UNIVERSITY OF CALIFORNIA

Los Angeles

Deep Learning Approaches for Assisting MR-guided Radiation Therapy

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Physics and Biology in Medicine

by

Jie Fu

2021

ABSTRACT OF THE DISSERTATION

Deep Learning Approaches for Assisting MR-guided Radiation Therapy

by

Jie Fu

Doctor of Philosophy in Physics and Biology in Medicine

University of California, Los Angeles, 2021

Professor John H. Lewis, Co-Chair

Professor Daniel A. Low, Co-Chair

Magnetic resonance-guided radiation therapy (MRgRT) has drawn enormous clinical and research interests. The superior soft-tissue contrast of magnetic resonance imaging (MRI) compared with computed tomography (CT) allows more accurate tumor and organ-at-risk (OAR) segmentation for brain, prostate, and abdominal cancer. Additionally, real-time target tracking ability and high-quality daily MR images offered by the online MRgRT system could further minimize treatment delivery uncertainties. However, the current MRgRT workflow has several limitations including the need to acquire an additional CT for treatment planning, slow tumor and OAR recontouring in the adaptive workflow, and underdeveloped tools for predicting

treatment response and survival outcome. In this dissertation, we developed and investigated several deep learning (DL) methods to address these three limitations.

First, 2D and 3D convolutional neural networks (CNNs) were proposed to generate pelvic synthetic CT (sCT) images from 1.5T MR images. Second, conditional generative adversarial network (cGAN) and cycle-consistent generative adversarial network (cycleGAN) were investigated for abdominal sCT generation based on 0.35T MR images. Third, a novel multi-path 3D DenseNet was proposed for automatic glioblastoma multiforme (GBM) segmentation based on multi-modal MR images and compared with the corresponding single-path DenseNet. For predicting neoadjuvant chemoradiation treatment (nCRT) response in patients with locally advanced rectal cancer (LARC), two logistic regression models were built using handcrafted radiomic features and DL-based radiomic features, respectively. These radiomic features were extracted from pre-treatment diffusion-weighted MR images based on manually delineated gross tumor volume. Additionally, an automatic radiomic workflow was proposed for GBM survival prediction based on multi-modal MR images. This workflow consisted of an automatic tumor segmentation CNN and a Cox regression model.

The proposed 3D CNN generated more accurate pelvic sCT images compared with the 2D CNN. Abdominal sCT images generated by both GANs achieved accurate dose calculation for liver radiotherapy plans. The multi-path DenseNet achieved more accurate GBM segmentation compared with the single-path DenseNet. The logistic regression model constructed using DL-based features achieved significantly better classification performance in predicting nCRT response compared with the model constructed using handcrafted features. The proposed automatic workflow demonstrated the potential of improving patient stratification and survival prediction in GBM patients.

The proposed DL methods could potentially address three limitations of the MRgRT workflow but were investigated across different cancer types due to limited data availability. Future work could be adapting these methods for one cancer type and conducting further investigation to translate them into clinics.

The dissertation of Jie Fu is approved.

Dan Ruan

Nzhde Agazaryan

Ann C. Raldow

John H. Lewis, Committee Co-Chair

Daniel A. Low, Committee Co-Chair

University of California, Los Angeles

2021

To Mengjia, Mom, and Dad

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would first like to express sincere gratitude to my advisor, Dr. John H. Lewis. When I first entered UCLA, I was very excited to start a new chapter of my life but had little knowledge about medical physics. It was John who introduced me to this field and provided direction for my research projects. He also gave me enough freedom to work on projects which I am interested in. My graduate journey would not come to end without his wisdom, guidance, and countless encouragements. I feel extremely grateful to be able to work with such a brilliant and passionate person.

I would also like to thank my committee, Dr. Dan Ruan, Dr. Daniel A. Low, Dr. Nzhde Agazaryan, and Dr. Ann C. Raldow, for their support and valuable suggestions. Additionally, I would like to thank other faculty members in the Department of Radiation Oncology including Dr. X. Sharon Qi, Dr. Yingli Yang, Dr. Minsong Cao, and Dr. James Lamb. All of these people taught me valuable knowledge about medical physics and gave me indispensable advice. I would also like to thank Dr. Michael McNitt-Gray for being such a kind and supportive program director. Many thanks to the program staff, Reth and Alondra, for their administrative help and support throughout my graduate study.

I would like to thank my labmates, Kamal, Geraldine, and Minghao, for the companionship and collaboration. Thank you to Qihui, Bao, Nyasha, Elizabeth, Alejandra, Vahid, Wenbo, Xinran, Meng, and Yucheng for your company and great friendship. I will miss hikes, karaoke, parties, and all the fun times we had.

Finally, I would like to thank my father Zhongwei, mother Qihua, and sister Mengjia for all the love and support you gave me throughout my life. I forget how many video calls we had

since I left home for college and studied abroad, but every single of them brought me your care and love.

VITA

EDUCATION

M.Sc.	University of California, Los Angeles, Physics and Biology in Medicine	2019
M.Sc.	University of British Columbia, Physics	2016
B.Sc.	Shandong University, Space Science and Technology	2014

HONORS AND AWARDS

Dissertation Year Fellowship, University of California, Los Angeles	2020
John R. Cameron Young Investigators Symposium Finalist, AAPM	2019
Editor's Choice of Medical Physics	2019/09
Faculty of Science Graduate Award, University of British Columbia	2014-16
President's Scholarship, Shandong University	2013
National Scholarship, Chinese Ministry of Education	2011-13

PUBLICATIONS

- **Fu, J.**, Singhrao, K., Qi, S., Yang, Y., Ruan, D., & Lewis, J. H. (2021). Three-dimensional multipath DenseNet for improving automatic segmentation of glioblastoma on pre-operative multi-modal MR images. *Medical Physics.* doi: 10.1002/mp.14800
- Gao, Y., Ghodrati, V., Kalbasi, A., **Fu, J.**, Ruan, D., Cao, M., … & Yang, Y. (2021). Prediction of Soft Tissue Sarcoma Response to Radiotherapy Using Longitudinal Diffusion MRI and A Deep Neural Network with Generative Adversarial Network-Based Data Augmentation. Medical Physics. doi: 10.1002/mp.14897
- Singhrao, K., **Fu, J.**, Parikh, N. R., Mikaeilian, A. G., Ruan, D., Kishan, A. U., & Lewis, J. H. (2020). A generative adversarial network-based (GAN-based) architecture for automatic fiducial marker detection in prostate MRI-only radiotherapy simulation images. *Medical Physics*, 47(12), 6405–6413. doi:10.1002/mp.14498
- Gao, Y., Kalbasi, A., Hsu, W., Ruan, D., **Fu, J.**, Shao, J., … Yang, Y. (2020). Treatment effect prediction for sarcoma patients treated with preoperative radiotherapy using radiomics features from longitudinal diffusion-weighted MRIs. *Physics in Medicine & Biology*, 65(17), 175006. doi:10.1088/1361-6560/ab9e58
- Singhrao, K., **Fu, J.**, Gao, Y., Wu, H. H., Yang, Y., Hu, P., & Lewis, J. H. (2020). A generalized system of tissue-mimicking materials for computed tomography (CT) and magnetic resonance imaging (MRI). *Physics in Medicine & Biology,* 65(13), 13NT01. doi:10.1088/1361-6560/ab86d4
- **Fu, J.**, Zhong, X., Li, N., Van Dams, R., Lewis, J., Sung, K., … Qi, X. S. (2020). Deep learning-based radiomic features for improving neoadjuvant chemoradiation response prediction in locally advanced rectal cancer. *Physics in Medicine & Biology*, 65(7), 075001. doi:10.1088/1361-6560/ab7970
- Singhrao, K., **Fu, J.**, Wu, H. H., Hu, P., Kishan, A. U., Chin, R. K., & Lewis, J. H. (2020).

A novel anthropomorphic multimodality phantom for MRI-based radiotherapy quality assurance testing. *Medical Physics,* 47(4), 1443–1451. doi:10.1002/mp.14027

- Singhrao, K., Ruan, D., **Fu, J.**, Gao, Y., Chee, G., Yang, Y., … Lewis, J. H. (2020). Quantification of fiducial marker visibility for MRI-only prostate radiotherapy simulation. *Physics in Medicine & Biology,* 65(3), 035015. doi:10.1088/1361-6560/ab65db
- **Fu, J.**, Singhrao, K., Cao, M., Yu, V., Santhanam, A. P., Yang, Y., … Lewis, J. H. (2020). Generation of abdominal synthetic CTs from 0.35T MR images using generative adversarial networks for MR-only liver radiotherapy. *Biomedical Physics & Engineering Express,* 6(1), 015033. doi:10.1088/2057-1976/ab6e1f
- **Fu, J.**, Yang, Y., Singhrao, K., Ruan, D., Chu, F., Low, D. A., & Lewis, J. H. (2019). Deep learning approaches using 2D and 3D convolutional neural networks for generating male pelvic synthetic computed tomography from magnetic resonance imaging. *Medical Physics*, 46(9), 3788–3798. doi:10.1002/mp.13672
- Guo, M., Chee, G., O'Connell, D., Dhou, S., **Fu, J.**, Singhrao, K., … Lewis, J. H. (2019). Reconstruction of a high-quality volumetric image and a respiratory motion model from patient CBCT projections. *Medical Physics*, 46(8), 3627–3639. doi:10.1002/mp.13595

SELECTED ORAL PRESENTATIONS

- **Fu, J.**, et al. Subregion-based radiomic analysis of preoperative multi-modal MR images for improving glioblastoma survival outcome prediction. Joint AAPM/COMP, Virtual, July 2020.
- **Fu, J.**, *et al.* Improved predictive performances using deep learning in assessment of neoadjuvant chemoradiation response in rectal cancer patients based on diffusion-weighted imaging. ASTRO, Chicago, IL, Sept 2019.
- **Fu, J.***, *et al.* Use of deep learning-based radiomic features from multimodal MRI in assisting survival prediction for glioblastoma multiforme patients. ASTRO, Chicago, IL, Sept 2019.
- **Fu, J.**, et al. Improved glioblastoma survival prediction using deep learning-based radiomic features from preoperative multimodal MR images. AAPM, San Antonio, TX, July 2019. **(John R. Cameron Young Investigator Symposium)**
- **Fu, J.**, et al. Performance comparison of conditional and cycle-consistent generative adversarial networks used for abdominal synthetic CT generation. AAPM, San Antonio, TX, July 2019.
- **Fu, J.**, et al. Abdominal synthetic CT generation for MR-only liver radiotherapy using conditional generative adversarial network. AAPM, Nashville, TN, July 2018.
- **Fu, J.**, et al. Deep learning approaches using 2D and 3D convolutional neural networks for genera ting male pelvic synthetic CT from MRI. AAPM, Nashville, TN, July 2018.

1 INTRODUCTION

1.1 MR-guided radiation therapy and its limitations

Radiation therapy uses ionizing radiation to kill cancer cells or slow their growth. It continues to be an essential component of effective cancer treatment. About 50% of cancer patients would receive radiation therapy for the treatment of localized disease, local control, and palliation[1,2]. Over the year, innovations in imaging guidance have been helping improve the precision of treatment delivery[3].

Magnetic resonance imaging (MRI) has been integrated into radiation therapy treatment planning, particularly for tumors in regions like brain, pelvis, and abdomen[3,4]. This integration, also known as offline MR-guided radiation therapy (MRgRT), provides superior soft-tissue contrast for better tumor and organ-at-risk (OAR) delineation compared with conventional computed tomography (CT)-based radiation therapy[5,6]. Additionally, the ability to acquire MR functional images may help achieve accurate treatment outcome prediction[7,8]. The first commercially available online MRgRT system, MRIdian (ViewRay, OH), uses a low-field MRI and 3 Cobalt sources[9]. This system allows real-time target tracking to achieve more precise

treatment delivery within each treatment fraction[10]. Besides intra-fraction organ motion, inter-fraction anatomy discrepancy may also lead to large uncertainties of treatment delivery. Online MRgRT system also allows the acquisition of high-quality daily MR images, which could achieve online adaptive MR-guided radiation therapy (oaMRgRT) for minimizing inter-fraction anatomy discrepancy. A phase I trial study showed that oaMRgRT increases target coverage and achieves better OAR sparing for abdominal cancer compared with non-adaptive radiation therapy[11].

There are three major limitations in the current MRgRT workflow. First, both offline and online MRgRT requires the acquisition of a planning CT for treatment planning. This is because that MR images, unlike CTs, cannot be directly used to generate electron density maps for dose calculation. However, acquiring an additional CT increases unwanted radiation exposure, clinical workload, and financial cost[12]. Additionally, co-registering MR image and CT is required for transferring delineation structures from the MR image to the CT. This process introduces a systematic uncertainty, on the order of 2-5 mm depending on the anatomical site, that propagates throughout the treatment[13]. Second, Lamb *et al.* reported that recontouring target and OARs in the oaMRgRT took up to 22 minutes even with the help of the autosegmentation tool provided by the MRIdian system[14]. This slow recontouring process would decrease patient comfort, extend treatment time, and decrease the effectiveness of adapted plans due to possible anatomy change. Third, the methods for predicting treatment response and survival outcome based on MR images are currently underdeveloped.

1.2 Synthetic CT generation for MR-only radiation therapy

Synthetic Hounsfield Unit (HU) maps, also known as synthetic CT (sCT) images, must be accurately generated from MR images to achieve MR-only radiation therapy. There are three types of methods developed for sCT generation: atlas-based, voxel-based, and hybrid methods. In atlas-based methods, one or multiple co-registered MR-CT images is deformably registered to a patient's MR image[15–17]. The resulting transformation can then be applied to the CT-atlas to generate the sCT. Atlas-based approaches can be time-consuming, especially when atlases are large. They may also easily fail if the patient has a very different anatomy from what is represented by the atlas. Voxel-based methods convert individual MR voxel intensities to HU values using bulk density assignments or machine learning models[18–21]. Bulk density assignments may lead to dose discrepancies and often have limited value in generating positioning reference images. Voxel-based machine learning methods are promising but the generation time is normally long. Hybrid methods combine elements of voxel-based and atlas-based approaches. Recently, deep learning (DL) methods including convolutional neural networks (CNNs) and generative adversarial networks (GANs) achieved state-of-the-art performance in image-to-image translation[22–25]. DL methods could rapidly generate sCT images and be integrated into the oaMRgRT.

1.3 CNNs for automatic tumor segmentation

Fast and accurate automatic tumor segmentation can speed up online adaptive planning. It is also essential for diagnosis, disease monitoring, and tumor characterization. Manually delineation is not only time-consuming but also sensitive to intra-observer and inter-observer variations.

Developing automatic tumor segmentation methods, which can generate reproducible and accurate segmentation, has drawn great research interests. Recently, 2D CNNs have been proposed and achieved great performance in nature scene semantic segmentation[26–29]. Studies have shown that 3D CNNs have achieved better performance in medical image segmentation than the corresponding 2D CNNs[23,30]

1.4 Radiomics for treatment outcome and survival prediction

It is beneficial for disease management to develop methods for predicting treatment response and survival outcome. Radiomic features extracted from the tumor using advanced mathematical algorithms may uncover tumor characteristics that fail to be appreciated by the naked eye[31,32]. Recent studies show radiomic features can assist tumor grading without biopsy, treatment outcome prediction, and survival prediction[32–35]. Commonly used radiomic features are acquired by explicitly designed, or "handcrafted", algorithm[33]. However, these handcrafted features are normally shallow and low-order image features and limited to the current human knowledge. Recent studies demonstrated that higher-order DL-based features, extracted using pre-trained CNNs, achieved better performance than handcrafted features for several classification tasks[38,39].

1.5 Specific aims

The goal of this proposal is to address those three limitations of the MRgRT workflow using DL methods. We hypothesize that the proposed DL methods can generate accurate sCTs for treatment planning, achieve fast and accurate automatic tumor segmentation, and provide accurate predictions of treatment response and survival outcome. The following specific aims

focus on developing key methodologies to accomplish the goal. Due to limited data availability, we investigated a wide range of cancer types across different tasks. But the developed methods could be adapted to other cancer types with enough training data.

- Specific aim 1 (SA1): Develop DL methods for generating pelvic and abdominal sCTs from MR images for MR-only radiation therapy.
- Specific aim 2 (SA2): Develop a DL method for automatic glioblastoma multiforme (GBM) segmentation based on multi-modal MR images.
- Specific aim 3 (SA2): Develop a machine learning method for predicting neoadjuvant chemoradiation treatment (nCRT) response in locally advanced rectal cancer (LARC) based on pre-treatment diffusion-weighted MR image.
- Specific aim 4 (SA4): Develop an automatic radiomic workflow for GBM survival prediction based on multi-modal MR image.

1.6 Overview

Chapters 2 through 6 contain versions of manuscripts written based on the core projects of this dissertation. Four manuscripts have been published[40–43], and one manuscript is currently under review[44]. Each chapter consists of an introduction section that thoroughly addresses the study motivation and background.

SA1 is addressed in Chapters 2 and 3. Chapter 2 describes a 3D CNN we proposed for generating pelvic sCT images from 1.5T T1-weighted MR images. When we started to work on abdominal sCT generation, we did a pilot study and found that a conditional GAN (cGAN) could generate more accurate abdominal sCT images compared with the corresponding CNN. So we investigated the cGAN and cycle-consistent GAN (cycleGAN) for generating abdominal sCT

images from 0.35T MR images in Chapter 3. Chapter 4 addresses SA2 and describes a novel multi-path 3D DenseNet that we proposed for automatic GBM segmentation. Chapter 5 and Chapter 6 address SA3 and SA4, respectively. They consist of two radiomic studies, one for early prediction of nCRT response in LARC, and the other for GBM survival prediction. Both radiomic studies compared DL-based radiomic features with conventional handcrafted radiomic features. The study on nCRT response prediction relied on manual tumor segmentation, while the GBM survival prediction study used an automatic segmentation model.

2 DEEP LEARNING APPROACHES USING 2D AND 3D CONVOLUTIONAL NEURAL NETWORKS FOR GENERATING MALE PELVIC SYNTHETIC CT FROM MRI[40]

2.1 Introduction

MRI is often integrated into radiation therapy treatment planning[45], particularly for tumors in regions like the brain, head and neck, and prostate[16]. The superior soft-tissue contrast of MR images facilitates precise delineations of tumor and OARs[5,6]. MR images can also provide guidance for adaptive radiation therapy. MR images can also provide guidance for adaptive radiation therapy[46,47]. The standard MRgRT workflow includes acquisition of a planning CT

image. The CT HU map, essentially a scaled linear attenuation map, is used to generate both electron density maps for dose calculation and digital reconstructed radiographs for subsequent patient positioning.

The need to acquire the CT image in MRgRT brings in several disadvantages. First, acquiring an additional CT scan increases unwanted radiation exposure, clinical workload, and financial cost. Second, co-registering CT and MR images is required for transferring delineation structures from the MR image to the CT image. This process introduces a systematic uncertainty, on the order of 2-5 mm depending on the anatomical site, that propagates throughout the treatment[13]. MR-only radiation therapy can avoid these pitfalls.

To achieve MR-only radiation therapy, sCT images must be accurately generated from the MR images. To date, there are three types of methods developed for this: atlas-based, voxel-based and hybrid[13]. In atlas-based methods, a set of one or multiple co-registered MR-CT images are deformably registered to a patient's MR image[15,17,48]. The resulting transformation can then be applied on the CT-atlas to generate the sCT image. Atlas-based approaches can be time-consuming, particularly when the atlases are large, and often fail if the patient has very different anatomy from what is represented by the atlas.

Voxel-based methods convert individual MR voxel intensities to HU values using bulk density assignments or machine learning models. Bulk density techniques assign the patient's electron density either to water or to pre-defined electron densities within selected MR-segmented tissue types[18,49–51]. These methods may lead to dose discrepancies and often have limited value in generating positioning reference images. Machine learning methods use paired MR-CT images to train models that associate MR intensities with HU values. It is challenging for models to distinguish air from bone in conventional MR images as both tissues exhibit weak

signals due to their small T2 values. Some learning methods required manual bone segmentation[19,52] in conventional MR images or require acquisition of specialized MR sequences like ultrashort echo time sequences[20,53,54] for separating bone and air.

Hybrid methods combine elements of voxel-based and atlas-based approaches[55]. A detailed summary of previous approaches can be found in the review paper by Karlsson *et al* [13].

Recently, DL methods[22] proposed to estimate sCT images from MR images have demonstrated promising results. Nie *et al.* presented a 3D CNN with three convolutional layers[56]. It was trained to convert 3D patches of pelvic MR images to corresponding 3D sCT patches. The sCT image was then generated by averaging the HU values of overlapping sCT patches. An updated model with an adversarial network was later proposed to improve the sCT quality[57]. Training on patches rather than whole volumes reduces the required number of CNN parameters and saves computational resources. However, using patches might miss larger scale (relative to patch size) image features. For example, the use of small patches could preclude the use of spatial context that could assist in differentiating between tissues with the similar appearance on MR image, but very different HU values (e.g., rib and lung tissues, or cortical bone and bowel gas). A 2D CNN with 27 convolutional layers was proposed by Han for brain sCT generation[58]. This more complex model could capture long-range information and generate brain sCTs slice by slice without dividing images into patches.

3D CNNs may have better performance than their corresponding 2D CNNs because they use entire image volumes rather than individual slices, allowing the exploitation of more information (e.g. relationships between consecutive slices). However, 3D models have some disadvantages. 3D models contain more parameters, potentially requiring more training data to achieve robust performance. 3D models can also be more difficult to implement on commonly-

available GPU cards due to their large memory consumption. It is reasonable to expect that a 3D model will have better performance when enough training data and sufficiently powerful computational hardware are available.

In this study, we investigated the performance of generating sCT images using CNNs in the male pelvis, which has greater anatomic variation than the brain. Several studies have shown that prostate target definition is more accurate on MR images than CT images[59–61]. For this reason, MR images for prostate cancer treatment are routinely acquired in many clinics. Implementing an MR-only workflow for prostate cancer radiation therapy would provide the benefit of removing a CT, as discussed above. As prostate cancer is one of the diseases most commonly treated with radiation, removing the need for CTs could have a large impact. We built a 2D CNN based on Han's 2D CNN, with three modifications implemented to save computational memory and prepare for the extension to 3D. The motivation for extending the 2D model to a 3D model is to test whether a small patient-cohort is enough to effectively train a 3D model. 2D and 3D model performances were compared with Han's model. In the training stage, we incorporated on-the-fly data augmentation and a modified loss function to enhance model performance. All three models were trained from scratch without implementing transfer learning. Geometric and voxel-wise metrics and patient alignment tests were used to evaluate and compare model performances.

This study contains a few novel elements. First, we proposed a 2D CNN that consumed less memory and achieved more accurate HU prediction than the state-of-art model proposed by Han. Second, to our knowledge, this is the first study that applies end-to-end 3D CNNs in sCT generation. While the previous study suggested that using a small dataset to train a 3D CNN may be infeasible due to overfitting[58], we demonstrated that a 3D CNN can be trained to generate

accurate pelvic sCT images using a 16-patient dataset. Also, our results show that the proposed 3D CNN achieved more accurate HU prediction than the corresponding 2D CNN or Han's model.

2.2 Materials and methods

2.2.1 Dataset

Retrospective analysis was performed using CT and MR images from 20 prostate cancer patients. Table 2-1 summarizes patient characteristics. All patients had intact prostates and no hip prosthesis. CT and MR images were acquired before radiation therapy. The CT images were acquired on a 64-slice CT scanner (Sensation, Siemens Medical Solutions, Erlangen, Germany) using the following settings: 120 kVp, 400 mA, and 1.5 mm or 3 mm slice thickness, with in-plane spatial resolutions varying from 0.85 × 0.85 mm^2 to 1.27 × 1.27 mm^2. The in-plane dimensions of CT images are 512 × 512. For each patient, an MR image was acquired on the same day as the CT image with a non-contrast T1-weighted 2D turbo spin echo sequence (echo time: 12 ms or 13 ms, repetition time: 523 ms to 784 ms, flip angle: 150^o) on a 1.5T MR scanner (Sonata, Siemens Healthcare, Erlangen, Germany). MR images had slice thickness of 5 mm and in-plane spatial resolutions ranging from 0.71 × 0.71 mm^2 to 0.94 × 0.94 mm^2. The in-plane dimensions of MR images range from 384 × 348 to 448 × 448.Thirty slices covering the prostate region were extracted from MR images.

Characteristic	Type	No. of patient
Age	Mean ± std (range)	69 ± 6 (61-80)
Prostate volume (cc)	Mean ± std (range)	56 ± 23 (20-107)
Cancer stage	T1c	17
	T2a	3

Table 2-1: Patient characteristics

2.2.2 Image preprocessing

Figure 2-1 (a) and (b) outline the image preprocessing and CNN training workflows, respectively. In the preprocessing stage, N4 bias field correction was applied on the MR images to remove inhomogeneity artifacts[62]. Histogram-based normalization was also performed to minimize the inter-patient MR intensity variation[63]. A body mask of each patient, which was used for restricting loss evaluation and sCT accuracy assessment, was generated from the bias-corrected MR image using Otsu's thresholding[64] followed by opening and closing morphological operations. To account for organ movement and patient setup variations between CT and MR images, the CT image was registered to the bias-corrected MR image using rigid and affine registrations, followed by a multi-resolution B-spline registration (Elastix[9]). Each deformed CT (dCT) was resampled to match the MR image resolution. An experienced clinical physicist reviewed the fusion of each dCT image with its paired MR image to ensure that there were no major qualitative errors in the registrations. The physicist assessed the alignment quality of pelvic bones and femurs in fused image displays. The Jacobian determinants of the deformation vector fields were also examined to confirm that all were greater than 0, and that no large local changes occurred. In the training stage, the sCT was generated by feeding the nMRI into the CNN. The loss was computed as the mean absolute error between the sCT and dCT within the

body mask and then minimized by updating variables of the CNN using backpropagation and stochastic gradient descent.

Both MR and dCT slices were down-sampled to 256 × 256, with pixel sizes varying from 1.25 × 1.11 mm^2 to 1.41 × 1.41 mm^2. Down-sampling reduces GPU memory consumption, enabling implementation on less expensive GPUs. However, down-sampling could result in some information loss. The tradeoffs associated with down-sampling are addressed in the Discussion section. The down-sampled in-plane resolution used in this study is within the typical clinically acceptable range for dose calculation and patient positioning in prostate cancer radiation therapy.

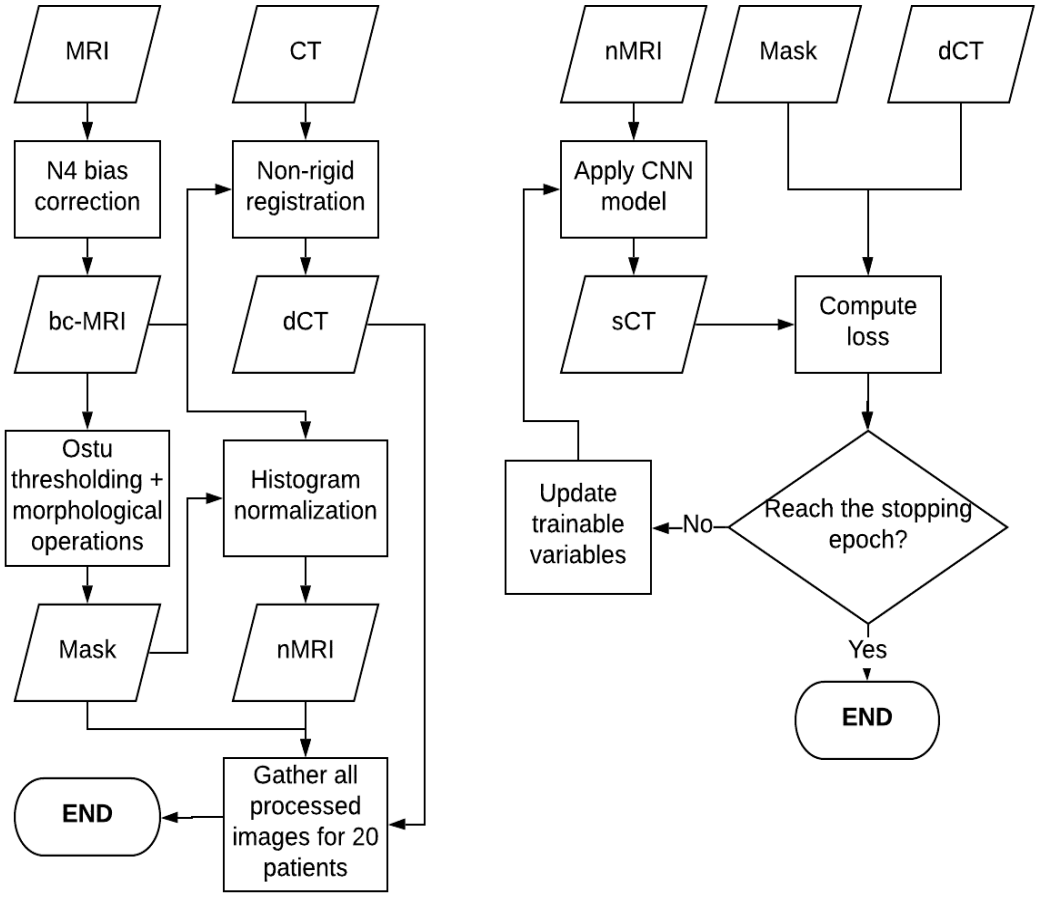

Figure 2-1: The overall workflow of sCT generation.

2.2.3 2D and 3D CNNs

The proposed 2D model was built based on Han's model and extended to 3D. 2D MR slices and 3D MR volumes were fed into the corresponding CNNs which were trained to output 2D sCT slices and 3D sCT volumes, respectively. Figure 2-2 shows the architecture of the 2D model. The 3D model shared the same architecture as the 2D model except that all 2D operations were replaced with their corresponding 3D counterparts. For example, all 3×3 2D convolution filters were replaced with 3×3×3 3D convolutional filters.

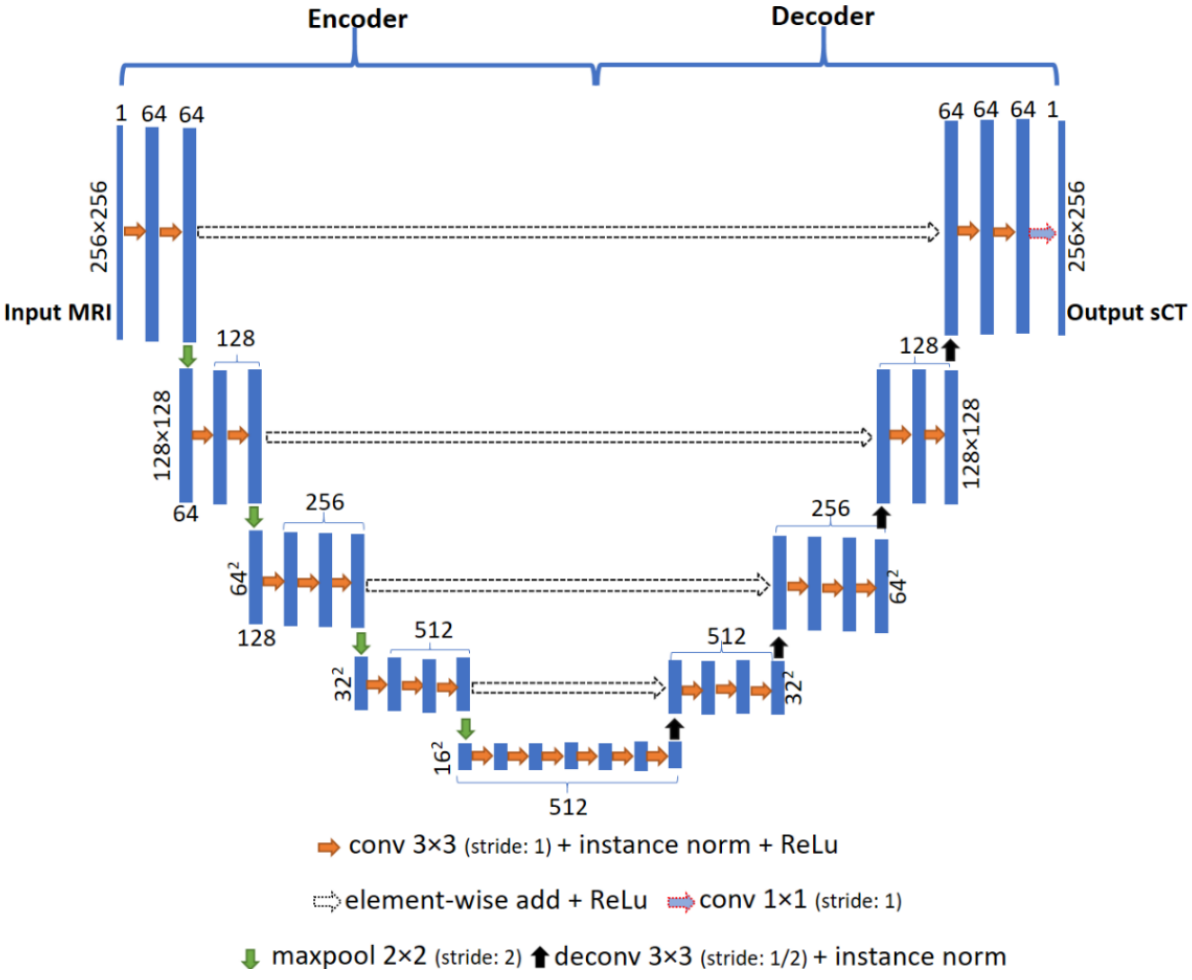

Figure 2-2: The overall 2D CNN architecture. Each filled box represents a set of feature maps, the numbers and dimensions of which are shown.

The 2D model has encoder and decoder networks. The encoder network, consisting of 13 convolutional layers, is identical to the convolutional layers in the VGG16 model[65], except that filters in the first convolutional layer have a depth of 1 rather than 3, because of the scalar nature of MR and CT. Each encoding convolutional layer performed convolution of its input with a set of 3 × 3 trainable filters at a stride of 1. Trainable filters have sets of trainable weights and biases that can be applied to input feature maps to produce deeper feature maps. Zero padding of 1 was used before convolution to ensure the produced deeper feature map had the same resolution as the input feature map. Feature maps were then normalized using instance normalization[66] to reduce internal covariate shifts and then operated by the element-wise activation function max(0,x), termed the Rectified Linear Unit (ReLU). The feature maps were downsampled by applying a maxpooling layer with a 2×2 window and a stride of 2. The sequence of several convolutional layers and maxpooling layers act to extract local and global features and increase translation invariance. The decoder network, consisting of a hierarchy of decoders, was used to upsample low-resolution feature maps and gradually reconstruct the sCT. Each decoding convolutional layer corresponded to an encoding convolutional layer, except for the final convolutional layer that had a set of 1 × 1 trainable filters with a stride of 1.

Three modifications to Han's model were made to develop the proposed models. First, batch normalization layers[67] were replaced with instance normalization layers[66]. Our tests, using all patients, showed that the model with instance normalization layers had better performance than the one with batch normalization layers when trained with a small batch size (which was limited by our GPU memory). Small batch sizes can cause less accurate mean and variance estimations, diminishing the effectiveness of batch normalization[68]. Second, the fractional-stride convolutional layers (also known as deconvolutional layers) were employed to replace the

unpooling layers. Unlike unpooling layers producing sparse feature maps, deconvolutional layers can be trained to produce dense feature maps[27,69], which should help to generate better quality sCT images. This modification also saves computational memory, which is critical for building the 3D model, because unpooling layers require more memory to keep track of maxpooling indices. Lastly, inspired by ResNet[70], U-Net skip connections[71] were replaced with residual shortcuts to further save computational memory. A residual shortcut adds encoder feature maps to corresponding upsampled feature maps (leaving the number of feature maps unchanged), while a skip connection concatenates encoder feature maps with upsampled feature maps (doubling the number of feature maps). To investigate and compare the effectiveness of these two types of shortcuts, we trained the 2D models with: 1) residual shortcuts; 2) skip connections; and 3) neither. Our tests, using all patients, showed that the models with residual shortcuts or skip connections yielded better MAE results than the model with neither, and residual shortcuts resulted in the similar performance as skip connections.

Weights and biases of trainable filters in the convolutional layers and deconvolutional layers were trained by minimizing a loss function. The loss function was defined as the mean absolute error (MAE) between the sCT and dCT within the body mask;

$$loss = \frac{1}{N}\sum_{i=1}^{N}|sCT_i - dCT_i|$$

Equation 2-1

where N was the number of voxels inside the body masks of MR images, and sCT_i and dCT_i represented the HU values of the i^{th} voxel in the sCT and dCT, respectively.

2.2.4 Model optimization

The proposed models and Han's model were implemented using Tensorflow packages[72] (V1.3.0, Python 2.7, CUDA 8.0) on Ubuntu 16.04 LTS system. All three models were trained with instance normalization and identical hyperparameters except for the batch size. At each iteration, a mini-batch of 2D images or 3D volumes was randomly selected from the training set. The batch size was limited by GPU memory. A mini-batch of 15 training slices was used to run the 2D model on an 8 GB NVIDIA GeForce GTX 1080 GPU, and Han's model on a 12 GB NVIDIA GeForce GTX Titan X GPU. The large memory GPU card was necessary for implementing Han's model due to its greater memory consumption compared with the proposed 2D model. The 3D model was run on a 12 GB Titan X GPU with a mini-batch of 1 training volume. The Adam stochastic gradient descent method[73] with default parameters ($\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 10^{-8}$), except for learning rate (0.01), was used for minimizing the loss function. On-the-fly data augmentation (random shift and rotation) was performed on each set of MR images, body masks, and dCT images to reduce overfitting. For all three models, the random translation was up to 15 pixels in the x and y directions, and the random rotation around the z-axis was confined within ±5°. Rotations with random angles within ±2° around the y-axis and x-axis were applied to the 3D images. The rotation and translation ranges are intended to roughly match patient positioning uncertainties. Weights of all models were initialized using He initialization[74], and the biases were initialized to 0. He initialization was selected based on literature reporting that it performs better than Xavier initialization for deep models with ReLu layers[74].

2.2.5 Model evaluation

Five-fold cross-validation was performed to evaluate model performance. The 20 patient-cohort was randomly divided into five groups. Each time validation was performed, four groups were used as the training set to train the model. The trained model was then used to generate sCT images of patients in the remaining group. For Han's and the proposed 2D models (3D model), four groups of four patients provided 480 (16) training samples. Using the batch size of 15 (1), it took 32 (16) iterations to go over all samples in the training set, which was considered as one epoch.

CNN accuracy was evaluated by using voxel-wise MAE between the sCT and dCT for three regions: 1) the whole body; 2) a soft tissue region generated by thresholding the dCT with a range [-100,150) HU; and 3) a bone region generated by thresholding the dCT at 150 HU, *i.e.* [150,+∞) HU.

CNN accuracy was also evaluated by calculating geometric metrics, such as the dice similarity coefficient (DSC), recall, and precision for the bone region. They were defined as:

$$DSC = \frac{2(V_{sCT} \cap V_{dCT})}{V_{sCT} + V_{dCT}}, recall = \frac{V_{sCT} \cap V_{dCT}}{V_{dCT}}, precision = \frac{V_{sCT} \cap V_{dCT}}{V_{sCT}}$$

Equation 2-2

where V was the bone region volume generated by thresholding the dCT or sCT at 150 HU.

The patient alignment tests based on bony structures were conducted to test whether generated sCT images can provide accurate patient positioning. To do this, for each patient, bone regions of sCT and dCT were rigidly aligned to the cone-beam CT (CBCT), acquired for

positioning in the first treatment fraction. The translation vector distances and absolute Euler angle differences were calculated for evaluation.

With the assumption of normal distributions for MAE metrics and transformation vector differences, analysis of variance (ANOVA) for repeated measures was carried out to compare these metrics among the three sCT generation models (2D model, 3D model, and Han's model). The normality assumption was imposed after assessing the data with quantile-quantile and frequency plots. A paired t-test was conducted as post-hoc analysis if a difference among three models was identified by ANOVA. While with assumption of non-normal distributions on bone geometric metrics, Friedman test for repeated measures was carried out to compare these metrics among the three models. A Wilcoxon signed-rank test[75] was conducted when a difference between three models was identified by the Friedman test. A p-value of 0.05 was considered significant, and the Bonferroni correction was used when applicable.

2.3 Results

All models were trained for 200 epochs. This epoch number was selected based on training and validation loss tracking. It required approximately 2 (4) hours to train an individual cross-validation 2D (3D) model. The time required for generating the whole sCT volume of a patient was approximately 5.5 s for both 2D and 3D models.

Figure 2-3 shows transverse slices of sCT images generated by the all three models along with the corresponding slices of the normalized T1-weighted MR images and dCT images from three patients. More transverse slices showing the bladder or the corpus cavernosum from these three patients are presented in Figure 2-4. As shown in the difference maps, all three models

gave accurate HU value predictions for most regions, especially soft tissues, but had difficulty generating accurate HU values near the body contour and bone borders.

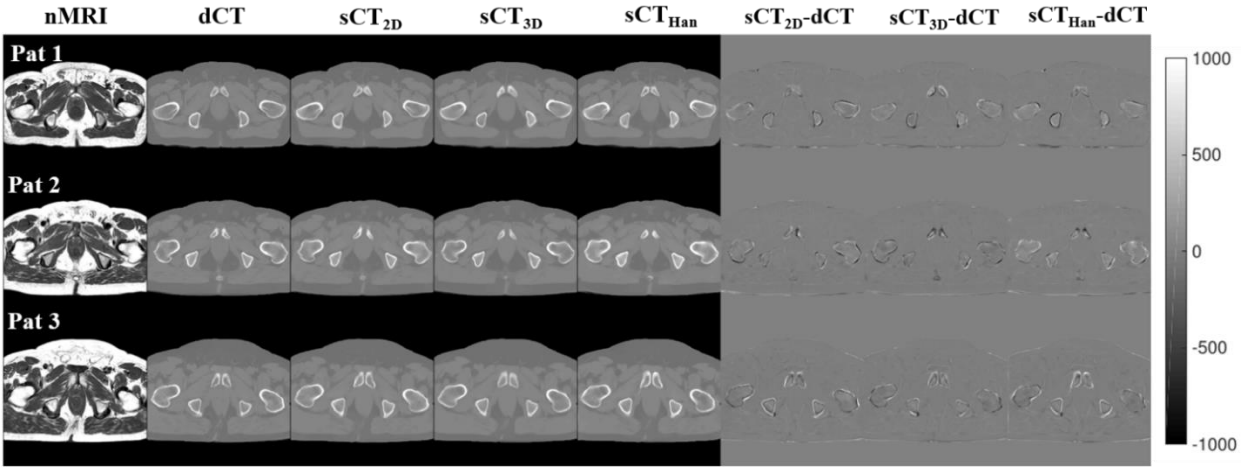

Figure 2-3: Transverse slices of the normalized MR images, the dCT images, the 2D model sCT images, the 3D model sCT images, and Han's model sCT images from three patients. The last three columns show the difference maps between the 2D model sCT images and the dCT images, and the difference maps between the 3D model sCT images and the dCT images, and the difference maps between Han's model sCT images and the dCT images. The color bar is associated with all images except normalized MR images.

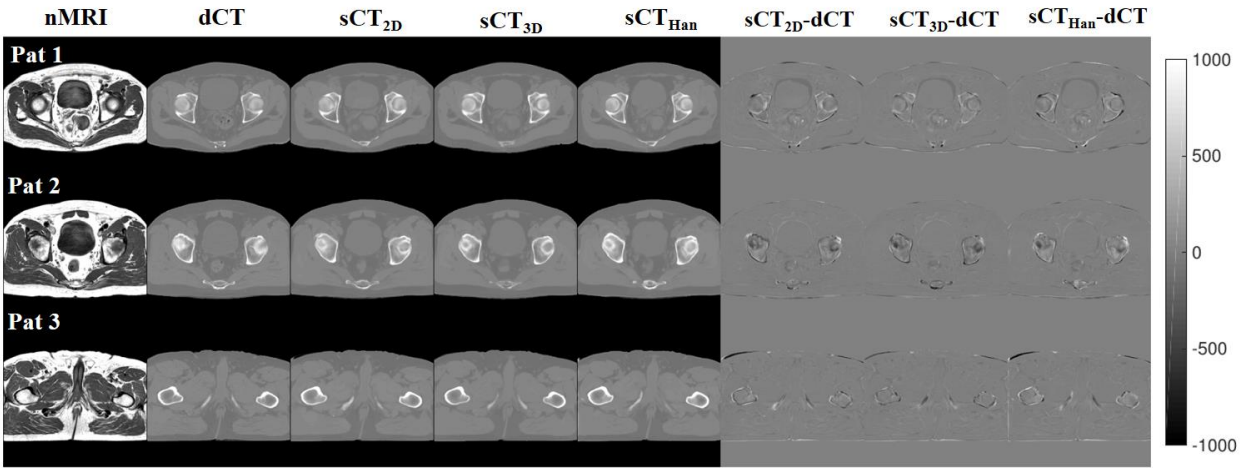

Figure 2-4: Additional transverse slices showing different anatomical regions for the patients shown in Figure 2-3.

The MAE, averaged across all patients, as a function of dCT values and relative frequency of CT HU values are shown in Figure 2-5 (a). The MAE was calculated in 25 HU bins. Similarly, mean error (ME, sCT-dCT) curves are shown in Figure 2-5 (b). All three models

have similar MAE curves for most HU values except that the Han's and the proposed 2D models yielded greater MAEs than the 3D model within (-650, -200) HU, and vice-versa within (850,1600) HU. ME curves suggest all three models underestimate absolute HU values.

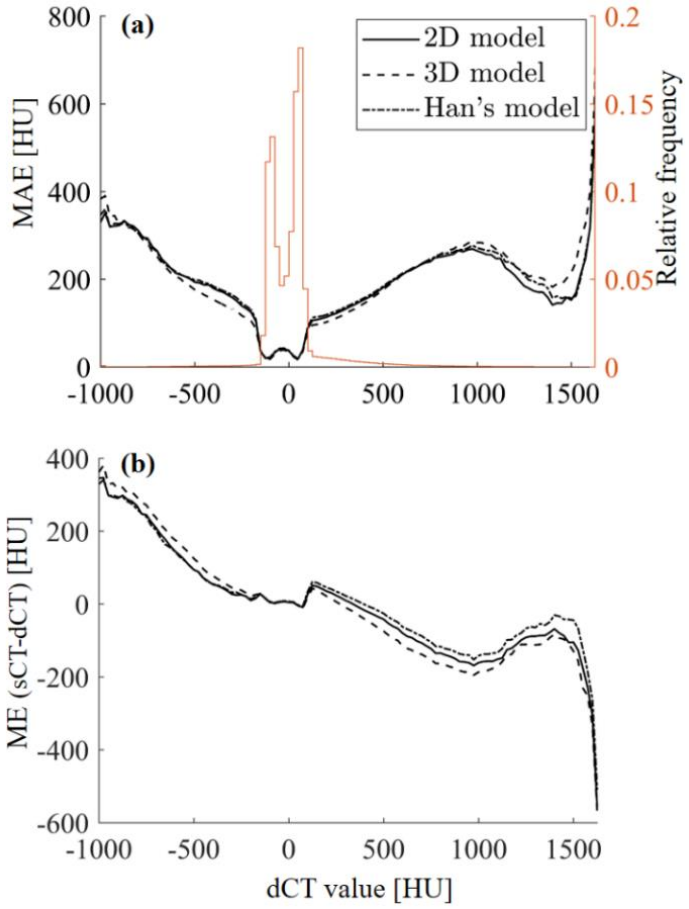

Figure 2-5: (a) Left axis: MAE of voxels within body masks from all patients as a function of dCT values, calculated in 25 HU bins. Right axis: relative frequency of voxels within each HU bin. (b) ME of voxels within body masks from all patients as a function of dCT values, calculated in 25 HU bins.

Table 2-2 summarizes the voxel-wise metrics, geometric metrics, and transformation vector differences, averaged across all patients. Table 2-3 shows the average whole-body MAE of four patients in each cross-validation fold across five folds. No indication of overfitting was observed during training. The maximum MAEs within the body were 56.5 HU, 53.1 HU, and 60.7 HU, for 2D, 3D and, Han's models, respectively. The minimum bone region DSCs were

0.70, 0.72, and 0.66, for 2D, 3D, and Han's models, respectively. The average translation vector distances are less than 0.6 mm for all three models. The average absolute differences of three Euler angles are less than 0.5° for all three models. Results of statistical tests are summarized in Table 2-4. Significant differences in MAEs for three regions and the bone precision were observed among three models.

		2D model	**3D model**	**Han's model**
MAE [HU]	whole body	40.5±5.4	37.6±5.1	41.9±6.5
	soft tissue	28.9±4.7	26.2±4.5	29.9±5.8
	bone	159.7±22.5	154.3±22.3	165.0±26.9
Bone DSC		0.81±0.04	0.82±0.04	0.80±0.05
Bone recall		0.85±0.04	0.84±0.04	0.86±0.04
Bone precision		0.77±0.09	0.80±0.08	0.76±0.10
$\left\|\vec{T}_{dCT} - \vec{T}_{sCT}\right\|_2$ [mm]		0.51±0.28	0.54±0.38	0.51±0.27
$\left\|\theta_{x\_dCT} - \theta_{x\_sCT}\right|$ [°]		0.47±0.74	0.47±0.72	0.47±0.76
$\left\|\theta_{y\_dCT} - \theta_{y\_sCT}\right|$ [°]		0.08±0.05	0.10±0.09	0.09±0.07
$\left\|\theta_{z\_dCT} - \theta_{z\_sCT}\right|$ [°]		0.08±0.09	0.07±0.07	0.09±0.10

Table 2-2: Results of voxel-wise metrics, geometric metrics, and transformation vector differences for sCT images generated by all three models. Results were averaged across the 20-patient cohort and shown in (mean ± SD) format.

Fold #	**1**	**2**	**3**	**4**	**5**
2D model	40.2	44.1	41.5	37.4	39.2
3D model	38.5	39.2	38.6	34.5	37.4
Han's model	40.5	46.7	45.2	36.3	40.8

Table 2-3: Results of single-fold-average MAEs for five cross-validation folds. Four patients are analyzed within each fold.

		ANOVA Or Friedman	Post-hoc analysis					
		p-value	2D vs 3D		2D vs Han		3D vs Han	
			95% CI	p-value	95% CI	p-value	95% CI	p-value
MAE	**whole body**	<0.001	(1.83,3.86)	<0.001	(-2.36, -0.48)	0.005	(-5.60, -2.93)	<0.001
	soft tissue	<0.001	(1.88,3.57)	<0.001	(-1.95, -0.12)	0.03	(-4.96, -2.57)	<0.001
	bone	<0.001	(1.01,9.83)	0.02	(-9.73, -0.81)	0.02	(-16.73, -4.65)	0.002
Bone DSC		0.09	NA	NA	NA	NA	NA	NA
Bone recall		0.35	NA	NA	NA	NA	NA	NA
Bone precision*		<0.001	-	<0.001	-	0.04	-	<0.001
$\left\|\vec{\mathbf{T}}_{\mathbf{dCT}} - \vec{\mathbf{T}}_{\mathbf{sCT}}\right\|_{\mathbf{2}}$		0.71	NA	NA	NA	NA	NA	NA
$\left\|\boldsymbol{\theta}_{\mathbf{x\_dCT}} - \boldsymbol{\theta}_{\mathbf{x\_sCT}}\right\|$		0.95	NA	NA	NA	NA	NA	NA
$\left\|\boldsymbol{\theta}_{\mathbf{y\_dCT}} - \boldsymbol{\theta}_{\mathbf{y\_sCT}}\right\|$		0.70	NA	NA	NA	NA	NA	NA
$\left\|\boldsymbol{\theta}_{\mathbf{z\_dCT}} - \boldsymbol{\theta}_{\mathbf{z\_sCT}}\right\|$		0.70	NA	NA	NA	NA	NA	NA

* Wilcoxon signed-rank test was used as post-hoc analysis for bone region precision. The values of mean differences were not reported here since this was not the aspect Wilcoxon signed rank test assessed.

Table 2-4: Statistical test results for comparing the three sCT generation models. In ANOVA or Friedman test, a p-value of <0.05 is considered significant. In post-hoc analysis, a p-value of < 0.0167 is considered significant as per the Bonferroni correction.

2.4 Discussion

In this study, 2D and 3D CNNs were proposed to generate pelvic sCT images from T1-weighted MR images. Our trained models are fully automated for sCT generation, requiring no deformable registration or manual segmentation of bone. Deformable registration was only applied on CT images in the training stage. During inference, all preprocessing steps on MR images can be achieved automatically. As shown by sCT and dCT difference maps in Figure 2-3 and Figure 2-4, the 2D and 3D CNNs generated accurate sCT images with HU values similar to their corresponding dCT images.

MAE and ME curves shown in Figure 2-5 indicated that the proposed models could precisely estimate soft-tissue HU values but had larger errors in reproducing air and bone. Trained models underestimate voxel absolute HU values as shown by ME curves. There are a few possible reasons. First, air and bone are both barely visible in T1-weighted MR images due to weak signals, making their HU prediction challenging. Second, registration errors between the MR and CT images would have more impact on the intensity mapping of end-of-range voxels than for soft tissue voxels. Misregistration can cause air- and bone-tissue boundaries to be shifted, introducing intensity mapping errors. As the total number of high absolute HU voxels is small, the proportion of misaligned labels is higher, which may introduce more perturbation. However, misregistration within the soft tissue itself does not have a large impact on the intensity mapping. Third, as suggested by the HU histogram in Figure 2-5 (a), most voxels within body contours have CT numbers in the low absolute HU range. This led to uneven sampling for training CNNs, which may result in the tendency of trained models to map voxels to the low absolute HU region in the testing stage. There are a few possible solutions for improving bone accuracy that may be worth future investigation. First, higher loss weights could

be assigned to bone structures for training. Second, we could develop a purpose-driven model (*e.g,* a model trained with bone-only images, so that only the accuracy of bone HU is considered). Such a model might be useful for generating images for bone-based patient alignment.

The average MAEs within the body contour across all patients were 40.5±5.4 HU and 37.6±5.1 HU for the 2D and 3D models, respectively. Our MAE results are comparable with published results presented in Table 2-5. Our CNN methods allow sCT to be generated quickly enough that it would be non-burdensome for most or all clinical tasks. The average bone region DSCs were 0.81±0.04 and 0.82±0.04 for the 2D and 3D models, respectively. Of the studies listed in Table 2-5, only Dowling et al. reported bone region DSC. This atlas-based method reported a bone region DSC of 0.91. However, they used a different method for computing bone region DSC. In their method, the DSC is calculated by comparing manually drawn MR bone contours and automatically computed bone contours, while we compared CT and sCT bone contours derived from HU thresholding. Considering the difference between ground truth images of two methods (manually drawn MR bone contours vs thresholded CT bone contours), a direct comparison between our DSC results is equivocal. A study conducted by Arabi *et al.* compared Han's model with four atlas-based methods (including Dowling's), showing that Han's model achieved the smallest MAE and similar bone DSC [76]. As shown in Table 2-2, sCT images generated by the proposed model can provide accurate patient positioning based on bony structures. ANOVA results shown in Table 2-4 indicated there is no significant difference in sCT patient positioning accuracy among three models.

	Kim *et al.*[77]	**Dowling *et al.*[17]**	**Andreasen *et al.*[78]**	**Andreasen *et al.*[79]**	**Siversson *et al.*[55]**	**Nie *et al.*[57]**
Method type	Voxel-based	Atlas-based	Hybrid	Voxel-based	Hybrid	Voxel-based
MAE [HU]	74.3±3.9	40.5±8.2	54.0±8.0	58.0±9.0	36.5±4.1	39.0±4.6
Generation time [min]	N.A. Requires bone contours	N.A.	20.8	N.A.	50 to 80	N.A.

Table 2-5: MAEs within the body contour published in previous pelvic sCT generation studies.

In assessing the clinical relevance of the results reported in Table 2-2, it may be useful to place the MAE results in the context of typical uncertainties observed in CT simulation images. While noise levels in CT images depend on numerous factors (imaging protocol, patient geometry, reconstruction algorithm, etc.), reported MAE results are beginning to approach typical HU variations observed during monthly or annual CT simulator QA (e.g., about ±15 HU for soft tissue, and about ±30 HU for bone). AAPM has recommended a tolerance of ±5 HU for field uniformity and for HU accuracy in water[80]. While these values leave some room for improvement in currently reported MAE results, improving sCT HU accuracies beyond the level of these other clinically acceptable uncertainties may not be practically useful.

It was feasible to train the proposed 3D model with 16 image volumes from scratch. The proposed 3D model shows better performance in generating sCT images compared with Han's and the proposed 2D models. Results of statistical tests shown in Table 4 demonstrated statistically significant improvements in MAEs for three regions and bone region precision of the 3D model compared with 2D and Han's model. Also, not only did the 2D model consume less

memory compared with Han's model, it generated sCT with a smaller average MAE as shown by paired t-test results.

There are a few factors that may affect model accuracy. First, MR intensities do not have a fixed tissue-specific numeric meaning. There is a large intensity variation across different subjects even with the same MR sequence and scanner. Although histogram-based intensity normalization was applied, the remaining variation might still be one of the largest error sources for training the mapping of MR intensities to CT HU values. Second, small training dataset size may limit the trained model to a small scope of anatomy variation. This potentially leads to large errors for abnormally large or small patients. As the number of training patients increases, CNN models are expected to be more robust and have better sCT generation performance. It should be noted that the dataset used in this study did not include patients with hip prostheses, radiation-induced fistulas, prostatectomies, or other such abnormalities. The performance of sCT generation for such cases was not explored, and the development of sCT methods for such patients is an interesting area for future work.

The MR and dCT slices were down-sampled to 256 x 256 so that the proposed models could be implemented on a single GPU (8 GB for 2D CNN or 12 GB for 3D CNN). Developing CNNs with smaller GPU memory consumption has several advantages for clinical implementation including easy model distribution/usage and better cost-effectiveness. While down-sampling can result in some information loss, sCT images generated in this study achieved accurate patient positioning and small MAEs.

We directly down-sampled the MR images to keep their original fields of view, resulting in resampled images with different in-plane spatial resolutions ranging from 1.25 ×1.11 mm^2 to 1.41 ×1.41 mm^2. To investigate the effect of small variation in spatial resolution, we down-

sampled the MR and CT slices to the same spatial resolution (1.41×1.41 mm^2) and the same dimension (256x256) with zero-padding, and re-retrained our models with updated images. No significant difference in MAE metrics was observed between the models trained with the same spatial resolution and the models trained with different spatial resolution. This suggests that our proposed models are robust to small resolution variations in our patient cohort. In our view, having small variations in the spatial resolution is comparable to having variations in the patient size. A robust model should be less sensitive to this type of variation.

A number of techniques could be investigated for improving model performance. Multiple MR images acquired with different sequences, like Dixon and UTE sequences, could be fed into models to provide more information for distinguishing different tissues, particularly air and bone. Nie *et al*. showed that introducing an additional adversarial discriminator improved overall sCT quality[57]. The same approach could be adapted in our proposed 2D and 3D CNN models. Non-rigid deformation could also be applied to both CT and MR images in the process of the on-the-fly data augmentation to produce more training pairs[23]. As more powerful computing hardware becomes more widely available, 3D models with deeper layers, larger training batch sizes, and images without down-sampling can be explored for possible model performance improvement.

2.5 Conclusion

We presented 2D and 3D CNNs for generating a pelvic sCT image from a T1-weighted MR image. In our study, both models successfully generated accurate sCT images for all 20 patients, with maximum MAEs of 56.5 HU and 53.1 HU for the 2D and 3D models, respectively. Statistical results of 20 patients showed that the 3D model generated sCT images with better

MAE and bone region precision compared with the 2D model and Han's model. Patient alignment tests indicated sCT images generated by the proposed models can provide accurate patient positioning using cone-beam CT based alignment. The fast speed and accurate HU mapping of the proposed 2D and 3D CNNs make them promising tools for generating pelvic sCT images for MR-only radiation therapy. Future work on dose calculation comparisons between the CT and sCT images is required before clinical implementation.

3 GENERATION OF ABDOMINAL SYNTHETIC CT_S FROM 0.35T MR IMAGES USING GENERATIVE ADVERSARIAL NETWORKS FOR MR-ONLY LIVER RADIATION THERAPY[41]

3.1 Introduction

The superior soft-tissue contrast of MRI, compared with that of CT, allows better tumor and healthy tissue differentiation in certain body areas, such as the brain, pelvis, and abdomen[4,81]. MR images are often acquired for tumor and OAR delineations in treatment planning workflows for pelvic or abdominal cancer radiation therapy[82–85]. Since there is no direct relationship between MR intensity values and electron densities, the standard MRgRT workflow still requires

the acquisition of a CT image for dose calculation. However, registration between CT and MR images for transferring target delineations introduces systematic uncertainties that propagate throughout the treatment[13]. Acquiring an additional CT image also increases unwanted radiation exposure, clinical workload, and financial cost[12]. MR-only radiation therapy can avoid these downsides.

A few methods have been proposed to generate sCT images from MR images. These methods include atlas-based methods, voxel-based methods, and hybrid methods[13]. In atlas-based methods[15,17], the target MR image was first deformably registered to atlas-MR images to acquire deformation vector fields. The acquired vector fields were then reversely applied on the atlas-CT images which were registered to atlas-MR images to generate the sCT image. Atlas-based approaches may not only take a long time to generate the sCT image but also fail if the target patient has substantially different anatomy compared with atlas-patients. Voxel-based methods used machine learning methods that were trained to covert voxel intensities of a single or multiple MR images to CT HUs[20,79]. Hybrid methods combined elements of voxel-based and atlas-based approaches[55,86].

Recently, DL[22], a subset of machine learning, has drawn great research interests for sCT generation mainly due to its fast generation speed and high accuracy. Han proposed a 2D CNN that achieved accurate brain sCT generation[58]. A study reported that the proposed 2D CNN generated the most accurate pelvic sCT images compared with four atlas-based methods[76]. Fu *et al.* proposed a 3D CNN[40] that generated more accurate pelvic sCT images than Han's 2D CNN. GANs were shown to have better performance in image-to-image translation tasks compared with the corresponding CNNs[24,25]. Two popular GANs, cGAN and cycleGAN, were investigated for generating pelvic and brain sCT images, respectively[87,88]. Results demonstrated that cGAN

could generate accurate pelvic images, and cycleGAN could generate brain sCT images. The pelvic sCT images generated by the cGAN achieved accurate dose calculation for pelvic radiation therapy[88]. A 3D patch-based dense cycleGAN proposed by Lei *et al.*[89] could generate accurate brain and pelvic sCT images. Another study showed that this model could also generate accurate abdominal sCT images and demonstrated its potential for MR-only liver SRBT planning[90]. Unlike cGANs, which require co-registered MR-CT image pairs for training, cycleGANs can be trained in an unsupervised manner. This could potentially enlarge the amount of data available for training cycleGANs. So far, no direct comparison of cGAN and cycleGAN for abdominal sCT generation has been made.

Although most studies showed that DL methods achieved promising performance in generating brain and pelvic sCT images, few studies on the application of DL methods to abdominal sCT generation have been published. Larger intra-scan and inter-patient anatomical variations, compared with those in the brain or pelvis, introduce significant challenges in the task of abdominal sCT generation. Interest in low-field MRgRT has grown rapidly in recent years. However, to our knowledge, no DL methods have been investigated for generating sCT images from low-field MR images. Compared with high-field MR images, low-field MR images have lower signal-to-noise ratios and more image artifacts caused by lower magnetic field homogeneity. This may result in poor image quality of the sCT images generated by DL methods.

This study provides the first investigation on applying DL methods for generating abdominal sCT images from low-field MR images in support of MR-only liver radiation therapy. We trained cGANs and cycleGANs to generate sCT images from 0.35 T abdominal MR images.

sCT HU accuracy was evaluated using voxel-based metrics. sCT-based dose calculation accuracy was also evaluated for liver cancer patients.

3.2 Materials and methods

3.2.1 Dataset

This study was conducted using 12 abdominal cancer patients (8 liver and 4 non-liver) who underwent MRgRT. Table 3-1 summarizes the patient characteristics and prescribed doses. All patients had MR and CT images acquired before the treatment. The average absolute time difference between acquisitions of MR and CT images is 61 mins. MR images were acquired with a true fast imaging with steady-state precession (TrueFISP) sequence on a 0.35T MRI scanner of the MRIdian system (ViewRay, OH, USA) during 25 s breath hold. MR slice thickness was 3 mm and in-plane resolution was 1.5 × 1.5 mm^2. Breath-hold CT images were acquired on a 16-slice CT scanner (Sensation Open, Siemens Medical Solutions, Erlangen, Germany) using 120 kVp and 360 mA. CT slice thickness was 1.5 mm and in-plane resolution was 0.98 × 0.98 mm^2. The target and OARs were delineated by radiation oncologists and medical physicists on the MR images. CT images were deformably aligned to MR images in the MRIdian treatment planning system to create deformed CT (dCT) images for treatment planning. Experienced physicists assessed the quality of the deformable registration in fused image display. No large local changes of the Jacobian determinants were observed.

3.2.2 GANs

We implemented two GANs, cGAN and cycleGAN, for abdominal sCT generation. Figure 3-1 shows the simplified architecture of the two networks.

Tumor location	Age	Gender	Tumor volume [cc]	Total dose [Gy]	No. of fraction
Liver	36	Female	23.6	45	3
	43	Male	44.1	60	3
	47	Female	1493.4	42	15
	54	Female	56.3	40	5
	54	Female	981.4	50	10
	55	Male	19.3	50	10
	58	Female	13.5	54	3
	71	Female	31.5	54	3
Pancreas	64	Male	25.7	40	8
Adrenal gland	71	Female	101.6	50	5
Middle abdomen	60	Male	534.7	50.4	28
	70	Female	132.6	40	20

Table 3-1: Patient characteristics and prescribed doses

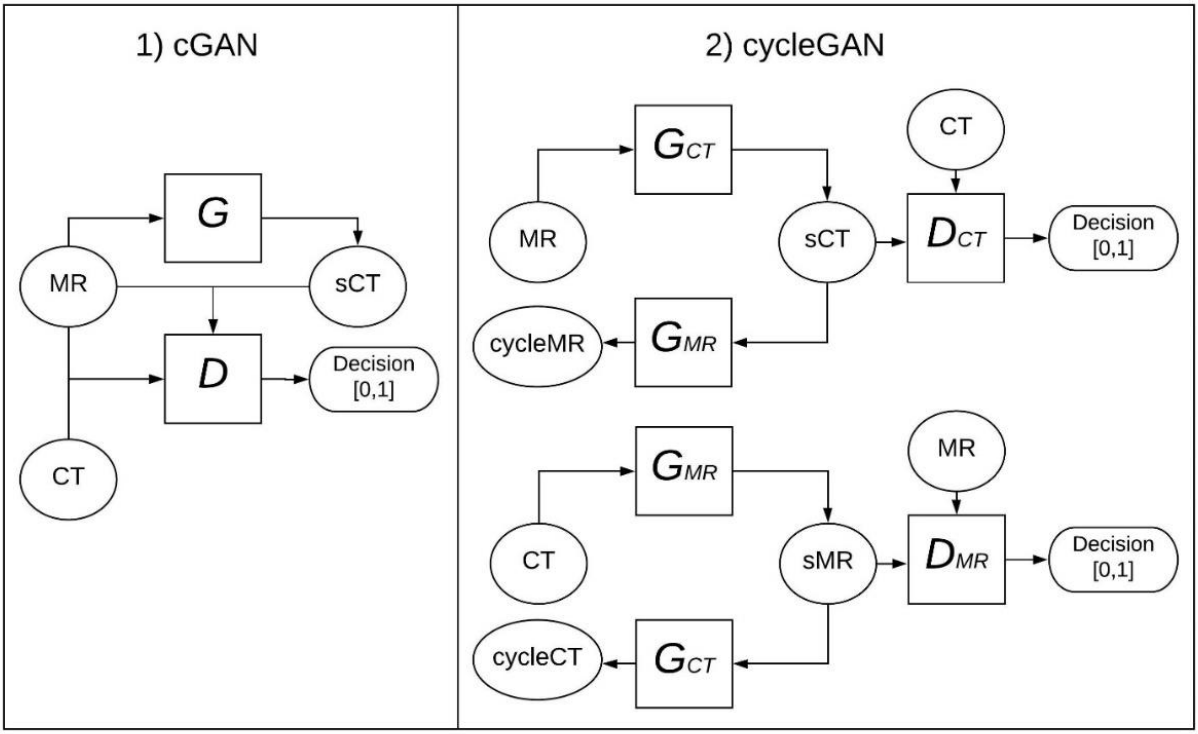

Figure 3-1: Simplified view of cGAN and cycleGAN architectures

The cGAN consisted of two DLs: a generator (*G*) and a discriminator (*D*). This network required paired MR and CT slices for training. *G* was trained to convert MR slices to sCT slices, while *D* was trained to distinguish the concatenated CT-MR slices from the concatenated sCT-MR slices. The adversarial goal was to generate sCT slices which not only had small L1 distance from CT slices but also could fool *D*. The *G* and *D* were trained by minimizing the losses

$$L_G = \lambda_{cGAN}\mathbb{E}_{MR,CT}[\| CT - G(MR) \|_1] - \mathbb{E}_{MR}[\log D(MR, G(MR))] \text{ and}$$

$$L_D = -\mathbb{E}_{MR,CT}[\log(D(MR, CT))] - \mathbb{E}_{MR}[\log(1 - D(MR, G(MR)))]$$

Equation 3-1

, respectively. λ_{cGAN} is the L1 loss regularization parameter.

The cycleGAN consisted of four CNNs: two generators (G_{CT} and G_{MR}) and two discriminators (D_{CT} and D_{MR}). G_{CT} (G_{MR}) was trained to convert MR (CT) slices to sCT (sMR) slices, and convert generated sMR (sCT) slices back to cycleCT (cycleMR) slices. D_{CT} (D_{MR}) was trained to distinguish real CT (MR) slices from sCT (sMR) slices. Unlike the cGAN, the cycleGAN was designed for unsupervised learning, i.e. training with unpaired MR and CT slices in this case. As L1 distance between unpaired CT and sCT slices is not valid, the adversarial goal is to generate cycleCT (cycleMR) slices that had small L1 distance from CT (MR) slices. The generators and discriminators were trained by minimizing the losses

$$L_{G_{CT}} = \lambda_{cycleGAN}(\mathbb{E}_{MR}[\| MR - cycleMR \|_1] + \mathbb{E}_{CT}[\| CT - cycleCT \|_1]) + \mathbb{E}_{MR}[(1 - D_{CT}(sCT))^2],$$

$$L_{G_{MR}} = \lambda_{cycleGAN}(\mathbb{E}_{MR}[\| MR - cycleMR \|_1] + \mathbb{E}_{CT}[\| CT - cycleCT \|_1) + \mathbb{E}_{CT}[\left(1 - D_{MR}(sMR)\right)^2],$$

$$L_{D_{CT}} = \mathbb{E}_{CT}[(1 - D_{CT}(CT))^2] + \mathbb{E}_{MR}[D_{CT}(sCT)^2], \text{ and}$$

$$L_{D_{MR}} = \mathbb{E}_{MR}[(1 - D_{MR}(MR))^2] + \mathbb{E}_{CT}[D_{MR}(sMR)^2].$$

Equation 3-2

$\lambda_{cycleGAN}$ is the L1 loss regularization parameter.

Both cGAN and cycleGAN could be trained to convert MR slices to sCT slices. Since our main goal is to test the feasibility of generating abdominal sCT images using GANs, we implemented the same network architectures presented by Isola *et al.*[24] and Zhu *et al.*[25]. The networks were modified to process and generate 16-bit single channel images.

3.2.3 sCT generation

N4 bias field correction was applied to all MR images to remove inhomogeneity artifacts[62]. Histogram-based intensity normalization was then performed to minimize the inter-patient MR intensity variation[63]. MR voxel intensities were clipped within the interval [0, 99th percentile], and dCT voxel intensities were clipped within the interval [-1000,1200] HU.

Four-fold cross-validation testing was conducted to generate sCT images for all 12 patients. The patient cohort was randomly divided into four groups. Three groups of 3 patients were used to train the network, the trained network was then applied on the MR images of the patients in the remaining group to generate their sCT images. The cGAN was trained with paired transverse MR and dCT slices, while the cycleGAN was trained with unpaired transverse MR and dCT slices. We adopted the same training protocols presented by Isola *et al.*[24] and Zhu *et al.*[25] for training cGAN and cycleGAN, respectively. Both models were implemented using Tensorflow packages[72] (V1.3.0, Python 2.7, CUDA 8.0) on Ubuntu 16.04 LTS system, and trained for 200 epochs with a batch size of 1 on a GeForce GTX 1080 Ti GPU (NIVIDIA, California, USA). The L1 loss regularization parameters were set as 100 for training.

3.2.4 Model evaluation

dCT and sCT image similarity was evaluated using mean absolute error (MAE) and peak-signal-to-noise-ratio (PSNR) within the MR body contour.

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |I_{sCT}(i) - I_{dCT}(i)| \text{ and}$$

$$PSNR = 20log_{10} \frac{4095}{\sqrt{\frac{\sum_{i=1}^{N}\left(I_{sCT}(i)-I_{dCT}(i)\right)^2}{N}}},$$

Equation 3-3

where N is the number of voxels inside the MR body contour; $I_{sCT}(i)$ and $I_{dCT}(i)_i$ represent the HU values of the ith voxel in the sCT and dCT, respectively; 4095, 2^{12}-1, is the maximum fluctuation in the 12-bit CT image. In general, lower MAE values and higher PSNR values indicate higher HU prediction accuracy.

Dosimetric evaluation was conducted using clinical plans from 8 liver patients. All plans were optimized on the dCT images according to the clinical guideline using the MRIdian treatment planning system. Dose distributions were calculated using the planning system's Monte Carlo algorithm with magnetic field corrections included. Clinical plans were copied to the corresponding sCT images, and the dose was recalculated using the same calculation protocol. dCT-based and sCT-based dose distributions were compared by gamma analysis[91] at 2%, 2mm and 3%, and 3mm within the volumes receiving at least 30%, 60%, and 90% of the prescribed dose. Mean dose and other clinically relevant dose-volume histogram (DVH) metrics were evaluated for the planning target volume (PTV) and OARs. Percentage differences ($\frac{D_{sCT}-D_{dCT}}{D\_prescribed}$) between these metrics calculated with sCT and dCT plans were computed.

Wilcoxon signed-rank tests[75] were conducted to test if the differences between evaluation metrics of cGAN and cycleGAN were significant. A *p*-value less than 0.05 was considered statistically significant.

3.3 Results

It took about 3 (15) hours to train an individual cross-validation cGAN (cycleGAN). On average, the time required for generating the sCT image of one patient was about 6 s using either model. Figure 3-2 shows the generated sCT slices along with corresponding MR and dCT slices from 3 representative liver cancer patients. Visual inspection reveals that sCT images generated by the cycleGAN are sharper than those generated by the cGAN. Both models achieved adequate performance in predicting HU values of air pockets, vertebral bodies, and soft tissues but had difficulties in reproducing ribs.

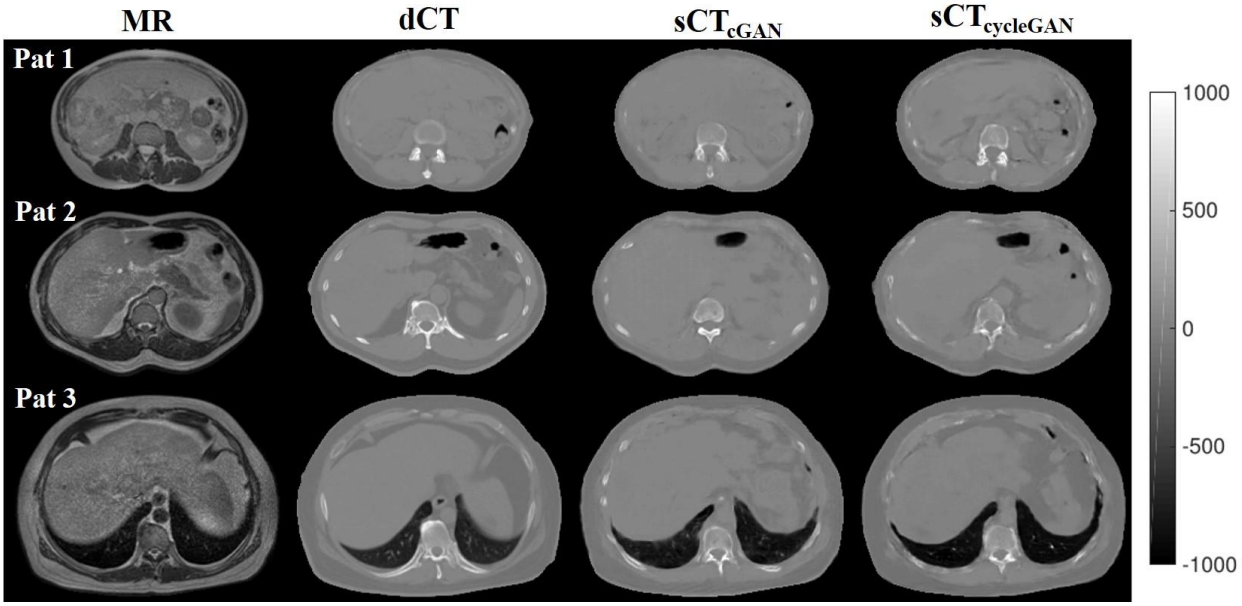

Figure 3-2: Transverse slices of the MR, dCT, sCT_{cGAN}, and $sCT_{cycleGAN}$ images from 3 liver cancer patients. The gray scale bar indicates the HU scale of the CT slices.

For all 12 abdominal cancer patients, MAEs and PSNRs between dCT and sCT images were computed using Equation 3-3. The statistical results are summarized in Table 3-2. On average, cGAN achieved smaller MAE and higher PNSR compared with the cycleGAN. The small patient number resulted in large standard deviations. The *p*-values of the corresponding Wilcoxon signed-rank tests were greater than 0.05.

	cGAN	cycleGAN	*p*-value
MAE [HU]	89.8±18.7	94.1±30.0	0.97
PSNR [dB]	27.4±1.6	27.2±2.2	0.62

Table 3-2: Statistics of MAE and PNSR between dCT and sCT images generated by cGAN or cycleGAN. Results were averaged across all 12 patients and shown in (mean ± SD) format. The *p*-values of the Wilcoxon signed-rank tests are shown.

For 8 liver cancer patients, the clinical plans optimized on dCT images were recalculated with the corresponding sCT_{cGAN} and $sCT_{cycleGAN}$ images, respectively. Figure 3-3 shows transverse slices of the dCT-based, sCT_{cGAN}-based, and $sCT_{cycleGAN}$ dose distributions along with corresponding difference maps for one liver patient. Both sCT images yielded dose distributions that were very similar to those calculated with the dCT image. The dCT-based and sCT-based dose distributions were compared using gamma analysis for the two models. As shown in Table 3-3, the average gamma passing rates within all evaluated volumes (receiving the dose greater than 30%, 60%, and 90% of the prescribed dose) were above 95% using 2%, 2 mm criterion, and 99% using a 3%, 3 mm criterion in both models. The sCT_{cGAN} plan achieved higher average gamma passing rates using a 2%, 2m criterion than the $sCT_{cycleGAN}$ plan. The *p*-values of the corresponding Wilcoxon signed-rank tests were greater than 0.05.

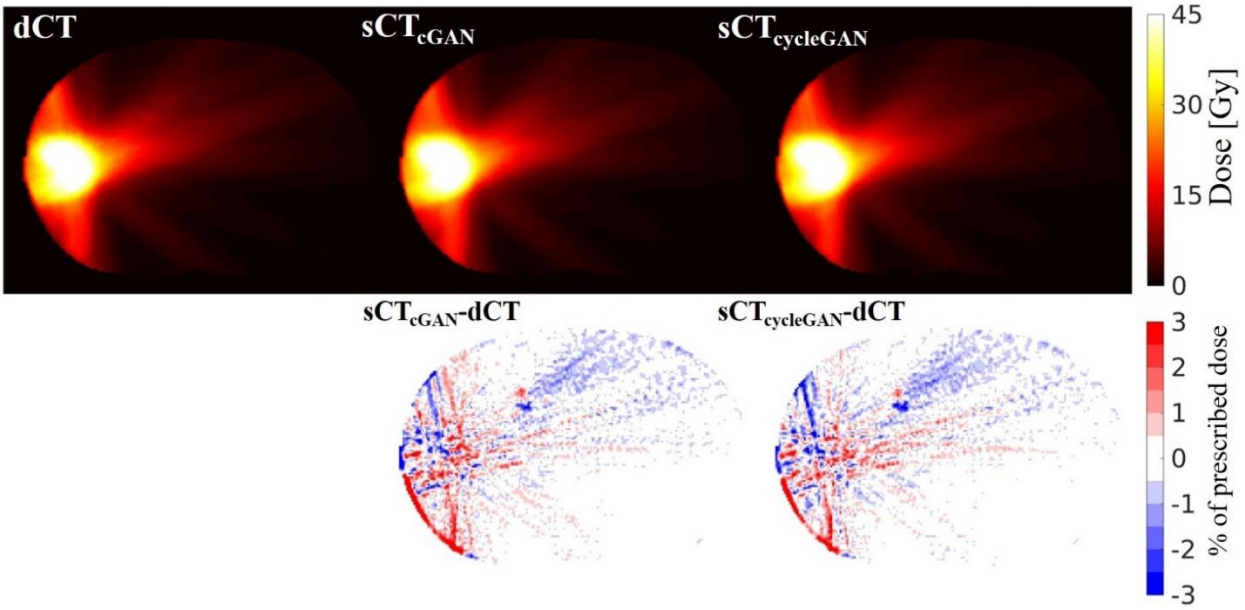

Figure 3-3: Comparison of dCT-based, sCT$_{cGAN}$-based, and sCT$_{cycleGAN}$-based dose distribution. The first column shows the transverse slices of three dose distributions for one liver cancer patient, and the corresponding dose difference maps between sCT and CT are presented in the second column as the percentage of the prescribed dose (45 Gy).

Gamma passing rate	Volume of interest	cGAN	cycleGAN	*p*-value
$\gamma_{3\%,3mm}$ [%]	D ≥ 30%	99.5±0.8	99.5±0.7	0.95
	D ≥ 60%	99.6±0.7	99.6±0.9	0.69
	D ≥ 90%	99.5±1.1	99.3±1.3	0.38
$\gamma_{2\%,2mm}$ [%]	D ≥ 30%	98.7±1.5	98.5±1.6	0.84
	D ≥ 60%	98.4±2.2	97.6±2.8	0.31
	D ≥ 90%	97.4±3.2	95.6±5.0	0.31

Table 3-3: Statistics of gamma passing rates within the volumes of interest. Results were averaged across 8 liver cancer patients and shown in (mean ± SD) format. The *p*-values of the Wilcoxon signed-rank tests are shown.

Mean doses and clinically relevant DVH metrics for the PTV and OARs were computed for dCT and sCT plans. Deviations of these metrics between dCT and sCT plans are shown in

Table **3-4**. In both models, the average deviations of all metrics were small, within ±0.6% for the PTV and within ±0.15% for all evaluated OARs. Figure 3-4 presents mean dose differences of the PTV and OARs for all 8 liver cancer patients. The maximum absolute differences of PTV mean doses were 0.4% for the cGAN and 1.0% for the cycleGAN. The cGAN achieved smaller deviation ranges (maximum-minimum) than the cycleGAN for all evaluated regions except the right kidney. The *p*-value of all Wilcoxon signed-rank tests are greater than 0.05, except the one for bowel mean dose.

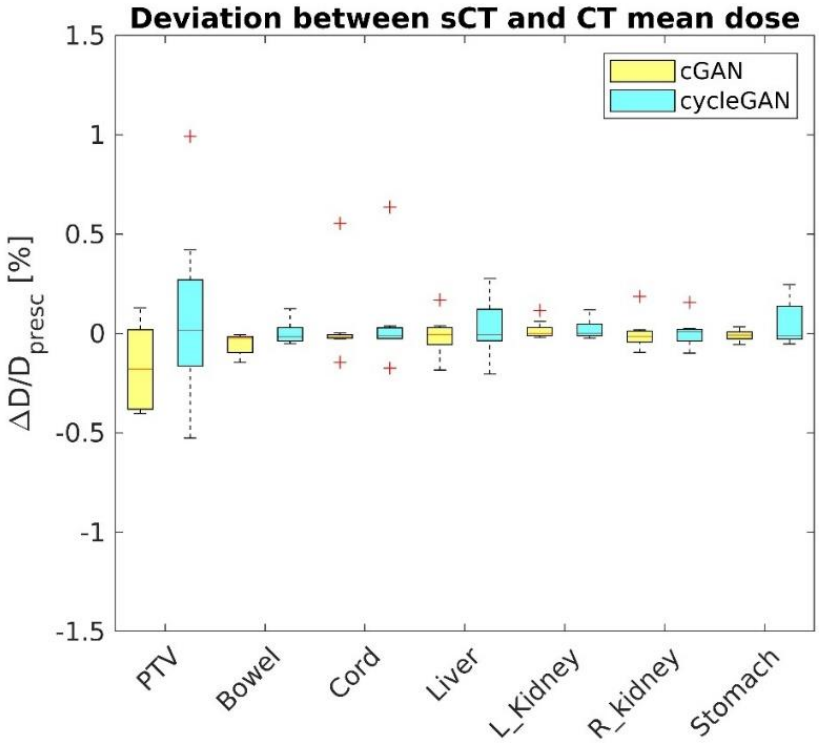

Figure 3-4: Box and whisker plot of deviations between sCT and CT mean dose within the PTV and OARs. The maximum (top line), 75% (top of box), median (central line), 25% (bottom of box), and minimum (bottom line) are shown. Outliers are drawn as red cross signs. cGAN and cycleGAN results are presented in yellow and cyan, respectively.

Regions	Metric	cGAN Deviation sCT vs dCT (% of prescribed dose or volume percentage difference)	cycleGAN Deviation sCT vs dCT (% of prescribed dose or volume percentage difference)	*p*-value
PTV	Mean	-0.17±0.22	0.09±0.46	0.08
	$D_{98\%}$	-0.30±0.31	-0.06±0.56	0.08
	$D_{95\%}$	-0.51±0.52	-0.09±0.94	0.08
	$D_{50\%}$	-0.27±0.48	0.19±0.81	0.11
	$D_{2\%}$	-0.39±0.63	0.17±0.99	0.11
Bowel	Mean	-0.05±0.05	0.00±0.06	0.04*
	V_{35Gy}	-0.05±0.15	-0.02±0.05	1.00
Cord	Mean	0.04±0.21	0.06±0.24	0.74
	Max	-0.03±0.29	0.01±0.17	0.84
Liver	Mean	-0.01±0.10	0.03±0.15	0.31
	D_{1000cc}	0.02±0.08	0.06±0.12	0.16
Left kidney	Mean	0.02±0.05	0.02±0.05	0.55
	Max	-0.08±0.14	-0.03±0.20	0.20
Right kidney	Mean	0.00±0.08	0.01±0.07	0.84
	Max	-0.03±0.30	0.07±0.34	0.55
Stomach	Mean	-0.01±0.03	0.05±0.12	0.20
	V_{35Gy}	0.03±0.09	0.15±0.43	1.00

* indicates statistically significant.

Table 3-4: Statistics of metric differences between between dCT and sCT plans. Differences are presented in percentage of prescribed dose (mean, maximum, $D_{2\%}$, $D_{50\%}$, $D_{95\%}$, $D_{98\%}$, D_{1000cc}) or volume percentage difference (V_{35Gy}). DVH metrics were chosen based on planning constraints for MR-guided stereotactic body radiation therapy (SBRT) requested by physicians. Results were averaged across 8 liver cancer patients and shown in (mean ± SD) format. The *p*-values of the Wilcoxon signed-rank tests are shown.

3.4 Discussion

In this study, for the first time, DL methods have been applied for generating abdominal sCTs images from low-field MR images. We trained two DL methods, cGAN and cycleGAN, to generate sCT images from 0.35T MR images for 12 abdominal cancer patients. sCT HU accuracy was evaluated using voxel-wise metrics. To evaluate the sCT dose calculation accuracy for liver radiation therapy, we compared the dCT-based and sCT-based dose distributions for 8 liver cancer patients. In both models, the average MAE between dCT and sCT images is of similar magnitude to that previously reported for sCT abdominal generation using high-field MR images[90]. The average gamma passing rates were above 99% using a 3%, 3 mm criterion in both models. Small deviations in the mean dose and clinically relevant DVH metrics between sCT- and dCT-based dose distributions were observed in both models as shown in
Table **3-4**. These results suggested that abdominal sCT images generated by either cGAN or cycleGAN achieved accurate dose calculation for liver radiotherapy planning in our patient cohort.

Although small differences between CT and sCT images were visible as shown in Figure 3-2, sCT images generated by both methods achieved accurate dose calculations. This may be caused by the relative insensitivity of photon dosimetry to small attenuation differences. Our results showed that sCTcGAN images had smaller average MAEs and higher average gamma passing rates (using a 2%, 2mm criterion) than sCTcycleGAN images. However, statistical tests suggested that no significant difference was observed between the two models in terms of all evaluation metrics except bowel mean dose. More patients are required to further compare the two models. Small dose calculation errors resulting from HU prediction errors, as shown in

Table 3-4, may be less significant compared with those introduced by other factors including patient setup errors and target delineation variations.

Several factors may contribute to differences between CT and sCT images. First, the model performance was limited by the small training dataset. For example, a DL method trained with patients having only small inter-patient anatomical variations may have difficulty in providing accurate HU prediction for patients with atypical anatomies. A larger training dataset may lead to more accurate and robust model performance. Second, the imperfect registration between MR and dCT images might introduce voxel mismatch and hence intensity mapping errors, which led to the perturbation of training the cGAN. The cycleGAN would not be affected by these mapping errors as it was not trained with paired MR and dCT images. However, unlike the cGAN which was trained using the L1 loss and adversarial loss between paired dCT and sCT images, cycleGAN was trained using unsupervised scheme where only adversarial loss between unpaired dCT and sCT images was computed. This looser loss constraint on sCT images for model training may lead to degraded sCT quality.

The average sCT generation time is less than 10 s using either cGAN or cycleGAN. Generation time can be affected by several factors including image dimension and GPU model. A phase I trial study suggested that oaMRgRT allowed PTV dose escalation and simultaneous OAR sparing compared with non-adaptive treatment[11]. The fast sCT generation speed of cGAN and cycleGAN makes it possible to achieve MR-only online adaptive workflow without extensively elongating the time required to adapt.

DL methods investigated in this work can also be trained to convert high-field MR images to sCT images. Using high-field MR images may result in better sCT quality since high-field MR images have higher signal-to-noise ratios (SNRs) and fewer image artifacts related to

the low magnetic homogeneity than low-field MR images. Future work includes acquiring high-field MR images and investigating the dose calculation accuracy of the sCT images generated from high-field MR images using other commercial treatment planning systems.

3.5 Conclusion

We implemented cGAN and cycleGAN to generate abdominal sCT images from 0.35T MR images. In this preliminary study, sCT images generated by both models enabled accurate dose calculations for liver radiotherapy planning. The fast generation speed and high dose calculation accuracy make both GANs promising tools for MR-only liver radiation therapy. More abdominal cancer patients will be enrolled in the future to further compare the dose calculation accuracy of the sCT images generated by cGAN and cycleGAN.

4 3D MULTI-PATH DENSENET FOR IMPROVING AUTOMATIC SEGMENTATION OF GLIOBLASTOMA ON PRE-OPERATIVE MULTI-MODAL MR IMAGES[42]

4.1 Introduction

Gliomas are tumors arising from glial cells, normally astrocytes and oligodendrocytes. Gliomas account for approximately 26% of all brain tumors and can be classified as grades I-IV based on histological characteristics[92,93]. GBMs, grade IV gliomas, are the most common malignant primary brain tumors with a median overall survival (OS) of only 15 months after diagnosis[94,95]. The gold standard treatment for GBM is a maximal safe resection followed by radiation therapy

with or without concurrent adjuvant chemotherapy[96,97]. As intensity-modulated radiation therapy (IMRT) and volumetric modulated arc therapy (VMAT) can deliver a high dose of radiation to the target, while providing better dose sparing of normal tissues compared with 3D conformal radiation therapy, they have been increasingly used for treating GBM.

Accurate target delineation is critical for the IMRT and VMAT treatment planning because both techniques have sharp dose gradients between the target and normal tissues. The Radiation Therapy Oncology Group trial recommends using multi-modal MR images, including a T2-weighted (T2w) images or a fluid-attenuated inversion recovery (FLAIR) image and a contrast-enhancing T1-weighted (CE-T1w) image, for GBM target delineation[98]. Manual segmentation is not only time-consuming but also sensitive to intra-observer and inter-observer variabilities. Hence, it is essential to develop automatic segmentation methods that can perform highly reproducible and accurate GBM segmentation.

Recently, many CNNs have achieved good performance in glioma segmentation based on multi-modal MR images. An ensemble method, called EMMA[99], earned first place in the 2017 Brain Tumor Segmentation (BraTS) challenge[100–102]. EMMA consisted of seven 3D CNNs including three 3D fully convolutional networks[103], two 3D U-Nets[23], and two DeepMedic models[104]. Every 3D CNN in the EMMA was built based on the encoder-decoder architecture and could achieve the end-to-end mapping from multi-model MR images to tumor contour. A novel 3D CNN with autoencoder regularization earned first place in the 2018 BraTS challenge and also had the encoder-decoder architecture[105]. Zhang *et al.* proposed a 3D DenseNet, the 3D CNN with several dense blocks, for acute ischemic stroke segmentation and showed it achieved better performance than a 3D U-Net with residual connections[106]. The dense block was proposed by Hung *et al.* to alleviate the vanishing-gradient problem, strengthen feature propagation, and

encourage feature reuse[107]. It could reduce the number of model parameters and achieved better performance on several object recognition tasks compared with the residual block[70]. However, all of these proposed 3D CNNs only employed a single-path architecture where the same encoding feature extraction filters were applied to the concatenation of multi-modal MR images. We hypothesized that a multi-path architecture, where each MR image has its own set of encoding filters, could achieve better segmentation performance than the single-path architecture by capturing the image-specific features.

In this study, we proposed a 3D multi-path DenseNet for automatically generating the GBM tumor contour from four multi-modal MR images. A 3D single-path DenseNet was built for comparison. Both DenseNets were trained, validated, and tested using a total of 258 GBM patients. Several evaluation metrics were used to compare the ground truth and autosegmented contours. The model performance of the two DenseNets was compared using Wilcoxon signed-rank tests[75].

4.2 Materials and methods

4.2.1 Dataset

The 2019 BraTS challenge training set, comprised of images from 259 GBM patients, was used in this study. Each patient had four pre-operative multi-modal MR images: T1w, CE-T1w, T2w, and FLAIR images. These images were acquired with different scanners and clinical protocols from multiple institutions. Preprocessing steps of co-registration and skull-stripping were applied to all MR images[101]. The image voxel size is 1.0 x 1.0 x 1.0 mm^3, and the image matrix size is 240 x 240 x 155. Labels of three tumor subregions (enhancing tumor core, non-enhancing tumor core, and edema) were manually delineated by one to four raters based on the same annotation

protocol. Manual delineations were approved by expert board-certified neuroradiologists to define the ground truth labels. One patient was removed because a portion of the FLAIR image was cut off, which resulted in a total number of 258 patients in this study.

4.2.2 Image preprocessing

The manual ground truth tumor contour of each patient was acquired by fusing three tumor subregion labels. The N4ITK algorithm was applied to all MR images, except the FLAIR image, to correct intensity inhomogeneity[62]. To save computational memory, all images and contours were cropped to exclude the background margin and resampled to have an isotropic voxel size of 1.5 x 1.5 x 1.5 mm^3. The final matrix size of the images and contours was 100 x 128 x 105. For each MR image, voxel intensity was normalized to z-score using the mean and standard deviation of the intensities of its brain voxels. Figure 4-1 shows the transverse slices of four preprocessed MR images along with the ground truth contour for one example patient.

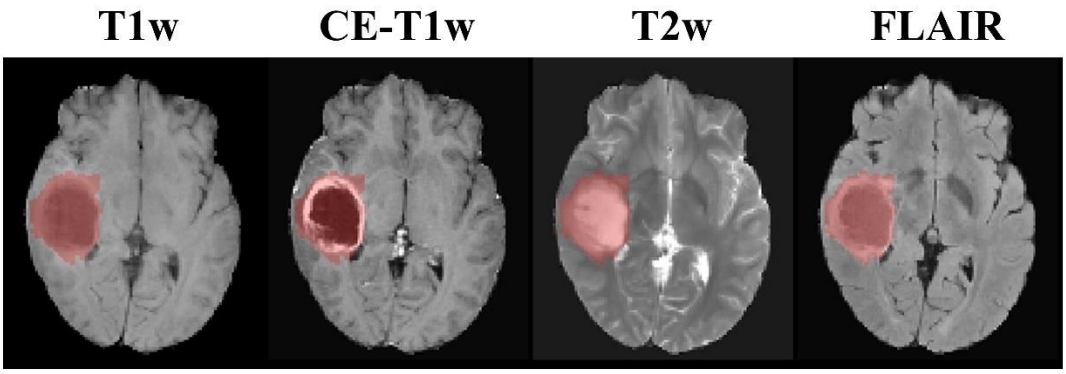

Figure 4-1: From left to right, transverse slices of the preprocessed T1w, CE-T1w, T2w, and FLAIR MR images, along with the ground truth tumor contour for one example patient. Z-score window [-4, 4] is used for image display.

4.2.3 3D CNNs

4.2.3.1 3D single-path DenseNet

Figure 4-2 shows the architecture of a 3-layer dense block used in the proposed 3D DenseNets. Each layer in the dense block contained one convolution layer with a filter size of 1 x 1 x 1 followed by one convolutional layer with a filter size of 3 x 3 x 3. The number of output feature maps after each 1 x 1 x 1 convolutional layer is the growth rate of the dense block. Instance normalization layers were used to reduce internal covariate shifts and speed up model optimization[108].

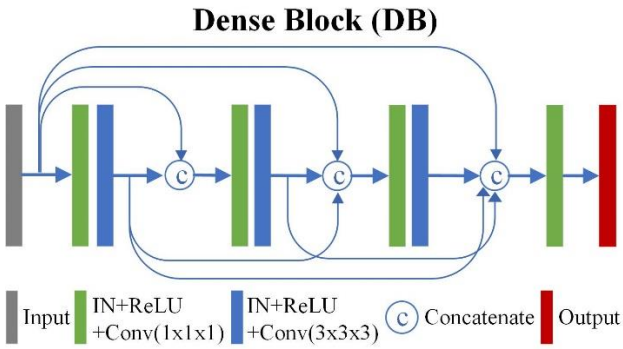

Figure 4-2: The architecture of the dense block. IN, instance normalization layer; Conv, convolutional layers. Color figure can be viewed online.

Figure 4-3 shows the architecture of the 3D single-path DenseNet for GBM segmentation. It contained 5 dense blocks forming an encoder-decoder architecture similar to a U-Net[23]. The encoder path extracted features from the concatenation of four MR images, while the decoder path gradually reconstructed the contour from the extracted features. Since a 3D MR image was represented using a 5D tensor with the shape of (batch size, depth, height, width, channels), the concatenation of four MR images means joining a sequence of four 5D tensors along the last dimension. Averaging pooling layers and deconvolutional layers were used to downsample and upsample the feature maps, respectively. At the end of the model, one

convolutional layer with a filter size of 1 x 1 x 1 followed by the Softmax layer[109] was used to generate the probability maps of background and tumor labels. The model can be trained to achieve an end-to-end mapping from the concatenation of four MR images to the autosegmented tumor contour.

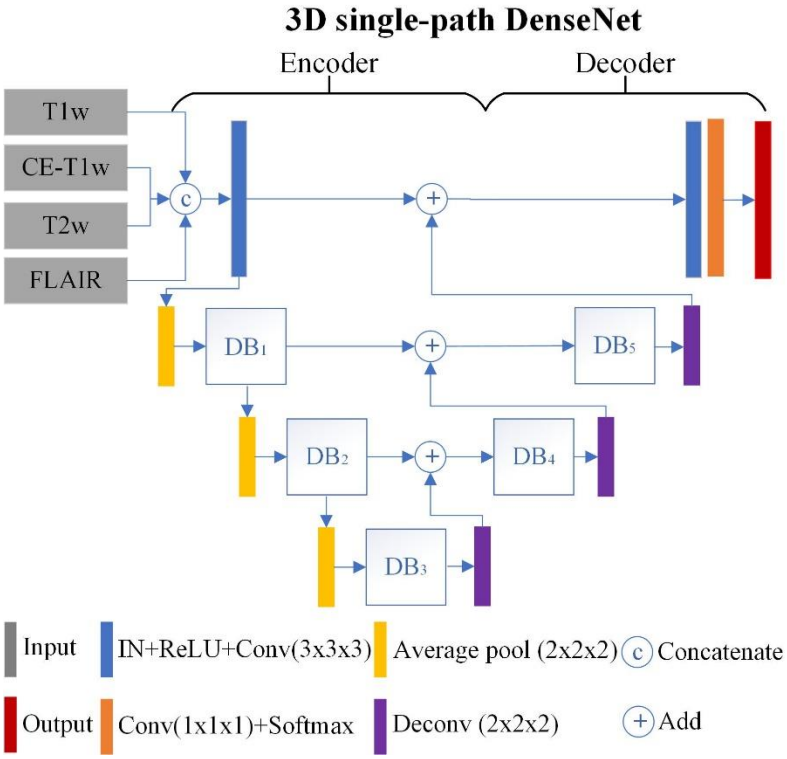

Figure 4-3: The architecture of the 3D single-path DenseNet. DB, dense block shown in Figure 2; IN, instance normalization layer; Conv, convolutional layers; Deconv, deconvolutional layer. Color figure can be viewed online.

4.2.3.2 3D multi-path DenseNet

Figure 4-4 (A) shows the architecture of the 3D multi-path DenseNet. It also had encoder and decoder paths. In contrast to the single-path DenseNet, where four MR images were concatenated and fed into the single encoder path, each MR image has its own encoder path in the multi-path DenseNet. The encoded feature maps from four different paths were concatenated and then fused by squeeze-and-excitation blocks (SEB) as shown in Figure 4-4 (B). Output

feature maps from the SEBs were fed into the same decoder path that was used in the single-path DenseNet. The SEB was proposed by Hu *et al.* to recalibrate the channel-wise feature response by modeling the inter-channel dependence[110]. Overall, the 3D multi-path DenseNet contains 14 dense blocks. In each SEB, the number of output feature maps after the convolutional layer and the number of nodes in two fully connected layers were set the same as the growth rate of the dense block.

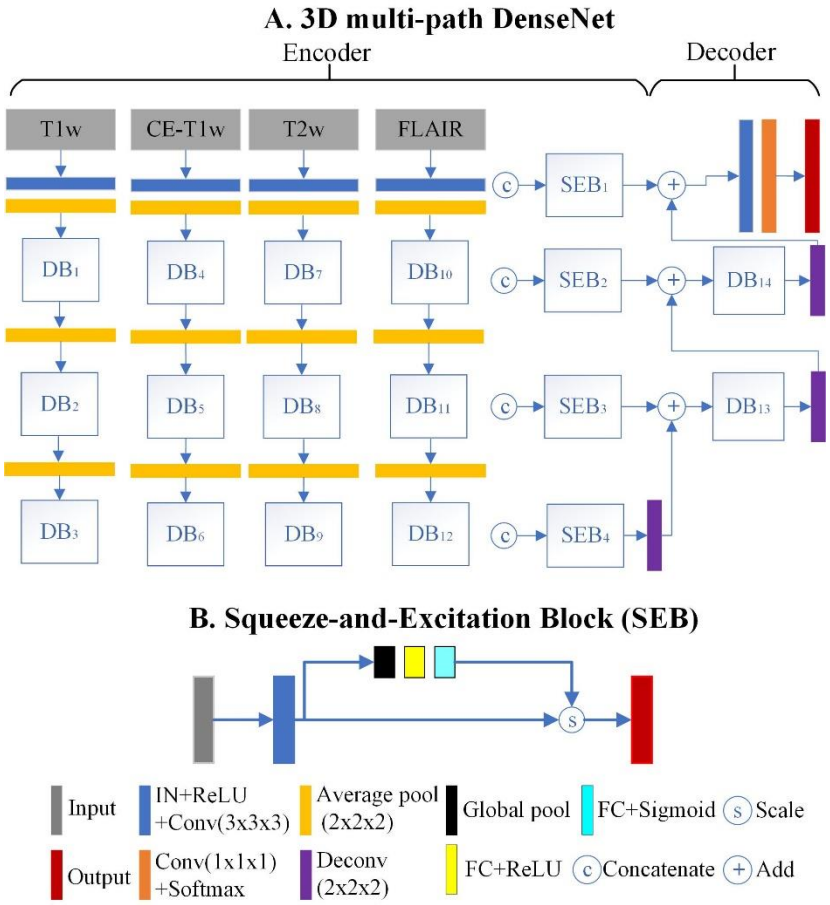

Figure 4-4: (a) The architecture of the 3D multi-path DenseNet. (b) The architecture of the squeeze-and-excitation blocks (SEB). DB, dense block shown in Figure 4-2; IN, instance normalization layer; Conv, convolutional layers; Deconv, deconvolutional layer; FC, fully connected layer. Color figure can be viewed online.

The number of trainable model parameters only depended on the growth rate of the dense block used in the model. The growth rate of the multi-path DenseNet was set as 16 based on

GPU memory limitations. The growth rate of the single-path DenseNet was set as 30 so that both DenseNets have a similar number of trainable parameters (about 0.46 million). Since the complexity of DL methods is normally measured by the number of trainable parameters, both DenseNets have similar complexity.

Tabel 4-1 and Table 4-2 show the layer configurations of the single-path DenseNet and multi-path DenseNet, respectively.

Layers	**Details**	**Output Size**
Convolution	3 x 3 x 3 conv, stride 1	128 x 105 x 100 x 30
Pooling	2 x 2 x 2 average pool, stride 2	64 x 53 x 50 x 30
DB_1	$\begin{bmatrix} 1 \times 1 \times 1 \text{ conv, stride } 1 \\ 3 \times 3 \times 3 \text{ conv, stride } 1 \end{bmatrix} \times 3$ 1 x 1 x 1 conv, stride 1	64 x 53 x 50 x 30
Pooling	2 x 2 x 2 average pool, stride 2	32 x 27 x 25 x 30
DB_2	Same as DB_1	32 x 27 x 25 x 30
Pooling	2 x 2 x 2 average pool, stride 2	16 x 14 x 13 x 30
DB_3	Same as DB_1	16 x 14 x 13 x 30
Deconvolution	2 x 2x 2 Deconv, stride 1/2	32 x 27 x 25 x 30
DB_4	Same as DB_1	32 x 27 x 25 x 30
Deconvolution	2 x 2x 2 Deconv, stride 1/2	64 x 53 x 50 x 30
DB_5	Same as DB_1	64 x 53 x 50 x 30
Deconvolution	2 x 2x 2 Deconv, stride 1/2	128 x 105 x 100 x 30
Convolution	3 x 3 x 3 conv, stride 1	128 x 105 x 100 x 30
Classification	1 x 1 x 1 Conv, stride 1 Softmax	128 x 105 x 100 x 2

Table 4-1: Layer configuration for the single-path DenseNet. Each "conv" layer shown in the table corresponds to the sequence IN-ReLU-Conv. Note that concatenations were not shown.

Layers	Details	Output Size
Convolution	3 x 3 x 3 conv, stride 1	128 x 105 x 100 x 16 (4 sets)
SEB_1	3 x 3 x 3 conv, stride 1 Global pool FC+ReLU FC+Sigmoid Scale	128 x 105 x 100 x 16 1 x 1 x 1 x 16 16 16 128 x 105 x 100 x 16
Pooling	2 x 2 x 2 average pool, stride 2	64 x 53 x 50 x 16 (4 sets)
DB_1, DB_4, DB_7, DB_{10}	$\begin{bmatrix} 1 \text{ x } 1 \text{ x } 1 \text{ conv, stride } 1 \\ 3 \text{ x } 3 \text{ x } 3 \text{ conv, stride } 1 \end{bmatrix}$ x 3 1 x 1 x 1 conv, stride 1	64 x 53 x 50 x 16 (4 sets)
SEB_2	Same as SEB_1	64 x 53 x 50 x 16
Pooling	2 x 2 x 2 average pool, stride 2	32 x 27 x 25 x 16 (4 sets)
DB_2, DB_5, DB_8, DB_{11}	Same as DB_1	32 x 27 x 25 x 16 (4 sets)
SEB_3	Same as SEB_1	32 x 27 x 25 x 16
Pooling	2 x 2 x 2 average pool, stride 2	16 x 14 x 13 x 16 (4 sets)
DB_3, DB_6, DB_9, DB_{12}	Same as DB_1	16 x 14 x 13 x 16 (4 sets)
SEB_4	Same as SEB_1	16 x 14 x 13 x 16
Deconvolution	2 x 2x 2 Deconv, stride 1/2	32 x 27 x 25 x 16
DB_{13}	Same as DB_1	32 x 27 x 25 x 16
Deconvolution	2 x 2x 2 Deconv, stride 1/2	64 x 53 x 50 x 16
DB_{14}	Same as DB_1	64 x 53 x 50 x 16
Deconvolution	2 x 2x 2 Deconv, stride 1/2	128 x 105 x 100 x 16
Convolution	3 x 3 x 3 conv, stride 1	128 x 105 x 100 x 16
Classification	1 x 1 x 1 Conv, stride 1 Softmax	128 x 105 x 100 x 2

Table 4-2: Layer configuration for the multi-path DenseNet. Each "conv" layer shown in the table corresponds to the sequence IN-ReLU-Conv. Note that concatenations were not shown.

4.2.4 Model training

The patient cohort was randomly split into a training set of 180 patients, a validation set of 39 patients, and a testing set of 39 patients. The Adam stochastic gradient descent method[73] was used to minimize the loss function,

$$loss = 1 - \sum_{i=1}^{N} \frac{2P_i \times L_i}{P_i + L_i},$$

Equation 4-1

where N is the number of image voxels, P_i is the Softmax probability[109] of the voxel i being a tumor voxel, and L_i is the ground truth tumor label (0: background, 1: tumor) of the voxel i.

Both DenseNets were implemented using the Tensorflow package (V1.10.0, Python 3.6.9, CUDA 10.0) and ran on an 11 GB GeForce GTX 1080 Ti. A batch size of 1 was used for training. The initial learning rate and the stopping epoch number were tuned using the validation set. For both DenseNets, the optimal initial learning rate and epoch number are 5×10^{-4} and 90, respectively.

4.2.5 Model evaluation

Trained models were applied to 39 testing patients to generate their autosegmented tumor contours. Model performance was evaluated using three metrics: DSC, average surface distance (ASD), and 95% Hausdorff distance ($HD_{95\%}$). These metrics are represented by the following equations:

$$DSC = \frac{2\,(V_{GT} \cap V_{Auto})}{V_{GT} \cup V_{Auto}},$$

$$ASD = \frac{1}{2}\left(\overline{\min_{x \in S_{GT}} d(x, S_{Auto})} + \overline{\min_{x \in S_{Auto}} d(x, S_{GT})}\right),$$

$$HD_{95\%} = \frac{1}{2}\left(K_{95}\left(\min_{x \in S_{GT}} d(x, S_{Auto})\right) + K_{95}\left(\min_{x \in S_{Auto}} d(x, S_{GT})\right)\right),$$

Equation 4-2

where V_{GT} and V_{Auto} refer to the volumes of the ground truth and autosegmented tumor contours, respectively; S_{GT} and S_{Auto} refer to the surfaces of the ground truth and autosegmented tumor contours, respectively; $\min_{x \in S_{GT}} d(x, S_{Auto})$ denotes the distance of the voxel *x*, on the tumor surface S_{GT}, to its closet voxel on the surface S_{Auto}; K_{95} refers to the 95th percentile of all distances.

Wilcoxon signed-rank tests[75] were conducted to compare the performance of the 3D single-path and multi-path DenseNets.

4.3 Results

Figure 4-5 shows the ground truth and autosegmented tumor contours for the three example patients. Autosegmented tumor contours generated by both DenseNets were similar to the corresponding ground truth tumor contour based on visual inspection. In Figure 4-5, white arrows point to the regions where there are larger differences between the ground truth and $Auto_{single\text{-}path}$ contours compared with those between ground truth and $Auto_{multi\text{-}path}$ contours.

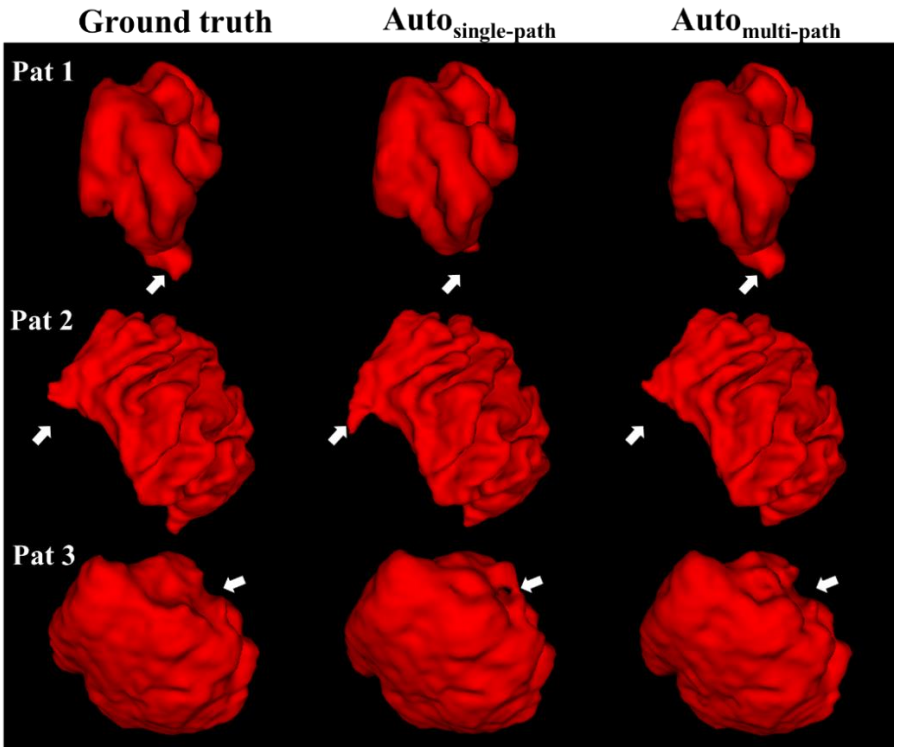

Figure 4-5: Ground truth tumor contours (left column) and the autosegmented tumor contours generated by the single-path DenseNet (middle column) and multi-path DenseNet (right columns) for the three example patients.

Table 4-3 summarizes the statistics of the evaluation metrics for the single-path and multi-path DenseNets. The multi-path DenseNet achieved a larger mean DSC of 0.922, a smaller mean ASD of 1.1 mm, and a smaller $HD_{95\%}$ of 3.9 mm compared with the single-path DenseNet. The *p*-values of all Wilcoxon signed-rank tests were less than 0.05, which indicates strong evidence against the null hypothesis that the median difference between the paired samples is 0.

Figure 4-6 shows box and whisker plots of the three evaluation metrics. The multi-path DenseNet generated more robust GBM tumor contours compared with the single-path DenseNet in terms of smaller box ranges (max-min) of all evaluation metrics.

Metric	Single-path DenseNet	Multi-path DenseNet	*p*-value
DSC	0.911±0.060	0.922±0.041	<0.001
ASD [mm]	1.3±0.7	1.1±0.5	0.002
$HD_{95\%}$ [mm]	5.2±7.1	3.9±3.3	0.046

Table 4-3: Statistics of DSC, ASD, and $HD_{95\%}$ between the ground truth contours and the autosegmented contours generated by the single-path DenseNet or multi-path DenseNet. Results were averaged across 39 testing patients and shown in (mean ± SD) format. The *p*-values of the Wilcoxon signed-rank tests are shown.

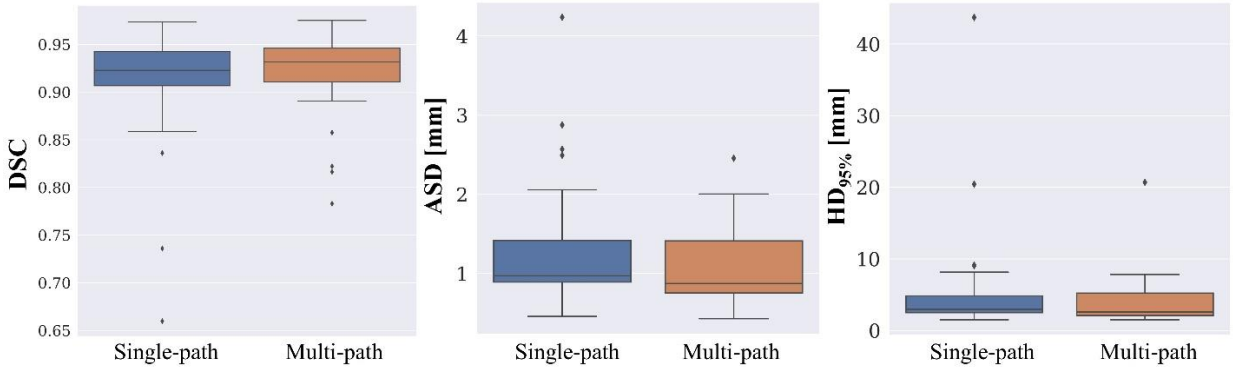

Figure 4-6: Box and whisker plots of DSC, ASD, and $HD_{95\%}$ for the single-path and multi-path DenseNets. The maximum (top line), 75th percentile (top of the box), median (central line), 25th percentile (bottom of the box), and minimum (bottom line) are shown. Outliers are drawn as diamond signs.

4.4 Discussion

In this study, we proposed a 3D multi-path DenseNet for generating the GBM tumor contour from four MR images. A 3D single-path DenseNet was also built for comparison. Both DenseNets were trained, validated, and tested using 180, 39, and 39 GBM patients, respectively. Autosegmented contours generated by both DenseNets were compared with the ground truth contours using DSC, ASD, and $HD_{95\%}$.

The multi-path architecture achieved better performance in GBM segmentation than the corresponding single-path architecture. The multi-path DenseNet achieved a larger mean DSC, a smaller mean ASD, and a smaller mean $HD_{95\%}$ compared with the single-path DenseNet. Results

of Wilcoxon signed-rank tests indicated significant differences in all three metrics. The autosegmented contours generated by the multi-path DenseNet were generally qualitatively more similar to the ground truth contours than those generated by the single-path DenseNet, as is illustrated by the examples in Figure 4-5. Figure 4-6 showed that the multi-path DenseNet achieved more robust segmentation compared with the single-path DenseNet.

The multi-path DenseNet achieved a mean DSC of 0.922 averaged across 39 testing patients. The proposed DenseNet could also be trained to generate the labels of three tumor subregions by changing the output channel number in the last convolutional layer to 4. We trained the multi-path DenseNet using average Dice loss for tumor subregion segmentation. The trained model achieved a mean DSC of 0.833 for enhancing tumor core, 0.876 for tumor core (enhancing and non-enhancing tumor core), and 0.913 for the whole tumor. Another approach for subregion segmentation is to train the CNN three times. Each CNN is trained to generate the label of one subregion or one fused subregion (tumor core or whole tumor).

The images and ground truth tumor contours were downsampled from the original voxel size of 1.0 x 1.0 x 1.0 mm^3 to the voxel size of 1.5 x 1.5 x 1.5 mm^3 to save computational memory. We upsampled the autosegmented contours generated by 3D DenseNets back to the original spatial resolution and compared them with original ground truth tumor contours. In this case, the single-path DenseNet achieved a mean DSC of 0.900, while the multi-path DenseNet achieved a mean DSC of 0.910 for GBM segmentation. The Wilcoxon signed-rank test suggested a significant difference (p-value<0.001). The EMMA achieved a mean DSC of 0.886 in the 2017 BraTS challenge[99]. The 3D CNN with autoencoder regularization achieved a mean DSC of 0.884 in the 2018 BraTS challenge[105]. A two-stage cascaded U-Net achieved a mean DSC of 0.888 in the 2019 BraTS challenge[111]. These results were included for a rough

comparison. It should be noted that a direct quantitative comparison is not appropriate because these models were trained and evaluated using different datasets.

We identified three outlier cases of the single-path model based on the DSC boxplot. After dropping these cases, the mean DSC results of the single-path and multi-path models are 0.924 and 0.929, respectively. The p-value of the Wilcoxon sign rank test is 0.03, which still indicates a significant difference. Similarly, we identified four outlier cases of the multi-path model based on the DSC boxplot. After dropping these cases, the mean DSC results of the single-path and multi-path models are 0.925 and 0.933, respectively. The p-value of the Wilcoxon sign rank test is 0.005.

The goal of our study was to test the hypothesis that the proposed 3D multi-path DenseNet could achieve better GBM segmentation than the corresponding 3D single-path DenseNet. Our proposed multi-path technique could be integrated into other 3D CNNs with different architectures for improving GBM segmentation. But this was not explored within the scope of this study. Also, our proposed multi-path technique may help achieve better performance in other image-transfer tasks, such as sCT generation and OAR segmentation from multi-modal MR images. The image output in the proposed DenseNets can be modified to a single channel for sCT generation, and multiple channels for OAR or tumor subregion segmentation. Future work will include integrating the multi-path technique into other 3D CNNs with different architectures to potentially improving GBM segmentation performance and investigating the performance of the multi-path technique in other image-transfer tasks.

4.5 Conclusion

We proposed a 3D single-path DenseNet and a 3D multi-path DenseNet for automatically generating GBM tumor contours from four multi-modal MR images. Both DenseNets generated accurate tumor contours that were in good agreement with the manually segmented contours. The single-path and multi-path DenseNets achieved DSCs of 0.911±0.060 and 0.922±0.041, respectively. Our study showed that the multi-path DenseNet generated more accurate GBM tumor contours than the single-path DenseNet.

5 Deep Learning-based Radiomic Features for Improving Neoadjuvant Chemoradiation Response Prediction in Locally Advanced Rectal Cancer[43]

5.1 Introduction

Colorectal cancer is the third most common cancer diagnosed and the second most common cause of cancer deaths in the US[112]. Rectal cancer accounts for about 30% of all colorectal cancer diagnoses[112]. Treatment for rectal cancer is based largely on the stage at diagnosis. LARC is commonly treated with nCRT followed by total mesorectal excision (TME) and adjuvant

chemotherapy[113,114]. Tumor response to nCRT is associated with recurrence and survival and can serve as a prognostic factor[115,116]. 15-27% of patients who undergo such treatment achieve pathologic complete response (pCR)[117]. TME is a highly invasive procedure with the potential risk of morbidity and functional complications. Achieving early prediction of tumor response using pre-treatment noninvasive approaches may allow for the design of individualized chemo-radiation treatment and potential avoidance of TME following nCRT for patients.

MRI is widely used in rectal cancer diagnosis and staging as it provides excellent soft-tissue contrast for tissue characterization. Specifically, increasing evidence has shown that diffusion-weighted MR images (DWIs), providing tissue cellularity information, aids the assessment of rectal cancer response to neoadjuvant treatment[118]. DWI is recommended to be routinely acquired in clinical guidelines[119]. The interpretation of DWI has gradually shifted from qualitative evaluation to quantitative assessment. For example, the apparent diffusion coefficient (ADC) map was one major quantitative map calculated from DWI. However, several studies showed that the mean pretreatment tumor ADC value was not a reliable indicator for predicting treatment response[120,121].

Radiomics is an emerging field of studies where a large number of medical image features are extracted in order to achieve better clinical diagnosis or decision support[122]. The conventional radiomic analysis typically involves extraction and analyzing quantitative imaging features from the previously defined region of interests (ROI) on one or multiple image modalities with the ultimate goal to obtain predictive or prognostic models. Previous studies showed that handcrafted, or explicitly designed, features extracted from the ADC ROI have predictive power for early nCRT response in LARC patients[123,124]. However, handcrafted features are lower-order image features and limited to current expert knowledge. Another type of

radiomic feature is DL-based extracted from the pre-trained CNNs via transfer learning[22,125]. Several studies have demonstrated that the DL-based features showed promising performance in breast cancer diagnosis, ovarian cancer recurrence prediction, and GBM survival prediction[38,126,127]. To our knowledge, no published study has investigated the DL-based features for managing LARC patients.

In this work, we first aimed to construct radiomics classifiers based on the handcrafted and DL-based radiomic features extracted from pre-treatment DWIs. Then, we compared the performance of the two classifiers to predict post-nCRT response in patients with LARC.

5.2 Materials and methods

5.2.1 Dataset

We identified forty-three consecutive patients with LARC treated from December 2015 to December 2016 at a single institution. All patients received concurrent capecitabine with a total prescription dose of 50 Gy in 25 fractions, followed by the TME surgery after 6-12 weeks of the nCRT completion. The resection specimens were evaluated by an expert pathologist. Patients were separated into good responders (GR) and non-GR groups based on the postoperative pathology report, MRI or colonoscopy. The GR group consisted of patients with either complete response (evaluated by pathology or MRI and colonoscopy) or partial response (assessed by pathology), and the non-GR group consisted of patients with stable disease (assessed by pathology) and progressive disease (confirmed by CT/MR).

All patients underwent pre-treatment DWIs before the nCRT. The DWI images were acquired using the single-shot echo-planar imaging (ssEPI) sequence on two 3T MR scanners (Discovery MR750 and Signa HDxt, GE Healthcare). MR imaging parameters are summarized

in Table 5-1. For each patient, the ADC map was computed using the equation $ADC = -\frac{1}{800}\ln(\frac{S}{S_0})$, where S_0 and S correspond to MR voxel intensities at b-values of 0 s/mm2 and 800 s/mm2. Gross tumor volume (GTV) of the primary tumor was manually delineated on the DWI image with the b-value of 800 s/mm^2 by a board-certified oncologist with 5-year experience.

Scanner model	Patient number	TR/TE (ms)	Matrix	FOV (mm^2)	Transvers spatial resolution(mm^2)	Slice thickness (mm)	b value (s/mm^2)
Discovery MR750	36	2600/74	256×256	380^2 or 400^2	1.48^2 or 1.56^2	5	0,500, 800,1000
Signa HDxt	7	4500-6000/64-67	256×256	320^2 - 400^2	1.25^2 - 1.56^2	5 or 6	0,800

Table 5-1: MR imaging parameters

5.2.2 Feature extraction

5.2.2.1 Handcrafted features

105 handcrafted features were extracted from the ADC map within the GTV contour for each patient using the PyRadiomics package[128] (version 2.1.2). Extracted features consisted of 14 shape-based features, 18 first-order statistic features, and 73 textural (second-order statistic) features. The methods used for extracting textural features were gray level co-occurrence matrix, gray-level size zone matrix, gray level run length matrix, gray level dependence matrix, and neighborhood gray-tone difference matrix. Shape-based features describe the shape characteristics of the GTV contour. First-order statistic features describe the distribution of voxel intensities within the GTV contour. Textural features describe the patterns or second-order spatial distributions of the voxel intensities.

5.2.2.2 DL-based features

The publicly-available pre-trained CNN, VGG19[65], was used to extract DL-based features. The network was trained using approximately 1.2 million images from the ImageNet database[129] for classifying nature images into 1000 objects. As the natural objects used for training VGG19 varied in their physical size, the extracted DL-based features using pre-trained VGG19 may be less sensitive to image spatial resolution (pixel size) compared with other factors like image gradients. Figure 5-1 shows the network architecture. It contained 16 convolutional layers followed by 3 fully-connected layers. 5 max-pooling layers were inserted across convolutional and fully-connected layers to reduce model parameter number for controlling overfitting and help achieve partial invariance to small translations. For each patient, a 2D square ROI was selected from the transverse slice that contains the largest tumor area. The ROI center was set as the center of the smallest bounding box covering the 2D tumor. The ROI size was set as the maximum dimension of the smallest bounding box. The ADC ROI was extrapolated to 224 by 224 for matching the original VGG19 design. The intensities of the ROI were converted to the range [0, 255]. Resampled ROI was copied into a 3-channel image and then inputted into the pre-trained model for feature extraction. We adopted the feature extraction method proposed by Antropova *et al.*[38]. As shown in Figure 5-1, five DL-based feature vectors were extracted by average-pooling the feature maps after max-pooling layers. Each feature vector was normalized with its Euclidean norm and then concatenated to one feature vector, which was normalized again to acquire the feature vector consisting of 1472 features. After extracting features for all patients, a cutoff on feature variance was used to pre-select 105 DL-based features out of 1472 features with the highest variance to train the prediction models.

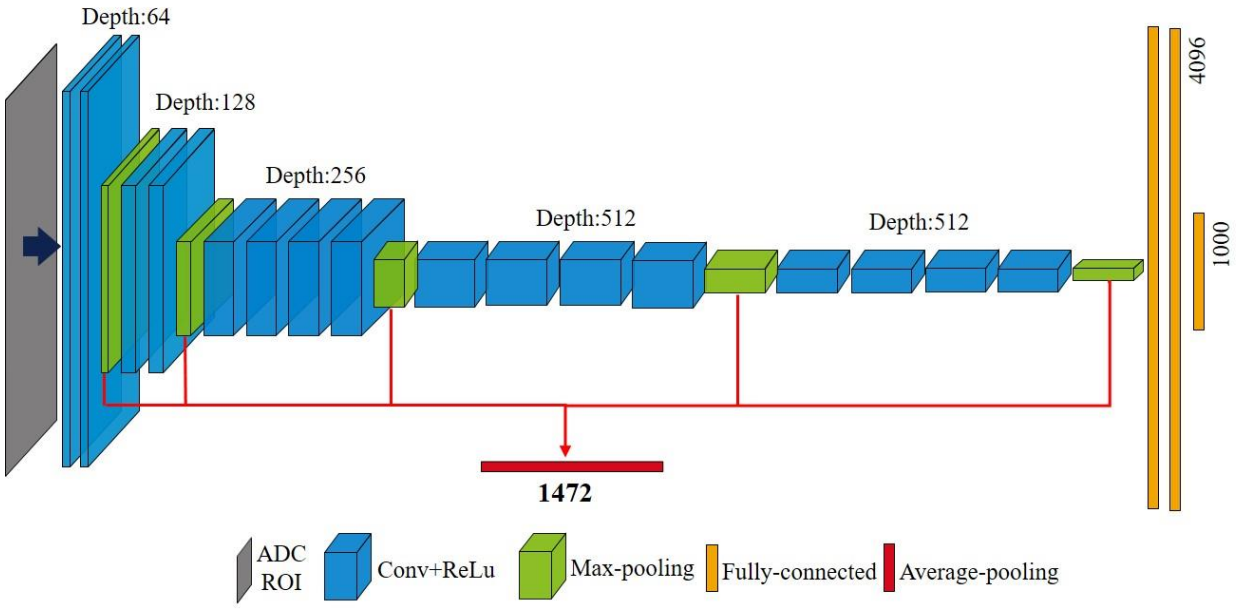

Figure 5-1: VGG19 architecture and feature extraction scheme. Feature maps and feature vectors, following each layer, are shown as cuboids and rectangles, respectively. The feature map depth and feature number are shown. For feature extraction, the network took an ADC ROI as input. 1472 DL-based features were extracted from max-pooling feature maps by average-pooling along the spatial dimensions. Conv, convolutional layer.

5.2.3 Classification and evaluation

The least absolute shrinkage and selection operator (LASSO) penalized logistic regression[130,131] was used for classification using radiomic features (Python version 2.7.13). The LASSO regularization was selected to handle the high feature dimension. The handcrafted classifier and DL-based classifier were trained using handcrafted features and DL-based features, respectively. The regularization parameter was optimized by grid searching with repeated 20 times stratified 4-fold cross-validation. For each cross-validation, stratified random sampling was used to split the patient cohort was into 4 folds, where 3 folds were used as the training set to train the classifier and the remaining one as the testing set for evaluation.

The performances of the handcrafted and DL-based classifiers were evaluated using the average area under the receiver operating curve (AUC) of 20 cross-validation repetitions. The

corrected paired t-test[132] was conducted to compare the AUC results for two classifiers. P-value <0.05 was considered to indicate a significant difference.

5.3 Results

Table 5-2 summarizes the clinical characteristics of our patient cohort. 22 (51.2%) patients achieved GR after nCRT. Among the 22 GR patients, there were 14 (63.6%) men and 8 (36.4%) women. Among 21 non-GR patients, there were 14 (66.7%) men and 7 (33.3%) women.

Characteristic	GR (n=22)	nGR (n=21)	Total (n=43)
Gender (male/female)	14/8	14/7	28/15
Age (mean, SD, in years)	53.7 (9.1)	54.9 (10.9)	54.3 (10.3)
Pre-nCRT TNM staging			
T stage (2/3/4)	1/18/3	1/16/4	2/34/7
N stage (0/1/2)	5/11/6	0/9/12	5/20/18

Table 5-2: Patient clinical characteristics; GR, good responder, nGR, non-good responder, SD, standard deviation.

Figure 5-2 shows the transverse slices of DWIs and ADC maps for the representative GR and non-GR patients. Both patients are male with rectal cancer at the same clinical stage of T3N1. No significant visual differences were observed.

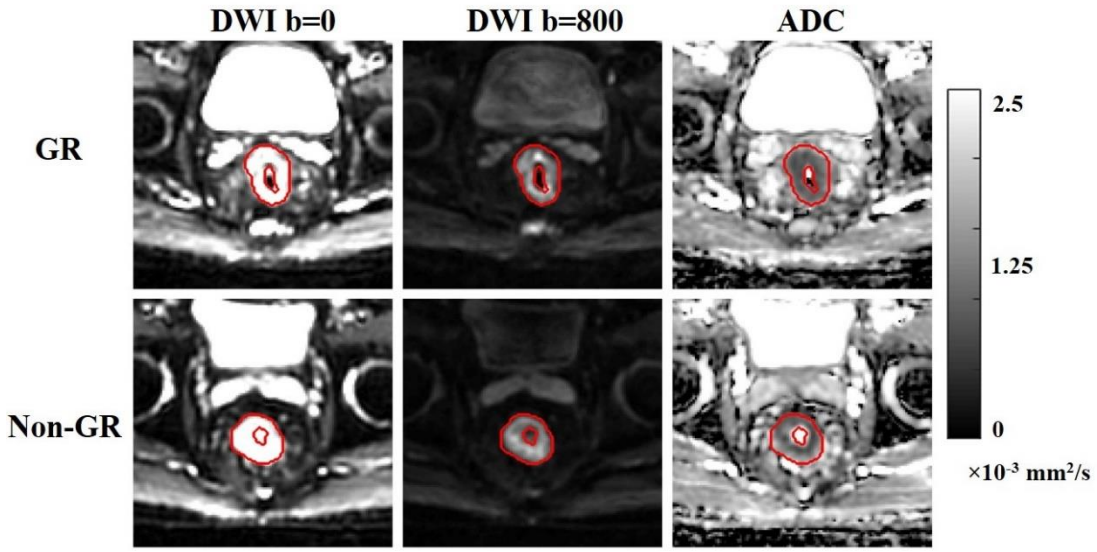

Figure 5-2: Comparison of the DWI (b=0,800 s/mm^2) slice and the ADC slice for the representative GR and non-GR patients. The GTV contours are demonstrated in red. The color bar of the ADC slices is shown.

Figure 5-3 (a) compares the boxplots of the mean AUC results of 20 cross-validation repetitions for two classifiers. Large deviations were observed due to the small sample size. The AUC of a single repetition varies from 0.51 to 0.73 for the handcrafted classifier, and from 0.58 to 0.80 for the DL-based classifier. The average ROC curves of the two classifiers are shown in Figure 5-3 (b). The handcrafted classifier achieved the mean AUC of 0.64 (standard error [SE], 0.08) using repeated 20 times 4-fold cross-validation, while the DL-based classifier achieved 0.73 (SE, 0.05). The p-value of the corrected paired t-test was 0.049, suggesting a significant difference in the AUC results for the handcrafted classifier and DL-based classifier.

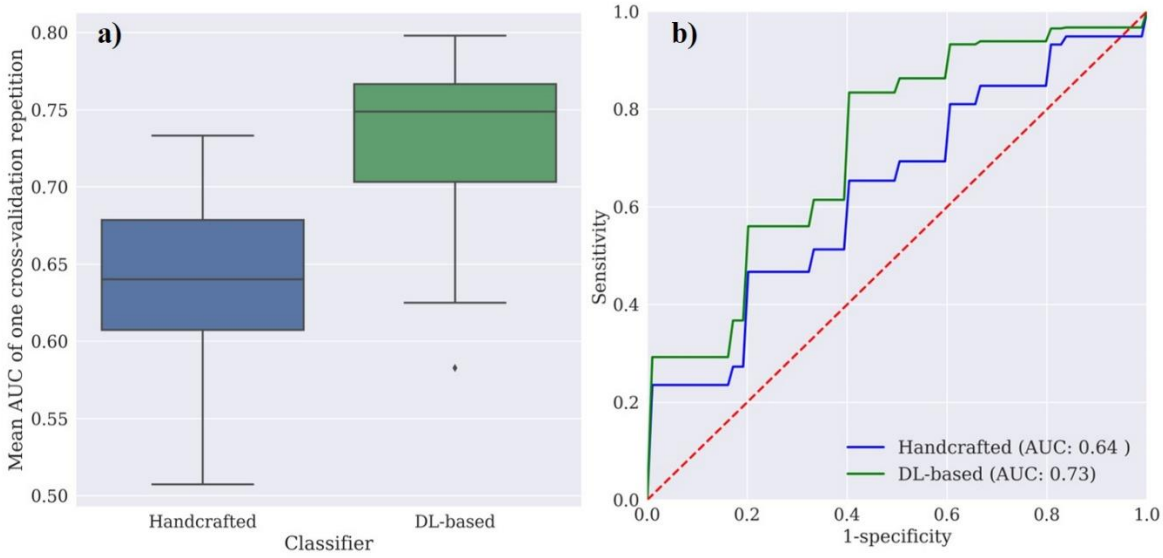

Figure 5-3: (a) Boxplots of the AUC results of 20 cross-validation repetitions for the handcrafted and DL-based classifiers. The minimum (bottom line), 25th percentile (bottom of the box), median (central line), 75th percentile (top of the box), and maximum (top line) are shown. An outlier is drawn as a diamond sign. (b) The ROC curves for two classifiers in predicting good response versus non-good response using repeated stratified 4-fold cross-validation. AUC results are averaged over 20×4 testing sets.

5.4 Discussion

In this study, we compared the performance of the classifiers built with the handcrafted and DL-based features, extracted from pre-treatment DWI, for predicting the post-nCRT response for a cohort of LARC patients. To our knowledge, this is the first study investigating DL-based features for this application. Compared with the handcrafted features, the DL-based features consisted of more abstract high-level information extracted from DWI images. Our study indicated that the DL-based classifier achieved a significantly better predictive performance than the handcrafted classifier in predicting nCRT response for rectal cancer. Studies showed that the DL-based features achieved better performance in breast cancer diagnosis and GBM survival prediction than the handcrafted features[38,127]. The DL-based features are expected to achieve

better performance and more generalizable results in diagnosis, recurrence and survival prediction for other sites as well.

We conducted repeated 4-fold cross-validation for evaluating the model performance as it stabilizes the accuracy estimation[132,133]. The handcrafted classifier achieved the mean AUC of 0.64 for predicting GR vs non-GR, while the DL-based classifier achieved an improved mean AUC of 0.73. Additionally, a fused classifier was constructed by averaging prediction scores of two classifiers. The fused classifier achieved the mean AUC of 0.71, which is better than that for the handcrafted classifier. Nie *et al.* using a single run of 4-fold cross-validation, reported the mean AUC of 0.73 for GR and non-GR prediction using DWI handcrafted features on a similar size cohort of 48 patients[123]. The standard error of the mean AUC was not reported. To investigate the cross-validation variation caused by the different data partitions for a small dataset, we conducted 20 independent cross-validation trials using our dataset. It should be noted that 20 independent cross-validation trials are different from the repeated 20 cross-validation since each cross-validation trial has its own optimal hyperparameters, while all 20 cross-validation repetitions need to have the same hyperparameters. The mean AUC of each cross-validation trial ranged from 0.56 to 0.79 for the classifier built with the handcrafted features, and from 0.63 to 0.82 for the one built with the DL-based features. Given a relatively small patient size, a single run of cross-validation may have a large bias. Also, different classification models, evaluation protocols, patient number, and response label ratio may result in different prediction accuracy.

We investigated the radiomic features extracted from a single imaging modality of DWI in this study. Several studies showed that including the handcrafted features from T2-weighted MR images and dynamic contrast-enhanced images improved predictive power[123,124]. The DL-

based feature extraction scheme can be applied to other MR imaging modalities and may further help improve the prediction accuracy. Comparing the handcrafted and DL-based features extracted from multiparametric MR images for treatment response prediction would be an interesting study to work on in the future.

The GTV contours used for feature extraction were manually delineated by a single radiation oncology. The effect of inter-observer delineation variabilities on the extracted features was not investigated in this work. Several studies suggested that the inter-observer delineation variability resulted in many unstable handcrafted radiomic features[134,135] and hence possibly less robust prediction models. Such delineation uncertainty may also lead to unstable DL-based features. The robustness of prediction models generated from the handcrafted and DL-based features can be investigated and compared using the intraclass correlation coefficient (ICC). A higher ICC indicates a better reproducibility. A cutoff on ICC could be used to select stable handcrafted and DL-based features that may result in more robust models. Alternatively, automatic tumor segmentation methods may be utilized to establish robust prediction models by reducing delineation variability.

The ROI used for extracting DL-based features was set based on tumor size, so the resampled ROIs would have different spatial resolutions across patients even if the spatial normalization was applied before the feature extraction. We believe it is unnecessary to conduct spatial normalization before extracting DL-based features in this study. To investigate the effect of the spatial normalization on handcrafted features, we resampled ADC maps to 1.56 x 1.56 x 5.00 mm^3 and re-extracted handcrafted features from the resampled ADC maps. The mean AUC achieved by the classifier trained using the updated handcrafted features was 0.65. Corrected

paired t-test showed that no significant difference in AUC was observed between the classifiers trained with original or updated handcrafted features (p=0.49).

Our study has several limitations. First, the study sample is relatively small, which may lead to unstable estimation and suboptimal model performance. A repeated cross-validation method was utilized to reduce the bias, and LASSO regularization was implemented to reduce overfitting. In this work, we investigated DL-based features extracted via transfer learning. Another popular DL approach for response prediction is to train CNNs from scratch or using finetuning. However, overfitting may become a major issue in this method especially when patient size is small. Our results, in concordance with other studies[127,136,137], showed that DL-based features extracted via transfer learning achieved promising results in various prediction tasks in the medical field. Second, our dataset only contained 9 patients with pCR. The pCR is defined as the absence of viable tumor cells in the primary and lymph nodes. The small number of pCR patients and unbalanced labels resulted in a large standard deviation on the AUCs using either handcrafted features or DL-based features for predicting pCR vs non-pCR. We chose to construct and evaluate the predictive model with the classification labels of GR and non-GR in this preliminary study. A larger dataset is desirable to provide a more reliable estimation for the AUC of pCR and non-pCR prediction. We expect to see better performance from the DL-based features than the handcrafted features in predicting pCR on a larger dataset. Lastly, the current study focused on the pre-treatment prediction of tumor response based on a single time point, due to the unavailability of during- and post-nCRT images for some patients. Given the primary focus of this work is mainly on comparing handcrafted features to DL-based features, we illustrated the earlier prediction for post-nCRT response, based on pre-treatment images, such early prediction will provide advantages for chemo-radiation treatment design and schedule. It

may be beneficial to assess the response by combining images at other time points, such as during- and post-nCRT images for higher prediction accuracy. We expect to see better performance from the DL-based features than the handcrafted features using other MR images acquired at different time points.

5.5 Conclusion

Our preliminary study showed that the DL-based radiomic features extracted via transfer learning from pretreatment DWIs achieved significantly better classification performance for predicting post-nCRT response in LARC patients, in comparison to the handcrafted radiomic features. Future work involves validation with a larger dataset and investigating the predictive power of the DL-based features extracted from multiparametric MR images (pre-, during-, and post-nCRT).

6 DL-BASED RADIOMIC FEATURES FOR IMPROVING GLIOBLASTOMA SURVIVAL PREDICTION IN AN AUTOMATIC WORKFLOW[44]

6.1 Introduction

Glioma is the most common type of primary brain tumor in adults. It arises from glial cells, normally astrocytes and oligodendrocytes. According to the World Health Organization guideline, glioma can be classified into grade I to IV based on the histological characteristics[93]. GBM is the most aggressive, grade IV, glioma. It accounts for 81% of malignant brain tumors[138]. Despite extensive efforts, GBM patient prognoses remain dismal. The median OS is only 15 months after diagnosis. The 5-year survival rate is below 5%[95]. It is beneficial to build survival prediction models for assisting therapeutic decisions and disease management in GBM patients.

MRI is the preferred imaging modality for GBM diagnosis and monitoring. Radiomic features extracted from MR images using advanced mathematical algorithms may uncover tumor characteristics that fail to be appreciated by the naked eye. Many studies have investigated the association of MRI radiomic features with the survival outcomes of GBM patients[127,139,140]. However, radiomic features were extracted from the manually drawn tumor contours in these studies. Manual tumor segmentation is not only time consuming but also sensitive to intra-observer and inter-observer variabilities. These segmentation variations could result in many inconsistent radiomic features[134,135], which introduces more challenges in constructing robust prediction models.

Developing an automatic GBM segmentation model could eliminate the manual contour variations and enable an automatic survival prediction workflow. CNNs have achieved state-of-the-art performance in medical image segmentation. Particularly, U-Net[71] and FCN[141] have been widely adopted. Shboul *et al.* used an ensemble of the 2D U-Net and the 2D FCN for GBM segmentation followed by an XGBoost based regression model to achieve automatic GBM survival prediction[142]. However, this study only investigated the handcrafted radiomic features that were extracted using explicitly designed algorithms. These features are normally low-level image features that are limited to current human knowledge. Another type of radiomic feature can be extracted using a pre-trained CNN using a pre-trained CNN[38,43,125]. We refer to these features as DL-based features in this study. These high-level features may have higher prognostic power than the handcrafted features.

In this study, we proposed an automatic workflow for GBM survival prediction based on four pre-operative MR images. A novel 3D CNN, VGG-Seg, was proposed and trained for automatic GBM segmentation. The handcrafted and DL-based radiomic features were extracted

from the autosegmented contours generated by the VGG-Seg and used to construct two separate Cox regression models for survival prediction. The prognostic power of the constructed signatures was evaluated and compared for improving prediction performance. To our knowledge, this is the first paper that investigated the DL-based radiomic features for automatic GBM survival prediction.

6.2 Materials and methods

6.2.1 Dataset

285 glioma patients were acquired from the BraTS 2018 challenge[100–102]. 210 patients had GBM, and the remaining 75 patients had low-grade (grade II-III) glioma (LGG). Each patient had four pre-operative MR images acquired. These included T1w, CE-T1w, T2w, and FLAIR MR images. Patient images were acquired with different clinical protocols and various scanners from multiple institutions. For each patient, MR images were co-registered, resampled to 1 mm^3 resolution using linear interpolation, and skull-stripped[100,101]. The final image dimension was 240×240×155. All patients have three ground truth tumor subregion labels (edema, enhancing tumor, and necrotic and non-enhancing tumor core) approved by experienced neuro-radiologists. OS data was available for 163 GBM patients.

We applied the N4 bias correction algorithm on all images, except the FLAIR images, to remove low-frequency inhomogeneity[62]. Each MR image was normalized to have zero mean and unit standard deviation in the brain voxels. Figure 6-1 hows the transverse slice of four preprocessed MR images and the corresponding tumor labels.

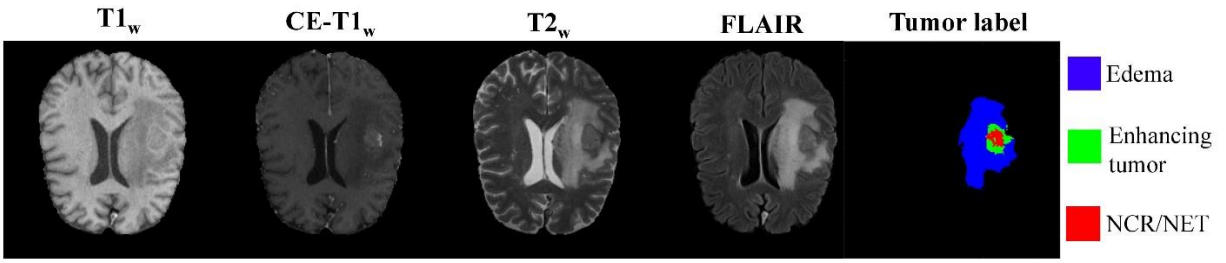

Figure 6-1: Transverse slices of preprocessed T1w, CE-T1w, T2w, FLAIR images along with the corresponding ground truth labels for edema, enhancing tumor, and necrotic and non-enhancing tumor core (NCR/NET) for a representative case.

6.2.2 VGG-Seg for automatic GBM segmentation

Figure 6-2 shows the architecture of the VGG-Seg proposed for automatic GBM segmentation. It contains 27 convolutional layers, forming an encoder and decoder architecture. The encoder network was constructed based on the VGG16 model[65] that achieved accurate performance in object detection. Instance normalization layers[66] and residual shortcuts[70] are implemented to improve model performance. The VGG-Seg can be trained to perform an end-to-end mapping, converting the concatenation of four preprocessed images to four probability maps for three tumor subregion labels and background labels.

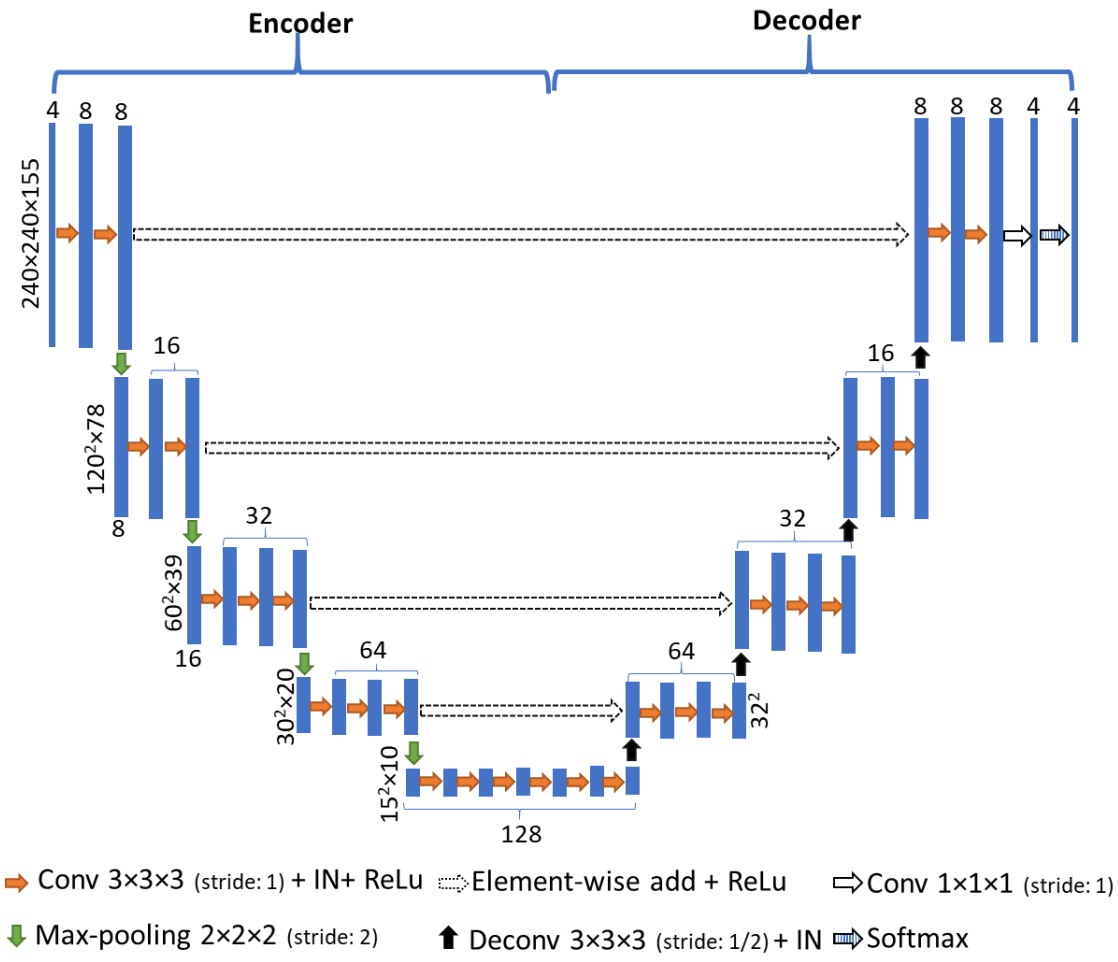

Figure 6-2: The overall VGG-Seg architecture. Each filled box represents a set of 4D feature maps, the numbers and dimensions of which are shown. The window size and the stride for convolutional, maxpooling, and deconvolutional layers are also presented. Conv, convolutional layer; IN, instance normalization layer; Maxpool, maxpooling layer; Deconv, deconvolutional layer.

In the model training stage, 122 patients without OS data were randomly split into a training set of 105 patients (75 LGG patients and 30 GBM patients) and a validation set of 17 GBM patients. The Adam stochastic gradient descent method[73] was used to minimize the multi-Dice loss,

$$loss = \frac{1}{4}\sum_{i=1}^{4}\left(1 - \frac{2\sum_{j=1}^{N} P_{ij} \times L_{ij}}{\sum_{j=1}^{N} P_{ij} + \sum_{j=1}^{N} L_{ij}}\right),$$

Equation 6-1

where P_{ij} is the probability, after Softmax layers, of the voxel *j* being the label *i*; four labels are background label and three tumor subregion labels; L_{ij} is the ground truth label, 0 or 1, of the voxel *j* being the label *i*; *N* is the voxel number. The validation set was used for tuning hyperparameters including the initial learning rate and the stopping epoch number and. A batch size of 1 was used for model training.

The trained VGG-Seg was applied to the remaining 163 GBM patients (all of which have corresponding OS data) to generate their tumor subregion labels. The autosegmented tumor contour was acquired by merging the three predicted subregion labels. Model accuracy was evaluated using the Dice coefficient,

$$Dice = \frac{2\left(V_{ground} \cap V_{auto}\right)}{V_{ground} + V_{auto}},$$

Equation 6-2

where V_{ground} and V_{auto} are the volumes of the ground truth tumor contour and autosegmented tumor contour, respectively.

6.2.3 Radiomic feature extraction

6.2.3.1 Handcrafted features

1106 handcrafted features were extracted from four MR images using the PyRadiomics[128] package (version 2.1.2) for all 163 GBM patients. These features were extracted from the autosegmented tumor contour and contained 14 shape-based features, 72 first-order statistical features, 292 second-order statistical (textural) features, and 728 high-order statistical features. Shape-based features represented the shape characteristics of the tumor contour. First-order

statistical features represented the characteristics of the tumor intensity distribution. Textural features were extracted based on gray level co-occurrence, gray level size zone, gray level run length, gray level dependence, and neighborhood gray-tone difference matrices. They represented the characteristics of the spatial intensity distributions. High-order statistic features were extracted from the images filtered using Laplacian of Gaussian (LoG) filters.

6.2.3.2 DL-based features

1472 DL-based features were extracted using a pre-trained classification CNN, VGG19 model[65], for all 163 GBM patients in the testing set. We used a pre-trained VGG19 that is available in the DL toolbox (Version 12.0) from MATLAB (Version 9.5, R2018b). It was trained on more than a million images from the ImageNet dataset[143]. Figure 6-3 shows the model architecture and feature extraction scheme. VGG19 contains 16 convolutional layers and 3 fully-connected layers. 5 max-pooling layers are used to achieve partial translational invariance, reduce model memory usage, and prevent overfitting. For each patient, we selected a square ROI from the transverse slice that had the largest tumor area. The size of the ROI was set as the maximum dimension of the tumor contour on the selected slice. We then resized the ROIs of FLAIR, T2w, and CE-T1w MR images to 224×224 using bilinear interpolation, mapped the pixel intensity to the range [0, 255], and concatenated them. The concatenation was input into the pre-trained VGG19 for feature extraction. As shown in Figure 6-3, DL-based features were extracted by average-pooling the 5 feature maps after max-pooling layers. Each feature map generated a vector after average-pooling. Five feature vectors were first normalized with their Euclidean norms and then concatenated to form a single feature vector. DL-based features were acquired by normalizing the single feature vector with its Euclidean norm.

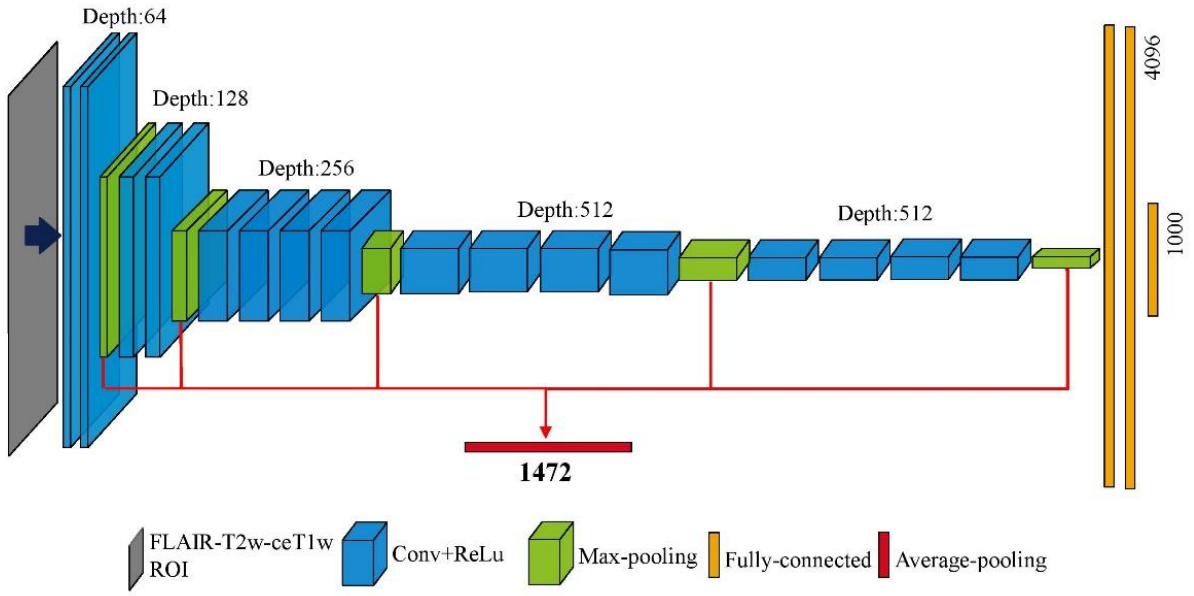

Figure 6-3: DL-based feature extraction scheme using VGG19. The average-pooling layers were used for extracting DL-based features. Feature maps and feature vectors after every layer are shown as cuboids and rectangles, respectively. The feature map depth and feature number are shown. A concatenation of FLAIR, T2w, and CE-T1w ROIs was input into the pre-trained VGG19 for feature extraction. 1472 DL-based features were extracted from max-pooling feature maps by average-pooling along the spatial dimensions. Conv, convolutional layer.

6.2.4 Survival prediction model

The 163 GBM patients with available OS data were randomly split into a training set of 122 patients and a testing set of 41 patients. Each feature was normalized using the mean and standard deviation of the training set. Since a large number of features may lead to overfitting, we pre-selected a subset of features having the highest univariate C-index. Higher C-index values indicate features with higher prognostic power. The Cox regression model with regularization was trained using the selected features to construct a radiomic signature for survival prediction in GBM patients. The radiomic signature is a linear combination of the features weighted by the Cox regression model coefficients. We tested three regularization techniques: Ridge, Elastic Net, and LASSO. The number of the pre-selected features, the

regularization technique, and the corresponding regularization parameters were chosen with 5-fold cross-validation using the training set. Two Cox regression models were trained using either handcrafted features or DL-based features. The resulting radiomic signatures are referred to as the handcrafted signature and the DL-based signature, respectively.

The prognostic power of the two constructed radiomic signatures was evaluated using the C-index. A paired t-test was conducted to test the significance of the differences in the C-index. A threshold on the radiomic signature can be set using the training set for patient stratification. We investigated two thresholds: one selected using the X-tile software[144], and the other defined by the median signature value of the training patients. The X-tile software selected the optimal threshold by selecting the highest X^2 value of the data divisions. The chosen thresholds were then used to stratify the testing patients into high-risk and low-risk groups. Log-rank tests were conducted to test the difference between the two risk groups for significance.

6.3 Results

6.3.1 OS statistics

The median and mean (standard deviation) of OS were 367.0 days and 416.5 (329.2) days in the training set, and 362.0 days and 442.1 (408.6) days in the testing set, respectively. A Mann-Whitney U test indicated that there was no significant difference in OS between two datasets (*p*-value=0.83).

6.3.2 Tumor segmentation

The VGG-Seg was trained using an initial learning rate of 5×10^{-4} for 150 epochs. These hyperparameters resulted in the minimum validation loss. Figure 6-4 compares the ground truth

tumor contours and the autosegmented tumor contours generated by the trained VGG-Seg for three patients in the testing set. The autosegmented contours were smoother than the corresponding ground truth contours. Both contours had similar shapes based on visual inspection.

The Dice coefficients of the whole tumor contours for the training, validation, and testing sets are summarized in Table 6-1. The autosegmented contours achieved the Dice coefficient of 0.86±0.09 on the whole tumor contour for 163 GBM patients in the testing set.

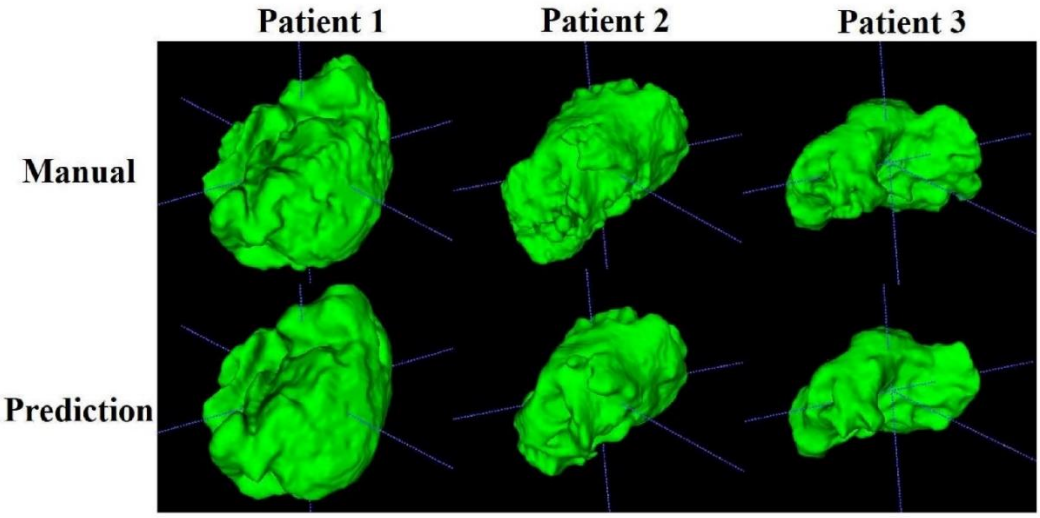

Figure 6-4: Ground truth contour (top) and autosegmented contour (bottom) for three GBM patients.

Dice	Training (75 LGG and 30 GBM)	Validation (17 GBM)	Testing (163 GBM)
Whole tumor	0.92±0.03	0.90±0.07	0.86±0.09

Table 6-1: Dice coefficients of the whole tumor contours for the training, validation, and testing sets. Results were averaged and showed in (mean ± SD) format.

6.3.3 Survival prediction

Table 6-2 shows the optimal pre-selected feature number, regularization technique, and regularization parameter that achieved the best cross-validation result for each feature set.

The handcrafted signature achieved a C-index of 0.64 (95% confidence intervals [CI]: 0.55-0.73) on the testing set, while the DL-based signature achieved a C-index of 0.67 (95% CI: 0.57-0.77). A paired t-test showed that there was no significant difference in the C-index (*p*-value=0.27).

	Number of pre-selected features	**Regularization technique**	**Regularization parameter (λ)**
Handcrafted features	50	Ridge	3.439
DL-based features	80	Ridge	1.813

Table 6-2: Optimal regularization technique and hyperparameters that were selected by 5-fold cross-validation for each feature set.

We split the testing patients into high-risk and low-risk groups based on signature thresholds. Figure 6-5 shows the Kaplan-Meier survival curves of the two risk groups. There was no significant association between the risk groups, stratified by thresholding the handcrafted signature, and the patient OS. (X-tile: *p*-value=0.31, hazard ratio [HR]=1.44, 95% CI: 0.71-2.91; Median: p-value=0.20, HR=1.51, 95% CI: 0.80-2.87). On the other hand, thresholds on the DL-based signature resulted in significant stratification of patients into two prognostically distinct groups (X-tile: *p*-value<0.01, HR=2.80, 95% CI: 1.26-6.24; Median: *p*-value=0.02, HR=2.16, 95% CI: 1.12-4.17).

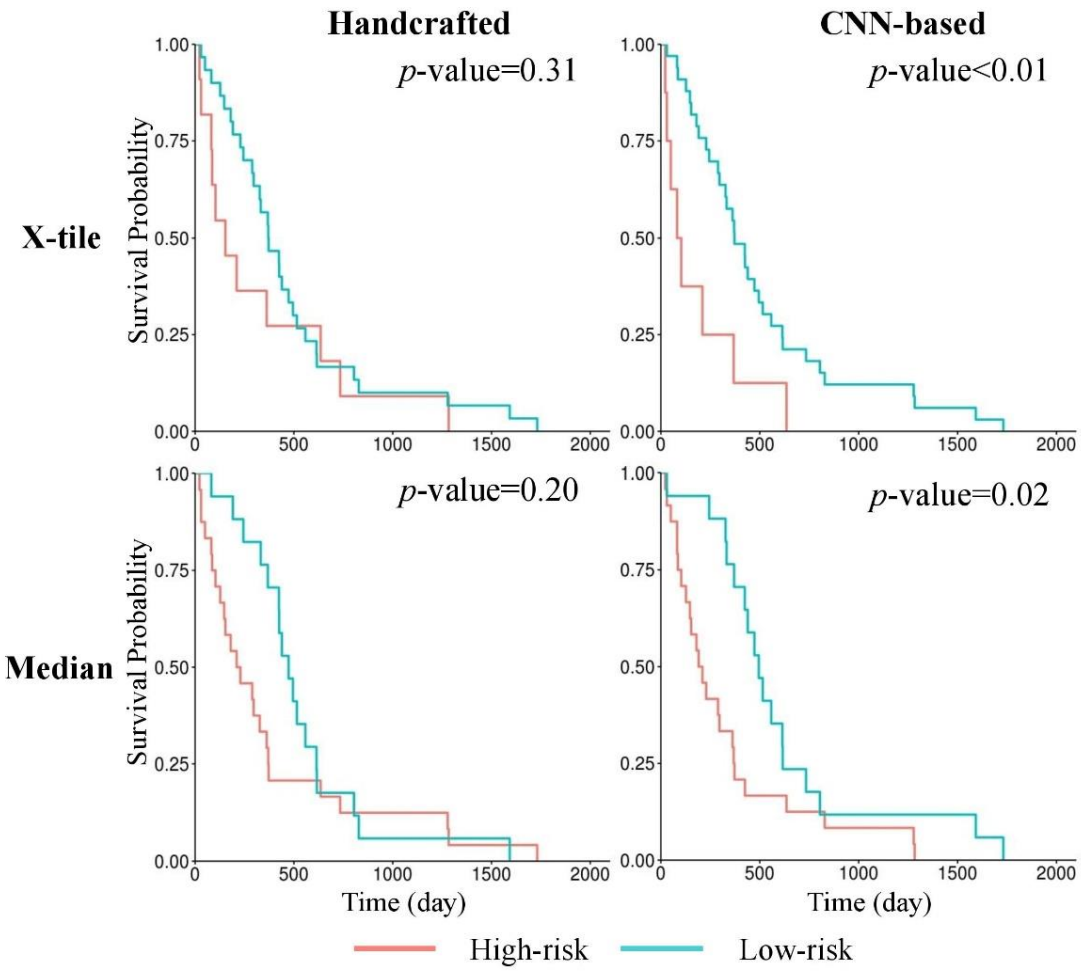

Figure 6-5: Kaplan-Meier survival curves of the testing patients. Patients were stratified into two risk groups based on thresholds of the handcrafted signature or the DL-based signature. The top row shows the stratification based on the threshold generated by X-tile software, and the bottom row shows the stratification based on the median signature value. *p*-values of the corresponding log-rank tests are shown.

6.4 Discussion

In this paper, we proposed an automatic workflow for GBM survival prediction based on four pre-operative MR images. The VGG-Seg was proposed and trained using 105 glioma patients for automatically generating GBM contours from four MR images. The trained VGG-Seg was applied to 163 GBM patients to generate their autosegmented tumor contours for survival analysis. We extracted handcrafted and DL-based radiomic features from the MR images using

the autosegmented contours for these patients. Two Cox regression models were trained using the extracted features to construct the handcrafted and DL-based signatures for survival prediction.

The DL-based signature had a better association with GBM OS than the handcrafted signature in terms of higher C-index and significant patient stratification. The handcrafted signature achieved a C-index of 0.64, while the DL-based signature achieved a C-index of 0.67. The DL-based signature, unlike the handcrafted signature, resulted in prognostically distinct groups using either X-tile generated or median threshold. Shboul *et al.* did not report the C-index but the accuracy of 0.52 in classifying GBM patients into three survival outcome groups[142]. However, DL-based radiomic features were not investigated in this study. It is also difficult to know whether significant patient stratification was achieved for the testing GBM patients in this study since log-rank tests were not conducted.

The VGG-Seg achieved accurate automatic GBM segmentation, with a mean Dice coefficient of 0.86 for the 163 GBM patients. A study showed that the mean Dice coefficient between the whole tumor contours drawn by two experts based on multi-modal MR images was 0.86[145]. Recently, many studies have proposed novel 3D CNN architectures for improving glioma segmentation accuracy[105,146]. The goal of this study is not to benchmark the best segmentation model but to develop an automatic workflow that can achieve accurate GBM survival prediction. Other automatic segmentation methods can be integrated into the proposed workflow but were not explored within the scope of this study. Potential future work includes selecting the best segmentation model and investigating whether more accurate autosegmented contours may result in a better survival prediction model.

We included 75 LGG patients for training the VGG-Seg because we found that the VGG-Seg trained with both 75 LGG patients and 30 GBM patients achieved better performance than the VGG-Seg trained with 30 GBM patients alone. This is expected as LGG and GBM have a similar appearance in MR images. The VGG-Seg could generate three tumor subregion labels. However, the accuracy of segmenting subregion labels using the VGG-Seg was low, with the mean Dice coefficients of the tumor subregions smaller than 0.75. Hence, we decided to use the whole tumor contours for feature extraction.

Our study has several limitations. First, the number of patients is limited so we only investigated the transfer learning method for survival prediction. A CNN trained from scratch for survival prediction could directly learn useful features from MR images. However, it could be easily overfitted and hence require more patient data to achieve robust performance. Other methods like training an autoencoder for feature extraction would also be valuable to explore. Second, the information provided by the MR images may be limited and not powerful enough for achieving more accurate models. Future work could be done to include genomic features and investigate whether the combination of genomic and radiomic features could improve prediction performance.

6.5 Conclusion

We proposed an automatic workflow for GBM survival prediction based on four pre-operative MR images. The proposed VGG-Seg generated accurate GBM contours. Our study showed that radiomic features, extracted from the autosegmented contours generated by the VGG-Seg, were associated with GBM OS. The DL-based radiomic signature resulted in a higher C-index than the handcrafted signature and helped achieve significant patient stratification. Our automatic

workflow based on DL-based radiomic features demonstrated the potential of improving patient stratification and survival prediction in GBM patients.

7 CONCLUSIONS

7.1 Summary of work

The studies presented in this dissertation provide solutions to address three major limitations in the current MRgRT workflow.

In Chapter 2, we investigated 2D and 3D CNNs to generate a male pelvic sCT using a T1-weighted MR image and compared their performance. A retrospective study was performed using CTs and T1-weighted MR images of 20 prostate cancer patients. Both models were trained and evaluated using a 5-fold cross-validation protocol. The average MAEs within the body contour were 40.5±5.4 HU (mean ± SD) and 37.6±5.1 HU for the 2D and 3D CNNs, respectively. For both CNNs, mean translation vector distances are less than 0.6 mm with mean absolute differences of Euler angles less than 0.5°. Our results showed that both models generated accurate pelvic sCTs for 20 patients. Statistical tests indicated that the proposed 3D CNN was able to generate sCTs with smaller MAE and higher bone region precision compared with the 2D CNN. Results of patient alignment tests suggested sCTs generated by the proposed CNNs can provide accurate patient positioning.

In Chapter 3, we investigated whether cGAN and cycleGAN can generate accurate abdominal sCT images from 0.35T MR images for MR-only liver radiation therapy. A retrospective study was performed using CT images and 0.35T MR images of 12 patients with liver (n=8) and non-liver abdominal (n=4) cancer. Both models were trained and evaluated using a 4-fold cross-validation protocol. $sCT_{cycleGAN}$ achieved the average MAE of 94.1 HU, while sCT_{cGAN} achieved 89.8 HU. In both GANs, the average gamma passing rates within all volumes of interest were higher than 95% using a 2%, 2 mm criterion, and 99% using a 3%, 3 mm criterion. The average differences in the mean dose and DVH metrics were within ±0.6% for the planning target volume and within ±0.15% for evaluated organs in both models. Our results demonstrated that abdominal sCT images generated by both models achieved accurate dose calculation for 8 liver radiation therapy plans.

Chapter 4 describes a novel 3D multi-path DenseNet for generating the accurate GBM tumor from four pre-operative multi-modal MR images. The corresponding 3D single-path DenseNet was also built for comparison. 258 GBM patients were included in this study. The patient cohort was randomly split into a training set of 180 patients, a validation set of 39 patients, and a testing set of 39 patients. Both DenseNets generated GBM contours in good agreement with the manually segmented contours from multi-modal MR images. The single-path DenseNet achieved the DSC of 0.911±0.060, ASD of 1.3±0.7 mm, and $HD_{95\%}$ of 5.2±7.1 mm, while the multi-path DenseNet achieved the DSC of 0.922±0.041, ASD of 1.1±0.5 mm, and $HD_{95\%}$ of 3.9±3.3 mm. The p-values of all Wilcoxon signed-rank tests were less than 0.05. The multi-path DenseNet achieved more accurate tumor segmentation than the single-path DenseNet.

In Chapter 5, we compared the handcrafted and DL-based radiomic features extracted from pre-treatment DWIs for predicting nCRT response in patients with LARC. 43 patients

receiving nCRT were included. The patient-cohort was split into the GR group (n=22) and nGR group (n=21) based on the post-nCRT response assessed by postoperative pathology. LASSO-logistic regression models were constructed using extracted features for predicting treatment response. The model performance was evaluated with repeated 20 times stratified 4-fold cross-validation. The model built with handcrafted features achieved the mean AUC of 0.64, while the one built with DL-based features yielded the mean AUC of 0.73. The corrected paired t-test on AUC showed P-value < 0.05. Our results showed that DL-based features achieved significantly better classification performance compared with handcrafted features for predicting nCRT response in patients with LARC.

Chapter 6 presents an automatic workflow for GBM survival prediction based on four pre-operative multi-modal MR images. 285 glioma (210 GBM and 75 low-grade glioma) patients were included. 163 of the GBM patients had OS data. A 3D CNN, VGG-Seg, was trained and validated using 122 glioma patients for automatic GBM segmentation. The trained VGG-Seg was applied to the remaining 163 GBM patients to generate their autosegmented tumor contours. The handcrafted and DL-based radiomic features were extracted from the autosegmented contours for these 163 patients who were randomly split into training (n=122) and testing (n=41) sets for survival analysis. Cox regression models were trained to construct the handcrafted and DL-based signatures. The handcrafted signature achieved a C-index of 0.64 (95% CI: 0.55-0.73), while the DL-based signature achieved a C-index of 0.67 (95% CI: 0.57-0.77). Unlike the handcrafted signature, the DL-based signature successfully stratified testing patients into two prognostically distinct groups. Our results showed that the DL-based signature resulted in better GBM survival prediction, in terms of higher C-index and significant patient stratification, than the handcrafted signature.

7.2 Future directions

The future work of each study was presented in the corresponding Discussion section. Besides those, there are a few other directions that are worthy of investigation.

First, we only trained 3D DenseNets for tumor segmentation in Chapter 4. However, OAR segmentation is more time-consuming than target segmentation in an online adaptive workflow. It would be meaningful to investigate the performance of DenseNets for OAR segmentation.

Second, the automatic radiomic workflow presented in Chapter 6 used the VGG-Seg that has worse GBM segmentation performance compared with the multi-path 3D DenseNet described in Chapter 4. The future work could be integrating the 3D DenseNet into the automatic radiomic workflow and re-evaluate its performance of GBM survival prediction.

Third, we had to study different cancer types in this dissertation due to the limited data availability. In future work, we could adapt the proposed methodologies for one cancer type, for example, hepatocellular carcinoma (HCC).

More than 800,000 new cases of liver cancer are diagnosed each year worldwide[147,148]. HCC is the most common type of primary liver cancer, accounting for about 75% of all liver cancers worldwide. As one of the leading causes of cancer-related mortality, HCC poses a significant economic burden on healthcare. The principal treatment for HCC is surgical resection or liver transplantation. However, more than 80% of patients are not eligible for surgical interventions[149]. For these patients, recommended curative treatments include percutaneous ethanol injection and radiofrequency ablation. SBRT allows the delivery of high doses in few fractions to the tumor. It has become an emerging treatment for inoperable HCC, particularly for

tumor down-staging before surgical procedures[150,151]. A study showed that MR-guided SBRT is a feasible and safe treatment option for HCC treatment.

We showed that GANs could generate accurate sCT images for MR-only liver SBRT in Chapter 3. However, more HCC patients should be enrolled to evaluate the feasibility of using GANs for MR-only SBRT in HCC. Most CNN-based segmentation models mainly focus on stable organs such as the brain and liver. Very few have been studied for liver tumor segmentation based on MR images. DenseNets proposed in Chapter 4 could be adapted for this in the future. A few studies suggested that standard image-based response assessment criteria, such as modified response evaluation criteria in solid tumors (mRECIST) and European association for the study of liver diseases (EASL), could lead to inaccurate evaluations of SBRT response in HCC[152,153] . Up to now, no study has been conducted to build an early treatment response prediction model for HCC treated with MR-guided SBRT. The radiomic workflow proposed in Chapter 6 could be adapted for this in the future.

8 REFERENCES

1. Atun R, Jaffray DA, Barton MB, et al. Expanding global access to radiotherapy. *Lancet Oncol*. 2015;16(10):1153-1186. doi:10.1016/S1470-2045(15)00222-3

2. Thompson MK, Poortmans P, Chalmers AJ, et al. Practice-changing radiation therapy trials for the treatment of cancer: where are we 150 years after the birth of Marie Curie? *Br J Cancer*. 2018;119(4):389-407. doi:10.1038/s41416-018-0201-z

3. Otazo R, Lambin P, Pignol JP, et al. MRI-guided Radiation Therapy: An Emerging Paradigm in Adaptive Radiation Oncology. *Radiology*. 2021;298(2):248-260. doi:10.1148/radiol.2020202747

4. Schmidt MA, Payne GS. Radiotherapy planning using MRI. *Phys Med Biol*. 2015;60(22):R323-R361. doi:10.1088/0031-9155/60/22/R323

5. Prabhakar R V., Julka PK, Malik M, et al. Comparison of contralateral breast dose for various tangential field techniques in clinical radiotherapy. *Technol Cancer Res Treat*. 2007;6(2):135-138. doi:10.1177/153303460700600210

6. Fiorentino A, Caivano R, Pedicini P, Fusco V. Clinical target volume definition for

glioblastoma radiotherapy planning: magnetic resonance imaging and computed tomography. *Clin Transl Oncol*. 2013;15(9):754-758. doi:10.1007/s12094-012-0992-y

7. Bowen SR, Yuh WTC, Hippe DS, et al. Tumor radiomic heterogeneity: Multiparametric functional imaging to characterize variability and predict response following cervical cancer radiation therapy. *J Magn Reson Imaging*. 2018;47(5):1388-1396. doi:10.1002/jmri.25874

8. Gao Y, Kalbasi A, Hsu W, et al. Treatment effect prediction for sarcoma patients treated with preoperative radiotherapy using radiomics features from longitudinal diffusion-weighted MRIs. *Phys Med Biol*. 2020;65(17):175006. doi:10.1088/1361-6560/ab9e58

9. Klein S, Staring M, Murphy K, Viergever MA, Pluim JPW. elastix: A Toolbox for Intensity-Based Medical Image Registration. *IEEE Trans Med Imaging*. 2010;29(1):196-205. doi:10.1109/TMI.2009.2035616

10. Wojcieszynski AP, Rosenberg SA, Brower J V., et al. Gadoxetate for direct tumor therapy and tracking with real-time MRI-guided stereotactic body radiation therapy of the liver. *Radiother Oncol*. 2016;118(2):416-418. doi:10.1016/J.RADONC.2015.10.024

11. Henke L, Kashani R, Robinson C, et al. Phase I trial of stereotactic MR-guided online adaptive radiation therapy (SMART) for the treatment of oligometastatic or unresectable primary malignancies of the abdomen. *Radiother Oncol*. 2018;126(3):519-526. doi:10.1016/J.RADONC.2017.11.032

12. Karlsson M, Karlsson MG, Nyholm T, Amies C, Zackrisson B. Dedicated magnetic resonance imaging in the radiotherapy clinic. *Int J Radiat Oncol Biol Phys*. 2009;74(2):644-651. doi:10.1016/j.ijrobp.2009.01.065

13. Edmund JM, Nyholm T. A review of substitute CT generation for MRI-only radiation therapy. *Radiat Oncol.* 2017;12(1):28. doi:10.1186/s13014-016-0747-y

14. Lamb J, Cao M, Kishan A, et al. Online Adaptive Radiation Therapy: Implementation of a New Process of Care. *Cureus.* 2017;9(8):e1618. doi:10.7759/cureus.1618

15. Sjölund J, Forsberg D, Andersson M, Knutsson H. Generating patient specific pseudo-CT of the head from MR using atlas-based regression. *Phys Med Biol*. 2015;60(2):825-839. doi:10.1088/0031-9155/60/2/825

16. Guerreiro F, Burgos N, Dunlop A, et al. Evaluation of a multi-atlas CT synthesis approach for MRI-only radiotherapy treatment planning. *Phys Med*. 2017;35:7-17. doi:10.1016/j.ejmp.2017.02.017

17. Dowling JA, Sun J, Pichler P, et al. Automatic Substitute Computed Tomography Generation and Contouring for Magnetic Resonance Imaging (MRI)-Alone External Beam Radiation Therapy From Standard MRI Sequences. *Int J Radiat Oncol Biol Phys*. 2015;93(5):1144-1153. doi:10.1016/j.ijrobp.2015.08.045

18. Chin AL, Lin A, Anamalayil S, Teo B-KK. Feasibility and limitations of bulk density assignment in MRI for head and neck IMRT treatment planning. *J Appl Clin Med Phys*. 2014;15(5):100-111. doi:10.1120/jacmp.v15i5.4851

19. Korhonen J, Kapanen M, Keyrilainen J, Seppala T, Tenhunen M. A dual model HU conversion from MRI intensity values within and outside of bone segment for MRI-based radiotherapy treatment planning of prostate cancer. *Med Phys*. 2014;41(1). doi:10.1118/1.4842575

20. Johansson A, Karlsson M, Nyholm T. CT substitute derived from MRI sequences with

ultrashort echo time. *Med Phys*. 2011;38(5):2708-2714. doi:10.1118/1.3578928

21. Hsu SH, Cao Y, Huang K, Feng M, Balter JM. Investigation of a method for generating synthetic CT models from MRI scans of the head and neck for radiation therapy. *Phys Med Biol*. 2013;58(23):8419-8435. doi:10.1088/0031-9155/58/23/8419

22. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521(7553):436-444. doi:10.1038/nature14539

23. Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 3D U-net: Learning dense volumetric segmentation from sparse annotation. In: *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol 9901 LNCS. Springer Verlag; 2016:424-432. doi:10.1007/978-3-319-46723-8_49

24. Isola P, Zhu J-Y, Zhou T, Efros AA. Image-to-Image Translation with Conditional Adversarial Networks. November 2016. http://arxiv.org/abs/1611.07004. Accessed June 11, 2019.

25. Zhu J-Y, Park T, Isola P, Efros AA. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. March 2017. http://arxiv.org/abs/1703.10593. Accessed June 11, 2019.

26. Long J, Shelhamer E, Darrell T. *Fully Convolutional Networks for Semantic Segmentation*. https://people.eecs.berkeley.edu/~jonlong/long_shelhamer_fcn.pdf. Accessed August 28, 2018.

27. Badrinarayanan V, Kendall A, Cipolla R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Trans Pattern Anal Mach Intell*.

2017;39(12):2481-2495. doi:10.1109/TPAMI.2016.2644615

28. Zhao H, Shi J, Qi X, Wang X, Jia J. *Pyramid Scene Parsing Network*. https://github.com/hszhao/PSPNet. Accessed September 11, 2018.

29. Chen L-C, Papandreou G, Schroff F, Adam H. *Rethinking Atrous Convolution for Semantic Image Segmentation*. https://arxiv.org/pdf/1706.05587.pdf. Accessed September 11, 2018.

30. Milletari F, Navab N, Ahmadi S-A. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. In: *2016 Fourth International Conference on 3D Vision (3DV)*. IEEE; 2016:565-571. doi:10.1109/3DV.2016.79

31. Yip SSF, Aerts HJWL. Applications and limitations of radiomics. *Phys Med Biol*. 2016;61(13):R150-R166. doi:10.1088/0031-9155/61/13/R150

32. Aerts HJWL. The Potential of Radiomic-Based Phenotyping in Precision Medicine. *JAMA Oncol*. 2016;2(12):1636. doi:10.1001/jamaoncol.2016.2631

33. Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images Are More than Pictures, They Are Data. *Radiology*. 2016;278(2):563-577. doi:10.1148/radiol.2015151169

34. Chaddad A, Sabri S, Niazi T, Abdulkarim B. Prediction of survival with multi-scale radiomic analysis in glioblastoma patients. *Med Biol Eng Comput*. June 2018. doi:10.1007/s11517-018-1858-4

35. Vallières M, Freeman CR, Skamene SR, El Naqa I. A radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities. *Phys Med Biol*. 2015;60(14):5471-5496. doi:10.1088/0031-9155/60/14/5471

36. Li Z-C, Li Q, Sun Q, Luo R, Chen Y. Identifying a radiomics imaging signature for prediction of overall survival in glioblastoma multiforme. In: *2017 10th Biomedical Engineering International Conference (BMEiCON)*. IEEE; 2017:1-4. doi:10.1109/BMEiCON.2017.8229098

37. Kickingereder P, Burth S, Wick A, et al. Radiomic Profiling of Glioblastoma: Identifying an Imaging Predictor of Patient Survival with Improved Performance over Established Clinical and Radiologic Risk Models. *Radiology*. 2016;280(3):880-889. doi:10.1148/radiol.2016160845

38. Antropova N, Huynh BQ, Giger ML. A deep feature fusion methodology for breast cancer diagnosis demonstrated on three imaging modality datasets. *Med Phys*. 2017;44(10):5162-5171. doi:10.1002/mp.12453

39. Nanni L, Ghidoni S, Brahnam S. Handcrafted vs. non-handcrafted features for computer vision classification. *Pattern Recognit*. 2017;71:158-172. doi:10.1016/j.patcog.2017.05.025

40. Fu J, Yang Y, Singhrao K, et al. Deep learning approaches using 2D and 3D convolutional neural networks for generating male pelvic synthetic computed tomography from magnetic resonance imaging. *Med Phys*. 2019;46(9):3788-3798. doi:10.1002/mp.13672

41. Fu J, Singhrao K, Cao M, et al. Generation of abdominal synthetic CTs from 0.35T MR images using generative adversarial networks for MR-only liver radiotherapy. *Biomed Phys Eng Express*. 2020;6(1):15033. doi:10.1088/2057-1976/ab6e1f

42. Fu J, Singhrao K, Qi XS, Yang Y, Ruan D, Lewis JH. 3D multi-path DenseNet for improving automatic segmentation of glioblastoma on pre-operative multi-modal MR

images. *Med Phys*. February 2021:mp.14800. doi:10.1002/mp.14800

43. Fu J, Zhong X, Li N, et al. Deep learning-based radiomic features for improving neoadjuvant chemoradiation response prediction in locally advanced rectal cancer. *Phys Med Biol*. 2020;65(7):075001. doi:10.1088/1361-6560/ab7970

44. Fu J, Singhrao K, Zhong X, et al. An automatic deep learning-based workflow for glioblastoma survival prediction using pre-operative multimodal MR images. *arXiv*. January 2020. http://arxiv.org/abs/2001.11155. Accessed March 20, 2021.

45. Dirix P, Haustermans K, Vandecaveye V. The Value of Magnetic Resonance Imaging for Radiotherapy Planning. *Semin Radiat Oncol*. 2014;24(3):151-159. doi:10.1016/j.semradonc.2014.02.003

46. Moffat BA, Chenevert TL, Lawrence TS, et al. Functional diffusion map: A noninvasive MR1 biomarker for early stratification of clinical brain tumor response. *Proc Natl Acad Sci U S A*. 2005;102(15):5524-5529. doi:10.1073/pnas.0501532102

47. Patterson DM, Padhani AR, Collins DJ. Technology Insight: water diffusion MRI - a potential new biomarker of response to cancer therapy. *Nat Clin Pract Oncol*. 2008;5(4):220-233. doi:10.1038/ncponc1073

48. Dowling JA, Lambert J, Parker J, et al. An atlas-based electron density mapping method for magnetic resonance imaging (MRI)-alone treatment planning and adaptive MRI-based prostate radiation therapy. *Int J Radiat Oncol Biol Phys*. 2012;83(1):e5-11. doi:10.1016/j.ijrobp.2011.11.056

49. Korsholm ME, Waring LW, Edmund JM. A criterion for the reliable use of MRI-only radiotherapy. *Radiat Oncol*. 2014;9. doi:10.1186/1748-717X-9-16

50. Doemer A, Chetty IJ, Glide-Hurst C, et al. Evaluating organ delineation, dose calculation and daily localization in an open-MRI simulation workflow for prostate cancer patients. *Radiat Oncol*. 2015;10. doi:10.1186/s13014-014-0309-0

51. Berker Y, Franke J, Salomon A, et al. MRI-based attenuation correction for hybrid PET/MRI systems: a 4-class tissue segmentation technique using a combined ultrashort-echo-time/Dixon MRI sequence. *J Nucl Med*. 2012;53(5):796-804. doi:10.2967/jnumed.111.092577

52. Kapanen M, Tenhunen M. T1/T2(star)-weighted MRI provides clinically relevant pseudo-CT density data for the pelvic bones in MRI-only based radiotherapy treatment planning. *Acta Oncol (Madr)*. 2013;52(3):612-618. doi:10.3109/0284186X.2012.692883

53. Hsu SH, Cao Y, Lawrence TS, et al. Quantitative characterizations of ultrashort echo (UTE) images for supporting air-bone separation in the head. *Phys Med Biol*. 2015;60(7):2869-2880. doi:10.1088/0031-9155/60/7/2869

54. Yang Y, Cao M, Kaprealian T, et al. Accuracy of UTE-MRI-based patient setup for brain cancer radiation therapy. *Med Phys*. 2016;43(1):262. doi:10.1118/1.4938266

55. Siversson C, Nordstrom F, Nilsson T, et al. Technical Note: MRI only prostate radiotherapy planning using the statistical decomposition algorithm. *Med Phys*. 2015;42(10):6090-6097. doi:10.1118/1.4931417

56. Nie D, Cao X, Gao Y, Wang L, Shen D. Estimating CT Image from MRI Data Using 3D Fully Convolutional Networks. *Deep Learn Data Label Med Appl*. 2016;2016:170-178. doi:10.1007/978-3-319-46976-8_18

57. Nie D, Trullo R, Lian J, et al. Medical image synthesis with context-aware generative

adversarial networks. In: *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol 10435 LNCS. Springer Verlag; 2017:417-425. doi:10.1007/978-3-319-66179-7_48

58. Han X. MR-based synthetic CT generation using a deep convolutional neural network method. *Med Phys*. 2017;44(4):1408-1419. doi:10.1002/mp.12155

59. Rasch C, Barillot I, Remeijer P, Touw A, Van Herk M, Lebesque J V. Definition of the prostate in CT and MRI: A multi-observer study. *Int J Radiat Oncol Biol Phys*. 1999;43(1):57-66. doi:10.1016/S0360-3016(98)00351-4

60. Khoo VS, Joon DL. New developments in MRI for target volume delineation in radiotherapy. *Br J Radiol*. 2006;79(SPEC. ISS.). doi:10.1259/bjr/41321492

61. McLaughlin PW, Evans C, Feng M, Narayana V. Radiographic and Anatomic Basis for Prostate Contouring Errors and Methods to Improve Prostate Contouring Accuracy. *Int J Radiat Oncol Biol Phys*. 2010;76(2):369-378. doi:10.1016/j.ijrobp.2009.02.019

62. Tustison NJ, Avants BB, Cook PA, et al. N4ITK: improved N3 bias correction. *IEEE Trans Med Imaging*. 2010;29(6):1310-1320. doi:10.1109/TMI.2010.2046908

63. Nyul LG, Udupa JK, Zhang X. New variants of a method of MRI scale standardization. *IEEE Trans Med Imaging*. 2000;19(2):143-150. doi:10.1109/42.836373

64. Otsu N. Threshold selection method from gray-level histograms. *Ieee Trans Syst Man Cybern*. 1979;9(1):62-66.

65. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. *arXiv Prepr arXiv14091556*. 2014.

66. Ulyanov D, Vedaldi A, Lempitsky V. Improved texture networks: Maximizing quality and

diversity in feed-forward stylization and texture synthesis. In: *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*. Vol 2017-January. Institute of Electrical and Electronics Engineers Inc.; 2017:4105-4113. doi:10.1109/CVPR.2017.437

67. Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: *32nd International Conference on Machine Learning, ICML 2015*. Vol 1. International Machine Learning Society (IMLS); 2015:448-456. https://arxiv.org/abs/1502.03167v3. Accessed January 5, 2021.

68. Ioffe S. Batch Renormalization: Towards Reducing Minibatch Dependence in Batch-Normalized Models. *Adv Neural Inf Process Syst*. 2017;2017-December:1946-1954. http://arxiv.org/abs/1702.03275. Accessed March 16, 2021.

69. Shelhamer E, Long J, Darrell T. Fully Convolutional Networks for Semantic Segmentation. *IEEE Trans Pattern Anal Mach Intell*. 2017;39(4):640-651. doi:10.1109/TPAMI.2016.2572683

70. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Vol 2016-December. IEEE Computer Society; 2016:770-778. doi:10.1109/CVPR.2016.90

71. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol 9351. Springer Verlag; 2015:234-241. doi:10.1007/978-3-319-24574-4_28

72. Abadi M, Agarwal A, Barham P, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv Prepr arXiv160304467*. 2016.

73. Kingma D, Ba J. Adam: A method for stochastic optimization. *arXiv Prepr arXiv14126980*. 2014.

74. He K, Zhang X, Ren S, Sun J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: *Proceedings of the IEEE International Conference on Computer Vision*. Institute of Electrical and Electronics Engineers Inc.; 2015:1026-1034. doi:10.1109/ICCV.2015.123

75. Wilcoxon F. Individual Comparisons by Ranking Methods. *Biometrics Bull*. 1945;1(6):80. doi:10.2307/3001968

76. Arabi H, Dowling JA, Burgos N, et al. Comparative study of algorithms for synthetic CT generation from MRI : Consequences for MRI -guided radiation planning in the pelvic region. *Med Phys*. 2018;45(11):5218-5233. doi:10.1002/mp.13187

77. Kim J, Glide-Hurst C, Doemer A, Wen N, Movsas B, Chetty IJ. Implementation of a Novel Algorithm For Generating Synthetic CT Images From Magnetic Resonance Imaging Data Sets for Prostate Cancer Radiation Therapy. *Int J Radiat Oncol Biol Phys*. 2015;91(1):39-47. doi:10.1016/j.ijrobp.2014.09.015

78. Andreasen D, Van Leemput K, Edmund JM. A patch-based pseudo-CT approach for MRI-only radiotherapy in the pelvis. *Med Phys*. 2016;43(8):4742-4752. doi:10.1118/1.4958676

79. Andreasen D, Edmund JM, Zografos V, Menze BH, Van Leemput K. Computed tomography synthesis from magnetic resonance images in the pelvis using multiple

random forests and auto-context features. In: Styner MA, Angelini ED, eds. Vol 9784. International Society for Optics and Photonics; 2016:978417. doi:10.1117/12.2216924

80. Mutic S, Palta JR, Butker EK, et al. Quality assurance for computed-tomography simulators and the computed-tomography-simulation process: Report of the AAPM Radiation Therapy Committee Task Group No. 66. *Med Phys*. 2003;30(10):2762-2792. doi:10.1118/1.1609271

81. Khoo VS, Joon DL. New developments in MRI for target volume delineation in radiotherapy. *Br J Radiol*. 2006;79(special_issue_1):S2-S15. doi:10.1259/bjr/41321492

82. Villeirs GM, L.Verstraete K, De Neve WJ, De Meerleer GO. Magnetic resonance imaging anatomy of the prostate and periprostatic area: a guide for radiotherapists. *Radiother Oncol*. 2005;76(1):99-106. doi:10.1016/J.RADONC.2005.06.015

83. Lim K, Small W, Portelance L, et al. Consensus Guidelines for Delineation of Clinical Target Volume for Intensity-Modulated Pelvic Radiotherapy for the Definitive Treatment of Cervix Cancer. *Int J Radiat Oncol*. 2011;79(2):348-355. doi:10.1016/j.ijrobp.2009.10.075

84. Heerkens HD, Hall WA, Li XA, et al. Recommendations for MRI-based contouring of gross tumor volume and organs at risk for radiation therapy of pancreatic cancer. *Pract Radiat Oncol*. 2017;7(2):126-136. doi:10.1016/J.PRRO.2016.10.006

85. Mittauer K, Paliwal B, Hill P, et al. A New Era of Image Guidance with Magnetic Resonance-guided Radiation Therapy for Abdominal and Thoracic Malignancies. *Cureus*. 2018;10(4):e2422. doi:10.7759/cureus.2422

86. Gudur MSR, Hara W, Le Q-T, Wang L, Xing L, Li R. A unifying probabilistic Bayesian

approach to derive electron density from MRI for radiation therapy treatment planning. *Phys Med Biol*. 2014;59(21):6595-6606. doi:10.1088/0031-9155/59/21/6595

87. Wolterink JM, Dinkla AM, Savenije MHF, Seevinck PR, van den Berg CAT, Išgum I. Deep MR to CT Synthesis Using Unpaired Data. In: Springer, Cham; 2017:14-23. doi:10.1007/978-3-319-68127-6_2

88. Maspero M, Savenije MHF, Dinkla AM, et al. Dose evaluation of fast synthetic-CT generation using a generative adversarial network for general pelvis MR-only radiotherapy. *Phys Med Biol*. 2018;63(18):185001. doi:10.1088/1361-6560/aada6d

89. Lei Y, Harms J, Wang T, et al. MRI-only based synthetic CT generation using dense cycle consistent generative adversarial networks. *Med Phys*. 2019;46(8):3565-3581. doi:10.1002/mp.13617

90. Liu Y, Lei Y, Wang T, et al. MRI-based treatment planning for liver stereotactic body radiotherapy: validation of a deep learning-based synthetic CT generation method. *Br J Radiol*. 2019;92(1100):20190067. doi:10.1259/bjr.20190067

91. Low DA, Harms WB, Mutic S, Purdy JA. A technique for the quantitative evaluation of dose distributions. *Med Phys*. 1998;25(5):656-661. doi:10.1118/1.598248

92. Ostrom QT, Gittleman H, Liao P, et al. CBTRUS Statistical Report: Primary brain and other central nervous system tumors diagnosed in the United States in 2010–2014. *Neuro Oncol*. 2017;19(suppl_5):v1-v88. doi:10.1093/neuonc/nox158

93. Louis DN, Perry A, Reifenberger · Guido, et al. The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary. *Acta Neuropathol*. 2016;3:803-820. doi:10.1007/s00401-016-1545-1

94. Koshy M, Villano JL, Dolecek TA, et al. Improved survival time trends for glioblastoma using the SEER 17 population-based registries. *J Neurooncol*. 2012;107(1):207-212. doi:10.1007/s11060-011-0738-7

95. TAMIMI AF, JUWEID M. Epidemiology and Outcome of Glioblastoma. In: *Glioblastoma*. Codon Publications; 2017:143-153. doi:10.15586/codon.glioblastoma.2017.ch8

96. Stupp R, Hegi ME, Mason WP, et al. Effects of radiotherapy with concomitant and adjuvant temozolomide versus radiotherapy alone on survival in glioblastoma in a randomised phase III study: 5-year analysis of the EORTC-NCIC trial. *Lancet Oncol*. 2009;10(5):459-466. doi:10.1016/S1470-2045(09)70025-7

97. Niyazi M, Brada M, Chalmers AJ, et al. ESTRO-ACROP guideline "target delineation of glioblastomas." *Radiother Oncol*. 2016;118(1):35-42. doi:10.1016/j.radonc.2015.12.003

98. Cabrera AR, Kirkpatrick JP, Fiveash JB, et al. Radiation therapy for glioblastoma: Executive summary of an American Society for Radiation Oncology Evidence-Based Clinical Practice Guideline. *Pract Radiat Oncol*. 2016;6(4):217-225. doi:10.1016/j.prro.2016.03.007

99. Kamnitsas K, Bai W, Ferrante E, et al. Ensembles of multiple models and architectures for robust brain tumour segmentation. In: *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol 10670 LNCS. Springer Verlag; 2018:450-462. doi:10.1007/978-3-319-75238-9_38

100. Menze BH, Jakab A, Bauer S, et al. The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE Trans Med Imaging*. 2015;34(10):1993-2024.

doi:10.1109/TMI.2014.2377694

101. Bakas S, Akbari H, Sotiras A, et al. Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features. *Sci Data*. 2017;4:170117. doi:10.1038/sdata.2017.117

102. Bakas S, Reyes M, Jakab A, et al. Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge. November 2018. http://arxiv.org/abs/1811.02629. Accessed October 23, 2019.

103. Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Vol 07-12-June-2015. IEEE Computer Society; 2015:3431-3440. doi:10.1109/CVPR.2015.7298965

104. Kamnitsas K, Ledig C, Newcombe VFJ, et al. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Med Image Anal*. 2017;36:61-78. doi:10.1016/j.media.2016.10.004

105. Myronenko A. 3D MRI brain tumor segmentation using autoencoder regularization. In: *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol 11384 LNCS. Springer Verlag; 2019:311-320. doi:10.1007/978-3-030-11726-9_28

106. Zhang R, Zhao L, Lou W, et al. Automatic Segmentation of Acute Ischemic Stroke From DWI Using 3-D Fully Convolutional DenseNets. *IEEE Trans Med Imaging*. 2018;37(9):2149-2160. doi:10.1109/TMI.2018.2821244

107. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*. Vol 2017-January. Institute of Electrical and Electronics Engineers Inc.; 2017:2261-2269. doi:10.1109/CVPR.2017.243

108. Ulyanov D, Vedaldi A, Lempitsky V. Instance Normalization: The Missing Ingredient for Fast Stylization. July 2016. http://arxiv.org/abs/1607.08022. Accessed March 24, 2020.

109. Bridle JS. Probabilistic Interpretation of Feedforward Classification Network Outputs, with Relationships to Statistical Pattern Recognition. In: *Neurocomputing*. Springer Berlin Heidelberg; 1990:227-236. doi:10.1007/978-3-642-76153-9_28

110. Hu J, Shen L, Sun G. Squeeze-and-Excitation Networks. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society; 2018:7132-7141. doi:10.1109/CVPR.2018.00745

111. Jiang Z, Ding C, Liu M, Tao D. Two-stage cascaded u-net: 1st place solution to brats challenge 2019 segmentation task. In: *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol 11992 LNCS. Springer; 2020:231-241. doi:10.1007/978-3-030-46640-4_22

112. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2019. *CA Cancer J Clin*. 2019;69(1):7-34. doi:10.3322/caac.21551

113. Kapiteijn E, Marijnen CAM, Nagtegaal ID, et al. Preoperative Radiotherapy Combined with Total Mesorectal Excision for Resectable Rectal Cancer. *N Engl J Med*. 2001;345(9):638-646. doi:10.1056/NEJMoa010580

114. van de Velde CJH, Boelens PG, Borras JM, et al. EURECCA colorectal: Multidisciplinary

management: European consensus conference colon & rectum. *Eur J Cancer*. 2014;50(1):1.e1-1.e34. doi:10.1016/J.EJCA.2013.06.048

115. Quah H, Chou JF, Gonen M, et al. Pathologic stage is most prognostic of disease-free survival in locally advanced rectal cancer patients after preoperative chemoradiation. *Cancer*. 2008;113(1):57-64. doi:10.1002/cncr.23516

116. Trakarnsanga A, Gönen M, Shia J, et al. Comparison of Tumor Regression Grade Systems for Locally Advanced Rectal Cancer After Multimodality Treatment. *JNCI J Natl Cancer Inst*. 2014;106(10). doi:10.1093/jnci/dju248

117. Maas M, Nelemans PJ, Valentini V, et al. Long-term outcome in patients with a pathological complete response after chemoradiation for rectal cancer: a pooled analysis of individual patient data. *Lancet Oncol*. 2010;11(9):835-844. doi:10.1016/S1470-2045(10)70172-8

118. Schurink NW, Lambregts DMJ, Beets-Tan RGH. Diffusion-weighted imaging in rectal cancer: current applications and future perspectives. *Br J Radiol*. 2019;92(1096):20180655. doi:10.1259/bjr.20180655

119. Beets-Tan RGH, Lambregts DMJ, Maas M, et al. Magnetic resonance imaging for clinical management of rectal cancer: Updated recommendations from the 2016 European Society of Gastrointestinal and Abdominal Radiology (ESGAR) consensus meeting. *Eur Radiol*. 2018;28(4):1465-1475. doi:10.1007/s00330-017-5026-2

120. Kim SH, Lee JY, Lee JM, Han JK, Choi BI. Apparent diffusion coefficient for evaluating tumour response to neoadjuvant chemoradiation therapy for locally advanced rectal cancer. *Eur Radiol*. 2011;21(5):987-995. doi:10.1007/s00330-010-1989-y

121. Amodeo S, Rosman AS, Desiato V, et al. MRI-Based Apparent Diffusion Coefficient for Predicting Pathologic Response of Rectal Cancer After Neoadjuvant Therapy: Systematic Review and Meta-Analysis. *Am J Roentgenol*. 2018;211(5):W205-W216. doi:10.2214/AJR.17.19135

122. Lambin P, Leijenaar RTH, Deist TM, et al. Radiomics: the bridge between medical imaging and personalized medicine. *Nat Rev Clin Oncol*. 2017;14(12):749-762. doi:10.1038/nrclinonc.2017.141

123. Nie K, Shi L, Chen Q, et al. Rectal Cancer: Assessment of Neoadjuvant Chemoradiation Outcome based on Radiomics of Multiparametric MRI. *Clin Cancer Res*. 2016;22(21):5256-5264. doi:10.1158/1078-0432.CCR-15-2997

124. Horvat N, Veeraraghavan H, Khan M, et al. MR Imaging of Rectal Cancer: Radiomics Analysis to Assess Treatment Response after Neoadjuvant Therapy. *Radiology*. 2018;287(3):833-843. doi:10.1148/radiol.2018172300

125. Afshar P, Mohammadi A, Plataniotis KN, Oikonomou A, Benali H. From Handcrafted to Deep-Learning-Based Cancer Radiomics: Challenges and opportunities. *IEEE Signal Process Mag*. 2019;36(4):132-160. doi:10.1109/MSP.2019.2900993

126. Wang S, Liu Z, Rong Y, et al. Deep learning provides a new computed tomography-based prognostic biomarker for recurrence prediction in high-grade serous ovarian cancer. *Radiother Oncol*. 2019;132:171-177. doi:10.1016/j.radonc.2018.10.019

127. Lao J, Chen Y, Li Z-C, et al. A Deep Learning-Based Radiomics Model for Prediction of Survival in Glioblastoma Multiforme. *Sci Rep*. 2017;7(1):10353. doi:10.1038/s41598-017-10649-8

128. van Griethuysen JJM, Fedorov A, Parmar C, et al. Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer Res*. 2017;77(21):e104-e107. doi:10.1158/0008-5472.CAN-17-0339

129. Russakovsky O, Deng J, Su H, et al. ImageNet Large Scale Visual Recognition Challenge. *Int J Comput Vis*. 2015;115(3):211-252. doi:10.1007/s11263-015-0816-y

130. Tibshirani R, Tibshirani R. Regression Shrinkage and Selection Via the Lasso. *J R Stat Soc Ser B*. 1994;58:267--288. http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.35.7574. Accessed August 19, 2019.

131. Wu TT, Chen YF, Hastie T, Sobel E, Lange K. Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*. 2009;25(6):714-721. doi:10.1093/bioinformatics/btp041

132. Bouckaert RR, Frank E. Evaluating the Replicability of Significance Tests for Comparing Learning Algorithms. In: Springer, Berlin, Heidelberg; 2004:3-12. doi:10.1007/978-3-540-24775-3_3

133. Kim J-H. Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Comput Stat Data Anal*. 2009;53(11):3735-3745. doi:10.1016/J.CSDA.2009.04.009

134. Pavic M, Bogowicz M, Würms X, et al. Influence of inter-observer delineation variability on radiomics stability in different tumor sites. *Acta Oncol (Madr)*. 2018;57(8):1070-1074. doi:10.1080/0284186X.2018.1445283

135. Fiset S, Welch ML, Weiss J, et al. Repeatability and reproducibility of MRI-based

radiomic features in cervical cancer. *Radiother Oncol*. 2019;135:107-114. doi:10.1016/j.radonc.2019.03.001

136. Huynh BQ, Li H, Giger ML. Digital mammographic tumor classification using transfer learning from deep convolutional neural networks. *J Med Imaging*. 2016;3(3):034501. doi:10.1117/1.JMI.3.3.034501

137. Deep Feature Transfer Learning in Combination with Traditional Features Predicts Survival Among Patients with Lung Adenocarcinoma. *Tomography*. 2016;2(4):388-395. doi:10.18383/j.tom.2016.00211

138. Ostrom QT, Bauchet L, Davis FG, et al. The epidemiology of glioma in adults: a "state of the science" review. *Neuro Oncol*. 2014;16(7):896-913. doi:10.1093/neuonc/nou087

139. Nicolasjilwan M, Hu Y, Yan C, et al. Addition of MR imaging features and genetic biomarkers strengthens glioblastoma survival prediction in TCGA patients. *J Neuroradiol*. 2015;42(4):212-221. doi:10.1016/j.neurad.2014.02.006

140. Sanghani P, Ang BT, King NKK, Ren H. Regression based overall survival prediction of glioblastoma multiforme patients using a single discovery cohort of multi-institutional multi-channel MR images. *Med Biol Eng Comput*. 2019;57(8):1683-1691. doi:10.1007/s11517-019-01986-z

141. Long J, Shelhamer E, Darrell T. *Fully Convolutional Networks for Semantic Segmentation*. https://people.eecs.berkeley.edu/~jonlong/long_shelhamer_fcn.pdf. Accessed September 11, 2018.

142. Shboul ZA, Alam M, Vidyaratne L, Pei L, Elbakary MI, Iftekharuddin KM. Feature-

Guided Deep Radiomics for Glioblastoma Patient Survival Prediction. *Front Neurosci*. 2019;13:966. doi:10.3389/fnins.2019.00966

143. Deng J, Dong W, Socher R, Li L-J, Kai Li, Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In: Institute of Electrical and Electronics Engineers (IEEE); 2010:248-255. doi:10.1109/cvpr.2009.5206848

144. Camp RL, Dolled-Filhart M, Rimm DL. X-tile: A new bio-informatics tool for biomarker assessment and outcome-based cut-point optimization. *Clin Cancer Res*. 2004;10(21):7252-7259. doi:10.1158/1078-0432.CCR-04-0713

145. Porz N, Bauer S, Pica A, et al. Multi-Modal Glioblastoma Segmentation: Man versus Machine. Strack S, ed. *PLoS One*. 2014;9(5):e96873. doi:10.1371/journal.pone.0096873

146. Ghosal P, Reddy S, Sai C, Pandey V, Chakraborty J, Nandi D. A Deep Adaptive Convolutional Network for Brain Tumor Segmentation from Multimodal MR Images. In: *TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON)*. IEEE; 2019:1065-1070. doi:10.1109/TENCON.2019.8929402

147. Akinyemiju T, Abera S, Ahmed M, et al. The Burden of Primary Liver Cancer and Underlying Etiologies From 1990 to 2015 at the Global, Regional, and National Level. *JAMA Oncol*. 2017;3(12):1683. doi:10.1001/jamaoncol.2017.3055

148. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2018;68(6):394-424. doi:10.3322/caac.21492

149. Poulou LS, Botsa E, Thanou I, Ziakas PD, Thanos L. Percutaneous microwave ablation vs radiofrequency ablation in the treatment of hepatocellular carcinoma. *World J Hepatol*.

2015;7(8):1054-1063. doi:10.4254/wjh.v7.i8.1054

150. Brunner T, Andratschke N, Gerum S, et al. OC-0424: SBRT for Primary Liver Cancer in Routine Clinical Practice: A Patterns-of-Care and Outcome Analysis. *Radiother Oncol*. 2017;123:S224. doi:10.1016/S0167-8140(17)30866-6

151. Daher S, Massarwa M, Benson AA, Khoury T. Current and Future Treatment of Hepatocellular Carcinoma: An Updated Comprehensive Review. *J Clin Transl Hepatol*. 2018;6(1):69-78. doi:10.14218/JCTH.2017.00031

152. Mendiratta-Lala M, Gu E, Owen D, et al. Imaging Findings Within the First 12 Months of Hepatocellular Carcinoma Treated With Stereotactic Body Radiation Therapy. *Int J Radiat Oncol.* 2018;102(4):1063-1069. doi:10.1016/j.ijrobp.2017.08.022

153. Schaub SK, Hartvigson PE, Lock MI, et al. Stereotactic Body Radiation Therapy for Hepatocellular Carcinoma: Current Trends and Controversies. *Technol Cancer Res Treat*. 17:1-19. doi:10.1177/1533033818790217