was used for volume delineation comparison among literature, which sometimes makes it difficult to directly compare studies. There are different quality aspects in 3D medical image segmentation according to which types of segmentation errors can be defined. Metrics are expected to discern some or all of these errors, depending on the data and the segmentation task. Comprehensive overviews of volumetric comparison methods have previously been published [31]. Generally speaking, the metrics are grouped into volumetric comparisons (volume measurements, volume ratios), measures of overlap (concordance, discordance, Dice similarity Coefficient), surface or dimension measures (dimensions, mean surface distance, Hausdorff distance) and center of volume or mass [32]. For the purpose of delineation/segmentation performance assessment, measures of overlap appear to be the most common.

The Dice coefficient (DSC), also called the overlap index, is the most used metric in validating medical volume segmentations. In addition to the direct comparison between

automatic and ground truth segmentations, it is common to use the DICE to measure reproducibility (repeatability). DSC is defined by

$$DSC = \frac{2|S_g \cap S|}{|S_g| + |S|}$$

where S_g is the "ground-truth" segmentation and S is the segmentation to be assessed. DSC is also proficient at dealing with situations where there is a strong imbalance between the number of foreground and background voxels, which is the case in tumor volume segmentation where the volumes of normal tissue far out weight that of the targets.

For surface-specific assessment, Karki et al. adopted the bidirectional local distance (BLD) as a measure for inter-observer variability [33]. BLD was shown to be more effective than conventional distance measures (e.g., minimum distance) in the presence of complex tumor contours by taking into account the bidirectional characteristics of minimum surface-to-point distance with both forward and backward search directions. BLD can be applied near round concave or folding regions for surface-specific analysis of segmentation performance.

The ultimate goal of a segmentation framework is clinical utility; therefore, a clinical acceptance task is essential. Gooding et al. presented a methodology inspired by the Turing Test [34] provided a framework for assessing automatic segmentation performance. This approach argues that if trained human observers are unable to distinguish the automatic segmentation from those manually-delineated by the clinicians, then the machine-generated segmentations are of sufficient quality for clinical use. This framework can be adapted into a clinical acceptability test where physicians will be asked to rate or select a collection of manual and automatic tumor contours.

A.4 Previous Automatic Lung Tumor Segmentation Methods

Previous publications have proposed a variety of PET and CT lung tumor co-segmentation methods using 1) traditional graphical models and variational methods, 2) supervised and unsupervised machine learning, or a mixture of 1) and 2). There are two main considerations: how to effectively incorporate the data from both PET and CT, and how to perform segmentation accurately and efficiently.

Traditional graph theory-based methods are intuitive and computationally efficient. More importantly, they are able to achieve competitive performance without consuming a large amount of data. Bagci et al. represented the PET and CT images in a product lattice, essentially making a hyper graph, and performed simultaneous delineation using the random walk (RW) algorithm [35]. Ju et al. formulated the RW segmentation on PET and graph cut segmentation on CT as a single energy minimization problem [36]. Under a Bayesian framework, Irace and Batatia [37] proposed to combine the PET and CT data using a bivariate Poisson mixture model. Markov random field optimization has also been previously employed for PET/CT co-segmentation [38]. More recently, Cui et al. [39] adopted a modified RW framework where a topology graph was generated from the PET images that applied spatial-topological constraints in addition to the local intensity changes from CT. The main disadvantage of the traditional graph theory is that in most of these approaches, at least one user-defined seed or ROI is needed to establish the baseline for segmentation, which means the process cannot be fully automated. Although a method to automatically select object/background seed was discussed in [35], the approach is prone to non-tumor uptake and requires a pre-defined signal threshold. Furthermore, graph-based theory tends to fare poorly at tumor and normal tissue interfaces when signal intensity changes are subtle due to the lack of complex spatial context. Decisions are

made only according to the information provided by intensity values of image voxels, while other imaging features such as texture are ignored.

Machine learning has seen increasing popularity in a wide range of medical image processing tasks including automatic tumor segmentation. Earlier works such as Kawata et al. [40] and Ikushima et al. [41] investigated the efficacy of automated frameworks based on traditional neural networks: fuzzy-c-means clustering method (FCM), artificial neural network (ANN), and support vector machine (SVM). These two studies followed the clinical protocol of radiotherapy planning by using the planning CT as well as the diagnostic PET/CT with the manual delineations by radiation oncologists as the ground truth. However, the databases were limited (< 20 patients), and the authors offered few details on the network designs, so it was difficult to determine the validity of these methods. It was also reported that the segmentation performance was highly dependent on tumor texture (solid, ground glass opacity, and part-solid ground glass opacity).

The majority of earlier reports on deep learning-based methods for lung tumor segmentation focused on a single imaging modality, i.e., CT,[42]–[46] omitting the complemental information provided by functional imaging modalities such as PET. In the last few years, dual-modality methods had begun to gain traction, and several methods were developed to automatically segment lung tumors using PET and CT inputs, with different approaches on how to effectively incorporate the data from both modalities. A summary of the databases and the DSCs of these works, including the two traditional ML methods, is provided in Table 2. Most of these works employed some variations of U-Net [47] and its 3D adaptation V-Net [48] due to the network architecture's ability to efficiently segment images using relatively small datasets, and the segmentation performance was shown to improve over previous non-machine learning-based methods [49], [50]. One main approach is to process the PET and CT images in parallel. Both Zhong et al. [51]

and Zhao et al. [49] utilized two independent V-Nets to extract features from PET and CT. While Zhong et al. performed a graph-cut co-segmentation using the two sub-network outputs as region costs, Zhao et al. fused the outputs of the two V-Nets through element-wise sum and fed the combined image into another CNN, functioning as a feature fusion module, to produce the final tumor mask. In another paper, Zhong et al. [52] proposed a network with two independent convolutional arms as well as two deconvolutional arms for simultaneous co-segmentation in the PET and CT images. Kumar et al. [53] constructed a variation of V-Net where the independently extracted PET and CT features were fused at each resolution level through elementwise multiplication with a co-learned fusion weight map. Arguing that the fusion of image features can be further improved through progressive phases, Bi et al. [54] proposed the addition of a recurrent fusion network (RFN) to the convolutional network backbones. Alternatively, some other works chose to utilize the two modalities sequentially. Fu et al. [55] used the U-Net to extract a spatial attention map from the PET image, which is then incorporated into another U-Net where the tumor segmentation is produced on the corresponding CT image. Li et al. [56] combined deep learning with a variational method by creating a tumor probability map from the CT image with a fully-convolutional network (FCN) and using a fuzzy variational model to segment the tumor from the combination of the probability map and the PET image.

On a completely different note with unsupervised learning, Lian et al. [50] devised a co-clustering algorithm where the data from PET and CT are modeled and fused based on the theory of belief function, a tool for representing and reasoning with uncertainty and probability.

While DSC values from 0.64 to 0.87 were reported in the previous deep learning-based works, the results of these studies were not directly comparable with either our method or among each other due to missing information on the quality of the databases,

the lack of sufficient validation, as well as the discrepancy in how the ground truth was defined in each study. All previous networks were investigated with limited cohorts of <100 cases [40], [41], [49]–[56] and used a variety of methods for to define the ground truth (SUV threshold-based [53]–[55], manual in PET and/or CT [49]–[52], [54], or the contours used in the clinical plans [40], [41]). In most reports, tumor specifics (size, stage, location, etc.) were not identified. In addition, there was a general lack of a clear validation scheme. The terms “validation” and “testing” were sometimes used interchangeably, and several papers used their testing datasets during the development of their networks [52], [56], [57]. Furthermore, no clinical validation or validation with an external database were conducted to our knowledge. As a result, it is difficult to compare the performance of different methods using the reported evaluation metrics. For example, Zhong et al. [51] reported DSCs of 0.87 ± 0.05 (vs. manual ground truth on CT) and 0.76 ± 0.09 (vs. manual ground truth on PET), the adaptation of this method by Kumar et al. [53] on their database only achieved a DSC of 0.63 (using 40% peak SUV on PET for initial contour with manual adjustments on CT).

Importantly, all previous works in deep learning-based lung tumor segmentation[49]–[53], [55] used the pre-registered diagnostic PET/CT images without incorporating the simulation CT that is the basic modality for radiation treatment planning, ignoring any uncertainties introduced by PET and simulation CT registration for a valid clinical scenario.

Table 2. Summary of the previous works in machine learning-based dual-modality automatic lung tumor segmentation.

Author (Year)	Modalities	Image Type	Database	Validation/Testing	Ground Truth	DSC
Kawata [40] (2017)	Planning CT*, PET/CT	3D	16 SBRT **	N/A	Manual delineation by radiation oncologists in the planning CT with reference to PET	0.79 ± 0.06
Ikushima [41] (2016)	Planning CT, PET/CT	3D	14 SBRT	Leave-one-out-by-patient	Manual delineation by radiation oncologists in the planning CT with reference to PET	0.78
Zhong [51] (2018)	PET/CT	3D	32	Training: 20 Testing: 12	Separate manual delineations by radiation oncologists in CT and PET	0.87 ± 0.05 (CT) 0.76 ± 0.09 (PET)
Zhong [52] (2019)	PET/CT	3D	60	Training: 38 Validation/Testing: 22	Separate manual delineations by radiation oncologists in CT and PET	0.87 ± 0.03 (CT) 0.85 ± 0.06 (PET)
Zhao [49] (2018)	PET/CT	3D	84	Training: 48 Testing: 36	Manual delineations by radiation oncologists in CT	0.85 ± 0.08
Kumar [53] (2019)	PET/CT	2D	50	5-fold Validation (Training/Testing)	40% SUV threshold in PET with manual adjustments in CT	0.64
Lian [50] (2019)	PET/CT	3D	21	N/A	Manual delineations by radiation oncologists in PET	0.86 ± 0.04 (CT) 0.87 ± 0.04 (PET)
Fu [55] (2020)	PET/CT	2D	50	5-fold Validation (Training/Testing)	Semi-automatic connected thresholding with manual adjustments	0.71
Li [56] (2020)	PET/CT	3D	84	Training: 48 Testing: 36	Manual delineations by radiation oncologists in CT	0.86 ± 0.05
Bi [54] (2021)	PET/CT	2D	50, 70 ***	5-fold Validation (Training/Testing)	40% SUV in PET with adaptive thresholding in CT (50); Manual delineations by radiation oncologists in CT (70)	0.68 ± 0.23

** Planning CT & PET/CT: Planning CT registered to PET/CT was used instead of the CT from PET/CT as in the other previous works.*

*** Stereotactic Body Radiation Therapy (SBRT) is a treatment technique that delivers higher dose in each fraction with fewer fractions. It is usually prescribed for relatively small and local tumors.*

**** The study used two databases separately.*

Similar multi-modality segmentation networks have been proposed for tumor sites other than lung or modalities other than PET and CT. An alternative network structure was tested in [58] where two V-Nets were cascaded to form a W-shaped network. The CT images were taken as input for the first U-Net, from which the outputs and the PET images were fed into the second U-Net. The argument for two sequential networks is that the extracted knowledge from the first network would help regularize the second learning process. However, in reverse, this would also mean that the performance of the second network is limited by the first because the segmentation error would propagate. Guo et al. [59], Havaei et al. [60], and Hou et al. [61] further investigated the effectiveness of different parallel and sequential, respectively, network architectures for brain tissue segmentation.

In general, adaptations of machine learning in medical image segmentation have to deal with challenges such as the class imbalance between normal and abnormal tissues, the high computational demand due to the handling of 3D images, the scarcity of labelled data for model training, and most importantly, the lack of validation. Many previous works do not have a large enough database size (Table 2) to train a robust model and draw statistically-significant conclusions. The parameters of the database (disease extent and complexity, tumor location and complexity, etc.) and the training/validation/testing protocols were also often not specified. A universal performance benchmark is difficult to implement due to the limitations on medical data sharing. Many of these works chose to demonstrate improved evaluation metrics, e.g., Dice similarity coefficient (DSC), over selected previous similar methods. However, the validity of these comparisons is often called into question when details on the database are not specified. In addition to performance validation on a good and sizable database representative of the lung cancer patient population, a clinical acceptability test is also crucial to testing the clinical usability of any proposed method.

A.5 Radiomics

A.5.1 General Background

Radiomics refers to the extraction of quantitative data from medical images using data characterization algorithms and the mining of these data for research or clinical decision guidance. The image features extracted often quantifies characteristics of the intensity patterns called biomarkers that are not readily apparent to the naked eye but are hypothesized to reflect the underlying physiology and shown to improve diagnosis, prognosis, treatment selection, and prediction of treatment response in many cancer-related applications [62], [63]. For lung cancer radiotherapy specifically, radiomics has been utilized to assess or predict overall survival, local or metastatic recurrence, radiation treatment response, lung toxicity, as well as staging [64].

A typical workflow of a radiomics study consists of the following steps: image acquisition, image segmentation, feature extraction, feature selection, and analysis/modeling [65]. In general, most radiomics studies are retrospective and use the standard-of-care medical images, which are CT, PET, and sometimes MRI in lung cancer treatment. While no sophisticated techniques are required, the specific image acquisition protocols and parameters affect the stability of the radiomics features, the extent of which has been widely discussed [66]. Following image acquisition, a pre-processing step is often added to homogenize the images. In the next crucial step, a region of interest (ROI) or volume of interest (VOI) is defined through image segmentation based on the study objectives. This can be achieved manually, semi-automatically, or automatically [67]. While most studies in lung cancer radiomics investigated the tumors, the metastatic lesions, or the entire area [68], there are some that analyzed the immediate peritumoral region [69]–[71]. This is relevant to our feature selection step and will be discussed in more detail in Section E.2. From the defined ROI or VOI, calculation of features can be

performed using a wide variety of algorithms and software. In theory, any mathematical rules or formulas can be applied to an image to extract a corresponding feature. In practice, guidelines for extracting standardized radiomics features are provided by the Image Biomarker Standardization Initiative (IBSI) [72], and the implementation is usually determined by the features available via the chosen extraction software. The types of radiomics features mostly commonly used will be discussed in the next section. Because of the availability of so many options, once the desired features were extracted, either as single values or feature maps, a feature selection or dimension reduction step is performed to narrow down the pool of candidates and rule out irrelevant features. This is a multi-step process that is highly specific to the study objectives. Further detail about the selection process for our work is presented in Section E.3. In the final step, a model is built according to the specific endpoint of the research using the most relevant features. In reality, feature selection and model development are often intertwined in an iterative process. It is important to note that a prominent issue preventing the clinical application of many radiomics-based methods is overfitting and the lack of generalizability [73], and therefore appropriate internal and external validations are essential.

A.5.2 Types of Radiomics Features

When radiomics was first introduced in 2012 [63], radiomics features were mostly semantic, a single characterizing number of a region of interest (ROI) such as volume, shape, or heterogeneity. These features have been widely studied as biomarkers in aiding clinical decision and outcome prediction. In recent years, the advent of computational resources has enabled the extraction of voxel-wise radiomics features, providing us with higher-order texture maps of the original image. Radiomics features in use nowadays can

be roughly divided into: 1) histogram-based, 2) texture-based, 3) model-based, 4) transform-based, and 5) shape-based [74].

1) Histogram-based features describe the first-order statistical distribution of the pixel or voxel intensity within the ROI. The most common members of this group include mean, min/max, variance, percentiles, etc. SUV_{max} in PET is another example of a histogram-based feature. Some of the more sophisticated features can semantically describe the shape or uniformity of the distribution, e.g., skewedness, entropy, and energy.

2) Texture-based features quantifies the spatial distribution of the neighboring pixels or voxels on a higher order and incorporates spatial information through the usage of matrices. Since the spatial context contained in these features are potentially useful in our work, a brief description of some of the most commonly investigated texture features is provided here, i.e., the Gray-Level Cooccurrence Matrix (GLCM), Gray-Level Run Length Matrix (GLRLM), Gray-Level Size Zone Matrix (GLSZM), Neighboring Gray Tone Difference Matrix (NGTDM), and Gray-Level Dependence Matrix (GLDM).

The GLCM quantifies the spatial relationships between pairs of pixels or voxels set distances apart in predefined directions [75]. To help demonstrate the general principle of how texture-based features are calculated, an illustrated example of GLCM calculation is shown in Figure 3. For an image matrix I, the GLCM is defined as the matrix $P(i,j|\delta,\theta)$ where the (i,j) element is the number of times two pixels in the image I with gray levels i and j appear to be δ pixels apart in the θ direction. Once the matrix P is generated, its normalized version p can then be used to calculate the GLCM features such as:

$$GLCM\ contrast\ = \sum_{i=1}\sum_{j=1}(i-j)^2 p(i,j)$$

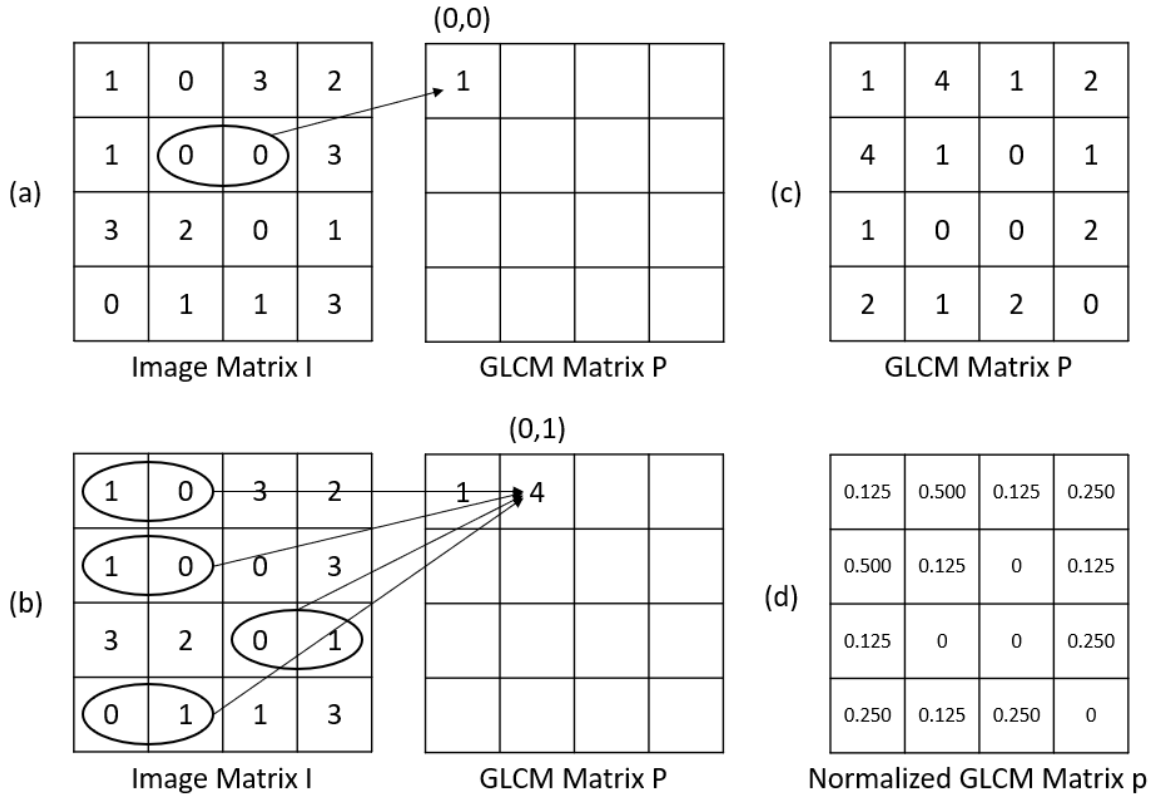


Figure 3. An example of GLCM calculation. Consider a 4x4 image I with discrete pixel values, i.e., gray levels from 0 to 3 (a), for distance $\delta = 1$ and direction $\theta = 0^o$ (for symmetrical GLCM, this mean the pixels can be to either the left or right of the center pixel), the (0,0) and (0,1) elements of GLCM P can be calculated as shown in (a) and (b), respectively. The fully calculated GLCM matrix is shown in (c) and the normalized version p in (d).

GLRLM quantifies gray level runs, i.e., the number of consecutive pixels or voxels with the same gray level [76]. Its corresponding features describe the properties of these runs, e.g., gray-level and run-length uniformity, long- or short-run emphasis, etc. Similarly, GLSZM counts the number of groups of neighboring pixels or voxels with the same gray level [77]. NGTDM is the sum of differences between the gray level of a center pixel and the mean gray level of its neighbors [78]. Some of the NGTDM features are colloquially named, e.g., coarseness and complexity, since the quantities they represent are similar to these image characteristics. In a similar fashion, GLDM counts the number of

neighboring pixels or voxels that are “dependent” on the center pixel or voxel, where dependency is defined as a gray level difference within a predefined threshold [79].

3) Model-based feature extraction apply a specific mathematical or geometrical model to the ROI to assess a specific quality of the ROI. For example, fractal analysis examines the structural details under increasing magnification [80].

4) Transform-based features are obtained through the application of transformation functions such as Gaussian, Gabor, wavelet, etc., most commonly as filters in the preprocessing step. The Gaussian function, for instance, smooth the image and blur the edge, while the Laplacian is an edge detector. Transform-based filters have been shown to be effective as image preprocessing techniques prior to tumor segmentation [50] [81].

5) Shape-based features are geometric properties such as diameter, sphericity, etc. They are not voxel-wise features but rather a single description of the ROI or VOI. Since this type of features is produced from the tumor segmentation that we are trying to predict, it is not applicable to our work.

A 5.3 Radiomics for Tumor Segmentation

To our knowledge, there has not been any previous work that used radiomics feature images as inputs to a deep learning network for the purpose of tumor segmentation. However, several methods have been proposed with radiomics and traditional machine learning or other non-machine learning segmentation algorithms. Woods et al. used 4D co-occurrence texture features and a four-layer artificial neural network to segment malignant breast lesions on dynamic contrast-enhanced MRI [82]. For lung tumor segmentation, Markel et al. tested a variety of features in combination with a decision tree and the K-nearest neighbors (KNN) classifiers on PET/CT images [83]. More recently, a

radiomics-based segmentation method using a region growing algorithm was proposed by Bundschuh et al [84]. While the goal of this study was to improve tumor visualization on PET, not radiotherapy planning, they found that segmentation produced from entropy-based features most closely matched the phantom studies. Furthermore, a study by Torrents-Barrena et al. proposed a segmentation method for fetal and maternal anatomy in MRI and ultrasound images through two parallel pipelines: one with radiomics and support vector machine (SVM) and the other with deep learning and found that the radiomics pipeline outperformed deep learning for some organs, while the opposite was true for the others [85]. In addition, several studies emerged in the last few years that demonstrated how integrating radiomics and deep learning could potentially improve the task of classification for lung [86], head and neck [87], and prostate lesions [88].

With these advances in mind, we hypothesized that the incorporation of radiomics feature images and deep learning could potentially improve the segmentation performance for lung tumors.

A.6 Overview of Dissertation

A.6.1 Problem Statement and Purpose

The purpose of this study is to develop an effective method to automatically segment lung tumors based on multi-modality medical images using machine learning and radiomics.

Accurate differentiation between cancerous and normal tissues through images is essential in the planning stage of external beam radiation therapy, so that the target and the organs at risk are precisely defined to achieve maximum tumor cell killing and minimum normal tissue damage. The most commonly employed imaging modalities in

radiotherapy for lung cancers are computed tomography (CT) and positron emission tomography (PET). In the current clinical practice, the radiation oncologists rely on the anatomical information from the former and the functional information from the latter, e.g., cell metabolism in the case of fluorodeoxyglucose (FDG)-PET, to create manual contours of the tumors. This procedure known as volume delineation is time-consuming, and the resulting contours are prone to inter- and intra-observer variability, one of the largest contributors to uncertainty in the radiotherapy process. This variation in tumor contours can have implications for not only the treatment outcome but also the comparison of radiotherapy plans and protocols among physicians and institutions. An effective automatic tumor segmentation method can potentially improve the clinical efficiency as well as address the observer variability stemming from differences in subjective image interpretation and improve the quality and consistency of lung cancer treatment.

Several strategies have been proposed for semi-automated or fully-automated tumor delineation, or segmentation as the more commonly used terminology in computerized image processing, using either single-modality or multi-modality images as input. Variational methods, graph cut, machine learning, or a combination of these methods has been proposed for a variety of anatomical sites. General limitations exist for non-machine learning methods such as the need for user-input seeds for graph cut and the arbitrary parameter selection in variational methods. While machine learning has shown great promise in solving computer vision tasks including image segmentation, the specific issue of lung tumor segmentation has not been sufficiently addressed. Not only does the segmentation accuracy need to be improved from previous works and adequately assessed using quantitative as well as clinical acceptability tests, but also should the results be validated against a large and diverse dataset that reflects the broad range of tumor size and case complexity encountered in the clinic. Previous machine leaning-

based frameworks showed promise but had not been able to produce sufficiently validated results for clinical utilization.

In light of these shortcomings, we aim to develop an automatic segmentation framework based on machine learning to utilize information from both PET and CT, the modalities that are the current clinical standards for lung tumor radiotherapy planning, optimizing the method with an expansive database representative of the lung cancer patient population, and validating it against ground truth defined through the standard treatment planning protocols. In addition to exploring dual-modality deep neural network structures, this project seeks to improve the segmentation performance by applying radiomics features as additional inputs. Other strategies such as tumor volume-based stratification of the datasets will also be investigated to address issues encountered during development and further improve performance. In addition to quantitative evaluation using multiple similarity metrics and external validation against a public dataset, an evaluation of the clinical acceptability of the segmentation results will be conducted with a group of qualified radiation oncologists.

A.6.2 Specific Aims

Based on preliminary work, this goal will be achieved through 3 specific aims:

Specific Aim 1: Develop a deep learning neural network that utilizes dual-modality images to automatically segment lung tumors for radiotherapy planning.

The goal of this specific aim is to establish a general framework from data curation and preprocessing to network training and testing to postprocessing and evaluation. A deep learning neural network will be constructed based on CNNs for 2 image inputs, specifically PET and CT. Once our network is capable of producing meaningful segmentation results, its performance will be evaluated through benchmark comparison

with a previously published deep learning network. Further measures will be implemented to enhance the network performance, such as the transition from 2D to 3D inputs, the optimization of the network parameters, as well as the adaptation of alternative architectures if time and resources allow.

Specific Aim 2: Evaluate the potential of incorporating radiomics features as a third input to the network to improve segmentation performance.

The network structure will first be modified to accommodate a third input of image feature map in addition to the original two inputs: PET and CT. We plan to identify and select useful features with the potential to improve the overall segmentation performance from a pool of radiomics features. The reliability of these features will be evaluated to assess the generalizability of our method.

Specific Aim 3: Improve segmentation performance by adding mechanisms to account for challenging tumor features and conduct quantitative evaluations and clinical acceptability tests of the segmentation results.

Based on preliminary results and previous studies, mechanisms will be introduced to address the specific issues that impact the segmentation performance. We will investigate the impact of tumor volume-based stratification of the dataset and adjustment in training/validation strategy, conduct a clinical acceptability test where the clinicians choose between a manual and a machine-generated contour in a blinded fashion, and assess our models on an external database. In addit6ion, case studies will be conducted to identify problematic tumor/tissue interfaces for future work in interface-specific improvement techniques.

A.6.3 Innovation

This thesis proposes a multi-modality deep learning network structure to simultaneously learn from planning CT, PET/CT dual-modality images, as well as the higher-level image features (e.g., radiomics), for automatic lung tumor segmentation. Based on preliminary results, a stratified training, validation, and testing strategy based on the tumor volume will be adopted to improve the segmentation performance. To our knowledge, no other work in automatic lung tumor segmentation has employed a stratified training/validation/testing strategy or used voxel-wise radiomics image features as additional inputs to improve segmentation performance.

Our access to a large and diverse patient database is essential to fully training, validating, and testing the proposed network model. Unlike many of the previous works, we will adhere closely to the current clinical protocols of lung cancer radiotherapy planning by using the planning CT and the diagnostic PET/CT and ground truth segmentations manually delineated by an experienced radiation oncologist. In addition to assessing the quality of the segmentation results using both volume and surface metrics, to address the clinical usability issue often seen in previous works, we plan to conduct a clinical acceptability test where clinicians will be asked to choose between the manual delineations and machine-generated segmentations. Furthermore, external validation will be conducted with a public database. Both validation measures have not been done by previous lung tumor segmentation studies.

B. Patient Database

B.1 Patient and Image Specifications

The patient database consists of the simulation CT and diagnostic PET images of 290 non-small cell lung cancer (NSCLC) cases, with a variety of tumor sizes (0.5 – 1036.4 mL with median = 11.5 mL), stages (I: 152 cases, II: 22 cases, III: 93 cases, IV: 23 cases), and locations, e.g., near mediastinum (110 cases), chest wall (164 cases), diaphragm (28 cases), etc. Patients were treated with either SBRT (166 cases) or conventionally fractionated (124 cases) radiation therapy. All patients were treated in the VCU Health Radiation Oncology clinic from 2008 to 2019. This study was approved by the institutional review board.

All simulation CTs were acquired on a dedicated Big Bore CT scanner (Brilliance, Philips Medical Systems, Best, The Netherlands) in a 512 x 512 matrix with an axial resolution of 0.98 mm to 1.37 mm and a typical slice thickness of 3 mm (with 5 exceptions ranging from 2 to 5 mm) as 4D-CTs in helical mode, with the respiratory cycle traced by either the Varian Real-Time Position Management (RPM) (Varian Medical Systems, Palo Alto, CA) or the Philips pneumatic bellows (Philips Medical Systems, Cleveland, OH). Breath-hold with the Active Breathing Coordinator device (Elekta, Stockholm, Sweden) was utilized for some patients who have large tumor motion and could comply with the procedure. The majority of the PET images were obtained in a 128 x 128 matrix (with seven 168 x 168) with an axial resolution of 3.91 mm to 5.47 mm and a slice thickness of 2 mm to 4.25 mm. While these images were acquired on a variety of PET/CT scanners: 52.5% GE Discovery 690, 29.3% GE Discovery LS, 12.5% GE Discovery ST (GE Medical Systems, Cleveland, OH), and 5.7% others, all the PET scanners followed the same institutional quality assurance (QA) guidelines. Typically, the images were acquired with 2.5 min per bed position and 8 bed positions per scan and reconstructed with ordered-

subset expectation maximization algorithm (OSEM) using 24 subsets. The noise characteristic in the PET images from different scans did not appear to be different. The PET images were typically obtained during staging within 6 weeks prior to the acquisition of the simulation CTs in 83% of the cases. In some instances, the PET images were found to be acquired up to 4 months before the simulation CT. Therefore, to ensure high quality input data, we compared the anatomy on the simulation CT and the CT acquired along with the PET on the PET/CT scanner and excluded the cases where significant anatomical changes were found. In total, 16 cases were excluded. The main causes for exclusion were development of atelectasis (6 cases) and drastic changes in tumor volume (4 cases).

For the majority of the cases, the planning CTs along with the contours used for the clinical plans were acquired from Eclipse (Varian Medical Systems, Palo Alto, CA), while some of the older cases were found on the Pinnacle archive (Philips Medical Systems, Cleveland, OH). It is worth noting that the planning protocol in our clinic switched from using the 30% breathing phase CT to the average intensity projection (AIP) in 2018. This study did not investigate them separately due to limitations in data availability as well as time and resources. The diagnostic PET/CTs were retrieved from the Philips IntelliSpace picture archiving and communication system (PACS). All the images were exported to MIM (MIM Software Inc., Cleveland, OH) for further processing.

All GTVs were initially delineated manually by the treating radiation oncologist at the time in the simulation CT with the visual guidance of the corresponding PET/CT and reviewed by an experienced radiation oncologist. Modifications were made mainly to exclude metastatic lymph nodes and other metastatic diseases since our current main focus is on the primary tumor only. Minor adjustments were also made in cases where an ITV was contoured instead of a GTV.

B.2 Preprocessing

The simulation CT and the diagnostic PET were rigidly registered in MIM (MIM Software Inc., Cleveland, OH) with the main focus on aligning the tumor volumes. In cases where nearby lung pathologies obscured the tumor boundaries which made direct alignment difficult, other anatomical landmarks in proximity to the tumor were included to improve registration accuracy. The registered PET was then resampled to match the resolution of the simulation CT.

As mentioned in the Lung Tumor Delineation and Observer Variability section, deformable registration did not show significant benefit in improving the quality of manual delineation. While its effect on machine learning-based automatic segmentation was unclear, we were reluctant to use deformable registration due to its potential to alter usable image features.

The DICOM images were preprocessed into 2D axial slices and 3D volumetric data and stored in the TIFF and NRRD formats, respectively. The GTV delineations were store in the DICOM RTst files as coordinates and was converted to masks with the voxels within the tumor region labeled as 1. Voxel value normalization was performed on the CT and PET images using the maximum and minimum voxel intensities of the CT and PET images, respectively, across all patients.

The 290 image pair database was shuffled and divided into 162 training cases (162 pair of 3D volumes and 2403 pair of 2D slices), 59 validation cases (691 slices), and 59 testing cases (811 slices) by an approximately 3:1:1 ratio. Moderate data augmentation was performed through rigid and affine transformation, i.e., horizontal flip, rotation in the range of -20 to 20 degrees, and scaling by a factor of 0.8 to 1 to boost the training database to 2-3 times its original size.

All codes were written in Python (Python Software Foundation. Version 3.6. Available at http://www.python.org), using packages including SimpleITK [89], OpenCV [90], and Pydicom [91]. For GTV-to-mask conversion, in-house algorithm was initially used, and later on the package RT_Utils (https://github.com/qurit/rt-utils) was adopted.

C. Specific Aim 1

Develop a deep learning neural network that utilizes dual-modality images to automatically segment lung tumors for radiotherapy planning.

C.1 Preliminary 2D Dual-Modality Network and Benchmarking

To our knowledge, there were no dual- or multi-modality deep learning software tools for medical images readily available in the public domain [92], [93]. In this section of our work, a novel dual-modality network architecture was constructed, the validity of which was established through performance benchmark against a state-of-the-arts dual-modality segmentation method published closed to the time of our network development.

C.1.1 Constructing the Dual-Modality Segmentation Network

The fundamental requirement for our network architecture was its ability to utilize information from two image modalities, PET and CT at this stage and provide the tumor segmentation as a single output. To achieve that goal, we needed a base structure to build on. Several previous works in multi-modality lung tumor segmentation (as mentioned in Section A.3) chose U-Net to be the basis of their networks, justifiably, as its simple but elegant architecture enables effective semantic image segmentation with relatively small datasets and offers flexibility for adaptations and modifications.

Based on previous works and theory of how convolutional neural networks and the U-Nets specifically function [93], our segmentation network was constructed based on concatenated subnetworks of 2D convolutional neural networks (CNNs) to simultaneously learn from both the planning CT and the PET images (Figure 3). Note that the architecture presented here was the one used for the preliminary work and benchmarking in this section, a similar but updated 3D version with more details of the network architecture will be presented in Section C.2.

The idea was to have two independent convolution arms for CT and PET so that the image features most relevant to each modality can be extracted separately as the images were downsampled and the resolution decreased. After each convolution block, the extracted features from the CT and the PET branches are concatenated by the feature channel and fed into a single deconvolution path at the corresponding resolution level through skip connections. The purpose of concatenation is to provide the deconvolution path with all the features available from both modalities at each resolution level and let the trained weights determine which features are important. The deconvolution path ends with an activation layer that produces a tumor/background probability map. The map was then thresholded at 0.5 (i.e., the pixel has a higher than 50% chance of being the tumor) to obtain the tumor mask. In training, to mitigate the class imbalance issue as the number of background pixels far outweigh the tumor pixels, we used the Dice similarity coefficient (DSC) as the training metric and negative DSC as the loss function.

The training process was monitored using the training and validation losses (both calculated as DSC) and, to prevent overfitting, was stopped when the validation loss appeared to plateau relative to the training loss. Each training session started with 40 epochs and continued in 10-epoch increments until the stopping conditions were met.

All codes were written in Python, utilizing the Keras library (Keras. Version 2.2) with the Tensorflow backend (Google Research. Version 1.14). The models were trained on a computer equipped with an NVIDIA GTX 980M GPU and an 8-GB memory. GPU acceleration was supported by CUDA (NVIDIA. Version 10.1.) and CuDNN (NVIDIA. Version 7.6.).

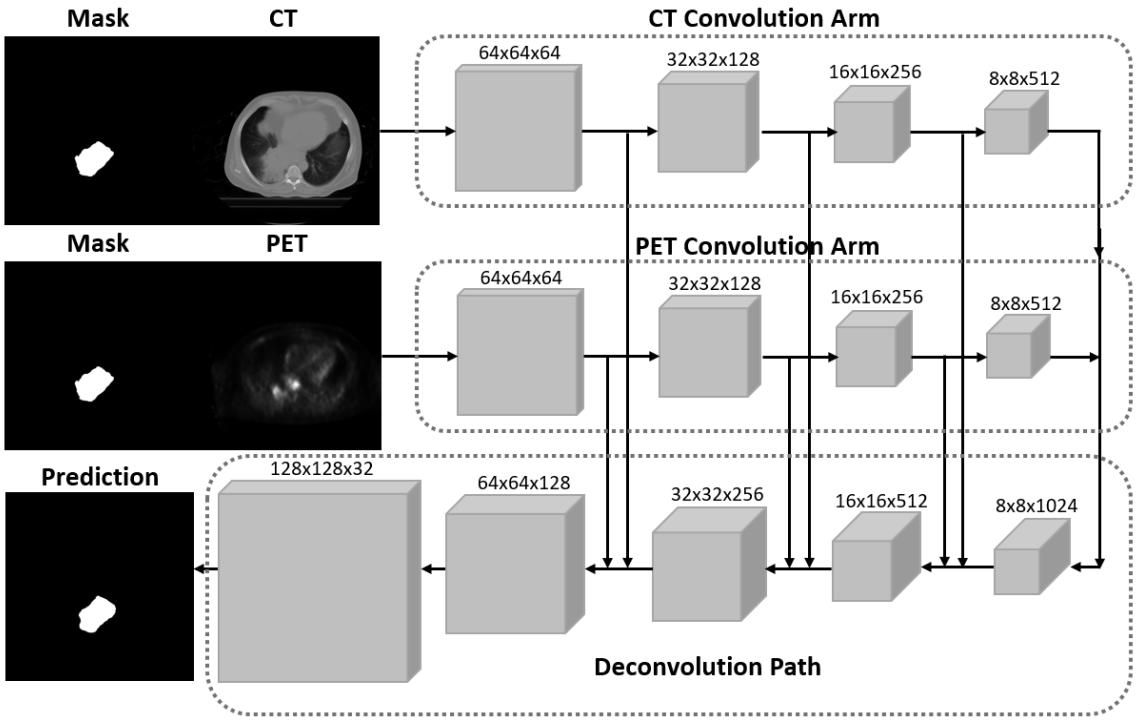


Figure 4. An illustration of our proposed dual-modality segmentation network architecture. The arrows indicate the direction of forward data flow. The numbers on top of the images indicate the resolution (first two) and the feature channel size (the third). Concatenation by the third dimension, i.e., *the feature channel, is performed every time the arrows from the 2 convolution arms flow into the deconvolution path.*

C.1.2 Performance Benchmark against a Selected Previous Work

A simple way to assess the validity of our proposed structure is through performance comparison with a state-of-the-art automatic segmentation method. For that purpose, another U-Net based architecture was adapted from a recently published article by Zhao et al. [49] and trained and evaluated with our dataset. Zhao et al had access to a database

of 84 lung cancer patients and compared their results with a variety of traditional graphical and variational methods as well as other deep learning-based methods, making it a suitable candidate as a performance benchmark.

The adapted architecture (Figure 4) employs two independent U-Nets (without the final activation layers) to extract the full feature maps from CT and PET and fuse them together via elementwise sum. The fused feature map was taken as the input to the third U-Net, which essentially acts as a feature fusion module that performs the segmentation task on the combined information of CT and PET and produce a tumor mask.

Figure 5. An illustration of the previously proposed network architecture by Zhao et al. [49]. ⊕ *represents the elementwise sum operation.*

Since we continued to build our dataset throughout this work, at the time when benchmarking was performed, a smaller database of 213 cases was used. The specifications of the database were similar to that of our eventual database with GTV sizes from 0.6 to 1025.3 mL (median = 16.5 mL) and stages from I to IV. The dataset was shuffled and divided into 129 training cases, 42 validation cases, and 42 testing cases by a roughly 3:1:1 ratio. For clarification, all performance evaluation in the preliminary work

was conducted using only the validation cases, as the testing cases were being reserved for the evaluation of the final model.

The DSC was calculated for each patient. Our model achieved a mean DSC of 0.772 ± 0.096, higher than the previous state-of-the-art method (DSC = 0.756 ± 0.075). While it was not conclusive that our model performs better (two-sample t-test, p = 0.22), the segmentation results are at least comparable. A DSC histogram comparison (Figure 5) indicates a decrease in patients with low DSC and an increase in DSC above 0.85.

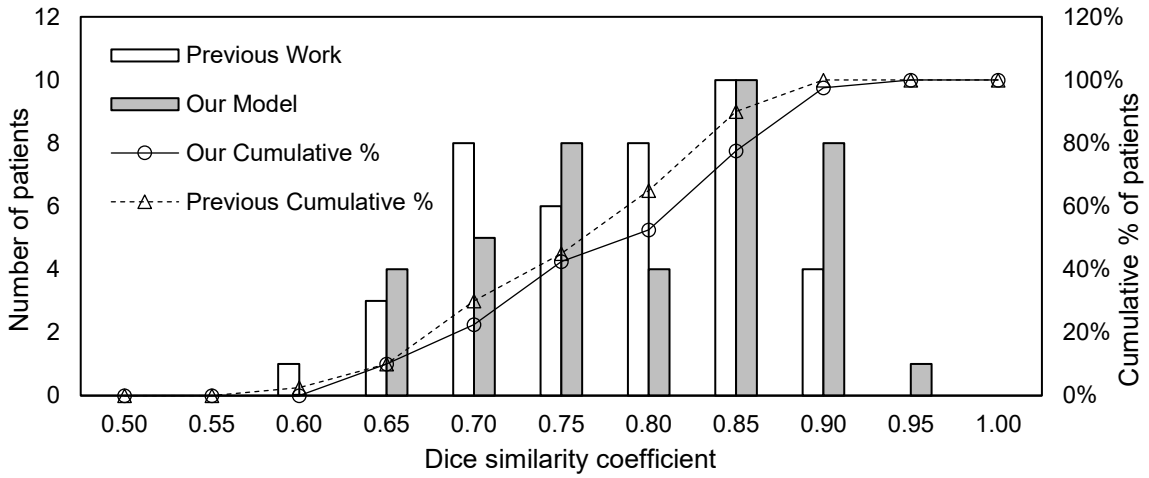


Figure 6. Histogram comparison of the DSC achieved by our preliminary network versus the previous study. The bin number represents the upper limit of the DSCs in that bin.

Visual comparison of some segmentation results (Figure 7) demonstrated our network's potential to perform better than the previous work when there are more than one lesions (or if the GTV is discontinuous on the axial slice) (Figure 7 (a-b)), or when there is significant CT or PET signal heterogeneity within the GTV (Figure 7 (c-d)).

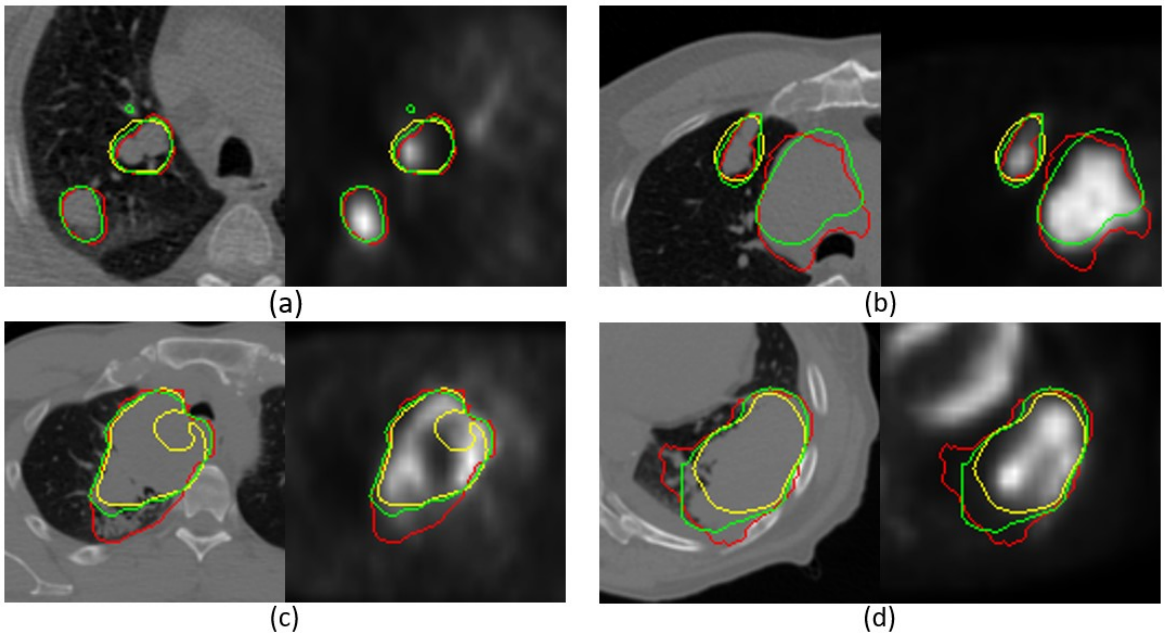


Figure 7. Case-specific comparison of the ground truth (Red), our work (Green), and previous work (Yellow). Case (a) and (b) showed that our network was able to detect multiple masses that were missed by the other network. Case (c) and (d) demonstrated our improvements in segmentation accuracy in the presence of significant PET heterogeneity and atelectasis.

C.2 3D Dual-Modality Network

Both CT and PET images in our database were acquired in 3D. Voxel information in adjacent slices provide additional inter-slice context that is missing when the images are processed slice-by-slice. In many segmentation tasks, direct training with 3D images were shown to achieve better accuracy at the cost of computational expense [94]. In this section, we explored converting our network to 3D to improve segmentation performance. In addition, for completeness of the work, single-modality networks using only CTs as inputs were also constructed to investigate the contribution of PET images to the dual-modality segmentation performance.

C.2.1 3D Data Preprocessing and Network Architecture

A framework to convert a network from 2D to 3D was presented in the article proposing V-Net [48], the 3D adaptation of U-Net. Keras also provided many references on the implementation of 3D CNNs, including an example of 3D image classification from CT scans (https://keras.io/examples/vision/3D_image_classification/) that we referenced.

The conversion involved two main aspects: data preprocessing and network architecture. While the 2D network took image slices in the TIFF format as inputs, the 3D network (Figure 7) used the entire volumetric images in the NRRD format. Because each scan had a different number of slices, the 3D volumes were resampled in the z-direction so that the input shapes are uniform. Modifications in the network architecture were mainly replacing the 2D functions with their 3D counterparts, e.g., Conv2D() to Conv3D() and MaxPool2D() to MaxPool3D().

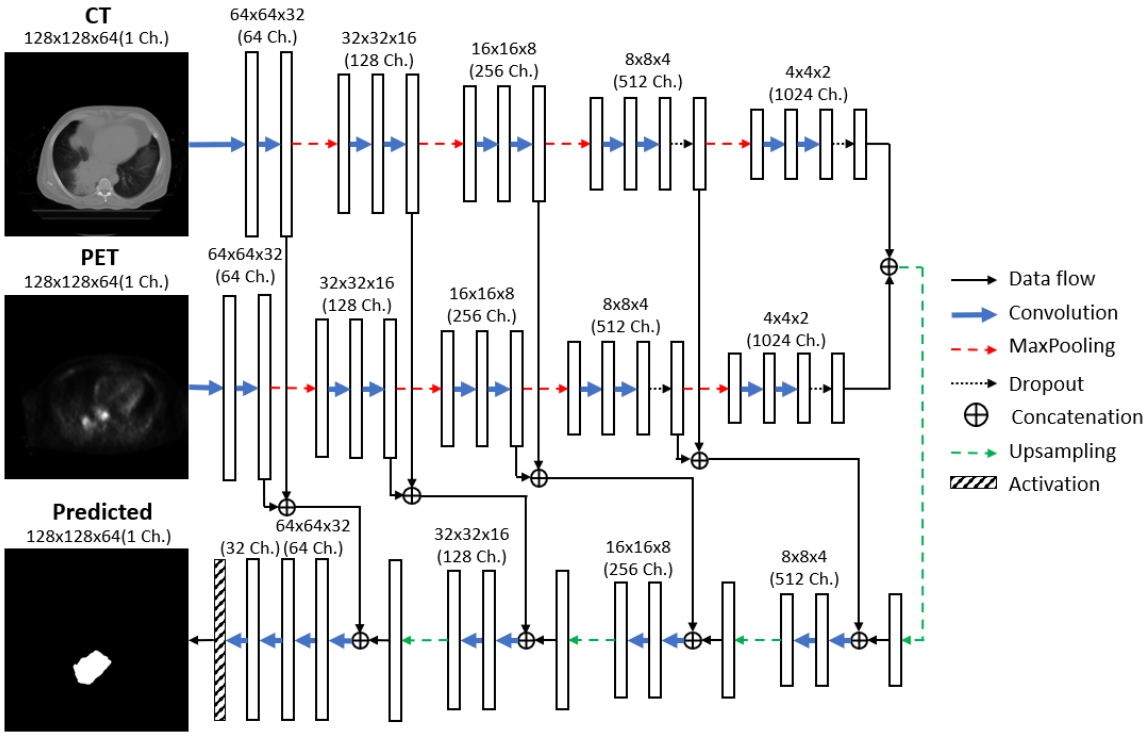


Figure 8. A diagram of the 3D dual-modality segmentation network architecture.

The challenges in this section of the work mostly arose from the increased need for computational resources. While training took 3-4 hours for the 2D dual-modality models on the computer with the NVIDIA 980M GPU, it would have taken significantly longer due to the restricted batch size to save RAM space and the increased computational burden of carrying out 3D convolution operations. Fortunately, we were able to utilize the Godel computing cluster at the VCU High Performance Research Computing (HPRC) Core Facility (https://hprc.vcu.edu). The godel.hprc.vcu.edu cluster is optimized for bioinformatics applications, with 1768 Intel and AMD 64 bit cores, each with at least 3 GB RAM/core, 4.8 TB of total RAM, 17 TB of /home space, tmp space of at least 180 GB/node, and 40 Gb/second Infiniband networking, 1.2TB of GPFS high-performance parallel file system storage.

C.2.2 Implementation of surface similarity evaluation metrics for 3D volumes

Before discussing the segmentation results, we would like to introduce a couple of surface evaluation metrics in addition to the Dice Similarity Coefficient (DSC) we had been using so far for a clearer understanding of the network performance: the Hausdorff Distance [95] and the mean Bi-directional Local Distance [96]. Whereas the DSC measures the general volumetric similarity, the surface metrics enables visualization of the global and local offset between the network-produced segmentation and the ground truth.

The Hausdorff distance is a measure of mutual proximity between two shapes by indicating the maximal distance between any point of one shape to the other. The original form of Hausdorff distance, $h(A,B)$, was formulated as a measure between the set of feature points in model A and those in model B, where

$$h(A,B) = \max_{a\in A} \min_{b\in B} \|a-b\|$$

and the double brackets || || is a norm of the points of A and B. Hausdorff average, a modified version, was shown to achieve superior performance over other variations [73]. The full algorithm of Hausdorff average was presented in Shapiro et al. The Python library SciPy [97] was utilized for HD calculation in this work.

By definition, the HD is demonstrably sensitive to outliers. Therefore, we employed the concept of the mean BLD. Whereas HD is a single value which represents the worst-case distance between two surfaces, BLD gives one local distance for each point in the reference surface, i.e., the maximum BLD is the HD. BLD especially excels in areas of folded or concave surfaces and has been used to assess the inter-observer variability in the manual segmentation of lung tumors.[33] The mean BLD thus provides a measure of the general distance between two surfaces. BLD is calculated in 3 steps [74]:

1. The forward minimum distance, FMinD, defined as the minimum distance from a point on a reference surface R to all points on a test surface T:

$$FMinD(p_{ref}, T) = d_{min}(p_{ref}, T) = \min_{p_i \in T} ||p_{ref} - p_i||_2$$

where $p_{ref} \in R$.

2. The backward maximum distance, BMaxD, at p_{ref} on the surface R:

$$BMaxD(T, p_{ref}) = \max_{p_i \in T}\{d | d = d_{min}(p_i, R), d = ||p_i - p_{ref}||_2\}$$

3. Finally, $BLD(p_{ref}, T) = max\{FMinD(p_{ref}, T), BMaxD(T, p_{ref})\}$.

In-house algorithms were written for the mean BLD calculation.

C.2.3 Performance Comparison with 2D and Single-Modality Alternatives

In addition to comparing the 2D with the 3D network, in order to investigate the role of PET in assisting segmentation, a single-modality network taking only CT images as input was also built based on the classic U-Net and its 3D variant, V-Net. The diagram of the 2D network is not shown here as it is not significantly modified from the standard U-Net.

In volume segmentation, 2D networks where the images are input by slice have the advantage of increased training sample size and faster training process. However, it only looks at a single slice each time and thus may lose the features related to the inter-slice correlation. 3D networks that learn from volumetric images, on the other hand, have fewer training samples and are slower to train, but are able to make use of the additional inter-slice features, which may benefit the segmentation of lung tumors for a wide range of tumor stages/sizes. Thus, both 2D and 3D versions of the segmentation network were investigated in this work to choose the network architecture which has the best performance.

The 3D PET & CT dual-modality model achieved mean DSC, HD, and mean BLD of 0.79 ± 0.10, 5.8 ± 3.2 mm, and 2.8 ± 1.5 mm, respectively. Comparatively, its 2D counterpart produced 0.78 ± 0.15, 7.6 ± 4.7 mm, and 3.1 ± 1.3 mm. Those same metrics for the single-modality CT-only models were 0.75 ± 0.16, 8.5 ± 5.8 mm, and 3.2 ± 1.4 mm for 2D, and 0.77 ± 0.12, 7.6 ± 4.7 mm, and 2.9 ± 1.4 mm for 3D.

A case study on 2D vs. 3D (Figure 8) illustrates how inter-slice context can help regulate the predicted segmentation and eliminate random outliers.

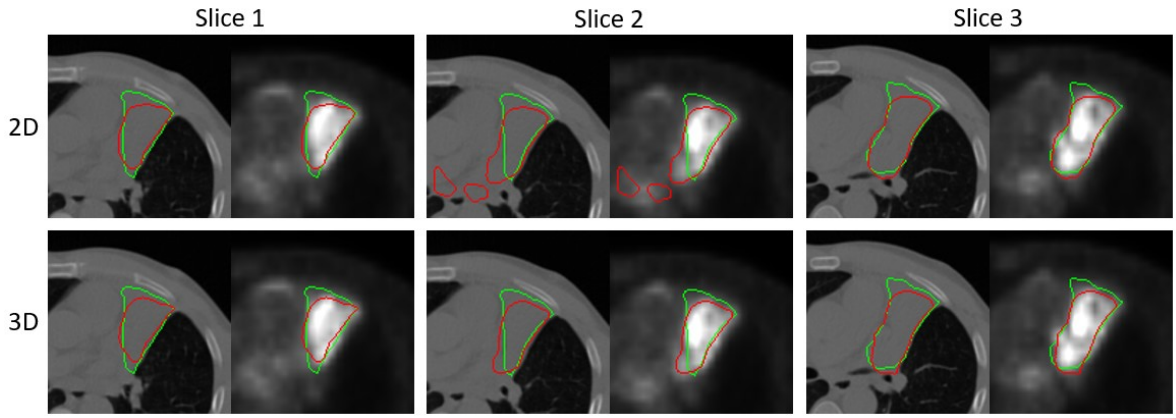


Figure 9. An example case comparing 2D versus 3D models. The top and bottom rows are the segmentations in 3 neighboring slices predicted by the 2D and 3D models, respectively. Green contour denotes the ground truth, and red denotes the network-predicted segmentation. The 3D model promoted inter-slice continuity and thus avoided misidentifying possible lymph nodes in the mediastinum in Slice 2.

Furthermore, Figure 9 shows how the dual-modality models achieved performance gain over their single-modality counterparts by incorporating the PET images. In case A (left), the PET showed activity in surrounding areas due to inflammation and tumor boundary on CT is obscured due to atelectasis. Case B is a small tumor adjacent to the chest wall with low to no PET signal. Visual assessment of these cases indicated that the tumor boundaries were not well defined on either PET or CT due to high density lung tissue (fibrosis or infectious changes) surrounding the tumor, or in some other cases the presence of atelectasis with inflammatory changes on PET between the acquisition of PET and CT made the accurate GTV definition impossible when the tumor boundary on CT was unclear to begin with.

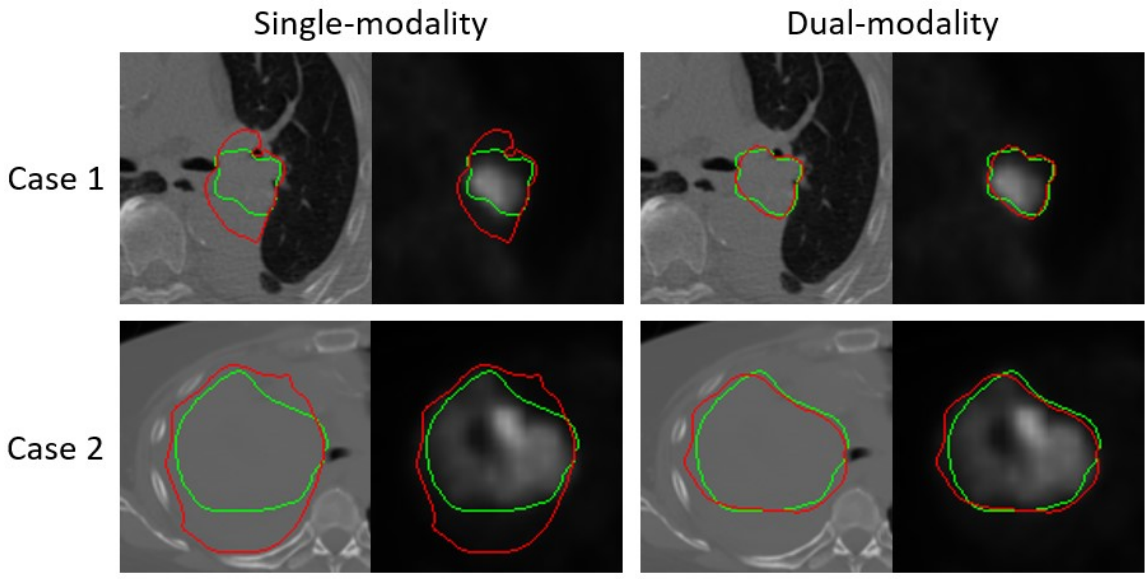


Figure 10. Two e*xample* case*s comparing single- versus dual-modality models. Green contour denotes the ground truth, and red denotes the network-predicted segmentation. The dual-modality model regulated the segmentation using the additional information from the PET images to better define the GTV when the tumor boundary on the CT images-only was not clear.*

Since the implementation of both 2D versus 3D and single- versus dual-modality were intertwined with the volume-based stratification that we will introduce later on in Section D.1, further discussion of these results will be included in Section D.2.

C.3 Optimization

To address the possibility of exploring alternative network architectures, while the literature of other deep learning-based segmentation studies using alternative network architectures were investigated, none stood out as a better alternative to U-Net. One popular choice for single-modality segmentation is the convolutional residual networks (CRNs), especially the ResNet [98]. The CRNs were created to deal with the problem of vanishing gradient and degrading accuracy as the depth of a network increases [99]. CRNs essentially utilizes skip connections to redirect information from a shallower layer to a deeper layer, enabling deep structures with minimal information degradation. Since

a deeper network theoretically has a higher capacity to learn, CRNs has the potential of obtaining more accurate segmentation results. However, the deeper a network is, the more parameters need to be trained, and therefore the more expensive it becomes computationally. Adapting the network to dual-modality would further increase the computational demand. Given that the established U-Net-based network had already produced meaningful results and our limited time and resources, we decided to focus on the current architecture and perform optimization through network functions and parameters as well as post-processing algorithms, which are discussed in this section.

C.3.1 Network Functions and Parameters

There are many functions and parameters in a network that can be tuned to improve segmentation performance, such as the number of convolution layers, the size of the convolution kernels, the number of features extracted at each layer, the rate and type of dropout functions, the choice of optimization algorithms and learning rate, etc. Without fundamentally changing the network architecture, all of these factors can be experimented with to achieve faster training or better results. While we investigated many different combinations of these factors throughout the development of the network, it was by no means comprehensive, as hyperparameter tuning in deep learning is an art of its own and a highly iterative and arduous process. Here we will elaborate on two functions and parameters that improved the network performance demonstrably: the optimizer and the dropout layers.

An optimizer is an algorithm by which the learnable parameters, i.e., weights and biases in a network are updated so as to minimizes an error function or a loss function that is defined with certain metrics chosen to measure the network performance. The two optimizers tested were the stochastic gradient descent (SGD) and Adam (Adaptive

Moment Estimation) [100] which is a updated version of stochastic gradient descent method that is based on adaptive estimation of first-order and second-order moments. According to Kingma et al., the method is "computationally efficient, has little memory requirement, invariant to diagonal rescaling of gradients, and is well suited for problems that are large in terms of data/parameters". We found that Adam converged much faster, while SGD tended to oscillate and sometimes did not reach a DSC as high as Adam did in the validation dataset (Figure 10). While there are arguments being made that SGD generalizes better and reaches an improved final performance [101], [102], we did not observe that effect, and therefore Adam was utilized for the remainder of this work.

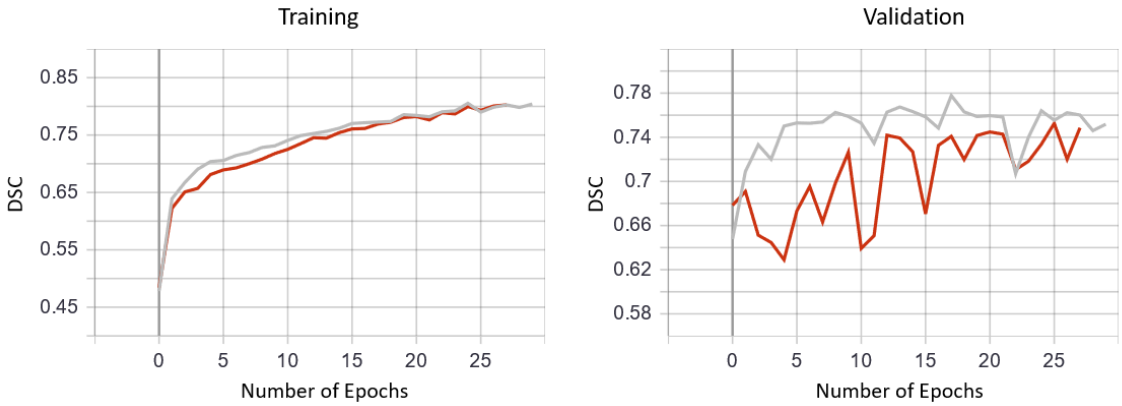


Figure 11. Comparison between two training sessions using Adam (Grey) versus SGD (Red) optimization algorithms.

The original dropout function was proposed to improve generalization and prevent overfitting by randomly “dropping-out” or zeroing neurons during training [103]. Tompson et. al. [104] argued that even after features from a certain central pixel were discarded, the features from adjacent pixels might retain the strong correlations that may cause overfitting. Therefore, they adopted the SpatialDropout function where the entire feature channels were dropped instead of individual neurons, and found that their network performance was improved, especially for small training sets. In our work, especially when the stratified training strategy was employed on the 3D dataset (Section D.1),

overtraining was observed where the training DSC quickly reached 0.9, while the validation DSC became stagnant just under 0.8, as shown in Figure 11. Replacing the Dropout function with the SpatialDropout function appeared to slow down the DSC increase in the training dataset and improve the final performance.

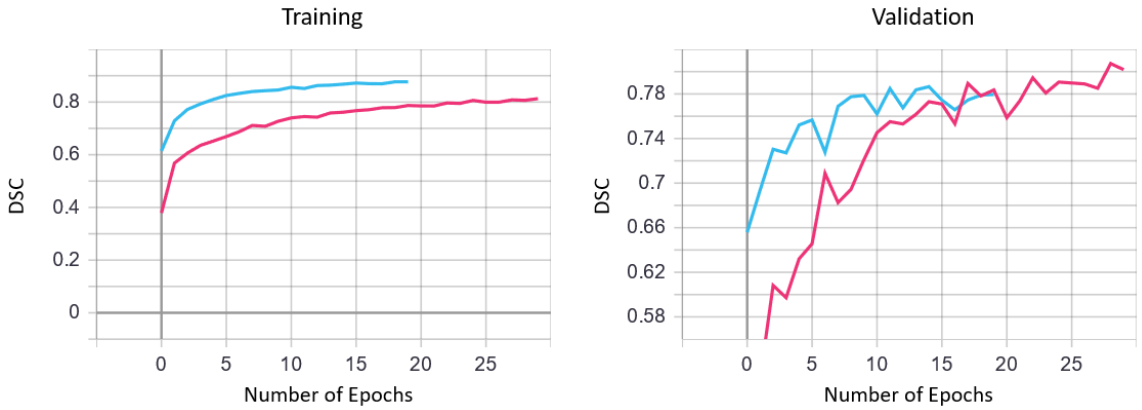


Figure 12. Comparison between two training sessions with the regular Dropout function (blue) versus the SpatialDropout (pink) function.

C.3.2 Conditional Random Field Post-Processing

The output of the network is a probability map with voxel values from 0 to 1, with 1 labeling the predicted GTV. As shown in the example of the edge of a GTV in Figure 12, most of the pixel values are close to either 0 or 1, so initially, the maps were simply thresholded at 0.5 to produce the final segmentation. A threshold value of 0.75 was also tested and was found to have negligible impacts on either the appearance of the segmentations or the DSC.

0.992297	1	1	1	1	1
4.72958e-07	1	1	1	1	1
1.46943e-17	1	1	1	1	1
4.38427e-29	2.1527e-05	1	1	1	1
8.46445e-38	9.76889e-19	0.999993	1	1	1
0	4.61141e-30	1.50246e-06	1	1	1
0	0	2.99326e-20	0.293104	1	1
0	0	2.63975e-33	3.84817e-17	0.983037	1
0	0	4.5611e-37	1.28075e-30	1.39565e-12	1
0	0	0	0	2.5543e-24	1.05368e-11
0	0	0	0	3.29637e-36	2.64634e-29
0	0	0	0	0	0

Figure 13. Edge pixels of a predicted segmentation.

However, upon closer inspection, some network-predicted segmentations have small “pockets” of mis-identified voxels (Figure 13) even after thresholding. To clean up these voxels, the output tumor probability maps were passed through a Conditional Random Field (CRF) filter [105]. Conditional random field (CRF) assesses areas of connectedness and reassigns voxels to the most probable voxel class with consideration of its neighboring voxels. Essentially, CRF looks at the surrounding voxel assignment and reassigns a voxel to its most probable class based on conditional probabilities (given that the surrounding voxels are class X, what is the probability this voxel is class Y). CRF has been shown to eliminate small, isolated segments which are far from the main area of the same class. While CRF post-processing did not have any significant impact of the DSC, it cleaned up the segmentation and prevented the HD from being hijacked by outliers.

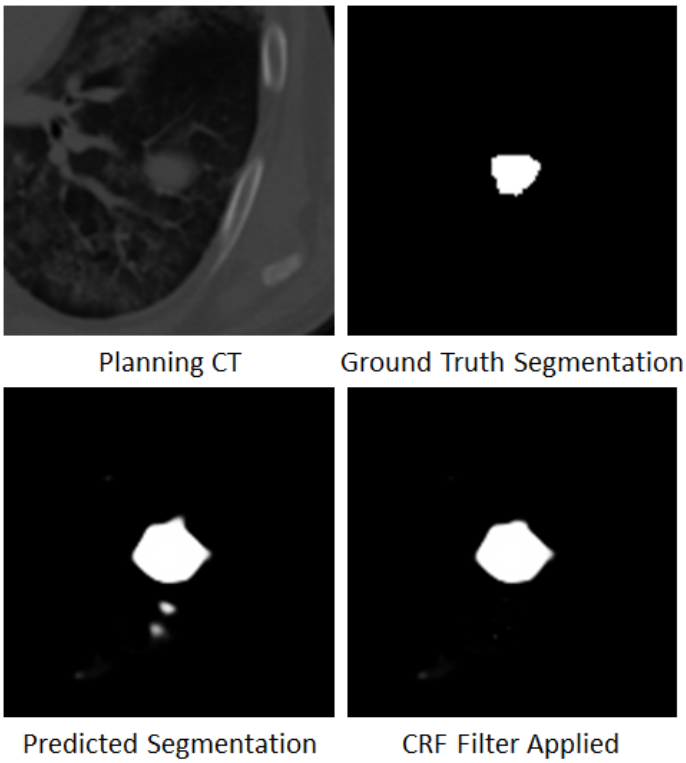


Figure 14. An example of a predicted segmentation with misidentified pixels and how a conditional random field filter can help erase them without modifying the main segmentation significantly.

C.5 Conclusion and Future Work

The preliminary dual-modality deep learning network we constructed was able to segment lung GTVs from planning CT and diagnostic PET images, performing comparatively, as assessed by the DSC, a state-of-the-art deep learning-based method when validated on a subset of our internal database. The conversion from 2D data and architecture to 3D, the optimization of network functions and parameters, and the introduction of conditional random field postprocessing further boosted the segmentation performance. The 3D dual-modality model outperformed its 2D and single-modality counterparts according to both volumetric and surface evaluation metrics.

In future work, with more time and resources, it would be interesting to explore other network architectures such as the previously mention CRNs. In addition, the recurrent

feature fusion network proposed by Bi et al. [54] is also of particular interest because of its unique approach to progressively fuse different image modalities.

D. Volume-Based Stratification (Sub-Aim 3.1)

Evaluate the potential of stratified training/validation/testing strategies to improve performance of the developed network.

D.1 Introduction

For clarification, this section of our work was originally a part of the Specific Aim 3 as one of the mechanisms implemented to improve the segmentation performance. However, throughout the course of this project, volume-based stratification of the datasets was shown to be effective in various stages, so it makes sense to the author that it should be placed here in its own section for the flow of the thesis.

After the initial segmentation network was trained and tested on all cases in Specific Aim 1, a visual inspection of the low-performing cases as well as the DSC versus GTV volume trends (Figure 15) showed that the initial network tended to perform worse at small volumes and overestimated the size of small GTVs. This could be a result of the inherently low resolution of PET as well as respiratory motion during the long acquisition period of PET. Small tumors are more prone to displacement due to respiratory movements and the signals from small tumors are more likely to be spread out in the neighboring voxels due to volume averaging effect [106]. In fact, differences in the automated segmentation performance of small and large lung tumors have been reported in a previous study using CT only, and a focal loss function and an attention U-Net network were explored to improve the segmentation accuracy for small tumors specifically [42]. Our intuition was to

encourage the network to treat small and large tumors differently and place a tighter restriction on the size of small tumors and put less weights on the low PET signal in the peripheral of the tumors. To that end, we investigated the effect of dataset stratification based on tumor size.

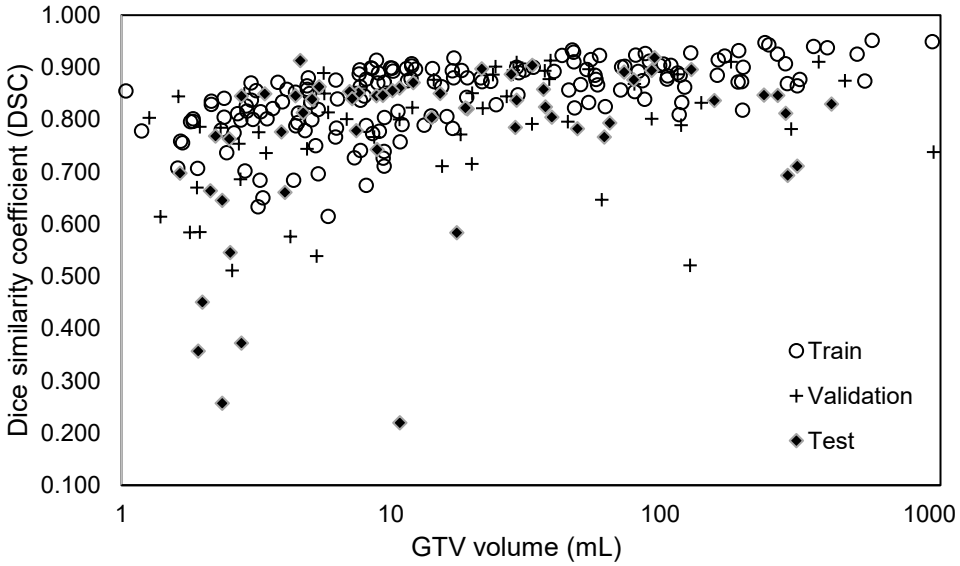


Figure 15. DSC vs volume for all patients. One case with a tumor volume of over 1000 mL was not shown in the graph since it can be considered a geometric outlier and makes the graph harder to read.

D.2 Stratified Training, Validation and Testing

A stratified training strategy based on the GTV volume was investigated. Separate models were trained to segment large and small tumors so that the individual models can better adapt to the image features specifically beneficial for the tumors in the different size group. The dataset was stratified into 2 groups: the small GTVs below 25 mL (183 cases with a median GTV of 7.0 mL) and the large GTVs above 25 mL (107 cases with a median GTV of 86.0 mL). Each group was divided into its own training/validation/test subsets with a ratio of approximately 3:1:1. The threshold volume for the stratification was chosen as the estimated GTV volume to separate the cases with hypofractionation from the ones

with conventional fractionation. The choice also ensures adequate data size in each training/validation/test subset.

In order to find the optimal model for each size group, all valid training/validation/testing combinations were evaluated, i.e., two stratified models were trained for the small GTV subset, one was trained on all cases combined regardless of GTV size and validated with the small GTV subset, the other was both trained and validated with the small GTV subsets. Two stratified models were trained for the large GTVs in similar fashion, i.e., one was trained on all cases combined regardless of GTV size and validated with the large GTV subset, the other was both trained and validated with the large GTV subsets. All combinations were shown in Table 1. The overall performance metrics for the entire case population will be calculated as the combined metrics from each test subset.

Paired t-tests were conducted in Microsoft Excel (Microsoft. Version 16.0.) to evaluate the differences in metrics, with a two-tailed p-value < 0.05 indicating statistically significant improvement. Comparisons were made between single- and dual-modality, 2D and 3D, stratified and un-stratified, and single-modality unstratified vs. dual-modality stratified networks.

Table 3. The mean Dice similarity coefficient (DSC), the Hausdorff distance (HD), and the mean bi-directional local distance (BLD) between the ground truth segmentation and the segmentation predicted by the 2D vs. 3D, single- vs. dual-modality networks trained/validated/tested with all (non-stratified) vs. stratified data subsets.

Mean DSC					
		2D		3D	
Training	Validation & Testing	Single-Modality (CT)	Dual-modality (PET & CT)	Single-Modality (CT)	Dual-modality (PET & CT)
All	All	0.749 ± 0.163	0.776 ± 0.140	0.772 ± 0.123	0.790 ± 0.103

All	Large	**0.821 ± 0.073**	**0.841 ± 0.047**	**0.830 ± 0.087**	**0.854 ± 0.052**
Large	Large	0.818 ± 0.080	0.830 ± 0.052	0.822 ± 0.074	0.837 ± 0.080
All	Small	0.679 ± 0.196	0.728 ± 0.170	0.714 ± 0.151	0.758 ± 0.128
Small	Small	**0.784 ± 0.101***	**0.797 ± 0.095***	**0.795 ± 0.092***	**0.821 ± 0.083***
Combined Overall		0.798 ± 0.093	0.814 ± 0.083	0.808 ± 0.090	0.833 ± 0.071
HD (mm)					
		2D		3D	
Training	Validation & Testing	Single-Modality (CT)	Dual-modality (PET & CT)	Single-Modality (CT)	Dual-modality (PET & CT)
All	All	7.95 ± 5.39	6.37 ± 3.42	7.59 ± 4.73	6.84 ± 3.22
All	Large	12.12 ± 6.29	**9.68 ± 3.99**	**10.89 ± 7.11**	**8.53 ± 3.79**
Large	Large	**11.79 ± 5.89**	11.19 ± 4.14	11.95 ± 5.56	10.30 ± 4.20
All	Small	5.44 ± 2.02	4.84 ± 1.99	5.95 ± 2.81	4.56 ± 1.77
Small	Small	**4.45 ± 1.59***	**4.49 ± 1.58**	**4.53 ± 1.63***	**4.41 ± 1.70**
Combined Overall		6.99 ± 5.08	6.28 ± 3.64	6.90 ± 3.67	5.95 ± 3.48
Mean BLD (mm)					
		2D		3D	
Training	Validation & Testing	Single-Modality (CT)	Dual-modality (PET & CT)	Single-Modality (CT)	Dual-modality (PET & CT)
All	All	3.93 ± 2.18	3.43 ± 2.06	3.59 ± 1.73	3.24 ± 2.52
All	Large	**4.62 ± 2.36**	**3.85 ± 1.44**	**4.49 ± 2.11**	**3.83 ± 1.79***
Large	Large	4.71 ± 2.62	4.11 ± 1.43	4.65 ± 2.56	4.30 ± 1.50
All	Small	3.95 ± 2.23	3.08 ± 2.11	3.55 ± 1.81	3.16 ± 1.77
Small	Small	**2.29 ± 0.91***	**2.40 ± 1.01***	**2.03 ± 0.63***	**2.18 ± 1.07***
Combined Overall		3.15 ± 1.96	2.94 ± 1.34	2.95 ± 1.18	2.80 ± 1.36

These results were obtained using the corresponding testing data sets. The data in **bold** were from the best performing stratified subsets and were used to calculate the combined overall metric. * denotes results that are statistically significant (paired t-test, $p < 0.05$) compared with their counterparts in the same cell blocks.

The stratified models produced significantly better overall metrics in all categories except for the mean HD for the 2D and 3D dual-modality networks. A closer inspection of the subsets showed that the small GTV subset appeared to clearly ($p = 0.0005$ to 0.05) benefit from separate training, i.e., training and validation by the small GTV subsets. While the 3D models do not appear to significantly improve the overall metrics, with the exception of the improved DSC by the 3D dual-modality stratified model over its 2D counterpart ($p = 0.05$), the large GTV subsets achieved better HDs in 3D models ($p = 0.05$ for the model trained by all cases and $p = 0.03$ for the model trained by the large training subset), potentially due to a reduction in outliers.

D.3 Conclusion

Our proposed stratification strategy was shown to be especially effective in improving segmentation accuracy for small GTVs. Separate training and validation of the small GTV cases improved all three evaluation metrics for all 2D/3D single-/dual-modality models. With the 3D dual-modality network architecture, stratification increased the mean DSC of the small GTV test cases from 0.758 ± 0.128 to 0.821 ± 0.083 with statistical significance.

It is important to note that the method conducted in this study was based on the GTV volume from the manual contours. However, in clinical deployment of our method, the volume is unknown since the GTV is yet to be segmented. We propose a two-step process: In the first step, cases will be passed through a segmentation model trained with all the training cases regardless of tumor volume to obtain the estimated GTV volume. Then in the second step, those cases will be assigned to its corresponding stratified model based on the volume calculated from its preliminary segmentation.

E. Specific Aim 2

Evaluate the potential of adding radiomics features as a third input to the network to improve segmentation performance.

E.1 Introduction

As discussed in the introduction (Section A.5.3), no previous work was published on using radiomics feature images as deep learning network inputs for tumor segmentation. The goal of this section of our work is to investigate the validity of this approach and whether one or more radiomics features can be identified to enhance the overall segmentation performance achieved so far. To that end, we introduced a novel pipeline unique to our needs to extract radiomics feature values from a customized VOI, to select features with the most potential to improve the segmentation performance, to incorporate these voxel-based feature images in building a predictive model, and to evaluate these models against the ones built without radiomics inputs.

All radiomics features evaluated in this work were extracted from the planning CT images, because they had been more widely studied than PET features and were found to be more stable [74]. As we mentioned in the introduction, radiomics features are sensitive to the imaging protocols and parameters. While all the planning CTs in our internal database were acquired on a dedicated CT simulator, the origins of the diagnostic PET/CTs were more heterogenous, which would in turn affect the quality of the PET radiomics features, the extent of which is beyond the scope of this work.

E.2 Feature Extraction

Codes to extract the voxel-based radiomics feature maps were written in Python with the open-source package Pyradiomics 2.2 [107]. The toolkit enables the extraction of both the feature value (a single quantitative descriptor of the entire ROI/VOI) and the voxel-based feature map of 120 features including 19 first-order, 26 shape-based (both 2D and 3D), and 75 texture-based features from 5 commonly used feature classes: GLCM, GLSZM, GLRLM, NGTDM, and GLDM. As previously discussed, the shape-based features rely on pre-established tumor segmentations and are therefore excluded from our work. The complete list of the 104 features investigated are shown in Table 4, and further information about their definitions and formulas can be found in the online documentations for Pyradiomics (https://pyradiomics.readthedocs.io/en/latest/features.html). The feature extraction process, especially for the voxel-based feature images, was computationally intensive and was executed through Google Colab Pro+.

Table 4. List of radiomics features investigated in this work.

First Order Statistics	Imc1	LargeAreaEmphasis
10Percentile	Imc2	LargeAreaHighGrayLevelEmphasis
90Percentile	InverseVariance	LargeAreaLowGrayLevelEmphasis
Energy	JointAverage	LowGrayLevelZoneEmphasis
Entropy	JointEnergy	SizeZoneNonUniformity
InterquartileRange	JointEntropy	SizeZoneNonUniformityNormalized
Kurtosis	MCC	SmallAreaEmphasis
Maximum	MaximumProbability	SmallAreaHighGrayLevelEmphasis
MeanAbsoluteDeviation	SumAverage	SmallAreaLowGrayLevelEmphasis
Mean	SumEntropy	ZoneEntropy
Median	SumSquares	ZonePercentage
Minimum	**GLRLM**	ZoneVariance
Range	GrayLevelNonUniformity	**NGTDM**
RobustMeanAbsoluteDeviation	GrayLevelNonUniformityNormalized	Busyness
RootMeanSquared	GrayLevelVariance	Coarseness
Skewness	HighGrayLevelRunEmphasis	Complexity
TotalEnergy	LongRunEmphasis	Contrast
Uniformity	LongRunHighGrayLevelEmphasis	Strength
Variance	LongRunLowGrayLevelEmphasis	**GLDM**

GLCM	LowGrayLevelRunEmphasis	DependenceEntropy
Autocorrelation	RunEntropy	DependenceNonUniformity
ClusterProminence	RunLengthNonUniformity	DependenceNonUniformityNormalized
ClusterShade	RunLengthNonUniformityNormalized	DependenceVariance
ClusterTendency	RunPercentage	GrayLevelNonUniformity
Contrast	RunVariance	GrayLevelVariance
Correlation	ShortRunEmphasis	HighGrayLevelEmphasis
DifferenceAverage	ShortRunHighGrayLevelEmphasis	LargeDependenceEmphasis
DifferenceEntropy	ShortRunLowGrayLevelEmphasis	LargeDependenceHighGrayLevelEmphasis
DifferenceVariance	**GLSZM**	LargeDependenceLowGrayLevelEmphasis
Id	GrayLevelNonUniformity	LowGrayLevelEmphasis
Idm	GrayLevelNonUniformityNormalized	SmallDependenceEmphasis
Idmn	GrayLevelVariance	SmallDependenceHighGrayLevelEmphasis
Idn	HighGrayLevelZoneEmphasis	SmallDependenceLowGrayLevelEmphasis

Two rounds of feature extraction were performed for:

1) Feature Selection:

To identify the features with the most potential to improve the segmentation performance, i.e., distinguish the tumoral tissue from its surroundings, it logically follows to compare the feature values in the tumoral region with the ones in the peritumoral region. Previous lung cancer radiomics studies in nodular classification [69], disease spread prediction [70], and general methodology proposal [71] analyzed the peritumoral region by dilating the manual tumor segmentations to create a rim or ring around the GTV. The thickness of the rings ranged from 8 mm to 30 mm. For our work, we took the rough average and created rings with a thickness of 20 mm by dilating the manual segmentations. Feature values of the VOIs, i.e., the tumoral and peritumoral regions, were then calculated for 104 features and the training dataset of 162 cases. Figure 8 shows an example of an CT image with its tumoral and peritumoral segmentations, as well as the corresponding feature images for visualization. In the feature selection step, only a single feature value is recorded from each VOI.

2) Network Input:

The feature images that were used as inputs to the deep learning network were extracted using masks with all voxels labeled as True, hence including all voxels from the CT images for voxel-wise feature extraction, with a kernel size of 3 x 3 and a bin width of 25. This was performed on all 290 including the training, validation, and testing datasets but only for the features selected for investigation in the next section, where examples of them will be shown. This step was time-consuming, as voxel-based extraction of the more complex feature classes such as GLCM often took hours for cases with large image volumes.

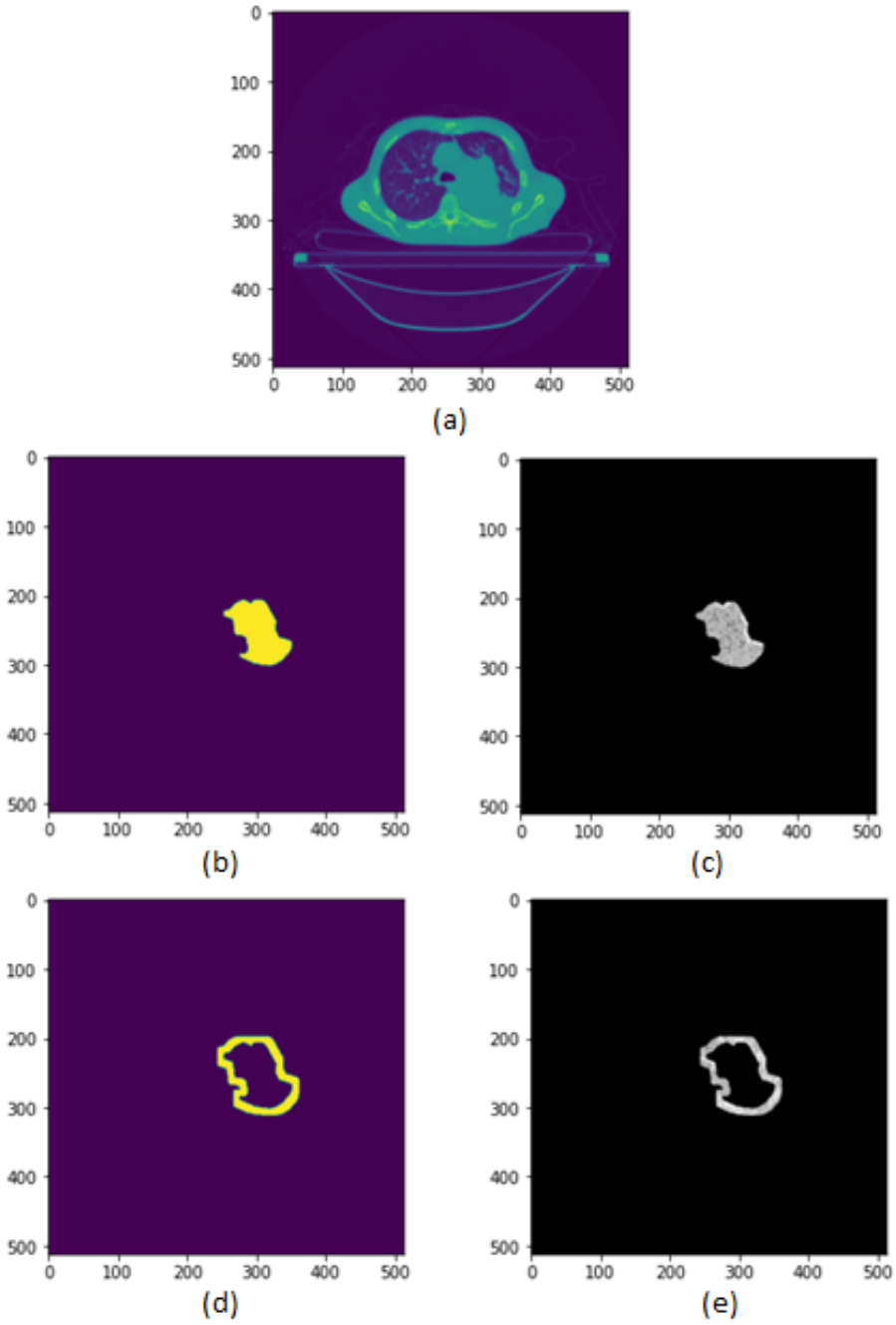


Figure 8: An example of a CT image (a), the tumoral segmentation (b) and its corresponding extracted GLDM Dependence Entropy image (c), the peritumoral segmentation (d) and its feature image (e). The GLDM Dependence Entropy feature value within the tumoral region is 7.42 versus 4.88 in the peritumoral region.

E.3 Feature Selection with SelectKBest

In total, the values of 104 feature were calculated for the tumoral and peritumoral regions of all 162 training cases. SelectKBest from the Python scikit-learn package [108] was utilized to reduce the candidate pool and eliminate redundant features. As its name suggests, SelectKBest works by removing all but the k highest scoring features based on the chosen univariate statistical test, for which we employed the classic ANOVA F-test, i.e., the function f_classif. F-test is a popular choice for classification tasks because it calculates the ratio between variances of different groups to measure how distinctly different they are. Essentially, features receiving higher F-Scores were better at distinguishing between the tumoral and peritumoral regions.

With k=10, the features selected are listed in Table 5.

*Table 5. Features selected by SelectKBest and ANOVA F-test from all cases. The features in **bold** were chosen for investigation.*

Feature Class	Feature Name	F-Score
GLDM	**DependenceEntropy**	430.6955
NGTDM	**Coarseness**	356.4575
GLSZM	**ZoneEntropy**	345.1818
GLCM	MCC	327.6402
First Order	RootMeanSquared	311.2521
GLCM	ClusterTendency	303.2483
GLCM	SumSquares	268.8430
GLDM	GrayLevelVariance	245.8914
First Order	Variance	245.8567
GLRLM	GrayLevelVariance	241.2145

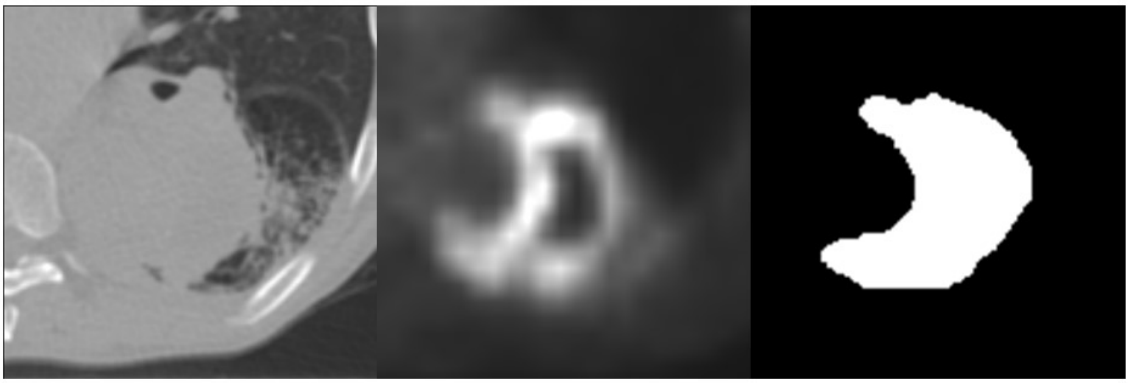
In addition, because the stratified training/validation/testing strategy improved the segmentation results, we were also interested in the features that would especially benefit the large or the small GTV subsets (Table 6).

*Table 6. Features selected by SelectKBest and ANOVA F-test from the large (top) and small (bottom) GTV subsets. The features in **bold** were chosen for investigation.*

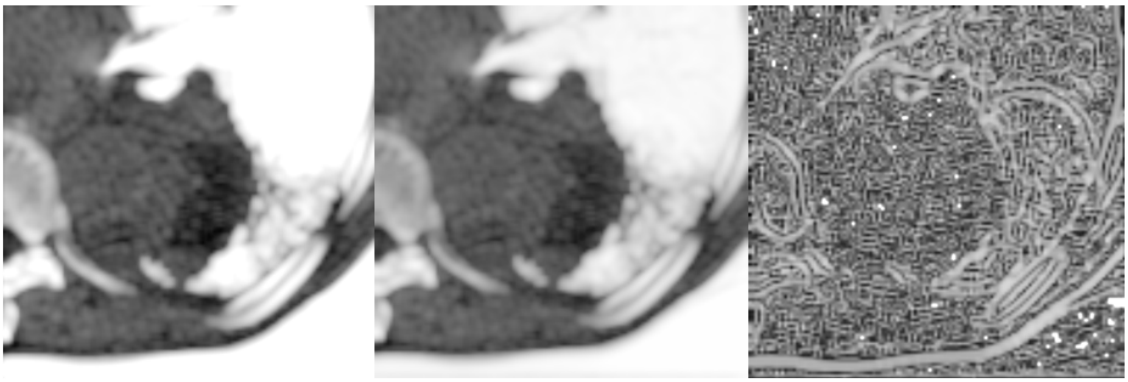
Feature Class	Feature Name	F-Score
GLDM	**DependenceEntropy**	312.7781
GLCM	**Correlation**	304.9258
GLCM	SumSquares	295.4277
First Order	Variance	253.5094
GLDM	GrayLevelVariance	253.3518
GLCM	SumEntropy	236.4214
GLRLM	GrayLevelVariance	231.0356
First Order	MeanAbsoluteDeviation	211.8688
GLCM	JointEntropy	201.1660
First Order	RootMeanSquared	200.7453

Feature Class	Feature Name	F-Score
First Order	**Energy**	350.4393
First Order	TotalEnergy	350.4393
First Order	**RootMeanSquared**	331.8582
NGTDM	Coarseness	294.7814
First Order	Median	268.6130
GLSZM	ZoneEntropy	261.7176
GLCM	Correlation	244.7554
GLCM	Idmn	239.0278
First Order	10Percentile	221.9056
GLSZM	ZonePercentage	216.1771

The top two scoring features from each list were chosen for investigation. Visual inspection of sample extracted feature images and analysis of the feature definitions were conducted to further assess the suitability (Figure 16.a-b). It is interesting to observe that the high-scoring features for the small GTVs were mostly the first order features that are closely related to the local voxel value of the CT image. This is consistent with the fact that most small GTVs were surrounded by the lung parenchyma, and the hypothesis that the voxel intensity could be the one of the most discriminative features.



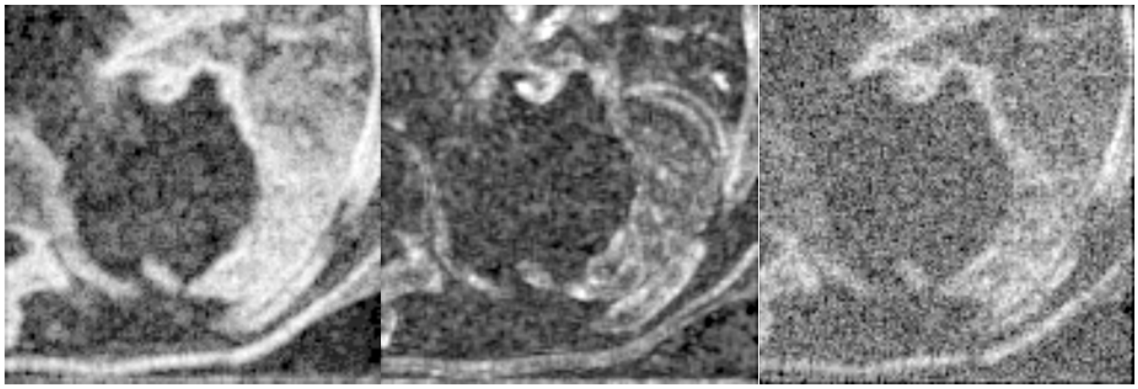
Planning CT **PET** **GTV Mask**



First Order Energy **First Order RMS** **GLCM Correlation**

$X^2(i,j)$	$\sqrt{\frac{1}{N_P}\sum_{i=1}^{N_P} X^2(i)}$	$\frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_g}(ij)p(i,j) - \mu_x\mu_y}{\sigma_x\sigma_y}$
$X(i,j)$ is the voxel value of the row i and column j element of the image. This is a measure of the magnitude of the local voxel value.	RMS is Root Mean Squared abbreviated for space. $X(i)$ is the element i in the calculation kernel of size N_P. This also measures the magnitude of the voxel value but averaged over the kernel.	$p(i,j)$ is the row i and column j element of the GLCM. N_g is the number of gray levels in the image. μ_x, μ_y and σ_x, σ_y are the mean and standard deviation of p_x and p_y of row x and column y. This measures the linear dependency of gray levels to the neighboring voxels.

Figure 16.a. Three of the six selected features with their sample voxel-based images and their definitions. The planning CT that the features were extracted from, the PET, and the GTV mask were also shown for reference.



GLSZM Zone Entropy	NGTDM Coarseness	GLDM Dependence Entropy
$-\sum_{i=1}^{N_g}\sum_{j=1}^{N_s} p(i,j)log_2(p(i,j))$	$\frac{1}{\sum_{i=1}^{N_g} p_i s_i}$	$-\sum_{i=1}^{N_g}\sum_{j=1}^{N_s} p(i,j)log_2(p(i,j))$
$p(i,j)$ is the row i and column j element of the GLSZM. N_g is the number of gray levels. N_s is the number of zone sizes. This measures the heterogeneity in the texture patterns.	s_i is the sum of absolute differences between the center voxel and the neighboring voxels. This measures the spatial rate of change, i.e., coarseness.	This is the same formula as the one for GLSZM Zone Entropy, as it also measures the entropy or heterogeneity among the matrix elements but for the GLDM.

Figure 16.b (Cont'd) Three of the six selected features with their sample voxel-based images and their definitions.

GLDM Dependence Entropy was the highest scoring feature on both the all-case list and the large-GTV list, so the third place feature on the all-case list was also tested. For the small-GTV list, Total Energy is similarly defined as Energy, hence it was deemed redundant and eliminated. The next place feature Root Mean Squared was selected instead.

We recognize that this feature selection process was by no means thorough, as studies on feature selection often employ multiple selection algorithms and more than one round of dimension reduction. However, this section of our work was intended to be a preliminary investigation into the validity of a novel method.

E.4 Incorporating Radiomics in Segmentation

E.4.1 Image Preprocessing and Normalization

The main challenge in preprocessing the voxel-based feature images is the unbalanced distribution of voxel values for three of the features (Table 7): First Order Energy, First Order Root Mean Squared (RMS), and NGTDM Coarseness.

Table 7. Specifications of the voxel values in each voxel-based feature image dataset.

Feature Name	Min	Max	Mean	Median
GLDM Dependence Entropy	-3.20e-16	4.75	3.71	3.78
GLSZM Zone Entropy	-3.20e-16	4.75	2.73	2.85
GLCM Correlation	-0.92	1.00	0.08	0.07
First Order Energy	0.00	4.31e8	5.67e6	8.22e5
First Order RMS	0.00	1.95e8	3.37e6	1.76e5
NGTDM Coarseness	0.00	1.00e6	1.50e4	0.17

The voxel values of GLDM Dependence Entropy, GLSZM Zone Entropy, and GLCM Correlation were constrained by their definitions and relatively evenly distributed (Figure 17), so the standard normalization procedure was applied using the minimum and maximum. However, the voxel values of the other three features: First Order Energy, First Order RMS, and NGTDM Coarseness are heavily concentrated near the bottom of the range (Figure 18). There are many reasons why data normalization is important in machine learning, among which the most important ones: First, it scales each feature to a similar range so that the network is not overly biased toward contributions from features of higher values; Second, it reduces what is called an Internal Covariate Shift where the distribution of the inputs to layers, especially those deep in the network, change after each input, which cause the learning algorithm to forever chase a moving target, and thus slowing down training [109]. Because the voxel value distributions for First Order Energy, First Order RMS, and NGTDM Coarseness are so bottom-heavy, simple normalization by the minima and the maxima would render the information in the

majority of the voxels trivial, while the issue of internal covariate shift remains. To mitigate this problem, Log transformation was applied to the three feature images before the standard normalization. The specifications of the voxel values post-transformation are shown in Table 8. As seen in Figure 18, the voxel value distributions for First Order Energy and First Order RMS became much more balanced. NGTDM Coarseness, on the other hand, was still not ideally distributed. More sophisticated normalization techniques might be necessary in future work.

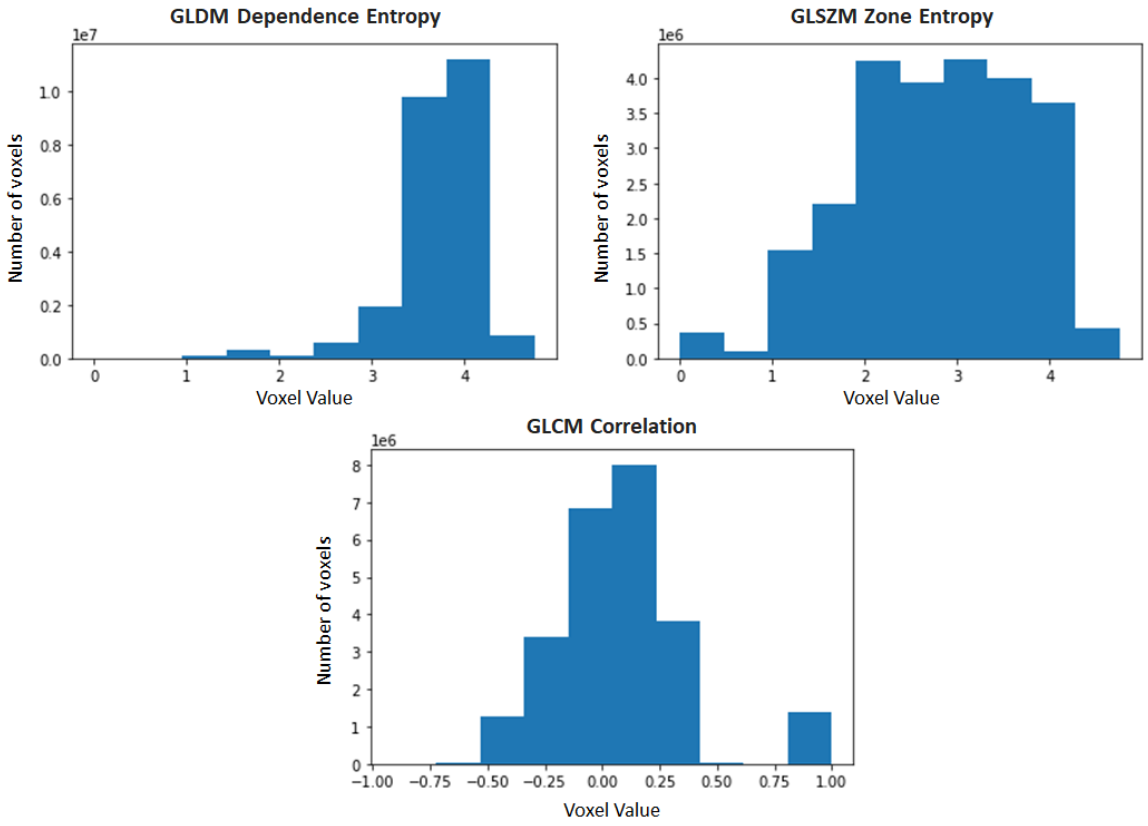


Figure 17. Voxel value distribution of GLDM Dependence Entropy, GLSZM Zone Entropy, and GLCM Correlation in the training datasets.

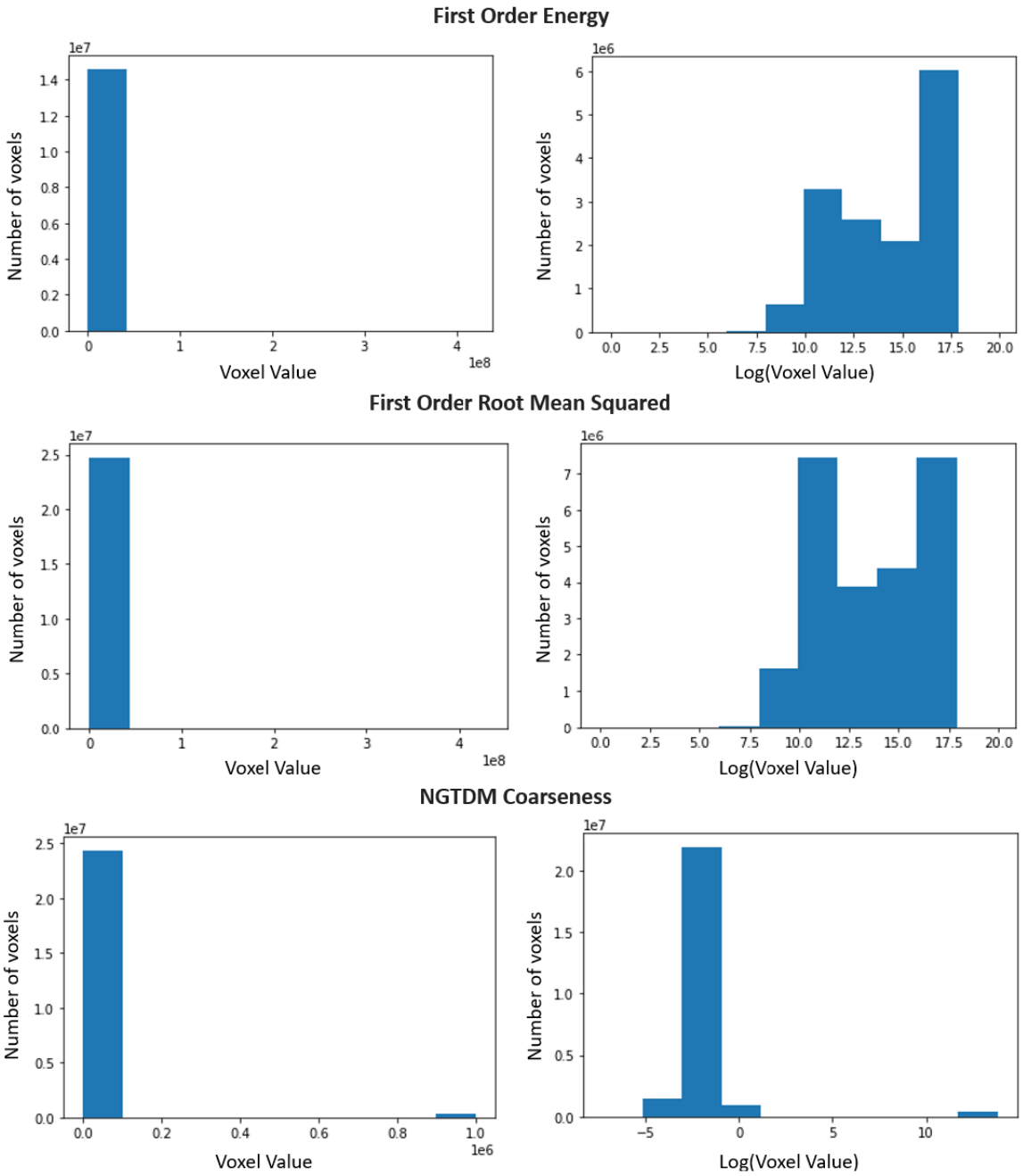


Figure 18. Voxel value distribution of First Order Energy, First Order Root Mean Squared, and NGTDM Coarseness in the training datasets.

Table 8. Specifications of the voxel values in each voxel-based feature image dataset after log transformation.

Feature Name	Min	Max	Mean	Median
First Order Energy	0.00	19.88	13.52	13.62
First Order RMS	0.00	18.51	10.78	11.38
NGTDM Coarseness	-7.24	13.82	-1.67	-1.77

E.4.2 Implementation and Results

A third convolutional arm was added to the 3D dual-modality network to learn from the feature image input (Figure 19). Other network parameters remained the same. Models were trained and tested for all six features with all cases as well as the stratified training/validation/testing datasets. Training was conducted on the Godel cluster at the VCU HPRC as well as Google Colab Pro+. The segmentation results, evaluated using the DSC, HD, and mean BLD metrics, are shown in Table 9.

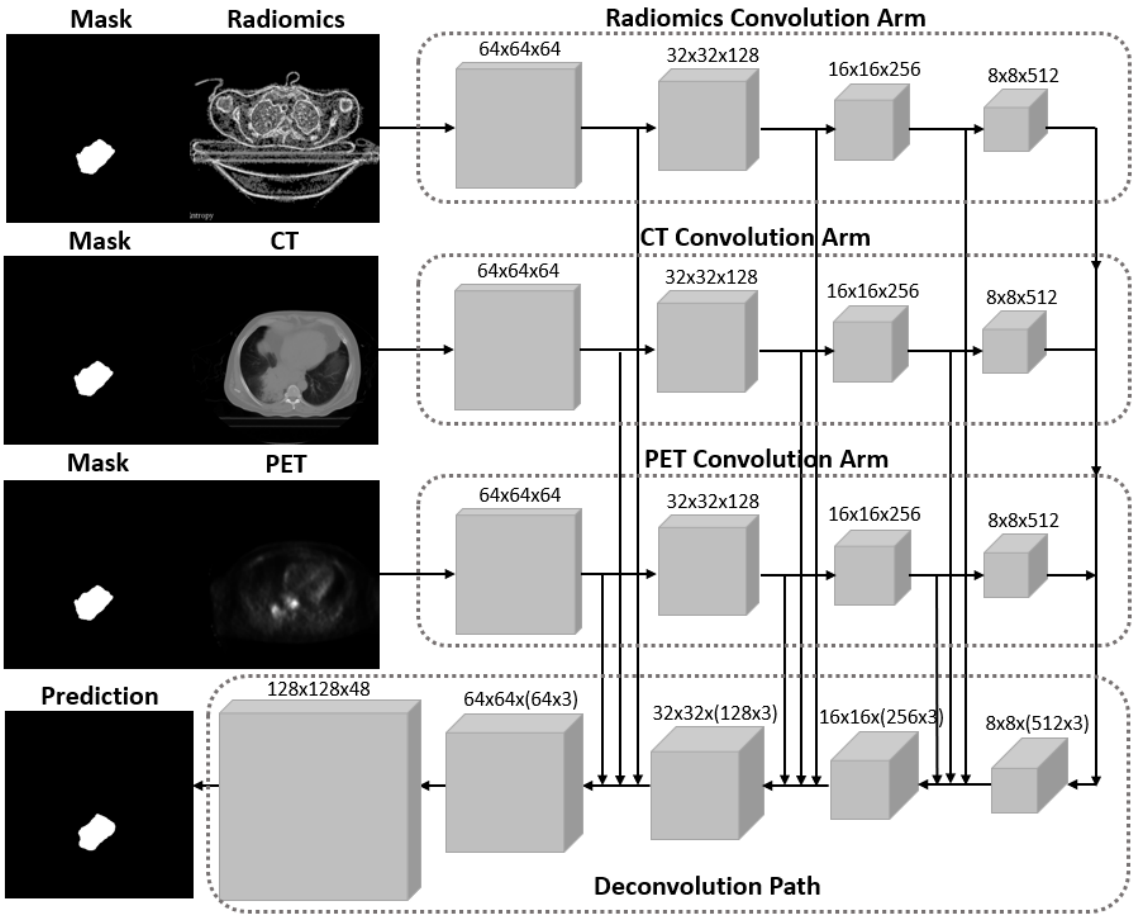


Figure 19. An illustration of our proposed multi-modality segmentation network architecture with 3 image inputs.

*Table 9. The mean Dice similarity coefficient (DSC), the Hausdorff distance (HD), and the mean bi-directional local distance (BLD) of the segmentations predicted by 3D networks using CT, PET, and radiomics feature images, trained/validated/tested with all (non-stratified) vs. stratified data subsets. The data in **bold** were from the best performing stratified subsets and were used to calculate the combined overall metric. * denotes results that are statistically significant (paired t-test, $p < 0.05$) compared with the counterpart without using radiomics features.*

Mean DSC								
			Feature Name					
Train	Val & Test	PET & CT	First Order Energy	First Order RootMeanSquared	GLDM DependenceEntropy	GLSZM ZoneEntropy	GLCM Correlation	NGTDM Coarseness
All	All	0.790 ± 0.103	0.803 ± 0.078	0.798 ± 0.145	0.796 ± 0.086	0.798 ± 0.093	0.805 ± 0.090	0.783 ± 0.145
All	Large	**0.854 ± 0.052**	**0.855 ± 0.055**	**0.839 ± 0.059**	**0.868 ± 0.042***	**0.847 ± 0.063**	**0.857 ± 0.054**	0.801 ± 0.109
Large	Large	0.837 ± 0.080	0.839 ± 0.086	0.825 ± 0.117	0.824 ± 0.091	0.833 ± 0.078	0.820 ± 0.108	**0.814 ± 0.152**
All	Small	0.758 ± 0.128	0.764 ± 0.104	0.773 ± 0.180	0.815 ± 0.073	0.749 ± 0.134	0.784 ± 0.098	0.739 ± 0.123
Small	Small	**0.821 ± 0.083**	**0.826 ± 0.062**	**0.806 ± 0.194**	**0.829 ± 0.064**	**0.820 ± 0.083**	**0.829 ± 0.066**	**0.764 ± 0.108**
Combined Overall		0.833 ± 0.063	0.836 ± 0.059	0.818 ± 0.144	0.844 ± 0.056*	0.830 ± 0.075	0.839 ± 0.062	0.782 ± 0.124
			HD (mm)					
			Feature Name					
Train	Val & Test	PET & CT	First Order Energy	First Order RootMeanSquared	GLDM DependenceEntropy	GLSZM ZoneEntropy	GLCM Correlation	NGTDM Coarseness
All	All	6.84 ± 3.22	6.97 ± 3.04	7.12 ± 2.32	6.74 ± 3.11	7.38 ± 4.01	6.65 ± 2.94	7.42 ± 5.44
All	Large	**9.53 ± 3.79**	**9.65 ± 3.81**	**10.03 ± 3.80**	**8.92 ± 2.46**	10.42 ± 3.69	**9.63 ± 4.09**	11.30 ± 5.90
Large	Large	10.30 ± 4.20	9.70 ± 4.53	10.71 ± 4.77	10.37 ± 4.35	**10.40 ± 3.60**	9.95 ± 4.58	**10.92 ± 5.84**
All	Small	4.56 ± 1.77	4.37 ± 1.60	**5.19 ± 2.43**	4.10 ± 1.53	4.65 ± 1.68	4.33 ± 1.33	**8.17 ± 6.00**
Small	Small	**4.41 ± 1.70**	**4.16 ± 1.47**	5.57 ± 2.83	**4.09 ± 1.56**	**4.12 ± 1.48**	**3.94 ± 1.29**	8.77 ± 6.92
Combined Overall		6.30 ± 2.47	6.18 ± 2.33	6.98 ± 2.94	5.87 ± 1.89	6.44 ± 2.29	6.04 ± 2.32	9.19 ± 5.94
			Mean BLD (mm)					
			Feature Name					
Train	Val & Test	PET & CT	First Order Energy	First Order RootMeanSquared	GLDM DependenceEntropy	GLSZM ZoneEntropy	GLCM Correlation	NGTDM Coarseness
All	All	3.24 ± 2.52	2.51 ± 1.96	3.17 ± 2.27	2.50 ± 1.94	2.60 ± 2.12	2.48 ± 1.99	3.58 ± 4.41
All	Large	**3.83 ± 1.79**	**3.36 ± 1.72**	**3.54 ± 1.70**	**3.10 ± 1.25**	**3.57 ± 1.85**	**3.32 ± 1.74**	**5.83 ± 4.10**
Large	Large	4.30 ± 1.50	3.90 ± 1.61	3.90 ± 1.60	3.98 ± 1.41	3.69 ± 1.21	4.09 ± 1.80	7.09 ± 4.07
All	Small	3.16 ± 1.77	3.00 ± 1.48	3.39 ± 1.68	2.65 ± 1.34	3.22 ± 1.71	3.14 ± 1.65	**3.90 ± 2.83**
Small	Small	**2.18 ± 1.07**	**1.90 ± 0.71**	**2.54 ± 1.82**	**1.98 ± 1.05**	**1.97 ± 1.11**	**1.87 ± 0.78***	4.24 ± 3.15
Combined Overall		2.80 ± 1.36	2.44 ± 1.08	2.84 ± 1.78	2.39 ± 1.13	2.56 ± 1.38	2.40 ± 1.14	4.82 ± 3.50

Analysis of Table 9 found statistically significant improvement in two results: The addition of GLDM Dependence Entropy feature images to the model improved the segmentation results from a DSC of 0.854 ± 0.052 to 0.868 ± 0.042 for large GTVs, and therefore the combined overall DSC; and GLCM Correlation improved the results from a mean BLD of 2.18 ± 1.07 to 1.87 ± 0.78 for small GTVs. Overall, these two features appeared to produce better metrics, i.e., higher DSC and lower HD and mean BLD, in other datasets as well, albeit these results were not statistically significant according to the t-test. One notable issue for using the standard t-test to measure significance in this case is the large standard deviations/variances in the HD and mean BLD values. It is possible that the assessed data violated one of the t-test assumptions, e.g., nonnormality, the existence of outliers, etc., in which case a modified t-test or a Wilcoxon rank-sum test may be more suitable tools. Further analysis should be conducted in future work when a more comprehensive feature selection process is developed.

Upon closer inspection, case studies demonstrated how the addition of the GLDM Dependence Entropy and GLCM Correlation feature images helped improving the tumor segmentations for a large and a small GTV, respectively (Figure 20-21).

Furthermore, the benefits of employing the volume-based stratification technique still applied even with the addition of radiomics features. There were a few exceptions, most noticeably NGTDM Coarseness. In fact, NGTDM Coarseness performed poorly over all metrics. This could potentially be a result of the normalization issue discussed in Section E.4.1. The example in Figure 22 shows that the area of extreme values in the feature image may have misdirected the network instead of helping it.

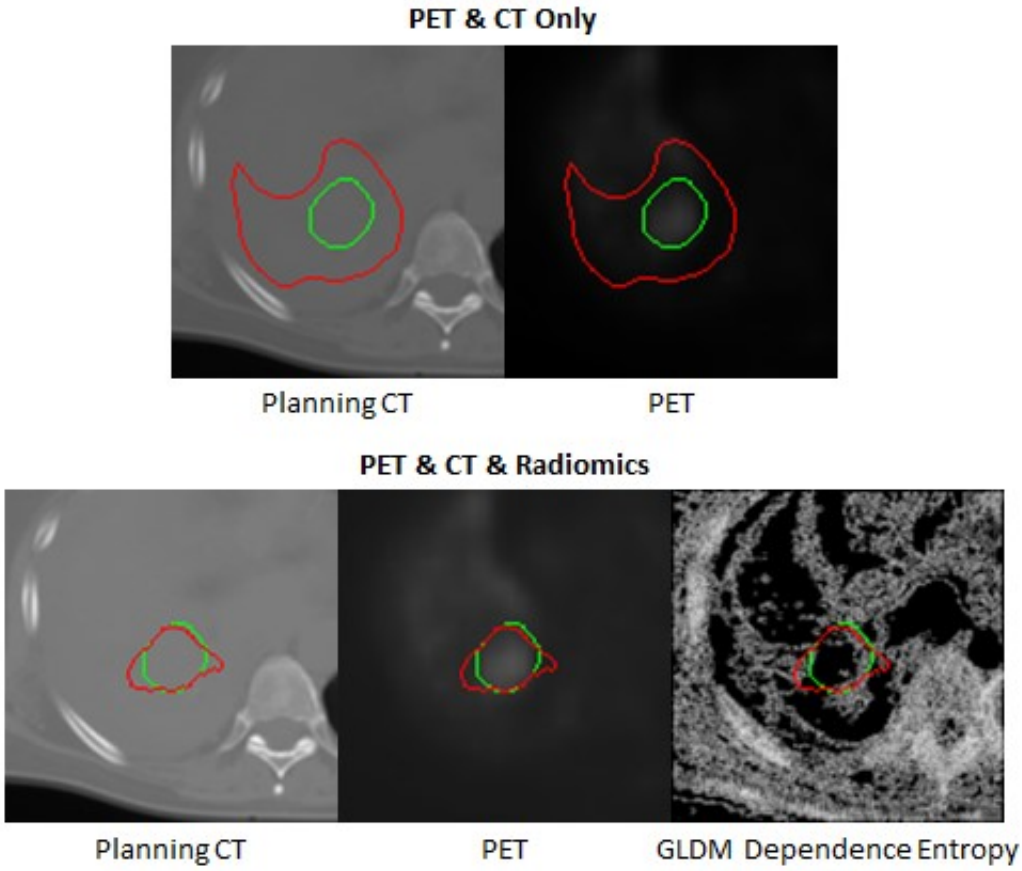


Figure 20. The automatic segmentation predicted by the model using PET and CT only (top row) was improved to the one predicted by the model using the GLDM Dependence Entropy feature images (bottom row) through the highlighting of the tissue structures that were not immediately apparent on the CT. Green contour denotes the ground truth, and red denotes the network-predicted segmentation.

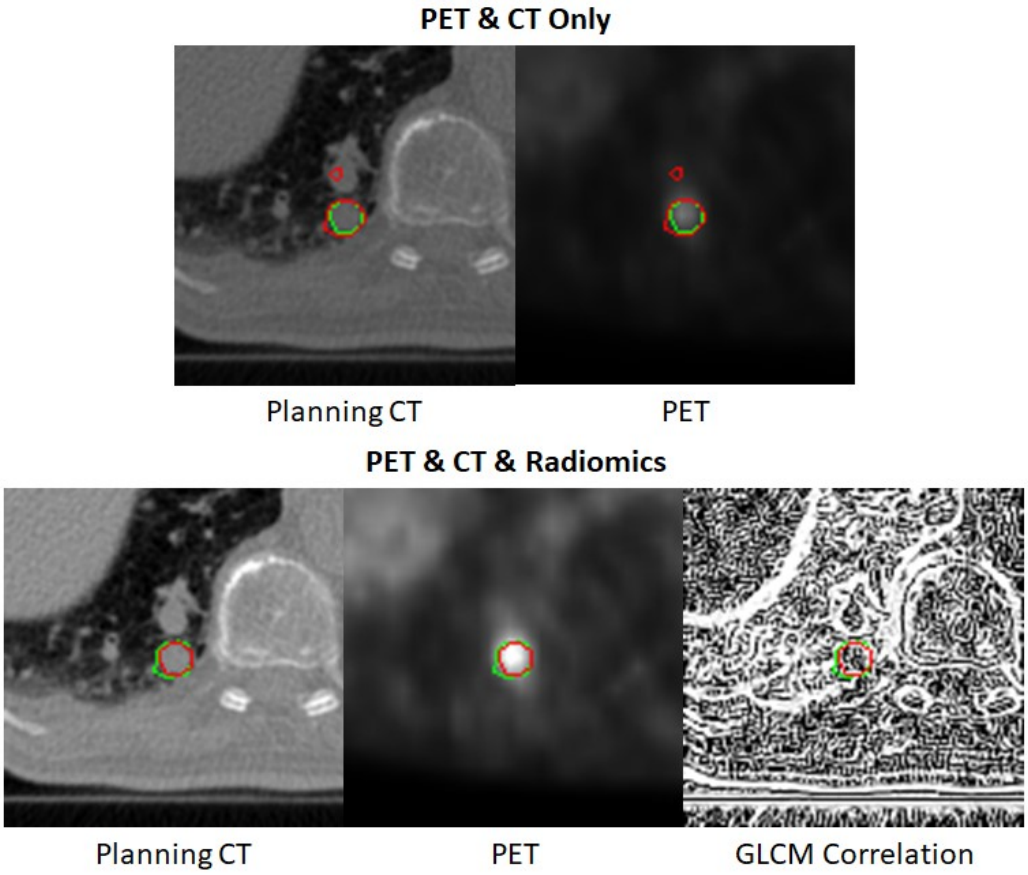


Figure 21. A case study of how GLCM Correlation (bottom row) helped with improving the segmentation by eliminating nearby normal tissue structures that was misidentified as part of the GTV in the PET & CT only (top row) segmentation. Green contour denotes the ground truth, and red denotes the network-predicted segmentation.

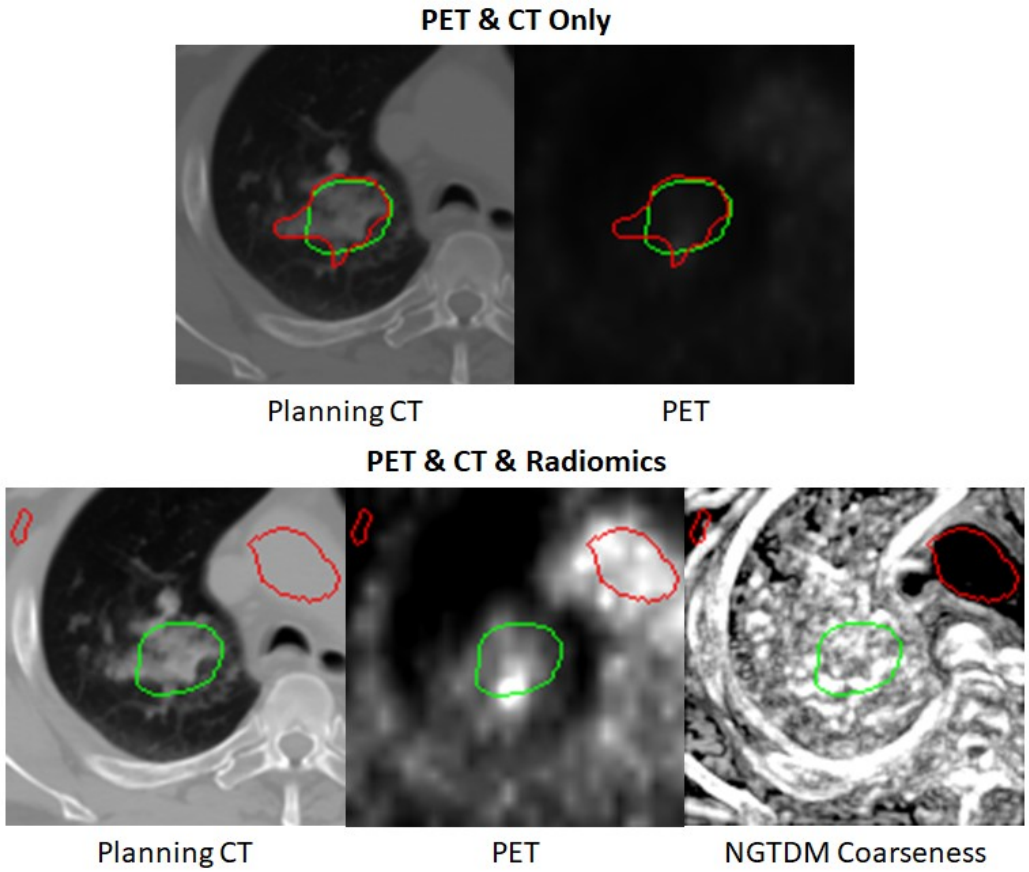


Figure 22. A case study where the introduction of NGTDM Coarseness feature images (bottom row) appeared to cause the network to misidentify normal structures as the GTV as compared with the PET and CT only model (top row). Green contour denotes the ground truth, and red denotes the network-predicted segmentation.

E.3 Conclusion and Future Work

A pipeline for the extraction, selection, and incorporation of radiomics features into the segmentation network was introduced in this section of our work. The validity of this method was tested with 104 first- and higher-order radiomics features. We were able to find two features that improved the model performance with statistical significance. However, it was also demonstrated that either choosing the wrong features or inadequately preprocessing the feature images can potentially degrade the segmentation performance. Further investigation into using radiomics feature images for lung tumor segmentation is warranted.

First, the candidate pool of radiomics features can be expanded. All features used in this work were extracted with the original Pyradiomics package. Modifications can be applied to the open-source codes to include more features introduced in subsequent studies [75], [110]. In addition, while we chose to only include the CT features in this study, PET radiomics has gained increasing traction in lung cancer-related applications and are worth investigating in future studies. Furthermore, an interesting study by Amini et al. proposed a methodology to extract the harmonized PET/CT features [111] that can potentially be applied to our work to not only incorporate PET radiomics but also utilize the information from the CT in the diagnostic PET/CT. In addition, within the functionalities of Pyradiomics, built-in filters such as the Laplacian of Gaussian (LoG) filter, the wavelet filter, the square filter, etc. can be applied prior to feature extraction as a preprocessing step. Previous studies in NSCLC metastasis prediction [112] and tumor subtype classification [113] demonstrated that filtered radiomics features made significant contributions to their results. Since voxel-based feature extraction is one of the most time-consuming processes in this work, we also propose for our future work the implementation of GPU-powered feature extraction [114], [115] toolkits.

For a more sophisticated feature selection process, numerous methods and combinations of methods have been employed in previous studies to reduce dimensionality through eliminating features of high collinearity or high correlation with each other and ranking the importance of features relative to the specific tasks [116], [117]. In addition, one of the biggest hurdles in translating radiomics studies to clinical practice is the robustness and reproducibility of the image features. As previously mentioned, factors from acquisition techniques to respiratory motion to image postprocessing can all influence the stability of radiomics features. Identifying and eliminating unstable features is the key to improve the robustness of any radiomics application. To that end, there are two main evaluation methods: 1) test-retest to assess

the temporal reproducibility of the features and 2) independent validation to address the issue of overfitting and evaluates clinical generalizability [118]. Test-retest data, i.e., images of the same subject taken under similar acquisition conditions but some period of time apart, are not routinely acquired in the clinical setting. To provide a clinically-feasible alternative, Larue et al. [119] found an effective substitute in CT images from different breathing phases in a 4DCT image set. Per our clinical protocol, 4DCT is acquired for lung tumor simulation, which provides us with a sizable “test-retest” database. Independent validation from external sources offers valuable insights into the cross-institutional generalizability of a method. A comprehensive feature selection process will be developed as part of the future work including common metrics to assess robustness and reproducibility such as the concordance correlation coefficients [120] and the intraclass correlation coefficients [121].

Finally, of the remaining features, voxel-wise feature images will be extracted and used to build models which will then be tested in a workflow that integrates the stratification technique and the radiomics feature images, as shown in Figure 23.

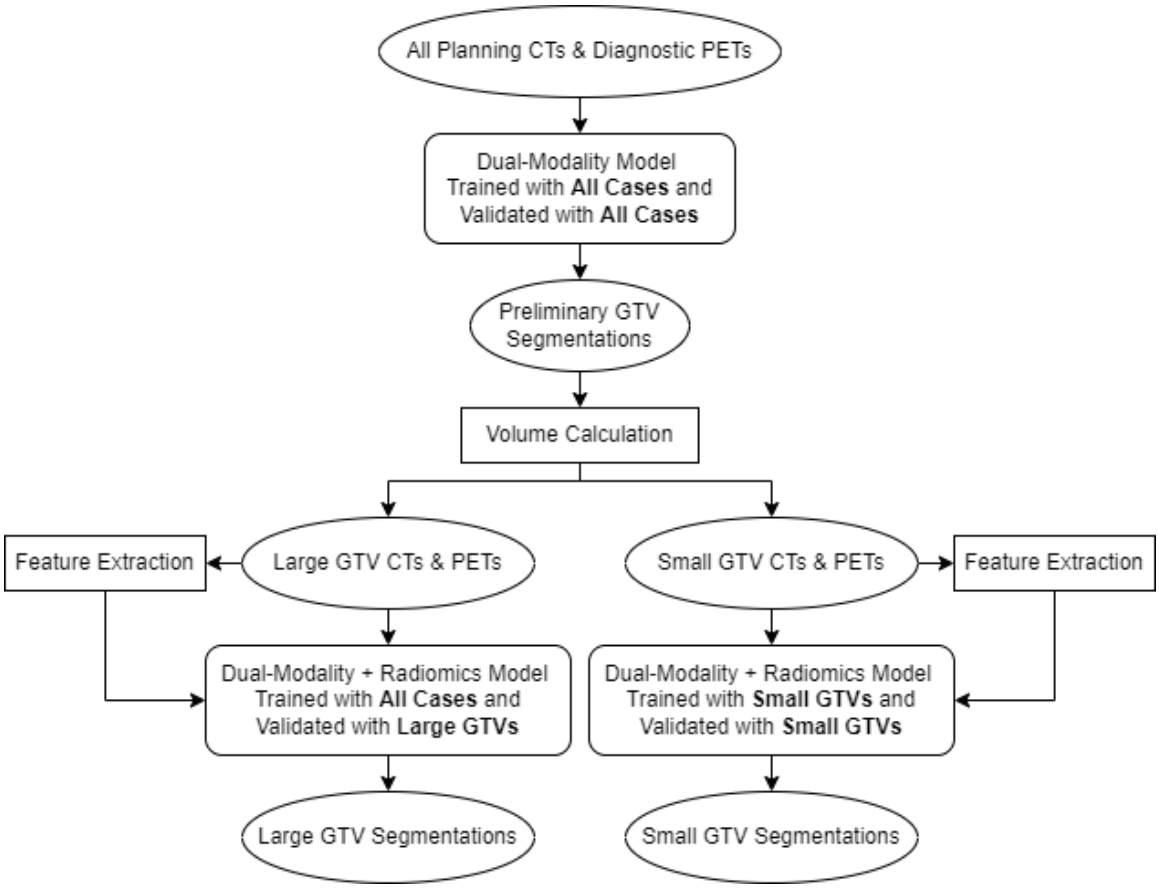


Figure 23. Workflow to incorporate the volume-based stratification technique and the radiomics feature images for lung tumor segmentation.

F. Specific Aim 3

Improve performance by adding mechanisms to account for challenging tumor features and assess the performance of our method clinically.

F.1 Volume-Based Stratification (Moved to Section D)

F.2 Clinical Evaluation

F.2.1 Introduction

In this section of our work, we recruited a group of four radiation oncologists to assess and modify the automatically-generated GTV segmentations and their manual counterparts of a 20-case subset of our database.

The clinical usability of tumor segmentations in radiotherapy planning is ultimately determined by the treating radiation oncologist. Even though we assessed the automatic segmentations using both volumetric and surface metrics, geometric measures alone may not translate to clinically meaningful endpoints [122], which is why a clinical acceptability test of the automatically-produced segmentations by radiation oncologists is essential. In practice, because of the numerous clinical duties already placed on physicians, it is often difficult to recruit participants for evaluation studies. As a result, we could not find any previous lung tumor automatic segmentation studies that implemented physician review. There have been publications on the evaluation of segmentations of other tumor sites [123], [124], organs-at-risk (OARs) [125], or the general methodologies [34] that we were able to reference. The evaluation protocol was designed to incorporate both a scoring scale and a modified Turing Test.

F.2.2 Implementation and Results

Four radiation oncologists who did not produce the original GTV segmentations independently assessed the clinical usability of the segmentations predicted by the models. The automatic segmentations used here were produced without radiomics but with stratification. 20 cases in total (10 from each size group) were randomly chosen. To avoid evaluation bias, both the automatic segmentations and their corresponding manual ground truth segmentations were anonymized and shuffled to form the 40-case database for review. The physicians were asked to categorize each segmentation as Accepted, Accepted with Modifications, or Rejected. In the latter two scenarios they were also asked to modify or recontour the GTV.

The evaluation guidelines sent to the physicians including the evaluation criteria and step-by-step instructions are attached in Appendix B.1, and the unprocessed evaluation results including reviewer comments in Appendix B.2.

On average, 91.25% of the manual vs. 88.75% of the automatic segmentations were Accepted or Accepted with Modifications, with a higher percentage of the manual segmentations being accepted as is (50% Accepted, 41.25% Accepted with Modifications and 6.25% Rejected for manual vs. 18.75%, 70%, and 8.75% for automatic). The distribution of the scores is shown in Figure 22.

The 20 cases selected for evaluation had a mean DSC of 0.820 ± 0.067 for automated contour vs. manual ground truth. Comparatively, the mean DSCs were 0.924 ± 0.046 between the manual segmentations and their modified versions and 0.918 ± 0.040 between the automatic segmentations and their modified versions, indicating that the extent of modifications were similar and relatively minor for both the manual and automatic segmentations. In addition, the mean DSC between the modified automatic segmentations and the corresponding manual ground truth contours were 0.829 ± 0.074,

similar to the results for automated versus ground truth manual contours for the evaluated cases.

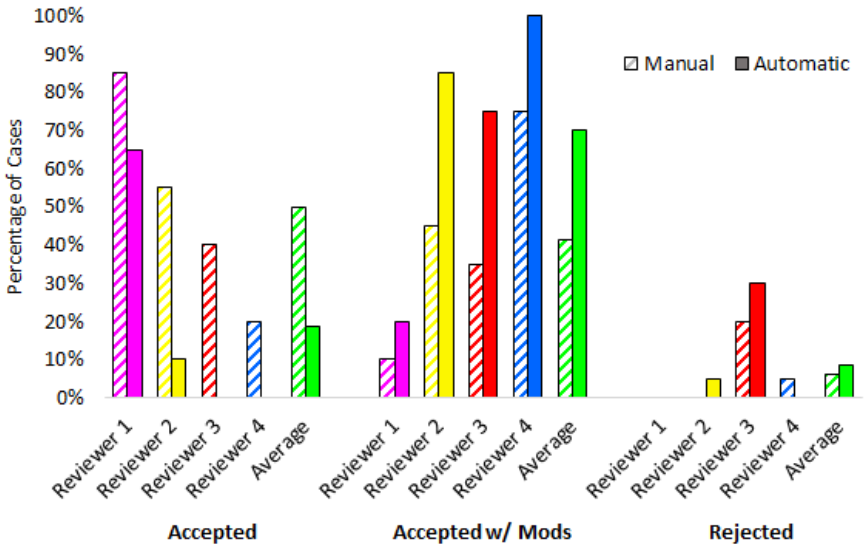


Figure 24. Percentage of cases (out of 20 Manual vs. Automatic segmentations): Accepted, Accepted with Modifications, and Rejected according to reviews by four radiation oncologists (distinguished by colors). The individual results were shown as well as the overall average in each category.

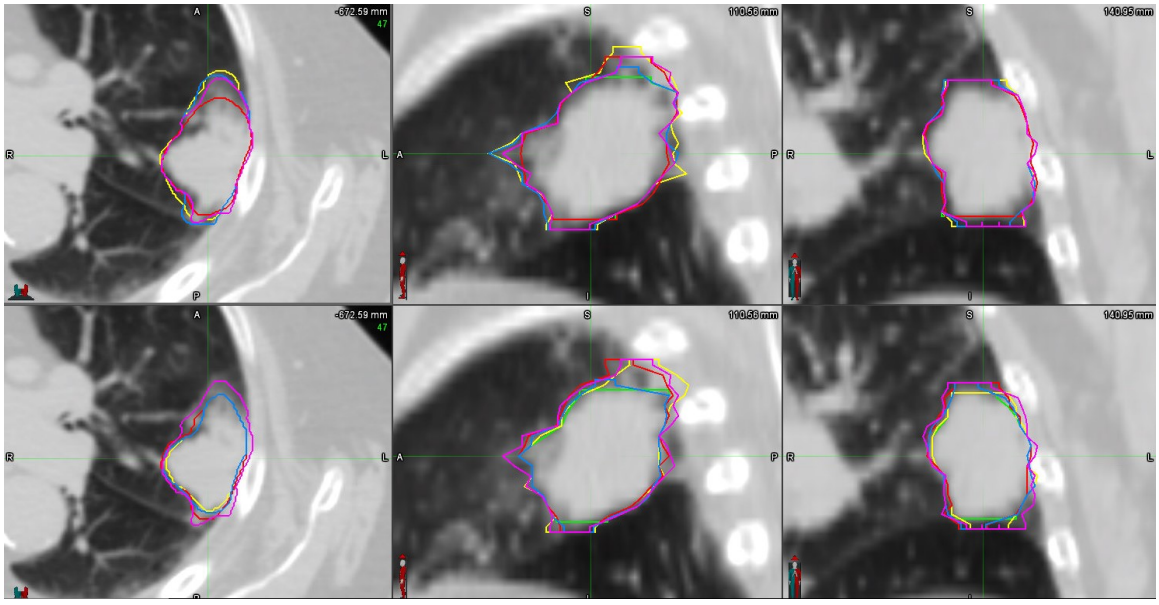


Figure 25. An example of the modifications on the manual segmentation (top row) and the automatic segmentation (bottom row). The axial, sagittal, and coronal views were displayed in the left, middle, and right columns, respectively. The manual or automatic contours to be reviewed are in green, and the contours of the same colors in the top and bottom rows were modifications made by the same reviewers.

F.2.3 Discussion

Independent clinical evaluation by four radiation oncologists showed that the vast majority of the segmentations produced through our method can be accepted for clinical use with minor modifications. While the difference in acceptance-as-is rates may indicate that the observers were able to identify that certain segmentations were of the machine learning origin, quantitatively speaking, the observers did not modify the automatic segmentations more than the manual ones. Furthermore, when modifications were made on the automatic segmentations without reference to the ground truth, the observer variations between the modified segmentations and the ground truth were comparable to the variations between the automatic segmentations and the ground truth. This clinical evaluation suggested that our method was able to produce clinically useful automatic lung cancer segmentations with high consistency.

The disparity between ratings from different reviewers was somewhat expected, as there was room for subjective interpretation in a scoring scale system. Furthermore, inter-observer variability in lung GTV segmentation is well observed and one of the main motivators of this study. It was, however, still interesting to see it clearly demonstrated in some of the more difficult cases (Figure 24) and at the superior and inferiors ends of the GTVs (Figure 25). Both the manual and automatic segmentations of the case shown in Figure 24 was rejected by two out of four reviewers. In addition, with contours by the same reviewer displayed in the same color, intra-observer variations can also be observed in the modifications made to the manual and automatic segmentations of the same case.

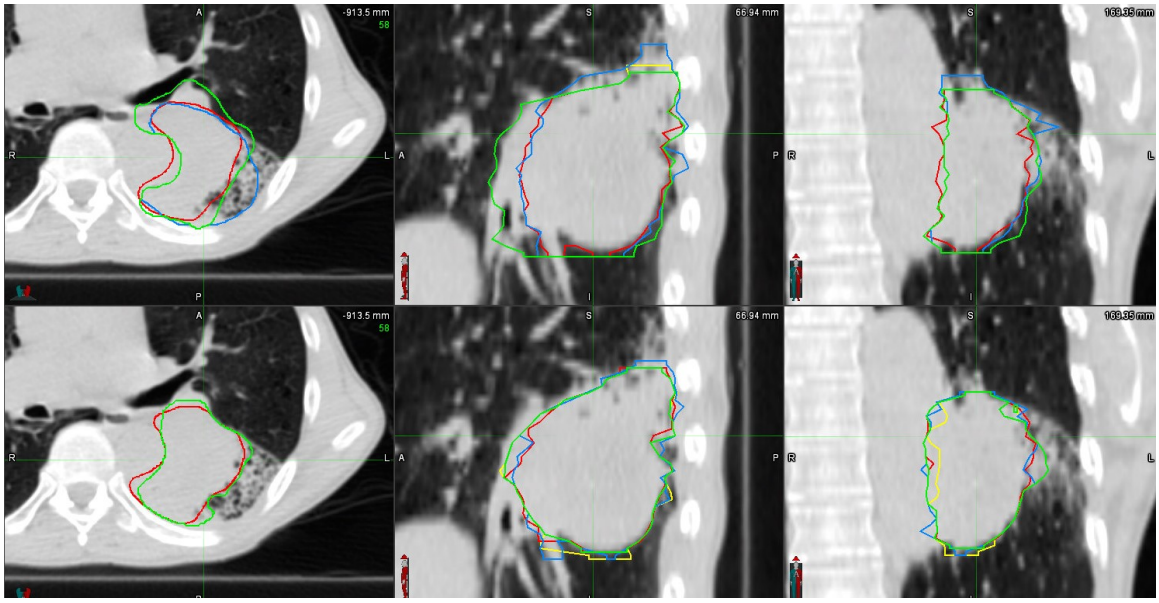


Figure 26. One of the cases that received the worst ratings. The reviewed segmentations (top: manual, bottom, automatic) were shown in green, and the modified segmentations were shown in other colors.

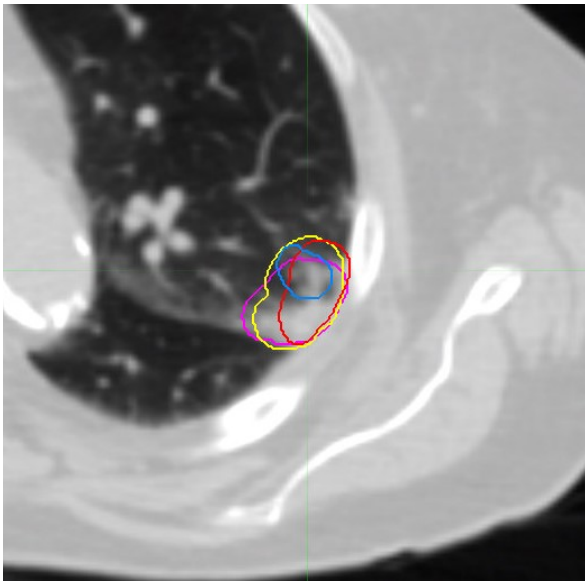


Figure 27. Variations in the modified segmentations by different physician reviewers at the superior end of a GTV.

F.3 External Validation

F.3.1 External Database

The external validation dataset is a 20-case subgroup of the NSCLC Radiogenomics data collection [126]–[128] on The Cancer Imaging Archive (TCIA) [129]. TCIA is a National Cancer Institute-funded public archive that hosts de-identified medical images of cancer and offers free access of most of its collections to the public.

The NSCLC Radiogenomics data collection is comprised of the diagnostic CTs, PET/CTs, tumor segmentations, and other radiogenomic data of a cohort of 211 patients. Out of the 144 cases with tumor segmentations, 78 were evaluated for suitability. Exclusion criteria include CT using contrast, local recurrence, image artifacts, and extended interval between CT and PET/CT. From the remaining cases, 20 were chosen with a variety of tumor sizes (1.1 – 361.8 mL with median = 20.0 mL) and stages (I: 11 cases, II: 4 cases, III: 5 cases, IV: 0 cases).

The preprocessing steps were similar to that of the internal dataset with the exception of tumor segmentation-to-mask conversion. The segmentation file format was incompatible with our original method, so 3D Slicer (https://www.slicer.org) [130] was employed instead and the segmentations were exported as masks manually.

F.3.2 Implementation and Results

The models for all cases and the stratified models with the best performance on our database, i.e., the model trained with all cases and validated with large GTVs, and the model trained and validated with small GTVs only, were tested on the external dataset. The models using only PET and CT as well as the ones with the additional feature image inputs for all six selected features were also tested. All results are listed in Table 10.

*Table 10. The mean Dice similarity coefficient (DSC), the Hausdorff distance (HD), and the mean bi-directional local distance (BLD) of the segmentations predicted by unstratified and stratified models with and without radiomics. The data in **bold** are the best overall (unstratified vs. stratified combined) result as evaluated by each metric.*

Mean DSC								
			Feature Name					
Train	Val & Test	PET & CT	First Order Energy	First Order RootMeanSquared	GLDM DependenceEntropy	GLSZM ZoneEntropy	GLCM Correlation	NGTDM Coarseness
All	All	**0.707 ± 0.083**	0.659 ± 0.146	0.678 ± 0.130	0.651 ± 0.156	0.669 ± 0.137	0.686 ± 0.113	N/A
All	Large	0.601 ± 0.193	0.579 ± 0.218	0.570 ± 0.222	0.597 ± 0.212	0.562 ± 0.223	0.667 ± 0.187	N/A
Small	Small	0.733 ± 0.115	0.744 ± 0.151	0.751 ± 0.146	0.738 ± 0.171	0.750 ± 0.157	0.738 ± 0.164	N/A
Combined Overall		0.667 ± 0.154	0.662 ± 0.185	0.661 ± 0.184	0.667 ± 0.192	0.656 ± 0.190	0.702 ± 0.175	N/A
HD (mm)								
			Feature Name					
Train	Val & Test	PET & CT	First Order Energy	First Order RootMeanSquared	GLDM DependenceEntropy	GLSZM ZoneEntropy	GLCM Correlation	NGTDM Coarseness
All	All	13.08 ± 5.90	12.76 ± 7.32	12.69 ± 6.76	12.5 ± 7.15	13.91 ± 15.05	12.27 ± 6.21	N/A
All	Large	19.54 ± 8.73	19.06 ± 8.92	20.96 ± 9.30	18.83 ± 7.86	21.39 ± 9.05	17.83 ± 6.70	N/A
Small	Small	7.16 ± 2.30	6.66 ± 1.97	6.48 ± 2.14	6.36 ± 2.44	6.32 ± 2.19	6.70 ± 2.07	N/A
Combined Overall		13.35 ± 5.52	12.86 ± 5.45	13.72 ± 5.72	12.6 ± 5.15	13.86 ± 5.62	**12.27 ± 4.39**	N/A
Mean BLD (mm)								
			Feature Name					
Train	Val & Test	PET & CT	First Order Energy	First Order RootMeanSquared	GLDM DependenceEntropy	GLSZM ZoneEntropy	GLCM Correlation	NGTDM Coarseness
All	All	**4.17 ± 3.20**	5.52 ± 4.80	5.61 ± 4.60	5.97 ± 4.74	6.72 ± 5.06	6.95 ± 5.36	N/A
All	Large	6.48 ± 3.59	9.54 ± 7.61	10.09 ± 3.79	7.14 ± 3.07	6.87 ± 3.23	6.82 ± 3.10	N/A
Small	Small	2.26 ± 1.01	2.94 ± 1.59	2.80 ± 1.43	3.02 ± 1.63	2.93 ± 1.53	2.96 ± 1.65	N/A
Combined Overall		4.37 ± 2.30	6.24 ± 4.60	6.45 ± 2.61	5.08 ± 2.35	4.90 ± 2.38	4.89 ± 2.37	N/A

F.3.3 Discussion

When tested on the external dataset, most of our models, with the exception of the ones using the NGTDM Coarseness feature, produced GTV segmentations that had reasonable appearances as judged by the author. However, comparison with the ground truth segmentations provided by the TCIA database yielded significantly worse metrics than our own internal database. Closer inspections revealed that there may be fundamental issues with some of the TCIA dataset segmentations (Figure 27). The TCIA dataset was created to facilitate radiogenomic and prognostic biomarker research, not to enable automatic lung tumor segmentation for the purpose of radiotherapy planning. There are a few important differences from our dataset:

1) The tumor segmentations in the TCIA dataset were automatically generated using unspecified in-house algorithm for the computation of 3D quantitative image features. Although the segmentations were reviewed by two radiologists, the criteria, albeit unspecified, are most likely not the same as those necessary for radiation treatment planning. To provide a preliminary assessment of the TCIA segmentations, our radiation oncologist reviewed two cases and made manual segmentations according to the clinical protocol for radiotherapy planning (Figure 28). Quantitative analysis of the two cases in Figure 28 (a) and 28 (b) showed that when compared to the physician segmentations, the automatic segmentations had DSCs of 0.874 and 0.860, respectively, but only 0.750 and 0.731 when compared to the TCIA segmentations.

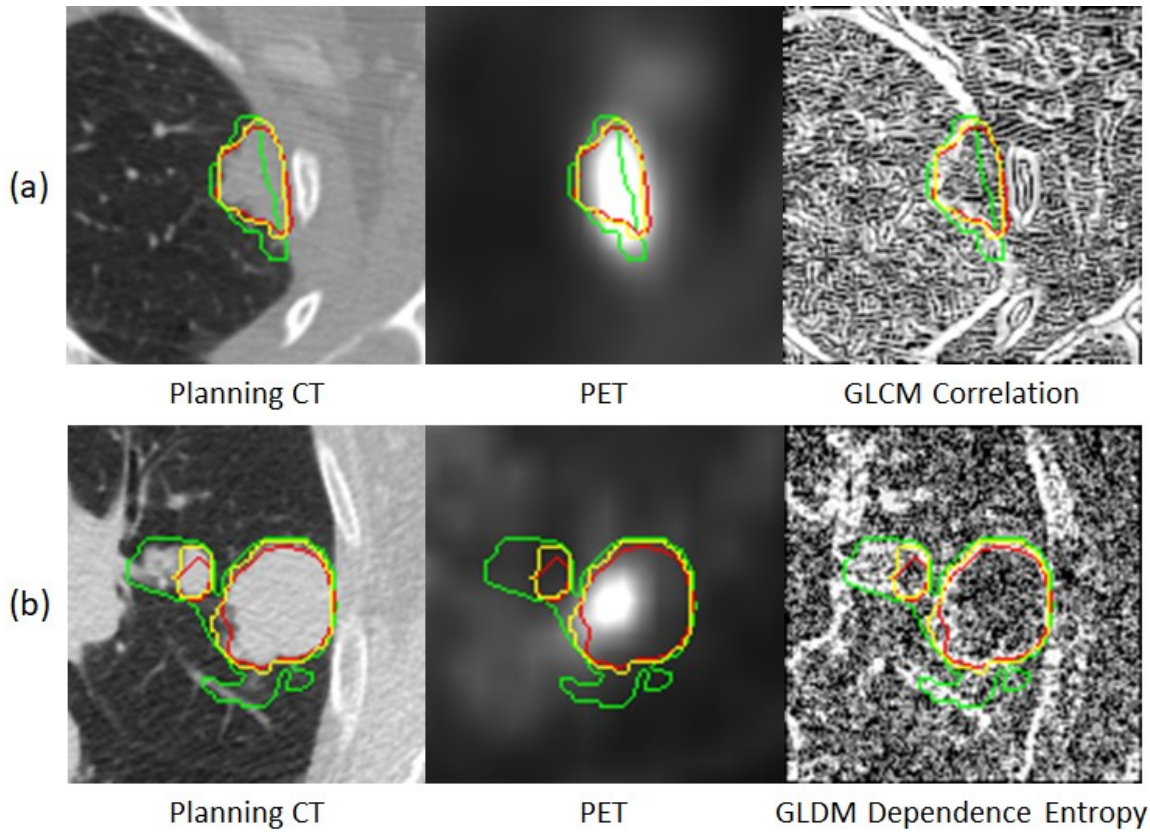


Figure 28. Two case studies comparing the TCIA segmentation (green), the manual segmentation by a radiation oncologist (yellow), and the automatic segmentation generated by our model (red). The automatic segmentations in cases (a) and (b) were predicted by models using GLCM Correlation and GLDM Dependence Entropy feature images, respectively.

2) All CTs were diagnostic CTs instead of simulation CTs and acquired on different scanners with different scanning protocols and parameters. The scanning conditions were similarly uncontrolled for the PET/CTs. External validation is an essential test of the generalizability of any prediction model. The more similar the external dataset is to the internal or development dataset, the more external validation assesses reproducibility. On the other end of the spectrum, dissimilarities between the external and the internal datasets assess the transportability of the model [131]. All these differences between the external dataset and the internal/development dataset most likely impacted the validation results negatively and measured the transportability more than the reproducibility against ground truths that were questionable.

While these caveats limited the validity of the quantitative assessment results, it is still encouraging to see that our models were able to produce reasonable segmentations with an external database of PET and CT images, especially in the two case studies where the automatic segmentations were shown to be more similar to the physician segmentations than those provided by the TCIA dataset. To our knowledge, there are no other publicly available databases with dual-modality PET and CT images of lung cancer that also provide tumor segmentations, not to mention ones produced for radiotherapy planning purposes. Therefore, in future work, there are two options that can be explored for a more reliable external validation, both requiring physician inputs: 1) direct clinical evaluation of the segmentations produced by our models, or 2) physician review and modification of the ground truth segmentations from the TCIA database.

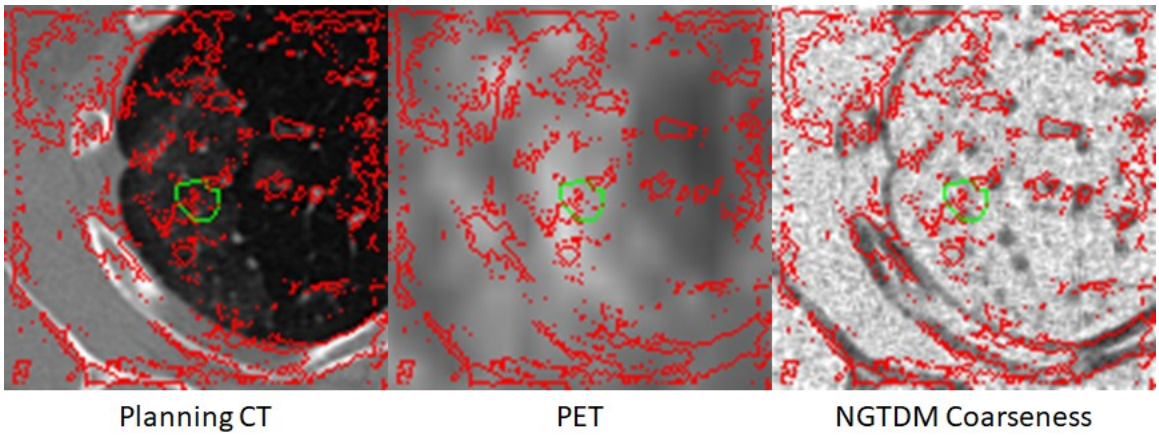


Figure 29. An example of a bad segmentation by the model using the NGTDM Coarseness feature images. Green contour denotes the ground truth, and red denotes the network-predicted segmentation.

F.4 Case Studies of Problematic Tumor/Tissue Interfaces

As discussed in the introduction, inter-observer variability in manual segmentation varies depending on the tumor location and the type of tumor/surrounding tissue interface. To assess how the automatic segmentation models perform at specific interfaces between the tumors and the surrounding tissues and to explore directions for

future work, we visually examined the segmentation results with focus on cases achieving one or more poor evaluation metrics. For each case, the segmentation by the best performing model, i.e., stratified or unstratified and with or without radiomics features, was presented.

F.4.1 High-Density Lung Tissue

The first type of problematic scenario observed in poor-performing cases is when the tumor boundary is partially or fully obscured by high-density lung tissue possibly due to fibrosis or infectious changes surrounding the tumor (Figure 30). This issue is further complicated by the presence of high signal areas on PET due to inflammation rather than tumoral metabolic activity. While we demonstrated in Figure 20 that certain features such as the GLDM Dependence Entropy was able to help visualize the structures inside of the high-density area and improve the segmentation performance in some cases, in others, it still proved to be difficult to accurately differentiate tumor and the surrounding area. As shown in the two cases in Figure 30, the automatic segmentations appeared to be mainly dependent on the PET signals.

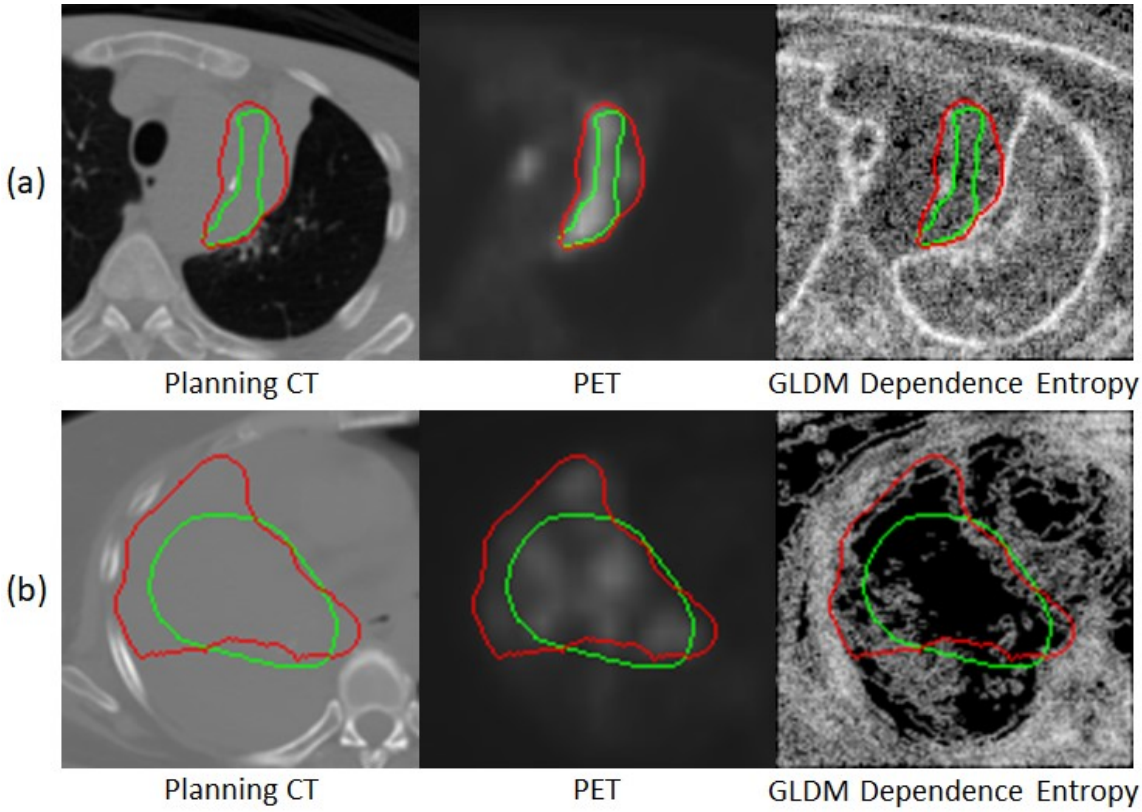


Figure 30. Two cases with high-density lung tissue surrounding the tumors. Both automatic segmentations in case (a) and (b) were predicted by models using the GLDM Dependence Entropy feature images. Green contour denotes the ground truth, and red denotes the network-predicted segmentation.

F.4.2 Lymph Nodes in Proximity

The study perimeters were established early on to include only the primary tumors, and therefore all ground truth manual segmentations were reviewed by a physician to exclude the lymph nodes. However, in several instances, the network still picked up some of the nearby lymph nodes, especially when there were pockets of high PET signal nearby. It would be interesting to expand the study perimeters to include lymph nodes in the ground truth in future work.

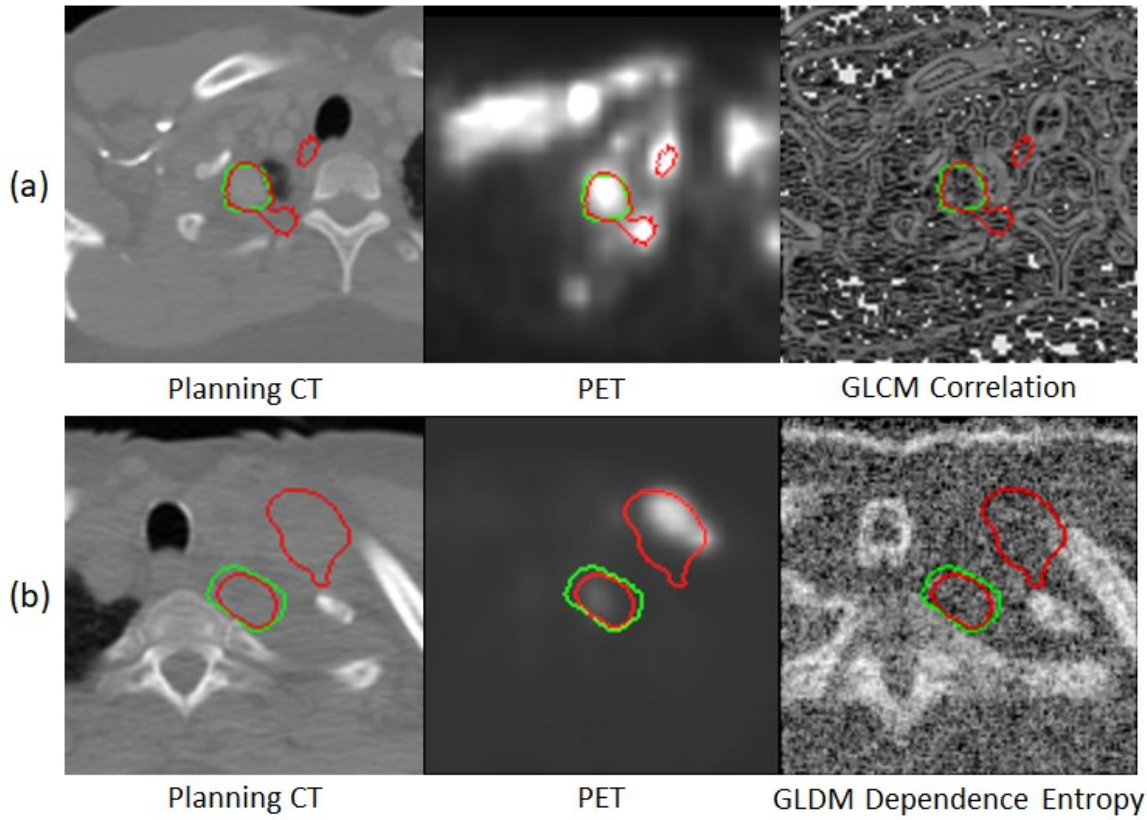


Figure 31. Two cases with lymph nodes in the proximity of the primary tumors. The automatic segmentations in case (a) and (b) were predicted by models using the GLCM Correlation and GLDM Dependence Entropy feature images, respectively. Green contour denotes the ground truth, and red denotes the network-predicted segmentation.

F.4.3 Diaphragm

Around 10% of the cases in our database had tumors adjacent to the diaphragms. In some of these cases, the automatic segmentation misidentified the diaphragm as part of the GTV, most likely as a result of diaphragm having a similar density as the tumor on CT and heightened metabolic activity and thus moderate amount of signal on PET. While this only affected the few inferior slices in large GTVs, a bigger impact on small GTVs, at least in terms of the DSC, was observed. For example, despite the significant portion of the diaphragm misidentified as the tumor, the larger GTV in Figure 32(a) still achieved an overall DSC of 0.928, while the smaller GTV in Figure 33 had a DSC of

0.764. Figure 32 also shows that some features such as the GLSZM Zone Entropy were specifically better at distinguishing the tumor and the diaphragm.

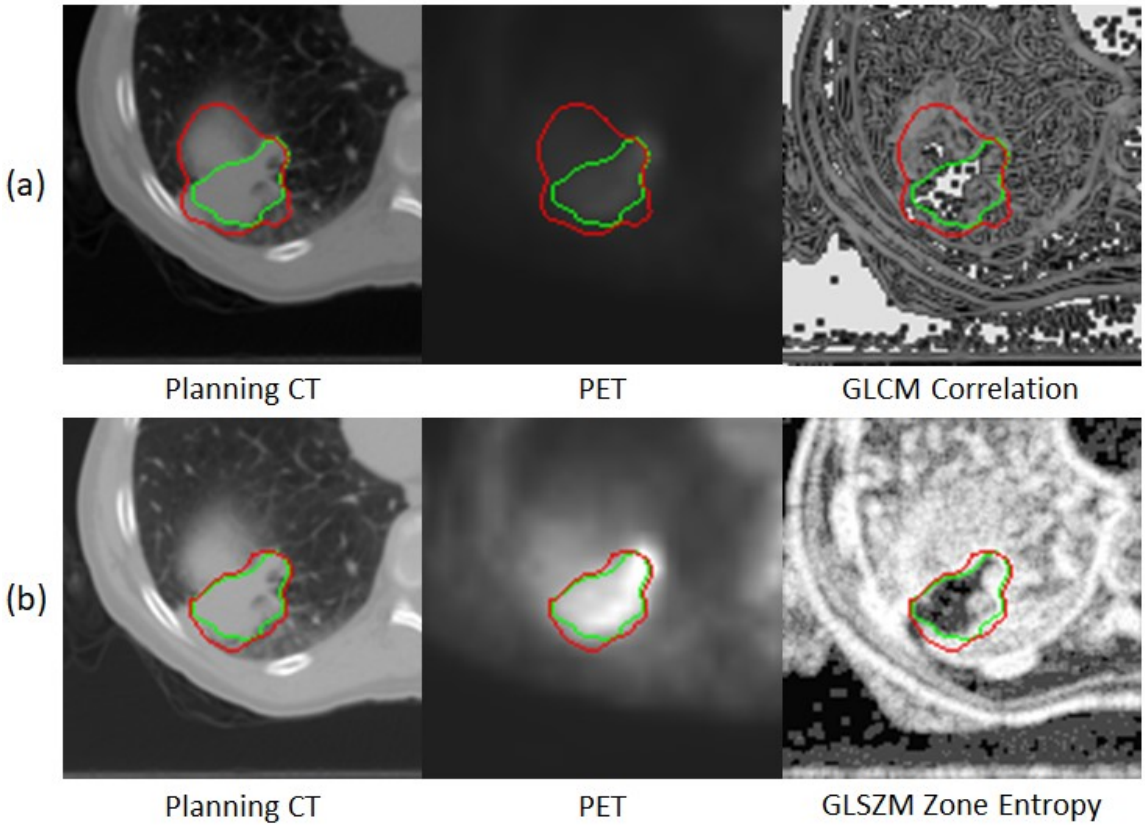


Figure 32. Automatic segmentations predicted by two different models using (a) GLCM Correlation and (b) GLSZM Zone Entropy feature images for a tumor adjacent to the diaphragm. Green contour denotes the ground truth, and red denotes the network-predicted segmentation.

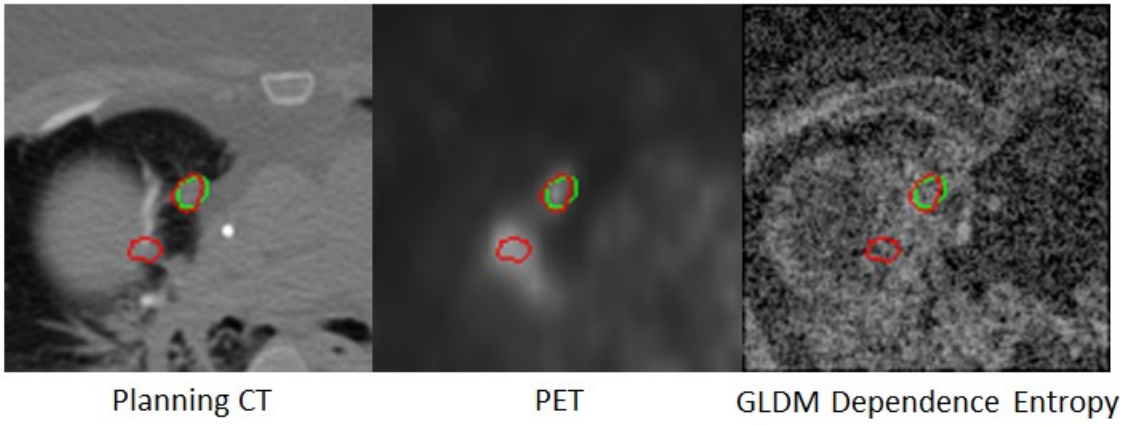


Figure 33. An example of a small tumor near the diaphragm. Green contour denotes the ground truth, and red denotes the network-predicted segmentation.

F.5 Conclusion and Future Work

In this section, clinical evaluation inspired by the Turing Test was conducted with four radiation oncologists to assess the clinical acceptability of the automatic segmentations produced by the stratified models without using radiomics, and the majority of the segmentations were accepted for clinical use with minor modifications. In addition, while the results of externally validating our models with a TCIA database were not conclusive due to potential issues with the ground truth definition, two case comparisons showed that despite using external image data that were different from our internal data in many aspects, the automatic segmentations by our models were consistent with the manual segmentations by an experienced radiation oncologist. Lastly, we were able to identify several specific tumor/normal tissue interfaces that can potentially be improved to further enhance the segmentation performance.

So far, all quantitative evaluations of the automatic segmentations were performed against the ground truth manual segmentations defined by a single radiation oncologist. We recognize that potential bias exists in this framework. In future work, with access to manual segmentations by multiple observers, an attractive alternative would be to create a consensus ground truth using what is called the STAPLE algorithm, i.e., Simultaneous Truth and Performance Level Estimation [132]. STAPLE is essentially a weighted voting algorithm that combine a collection of segmentations in an iterative process into a probabilistic estimate of the “true” segmentation, which can then be used to evaluate the automatic segmentations. In addition, since one of the main goals of an automatic segmentation method is to reduce the observer-related uncertainty, the inter-observer variability can also be assessed using the consensus segmentation. If our method is able to produce consistent tumor segmentations whose deviation from the consensus is smaller than the mean observer variation, then our method is effective in reducing

observer-related uncertainty. Furthermore, STAPLE could potentially be adapted to leverage the different automatic segmentations produced by the different models in our work to generate an optimal segmentation.

As previously discussed, a more reliable external validation could be carried out with physician inputs to modify the TCIA segmentations or with an alternative database if we can identify and access one with segmentations better suited to our purpose.

There are a few options to be explored for the interface-related issues we identified: 1) For the high-density lung tissue that is pathological but not tumoral, several radiomics features have already demonstrated their potentials to accurately identity the tumor/tissue interface. Continued research into incorporating radiomics feature images for segmentation is warranted. 2) For the nearby lymph nodes, as previously discussed, we can consider expanding the study perimeters to include lymph nodes in proximity of the primary tumor or to segment them separately. 3) For the diaphragm or other normal tissues and structures, existing automatic segmentation methods for normal organs can be applied to discourage or exclude them from consideration by the tumor segmentation network.

G. Dissertation Conclusion

The goal of this dissertation was to develop an automatic framework based on deep learning and dual-modality images for lung tumor segmentation, to investigate the feasibility of incorporating radiomics and other mechanisms to improve the segmentation performance, and to evaluate the developed methodology both quantitatively and clinically.

To facilitate the development and validation of our methodology, an expansive database of planning CTs, diagnostic PET/CTs, and manual tumor segmentations was curated, and an image registration and preprocessing pipeline was established.

The first specific aim of this work was to build and optimize a dual-modality deep learning network for lung tumor segmentation. A preliminary 2D convolutional neural network was constructed to segment the GTVs from the planning CTs and the diagnostic PETs and benchmarked against a state-of-the-art dual modality deep learning method. The segmentation performance was optimized through the conversion to 3D, the tuning of network functions and parameters, and the addition of conditional random field postprocessing. The 3D dual-modality model outperformed all other models based on the volumetric evaluation metric DSC and the surface metrics HD and mean BLD.

The second specific aim of this work was to investigate the potential of incorporating radiomics feature images to improve the segmentation performance. We introduced a workflow to extract quantitative feature values from the tumoral and peritumoral regions, to select relevant features, to extract voxel-wise feature images, and to incorporate those images as the third input in a modified segmentation network. Two radiomics features, GLDM Dependence Entropy and GLCM Correlation, were determined to enhance the segmentation performance with statistical significance.

The third and last specific aim was to further improve the method by adding mechanisms to account for challenging cases and to evaluate the developed method both quantitatively and clinically. Based on the observation that the initial segmentation models performed worse for smaller tumors, we introduced a GTV volume-based stratification technique to divide the training, validation, and testing datasets into their respective small and large GTV subsets. With or without radiomics, the models trained with the stratification strategy demonstrated improved segmentation performance, especially for the small GTVs. In addition to the quantitative evaluation metrics, a clinical

acceptability test was conducted with multiple radiation oncologists, and the majority of the automatic segmentations were found to be acceptable for clinical use with minor modifications. Furthermore, external validation was carried out using a public database. While the validation results were not conclusive, our models were able to produce automatic segmentations in case studies that were comparable with the manual segmentations by a radiation oncologist. Lastly, we identified several problematic tumor/tissue interfaces that could potentially be improved in future work.

H. References

[1] Centers for Disease Control and Prevention, “An Update on Cancer Deaths in the United States,” Atlanta, GA, 2022.

[2] “Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) SEER*Stat Database: Survival - Lung and Bronchus SEER 5-Year Relative Survival Rates (2012-2018).”

[3] N. C. Insitutue, “Non-Small Cell Lung Cancer Treatment (PDQ ®) Health Professional Version,” 2022. .

[4] International Atomic Energy Agency, “Accuracy Requirements and Uncertainties in Radiotherapy,” Vienna, 2016.

[5] C. F. Njeh, “Tumor delineation: The weakest link in the search for accuracy in radiotherapy,” *J. Med. Phys.*, vol. 33, no. 4, pp. 136–140, 2008.

[6] J. C. Stroom and B. J. M. Heijmen, “Geometrical uncertainties, radiotherapy planning margins, and the ICRU-62 report,” *Radiother. Oncol.*, vol. 64, no. 1, pp. 75–83, 2002.

[7] H. Vorwerk *et al.*, “Protection of quality and innovation in radiation oncology: The prospective multicenter trial the German Society of Radiation Oncology (DEGRO-QUIRO study): Evaluation of time, attendance of medical staff, and resources during radiotherapy with IMRT,” *Strahlentherapie und Onkol.*, vol. 190, no. 5, pp. 433–443, 2014.

[8] International Commission on Radiation Units and Measurements, “Prescribing, Recording and Reporting Photon Beam Therapy (Report 62),” 2016.

[9] C. S. Hamilton and M. A. Ebert, “Volumetric uncertainty in radiotherapy,” *Clin. Oncol.*, vol. 17, no. 6, pp. 456–464, 2005.

[10] B. Hochhegger *et al.*, “PET/CT imaging in lung cancer: indications and findings,” *J. Bras. Pneumol.*, vol. 41, no. 3, pp. 264–274, 2015.

[11] J. Bradley *et al.*, “Impact of FDG-PET on radiation therapy volume delineation in non-small-cell lung cancer,” *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 59, no. 1, pp. 78–86, 2004.

[12] C. Greco, K. Rosenzweig, G. L. Cascini, and O. Tamburrini, “Current status of PET/CT for tumour volume definition in radiotherapy treatment planning for non-small cell lung cancer (NSCLC),” *Lung Cancer*, vol. 57, no. 2, pp. 125–134, 2007.

[13] U. Nestle, S. Kremp, and A. L. Grosu, “Practical integration of [18F]-FDG-PET and PET-CT in the planning of radiotherapy for non-small cell lung cancer (NSCLC): The technical basis, ICRU-target volumes, problems, perspectives,” *Radiother. Oncol.*, vol. 81, no. 2, pp. 209–225, 2006.

[14] T. Konert *et al.*, “PET/CT imaging for target volume delineation in curative intent radiotherapy of non-small cell lung cancer: IAEA consensus report 2014,” *Radiother. Oncol.*, vol. 116, no. 1, pp. 27–34, 2015.

[15] L. J. Yin, X. Bin Yu, Y. G. Ren, G. H. Gu, T. G. Ding, and Z. Lu, “Utilization of PET-CT in target volume delineation for three-dimensional conformal radiotherapy in patients with non-small cell lung cancer and atelectasis,” *Multidiscip. Respir. Med.*, vol. 8, no. 3, pp. 1–7, 2013.

[16] A. Fiorentino *et al.*, “Positron emission tomography with computed tomography imaging (PET/CT) for the radiotherapy planning definition of the biological target volume: PART 2,” *Crit. Rev. Oncol. Hematol.*, vol. 139, no. January, pp. 117–124, 2019.

[17] G. A. Ulaner, “Fundamentals of Oncologic PET/CT,” G. A. B. T.-F. of O. P. Ulaner, Ed. Elsevier, 2019, pp. 87–98.

[18] K. J. Biehl *et al.*, “18F-FDG PET definition of gross tumor volume for radiotherapy of non-

small cell lung cancer: Is a single standardized uptake value threshold approach appropriate?," *J. Nucl. Med.*, vol. 47, no. 11, pp. 1808–1812, 2006.

[19] G. J. Weiss and R. L. Korn, "Interpretation of PET scans: Do not take SUVs at face value," *J. Thorac. Oncol.*, vol. 7, no. 12, pp. 1744–1746, 2012.

[20] Y. E. Erdi *et al.*, "The CT motion quantitation of lung lesions and its impact on PET-measured SUVs," *J. Nucl. Med.*, vol. 45, no. 8, pp. 1287–1292, 2004.

[21] S. Jelercic and M. Rajer, "The role of PET-CT in radiotherapy planning of solid tumours," *Radiol. Oncol.*, vol. 49, no. 1, pp. 1–9, 2015.

[22] D. De Ruysscher *et al.*, "Selective mediastinal node irradiation based on FDG-PET scan data in patients with non-small-cell lung cancer: A prospective clinical study," *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 62, no. 4, pp. 988–994, 2005.

[23] E. D. Soykut *et al.*, "The use of PET/CT in radiotherapy planning: Contribution of deformable registration," *Front. Oncol.*, vol. 3 APR, no. April, pp. 1–4, 2013.

[24] D. Fortin, P. S. Basran, T. Berrang, D. Peterson, and E. S. Wai, "Deformable versus rigid registration of PET/CT images for radiation treatment planning of head and neck and lung cancer patients: A retrospective dosimetric comparison," *Radiat. Oncol.*, vol. 9, no. 1, pp. 1–7, 2014.

[25] R. J. H. M. Steenbakkers *et al.*, "Reduction of observer variation using matched CT-PET for lung cancer delineation: A three-dimensional analysis," *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 64, no. 2, pp. 435–448, 2006.

[26] B. Segedin and P. Petric, "Uncertainties in target volume delineation in radiotherapy - Are they relevant and what can we do about them?," *Radiol. Oncol.*, vol. 50, no. 3, pp. 254–262, 2016.

[27] J. L. Fox *et al.*, "Does registration of PET and planning CT images decrease interobserver and intraobserver variation in delineating tumor volumes for non-small-cell lung cancer?," *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 62, no. 1, pp. 70–75, 2005.

[28] A. V. Louie *et al.*, "Inter-observer and intra-observer reliability for lung cancer target volume delineation in the 4D-CT era," *Radiother. Oncol.*, vol. 95, no. 2, pp. 166–171,

2010.

[29] I. S. F. Fitton *et al.*, “Impact of anatomical location on value of CT-PET co-registration for delineation of lung tumors,” *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 70, no. 5, pp. 1403–1407, 2008.

[30] S. K. Vinod, M. Min, M. G. Jameson, and L. C. Holloway, “A review of interventions to reduce inter-observer variability in volume delineation in radiation oncology,” *J. Med. Imaging Radiat. Oncol.*, vol. 60, no. 3, pp. 393–406, 2016.

[31] G. G. Hanna, A. R. Hounsell, and J. M. O’Sullivan, “Geometrical Analysis of Radiotherapy Target Volume Delineation: A Systematic Review of Reported Comparison Methods,” *Clin. Oncol.*, vol. 22, no. 7, pp. 515–525, 2010.

[32] M. G. Jameson, L. C. Holloway, P. J. Vial, S. K. Vinod, and P. E. Metcalfe, “A review of methods of analysis in contouring studies for radiation oncology,” *J. Med. Imaging Radiat. Oncol.*, vol. 54, no. 5, pp. 401–410, 2010.

[33] K. Karki *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. Biol. Phys.*, vol. 99, no. 1, pp. 80–89, 2017.

[34] M. J. Gooding *et al.*, “Comparative evaluation of autocontouring in clinical practice: A practical method using the Turing test,” *Med. Phys.*, vol. 45, no. 11, pp. 5105–5115, 2018.

[35] and D. J. M. Ulas Bagci, Jayaram K. Udupa, Neil Mendhiratta, Brent Fosterm Ziyue Xu, Jianhua Yao, Xinjian Chen, “Joint Segmentation of Anatomical and Functional Images: Applications in Quantification of Lesions from PET, PET-CT, MRI-PET, and MRI-PET-CT Images,” vol. 17, no. 5, pp. 213–223, 2015.

[36] W. Ju, D. Xiang, B. Zhang, L. Wang, I. Kopriva, and X. Chen, “Random Walk and Graph Cut for Co-Segmentation of Lung Tumor on PET-CT Images,” *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5854–5867, 2015.

[37] Z. Irace and H. Batatia, “Bayesian spatiotemporal segmentation of combined PET-CT data using a bivariate poisson mixture model,” *Eur. Signal Process. Conf.*, pp. 2095–2099,

2014.

[38] Q. Song *et al.*, “Optimal Co-segmentation of tumor in PET-CT images with context information,” *IEEE Trans. Med. Imaging*, vol. 32, no. 9, pp. 1685–1697, 2013.

[39] H. Cui *et al.*, “Primary lung tumor segmentation from PET–CT volumes with spatial–topological constraint,” *Int. J. Comput. Assist. Radiol. Surg.*, vol. 11, no. 1, pp. 19–29, 2016.

[40] Y. Kawata *et al.*, “Impact of pixel-based machine-learning techniques on automated frameworks for delineation of gross tumor volume regions for stereotactic body radiation therapy,” *Phys. Medica*, vol. 42, pp. 141–149, 2017.

[41] K. Ikushima *et al.*, “Computer-assisted framework for machine-learning-based delineation of GTV regions on datasets of planning CT and PET/CT images,” *J. Radiat. Res.*, vol. 58, no. 1, pp. 123–134, 2017.

[42] X. Li, H. Chen, X. Qi, Q. Dou, C. W. Fu, and P. A. Heng, “H-DenseUNet: Hybrid Densely Connected UNet for Liver and Tumor Segmentation from CT Volumes,” *IEEE Trans. Med. Imaging*, vol. 37, no. 12, pp. 2663–2674, 2018.

[43] U. Kamal, A. M. Rafi, R. Hoque, J. Wu, and M. K. Hasan, “Lung Cancer Tumor Region Segmentation Using Recurrent 3D-DenseUNet,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 12502 LNCS, pp. 36–47, Dec. 2018.

[44] G. Pezzano, V. Ribas Ripoll, and P. Radeva, “CoLe-CNN: Context-learning convolutional neural network with adaptive loss function for lung nodule segmentation,” *Comput. Methods Programs Biomed.*, vol. 198, p. 105792, 2021.

[45] G. Aresta *et al.*, “iW-Net: an automatic and minimalistic interactive lung nodule segmentation deep network,” *Sci. Rep.*, vol. 9, no. 1, pp. 1–9, 2019.

[46] X. Huang, W. Sun, T. L. (Bill) Tseng, C. Li, and W. Qian, “Fast and fully-automated detection and segmentation of pulmonary nodules in thoracic CT scans using deep convolutional neural networks,” *Comput. Med. Imaging Graph.*, vol. 74, pp. 25–36, 2019.

[47] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical

image segmentation," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 9351, pp. 234–241, 2015.

[48] F. Milletari, N. Navab, and S. Ahmadi, "V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation," in *2016 Fourth International Conference on 3D Vision (3DV)*, 2016, pp. 565–571.

[49] X. Zhao, L. Li, W. Lu, and S. Tan, "Tumor co-segmentation in PET/CT using multi-modality fully convolutional neural network," *Phys. Med. Biol.*, vol. 64, no. 1, p. 015011, Dec. 2018.

[50] C. Lian, S. Ruan, T. Denœux, H. Li, and P. Vera, "Joint Tumor Segmentation in PET-CT Images Using Co-Clustering and Fusion Based on Belief Functions," *IEEE Trans. Image Process.*, vol. 28, no. 2, pp. 755–766, 2019.

[51] Z. Zhong *et al.*, "3D fully convolutional networks for co-segmentation of tumors on PET-CT images," *Proc. - Int. Symp. Biomed. Imaging*, vol. 2018-April, no. Isbi, pp. 228–231, 2018.

[52] Z. Zhong *et al.*, "Simultaneous cosegmentation of tumors in PET-CT images using deep fully convolutional networks," *Med. Phys.*, vol. 46, no. 2, pp. 619–633, 2020.

[53] A. Kumar, M. Fulham, D. Feng, and J. Kim, "Co-Learning Feature Fusion Maps from PET-CT Images of Lung Cancer," *IEEE Trans. Med. Imaging*, vol. 39, no. 1, pp. 204–217, 2020.

[54] L. Bi *et al.*, "Recurrent feature fusion learning for multi-modality pet-ct tumor segmentation," *Comput. Methods Programs Biomed.*, vol. 203, p. 106043, 2021.

[55] X. Fu, L. Bi, A. Kumar, M. Fulham, and J. Kim, "Multimodal Spatial Attention Module for Targeting Multimodal PET-CT Lung Tumor Segmentation," *IEEE J. Biomed. Heal. Informatics*, pp. 1–1, 2021.

[56] L. Li, X. Zhao, W. Lu, and S. Tan, "Deep learning for variational multimodality tumor segmentation in PET/CT," *Neurocomputing*, vol. 392, pp. 277–295, 2020.

[57] X. Zhao, Y. Wu, G. Song, Z. Li, Y. Zhang, and Y. Fan, "A deep learning model integrating FCNNs and CRFs for brain tumor segmentation," *Med. Image Anal.*, vol. 43, pp. 98–111, 2018.

[58] L. Xu *et al.*, “Automated Whole-Body Bone Lesion Detection for Multiple Myeloma on 68 Ga-Pentixafor PET/CT Imaging Using Deep Learning Methods,” *Contrast Media Mol. Imaging*, vol. 2018, 2018.

[59] Z. Guo, X. Li, H. Huang, N. Quo, and Q. Li, “Medical image segmentation based on multi-modal convolutional neural network: Study on image fusion schemes,” *Proc. - Int. Symp. Biomed. Imaging*, vol. 2018-April, pp. 903–907, 2018.

[60] M. Havaei *et al.*, “Brain tumor segmentation with Deep Neural Networks,” *Med. Image Anal.*, vol. 35, pp. 18–31, 2017.

[61] B. Hou, G. Kang, N. Zhang, and C. Hu, “Robust 3d convolutional neural network with boundary correction for accurate brain tissue segmentation,” *IEEE Access*, vol. 6, pp. 75471–75481, 2018.

[62] R. J. Gillies, P. E. Kinahan, and H. Hricak, “Radiomics: Images Are More than Pictures, They Are Data,” *Radiology*, vol. 278, no. 2, pp. 563–577, 2016.

[63] P. Lambin *et al.*, “Radiomics: The bridge between medical imaging and personalized medicine,” *Nat. Rev. Clin. Oncol.*, vol. 14, no. 12, pp. 749–762, 2017.

[64] I. Fornacon-wood, C. Faivre-finn, J. P. B. O. Connor, and G. J. Price, “Lung Cancer Radiomics as a personalized medicine tool in lung cancer : Separating the hope from the hype,” *Lung Cancer*, vol. 146, no. March, pp. 197–208, 2020.

[65] J. E. van Timmeren, D. Cester, S. Tanadini-lang, H. Alkadhi, and B. Baessler, “Radiomics in medical imaging —‘ how-to ’ guide and critical reflection,” *Insights Imaging*, vol. 11, p. 91, 2020.

[66] A. Traverso, L. Wee, A. Dekker, and R. Gillies, “Repeatability and Reproducibility of Radiomic Features: A Systematic Review,” *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 102, no. 4, pp. 1143–1158, 2018.

[67] A. Ibrahim *et al.*, “Radiomics for precision medicine : Current challenges , future prospects , and the proposal of a new framework,” *Methods*, vol. 188, pp. 20–29, 2021.

[68] I. Tunali, R. J. Gillies, and M. B. Schabath, “Application of Radiomics and Arti fi cial Intelligence for Lung Cancer Precision Medicine,” *Cold Spring Harb. Perspect. Med.*, vol.

11, p. a039537, 2021.

[69] N. Beig, N. Braman, J. Ginsberg, and P. Tiwari, “Perinodular and Intranodular Radiomic Features on Lung CT Images Distinguish Adenocarcinomas from Granulomas,” *Radiology*, vol. 290, pp. 783–792, 2019.

[70] G. Liao *et al.*, “Preoperative CT-based peritumoral and tumoral radiomic features prediction for tumor spread through air spaces in clinical stage I lung adenocarcinoma,” *Lung Cancer*, vol. 163, pp. 87–95, 2022.

[71] D. Vuong *et al.*, “Radiomics Feature Activation Maps as a New Tool for Signature Interpretability ,” *Frontiers in Oncology* , vol. 10. 2020.

[72] A. Zwanenburg, M. Vallieres, M. A. Abdalah, H. J. W. L. Aerts, A. Apte, and E. Al., “The Image Biomarker Standardization Initiative : Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping,” *Radiology*, vol. 295, no. 2, 2020.

[73] G. Lee, H. Park, S. H. Bak, and H. Y. Lee, “Radiomics in Lung Cancer from Basic to Advanced : Current Status and Future Directions,” *Korean J. Radiol.*, vol. 21, no. 2, pp. 159–171, 2020.

[74] M. E. Mayerhoefer *et al.*, “Introduction to Radiomics,” *J. Nucl. Med.*, vol. 61, no. 4, pp. 488–495, 2020.

[75] R. M. Haralick and K. Shanmugam, “Textural Features for Image Classification,” *IEEE Trans. Syst. Man. Cybern.*, vol. SMC-3, no. 6, pp. 610–621, 1973.

[76] M. M. Galloway, “Texture analysis using gray level run lengths,” *Comput. Graph. Image Process.*, vol. 4, no. 2, pp. 172–179, Jun. 1975.

[77] G. Thibault *et al.*, *Texture Indexes and Gray Level Size Zone Matrix Application to Cell Nuclei Classification*. 2009.

[78] M. A. Member Ieee and R. King, “Textural features corresponding to textural properties,” 1989.

[79] C. Sun and W. G. Wee, “Neighboring gray level dependence matrix for texture classification,” *Comput. Vision, Graph. Image Process.*, vol. 23, no. 3, pp. 341–352, Sep. 1983.

[80] T. G. Smith, G. D. Lange, and W. B. Marks, “Fractal methods and results in cellular morphology — dimensions, lacunarity and multifractals,” *J. Neurosci. Methods*, vol. 69, no. 2, pp. 123–136, Nov. 1996.

[81] S. Alzubi, N. Islam, and M. Abbod, “Multiresolution analysis using wavelet, ridgelet, and curvelet transforms for medical image segmentation,” *Int. J. Biomed. Imaging*, vol. 2011, 2011.

[82] B. J. Woods *et al.*, “Malignant-lesion segmentation using 4D co-occurrence texture analysis applied to dynamic contrast-enhanced magnetic resonance breast image data,” *J. Magn. Reson. Imaging*, vol. 25, no. 3, pp. 495–501, Mar. 2007.

[83] D. Markel *et al.*, “Automatic Segmentation of Lung Carcinoma Using 3D Texture Features in 18-FDG PET/CT,” vol. 2013, 2013.

[84] L. Bundschuh, V. Prokic, M. Guckenberger, S. Tanadini-Lang, M. Essler, and R. A. Bundschuh, “A Novel Radiomics-Based Tumor Volume Segmentation Algorithm for Lung Tumors in FDG-PET/CT after 3D Motion Correction—A Technical Feasibility and Stability Study,” *Diagnostics*, vol. 12, no. 3, p. 576, Feb. 2022.

[85] J. Torrents-Barrena *et al.*, “Assessment of Radiomics and Deep Learning for the Segmentation of Fetal and Maternal Anatomy in Magnetic Resonance Imaging and Ultrasound,” *Acad. Radiol.*, vol. 28, no. 2, pp. 173–188, 2021.

[86] X. Wang *et al.*, “Deep learning combined with radiomics may optimize the prediction in differentiating high-grade lung adenocarcinomas in ground glass opacity lesions on CT scans,” *Eur. J. Radiol.*, vol. 129, no. June, p. 109150, 2020.

[87] A. Bizzego *et al.*, “Integrating deep and radiomics features in cancer bioimaging,” in *2019 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, 2019, pp. 1–8.

[88] L. Donisi *et al.*, “A combined radiomics and machine learning approach to distinguish clinically significant prostate lesions on a publicly available mri dataset,” *J. Imaging*, vol. 7, no. 10, 2021.

[89] B. C. Lowekamp, D. T. Chen, L. Ibáñez, and D. Blezek, “The design of simpleITK,” *Front.*

Neuroinform., vol. 7, no. DEC, pp. 1–14, 2013.

[90] G. Bradski, “The OpenCV library,” *Dr. Dobb’s J. Softw. Tools*, 2000.

[91] D. Mason, “SU-E-T-33: Pydicom: An Open Source DICOM Library,” *Med. Phys.*, vol. 38, no. 6Part10, pp. 3493–3493, Jun. 2011.

[92] H. Seo *et al.*, “Machine learning techniques for biomedical image segmentation: An overview of technical aspects and introduction to state-of-art applications,” *Med. Phys.*, vol. 47, no. 5, pp. 139–148, May 2020.

[93] N. Siddique, S. Paheding, C. P. Elkin, and V. Devabhaktuni, “U-net and its variants for medical image segmentation: A review of theory and applications,” *IEEE Access*, pp. 82031–82057, 2021.

[94] T. Zhou, S. Ruan, and S. Canu, “A review: Deep learning for medical image segmentation using multi-modality fusion,” *Array*, vol. 3–4, no. August, p. 100004, 2019.

[95] D. P. Huttenlocher, W. J. Rucklidge, and G. A. Klanderman, “Comparing images using the Hausdorff distance under translation,” *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 1992-June, no. 9, pp. 654–656, 1992.

[96] H. S. Kim, S. B. Park, S. S. Lo, J. I. Monroe, and J. W. Sohn, “Bidirectional local distance measure for comparing segmentations,” *Med. Phys.*, vol. 39, no. 11, pp. 6779–6790, 2012.

[97] P. Virtanen *et al.*, “SciPy 1.0: fundamental algorithms for scientific computing in Python,” *Nat. Methods*, vol. 17, no. 3, pp. 261–272, 2020.

[98] X. Liu, L. Song, S. Liu, and Y. Zhang, “A review of deep-learning-based medical image segmentation methods,” *Sustain.*, vol. 13, no. 3, pp. 1–29, 2021.

[99] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” *Indian J. Chem. - Sect. B Org. Med. Chem.*, vol. 45, no. 8, pp. 1951–1954, Dec. 2015.

[100] D. P. Kingma and J. L. Ba, “Adam: A method for stochastic optimization,” *3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc.*, pp. 1–15, 2015.

[101] M. Hardt, B. Recht, and Y. Singer, “Train faster, generalize better: Stability of stochastic gradient descent,” *Emot. Labor 21st Century Divers. Perspect. Emot. Regul. Work*, vol.

48, pp. 57–78, Sep. 2015.

[102] A. C. Wilson, R. Roelofs, M. Stern, N. Srebro, and B. Recht, “The marginal value of adaptive gradient methods in machine learning,” *Adv. Neural Inf. Process. Syst.*, vol. 2017-Decem, no. Nips, pp. 4149–4159, 2017.

[103] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, “Improving neural networks by preventing co-adaptation of feature detectors,” pp. 1–18, 2012.

[104] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler, “Efficient object localization using Convolutional Networks,” *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 07-12-June, pp. 648–656, 2015.

[105] P. Krähenbühl and V. Koltun, “Efficient inference in fully connected crfs with Gaussian edge potentials,” *Adv. Neural Inf. Process. Syst. 24 25th Annu. Conf. Neural Inf. Process. Syst. 2011, NIPS 2011*, pp. 1–9, 2011.

[106] A. Chi and N. P. Nguyen, “4D PET/CT as a strategy to reduce respiratory motion artifacts in FDG-PET/CT,” *Front. Oncol.*, vol. 4 JUL, no. August, pp. 1–4, 2014.

[107] J. J. M. Van Griethuysen *et al.*, “Computational radiomics system to decode the radiographic phenotype,” *Cancer Res.*, vol. 77, no. 21, pp. e104–e107, 2017.

[108] F. Pedregosa *et al.*, “Scikit-learn: Machine Learning in Python,” *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[109] S. Ioffe and C. Szegedy, “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift,” *32nd Int. Conf. Mach. Learn. ICML 2015*, vol. 1, pp. 448–456, Feb. 2015.

[110] M. Bevk and I. Kononenko, “A statistical approach to texture description of medical images: a preliminary study,” in *Proceedings of 15th IEEE Symposium on Computer-Based Medical Systems (CBMS 2002)*, 2002, pp. 239–244.

[111] M. Amini *et al.*, “Overall Survival Prognostic Modelling of Non-small Cell Lung Cancer Patients Using Positron Emission Tomography/Computed Tomography Harmonised Radiomics Features: The Quest for the Optimal Machine Learning Algorithm,” *Clin.*

Oncol., vol. 34, no. 2, pp. 114–127, 2022.

[112] T. P. Coroller *et al.*, “CT-based radiomic signature predicts distant metastasis in lung adenocarcinoma,” *Radiother. Oncol.*, vol. 114, no. 3, pp. 345–350, Mar. 2015.

[113] J. Liu, J. Cui, F. Liu, Y. Yuan, F. Guo, and G. Zhang, “Multi-subtype classification model for non-small cell lung cancer based on radiomics: SLS model,” *Med. Phys.*, vol. 46, no. 7, pp. 3091–3100, Jul. 2019.

[114] Y. Jiao, O. M. Ijurra, L. Zhang, D. Shen, and Q. Wang, *Curadiomics: A GPU-based radiomics feature extraction toolkit*, vol. 11991 LNCS, no. February. Springer International Publishing, 2020.

[115] L. Rundo *et al.*, “A CUDA-powered method for the feature extraction and unsupervised analysis of medical images,” *J. Supercomput.*, vol. 77, no. 8, pp. 8514–8531, 2021.

[116] G. Kothari *et al.*, “A systematic review and meta-analysis of the prognostic value of radiomics based models in non-small cell lung cancer treated with curative radiotherapy,” *Radiother. Oncol.*, vol. 155, pp. 188–203, 2021.

[117] M. Avanzo, J. Stancanello, G. Pirrone, and G. Sartor, “Radiomics and deep learning in lung cancer,” *Strahlentherapie und Onkol.*, vol. 196, no. 10, pp. 879–887, 2020.

[118] 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.*, vol. 5, no. 4, 2016.

[119] R. T. H. M. Larue *et al.*, “4DCT imaging to assess radiomics feature stability: An investigation for thoracic cancers.,” *Radiother. Oncol. J. Eur. Soc. Ther. Radiol. Oncol.*, vol. 125, no. 1, pp. 147–153, Oct. 2017.

[120] L. I.-K. Lin, “A Concordance Correlation Coefficient to Evaluate Reproducibility,” *Biometrics*, vol. 45, no. 1, pp. 255–268, May 1989.

[121] P. E. Shrout and J. L. Fleiss, “Intraclass correlations: uses in assessing rater reliability.,” *Psychol. Bull.*, vol. 86, no. 2, pp. 420–428, Mar. 1979.

[122] M. V. Sherer *et al.*, “Metrics to evaluate the performance of auto-segmentation for radiation treatment planning: A critical review,” *Radiother. Oncol.*, vol. 160, pp. 185–191, Jul. 2021.

[123] S.-L. Lu *et al.*, “Randomized multi-reader evaluation of automated detection and segmentation of brain tumors in stereotactic radiosurgery with deep neural networks,” *Neuro. Oncol.*, vol. 23, no. 9, pp. 1560–1568, Sep. 2021.

[124] M. S. Choi *et al.*, “Clinical evaluation of atlas- and deep learning-based automatic segmentation of multiple organs and clinical target volumes for breast cancer.,” *Radiother. Oncol. J. Eur. Soc. Ther. Radiol. Oncol.*, vol. 153, pp. 139–145, Dec. 2020.

[125] T. Lustberg *et al.*, “Clinical evaluation of atlas and deep learning based automatic contouring for lung cancer.,” *Radiother. Oncol. J. Eur. Soc. Ther. Radiol. Oncol.*, vol. 126, no. 2, pp. 312–317, Feb. 2018.

[126] S. Bakr *et al.*, “Data descriptor: A radiogenomic dataset of non-small cell lung cancer,” *Sci. Data*, vol. 5, pp. 1–9, 2018.

[127] S. Bakr, Shaimaa; Gevaert, Olivier; Echegaray, Sebastian; Ayers, Kelsey; Zhou, Mu; Shafiq, Majid; Zheng, Hong; Zhang, Weiruo; Leung, Ann; Kadoch, Michael; Shrager, Joseph; Quon, Andrew; Rubin, Daniel; Plevritis, Sylvia; Napel, SandyBakr, Shaimaa; Gevaert, Ol, “Data for NSCLC Radiogenomics Collection,” *The Cancer Imaging Archive*, 2017. [Online]. Available: http://doi.org/10.7937/K9/TCIA.2017.7hs46erv.

[128] O. Gevaert *et al.*, “Non-small cell lung cancer: Identifying prognostic imaging biomarkers by leveraging public gene expression microarray data - Methods and preliminary results,” *Radiology*, vol. 264, no. 2, pp. 387–396, 2012.

[129] K. Clark *et al.*, “The cancer imaging archive (TCIA): Maintaining and operating a public information repository,” *J. Digit. Imaging*, vol. 26, no. 6, pp. 1045–1057, 2013.

[130] R. Kikinis, S. Pieper, and K. Vosburgh, “3D Slicer: a platform for subject-specific image analysis, visualization, and clinical support.,” in *Intraoperative Imaging Image-Guided Therapy*, 3rd ed., 2014, pp. 277–289.

[131] E. W. Steyerberg and F. E. Harrell, “Prediction models need appropriate internal, internal–external, and external validation,” *J. Clin. Epidemiol.*, vol. 69, no. 5, pp. 245–247, Jan. 2016.

[132] S. K. Warfield, K. H. Zou, and W. M. Wells, “Simultaneous truth and performance level

estimation (STAPLE): an algorithm for the validation of image segmentation.," *IEEE Trans. Med. Imaging*, vol. 23, no. 7, pp. 903–921, Jul. 2004.

Appendix A Best- and Worst-Performing Case Examples

In the following figures, green contour denotes the manually delineated ground truth, and red denotes the network-predicted segmentation. The specifications of the 3D dual-modality model that produced the segmentation, i.e., the stratification scheme and the radiomics feature used, are also listed. For the Hausdorff distance and the mean bi-directional local distance, examples were shown separately for the large and the small GTVs, since these surface metrics were volume-sensitive. Note that the NGTDM Coarseness models were excluded from this presentation, because they performed significantly worse than all other models possibly due to inadequate normalization.

Appendix A.1 Dice Similarity Coefficient

- Best-Performing Case:

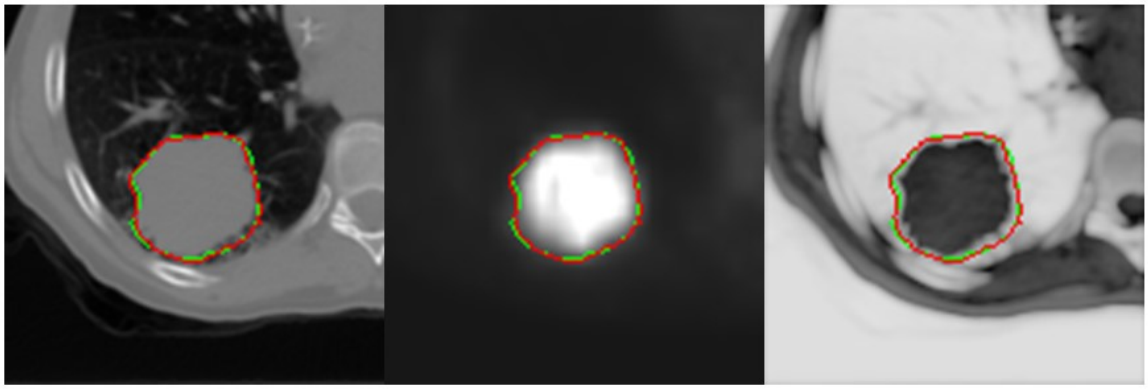
Case ID: Subject X

DSC = 0.958

Training Dataset: Large

Validation and Testing Datasets: Large

Radiomics Feature: First Order Energy



Planning CT PET First Order Energy

- Worst-Performing Case:

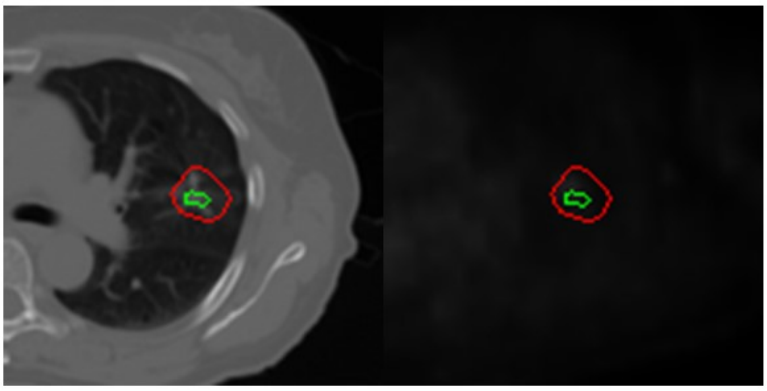
Case ID: CS009

DSC = 0.357

Training Dataset: All

Validation and Testing Datasets: All

Radiomics Feature: None



Planning CT PET

Appendix A.2 Hausdorff Distance

- Best-Performing Large GTV Case:

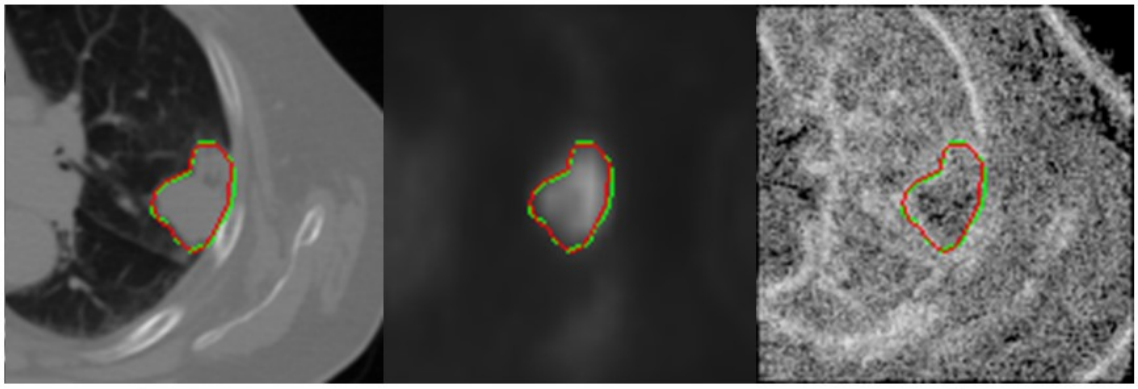
Case ID: PX051

HD = 3.31 mm

Training Dataset: All

Validation and Testing Datasets: Large

Radiomics Feature: GLDM Dependence Entropy



Planning CT PET GLDM Dependence Entropy

- Worst-Performing Large GTV Case:

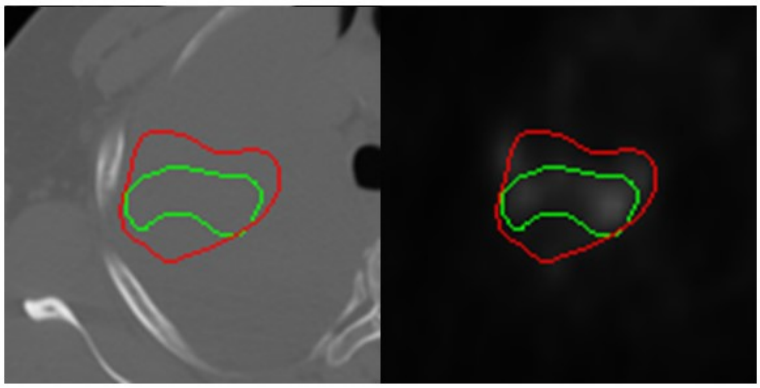
Case ID: Subject I

HD = 26.51 mm

Training Dataset: All

Validation and Testing Datasets: All

Radiomics Feature: None



Planning CT PET

- Best-Performing Small GTV Case:

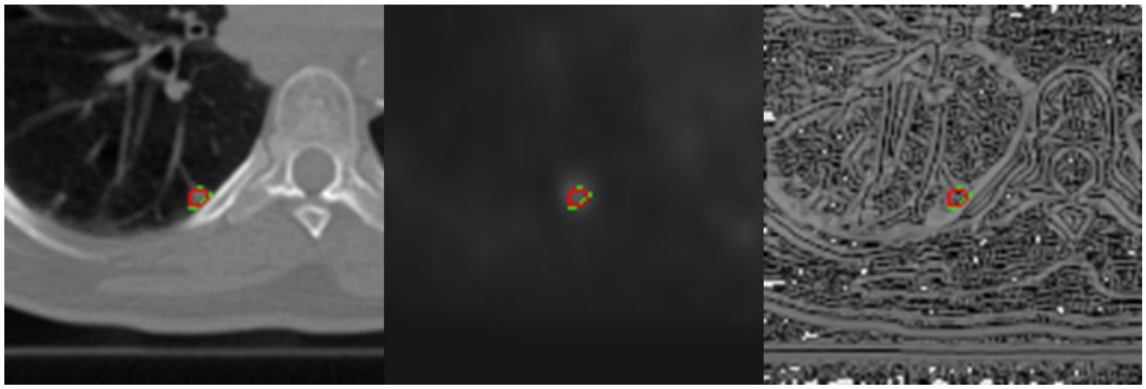
Case ID: Subject X

HD = 0.78 mm

Training Dataset: Large

Validation and Testing Datasets: Large

Radiomics Feature: First Order Energy



Planning CT PET GLCM Correlation

- Worst-Performing Small GTV Case:

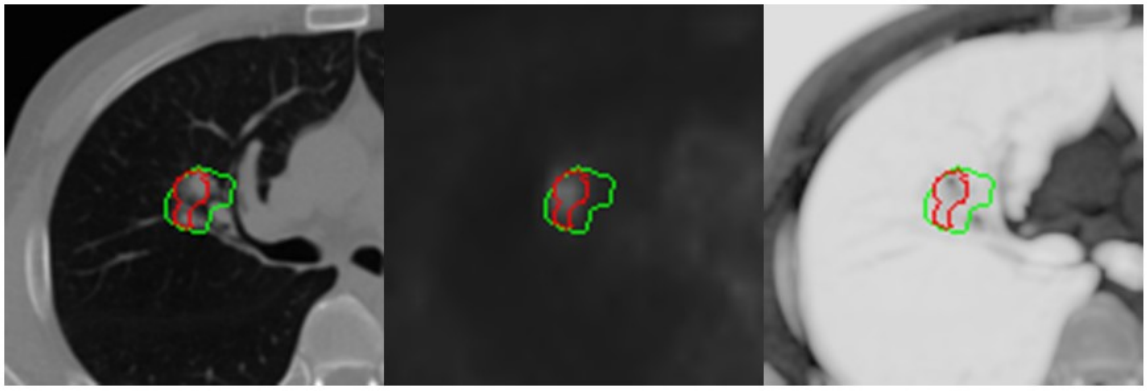
Case ID: SX0102

HD = 11.21 mm

Training Dataset: Small

Validation and Testing Datasets: Small

Radiomics Feature: First Order Root Mean Squared



Planning CT PET First Order RMS

Appendix A.3 Mean Bi-directional Local Distance

- Best-Performing Large GTV Case:

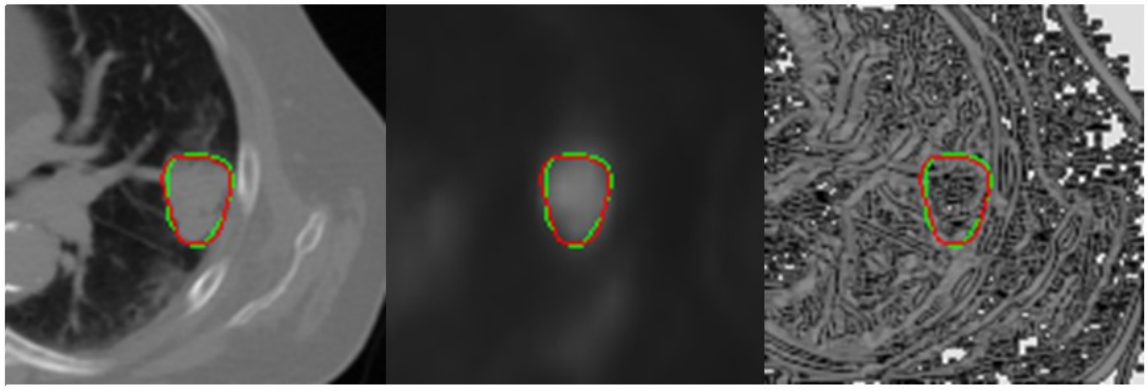
Case ID: PX051

Mean BLD = 1.35 mm

Training Dataset: All

Validation and Testing Datasets: Large

Radiomics Feature: GLCM Correlation



Planning CT PET GLCM Correlation

- Worst-Performing Large GTV Case:

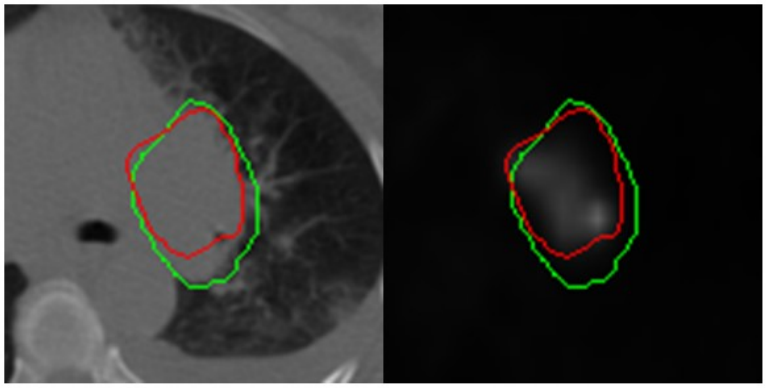
Case ID: Subject I

Mean BLD = 8.24 mm

Training Dataset: All

Validation and Testing Datasets: All

Radiomics Feature: None



Planning CT PET

- Best-Performing Small GTV Case:

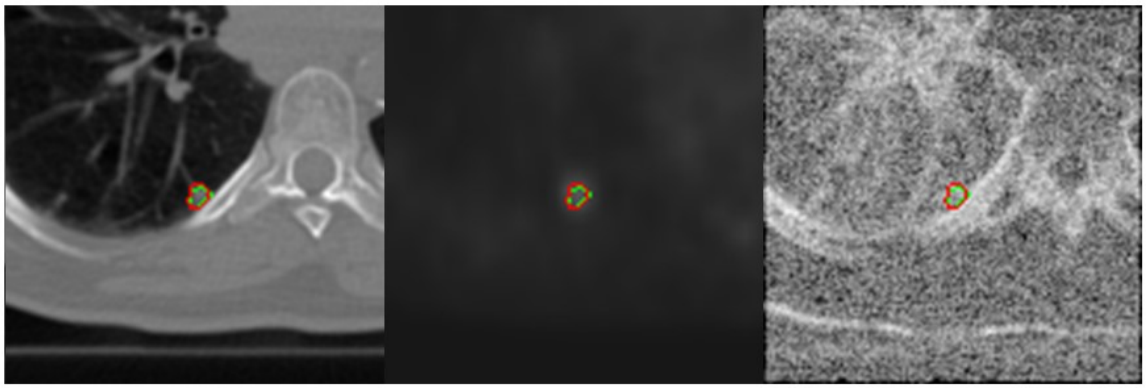
Case ID: SX0144

Mean BLD = 0.83 mm

Training Dataset: All

Validation and Testing Datasets: All

Radiomics Feature: GLDM Dependence Entropy



Planning CT PET GLDM Dependence Entropy

- Worst-Performing Small GTV Case:

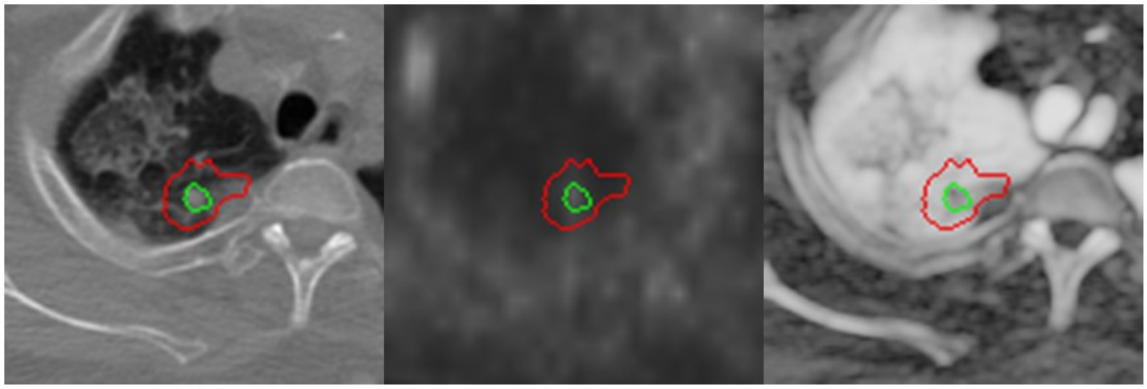
Case ID: SX0145

Mean BLD = 12.21 mm

Training Dataset: All

Validation and Testing Datasets: All

Radiomics Feature: First Order Root Mean Squared



Planning CT PET First Order RMS

Appendix B.1 Clinical Evaluation Guidelines

Project Description:

Manual lung cancer delineation for radiotherapy is time-consuming and suffers from observer variabilities that can result in suboptimal tumor control or increased risk of treatment-related side effects. This study aims to devise a robust and effective automatic lung tumor segmentation method using deep learning with the goal to minimize the interobserver contouring variability and improve the clinical efficiency.

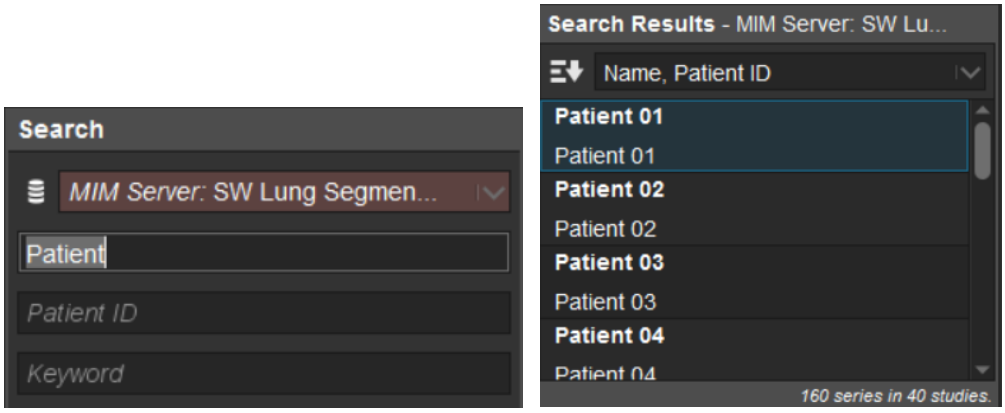
With the purpose of mimicking the current clinical workflow of lung GTV contouring, we developed a lung tumor segmentation neural network that is capable of learning simultaneously from both the PET from the diagnostic PET/CT and the simulation CT, with the manual GTV contour by an experienced radiation oncologist as the ground truth. The network produces a tumor probability map that was post-processed to obtain the GTV segmentation. In addition, we devised a tumor-volume-based training strategy that further improved the network performance. Our network was trained/validated/tested on a dataset of 290 pairs of PET and CT from the lung cancer patients treated at our clinic. The network-produced segmentations are evaluated both statistically (DICE, Hausdorff Distance, and bi-directional local distance) and clinically (by you!).

Project Guidelines:

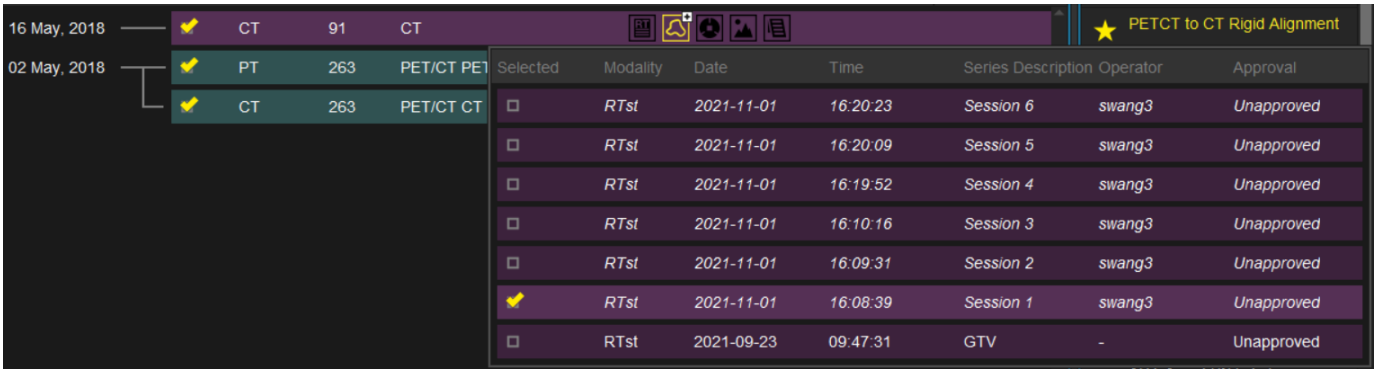
This part of the project is the clinical evaluation of 40 cases with GTV contours for lung cancer radiotherapy. The GTV could be either manually contoured by a radiation oncologist or automatically segmented through the machine learning method. All patients were anonymized and randomized to reduce bias. Each case should be rated in terms of acceptability (accepted, accepted w/ modifications, or rejected), and the evaluator can modify the contour if it is accepted with modifications, or completely redo if the contour was rejected.

MIM Server: SW Lung Segmentation

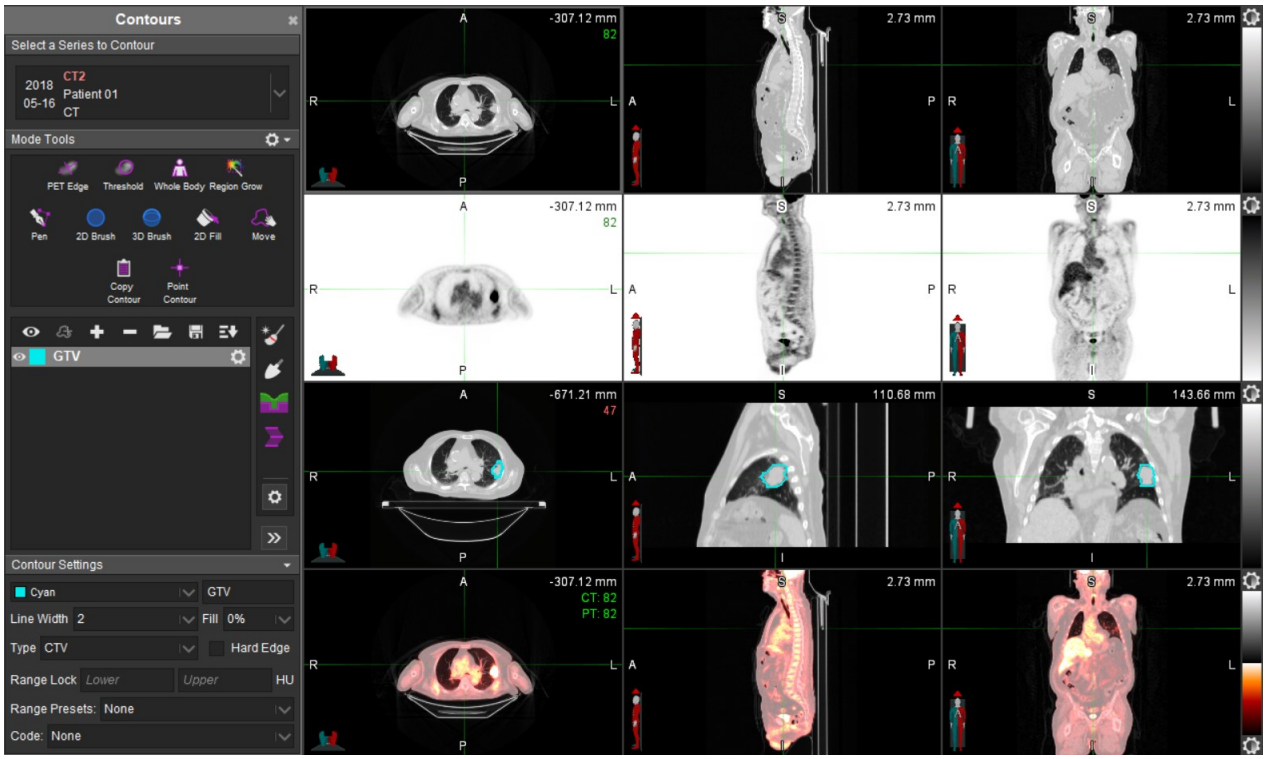
1. **Patient Name/ID:** Patient XX (01-40)



2. **Select** all modalities (CT, PET/CT PET, and PET/CT CT) and your assigned RTst **Session** X (1-6), then **Open Series**



3. **Evaluate** the GTV contour according to the clinical protocols for lung radiotherapy:
    - Window/Level: mediastinum or lung (preset at lung)
    - PET/CT for reference, etc.



4. **Record** the results:
   - Open the spreadsheet **SW Contour Evaluation Results.xlsx**
   - Select **Contour Acceptability**: Accepted, Accepted w/ modifications, or Rejected
   - If Accepted w/modifications: modify the contour
   - If Rejected: recontour
   - Check the column in the spreadsheet **Modified?**
   - Please feel free to leave comments about any cases. I would very much appreciate your input (e.g., why you rejected the contour, what your thoughts were about the modifications that needed to be made, general musings, etc.)
   - Example:

	Contour Acceptibility				
Patient ID	Accepted	Accepted w/ modifications	Rejected	**Modified?**	**Comments**
Patient 01		x		x	Meh.
Patient 02	x				Good job, contouring human or robot!
Patient 03			x		That was terrible.

5. **Save** the session

Appendix B.2 Evaluation Results

Reviewer 1:

	Contour Acceptability				
Patient ID	Accepted	Accepted w/ modifications	Rejected	**Modified?**	**Comments**
Patient 01		x		x	some part not included
Patient 02	x				
Patient 03	x				
Patient 04				x	some part not included
Patient 05	x				
Patient 06	x				
Patient 07	x				
Patient 08	x				
Patient 09	x				
Patient 10					
Patient 11	x				
Patient 12	x				
Patient 13	x				
Patient 14	x				
Patient 15	x				
Patient 16	x				
Patient 17	x				
Patient 18		x		x	Modified according to the layout on PET
Patient 19	x				
Patient 20	x				
Patient 21		x		x	
Patient 22		x		x	
Patient 23	x				
Patient 24	x				
Patient 25					
Patient 26		x		x	
Patient 27	x				
Patient 28	x				
Patient 29	x				
Patient 30	x				
Patient 31		x		x	
Patient 32	x				
Patient 33	x				
Patient 34					
Patient 35	x				
Patient 36	x				
Patient 37	x				
Patient 38	x				
Patient 39	x				
Patient 40	x				

Reviewer 2:

	Contour Acceptability				
Patient ID	Accepted	Accepted w/ modifications	Rejected	**Modified?**	**Comments**
Patient 01		x		x	pretty good. added slice inferior and several superior. Expanded GTV to cover more of the peripheral spiculations/fluffiness. Erased off of rib
Patient 02	x				no change needed. Good GTV
Patient 03		x		x	overall good. Chased the projections a little more in my GTV
Patient 04		x		x	overall good, only minor adjustments on a few slices. Added one slice superiorly
Patient 05		x		x	difficult to distinguish tumor from effusion inferiorly, but based on PET inferior portion is effusion, so no need to add to GTV - which they didn't. Added about 2 cm to last inferior slice. No other changes…overall good GTV on a tough target
Patient 06		x		x	overall good contour. Added 1 small slice superiorly. Very minor tweaks to the peripherally of GTV. Probably would have been fine without. Stylistic more than anything
Patient 07	x				Not an easy target to delineate, but with the use of PET, lung/mediastinal window, I think it's a good overall GTV. No change
Patient 08	x				Weird tumor. Little activity on PET, so not helpful. Overall, the GTV encompasses what I would consider the tumor. No changes made
Patient 09	x				No change needed
Patient 10		x		x	SCV node should be covered…likely in a separate contour. Brachiocephalic vessels were in GTV, I trimmed off of them. Ideally would definitely want PET/CT registration for this contour. Difficult to assess tumor vs. collapsed lung/atelectasis probably not a realistic target given large volume
Patient 11		x		x	Minimal changes
Patient 12		x		x	minimal changes - added 1 slice superior, more generous inferiorly because difficult to assess effusion vs. tumor
Patient 13	x				good contour
Patient 14	x				a little generous in places, but acceptable especially since such a small volume
Patient 15	x				good contour
Patient 16	x				good contour
Patient 17		x		x	overall good contour. Added some along the chest wall
Patient 18		x			overall good coverage of the primary solid tumor. Minor changes
Patient 19		x		x	added some to GTV to cover spiculations better. Erased one errant contour that went along the chest wall where there was not tumor
Patient 20		x		x	overall good. Minor changes.
Patient 21			x	x	Definitely felt like a AI contour. GTV included descending aorta which I cropped off of. Made lots of other changes. Probably should have started from scratch, but just modified
Patient 22		x		x	periphery of tumor was not covered by GTV on some slices. I extended
Patient 23		x		x	made GTV larger - sup/inf. Extended to cover more chest wall
Patient 24	x				good contour - no change
Patient 25		x		x	made general expansion of GTV. Erased errant contour that extended along peripheral vessel
Patient 26		x		x	added 1 slice sup, erased 1 slice inf. Otherwise no other significant issue
Patient 27	x				no issues. I would accept
Patient 28	x				maybe a little generous of a GTV, but difficult to determine tumor vs. atelectasis, collapse, etc. so I think it is a reasonable GTV. Would not change
Patient 29		x		x	minor changes
Patient 30		x		x	minor changes, stylistic and probably wasn't needed

Patient 31	x				no change
Patient 32		x		x	added 1 slice sup. Enlarged GTV on the periphery on a couple slices
Patient 33		x		x	very minor changes. Probably would have been acceptable as is
Patient 34		x		x	added 1 slice inf. Covered the lateral aspect 1 slice before original contour
Patient 35	x				really tight GTV around the tumor, I was tempted to extend slightly to cover the equivocal hazy area outside the solid component, but ultimately left it as is
Patient 36		x		x	minor changes. Probably could have been left alone
Patient 37		x			added 1 slice sup. Extended contour into the carina one slice earlier
Patient 38		x		x	GTV was too small. Did not cover solid component of tumor. Also extended into chest wall further
Patient 39		x		x	inferior two slices include dome of liver. Erased. Otherwise, no other changes
Patient 40		x		x	minor changes to the superior/anterior portion of the tumor.

Reviewer 3:

	Contour Acceptability				
Patient ID	Accepted	Accepted w/ modifications	Rejected	**Modified?**	**Comments**
Patient 01		x		x	Added some expansion and smoothing between slices
Patient 02	x				Good inclusion of axial inferior and superior slices with haziness (representing the lower and upper limits of disease)
Patient 03	x				Excellent distinction of tumor and vessel
Patient 04			x	x	Missing obvious tumor in portions of GTV
Patient 05	x				
Patient 06		x		x	Very tight contours with tumor slightly outside GTV on some slices
Patient 07			x	x	Way too generous in some areas, not generous enough in others
Patient 08			x	x	**Wrong lesion contoured; PET avid lesion more superior**
Patient 09	x				Perfect
Patient 10			x		Way too generous in some areas, not generous enough in others
Patient 11		x		x	Good GTV coverage, expanded margins slightly, added inferior and superior slices
Patient 12		x		x	Tough case with differences between PET and CT Sim, infection at time of PET?
Patient 13		x		x	All GTV was covered, but decreased total coverage based on PET
Patient 14	x				Good coverage
Patient 15	x				Good coverage
Patient 16		x		x	Added one superior slice
Patient 17		x			Very small modifications
Patient 18		x		x	Tough case with? progression since PET, would benefit from PET/CT sim fusion
Patient 19		x		x	Very small modifications, mostly superior
Patient 20		x		x	would benefit from PET/CT sim fusion
Patient 21			x	x	Tough case, likely infection showing on PET, contours missing some GTV
Patient 22		x		x	Good coverage, but contours too tight
Patient 23		x		x	Good coverage, but contours too tight
Patient 24		x		x	Very small modifications, mostly superior
Patient 25		x		x	Small modifications, missing superior coverage
Patient 26		x		x	Not great coverage
Patient 27			x	x	**Missing mediastinal LN coverage** Otherwise, small modifications, missing inferior slice

Patient 28		x		x	Tough case
Patient 29		x		x	Difficult to distinguish pleural fluid and tumor
Patient 30	x				Good coverage
Patient 31			x	x	Over-coverage
Patient 32			x		**Wrong lesion contoured; PET avid lesion more superior**
Patient 33		x		x	Small modifications
Patient 34		x		x	
Patient 35	x				Great coverage
Patient 36		x		x	Good coverage, but contours too tight
Patient 37		x		x	Good coverage, but contours too tight; GTV covers some esophagus at inferior
Patient 38			x	x	Poor coverage of GTV in middle of tumor
Patient 39			x	x	Starts covering into liver at inferior slice; Otherwise, small modifications due to tight contours
Patient 40		x		x	Tough case, would benefit from PET/CT sim fusion, small modifications made

Reviewer 4:

	Contour Acceptability				
Patient ID	Accepted	Accepted w/ modifications	Rejected	Modified?	Comments
Patient 01		X		X	Tweaked edges, added a slice
Patient 02	X				
Patient 03	X				
Patient 04		X			added a slice sup and inf, tweaked edges
Patient 05		X			
Patient 06		X			
Patient 07			X		Was able to adjust the initial contour, but needed a lot of changes
Patient 08	X				Contour is fine, though area contoured not PET avid, PET avid area in lateral lung.
Patient 09	X				Good
Patient 10		X		X	
Patient 11		X		X	Need a number of tweaks to edges
Patient 12		X		X	
Patient 13		X		X	
Patient 14		X		X	trimmed off some edges
Patient 15		X		X	only one or two minor tweaks
Patient 16		X		X	Good.. Just added one slice sup and inf
Patient 17		X		X	A few minor changes
Patient 18		X		X	moderate modifications
Patient 19		X		X	pretty good, minimal changes
Patient 20		X		X	moderate changes
Patient 21		X		X	moderate changes
Patient 22		X		X	minor changes
Patient 23		X		X	
Patient 24		X		X	some changes, overall pretty good
Patient 25		X		X	minor changes
Patient 26		X		X	volume felt ok, just shifted a bit off target, so tweaked
Patient 27		X		X	good, minor changes
Patient 28		X		X	hard to contour without actual fusion, would fuse for real planning
Patient 29		X		X	hard to contour without actual fusion, would fuse for real planning

Patient 30		X		X	minor changes
Patient 31		X		X	
Patient 32					*tweaked contoured lesion, though no PET-avid, small avid lesion not contoured
Patient 33		X		X	very minimal tweaks
Patient 34		X		X	minor changes
Patient 35		X		X	very minimal tweaks
Patient 36		X		X	minor changes
Patient 37		X		X	minor changes
Patient 38		X		X	minimal changes
Patient 39		X		X	took off bottom slice, added slice superiorly, tweaked a few edges
Patient 40		X		X	needed a good bit of tweaking