

Spring 2019

Machine Learning Solution to Organ-At-Risk Segmentation for Radiation Treatment Planning

Brie Goo

Katrina May

Haobo Zhang

James Olivas

Follow this and additional works at: https://scholarcommons.scu.edu/idp_senior

 Part of the [Biomedical Engineering and Bioengineering Commons](#), and the [Computer Engineering Commons](#)

SANTA CLARA UNIVERSITY

Department of Bioengineering
& Department of Computer Engineering

I HEREBY RECOMMEND THAT THE THESIS PREPARED
UNDER MY SUPERVISION BY

Brie Goo, Katrina May, Haobo Zhang, and James Olivas

ENTITLED

**MACHINE LEARNING SOLUTION TO ORGANS-AT-RISK
SEGMENTATION FOR RADIATION TREATMENT
PLANNING**

BE ACCEPTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF

**BACHELOR OF SCIENCE
IN
BIOENGINEERING
AND
BACHELOR OF SCIENCE
IN
COMPUTER SCIENCE & ENGINEERING**

Thesis Advisor(s) (use separate line for each advisor)

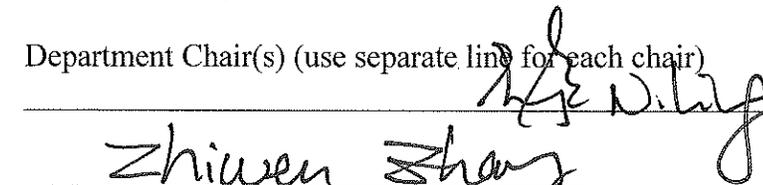
date



6/12/19
6/13/19

Department Chair(s) (use separate line for each chair)

date



6/13/19
6/13/19

MACHINE LEARNING SOLUTION TO ORGANS-AT-RISK SEGMENTATION FOR RADIATION TREATMENT PLANNING

By

Brie Goo, Katrina May, Haobo Zhang, and James Olivas

SENIOR DESIGN PROJECT REPORT

Submitted to
the Department of Bioengineering
and
the Department of Computer Science & Engineering

of

SANTA CLARA UNIVERSITY

in Partial Fulfillment of the Requirements
for the degree of
Bachelor of Science in Bioengineering
and
Bachelor of Science in Computer Science & Engineering

Santa Clara, California

Spring 2019

Abstract

In the treatment of cancer using ionizing radiation, it is important to design a treatment plan such that dose to normal, healthy organs is sufficiently low. Today, segmentation requires a trained human to carefully outline, or segment, organs on each slice of a treatment planning computed tomography (CT) scan but it is laborious, time-consuming, and contains intra- and inter-rater variability. Currently, existing clinical automation technology relies on atlas-based automation, which has limited segmentation accuracy. Thus the auto-segmentations require post process editing by an expert. In this paper, we propose a machine learning solution that shortens the segmentation time of organs-at-risk (OARs) in the thoracic cavity. The overall system will include preprocessing, model processing, and postprocessing steps to make the system easily integratable into the radiotherapy planning process. For our model, we chose to use a 3D deep convolutional neural network with a U-net based architecture because this machine learning strategy takes into account local spatial relationships, will restore the original image resolution and has been utilized in image segmentation, especially in medical image analysis. Training and testing were done with a 60 patient dataset of thoracic CT scans from the AAPM 2017 Grand Challenge. To assess and improve our system we calculated accuracy metrics (Dice similarity coefficient (DSC), mean surface distance (MSD)) and compared our model's segmentation performance to that of an expert and the top two performing machine learning methods of the challenge. We explored using preprocessing steps such as cropping and image enhancement to improve the model segmentation accuracy. Our final model was able to segment the lungs as accurately as a dosimetrist and the heart and spinal cord within acceptable DSC ranges. All DSC values of the OARs from our method were as accurate as other machine learning methods. The DSC for the esophagus was below tolerable error for radiotherapy planning, but our mean surface distance was superior to other auto-segmentation methods. We were successful in significantly reducing manual segmentation time by developing a machine learning system. Though our approach still necessitates a single preparatory step of manually cropping anatomical regions to isolate segmentation volume, a general hospital technician could complete this task which removes the need of an expert for one time-consuming step of radiotherapy planning. Implementation of our methods to provide radiotherapy in lower-middle income countries brings us closer to accessibility of treatment for a wider population.

Acknowledgments

We would like to thank our advisors, Dr. Julia A. Scott with the Department of Bioengineering and Dr. Ying Liu with the Department of Computer Engineering, for their support and guidance throughout this project. We also extend our thanks to Dr. Prashanth Asuri with the Bioinnovation Lab at Santa Clara University for connecting us with Varian Medical Systems and their representatives Dr. Daren Sawkey and Dr. Anthony Lujan. Thank you Dr. Daren Sawkey, Dr. Anthony Lujan, and Dr. Prashanth Asuri for helping us brainstorm the project and it's needs to bring the project goal to fruition. Lastly, we'd like to thank the undergraduate students who participated in the development of our solution.

Table of Contents

| | |
|---|-----------|
| Abstract | 3 |
| Acknowledgments | 4 |
| 1. Introduction | 11 |
| 1.1 Problem Statement & Goal | 11 |
| 1.2 Motivations | 11 |
| 1.3 Background | 12 |
| 1.4 Objectives | 13 |
| 2. System Overview | 13 |
| 2.1 Conceptual Model | 13 |
| 2.2 Customer Needs | 13 |
| 2.3 System Level Requirements | 14 |
| 2.4 Use Case | 14 |
| 2.4.1 Perform CT scans of patients | 15 |
| 2.4.2 Run data through algorithm to output segmented labels | 15 |
| 2.4.3 Radiotherapy planning | 15 |
| 3. Main Function | 15 |
| 3.1 DCNN Model: 3D U-Net | 15 |
| 3.2 Model Parameters | 16 |
| 3.2.1 Number of Model Features | 16 |
| 3.2.2 Number of Epochs | 17 |
| 3.2.3 Preprocessing Filter | 17 |
| 3.2.4 Cropping | 17 |
| 3.3 Dataset | 17 |
| 4. Subsystem Functions | 18 |
| 4.1 Hardware Configuration | 18 |
| 4.2 Software Functions | 19 |
| 4.2.1 File Conversion | 19 |
| 4.2.2 Cropping | 19 |
| 4.2.3 Filters | 20 |
| Bilateral Mean and Contrast Stretching | 20 |
| Local Equalization | 20 |
| 4.3.2 Producing Label Map | 21 |
| 4.4.3 Save in DICOM Format | 21 |

| | |
|--|-----------|
| 5. Testing | 21 |
| 5.1 Test Phase 1: Downsampled Images | 22 |
| 5.1.1 Data Input | 22 |
| 5.1.2 Calculating Accuracy | 22 |
| 5.1.3 Visualization | 22 |
| 5.2 Test Phase 2: Cropping Organs | 23 |
| 5.2.1 Cropping | 23 |
| 5.2.3. Visualization | 23 |
| 5.3 Test Phase 3: Image Enhancement using Local Equalization Filter | 23 |
| 5.3.1 Local Equalization Filter | 24 |
| 5.4 Test Phase 4: Image enhancement using Bilateral Mean and Contrast Stretching | 24 |
| 5.4.1 Bilateral Mean and Contrast Stretch Filter | 24 |
| 5.4 System Level Issues | 25 |
| 5.5 Options and Trade-Offs | 25 |
| 5.6 Test Phase Parameters | 26 |
| 6. Accuracy Metrics | 26 |
| 6.1 Dice Similarity Coefficient (DSC Score) | 26 |
| 6.2 Hausdorff Distance 95% (HD95) | 27 |
| 6.3 Mean Surface Distance (MSD) | 27 |
| 7. Results | 27 |
| 7.1 Comparative Quantitative Performance to External Methods | 27 |
| 7.2 Subsystem to Subsystem Comparison | 28 |
| 7.3 Qualitative Performance | 30 |
| 8. Discussion | 31 |
| 8.1 Meeting our Requirements | 31 |
| 8.1.1 Segmentation Time | 31 |
| 8.1.2 Accuracy | 31 |
| 8.1.3 Functional Requirements | 33 |
| 8.2 Project Challenges and Constraints | 33 |
| 8.3 Risks and Mitigations | 34 |
| 8.4 Societal Issues | 34 |
| 9. Conclusion | 35 |
| Appendix | 37 |
| Team Approach | 37 |

| | |
|-------------------|-----------|
| Key Lessons | 37 |
| Budget | 37 |
| Timeline | 39 |
| References | 39 |

List of Figures

| | | |
|-----|--|----|
| 2.1 | Conceptual model of the workflow..... | 13 |
| 2.4 | Example use case..... | 14 |
| 3.1 | U-Net architecture as proposed by Ronneberger et al., 2015..... | 16 |
| 4.1 | Chassis set-up for all of our team’s computing needs..... | 19 |
| 4.2 | Image of cropping GUI..... | 20 |
| 5.1 | A screenshot of a random transverse plane from Test 1..... | 22 |
| 5.3 | Local histogram equalization filter..... | 24 |
| 5.3 | Histogram of bilateral mean and contrast stretching filters..... | 25 |
| 7.2 | Box plots of internal accuracy tests..... | 29 |
| 7.2 | Training DSC loss as a function of epochs versus relative DSC score..... | 30 |
| 7.3 | Contouring Results displayed at 4 axial slices..... | 31 |
| 8.1 | 3D visualization of segmentation in 3D Slicer..... | 33 |

List of Tables

| | | |
|-----|---|----|
| 1.2 | Cancer incidence predicted in higher-income countries versus LMICs..... | 12 |
| 2.3 | Lists of requirements and constraints..... | 14 |
| 3.3 | Characteristics of data set provided by the 2017 AAPM Grand Challenge..... | 18 |
| 5 | Summary of subsystem item changes from Tests 1-4..... | 22 |
| 5.6 | Summary of test phase attributes for all organs and for the esophagus..... | 26 |
| 7.1 | Metric results for models in each testing phase with benchmark..... | 28 |
| 7.1 | Interrater differences in segmentation of organs at risk (OARs)..... | 28 |
| 8.1 | Comparison of our methods (SCU), University of Virginia (UV), and Elekta..... | 32 |

Glossary of Terms

1. LMIC: Lower and middle income countries
2. RT: Radiotherapy
3. OARs: Organs-at-risk
4. DCNN: Deep Convolutional Neural Net
5. Dosimetrist: The medical dosimetrist is responsible for developing a radiotherapy treatment plan by means of computer and/or manual computation to determine a treatment field technique that will deliver that prescribed radiation dose. When designing that plan, also taken into consideration are the dose-limiting structures.
6. AAPM Challenge: American Association for Physical Medicine Grand Challenge: Auto-Segmentation for Thoracic Radiation Treatment Planning
7. DICOM: Digital Imaging and Communications in Medicine
8. RTSTRUCT: Radiotherapy structure set
9. DSC: Dice Similarity Coefficient
10. HD95: Hausdorff distance
11. MSD: Mean surface distance
12. BM&CS: Bilateral mean and contrast enhancement
13. LE: Local histogram equalization filter

1. Introduction

1.1 Problem Statement & Goal

Radiation therapy, a key component of cancer management, is required in more than half of new cancer patients, particularly in low- and middle-income countries (LMICs)[1]. For safe and effective radiotherapy (RT) treatment, it is crucial to accurately segment organs-at-risks (OARs) to minimize radiation exposure to these healthy tissues. Current practice necessitates expert manual delineation of OARs, an arduous and labor intensive task with variable accuracy. RT is not an accessible treatment option in many LMICs because the lack of trained professionals or radiotherapy units. To address these issues, we propose a deep convolutional neural network (DCNN) machine learning-based algorithm to automate the segmentation of OARs in thoracic CT images. We aim to (i) segment the organs quicker than the average manual segmentation time, (ii) segment as accurately as a dosimetrist, and (iii) fully automate the segmentation process for the thoracic cavity. Successful demonstration of this method for thoracic CT will lay a foundation for a generalizable machine learning strategy of OAR segmentation integrated into radiotherapy planning.

1.2 Motivations

With a growing number of cancer incidences, there is an increasing need for access to radiation treatment. For effective treatment and to minimize post-treatment complications, OARs, such as lungs, heart, esophagus and spinal cord, must be accurately delineated. Currently, manual segmentation by high-level expertise is the gold standard for OAR segmentation. However, the complexity of OARs morphology and imperfection of imaging devices make manual delineation prone to errors and time-consuming--an expert can spend two or more hours on a single case [2]. This can cause clinically significant delays to treatment commencement which has shown to be associated with increased risk of both local recurrence and overall mortality [3]. Large inter- and intra-rater variability of manual segmentation impacts the measurement of radiation an expert (dosimetrist) calculates to administer to the patient [3,4]. Therefore there is a high demand for reliably accurate OAR delineation and to considerable reduce the amount of manual labor in treatment planning [4].

Relying on manual segmentation is especially an issue in developing nations that do not have access to expertise. In select international partnerships, RT planning is outsourced to regions with expertise for treatment locally [5]. While a charitable model, it cannot be effectively scaled to meet the growing need for RT worldwide. Populations in LMIC face an expected rise in annual cancer incidence of nearly 70% by 2030 over the 2010 rates [6] (Table 1). By 2020,

these LMICs would need an additional 9,169 teletherapy units, 12,149 radiation oncologists, 9,915 medical physicists, and 29,140 radiation therapy technologists [1]. Automating the segmentation process is a viable cost effective solution to feasibly upscale access to RT worldwide to improve survival rates and provide the treatment millions of people deserve.

Table 1. Cancer incidence predicted in higher-income countries versus LMICs [6].

| | 2010 | 2020 | 2030 |
|-------------------------|-----------|-----------|------------|
| Higher-income Countries | 5,719,728 | 6,583,577 | 7,425,611 |
| LMICs | 7,521,150 | 9,917,509 | 12,876,263 |

1.3 Background

Automated methods for multi-organ segmentation has shown its potential for clinical use with high efficiency [7]. However, current automation methods still have their drawbacks. Atlas-based automation has become a standard paradigm in medical image segmentation for exploiting prior anatomical knowledge. The atlas is a reference image in which structures of interest have been carefully segmented, usually by hand. One of the main advantages of atlas-based methods compared to manual segmentation, is that it easily estimates, in the patient image, the position of structures with fuzzy or not visible contours. This saves considerable time during RT planning. That being said, this approach is not accurate enough to fully automate segmentation. It still requires editing and review by an expert to avoid risk of incorrect dosage [8,9].

DCNNs are another method of automation. The increasing computational power of modern hardware platforms, including GPUs, has allows auto-segmentation to be typically done in a range of a few minutes. Studies have shown that DCNNs provide significantly better accuracy than atlas-based methods [10]. Machine learning methods are competitive with standard image processing algorithms in the field of organ segmentation [11].

One of the main limitations of using DCNN auto-segmentation methods is the lack of sufficient soft tissue contrast that compromises accurate segmentation of critical anatomical structures in the path of radiation beams [4]. This is particularly an issue for soft tissue with irregular morphology. Denoising filters and contrast enhancement can provide better visibility of soft tissue boundaries. DCNN models show promising accuracy results with the addition of image enhancing filters to be able to contour smaller irregular soft tissues.

1.4 Objectives

Here we address these challenges by following our study design which includes (i) building a GPU computing system that is capable of processing our large dataset, (ii) create a 2D CNN followed by a 3D CNN, (iii) submit our results to 2017 American Association for Physical Medicine (AAPM) challenge to compare accuracy metrics to other auto-segmentation models, and (iv) modify our model to improve our accuracy results.

2. System Overview

2.1 Conceptual Model

The optimal model for the implemented system to have is an automated workflow for medical image segmentation (Figure 1). The input to the system would be a CT scan in DICOM (Digital Imaging and Communications in Medicine) a standard for handling, storing, printing, and transmitting information in medical imaging. The system would perform testing on the image which produces a label map of the five different structures in the thoracic cavity: the right lung, left lung, heart, spinal cord and esophagus. The label map is combined with the original CT scan and saved in DICOM format. The segmentation can then be used in the next step of RT planning.

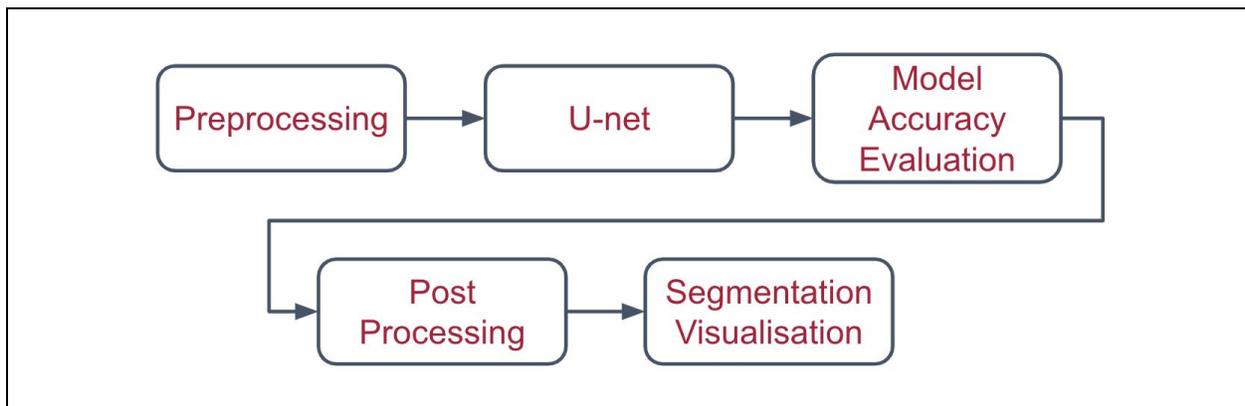


Figure 1. Conceptual Model for auto-segmentation with machine learning.

2.2 Customer Needs

Our customer, Varian Medical Systems, has requested that we achieve auto-segmentation of OARs using a machine learning approach. The processing per scan must be significantly faster than a dosimetrist. Accelerating the OAR segmentation step of RT planning will help achieve their goal of a comprehensive, one-day process from scan to treatment.

2.3 System Level Requirements

The system needed to be able to segment OARs at an accuracy similar to that of a dosimetrist, which minimizes peripheral radiation damage to healthy organs. The output needs to be saved as a DICOM file format compatible with existing radiotherapy planning software.

Table 2. Lists of requirements from most least importance and constraints of our project.

| Functional Requirements | Non-Functional Requirements | Constraints |
|--|--|--|
| <ul style="list-style-type: none"> • Segment OARs • Compute accuracy metrics • Save in DICOM file format • Visualise results | <ul style="list-style-type: none"> • < 30 min segmentation time per case • Segmentation as accurate as an expert • Simple user interface | <ul style="list-style-type: none"> • Number of training and test images • Memory limitations of our computer • Time |

2.4 Use Case

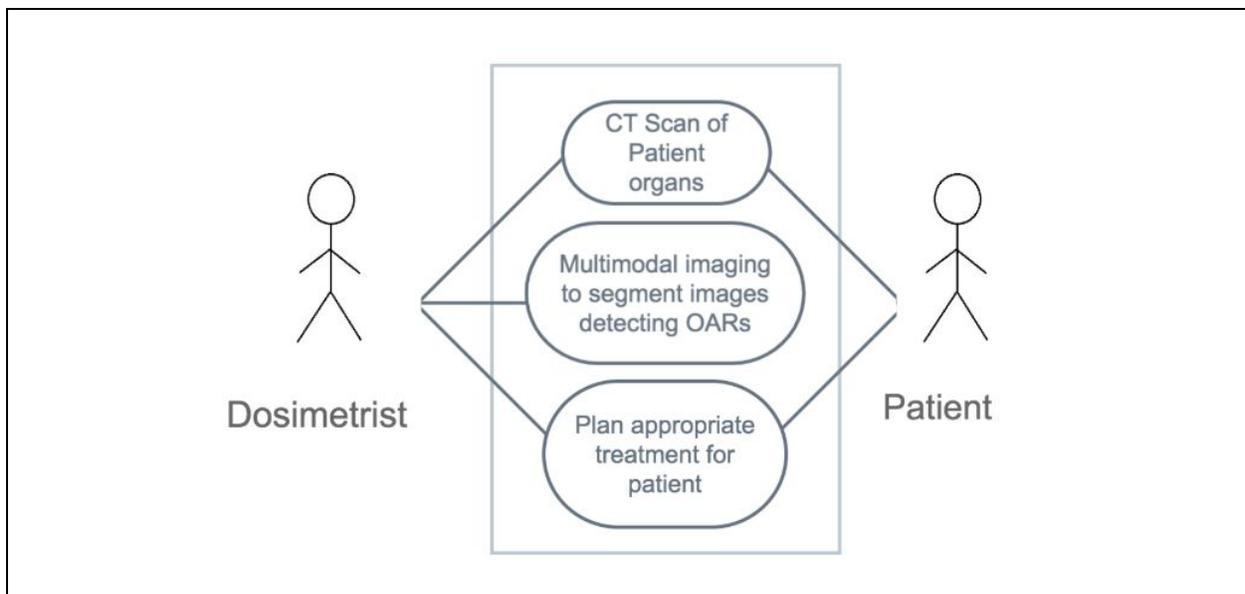


Figure 2. Example use case for our machine learning solution.

2.4.1 Perform CT scans of patients

The patient is scanned in the Radiology department to perform the scans from the CT machine and receive the scans in a DICOM format. The data is then used by the dosimetrist for further processing.

2.4.2 Run data through algorithm to output segmented labels

The images are then run through the machine learning algorithm to produce an auto-segmented image. These images will have the identified organs labeled in a separate file that can be converted back to a DICOM or another medical imaging format. The organs are cropped and each structure is individually ran through the model, each forming a separate binary anatomical label. These labels are then combined and overlaid on the original CT image in a visualization software.

2.4.3 Radiotherapy planning

With correctly labeled organs from the model, the dosimetrist can use DICOM to plan an appropriate treatment for the patient. The OAR labels are able to determine which regions to minimize radiation exposure. Because the organs are now labeled, the dosimetrist is able to calculate the intensity of radiation beam and orientation to only the tumor area. The patient will be able to receive treatment sooner by reducing the time required for this critical step of RT planning.

3. Main Function

3.1 DCNN Model: 3D U-Net

The main function of our workflow is based on the U-Net architecture model of CNN as proposed by Ronneberger et al., 2015 (Fig. 3). CNNs convert an image into a vector volume that is convolved by kernels in each layer that creates activation maps from the image. The U-Net up samples the activation maps to the original resolution. We can then return the vector to image form.

The U-Net model consists of three steps: contraction, bottleneck, and expansion. The contraction process involves a long series of contraction blocks, each of which applies two 3x3 CNN layers and doubles the number of feature maps. This is performed until the image has been compressed into a vector. This leads into the bottleneck stage of the process, which is where data is fed to the next step. The expansion step has the vector be put through a number of expansion blocks, one for each contraction block in the first step, which each apply two 3x3 CNN layers and then a 2x2 upsampling layer. The corresponding original image from the contraction block is

input to the expansion block to form a complete label. This overall architecture can be represented by a “U”, which is where its name comes from.

CNNs are a type of supervised machine learning method in which the features are automatically extracted without the need of any pre-processing. Comprised of several neural network layers, each layer is convolved with a set of kernels $W = \{W_1, W_2, \dots, W_K\}$ and added biases each generating a new feature map X_k . These features then are put through a non-linear transform $\sigma(\cdot)$ and the same process is repeated for each convolutional layer:

$$X_k^l = \sigma(W_k^{l-1} * X^{l-1} + b_k^{l-1}) \quad (1)$$

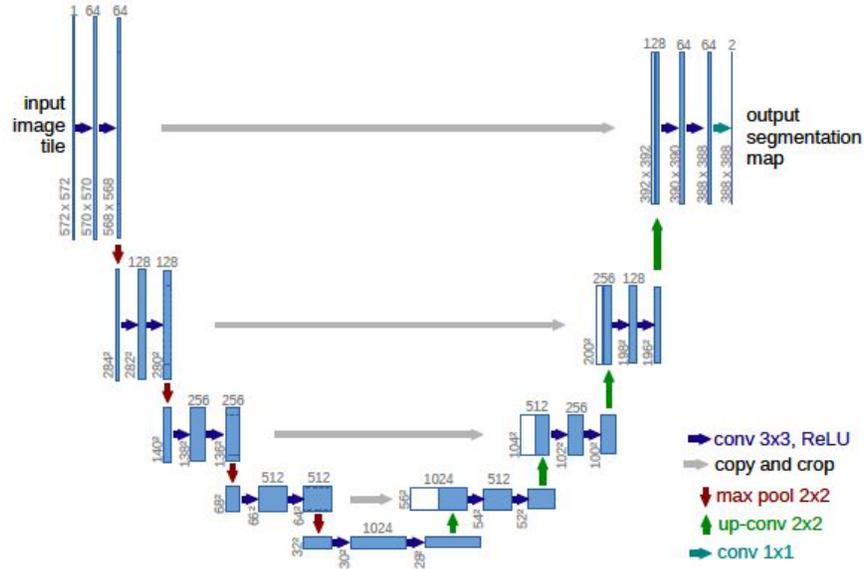


Figure 3. U-Net architecture as proposed by Ronneberger et al., 2015. Each blue box corresponds to a multi-channel feature map. The number of channels is denoted on top of the box. The x-y-size is provided at the lower left edge of the box. White boxes represent copied feature maps. The arrows denote the different operations.

3.2 Model Parameters

3.2.1 Number of Model Features

Using a higher number of features allows the model to extract more detailed information and patterns about the input image. However, as a trade-off, the more features added the more memory within our computer is used up, which can prove problematic when memory is limited.

3.2.2 Number of Epochs

One Epoch is when an entire dataset is passed forward and backward through the neural network only once. As the number of epochs increases, the accuracy should increase. However,

this will increase the time it takes to train our model. Also, there is a limit to how much the accuracy will increase; overfitting may occur if too many epochs are used which causes a decrease in accuracy. Therefore the optimal number of epochs needs to be calculated.

3.2.3 Preprocessing Filter

Image filters can be added to the image beforehand in order to enhance the features like contrast before training. Filters can also remove features that are deemed important in the model, so they need to be carefully selected.

3.2.4 Cropping

Organs can be cropped individually and selected as a portion to be trained. This will reduce memory usage so that a higher quality image can be used rather than downsampling the image resolution to fit the entire image through the model. The downside is that cropping for non-labeled data will need to either be done manually. Cropping inaccurately may remove some necessary features or cut off larger organs.

3.3 Dataset

The dataset used for the training and testing was accessed from the AAPM Grand Challenge of 2017. The dataset consists of 36 training images and 24 testing images; meaning one set of training data and one set of testing data, more information can be found in Table 2. The training images consisted of segmented data while the testing only contained the scans. We chose to use this data set because:

- i. Every image was of high quality eliminating the chance of low label prediction accuracy because of image quality.
- ii. It contained manual segmentations on the thoracic cavity CT scans, the exact region we selected to work on.
- iii. It was one of the few datasets that were relatively small in size making it better for experimentation within our time restrictions.
- iv. The data set was used with other companies and university's automation algorithms giving us the ability to compare our system results against other teams as well as the manual segmentations.

Table 3. Characteristics of data set provided by the 2017 AAPM Grand Challenge.

| Collection Statistics | Updated 2017/05/17 |
|-----------------------|--------------------|
| Modalities | CT, RT |
| Number of Patients | 60 |
| Number of Studies | 60 |
| Number of Series | 96 |
| Number of images | 9569 |
| Image Size (GB) | 4.8 |

4. Subsystem Functions

4.1 Hardware Configuration

Figure 4 shows the hardware configuration that was used to run the system. A GPU is capable of processing large data more efficient than running on the CPU making it a necessary part of the system. This is because the machine learning model is performed on CUDA cores of the GPU. The GPU we purchased, NVIDIA GeForce RTX 2080 GPU, contains 11 GB of RAM and 4352 CUDA cores for processing. The memory bandwidth is 616 GB/s, so this allows very fast memory transfers when running tests. Our operating system needed to be reliable and secure. For this, we were able to download the Windows 10 operating system through the University's resources. The PyCharm programming environment incorporated Git, Tensorflow with Keras, and Python for our model's needs.

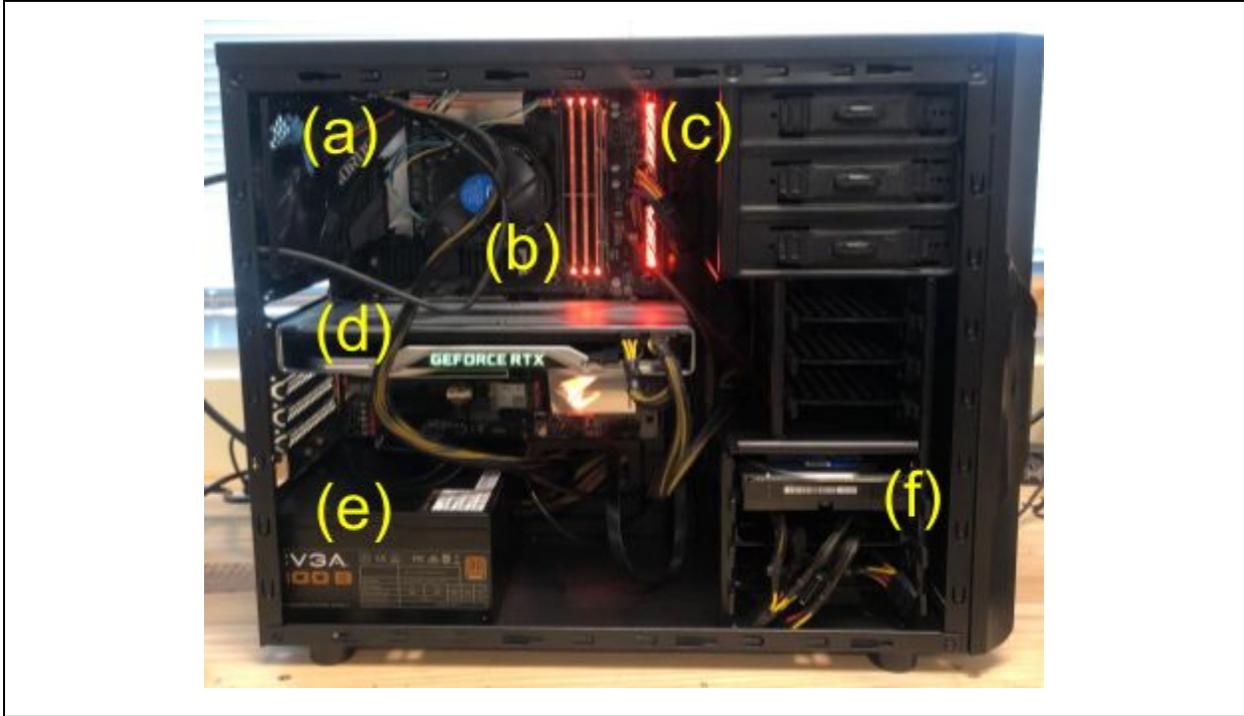


Figure 4. Chassis set-up for all of our team’s computing needs. For component budget and final costs see Appendix A. Parts are as follows: (a) GIGABYTE H370 AORUS Motherboard; (b) Intel Core i7-8700 Desktop Processing Unit (CPU); (c) Ballistix Sport LT 16GB Single DDR4 RAM; (d) NVIDIA GeForce RTX 2080 GPU; (e) EVGA 700 B1 Power Supply; (f) WD Blue 2TB Hard Drive (HDD).

4.2 Software Functions

4.2.1 File Conversion

The system needs to convert files from DICOM to a raw array format (numpy) that can be directly input to the model. After training and applying models, the result is also a raw array, so it will also need to be converted to a medical image format. There was not a direct method to convert the array to DICOM format. To resolve this problem, the array was first converted to MHA metaimage format, which could be done directly. The MHA files can be loaded into 3D Slicer that converts the file into the label map in DICOM format.

4.2.2 Cropping

To crop and separate the structures we coded a GUI (Figure 5) that allows a general technician to:

- a) Select an image to process,
- b) Select an organ to crop,
- c) Maneuver the cropping box by clicking on the image or using the slide bar in the x and y direction, and

- d) Select the slices that contain the organ by sliding the bottom bar. This correlates selection of the z-axis.

Each organ needed to be manually centers in the x, y, and z direction to attain better model results.

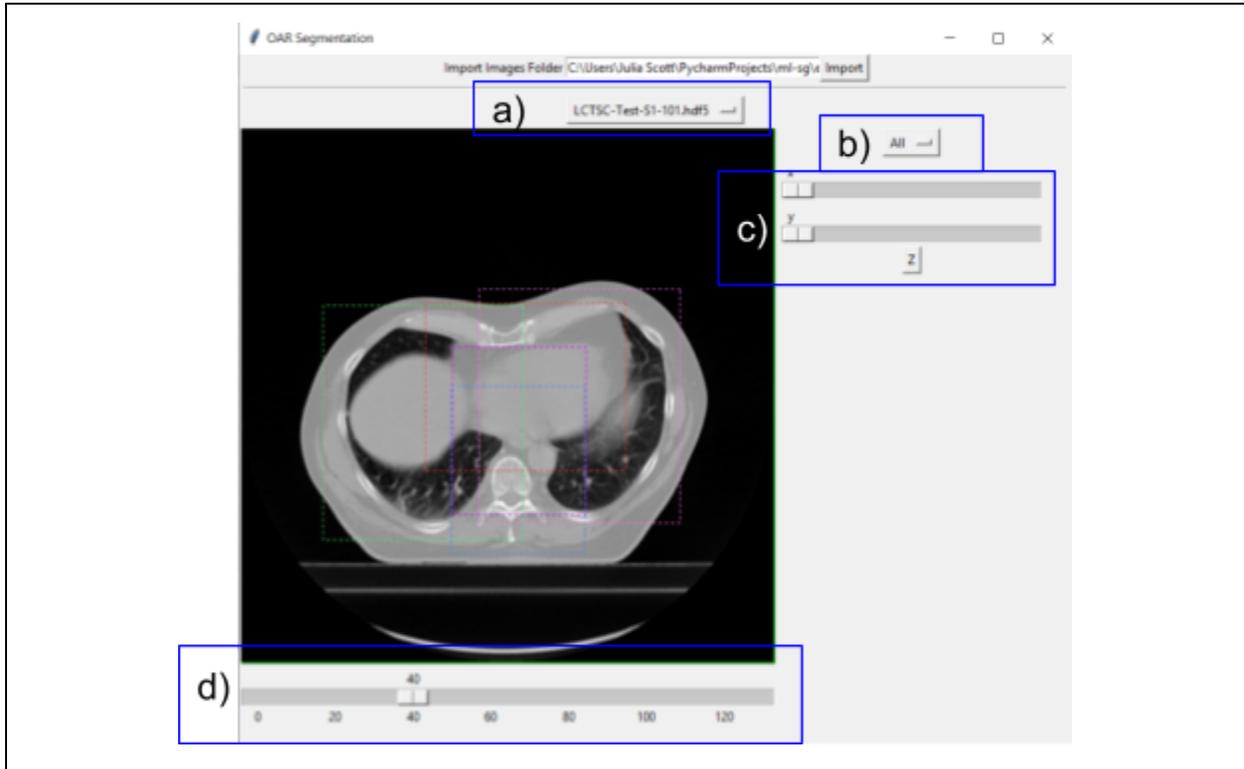


Figure 5. Image of cropping GUI. The boxes outlined in dotted lines are the new boundaries of the image being put through the model. Purple box holds the esophagus, blue box the spinal cord, red box the heart, green box the right lung and pink box the left lung.

4.2.3 Filters

Bilateral Mean and Contrast Stretching

A bilateral mean filter restricts the local neighborhood to have a gray-level similar to the central one as a strategy to denoise. Contrast stretching applies a linear scaling of a set intensity range to maximum gray-level range [12].

Local Equalization

One of the filters chosen was local histogram equalization, in which a low contrast image has each point spread out at the most frequent intensity values [12] to equalize the different parts of the image. The result has the light and dark gray parts of the image that are adjacent to each other increase contrast and be more identifiable.

4.3.2 Producing Label Map

Individual images are created for each OAR and their coordinates and size are recorded in preprocessing. Based on that data, it is then combined together based on which voxel has the largest chance of being a specific OAR. The result is then written to an MHA file.

4.4.3 Save in DICOM Format

After obtaining our label map, we were required to convert segmentations back into DICOM format. We were able to complete this task using 3D Slicer, moving the new radiotherapy structure set (RTSTRUCT) into the original patient image files. This could then be exported into the correct file type, combined into a compressed zip, and submitted to the AAPM challenge. We followed the following instructions:

1. Download mha files
2. Right-click the files and select, “Convert labelmap to segmentation node”
3. Re-name structures in segmentation to exact names required for submission:
 - a. “Lung_R,” “Lung_L,” “Esophagus,” “SpinalCord,” “Heart”
4. Download all patient images
5. Delete original RTSTRUCT from patient images
6. Move newly named segmentations into respective patient files
7. Right click on patient information and select “Export to DICOM...”
8. Go into downloads and name file according to requirements:
“LCTSC-Test-SX-XXX.dcm”

5. Testing

We tested two variables: 1) Tests #1-4 (Table 4) was focused on increasing accuracy to be as accurate as a dosimetrist segmentation. We also submitted our outputted labels from tests 3 and 4 to the AAPM challenge to compare our method’s results to other teams. 2) Test #5 was focused on testing the effect of image filters on our accuracy metrics. This was an internal test conducted with the same batch seed.

Table 4. Summary of subsystem item changes from Tests 1-4.

| Subsystem | Test 1 | Test 2 | Test 3 | Test 4 |
|-----------------------------------|----------------------------------|-------------------------------------|---------------------------|---|
| Preprocessing | None | Cropped organs | Local Equalization Filter | Bilateral Mean and Contrast Stretching filter |
| Image Input to Model | Resized smaller images (256x256) | Original image resolution (512x512) | Original image resolution | Original image resolution |
| Esophagus DSC | 0 | 0.66 | 0.66 | 0.69 / 0.71 |
| Post Processing | None | Combined organs | Combined organs | Combined organs |
| Segmentation Visualization | GIF | DICOM | DICOM | DICOM |

5.1 Test Phase 1: Downsampled Images

The purpose of this phase was to test our model and make sure our GPU was running our model with the given data correctly.

5.1.1 Data Input

The complete organ set with each of the binary labels stacked above each other was ran through our model. We downsampled our images from 512x512 to 256x256 before running them through the model because when running the model with the full image resolution, the model crashed due to insufficient memory storage.

5.1.2 Calculating Accuracy

The metric values was calculated for the aggregate of the OARs since only one model was produced. Thus a single collective DSC loss value was produced.

5.1.3 Visualization

Since we knew we were not submitting our results to the AAPM Challenge due to using downsampled images, we outputted the labels and CT scan as a GIF (Figure 6). Both smaller OARs, the esophagus (low contrast boundaries) and the spinal cord (high contrast boundaries), were missing in the outputted predicted labels. This is because there are an insufficient number of voxels that represent the esophagus/spinal cord at the lower resolution condition. Therefore their features are lost through the kernel filters as the U-net gets deeper through all the layers.

because their **left because of Dr. Scott's notes

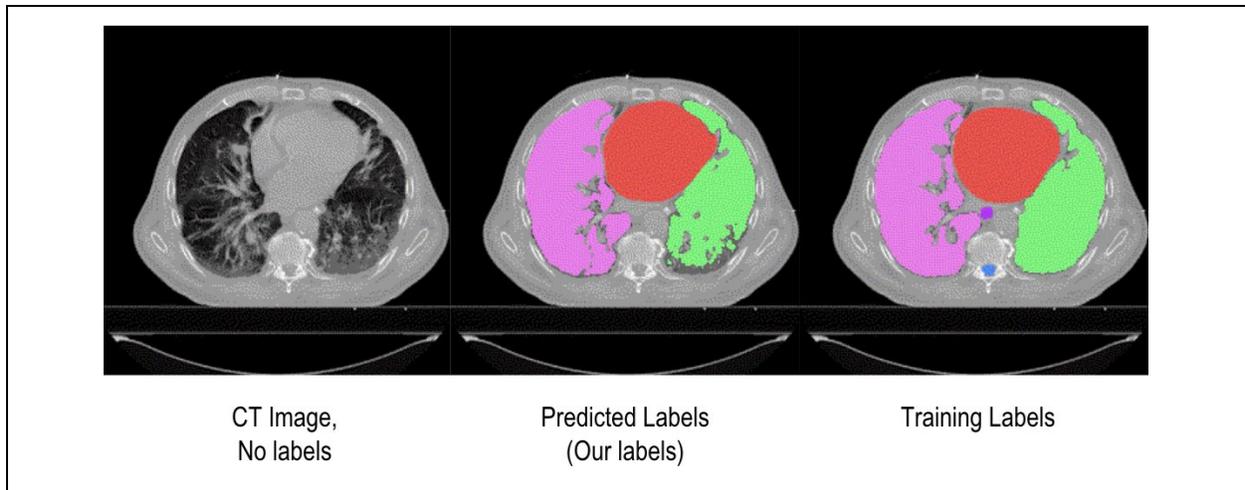


Figure 6. A screenshot of a random transverse plane within our GIF. The pink and green structures are the lungs, the red is the heart, the blue is the spinal cord, and the purple is the esophagus.

5.2 Test Phase 2: Cropping Organs

In this test we aimed to use the original image resolution, convert to DICOM, and submit the results to the AAPM Challenge.

5.2.1 Cropping

A GUI was made to manually create specific bounding boxes that were minimally inclusive of the target organ such that only a part of the image is used for each model (Figure 5). We used the already segmented organs in the training dataset as cropping boundary references. Each individual organ is cropped separately so each can create their own training model allowing the original resolution of the image, 512x512, to be kept. This step was done to improve segmentations of smaller OARs that were not identified in Test Phase I.

5.2.3. Visualization

Since we were more confident in our label map results, we wanted to submit our results to the AAPM challenge. In order to do so we converted the label map and image data into a DICOM file, the approved file type for submission.

5.3 Test Phase 3: Image Enhancement using Local Equalization Filter

The goal of the third test was to achieve higher accuracy results and submit again to AAPM. To reach this goal we decided to add image enhancement features before running them through the model. We decided that preprocessing the images would be quicker than editing our model so we would be able to test within our time constraint.

5.3.1 Local Equalization Filter

A local equalization (LE) filter was added to our images before running the images through our model. The filter enhanced contrast of all tissues dramatizing the contrast between all tissues and blood vessels. An example of the filter on our CT scans is shown below in Figure 7.

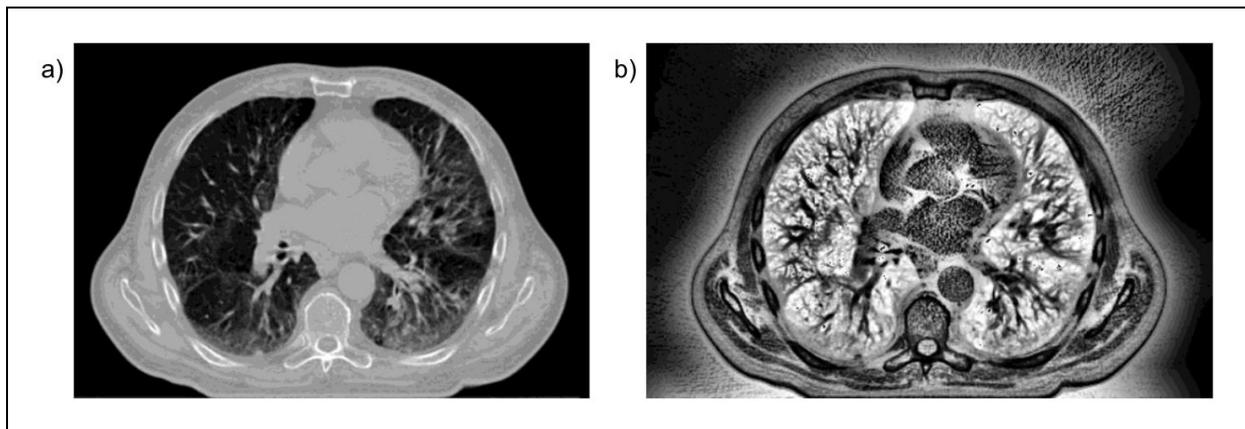


Figure 7. Local histogram equalization filter. Images of the same thoracic cavity scan along the same transverse plane. (a) image of the CT scan without a filter. (b) image of the CT scan with the local equalization filter applied.

5.4 Test Phase 4: Image enhancement using Bilateral Mean and Contrast Stretching

The results from test phase 3 did not improve our esophagus DSC so we decided to run our model with images with no filter and with a bilateral mean filter layered below a contrast stretching filter.

5.4.1 Bilateral Mean and Contrast Stretch Filter

Our aim of image enhancement was to better the contrast between tissues boundaries. We tested two different enhancement approaches: one using a local histogram equalization filter and another using a bilateral mean filter followed by a contrast enhancement filter. A bilateral means filter (BMF) denoised the image first. Followed by a contrast stretching filter (Stretch) to expand the relevant parts of the intensity histogram. Sample images and histograms in the preprocessing step are shown in Figure 8.

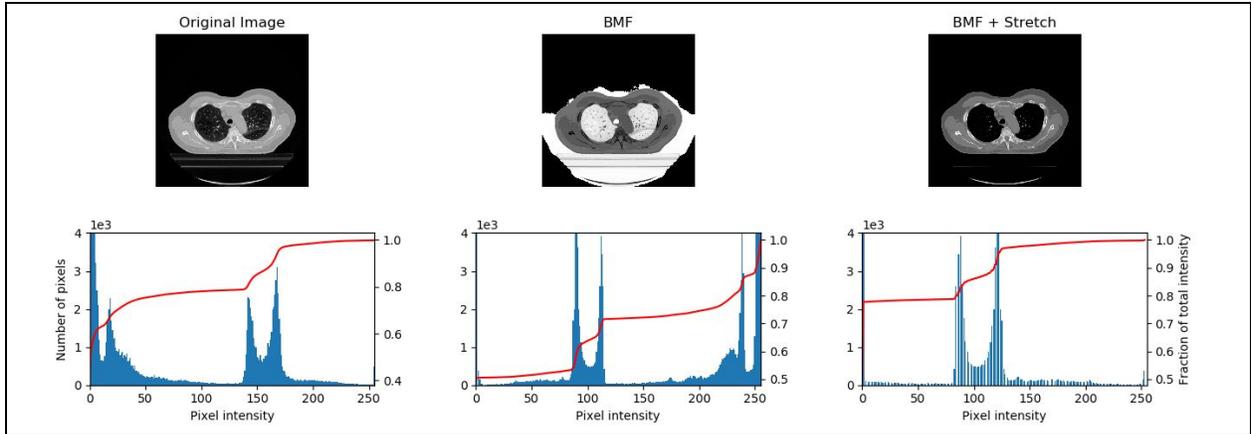


Figure 8. Preprocessing filters. Representing changes in contrast of our (a) original image, (b) image after the application of a bilateral mean filter, and (c) image after the application of bilateral mean filter followed by contrast stretching; (d) histogram evaluation of visual image differences and quantitative contrast analysis.

5.4 System Level Issues

One of the main limitations was the amount of memory available on the GPU. The GPU bought for the system only contained 11 GB of RAM, so the model could not exceed that size.

5.5 Options and Trade-Offs

Our project constraints aided us in selecting the parameters to adjust for each test phase. For most of the training, we chose to use 50 epochs for the base number to train. This was tested when running the first test, 50 epochs was the value at which there was little to no increase in the accuracy of the model. A smaller number of epochs allowed us to run the model quickly so that more iterations could be tested.

Cropping the test data could be executed either manually or be automated. Because the test data did not have labels, an automated approach would involve prior image registration. This has a risk of being extremely inaccurate and given the time constraint and the small dataset size, manual cropping was a better option. We also considered that for actual dosimetrists, the manual cropping would be fast and less laborious because only a general boundary around the organ needs to be found rather than the exact tissue boundaries.

Image enhancement was intended to improve the features of the image for boundary detection. Since there is not an established set of preprocessing filters optimized for machine learning image segmentation, we chose filters that had proven to improve manual and atlas-based segmentation. We chose to apply a bilateral means filter denoised the image followed by a contrast stretching filter to remove extreme intensity values and distribute the most common intensity over a greater range. We also tested a LE filter as a non-traditional option. The resulting image improved contrast, but was noisier than the original image.

5.6 Test Phase Parameters

Tests were performed at 50 epochs because large organs like the heart and lung had already stabilized at that point and more training would not produce significant gain in accuracy. However, the spinal cord and esophagus models were far less stable, and more training to find the minimum validation value was productive. For those OARs, 200 epochs effectively reached that minimum value. The esophagus was prioritized to improve model training and accuracy because it is more vulnerable to radiation exposure due to its proximity to the lungs.

Table 5. Summary of test phase attributes for all organs and for the esophagus.

| Test # | OAR Test (i.e. all organs, esophagus) | Epochs | Filter Applied |
|--------|---------------------------------------|--------|--------------------|
| 3 | All | 50 | Local Equalization |
| 4 | All | 50 | BM+CS |
| 4 | All | 50 | No Filter |
| 5 | Esophagus | 200 | Local Equalization |
| | Esophagus | 200 | BM+CS |
| | Esophagus | 200 | No Filter |

6. Accuracy Metrics

A total of 12 patients were used to assess the performance of the model compared to external methods. Manual segmentations were defined as the reference segmentations from the AAPM challenge. The input was the 3D CT image and the final output was one to five organ segmentation labels. Performance of the proposed methods were tested and compared with the manual segment and of the top two performing teams of the AAPM challenge. The DSC, Hausdorff distance 95 (HD95), and mean surface distance (MSD) were used to quantify the results.

6.1 Dice Similarity Coefficient (DSC Score)

Given two sets X and Y, this metric measures relative overlap. In our case, X represents ground truth and Y represents the submitted segmentations. The DSC is defined as shown in Eq. 2 as follows:

$$DSC = \frac{2|X \cap Y|}{|X| + |Y|} \quad (2)$$

6.2 Hausdorff Distance 95% (HD95)

The directed 95% Hausdorff measure is the 95th percentile distance over all distances from points in X to their counterparts in Y when X and Y are subsets in a metric space (M, d) . HD95 is defined as shown in Eq. 3. In this equation, \sup is the supremum, \inf is the infimum.

$$d_H(X, Y) = \max \left\{ \sup_{x \in X} \inf_{y \in Y} d(x, y), \sup_{y \in Y} \inf_{x \in X} d(x, y) \right\} \quad (3)$$

6.3 Mean Surface Distance (MSD)

Given two sets X and Y , MSD measures the average distance of point X to its closest point in Y . In the equation for MSD (Eq. 4), \bar{d} is the mean of the distances between the points.

$$d_{mean} = \frac{1}{2} \left[\bar{d}(X, Y) + \bar{d}(Y, X) \right] \quad (4)$$

7. Results

7.1 Comparative Quantitative Performance to External Methods

Segmentation accuracy results for each organ, metric, and method are summarized in Table 6. Manual segmentation accuracy metrics are also summarized in Table 7 to use as values of an acceptable automated segmentation result. The manually segmentations are considered ground truth, despite the fact that these boundaries are not the definitive truth for the organs, just the accepted guidelines as edited by a group of dosimetrists.

Regarding larger structures like the lungs, there was very little difference in the performance of our model compared to the top two scoring teams of the 2017 AAPM Grand Challenge. The LE filter system scored the highest in this category out of our 3 systems. In the LE filter model, right and left lung segmentation DSC was 0.98 and 0.97 respectively, very similar numbers to University of Virginia (UV) (0.97 and 0.98) and Elekta (0.97 and 0.97). All three of our systems were equal to or above the intra-rater DSC of 0.95.

The heart, a medium sized structure, and the spinal cord, soft tissues surrounded by bone, segmentation predictions were also comparable to the UV and Elekta teams. The unfiltered system scored the highest DSC for the heart (0.90) and the LE filter for the spinal cord (0.86) out of our systems, which were also comparable results to the UV and Elekta methods. The LE filter system DSC was equal to the intra-rater DSC. All of our systems scored slightly below the intra-rater DSC of 0.93.

Lastly the esophagus, a narrow and long structure with poor tissue boundary contrast, segmentation showed the largest difference between automated and manual segmentation. The unfiltered system scored the highest with a DSC of 0.71, which was significantly different than the intra-rater DSC of 0.818. However, our score was better than the UV team (0.64) and comparable to the Elekta team (0.72). Also, when looking at the other metrics the original image inputs (a) scored the best HD95 and MSD out of all the methods (6.54, 1.93).

Table 6. Metric results for models in each testing phase with benchmark.

| OAR | DSC | | | | | HD95 (mm) | | | | | MSD (mm) | | | | |
|----------------------------|------|------|------|------|------|-----------|------|------|------|------|----------|------|------|------|------|
| | RL | LL | H | E | SC | RL | LL | H | E | SC | RL | LL | H | E | SC |
| (a) No filters | 0.95 | 0.94 | 0.90 | 0.71 | 0.84 | 5.47 | 3.99 | 11.0 | 6.54 | 4.21 | 1.48 | 1.43 | 3.44 | 1.93 | 1.16 |
| (b) LE Filter | 0.98 | 0.97 | 0.86 | 0.66 | 0.86 | 3.87 | 3.68 | 15.3 | 18.7 | 2.74 | 0.92 | 0.97 | 4.37 | 3.41 | 0.81 |
| (c) BM & CS Filters | 0.95 | 0.94 | 0.88 | 0.69 | 0.8 | 5.53 | 4.06 | 13.9 | 7.39 | 5.48 | 1.49 | 2.08 | 1.39 | 4.44 | 1.62 |
| (d) Elekta | 0.97 | 0.97 | 0.93 | 0.72 | 0.88 | 4.7 | 2.9 | 5.8 | 7.3 | 0.2 | 1.08 | 0.74 | 2.05 | 2.23 | 0.73 |
| (e) University of Virginia | 0.97 | 0.98 | 0.92 | 0.64 | 0.89 | 3.6 | 2.2 | 7.1 | 19.7 | 1.9 | 0.93 | 0.61 | 2.24 | 6.3 | 0.69 |

Note. Table represents Dice similarity coefficient (DSC), average Hausdorff distance (HD95) (mm), and mean surface distance (MSD) (mm) of different thoracic organs such as the right lung (RL), left lung (LL), heart (H), esophagus (E), spinal cord (SC). Results are compared internally with enhanced or unenhanced images: (a) no filters, (b) local equalization filter, (c) bilateral mean and contrast stretching. Results are also compared to external teams: (d) Elekta and (e) University of Virginia.

Table 7. Interrater differences in segmentation of OARs for the analyzed metrics.

| OAR | DSC | HD95 (mm) | MSD (mm) |
|-------------|---------------|-------------|-------------|
| Left Lung | 0.956 ± 0.019 | 5.17 ± 2.73 | 1.51 ± 0.67 |
| Right Lung | 0.955 ± 0.019 | 6.71 ± 3.91 | 1.87 ± 0.87 |
| Heart | 0.931 ± 0.015 | 6.42 ± 1.82 | 2.21 ± 0.59 |
| Esophagus | 0.818 ± 0.039 | 3.33 ± 0.90 | 1.07 ± 0.25 |
| Spinal cord | 0.862 ± 0.038 | 2.38 ± 0.39 | 0.88 ± 0.23 |

7.2 Subsystem to Subsystem Comparison

Tests were done internally to compare the accuracy of subsystem tests. The results from running our proposed U-net CNN auto-segmentation model with the original images, bilateral mean and contrast stretching (BM&CS) filters, and local equalization (LE) filter is shown in Figure 9. The model was run with the same random seed to eliminate the variability of the machine learning model. The BM&CS filters showed a higher median DSC value for the lungs and slightly higher median DSC for the spinal cord and esophagus. The LE filter showed a

slightly higher median DSC for the heart. Neither filter outperformed no filtering for all OARs. Extreme outlier points are consistently from the same two cases, suggesting the individual anatomy varies considerably from the average anatomy.

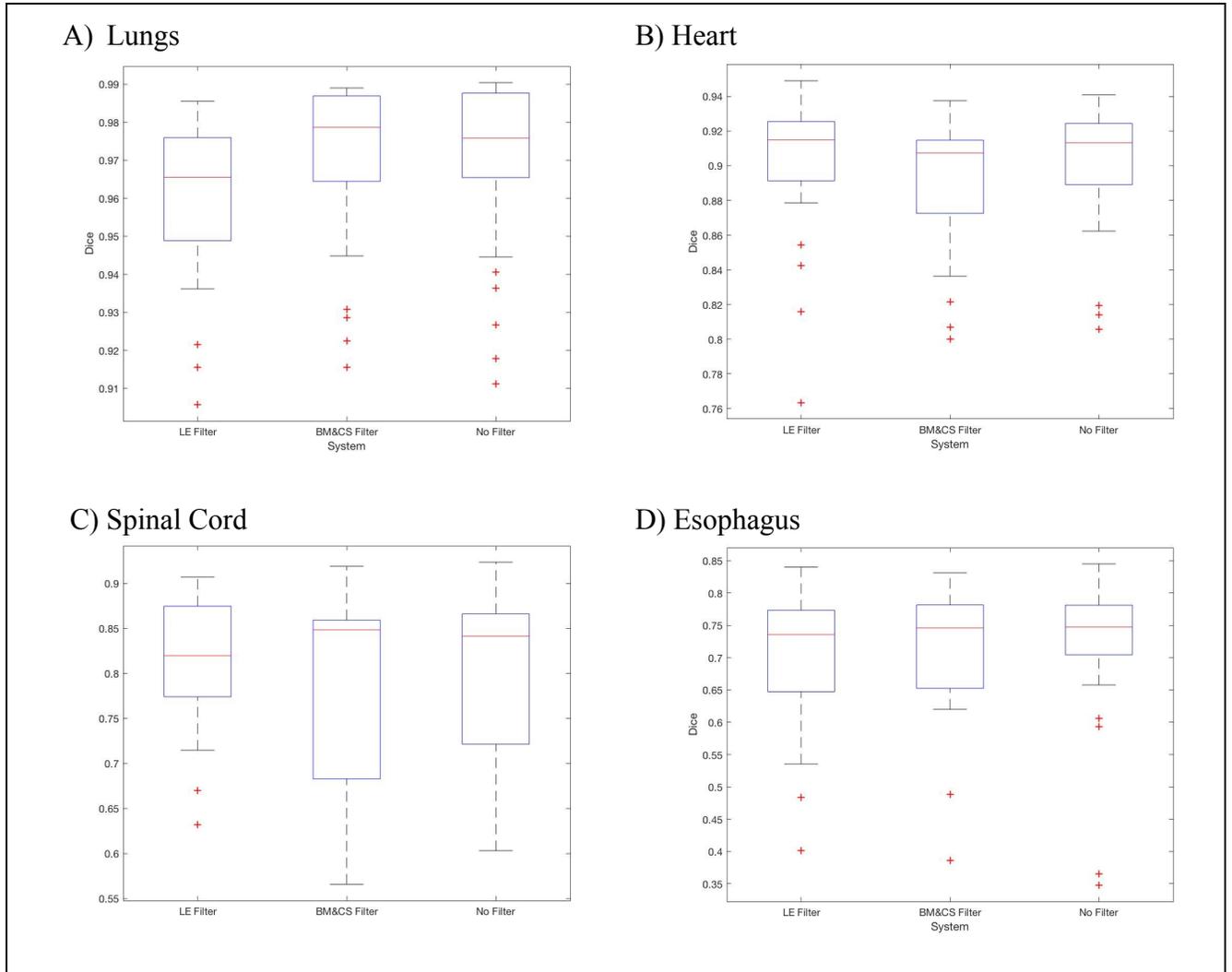


Figure 9. Boxplots representing the DSC achieved from 24 test cases and various image filtering with the same seed. The lungs, heart, and spinal cord was ran with 50 epochs, and the esophagus with 200. The maximum whisker length specified as 1.0 times the interquartile range. Data points beyond the whiskers are displayed using +. (a) right and left lungs; (b) heart; (c) spinal cord; and (d) esophagus.

The accuracy metric results after running the cropped esophagus images through each model with 200 epochs are summarized in Figure 10. These results show the esophagus DSC loss for three different models, each running at 200 epochs. At this point, the DSC value for the validation set, or the testing data began to plateau. However, local histogram equalization filter test showed far less random factors that affected the validation. The overall variation was much less than having no filter or having BM&CS filter.

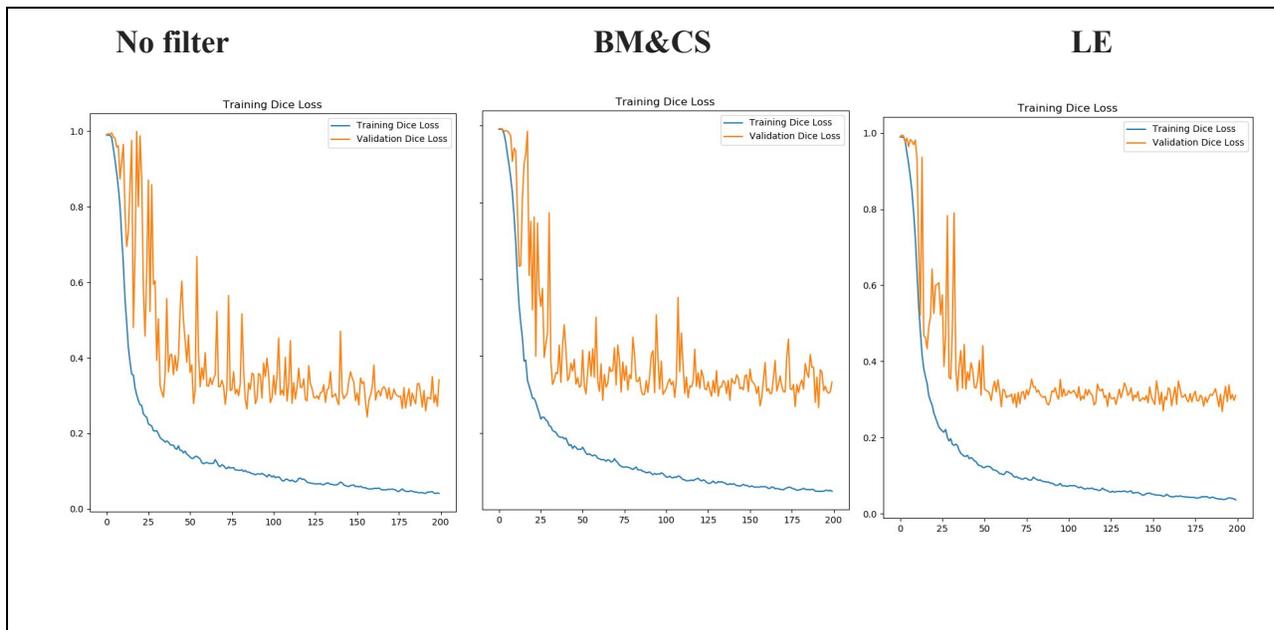


Figure 10. Training DSC loss as a function of epochs versus relative DSC score. The spikes in the validation DSC are only the result of randomness in training. Final DSC scores comparable between three image filters.

7.3 Qualitative Performance

Visual examination of the segmentation labels informed us of where the agreement and disagreement lay between the manual and automated methods (Fig. 11). For example, at $z = +20$ and $z = +40$, patches of esophagus label appear adjacent to the actual esophagus. The heart label overestimated the anterior boundary compared to the manual contour, as shown in $z = -20$ and $z = -40$. The lung show nearly complete overlap, with the exception of gaps in labels over branching bronchi which are hyperintense compared to the air filled the majority of the lung.

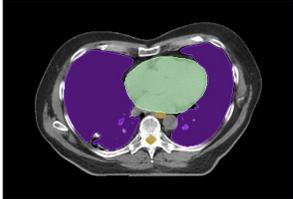
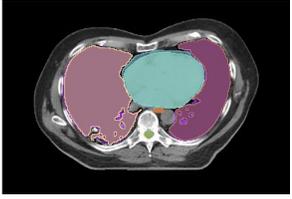
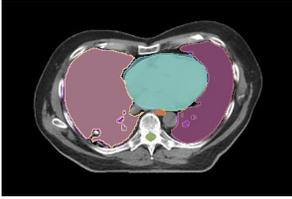
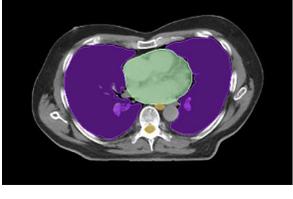
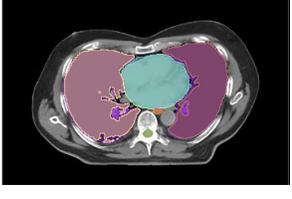
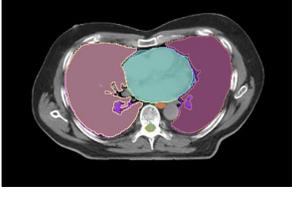
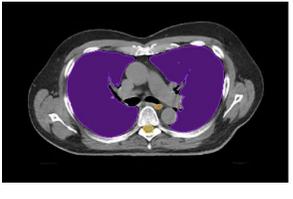
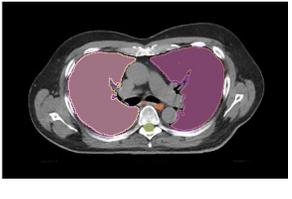
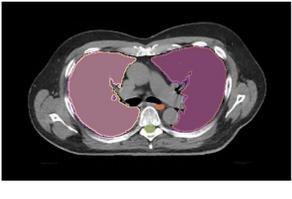
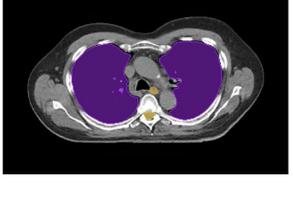
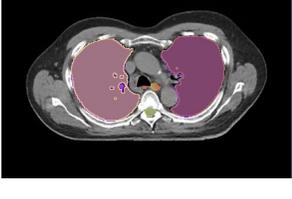
| CT Image | Manual Contour | Organ Cropping | BM&CS Filters |
|----------|--|---|--|
| z = -40 |  |  |  |
| z = -20 |  |  |  |
| z = 20 |  |  |  |
| z = 40 |  |  |  |

Figure 11. Results displayed at 4 axial slices from a randomly selected case from test set. Manual segmentation, Test 2 (Crop), Test 3 (Filters) results are shown. Left Lung--light pink, Right Lung--dark pink, Heart--blue, Esophagus--orange, Spinal Cord--green.

8. Discussion

8.1 Meeting our Requirements

8.1.1 Segmentation Time

We were successful in significantly reducing manual segmentation time by developing a machine learning system. It took 2 minutes to segment once patient. With an additional 30 secs to crop the organs, the total amount of time it would take to segment a patient would be 2 minutes and 30 seconds which is 48 times quicker than manual segmentation.

8.1.2 Accuracy

Within our particular model, an additional pre-processing step was introduced to a 3D U-Net to improve the accuracy of the esophagus structure. We found that the BM&CS filter produced more favorable results for the lungs, an organ with higher contrast boundaries. The BM&CS filter also produced slightly higher median DSC values for the spinal cord and esophagus. LE filter showed a slightly higher median DSC for the heart. Outlier points are consistent, suggesting the individual anatomy variability.

Although our methods did not produce esophagus DSC values as accurate as manual segmentation, we were able to achieve a score higher than the UV team’s DSC value. The low contrast boundaries of the esophagus makes this organ difficult for both software and humans to distinguish tissue boundaries for segmentations [13]. Adding an image cropping step, which isolated the esophagus, resulted in this organ being detected and segmented. The cropping aids the deep learning methods to contour the esophagus and may be a necessary preprocessing step for deep learning automation methods.

Table 8. Table comparison of our methods (SCU), University of Virginia (UV), and Elekta.

| Method | Model | Layers | Preprocessing | Input | Framework |
|--------|--|--------|--|---|---------------|
| SCU | 3D U-Net | 18 | Cropping and image enhancement | 512x512 for all organs | Keras |
| UV | Deep learning VGGNet model based on 3D U-Net | 7 | Intensity normalization and image resizing | All OAR model: down-sampled images Single OAR models: cropped full resolution images | Tensorflow |
| Elekta | DCNN model that was modified from the U-Net architecture. Two models were trained and applied in sequence. | 27 | None | 2.5D model: 5×360×360 voxels was trained to segment lungs 3D model: 32 × 128 × 128 voxels was trained to segment heart, esophagus, and spinal cord | Caffe Package |

Using a machine learning approach potentially minimizes the impact of intra- and inter-rater variability of manual delineations on the final segmentation. It needs to be kept in mind that these results are not a true representation of the segmentation accuracy of the tissues themselves, but only a comparison to the ground truth of the manual segmentations. This is because there are rules and regulations that dosimetrist have to follow that do not segment the organs exactly. For example, Figure 11 is an example 3D visualization of a manually segmented

scan and the heart segmentation (in blue) bluntly stops at the superior boundary. Dosimetrists are required to segment the heart only up to that point because the superior region has connecting veins and arteries that are difficult to accurately segment by hand. Our model does not have these requirements to follow, therefore there will always be error between manual segmentations and automated segmentations [4].

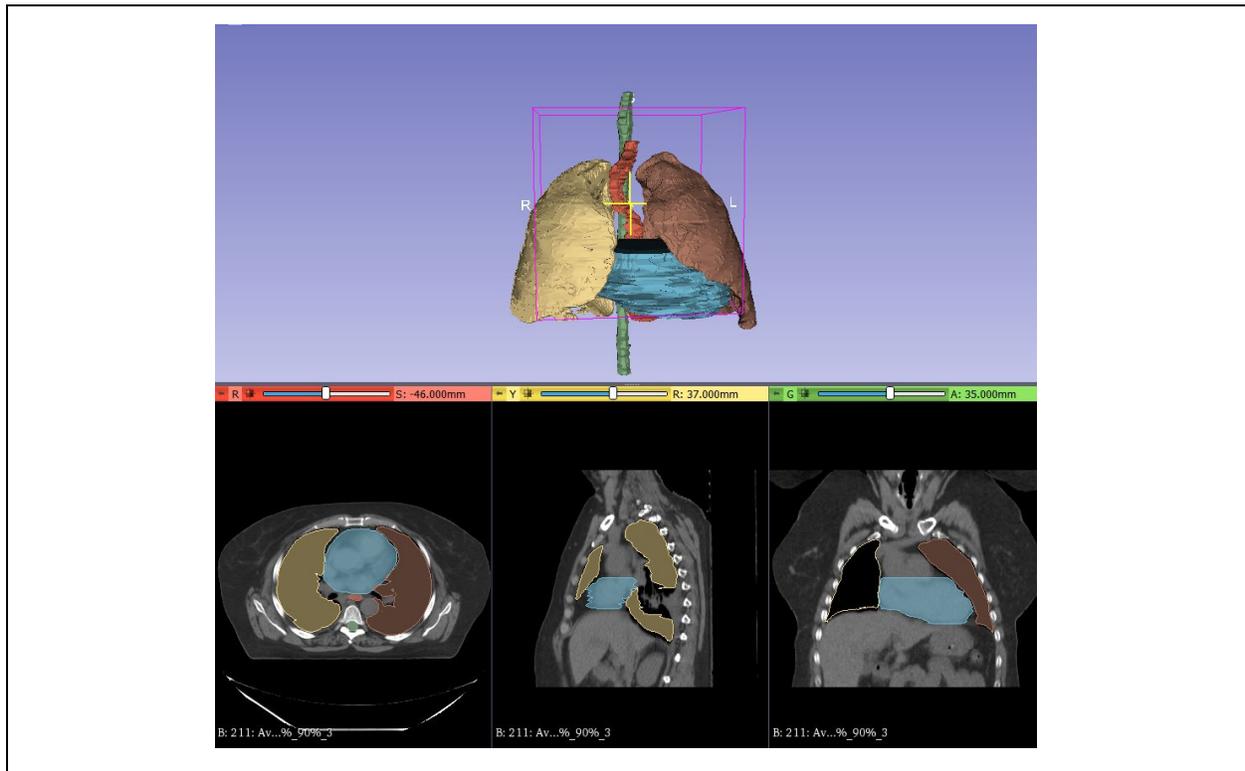


Figure 11. 3D visualization of segmentation produced by manual segmentation in 3D Slicer.

8.1.3 Functional Requirements

We successfully created an automated method that when given a patient's CT scan will segment the OARs and output a file which is in the correct format (DICOM) to use for radiation dosage planning. Visualization of the DICOM format is displayed in Figure 11. Our approach still necessitates a single preparatory step of manually cropping anatomical regions to isolate segmentation volume shown in Figure 5. However, a general hospital technician could complete this task, which only takes approximately 30 seconds per case. Therefore, we were able to remove the need of an expert for the time-consuming step of segmenting OARs in RT planning.

8.2 Project Challenges and Constraints

Engineering a machine learning solution brings a number of risks, challenges, and ethical considerations. Our approach used a dataset from the 2017 AAPM Grand Challenge, which

allowed us to compare our results to other machine learning approaches, yet also challenged us to innovate outside of established machine learning models.

The first challenge we approached at the start of our project was determining what skills we needed as members of an interdisciplinary team. Bioengineering students were tasked with learning the fundamentals of Python coding and collaborative programming, and all members took a course in Machine Learning through Coursera. This knowledge step, in congruence with research into previous literature, took up the first ten weeks of the project. Our next step, building the computing system, caused delays. The central processing unit (CPU) recommended to us was incompatible with the motherboard, requiring us to order a new CPU and wait a week for delivery. After training our model, our focus on data conversion took another significant portion of available time. Medical images are typically processed through a file type called DICOM (Digital Imaging and Communications in Medicine). However, Tensorflow requires a conversion into hdf5 files and then back through to MHA files for use in 3D Slicer to obtain the DICOM files for final testing on the AAPM website. This detailed process took time to research, discuss, and implement successfully. When reviewing our segmentations, we also recognized variability in our patient dataset. Certain patients had collapsed lungs or tumors, and interobserver consistency was low for the editing of segmentations, causing skewed tracings [14]. We experienced our last problem while submitting results for the model featuring cropping and pre-processing. The AAPM website had accepted our initial request to join the competition, but after a few days, closed the website to any new submissions. We relied on their system to calculate our metrics and compare to previous teams, and asked that they reopen the competition for us to continue submissions. We were fortunately able to have the website reopened, and collected the necessary data.

8.3 Risks and Mitigations

The first risk with this was our limited dataset. Because we did not have a dosimetrist to segment sample thoracic CT images, we were constrained to the 60 patients provided by the challenge. In the future, models created with larger, more diverse datasets may provide more accurate results. Another risk was limited time, which became more of a challenge as we neared the end of the last quarter. Our first quarter was focused primarily on learning the technical skills necessary to build and test a machine learning model, our second on building our computing system and training the model, and our third was modifying this model and obtaining results. An additional risk was in protecting our data and preventing overloading the physical GPU RAM. We were able to purchase and build our own computing system so that we had a dedicated local machine for processing and data storage. Code developed for our system was backed up through Github. If we faced insurmountable roadblocks with our computing system, we planned to use the University Engineering Computer Center resources as a back-up. Our dedicated workstation

was reliable and met the specifications of our system, such that we did not need to rely on other resources.

8.4 Societal Issues

As we created our solution for manual segmentation, it was necessary to reflect on its implications on society's health and safety, the national and global economy, and usability in target regions. There is high variability of manual segmentation between dosimetrists in different regions, and by the dosimetrist themselves [15], so we were required to ensure that a machine learning solution segmented as well as a dosimetrist with this variability. We were able to compare a normalized score of the model's three metric outputs to the mean for interrater differences. If we were to reach a mean of 50 from a scale of 0 to 100, our model would be performing as well as a dosimetrist would in any given region. Our model with filters had a score of 43.49, and our model without filters reached a normalized score of 48.853. Excluding low-performing organ structures from the model could improve these final scores to be at an acceptable clinical level.

Another consideration outside of the scope of our project, but coinciding with machine learning in healthcare, is adherence to clinical guidelines across international borders and in patient data protection or privacy. Enabling access to RT would be greatly beneficial to LMICs, however, for RT to truly be beneficial in these areas, dosimetry auditing is essential. Worldwide auditing for RT planning is currently insufficient, with only two-thirds of RT centers receiving some level of auditing [16]. Advancements in cloud computing and international data transfers may allow for more centralized auditing to be instituted.

One drawback to our solution is that this is only semi-automated. A dosimetrist still needs to oversee the RT planning steps, but the manual labor is reduced. With a reduction in the length of time needed for segmentation, a dosimetrist would be able to complete more RT plans per day. The target total RT planning time is 30 minutes or less--a fourth of the current average time. While reducing reliance on human expertise, this strategy increases reliance on computing infrastructure. Treatment centers installing new machines with machine learning solutions like ours would still require a trained individual to occasionally audit the system. However, the need for dosimetrists would decrease overall in areas where they are difficult to find.

For patients in LMICs, RT would become more affordable as patients' travel costs are reduced for those few weeks radiotherapy planning requires. We worked on this project with an abundance of compassion, because the development of a single day consult-to-treatment plan would enable more patients in more regions to have access they need to a treatment with fewer side effects, so they can ultimately get back to enjoying life with their families.

9. Conclusion

Over the course of this project, our team learned new technical and interpersonal skills to create a solution for the tedious nature of organ-at-risk segmentation in the radiotherapy planning process. Both the computer engineering and bioengineering sides of our team started with extensive research into artificial intelligence and machine learning, followed by a self-directed lesson of computer building. We were ultimately able to develop a system that segmented organs to improve the quality of care for radiotherapy, the consistency of segmentations, overall reducing the time to treatment and hopefully expanding access to the developing world. The next challenge to address is how this would integrate into a radiotherapy planning software, and what training would be required to make this operational in any region for a one-day consult-to-treatment system. In the time we were given for this project, our outcome was a sufficient start to this goal, but future teams have the potential to target the esophagus for metric improvement, find a larger data set to train and test with, segment an entirely different organ system like the abdomen or head and neck, or fully automate the algorithm. Not only did we improve upon our project management abilities, learn how to build a computer and code, understand the drawbacks to our approach, and problem solve when challenges arise, but we have also concluded that a machine learning approach is a viable solution for radiotherapy planning.

Appendix

Team Project Management Summary

Team Approach

To account for gaps in knowledge between majors, throughout the project, computer engineering students gained anatomical knowledge to understand and strive to meet project requirements. Bioengineering students in our team took on a project management-oriented role when faced with computer engineering tasks outside of their skill range. As liaisons with each department for funding and project questions, our bioengineering students were able to organize our computer components, timeline, and ensure we were always moving forward towards the next steps.

Key Lessons

- We learned to save all files and result data, no matter how advantageous the results are.
- We learned that machine learning models have a variety of applications.
- We learned that building a computer requires research and meticulous technical skill.
- We learned that data conversion and protection is just as important as the model.

Budget

This design team was provided a budget of \$2,000, or \$500 per team member, from Engineering Undergraduate Programs. Our team estimated that the maximum total cost to develop an efficient and accurate machine learning approach to OAR segmentation would total to \$2155. This covered materials for a new desktop, CPU, and GPU. The timeline for the Engineering Computer Center transition and the timeline to develop our methods overlapped so we could not solely depend on the usage of their equipment because it would significantly postpone the development phase of our methods, potentially detrimentally affecting our senior design experience.

We planned to adopt a deep neural network (DNN) for the segmentation task. DNNs must consider many training parameters, which has high computational complexity and requires powerful computing resources. Data storage and visualization cannot be done on a typical student laptop. Cloud computing storage services are insufficient to support the size of our data set. Amazon AWS subscription holds up to 100GB and our raw image data set alone is 124GB. This is what the GPU could do to meet our needs:

- GPU's many-core architecture has produced significant speedups in DNN training, because of the suitability of its processing architecture for matrix and vector computations.
- This workstation will be handed down to future work on this project. In order to prepare for larger data sets in the future, it is better to invest in a powerful computer for continuous use than upgrade it constantly.

To create estimations for our budget and prepare our final buylist, we consulted with our Computer Engineering advisor Dr. Liu to create the estimations for our budget. We performed cost analysis estimations by looking at material prices from Amazon, Best Buy, and NVIDIA.

Table A1. Final spending estimates for senior design project.

| Project Costs | | |
|----------------------|------------------|---------|
| Hardware Materials | | |
| Desktop: | Item | Price |
| | CPU | \$330 |
| | Motherboard | \$120 |
| | RAM | \$155 |
| | GPU | \$1200 |
| | Hard drive | \$60 |
| | Chassis | \$40 |
| | Power Supply | \$80 |
| | Monitor | \$90 |
| | Keyboard & Mouse | \$20 |
| Est. Project Total | | \$2,095 |

Timeline

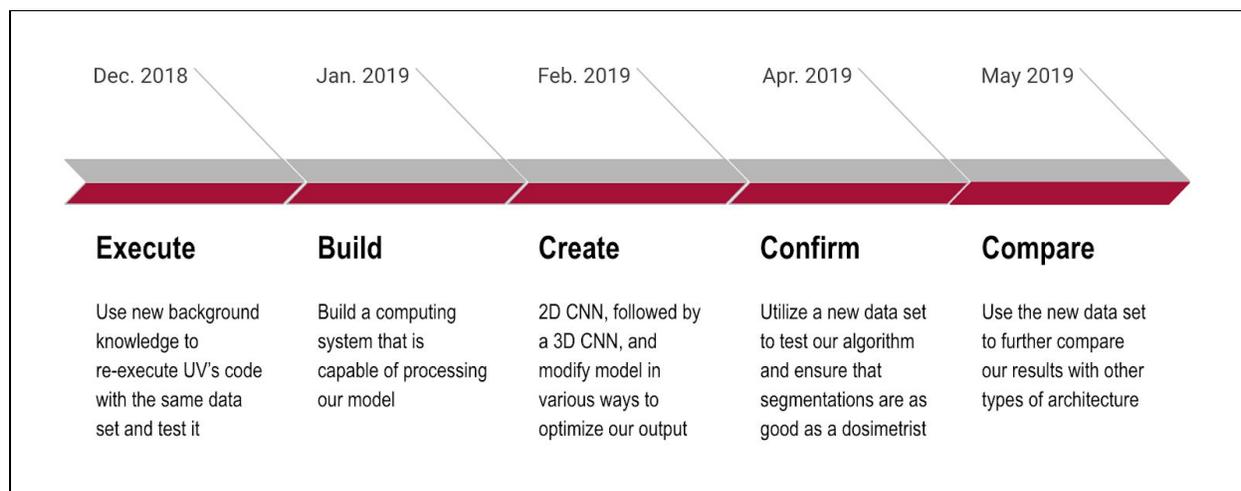


Figure A1. Estimated timeline for OAR Segmentation project team. Project starting date in September of 2018, extending to the start of June 2019.

References

1. Datta NR, Samiei M, Bodis S. (2014). Radiation therapy infrastructure and human resources in low- and middle-income countries: present status and projections for 2020. *Int J Radiat Biol Phys.* 2014;89(3): 448-457.
2. Chen Z, King W, Pearcey R, et al. (2008). The relationship between waiting time for radiotherapy and clinical outcomes: a systematic review of the literature. *Radiother. Oncol.*, vol. 87, no. 1, pp. 3–16.
3. Nelms BE, Tomé WA, Robinson G, Wheeler J. (2012). Variations in the contouring of organs at risk: Test case from a patient with oropharyngeal cancer. *Int J Radiat Oncol Biol Phys.* 2012;82:368–378.
4. Sharp G, Fritscher KD, Pekar V, et al. (2014). Vision 20/20: perspectives on automated image segmentation for radiotherapy. *Med Phys.* 2014;41: 50902.
5. Litjens G, Kooi T, Bejnordi BE, et al. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis.* 2017;42:60-88.
6. P. Kanavos. (2006). The rising burden of cancer in the developing world. *Annals of Oncology.* Volume 17. Issue suppl_8. Pages viii15–viii23.
7. Hu P, Wu F, Peng J, et al. (2016). Automatic abdominal multi-organ segmentation using deep convolutional neural network and time-implicit level sets. *Int J CARS.* 2017;12(3):399-411.

8. Voet P, Dirkx M, Teguh DN., et al. (2011). Does atlas-based autosegmentation of neck levels require subsequent manual contour editing to avoid risk of severe target underdosage? A dosimetric analysis. *Radiotherapy and Oncology*. 98(3):373-377.
9. Daisne J, Blumhofer A. (2013). Atlas-based automatic segmentation of head and neck organs at risk and nodal target volumes: A clinical validation. *Radiation oncology* (London, England). 8(1):154.
10. Han X. (2017). MR-based synthetic CT generation using a deep convolutional neural network method. *Medical Physics*. 44(4):1408-1419.
11. Greenham S, Dean J, Fu CKK, et al. (2014) Evaluation of atlas-based auto-segmentation software in prostate cancer patients. *Journal of Medical Radiation Sciences*. 61(3):151-158.
12. Scikit-image Development Team. Local histogram equalization. https://scikit-image.org/docs/dev/auto_examples/color_exposure/plot_local_equalize.html Web site. Updated 2019. Accessed June 1, 2019.
13. Lustberg T, van Soest J, Gooding M, et al. (2018). Clinical evaluation of atlas and deep learning based automatic contouring for lung cancer. *Radiotherapy and Oncology*. 126(2):312-317
14. Yang J, Veeraraghavan H, Armato SG, et al. (2018). Autosegmentation for thoracic radiation treatment planning: A grand challenge at AAPM 2017. *Medical Physics*, 45(10), 4568-4581.
15. McCall R, et al. (2016). Anatomical contouring variability in thoracic organs at risk. *Medical Dosimetry*, 41(4), 344-350.
16. Izewska J, Lechner W, & Wesolowska P. (2018). Global availability of dosimetry audits in radiotherapy: The IAEA dosimetry audit networks database. *Physics and Imaging in Radiation Oncology*, 5, 1-4.