Computer-Aided Image Analysis and Decision Support System for Bladder Cancer

by

Kenny Heekon Cha

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biomedical Engineering)
in The University of Michigan
2017

Doctoral Committee:

Professor J. Brian Fowlkes, Co-Chair
Professor Lubomir M. Hadjiiski, Co-Chair
Professor Heang-Ping Chan
Professor Richard H. Cohan
Professor Jeffrey A. Fessler
Professor Douglas C. Noll
Associate Professor Alon Z. Weizer

Kenny H. Cha

heekon@umich.edu

ORCID iD: 0000-0003-3847-7448

DEDICATION

This dissertation is dedicated to my family, especially to my parents.

ACKNOWLEDGMENTS

I have had the great privilege of working with my advisor, Dr. Lubomir Hadjiiski. As I was at a point in my life where I wasn't sure what I was going to do after receiving my Bachelor's degree, I applied to work with Dr. Hadjiiski on this project of computer-aided diagnosis of bladder cancer. With his guidance, and the guidance of the lab director Dr. Heang-Ping Chan, I have come to where I am today. Their influence has allowed me to be successful, as they have imparted in me the correct way to perform experiments and analyze results. I would like to thank my co-advisor to Biomedical Engineering, Dr. Brian Fowlkes. Every time we have a discussion about research, I've gotten new ideas that were insightful. Dr. Richard Cohan is a face that I have gotten used to seeing about once a week in the last few years. His expertise helped me design the methods so that they may be useful for the clinicians. His input was always appreciated when we had a clinical question regarding a case. I would also like to thank Dr. Alon Weizer; his input as the physician performing the treatment of the patients with bladder cancer helped me understand the clinical problem. I would like to thank Dr. Doug Noll, who I learned a great deal from during my undergraduate Biomedical Engineering Lab, and Dr. Jeff Fessler, whose image processing class helped build the foundation for my work. I would also like to thank Dr. Elaine Caoili for being one of the radiologists to define the reference standard for the studies that we have performed.

I would like to thank the other members of the lab for their input and discussion, as they helped develop my skills. I was always able to bounce some ideas off of Dr. Ravi Samala, and him along with Drs. Jun Wei and Chuan Zhou have always given good advice.

I would also like to thank others from the lab: Sankeerth Garapati, Dhanujg Girish, Marshall Gordan, and Jordan Noey for their support with the research as this work would not have been completed without their help.

This research is supported by National Institutes of Health grant numbers R01CA134688 and U01CA179106. Chapter II was published in *Physics in Medicine and Biology*[1], Chapters III

and IV were published in *Medical Physics*[2, 3], while Chapter VIII was published in *Tomography*[4]. Chapter VII has been accepted for publication in *Medical Physics*, and Chapter IX was published in *Scientific Reports*[5] at the time of this dissertation.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Bladder cancer is a common type of neoplasm that can cause substantial morbidity and mortality among patients. Bladder cancer causes 16,870 deaths per year in the United States. It is expected that 76,030 new bladder cancer cases will be diagnosed in 2017. Multi-detector row CT (MDCT), and specifically MDCT when used for urography (CTU), has become the imaging modality of choice for evaluation of most urinary tract abnormalities. Interpretation of MDCT urograms that commonly exceeds 400 images. This is a demanding task for radiologists who have to visually track the entire upper and lower urinary tract and look for lesions that usually are small in size. Using a computer-aided diagnosis (CAD) system as an adjunct for the radiologist may reduce the number of lesions that are missed by the radiologists. To create a CAD system for detection of bladder lesions, we have developed methods and software to perform the following specific tasks: (1) segment the bladder; (2) detect bladder lesion candidates; (3) segment the bladder lesion candidates; (4) extract features from the segmented lesion candidates; (4) classify lesion candidates as true lesion findings or false positives using the extracted features; (5) detect bladder wall thickenings.

Correct staging of bladder cancer is crucial in determining the need for neoadjuvant chemotherapy treatment and minimizing the risk of under-treatment or over-treatment. Also, reliable assessment of the response to neoadjuvant therapy at an early stage is vital for identifying patients who do not respond to this treatment and allows the physician to discontinue ineffective treatment and its undesired adverse effects on a patient's physical condition. We have developed prototype predictive models for both bladder cancer staging and bladder cancer response to neoadjuvant treatment by adapting the methodology of feature classification that merges image-based biomarkers. The bladder lesion segmentation modules were used to extract the image-based biomarkers as input to the models. Early detection of bladder cancer, accurate tumor staging, and early prediction of treatment response could reduce mortality and morbidity, and improve quality of life for surviving patients.

This dissertation presents the methods that we developed to automatically segment the bladder on CTU, and automatically detect masses and wall thickenings with sensitivity near 90%

with a relatively low number of false positive findings. We have developed a system that distinguishes between muscle-invasive and non-muscle-invasive cancer, which is a clinical threshold used to determine treatment. We have also developed a system that uses the pre-treatment and post-treatment CTU scans to estimate the likelihood that the patient has fully responded to treatment. This system has achieved performance comparable to the radiologists, with area under the receiver operator characteristic curve (AUC) values in the range of 0.69 to 0.77. We performed an observer performance study where we saw that our system of predicting complete response to treatment improves the performance of the clinicians when they read the pre- and post-treatment scans with the aid of the system. On average, the clinician AUC increased with statistical significance from 0.74 to 0.77.

Early detection of bladder cancer, accurate staging of tumors, and early prediction of treatment response can reduce mortality and morbidity, and improve the quality of life for surviving patients. These studies show possible methods for performing those tasks.

Chapter I
Introduction

1.1 Motivation

Bladder cancer is a common type of neoplasm that can cause substantial morbidity and mortality among patients. Bladder cancer causes 16,870 deaths per year in the United States. It is expected that 76,030 new bladder cancer cases will be diagnosed in 2017. Multi-detector row CT (MDCT), and specifically when used in CT urography (CTU), has become the imaging modality of choice for most urinary tract abnormalities. For my PhD dissertation, I have developed a computer-aided image analysis and decision support system for bladder cancer. This system will be useful for two major applications: (1) computer-aided diagnosis (CAD) system for bladder lesion detection, and (2) computerized bladder cancer staging and monitoring of treatment response.

Interpretation of MDCT urograms usually requires reviewing more than 400 reconstructed images. This is a demanding task for radiologists who have to visually track the upper and lower urinary tract and look for lesions that usually are small in size. Using a CAD system as an adjunct for the radiologist may reduce the number of lesions that are missed by the radiologists. To develop a CAD system for detection of bladder lesions, we will develop methods and software to perform the following specific tasks: (1) segment the bladder; (2) detect bladder lesion candidates within the contrast-enhanced and non-contrast enhanced regions of the bladder; (3) segment the bladder lesion candidates; (4) extract features from the segmented lesion candidates; (4) classify lesion candidates as true lesion findings or false positives using the extracted features; (5) detect bladder wall thickenings. In our preliminary studies, we have demonstrated the feasibility of automatic segmentation of the bladder and automatic detection of bladder masses within the contrast-enhanced region. If successfully developed, the CAD system, used for decision support by radiologists, can potentially improve the performance of the radiologists in detecting bladder lesions.

Correct staging of bladder cancer is crucial in determining the need for neoadjuvant chemotherapy treatment and minimizing the risk of under-treatment or over-treatment. A reliable assessment of the response to neoadjuvant therapy at an early stage is vital for identifying patients who do not respond, allowing a clinician to cease ineffective treatment that can have adverse effects on a patient's physical condition. Decision support systems that can combine available data with statistical models will help clinicians determine more objectively and quantitatively bladder cancer stage and monitor treatment response. Bladder lesion segmentation and feature classification play a central role in such decision support systems. Accurate bladder lesion segmentation defines the region where the image-based biomarkers are extracted and used as input to build the predictive models for decision support. We have developed prototype predictive models for bladder cancer stage and response to neoadjuvant treatment by adapting the methodology of feature classification to merge image-based biomarkers (tumor characteristics such as tumor volume, morphology, textures, etc., and their changes) and non-image-based clinical biomarkers (examination under anesthesia). Early detection of bladder cancer, accurate staging of tumors, and early prediction of treatment response would reduce patient mortality and morbidity, and improve quality of life for surviving patients.

1.2 Problem Statement

The main goal of this dissertation is to develop a computer-aided image analysis and decision support system for bladder cancer using advanced computer vision and machine learning techniques. The development of this new system for CTU is built upon the knowledge and experiences in the development of CAD techniques for breast cancer, lung cancer, and other diseases in the CAD Research Laboratory. This system is useful for two major applications: (1) computer-aided diagnosis (CAD) system for detection of bladder lesion, and (2) computerized bladder cancer staging and monitoring of treatment response.

The CAD system includes modules for bladder segmentation, prescreening of lesion candidates, lesion segmentation, feature extraction, and feature classification. The CAD system is intended to provide a second opinion to radiologists in the detection of bladder lesions on multi-detector row CT urography (CTU).

The decision support systems for bladder cancer staging and monitoring of treatment response share some of the image analysis and machine learning methods developed for the

CAD system. We modified the segmentation and feature extraction modules developed for the CAD system and adapted them to extract image-based biomarkers. The classification module was adapted to combine image-based biomarkers and non-image-based biomarkers to predict bladder cancer staging and treatment response assessment.

We aim to build (1) a CAD system that may assist radiologists in detecting bladder lesions, and (2) a decision support system for computerized bladder cancer staging and monitoring of treatment response.

1.3 Summary of Contributions

The main contributions of this dissertation are summarized below:

- Refining a previously developed bladder segmentation method using image processing techniques and a new method for propagating segmentation contours (Chapter 2)[1].
- Using deep-learning convolutional neural network (DL-CNN) to perform bladder segmentation, which improves the segmentation performance while reducing the number of user inputs required (Chapter 3)[2].
- Creating a system for the detection of bladder masses within the contrast-enhanced region of the bladder, where the contrast between the contrast material and the bladder masses makes it easier to detect the masses (Chapter 4)[3].
- Creating a system for the detection of the bladder masses within the entire bladder, including the subtle lesions within the non-contrast-enhanced region of the bladder, while reducing the number of false positive findings resulting from the added detection task (Chapter 5)[6].
- Creating a system for detection of bladder wall thickening that may also be indicative of bladder cancer (Chapter 6).
- Creating a system for automatic staging of bladder lesions, where muscle-invasive bladder cancers are distinguished from non-muscle-invasive bladder cancers (Chapter 7)[7].
- Creating a system to segment bladder lesions using DL-CNN and testing if such a system can predict whether a patient will completely respond to neoadjuvant chemotherapy using the change in gross tumor volume (Chapter 8)[4].

- Creating a system for identifying lesions that completely respond to neoadjuvant chemotherapy using pre- and post-treatment CTU images (Chapter 9)[5].
- Conducting an observer performance study with 11 clinicians to demonstrate that the decision support system can identify cases that completely respond to treatment and improve the performance of the clinicians who use the CAD system as a second opinion (Chapter 10).
- Performing a simulation study that shows that the performance of a trained classifier on a test data set with missing data replaced with the average value of the training set decreases on average compared to that without missing data, but the average performance is higher than retraining the classifier without the missing feature when the proportion of test cases with missing data is small (Chapter 11).

1.3.1 Other contributions

Other contributions not detailed in this dissertation include the use of the developed methods for organs other than the bladder. One study estimated the tumor features of head and neck cancers on CT[8], and another study examined using different biomarkers, including imaging biomarkers, for accurate detection of tumor progression of oropharyngeal cancers[9]. We also attempted to apply pre-trained networks and transferring the learned weights of the deep-learning network to the task of pre- and post-treatment response assessment[10].

Chapter II
CT Urography: Segmentation of Urinary Bladder using Conjoint Level Set Analysis and Segmentation System with Local Contour Refinement

2.1 Abstract

We are developing a computerized system for bladder segmentation on CT urography (CTU), as a critical component for computer-aided detection of bladder cancer. The presence of regions filled with intravenous contrast and without contrast presents a challenge for bladder segmentation. Previously, we proposed a Conjoint Level set Analysis and Segmentation System (CLASS)[11]. In case the bladder is partially filled with contrast, CLASS segments the non-contrast (NC) region and the contrast-filled (C) region separately and automatically conjoins the NC and C region contours; however, inaccuracies in the NC and C region contours may cause the conjoint contour to exclude portions of the bladder. To alleviate this problem, we implemented a local contour refinement (LCR) method that exploits model-guided refinement (MGR) and energy-driven wavefront propagation (EDWP). MGR propagates the C region contours if the level set propagation in the C region stops prematurely due to substantial non-uniformity of the contrast. EDWP with regularized energies further propagates the conjoint contours to the correct bladder boundary. EDWP uses changes in energies, smoothness criteria of the contour, and previous slice contour to determine when to stop the propagation, following decision rules derived from training. A data set of 173 cases was collected for this study: 81 cases in the training set (42 bladders with lesions, 21 bladders with wall thickenings, 18 normal bladders) and 92 cases in the test set (43 bladders with lesions, 36 bladders with wall thickenings, 13 normal bladders). For all cases, 3D hand segmented contours were obtained as reference standard and used for the evaluation of the computerized segmentation accuracy. For CLASS with LCR, the average volume intersection ratio, average volume error, absolute average volume error, average minimum distance and Jaccard index were 84.2±11.4%, 8.2±17.4%, 13.0±14.1%, 3.5±1.9 mm, 78.8±11.6%, respectively, for the training set and 78.0±14.7%, 16.4±16.9%, 18.2±15.0%, 3.8±2.3 mm, 73.8±13.4% respectively, for the test set. With CLASS only, the corresponding values were 75.1±13.2%, 18.7±19.5%, 22.5±14.9%, 4.3±2.2 mm,

71.0±12.6%, respectively, for the training set and 67.3±14.3%, 29.3±15.9%, 29.4±15.6%, 4.9±2.6 mm, 65.0±13.3%, respectively, for the test set. The differences between the two methods for all five measures were statistically significant ($p<0.001$) for both the training and test sets. The results demonstrate the potential of CLASS with LCR for segmentation of the bladder. The results presented in this chapter have been published[1].

2.2 Introduction

Bladder cancer is the fourth most common cancer diagnosed in men. The American Cancer Society estimates that bladder cancer will cause 16,870 deaths (12,240 in men, 4,630 in women) in the United States in 2017, with 76,030 new cases being diagnosed (60,490 in men, 18,540 in women)[12]. Early detection and treatment of bladder cancer increases patient survivability. If bladder cancers are detected and treated while the cancer is confined within the bladder's inner lining but has not invaded the muscular bladder wall, the 5-year survival rate is 88%. If the cancer is detected after it has invaded the bladder wall but is still confined to the bladder, the 5-year survival rate drops to 63%[12].

Multi-detector row CT (MDCT) urography has shown promise of detecting bladder lesions and has become the imaging modality of choice for evaluation of most urinary tract abnormalities, since a single exam can be used to evaluate the kidneys, intrarenal collecting systems, and ureters. CT urography (CTU), therefore, may spare the patients from having to undergo other imaging studies (intravenous pyelogram (IVP), ultrasound, conventional abdominal CT, and even MRI), thereby reducing health care costs[13-17].

Interpretation of a CTU study requires thorough image analysis, often requiring extensive time. On average, 400 slices are generated for each CTU scan at a slice interval of either 1.25 mm or 0.625 mm (range: 200 to 600 slices). The radiologists interpreting the study have to visually determine whether or not lesions are present within the urinary tracts, frequently needing to adjust the brightness and contrast of the images and use zooming from a display workstation. The possibility that multiple lesions may be present requires that the radiologists pay close attention throughout the entire urinary tract. In addition, many different urinary anomalies may be found in a single CTU study. Not only do the radiologists have to identify these anomalies, they must also determine how likely each of them is to be a urothelial neoplasm. The challenges of analyzing a CTU study leads to a substantial variability among radiologists in detection of

bladder cancer, with reported sensitivities ranging from 59% to 92%[18, 19]. Due to the workload of interpreting CTU studies, the likelihood for the radiologists' to miss a subtle lesion may not be negligible, thus any technique that may help radiologists with identification of urothelial neoplasms within the urinary tract will be useful. Computer-aided diagnosis (CAD) system, used as an adjunct may be the tool that reduces the chance of oversight by the radiologists. We are developing a CAD system that detects bladder cancer in CTU, and a critical part of this system is accurate bladder segmentation that isolates the bladder from the surrounding anatomical structures.

Li et al.[20] segmented the bladder wall from magnetic resonance (MR) cytoscopy in 6 patients and analyzed it for suspected lesions using a partial volume segmentation algorithm. Duan et al.[21] used two collaborative level set functions and clustering to segment the bladder, also in MR cytoscopy of 6 patients. In a different study, Duan et al.[22] developed a segmentation method using MR images of 10 patients that used an adaptive window-setting scheme to detect tumor surfaces. Recently, Han et al.[23] of the same group segmented the bladder wall in T1-weighted MR images using an adaptive Markov random field model and coupled level set information in 6 patients. Li et al.[20] did not report quantitative results for the bladder segmentation. Duan et al.[21] and Han et al.[23] evaluated the segmentation performances using radiologists' subjective ratings without reporting quantitative results. Chai et al.[24] presented a method for semiautomatic bladder segmentation on cone beam CT by using a population-based statistical bladder shape calculated using spherical harmonics description, then applying principal component analysis. They used 95 scans from 8 patients to train their method, and validated their method using 233 scans from 22 patients. Segmentation performance was measured using the Jaccard index comparing the segmentation with manual segmentation done slice-by-slice, which was 70.5% at the automatic stage and increased to 77.7% at the following semiautomatic stage. Hadjiiski et al. developed preliminary bladder segmentation methods for CTU using active contour with 15 patients[25] without quantitative results, and level sets with 70 patients, which was evaluated using quality ratings[26].

There are challenges to segmenting bladders in CTU. Bladders may be partially or fully filled with excreted intravenous (IV) contrast material that opacifies a portion of the bladder. The boundaries between the bladder wall and the surrounding soft tissue have very low contrast such that they are often difficult to delineate. In addition, bladders may exist in a

variety of shapes and sizes. To address these challenges, Hadjiiski et al.[27] developed a segmentation package specifically designed based on the characteristics of the bladder in CTU images referred to as Conjoint Level set Analysis and Segmentation System (CLASS) (frequently used acronyms are listed in Table 2.1). The segmentation performance was qualitatively evaluated using 81 bladders and quantitatively evaluated using 30 bladders by comparing the computer-segmented contours to hand-segmented reference contours and obtained promising results.

Table 2.1: Definitions of frequently used acronyms.

AVDIST	Average minimum distance between two contours
C	Contrast-filled
CAD	Computer-aided detection
CCP	Contour conjoint procedure
CLASS	Conjoint level set analysis and segmentation system
CTU	CT urography
EDWP	Energy-driven wavefront propagation
IV	Intravenous
L	Contour of the contrast-filled region of the bladder
LCR	Local contour refinement
MDCT	Multi-detector row CT
MGR	Model-guided refinement
NC	Non-contrast
ROI	Region of Interest
CLASS C Contour	CLASS contour of the contrast-filled region of the bladder
CLASS NC Contour	CLASS contour of the non-contrast-filled region of the bladder

In this study, we further developed the CLASS method by incorporating a local contour refinement procedure to improve the segmentation accuracy and evaluated its performance using a moderately-sized data set, although it was the largest data set of independent subjects compared to those used in other reported studies. The improvement in the segmentation performance of CLASS with the new refinement technique compared to that of CLASS alone was quantified using manual segmentation as reference standard.

2.3 Materials and Methods

2.3.1 Bladder segmentation using CLASS

Our previous work on CLASS[27] is described briefly as follows. An axial CTU scan of the bladder is shown in Figure 2.1. Figures 2.2 and 2.3 show two regions of interest (ROI) from different CTU slices that contain the bladder. The bladders shown are partially filled with IV contrast material, and malignant lesions of different sizes can be identified in the lower, contrast-enhanced portion of the bladder. The presence of the two distinct areas that have very different attenuation values, an area filled with excreted IV contrast material and an area without contrast material (Figures 2.2 and 2.3), poses a challenge for segmentation that needs to go across the strong boundary between the two areas.

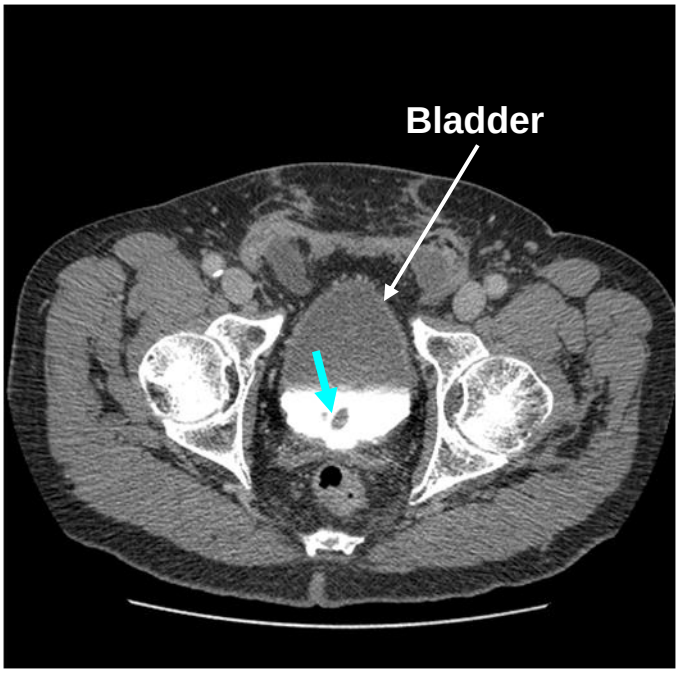


Figure 2.1: An axial slice of a CTU scan in which the bladder is partially filled with IV contrast material. A malignant lesion is present in the contrast-filled region of the bladder, indicated by the light blue arrow.

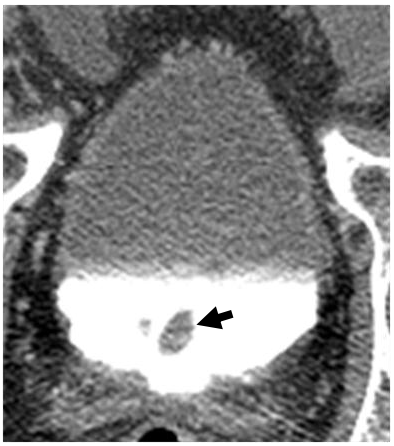


Figure 2.2: ROI of a bladder partially filled with IV contrast material, showing the two distinct areas. A malignant lesion is present in the contrast-filled region of the bladder (black arrow).

Figure 2.3: A large malignant lesion is located in the contrast-filled portion of the bladder (black arrow).

We are developing a software package, referred to as the Conjoint Level set Analysis and Segmentation System (CLASS), for segmenting the bladder in CTU. We introduced CLASS previously[27] with limited quantitative assessment of its performance including comparison of its segmentation accuracy to radiologist's hand outlines in a small data set and their qualitative visual judgment on a larger data set.

CLASS consists of four stages: (1) preprocessing and initial segmentation, (2) 3D level set segmentation, (3) 2D level set segmentation, and (4) post-processing. CLASS segments the non-contrast (NC) region and the contrast-filled (C) region by applying the level sets to each region separately, then automatically conjoins them during post-processing using a procedure called Contour Conjoint Procedure (CCP). More details of CLASS can be found in the literature[27].

2.3.2 Local Contour Refinement

Although CLASS performs reasonably well in comparison to hand segmentations, achieving a large volume intersection ratio for more than 70% of bladders[27], a number of cases failed because the level set propagation in the C region may stop prematurely due to either substantial non-uniformity of the contrast, or a lesion present in the region. In this study, we designed a new Local Contour Refinement (LCR) method to further improve the segmentation

accuracy, especially to improve the segmentation of the C region, and including lesions that may be present.

Local contour refinement consists of two main processes: model-guided refinement (MGR) and energy-driven wavefront propagation (EDWP). MGR propagates the contour of the C region (*L*) (See Figure 2.4) using local moving windows if the level set segmentation stops prematurely. EDWP uses regularized local energies to propagate the occasionally imperfect conjoining contour segments that conjoin the contours of NC and C regions, to the correct bladder boundary. The block diagram of CLASS with LCR is presented in Figure 2.4. The parameter values for the proposed methods were determined experimentally using cases from the training set. By altering one parameter at a time, we studied the sensitivity of the parameters (i.e., the effect of the change of each parameter on the quality of the automatic 3D contour), both by visual inspection and by the performance measures. We fixed the less sensitive parameters first, and altered the parameters with higher sensitivity to achieve the best segmentation. Once the parameters were fixed based on the training set, the CLASS with LCR was applied to the test set with that fixed parameter set without further changes.

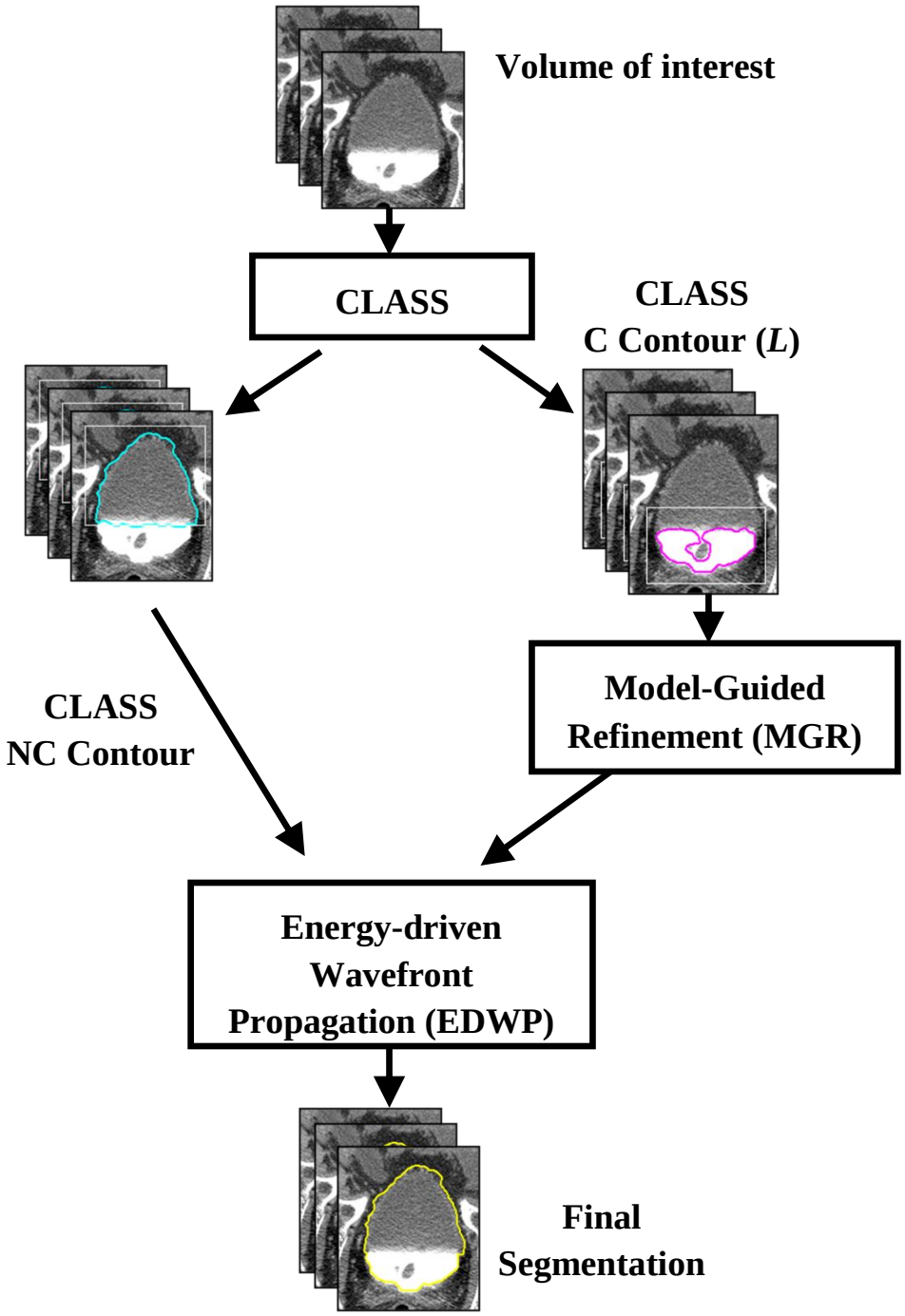


Figure 2.4: Block diagram of the CLASS with LCR. MGR is performed on the CLASS contour of the contrast-filled region. The CLASS contour of the non-contrast region and the contour *L* of contrast-filled region after MGR are joined and refined by EDWP to obtain the final contour of the bladder.

2.3.2.1 Model-guided refinement

The model-guided refinement (MGR) method was designed to improve the CLASS C contour in 2D for every slice. Inaccuracies in the NC and C contours may cause the Contour Conjoint Procedure (CCP) to exclude portions of the bladder. To alleviate this problem, we implemented an MGR method to propagate the C region contour, *L*, if the level set propagation in the C region stops prematurely due to substantial non-uniformity of the contrast. MGR uses the level set contour, the local gradient and contrast as input and incorporates adaptive thresholding to propagate the *L* contour in 2D for every slice to the correct bladder boundary. The parameters used for the MGR method was determined experimentally using the training set.

MGR propagates the level set *L* contour by analysis of three moving windows, all of 3 X 3 pixels, at each point L_i along the *L* contour: one centered at the point L_i, the other two positioned at an inner and outer neighborhood of the contour, respectively. The center of the outer window (*WO)* is located outside the contour in the normal direction, two pixels away from L_i. The center of the inner window (*WI)* is located inside the contour, also in the normal direction, two pixels away from L_i (Figure 2.5). The average intensities of the three windows are calculated as follows:

$$I_W(L_i) = \frac{\sum_{P \in W} I(P)}{N_W}, \tag{2.1}$$

$$I_{WO}(L_i) = \frac{\sum_{P \in WO} I(P)}{N_{WO}}, \tag{2.2}$$

$$I_{WI}(L_i) = \frac{\sum_{P \in WI} I(P)}{N_{WI}}, \tag{2.3}$$

where $I_W(L_i)$ is the average intensity of the window *W* centered at L_i, obtained by taking the average of the intensity of the pixels *I*(*P*) within the window *W*. $I_{WO}(L_i)$ and $I_{WI}(L_i)$ are the average intensity of the *WO* and *WI* windows, respectively, calculated in a similar way as $I_W(L_i)$. N_W, N_{WO}, and N_{WI} are the areas of the *W*, *WO*, and *WI* windows, respectively, and $N_W = N_{WO} = N_{WI} = 9$ pixels.

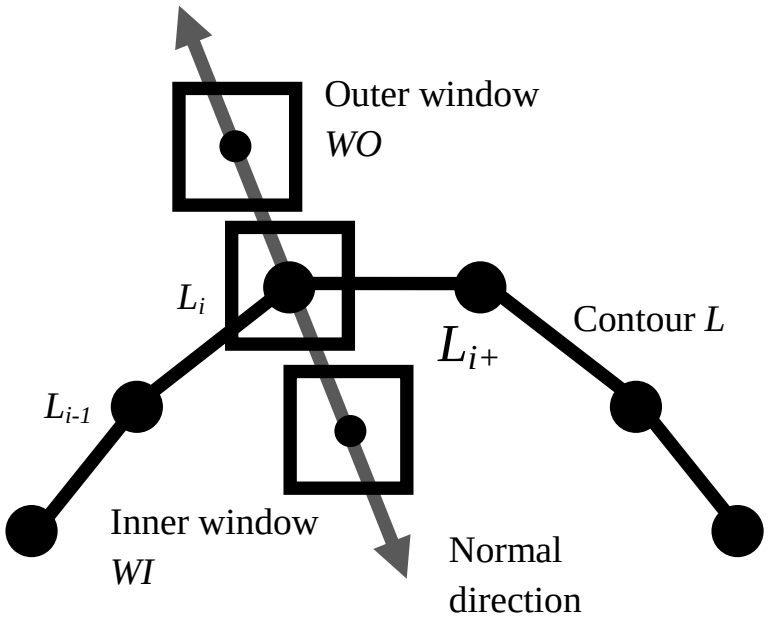


Figure 2.5: Diagram of inner and outer window used for MGR. L_i is the given point on contour L that may be propagated. L_{i+1} and L_{i-1} are the next and previous points, respectively, on the contour neighboring the point L_i. The inner and outer windows are located two pixels away from L_i and centered along the normal. The inner window *WI* is located towards the centroid of the contour L. The outer window *WO* is located outside the contour.

The normal direction is determined by the normal angle θ defined as:

$$\theta = \frac{1}{2}\left(\tan^{-1}\left(\frac{y_i - y_{i-1}}{x_i - x_{i-1}}\right) + \tan^{-1}\left(\frac{y_{i+1} - y_i}{x_{i+1} - x_i}\right)\right), \tag{2.4}$$

where (x_i, y_i) is the coordinate of the given point L_i on the contour, and (x_{i+1}, y_{i+1}) and (x_{i-1}, y_{i-1}) are the coordinates of the next L_{i+1} and previous L_{i-1} neighboring points on the contour to the point L_i, respectively (Figure 2.5).

The point L_i is propagated if the condition in either Equation (2.5) or Equation (2.6) is satisfied:

$$I_{WO}(L_i) - I_{WI}(L_i) > 200, \tag{2.5}$$

$$\begin{cases} I_W(L_i) > Th \\ I_{WO}(L_i) - I_{WI}(L_i) > -400, \end{cases} \tag{2.6}$$

where *Th* is the optimal threshold determined at the beginning of the MGR by the process explained below. *Th* is kept constant during the propagation of all L_i points for all contours in a case. Equation (2.5) propagates the contour while the contour is within the C region of the bladder, propagating the contour to areas with positive gradient. Equation (2.6) propagates the contour even if the non-uniformity of the contrast creates areas of negative gradient. Equation

(2.6) also propagates the contour to encompass the bladder wall once the contour reaches the edge of the C region. If one of these conditions is met, point L_i propagates to a point one pixel away in the normal direction, determined by Equation (2.4) in a single iteration. Multiple iterations are carried out to propagate the contour. The propagation of the contour stops when all points L_i stop moving. Smoothness of the contour is maintained by automatically adding points to the contour if the distance between consecutive points on the contour exceeds 6 pixels where the contour becomes too sparse, and removing points from the contour that cause sharp angles, defined as angle less than 90 degrees.

The optimal threshold *Th* for the MGR is selected based on an adaptive method and is specific for every case. The *Th* is selected at the beginning of MGR. First, the mean (μ) and the standard deviation (σ) of the voxel intensities within the volume defined by the 3D level set CLASS contours of the C region are calculated. A set of candidate thresholds Th_m are estimated by $Th_m = \mu - n_m\sigma$, where n_m = 0, 0.5, 1, …, 10. For every candidate threshold, the level set contours on the best slice *b* (the slice that best represents the bladder region, where the bladder is seen the largest, which was manually selected when the ROI was defined), the slice before (*b-1*), and the slice after (*b+1*), are propagated. A given point L_i on the 2D contours of the three slices is propagated if it satisfies both Equation (2.5) and Equation (2.6):

To further prevent leaking, a criterion is set to determine adaptively how much a subsequent contour can expand from the current contours. We first calculated the average minimum distance between two contours (*AVDIST*) as:

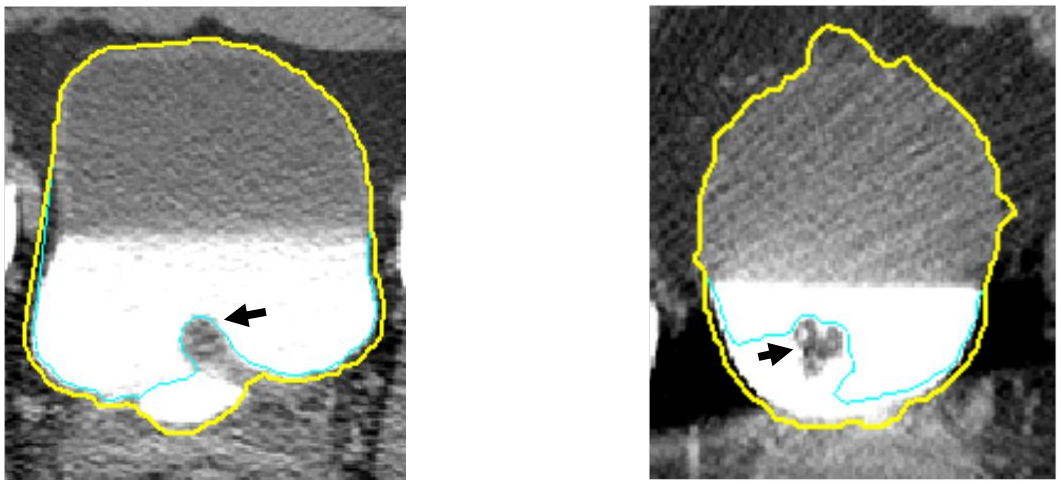
$$AVDIST(G,U) = \frac{1}{2}\left(\frac{\sum_{x\in G}\min\{d(x,y): y \in U\}}{N_G} + \frac{\sum_{y\in U}\min\{d(x,y): x \in G\}}{N_U}\right), \quad (2.7)$$

where *G* is the current contour, and *U* is the subsequent contour. N_G and N_U denote the number of voxels on *G* and *U*, respectively. The function *d* is the Euclidean distance. For a given voxel along the contour *G*, the minimum distance to a point along the contour *U* is determined. The minimum distances obtained for all points along *G* are averaged. This process is repeated by switching the roles of *G* and *U*. *AVDIST* is then calculated as the average of the two average minimum distances.

For every candidate threshold Th_m, the relative contour propagation increase (*RCPI*) is defined as:

$$RCPI = \frac{1}{2}\left(AVDIST\left(S^{(b)}, S^{(b-1)}\right) + AVDIST\left(S^{(b)}, S^{(b+1)}\right)\right), \quad (2.8)$$

where $S^{(b)}, S^{(b-1)}, S^{(b+1)}$ are the contours obtained by propagating level set *L* contours on the best slice *b*, the slice before *b-1*, and the slice after *b+1*, respectfully, using the candidate threshold Th_m and Equations (2.5) and (2.6). The optimal threshold *Th* is selected to be the largest Th_m that provides less than 40% increase in *RCPI* from the *RCPI* obtained from the previous threshold, Th_{m-1}. *Th* is fixed for the entire case, and is used to obtain the *L* contours on other slices of the case by propagating their respective level set *L* contours using Equations (2.1) through (2.6), applying methods as mentioned above. Obtaining the optimal threshold by this method helps prevent MGR from leaking. Figure 2.6 shows an example comparing segmentation results with and without MGR.



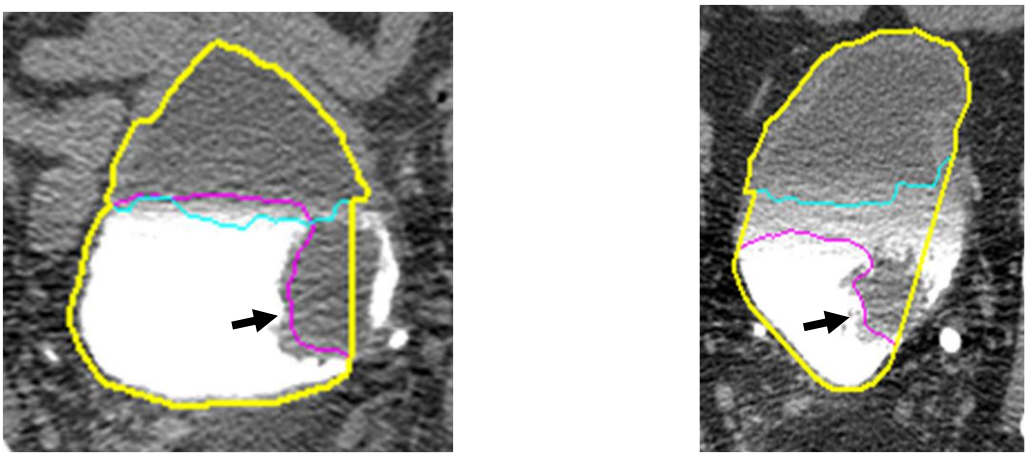
(a) (b)

Figure 2.6: Bladder segmentation using CLASS with and without MGR. (a) CLASS excluded portions of the bladder due to a malignant lesion (black arrow) attached to the bladder wall, but MGR propagated the *L* contour through the lesion to the correct bladder boundary. (b) MGR resolved a similar problem with a malignant mass lesion preventing the *L* contour from correctly segmenting the bladder. The yellow contour represents CLASS with MGR. The light blue contour shows CLASS without MGR.

2.3.2.2 Energy-driven wavefront propagation (EDWP)

The EDWP method was designed to smooth the transition between the conjoint C and NC contours and further pushes the contour towards the true bladder boundary, as described in the following. Inaccuracies in the contours of the NC and C regions may cause Contour Conjoint Procedure (CCP) to exclude portions of the bladder. In cases where a lesion is present in the C region, MGR may not propagate through the lesion due to the edge of the lesion having similar edge properties as the bladder wall. When the level set contour of the NC region and the contour *L* of C region after MGR are conjoined, CCP may create erroneous connection between

the two contours, as seen by the straight connections between the contours of NC and C regions in Figure 2.7. This conjoining segment may cut across a lesion or the bladder, excluding other portions of the bladder in the process. The EDWP method combines local energies with *a priori* knowledge on contrast-enhanced bladder on CT images to guide the contour to search for the true bladder boundary.



(a) (b)

Figure 2.7: Bladder segmentation using CLASS with MGR. Inaccuracies in the NC region contour and C region contour *L* may cause CCP to exclude portions of the bladder. MGR was unable to propagate past the lesions (black arrow) in the C region to the correct bladder wall in both (a) and (b). (a) The malignant lesion extends to the NC area, thus MGR did not propagate the *L* contour through the lesion. (b) The inhomogeneous nature of the boundary between the NC and C regions of the bladder prevented MGR from propagating the *L* contour fully, missing the malignant lesion. The light blue contour represents the CLASS NC region contour. The pink contour represents the *L* contour with MGR. The yellow outer contour is the result of conjoining the NC region contour and *L* contour without LCR.

EDWP propagates the points on the conjoining contour segment to the bladder boundary by calculating the local energies $E_R(Z_i)$, $E_{OutR}(Z_i)$, and $E_{OrigR}(Z_i)$. The local energy $E_R(Z_i)$ is the energy of a region R_i centered at a given point Z_i along the conjoining segment. R_i includes Z_i and the points on the conjoining segment that are a maximum of two points away from Z_i, i.e. $Z_{i\pm2}$, $Z_{i\pm1}$, therefore Zi, $Zi_{\pm1}$, $Z_{i\pm2} \in R_i$. The energy term $E_R(Z_i)$ is defined as:

$$E_R(Z_i) = \frac{\sum_{P \in R_i} \alpha I_w(P)}{T_\alpha}, \tag{2.9}$$

where $I_W(P)$ is the average intensity of a window defined by Equation (2.1), but a window size of 5 X 5 pixels is used here. α is the weight assigned to $I_W(P)$: α is 4 for the energy of Z_i and 1 for the rest of the reference points in R_i. T_α is the total weight, which is 8.

The local energy $E_{OutR}(Z_i)$, is the energy of the region R_i^{out} obtained by propagating all points in R_i one pixel outside of the contour in their respective normal direction, estimated by the normal angle (Equation (2.4)). $E_{OutR}(Z_i)$ is defined as:

$$E_{OutR}(Z_i) = \frac{\sum_{P \in R_i^{out}} \alpha I_w(P)}{T_\alpha}, \tag{2.10}$$

where $I_W(P)$, α , and T_α are defined as in Equation (2.9) above.

$E_{OrigR}(Z_i)$ is the local energy of the region R_i^{orig} centered at the original Z_i in the conjoint segment before propagation with EDWP. $E_{OrigR}(Z_i)$ is defined as:

$$E_{OrigR}(Z_i) = \frac{\sum_{P \in R_i^{orig}} \alpha I_w(P)}{T_\alpha}, \tag{2.11}$$

where $I_W(P)$, α , and T_α are defined as in Equation (2.9) above. Before the first iteration, $E_R(Z_i) = E_{OrigR}(Z_i)$.

The local energies $E_R(Z_i)$, $E_{OutR}(Z_i)$, and $E_{OrigR}(Z_i)$ are calculated for every iteration of the propagation. The EDWP monitors the changes in the energy to determine when to stop the propagation. The point Z_i is propagated if both the following conditions are satisfied:

$$E_{OutR}(Z_i) - E_R(Z_i) > -20, \tag{2.12}$$

$$E_{OutR}(Z_i) - E_{OrigR}(Z_i) > -20, \tag{2.13}$$

Equation (2.12) propagates the conjoining segment through lesions and C region of the bladder until the drop in $E_{OutR}(Z_i)$ energy exceeds the criterion, usually the darker tissue background around the bladder wall. Equation (2.13) stops the propagation in cases where Equation (2.12) is still satisfied when the conjoint segment is at the bladder wall. Using these conditions allows the segment to propagate through regions of higher energy such as contrast material (Figure 2.8), and eventually stop at a region of low energy (the darker tissue background around the bladder wall (Figure 2.8)). If both criteria for propagation are satisfied, Z_i propagates to the pixel located one pixel outside of the contour in the normal direction, estimated by Equation (2.4), to the conjoining segment. After propagating all points along the conjoint segment for a given iteration, smoothness of the contour is maintained by limiting the smoothing energy terms in 2D and 3D as follows.

The 2D smoothness criterion is applied sequentially to all regions R_i, along the conjoint segment. The 2D smoothing energy is defined as:

$$E_{Smooth\ 2D} = \max\{\gamma(Z_i, P): P \in Z_{i\pm1}, Z_{i\pm2}\}, \tag{2.14}$$

where the function γ provides the number of iterations Z_i propagated after a point P among $Z_{i\pm1}$ and $Z_{i\pm2}$ stopped propagating. Within a region R_i, Z_i stops propagating if $E_{Smooth\ 2D}$ is greater than 4, even if the energy criteria is met. Using this method ensures that Z_i does not over-propagate by accounting for the changes in energy at the reference points in R_i.

3D smoothing energy is defined as:

$$\mathrm{E}_{Smooth\ 3D} = \min\{d(Z_i, P): P \in PS\}, \tag{2.15}$$

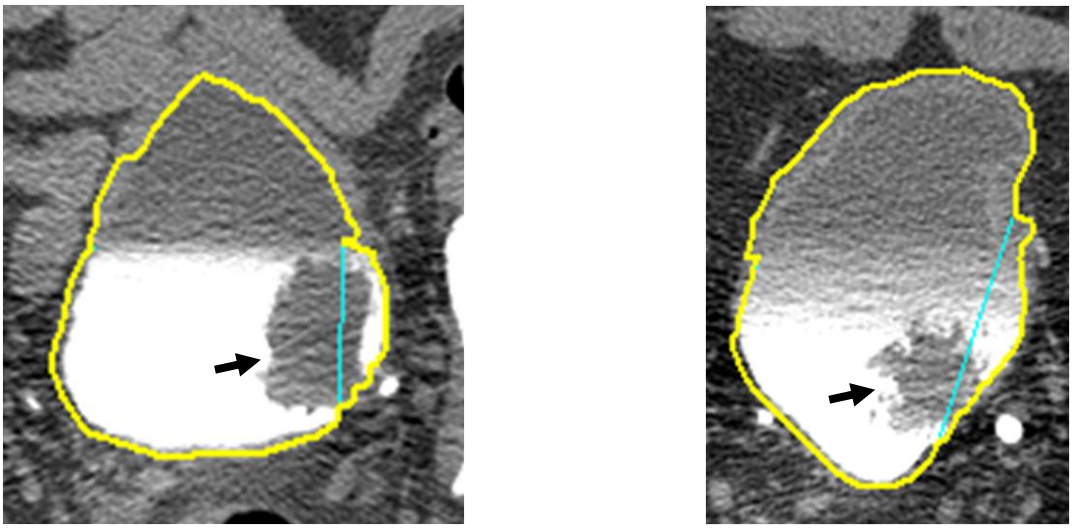
where the function d is the Euclidean distance. PS is the previous slice (j-1) located in the cephalic aspect of the current slice (j). $E_{Smooth\ 3D}$ measures the minimum distance between Z_i on slice j and the points on the previous contour on slice j-1. If this value is greater than 5 voxels, Z_i stops propagating. Additional steps to maintain smoothness include adding points to the contour if the contour becomes too sparse defined by the same criteria as MGR, and removing points from the contour that cause sharp angles, defined as angle less than 90 degrees.

Once propagation stops the energy of the conjoining segment follows the relationships below:

$$\sum_{i=1}^{Q} E_{OutR}(Z_i) < \sum_{i=1}^{Q} E_R(Z_i) \tag{2.16}$$

$$\sum_{i=1}^{Q} E_{OutR}(Z_i) < \sum_{i=1}^{Q} E_{OrigR}(Z_i) \tag{2.17}$$

where Q is the number of points along the conjoining segment. Figure 2.8 shows examples after improvement by applying EDWP to the cases in Figure 2.7.



(a) (b)

Figure 2.8: Bladder segmentation using CLASS with LCR. (a) Energy-driven wavefront propagation (EDWP) propagated the conjoint contour past the malignant lesion (black arrow) to the correct bladder boundary. (b) Inaccuracy with *L* contour after MGR caused CCP to exclude portions of the bladder, but EDWP propagated the contour to the correct bladder boundary. The light blue contour represents contour without EDWP. The yellow contour represents the contour with EDWP.

2.3.4 Data set

In this study, a data set of 173 patients undergoing CTU who subsequently underwent cystoscopy and biopsy was utilized. The cases were collected retrospectively from the Abdominal Imaging Division of the Department of Radiology at the University of Michigan with approval of the Institutional Review Board. We designated 81 of these cases as the training set, and the other 92 cases as the test set. The cases were assigned to the training or the test sets by balancing the difficulty of the cases between the two sets by researcher who was trained by an experienced abdominal radiologist.

Of the 81 training set bladders, 42 bladders contained focal mass-like lesions (40 malignant and 2 benign), 21 bladders had wall thickening (16 malignant and 5 benign) and 18 were normal. In these cases, 61 bladders were partially filled with IV contrast material, 8 were completely filled with contrast material, and 12 had no visible contrast material. Of the 92 test set bladders, 43 bladders contained focal mass-like lesions (42 malignant and 1 benign), 36 bladders had wall thickening (23 malignant and 13 benign) and 13 were normal. In these cases, 85 bladders were partially filled with IV contrast material, 4 were completely filled with contrast material, and 3 had no visible contrast material. The bladder conspicuity was medium to high.

The MDCT urography scans used in this study were acquired with GE Healthcare LightSpeed MDCT scanners. Excretory phase images, obtained 12 minutes after the initiation of

intravenous contrast injection at a concentration of 300 mg iodine per ml, were utilized. The images used were acquired at an interval and slice thickness of 1.25 mm or 0.625 mm using 120 kVp and 120–280 mA. Since patients were not turned prior to image acquisition, dependently layering IV contrast material that had been excreted into the renal collecting systems partially filled the bladder on the CTU images.

2.3.5 Evaluation methods

Segmentation performance was evaluated by quantitative methods comparing the computer segmentation result to the reference standard. The 3D hand-segmented contours for all 173 cases were obtained as reference standard in this study. An experienced radiologist provided manual outlines on the CT slices for all cases using an in-house developed graphic user interface (GUI). The radiologist outlined the bladder on every 2D CT slice on which the bladder was visible, resulting in a 3D surface contour. There were a total of 16,197 slices manually outlined by the radiologist for the 173 bladders. Several performance metrics [28] that quantify the similarity of a pair of contours were used for evaluating the system, including the volume intersection ratio, the volume error, the average minimum distance, and the Jaccard index [29], between the hand-segmented contours and computer segmented contours.

The volume intersection ratio is the ratio of the intersection between the reference volume and the given volume to the reference volume:

$$R^{3D} = \frac{V_G \cap V_U}{V_G}, \tag{2.19}$$

where V_G is the volume enclosed by the reference standard contour G and V_U is the volume enclosed by the contour U being evaluated. A value of 1 indicates that V_U completely overlaps with V_G, while a value of 0 means V_U and V_G do not overlap.

The volume error is the ratio of the difference between the reference volume and the given volume to the reference volume:

$$E^{3D} = \frac{V_G - V_U}{V_G}, \tag{2.20}$$

where positive error indicates under-segmentation and negative error indicates over-segmentation. Because the over- and under-segmentation tend to mask the actual deviations from the reference standard when the average is taken, the absolute error $|E^{3D}|$ is also calculated.

The average distance, *AVDIST*, is the average of the distances between the closest points of the two contours already defined in Section 2.2.1, Equation (8). For this calculation, *G* is the 3D reference contour marked by the radiologist and *U* is the 3D contour being evaluated.

The Jaccard index is defined as the ratio of the intersection between the reference volume and the segmented volume to the union of the reference volume and the segmented volume:

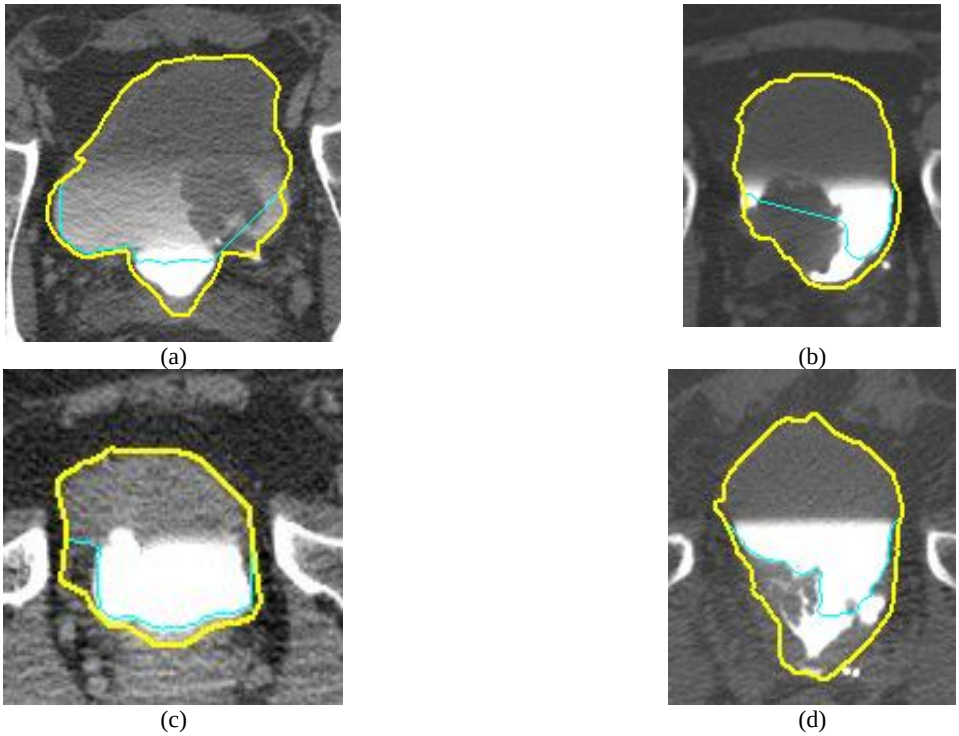
$$JACCARD^{3D} = \frac{V_G \cap V_U}{V_G \cup V_U}, \tag{2.21}$$

A value of 1 indicates that V_U completely overlaps with V_G, whereas a value of 0 implies V_U and V_G are disjoint.

No single measure can completely describe the agreement between the two volumes; however, by combining two performance measures, different aspects of the performance can be assessed. For example, the Jaccard index, the overlap and non-overlap fractions with the reference standard, can be derived from the volume intersection ratio and the volume error [30].

2.4 Results

Examples of CLASS NC and C region contours and CLASS with LCR segmentation are shown in Figure 2.9. The segmentation performance measures averaged over the cases in the training and test sets, respectively, are presented in Tables 2.2 and 2.3.



(a) (b) (c) (d)

Figure 2.9: Bladder segmentation of test cases using CLASS with and without LCR. (a) CLASS missed portions of the malignant lesion and failed to segment the bladder boundary in the C region, whereas LCR segmented the bladder more accurately. (b) The large malignant lesion in the C region was mostly missed by CLASS; however, LCR fully segmented the lesion. (c) Benign wall thickening was missed by CLASS, but LCR segmented the bladder more accurately. (d) A difficult malignant lesion was missed with CLASS, but LCR propagated the contour through the lesion to the bladder boundary. The light blue contour represents segmentation results using CLASS. The yellow contour represents segmentation result from CLASS with LCR.

Table 2.2: Segmentation results from CLASS with and without LCR, averaged over the 81 bladders in the training set. The value for the method that performed better is in bold.

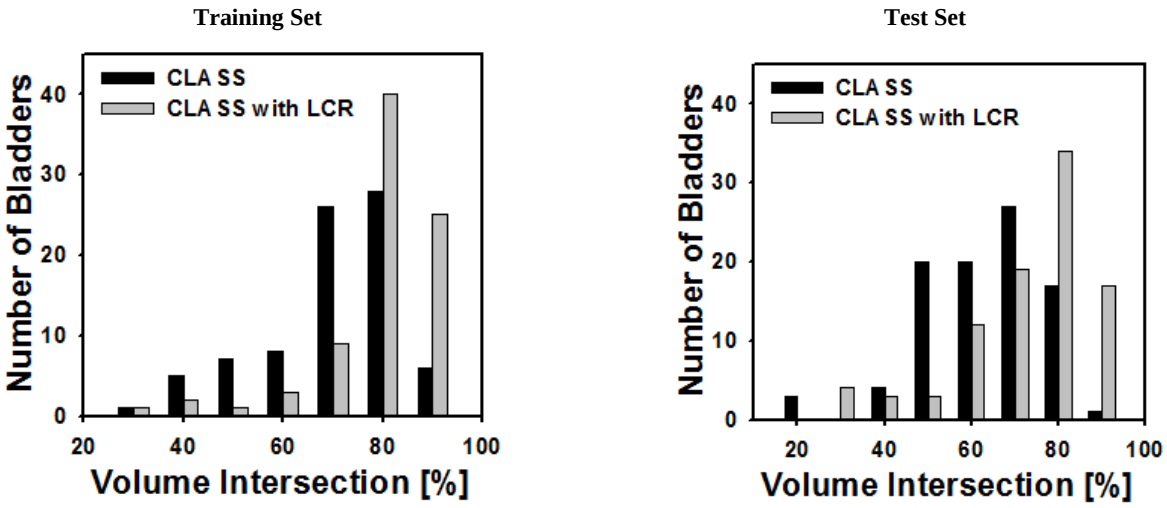
Segmentation Method (Training set)	Volume intersection ratio R^{3D}	Volume error E^{3D}	Absolute volume error $\lvert E^{3D} \rvert$	Average minimum distance *AVDIST*	Jaccard index $JACCARD^{3D}$
CLASS with LCR	**84.2±11.4%**	**8.2±17.4%**	**13.0±14.1%**	**3.5±1.9 mm**	**78.8±11.6%**
CLASS without LCR	75.1±13.2%	18.7±19.5%	22.5±14.9%	4.3±2.2 mm	71.0±12.6%

Table 2.3: Segmentation results from CLASS with and without LCR, averaged over the 92 bladders in the test set. The value for the method that performed better is in bold.

Segmentation Method (Test set)	Volume intersection ratio R^{3D}	Volume error E^{3D}	Absolute volume error $\lvert E^{3D} \rvert$	Average minimum distance *AVDIST*	Jaccard index $JACCARD^{3D}$
CLASS with LCR	**78.0±14.7%**	**16.4±16.9%**	**18.2±15.0%**	**3.8±2.3 mm**	**73.8±13.4%**
CLASS without LCR	67.3±14.3%	29.3±15.9%	29.4±15.6%	4.9±2.6 mm	65.0±13.3%

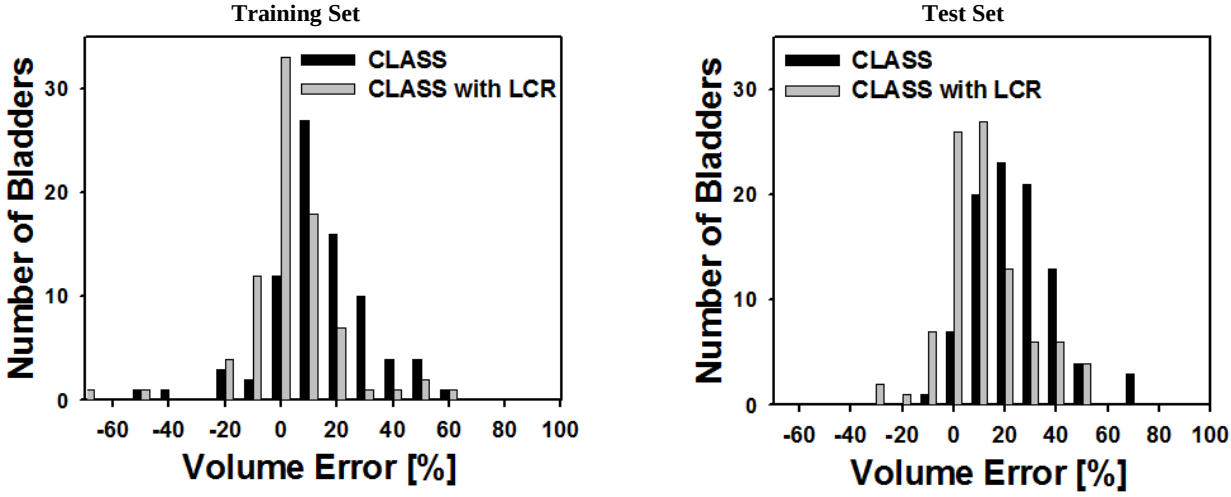
CLASS with LCR consistently performed better than CLASS for the data set used. In comparison to CLASS only, CLASS with LCR resulted in larger volume intersection ratio, smaller volume error, smaller absolute volume error, smaller average distance error, and larger Jaccard index for both training and test sets. The standard deviations for all 5 measures were smaller with LCR than without LCR for the training set. For the test set, the standard deviations were smaller with LCR than without LCR for the absolute volume error and average distance error. The differences for all of the performance measures between the two methods were statistically significant for both the training and test sets ($p<0.001$ by two-tailed paired t-test) at alpha level of 0.01 after the Bonferroni correction for the 5 comparisons.

The histograms for volume intersection ratio, volume error, and average distance for both the training and test sets are shown in Figures 2.10, 2.11, and 2.12, respectively.



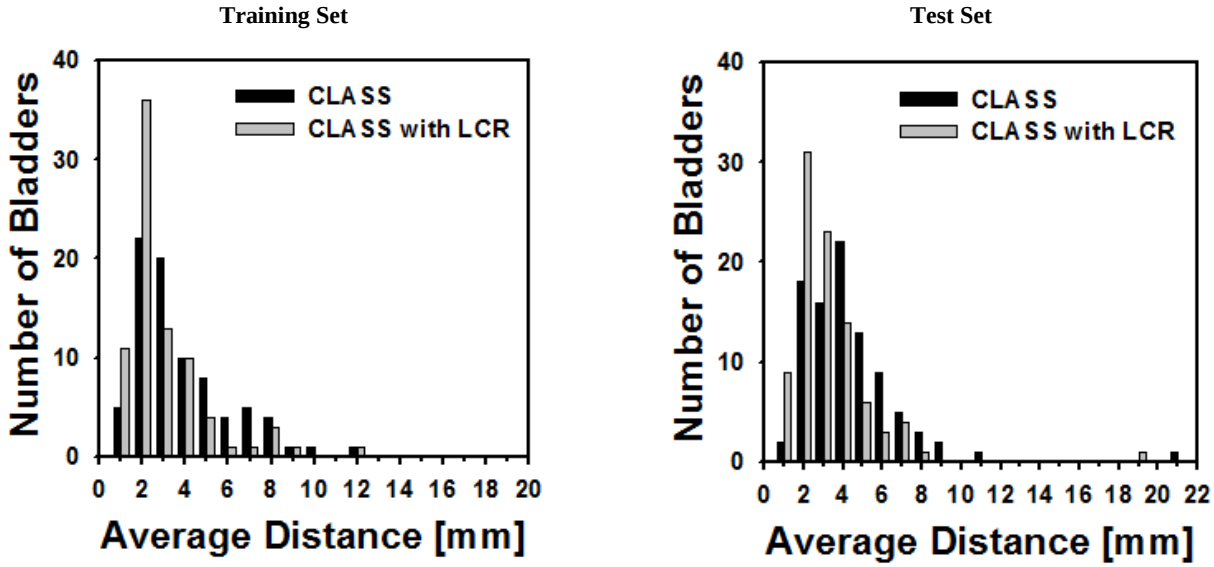
(a) (b)

Figure 2.10: Histogram of percent volume intersection ratio for (a) the training set and (b) the test set for CLASS and CLASS with LCR. The improvement by LCR was statistically significant ($p<0.001$) for both the training and test sets.



(a) (b)

Figure 2.11: Histogram of the volume error for (a) the training set and (b) the test set for CLASS and CLASS with LCR. The improvement by LCR was statistically significant ($p<0.001$) for both the training and test sets.



(a) (b)

Figure 2.12: Histogram of the average minimum distance for (a) the training set and (b) the test set for CLASS and CLASS with LCR. The improvement by LCR was statistically significant ($p<0.001$) for both the training and test sets.

Of the 81 cases in the training set, CLASS with LCR had 65 bladders with a volume intersection ratio greater than 80% whereas CLASS had 34 bladders (Figure 2.10(a)). There were 63 bladders whose absolute volume error for the training set was less than 20% for CLASS with LCR, compared to 41 bladders for CLASS only (Figure 2.11(a)). Forty-seven bladders in the training set had an average distance less than 3 mm for CLASS with LCR, compared to 27 bladders for CLASS only (Figure 2.12(a)).

Of the 92 test cases, CLASS with LCR had 51 bladders with a volume intersection ratio greater than 80% whereas CLASS had only 18 bladders (Figure 2.10(b)). There were 60 bladders whose absolute volume error for the test set was less than 20% for CLASS with LCR, compared to 28 bladders for CLASS only (Figure 2.11(b)). Forty bladders in the test set had an average distance less than 3 mm for CLASS with LCR, compared to 20 bladders for CLASS only (Figure 2.12(b)).

2.5 Discussion

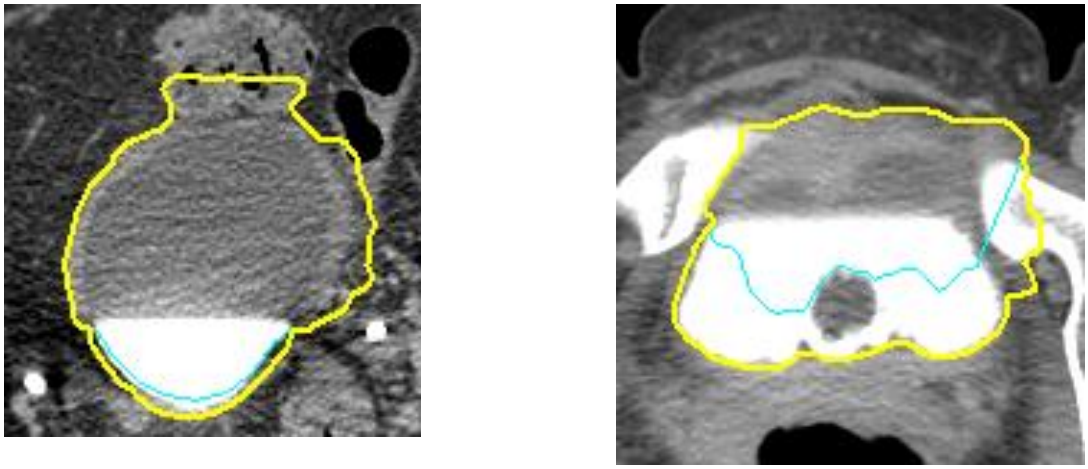
In this study, CLASS with the new LCR method was applied to a data set containing bladders in CTUs having a wide range of image quality. Most of the bladders were partially filled with excreted contrast material; however, some bladders were entirely filled and others did

not contain any contrast-enhanced urine due to variation in the rate of excreted contrast material accumulation in the bladder. CLASS with LCR performed better than CLASS for every performance measure. The LCR more than doubled the number of bladders with volume intersection ratio greater than 80%, absolute volume error less than 20%, and average distance less than 3 mm for the test set. The number of bladders with the above mentioned performance measures also increased for each measure in the training set. Figures 2.8 and 2.9 show examples of difficult bladder cases successfully improved by LCR. The bladders in Figure 2.8 and Figures 2.9(a) and 2.9(b) contained large malignant lesions in the contrast-enhanced area that CLASS had missed. The benign wall thickening in Figure 2.9(c) was missed by CLASS. The case in Figure 2.9(d) contains an inhomogeneous malignant lesion that was also missed by CLASS. In each of these cases, CLASS could not accurately segment the C portion of the bladder due to the strong edge between the large abnormality and the contrast agent, resulting in segmentation excluding portions of the bladder and the lesion. LCR propagated the conjoint contour to the proper bladder boundary, enclosing the previously mentioned features, thus achieving a more accurate segmentation.

It is difficult to perform direct comparison to the previous methods by other investigators summarized in the Introduction due to the differences in the data sets and their varying difficulty. A relative comparison was performed to only one of the studies[24], in which quantitative results were reported. Chai et al.[24] achieved Jaccard indices of 70.5% and 77.7% for their automatic and semiautomatic methods, respectively, using 95 scans of 8 patients for training, and using 233 scans of 22 patients for testing. In comparison, the Jaccard indices of our CLASS with LCR method were 78.8% and 73.8% for the training (81 patients) and test set (92 patients), respectively. CLASS with LCR achieved accuracy comparable to that of Chai et al.[24] while using a larger independent data set. CLASS with LCR also does not require initial manual delineation of the bladder, and also does not require manual correction of the contours, unlike the automatic and semiautomatic methods, respectively, of Chai et al.[24].

There are cases that LCR still could not segment the bladder accurately. LCR did not refine the non-contrast filled region contour from CLASS; therefore, the method was not able to reliably stop the NC region contour at the bladder boundary when a complex background is present, as shown in the upper boundary of the example in Figure 2.13(a). LCR did, however, propagate the C region contour L properly to the correct lower bladder boundary, compared to

CLASS. In addition, LCR does not solve the problem if errors in the NC region contour cause the initial conjoining segment used for EDWP to cut across the bone. As shown in Figure 2.13(b), LCR propagated the contour *L* to include a lesion and portions of the contrast-filled region of the bladder that CLASS had missed; however, the inaccuracy in the NC region contour created an initial conjoint segment through the bone. LCR was not able to identify this error, and propagated the contour segment within the bone, due to the similar appearance of the bone and the bone marrow to a C region containing a lesion, causing leaking in the segmentation. Better criteria to prevent leakage into adjacent normal tissue and bone are needed. It may also be necessary to develop additional local refinement methods to improve segmentation accuracy.



(a) (b)

Figure 2.13: Bladder segmentation using CLASS with LCR. (a) shows leaking in the NC contour due to complex boundary. (b) shows segmentation leaking into the bone. The light blue contour represents segmentation results using CLASS. The yellow contour represents segmentation result from CLASS with LCR.

2.6 Conclusion

The results show that the LCR method can significantly improve the segmentation of bladders by CLASS on CTU scans. LCR propagates the CLASS contour of the contrast-filled region of the bladder and propagates the conjoint contour of the contrast and non-contrast filled region to the correct bladder boundaries using both 2D and 3D information. Further work to prevent over-segmentation into adjacent organs or bone edges is discussed in Chapter III. This study is a step toward the development of a reliable and efficient system for segmentation of bladders, which is a critical component of a CAD system for detection of urothelial lesions imaged with CT urography.

Chapter III
Urinary Bladder Segmentation in CT Urography using Deep-Learning Convolutional Neural Network and Level Sets

3.1 *Abstract*

A Deep-Learning Convolutional Neural Network (DL-CNN) was trained to distinguish between the inside and the outside of the bladder using 160,000 regions of interest (ROI) from CTU images. The trained DL-CNN was used to estimate the likelihood of an ROI being inside the bladder for ROIs centered at each voxel in a CTU case, resulting in a likelihood map. Thresholding and hole-filling were applied to the map to generate the initial contour for the bladder, which was then refined by 3D and 2D level sets. The segmentation performance was evaluated using 173 cases: 81 cases in the training set (42 bladders with lesions, 21 bladders with wall thickenings, 18 normal bladders), and 92 cases in the test set: (43 bladders with lesions, 36 bladders with wall thickenings, 13 normal bladders). The computerized segmentation accuracy using the DL likelihood map was compared to that using a likelihood map generated by Haar features and a random forest classifier, and that using our previous Conjoint Level set Analysis and Segmentation System (CLASS) without using a likelihood map. All methods were evaluated relative to the 3D hand-segmented reference contours. With DL-CNN-based likelihood map and level sets, the average volume intersection ratio, average percent volume error, average absolute volume error, average minimum distance, and the Jaccard index for the test set were 81.9±12.1%, 10.2±16.2%, 14.0±13.0%, 3.6±2.0 mm, and 76.2±11.8%, respectively. With the Haar-feature-based likelihood map and level sets, the corresponding values were 74.3±12.7%, 13.0±22.3%, 20.5±15.7%, 5.7±2.6 mm, and 66.7±12.6%, respectively. With our previous CLASS with LCR method, the corresponding values were 78.0±14.7%, 16.4±16.9%, 18.2±15.0%, 3.8±2.3 mm, 73.8±13.4%, respectively. We demonstrated that the DL-CNN can overcome the strong boundary between two regions that have large difference in gray levels and provides a seamless mask to guide level set segmentation, which has been a problem for many gradient-based segmentation methods. Compared to our previous CLASS with LCR method that

required two user inputs to initialize the segmentation, DL-CNN with level sets achieved better segmentation performance while using a single user input. Compared to Haar-feature-based likelihood map, the DL-CNN-based likelihood map could guide the level sets to achieve better segmentation. The results demonstrate the feasibility of our new approach of using DL-CNN in combination with level sets for segmentation of the bladder. The results presented in this chapter have been published[2].

3.2 Introduction

As there are still challenges to segmenting the bladders in CTU, we have studied new techniques for improved performance and reduced requisite user input. Bladders may be filled with excreted intravenous (IV) contrast material that partially or fully opacifies the bladder. The boundaries between the bladder wall and the surrounding soft tissue have very low contrast such that they are often difficult to delineate. In addition, imaged bladders may have a variety of shapes and sizes. To address these challenges, Hadjiiski et al.[25, 26] developed preliminary bladder segmentation methods for CTU using active contour with 15 patients and level sets with 70 patients. Hadjiiski et al.[27] also developed a segmentation package specifically designed based on the characteristics of the bladder in CTU images, referred to as Conjoint Level set Analysis and Segmentation System (CLASS), that segments the contrast-enhanced and non-contrast regions of the bladder separately, using two input bounding boxes, and then joins the regions together. They qualitatively evaluated the segmentation performance of 81 bladders and performed quantitative evaluation of 30 bladders comparing the computer-segmented contours to hand-segmented reference contours and obtained promising results. The CLASS method was further developed by Cha et al.[1] as described in Chapter III to improve the segmentation accuracy. Model-guided refinement was used to propagate the contours of the contrast-enhanced region if the level set propagation stopped prematurely due to substantial non-uniformity of the contrast, as described in Chapter II, section 2.3.2.1. An energy-driven wavefront propagation that used changes in energies, smoothness criteria of the contour, and a stop criterion determined by the previous slice contour was designed to further propagate the conjoint contours to the correct bladder boundary. The segmentation performance was evaluated using 81 training cases and 92 independent test cases.

Convolutional neural networks (CNN) have been used previously to classify patterns in medical images for use with computer-aided detection and specifically for microcalcification detection in mammograms[31-38]. In these applications, the training sets were typically small, generally using less than 500 samples. As computational power grows, CNNs with very complex architectures that require training with massive data become practical. The deep-learning CNN (DL-CNN) using graphics processing units (GPU) has been shown to be able to classify natural images using a large training set. Krizhevski et al.[39, 40] have shown that by using DL-CNN, they are able to achieve relatively low error rates and good classification accuracy on the ImageNet ILSVRC-2010 and ILSVRC-2012 data sets[41], and the CIFAR-10 data set[42].

In this study, we explored the application of the DL-CNN to bladder segmentation. The DL-CNN was trained to recognize the patterns inside and outside the bladder and generated a bladder likelihood map to guide the level set segmentation. For comparison, we also generated a bladder likelihood map by using Haar features[43, 44], to differentiate bladder region from the surrounding structures as classified by a random forest classifier. To evaluate the effectiveness of the template-based approach, their performances were compared to our CLASS with local contour refinement (LCR) method described in Chapter 2.

The chapter is organized as follows. First, the data set used in the study is described. Second, the method of generating the bladder likelihood map using DL-CNN is presented. Third, the level set segmentation method using the likelihood map is described. Fourth, the method of generating the likelihood map using Haar features is designed as a comparison to the DL-CNN approach. Finally, the segmentation results are presented and discussed.

3.3 Materials and Methods

A DL-CNN was trained to distinguish between regions of interest (ROI) that are inside and outside of the bladder. The DL-CNN outputs the likelihood that an input ROI is inside the bladder that is used to form the bladder likelihood map. The map is used to generate the initial contour for level-set-based bladder segmentation. A flowchart of the segmentation method is shown in Figure 3.1.

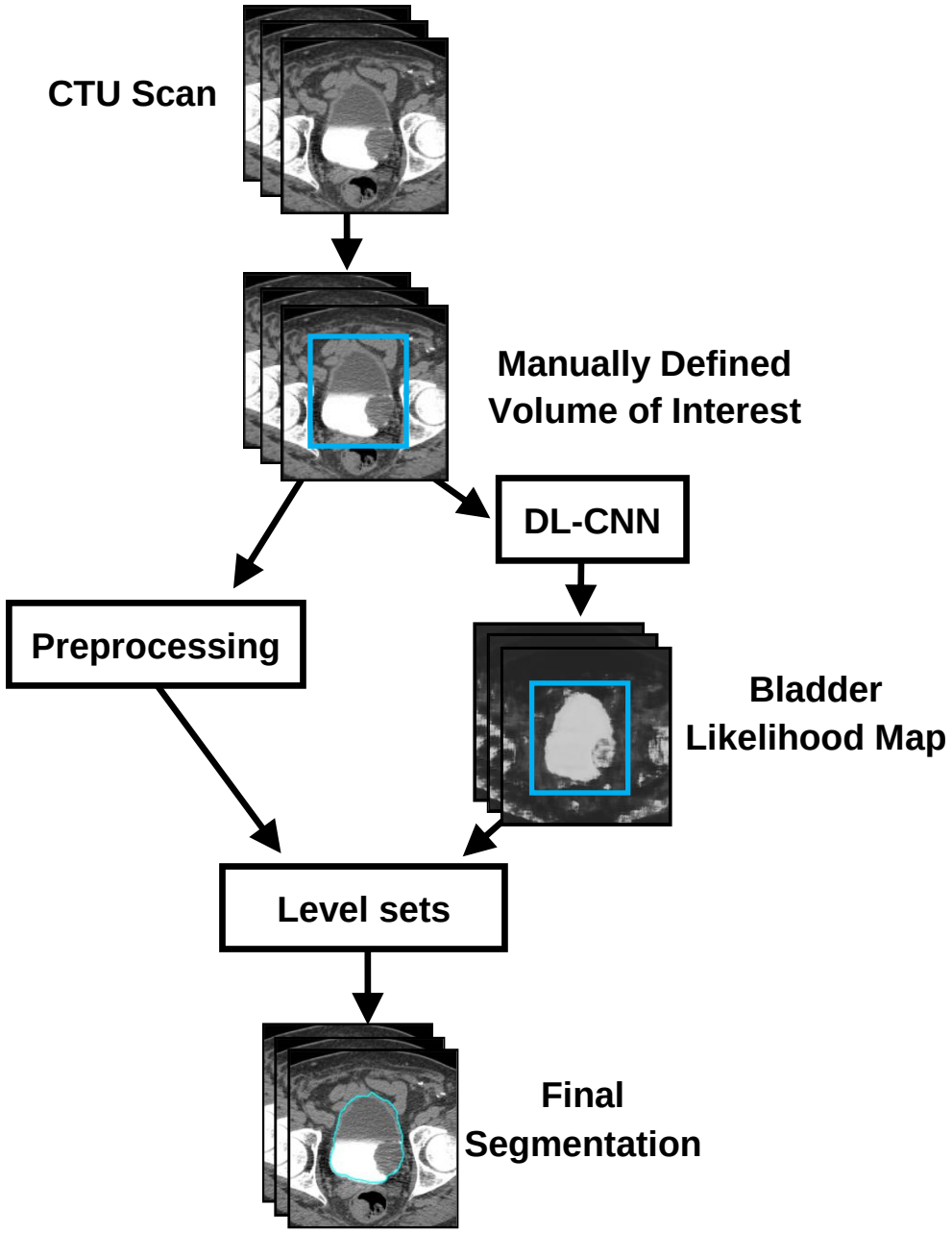


Figure 3.1: Flowchart of the template-based segmentation method.

3.3.1 Data set

In this study, a data set of 173 patients undergoing CTU who subsequently underwent cystoscopy and biopsy was utilized. The cases were collected retrospectively from the Abdominal Imaging Division of the Department of Radiology at the University of Michigan with approval of the Institutional Review Board. We designated 81 of these cases as the training set, and the other 92 cases as the test set. The cases were assigned to the training or the test sets by

balancing the difficulty of the cases between the two sets by researcher who was trained by an experienced abdominal radiologist. The data set used in this study was described in Chapter II, section 2.3.4.

Three-dimensional (3D) hand-segmented contours for all 173 cases were obtained as reference standard (RS1) in this study. An experienced radiologist provided manual outlines on the CT slices for all cases using a graphical user interface. The bladder was outlined on every 2D CT slice on which the bladder was visible, resulting in a 3D surface contour. There were a total of 16,197 slices for the 173 bladders. A subset of cases that contained lesions (41 training set cases, 50 test set cases, a total of 8,420 slices) was outlined by a different reader experienced in bladder segmentation to provide a second reference standard (RS2). The two sets of independent manual outlines allowed us to study the inter-observer variability and to evaluate the difference in the computer segmentation performance relative to the two sets of hand outlines.

3.3.2 Bladder likelihood map generation using deep-learning convolutional neural network (DL-CNN)

We applied the DL-CNN developed by Krizhevski et al. called cuda-convnet[39, 40] to the classification of ROIs on 2D slices as being inside or outside of the bladder. The neural network is trained using labeled ROIs of the same size extracted from the CTU slices in the training cases. Each of the extracted ROIs is input into the DL-CNN that outputs the likelihood of the ROI to be inside the bladder. To use the trained DL-CNN to generate a bladder likelihood map, it is applied to ROIs centered at each pixel on an axial slice in a CTU scan that contains the bladder and the likelihood value for the ROI is assigned to the center pixel. The resulting output over all pixels on the slice forms a bladder likelihood map, and the 2D maps over the consecutive CT slices constitutes a 3D likelihood map.

3.3.2.1 DL-CNN components

Components of the DL-CNN are briefly described in the following. More information about this network can be found in literature[39,40].

Neurons: A DL-CNN neuron consists of two functional parts: (1) summation of the weighted inputs to the neuron and (2) application of an activation function to the sum. The

activation function used in this DL-CNN is a non-saturation nonlinear function, defined by the following equation:

$$f(x) = \max(0, x) \tag{3.1}$$

The output of a neuron generally is obtained by a sigmoid activation function; however, it was shown that networks trained with gradient descent can converge much faster when neurons with the activation function in Equation (3.1) are used, which were named Rectified Linear Units, following Nair et al.[40, 45].

Convolution layer: In the convolution layer, the input ROI is convolved with the convolution kernels. The resulting values are collected into corresponding kernel maps in the convolution layer (Figure 3.2). The kernel maps are transformed by the activation function given by Equation (3.1), to give the output signal.

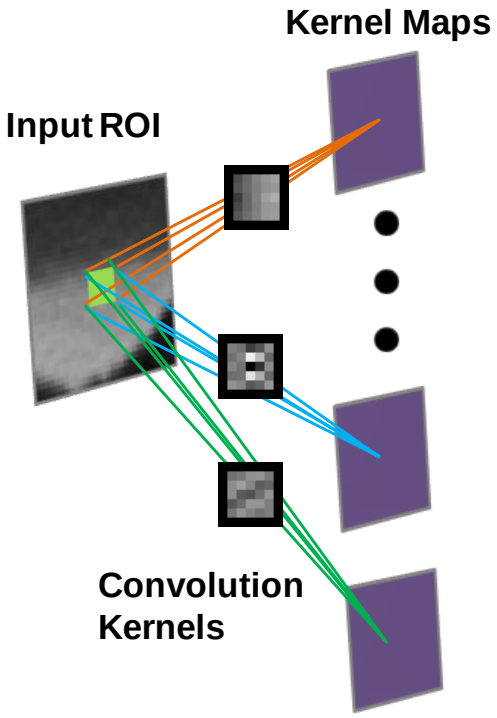


Figure 3.2: Diagram of the convolution layer. An input ROI is convolved with multiple convolution kernels, and the resulting values are collected into corresponding neurons in the kernel maps.

Pooling layer: The pooling layers summarize the outputs of neighboring groups of neurons within the same kernel map. We compared two commonly used overlapping pooling layers for our application; one used the maximum values and the other used average values within 3 x 3 groups of pixels centered at the pooling unit, with the distance between pooling set

to two pixels. Previous study had found that using overlapping pooling was less prone to overtraining[40].

Local Response Normalization layer: Using local normalization scheme aids in the generalization of the training. The activity of a neuron in the layer pervious to the normalization layer was normalized using the following equation[40]:

$$b_{x,y}^{i} = \frac{a_{x,y}^{i}}{\left(1 + \frac{\tau}{N}\sum_{j=\max\left(0,i-\frac{N}{2}\right)}^{\min\left(n,i+\frac{N}{2}\right)}\left(a_{x,y}^{j}\right)^{2}\right)^{\varepsilon}} \tag{3.2}$$

where $b_{x,y}^{i}$ is the response-normalized neuron activity, $a_{x,y}^{i}$ is the neuron activity computed by applying the kernel i at the coordinates (x,y), n is the number of kernel maps of the previous layer, and N, τ, and ε are constants. For our implementation of the DL-CNN, we used $N = 9$, $\tau = 0.001$, and $\varepsilon = 0.75$, following the study by Krizhevsky et al.[40].

3.3.2.2 DL-CNN architecture

A block diagram of the network architecture used in this study is shown in Figure 3.3. The network consists of five main layers: two convolution layers, two locally-connected layers, and one fully-connected layer. The locally-connected layers perform the same operation as the convolution layer, except that instead of applying a single convolution kernel to every location of the input image to obtain a kernel map, different convolution kernels are applied at every location of the input image, and the resulting values are collected into the corresponding neurons within the corresponding kernel map. The fully-connected layer uses every kernel map element multiplied by a weight as input. All of the inputs are summed, and the activation function (Equation 3.1) is applied to generate output values.

The first convolution layer filters the input images with 64 kernels of size 5 x 5. The output of the layer is pooled and normalized using the pooling and local response normalization, and is input into the second convolution layer that filters the output with an additional 64 kernels of size 5 x 5. The first locally-connected layer takes as input the pooled and normalized output of the second convolution layer and filters it with 64 kernels of size 3 x 3. The second locally-connected layer has 32 kernels of size 3 x 3 connected to the normalized, pooled output of the first locally-connected layer. The fully-connected layer outputs two values. The outputs from the fully-connected layer are input into a Softmax layer that computes the following function:

$$f(x_i) = \frac{e^{x_i}}{\sum_j e^{x_j}} \tag{3.3}$$

where x_i is each input value to the layer. The output of this layer ranges from 0 to 1, which can be interpreted as the likelihood of the input ROI being classified into the one of the given categories. The negative log likelihood was used for the loss function. The network structure including the number of kernels and kernel sizes, and parameters were determined using the training set. For example, the kernel size combination of 5 x 5 and 3 x 3 for the convolution layers and the locally-connected layers, respectively, performed best on the training set compared to other kernel size combinations.

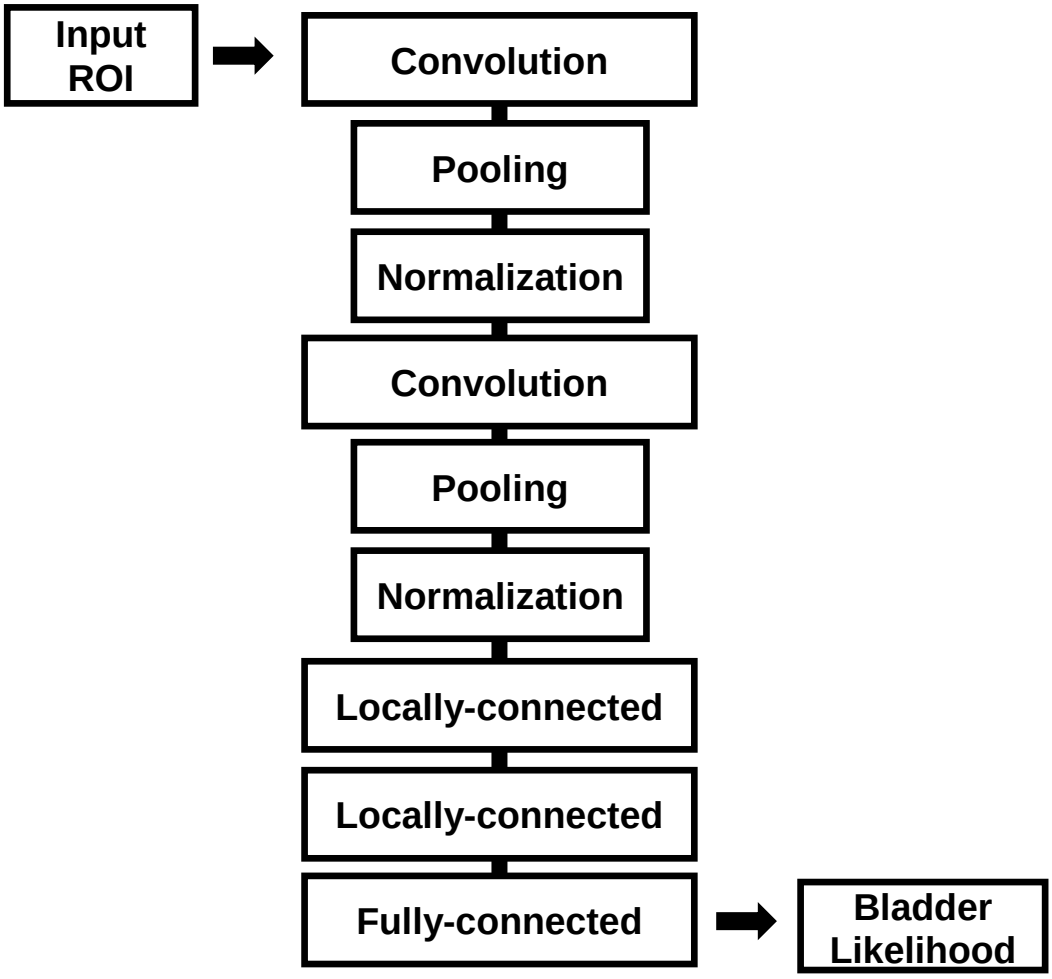


Figure 3.3: Block diagram of the DL-CNN architecture used in this study.

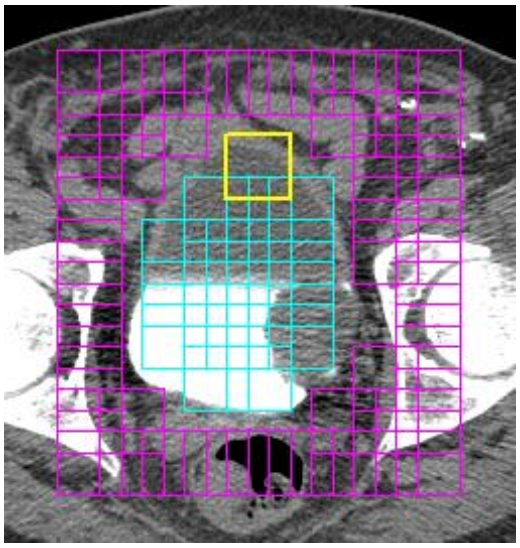
3.3.2.3 DL-CNN training

The DL-CNN was trained using the cases in the training set. A cropped CTU slice of a bladder case is shown in Figure 3.4(a). For each axial slice of the cases in the training set, ROIs

of N x N-pixels inside and outside the bladder were extracted using hand-outlines provided by an experienced radiologist (Figure 3.4(b)). Three ROI sizes, N = 16, 32, 64 were studied but the size of 32 x 32 pixels was used in the following discussion. Each bladder ROI was labeled as being inside or outside of the bladder as follows. If over 90% of an ROI was within the hand-outlined bladder, the ROI was labeled as being inside the bladder. A value of 90% was chosen to ensure a sufficient number of ROIs was identified as being inside the bladder. If less than 5% of an ROI was within the hand-outlined bladder, the ROI was labeled as being outside the bladder to avoid most of the bladder and the bladder wall while including the background regions that surround the bladder. ROIs not labeled as being inside or outside of the bladder were excluded. Figure 3.4(c) shows examples of ROIs that were extracted from a slice.

(a)

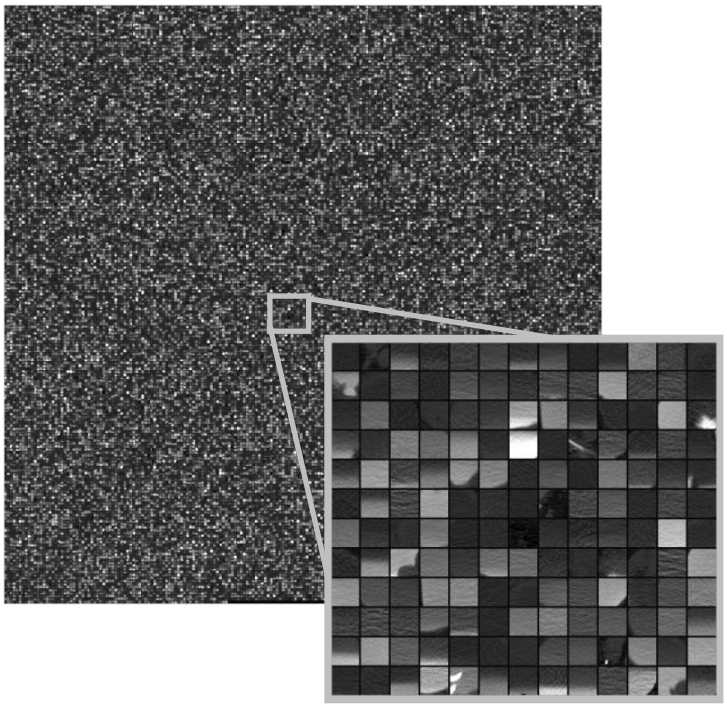
(b)



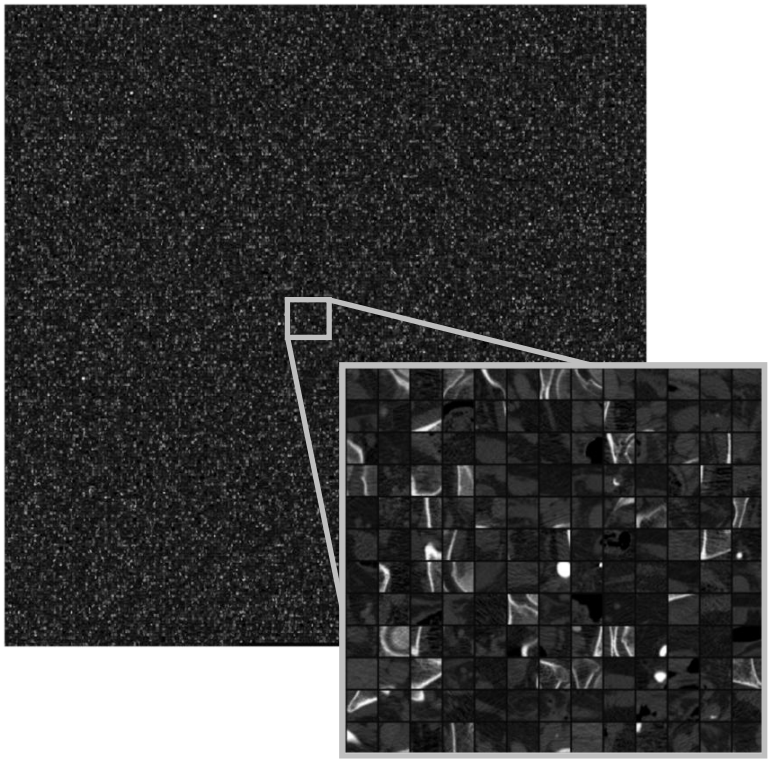
(c)

Figure 3.4: Images of a CTU slice from a training case. (a) Cropped CTU slice centered at the bladder. (b) The CTU slice shown with radiologist's hand outline of the bladder. (c) Example of ROIs that were extracted from the CTU slice to train the DL-CNN. The yellow ROI at the top of the bladder shows the size of a 32x32-pixel ROI. The ROIs are partially overlapping. The pink ROIs are ones marked as outside of the bladder. The light blue ROIs are ones marked as inside of the bladder.

Approximately 160,000 ROIs were generated from the cases in the training set after balancing the number of ROIs that were inside and outside of the bladders. Figure 3.5(a) and 3.5(b) show examples of the ROIs inside and outside the bladder, respectively, used to train the DL-CNN.



(a)



(b)

Figure 3.5: Images of the 160,000 ROIs used to train the DL-CNN using the cases in the training set. Each ROI is 32x32 pixels. (a) ROIs that are labeled as inside the bladders. (b) ROIs that are labeled as outside the bladders. A small subset of the ROIs in each class is zoomed in to illustrate the content of typical ROIs.

The neural network was trained for 1500 iterations, but the DL-CNN trained for 1000 iterations was selected to generate the bladder likelihood maps. We observed that a network trained with up to 1000 iterations had similar training classification error rates as a network trained with up to 1500 iterations. Classification error rate is defined as the ratio of the number of incorrectly identified ROIs to the total number of ROIs. Figure 3.6 shows the classification error rate of the DL-CNN training for the entire training set as the number of iterations increased. In addition, we observed that bladder likelihood maps generated using DL-CNN trained for 1000 iterations were better or comparable to maps generated using network trained for 1500 iterations for representative cases from a range of difficulties in the training set, thus 1000 iterations was used to generate the likelihood maps. Training the network using 160,000 ROIs and 1000 iterations took approximately 5.5 hours using a Tesla C2075 GPU.

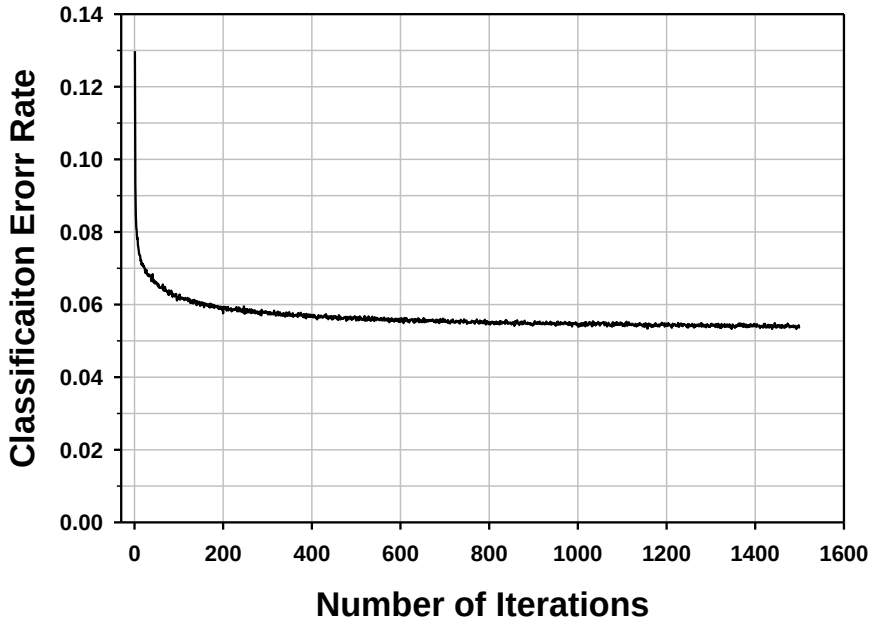


Figure 3.6: Plot of the classification error rate of DL-CNN training for the entire training set as the number of iterations increase. The error rates at iterations 1000 and 1500 were very similar. The training results from iteration 1000 were used to generate the bladder likelihood maps.

3.3.2.4 Bladder likelihood map generation with DL-CNN

For every axial slice in a CTU scan that contains the bladder, a bladder likelihood map was generated. Our current segmentation system uses a single box, or volume of interest (VOI) that approximately encloses the bladder as input. The bladder likelihood map is therefore generated within this VOI. The trained DL-CNN is applied to each voxel within the VOI. At each voxel, a 32 x 32-pixel ROI on the axial slice is extracted and input to the DL-CNN that outputs the likelihood that the input ROI is inside the bladder. The likelihood score for the ROI is assigned to the center pixel of the ROI. The collection of voxel-wise likelihood scores forms a bladder likelihood map. Figure 3.7 shows the bladder likelihood map of the CTU slice shown in Figure 3.4.

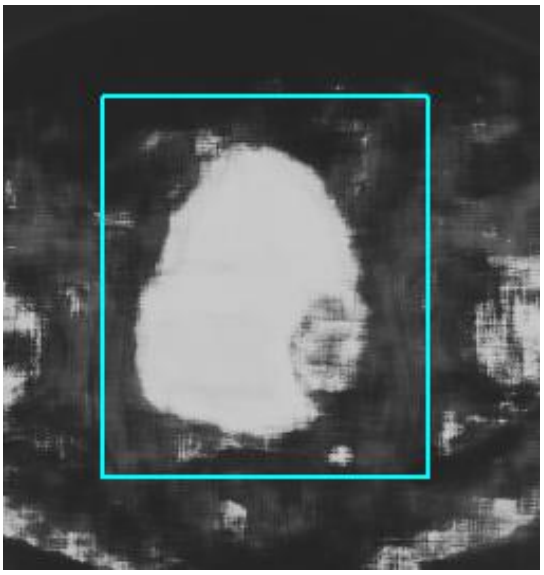


Figure 3.7: Bladder likelihood map of the CTU slice shown in Figure 3.4. High intensity represents high likelihood of the voxel being inside the bladder. In this example, for demonstration purposes, the bladder likelihood map was generated for an area larger than the VOI. The VOI is shown by the box around the bladder.

3.3.3 Bladder segmentation using DL-CNN bladder likelihood map

We have developed a software package that uses the DL-CNN bladder likelihood map and level sets to segment the bladder from the surrounding tissue. The system was initialized by the same VOI that encloses the bladder within which the bladder likelihood map was generated. The system consisted of four stages: (1) preprocessing (2) initial segmentation, (3) 3D level set segmentation, and (4) 2D level set segmentation.

In the first stage, preprocessing techniques were applied in 3D to the VOI. Smoothing, anisotropic diffusion, gradient filters and the rank transform of the gradient magnitude were applied to the slices within the VOI to obtain a set of gradient magnitude images and a set of gradient vector images that were used during level set propagation in the third stage.

In the second stage, the initial segmentation surface was generated using the DL-CNN bladder likelihood maps. First, a binary bladder mask, DL_{Mask}, was generated by applying the following criterion to every pixel on all slices of the bladder likelihood map:

$$DL_{Mask}(x,y) = \begin{cases} 1, & DL_{Score}(x,y) \geq \theta \\ 0, & DL_{Score}(x,y) < \theta \end{cases} \tag{3.4}$$

where $DL_{Mask}(x,y)$ is the pixel value on the bladder mask at the coordinates (x,y), $DL_{Score}(x,y)$ is the bladder likelihood score at the coordinates (x,y), and θ is the threshold

imposed on the bladder likelihood score. The value for θ was determined by histogram analysis. A histogram of the DL-CNN likelihood score for the pixels inside and outside of the bladder within the VOIs in the training cases was generated (Figure 3.8). We observed that the likelihood score of 0.85 provided a good separation of the two classes (e.g. inside the bladder and outside the bladder), with a large number of pixels correctly identified as being inside the bladder. Thresholding the likelihood maps at the score of 0.85 gave the best contour that did not leak to the outside of the bladder while closely approaching the hand segmentation for cases in the training set. For these reasons, 0.85 was chosen as the threshold.

Second, an ellipsoid whose minor and major axes are 1.5 of the width and height of the VOI, respectively, centered at the centroid of the bladder mask, is placed on the DL_{Mask}. The intersection of the bladder mask and the ellipsoid is labeled as the object region. The ellipsoid is used to prevent the object region from leaking into the organs above the bladder and the pelvic bones, as these structures can also obtain high likelihood scores from the DL-CNN. Finally, a morphological dilation filter with a spherical structuring element of 2 voxels in radius, 3D flood fill algorithm, and a morphological erosion filter with a spherical structuring element of 2 voxels in radius are applied to the object region to connect neighboring components and extract an initial segmentation surface, $\phi_0(x)$.

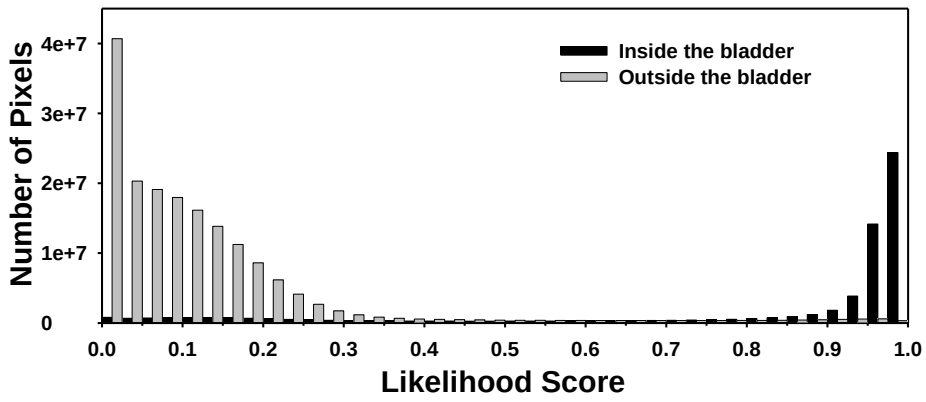


Figure 3.8: Histogram of the DL-CNN likelihood score for the pixels in the training set. Higher likelihood score indicates that the pixel is more likely to be inside the bladder.

In the third stage, the initial segmentation surface is propagated towards the bladder boundary using cascading level sets. Our chosen level set implementation evolves according to the equation:

$$\begin{cases} \frac{\partial}{\partial t}\Psi(x) = -\alpha A(x)\nabla\Psi(x) - \beta P(x)|\nabla\Psi(x)| + \gamma\kappa(x)|\nabla\Psi(x)| \\ \Psi(x, n=0) = \phi_0(x) \end{cases} \quad (3.5)$$

where $\alpha, \beta,$ and γ are the coefficients for the advection, propagation, and curvature terms, respectively, $A(x)$ is a vector field image (assigning a vector to each voxel in the image) that drives the contour to move towards regions of high gradient, $P(x)$ is a scalar speed term between 0 and 1 causing the contour to expand at the local rate, and $\kappa(x) = div\left(\frac{\nabla\Psi(x)}{|\nabla\Psi(x)|}\right)$ is the mean curvature of the level set at point x. The symbol ∇ denotes the gradient operator and div is the divergence operator[28]. $\phi_0(x)$ is the initial segmentation surface, and n is the number of iterations.

Three 3D level sets with predefined sets of parameters are applied in series to the initial segmentation surface. The corresponding parameters of the 3 level sets are presented in Table 3.1.

Table 3.1: Parameters for the level sets.

Level set:	α	β	γ	n
First	1	2	1	10
Second	1	0.6	q	150
Third	0	1.0	0	10
2D slices	4.0	0.2	0.5	100

The first 3D level set slightly expands and smoothes the initial contour. The second 3D level set brings the contour towards the sharp edges, but also expands it slightly in regions of low gradient. The parameter "q" in Table 3.1 is defined to be a linear function $\sigma M + \phi$ of the 2D diagonal distance M of the VOI box in millimeters (mm), where $\sigma = 0.06, \phi = -0.11$ as shown previously[28]. The third 3D level set further draws the contour towards sharp edges. As a final step, a 2D level set is applied to every slice of the segmented object to refine the 3D contours using the 3D level set contours as the initial contour. Further details on the level sets used can be found in the literature[28]. An example of the segmented bladder for CTU slice shown in Figure 3.4 using the DL-CNN bladder likelihood map with level sets (DL-CNN with level sets) is shown in Figure 3.9.

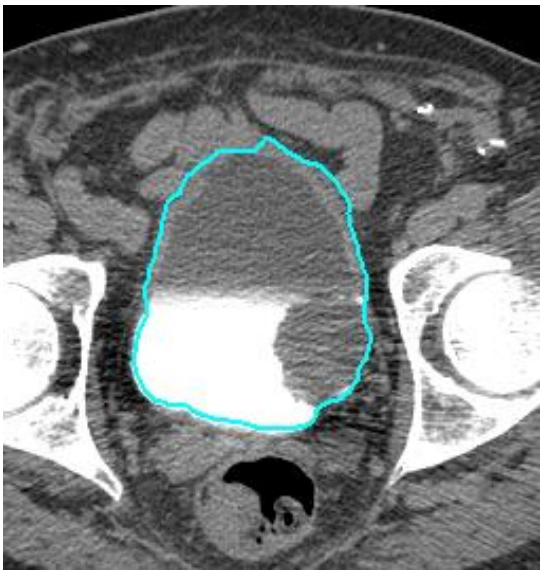


Figure 3.9: Bladder segmentation of the CTU slice shown in Figure 3.4 using the DL-CNN bladder likelihood map with level sets.

3.3.4 Bladder likelihood map generation using Haar features and random forest classifier

To compare the performance of DL-CNN for bladder likelihood map generation, the maps were also generated using Haar features and random forest classifier. Fifty-nine Haar features were extracted from the 32 x 32-pixel ROIs used to train the DL-CNN. A large number of Haar features can be extracted from a 32 x 32-pixel ROI. Using every possible Haar feature would be difficult due to the enormous number of features that would be generated; therefore, we considered the representative shapes for the bladder boundaries, and after experimenting on the training cases, we selected 59 different Haar features to generate the bladder likelihood maps, which are described in Table 3.2.

Table 3.2: Number of features extracted for different Haar filter sizes and filter types as described by Viola et al.[43] and Lienhart et al.[43].

	8 x 8-pixel	16 x 16-pixel	16 x 18-pixel	18 x 16-pixel	16 x 32-pixel	32 x 16-pixel	32 x 32-pixel
Edge Features	10	8	0	0	2	2	2
Line Features	10	0	4	4	0	0	2
Four-Rectangle Features*	9	5	0	0	0	0	1

*A single filter of this Four-Rectangle feature filter consists of 4 smaller, equal-sized rectangles arranged in a checkerboard pattern.

The extracted features were used to train a random forest classifier that combined the features together to generate a score that corresponds to an ROI's likelihood of being inside the bladder. The random forest classifier with 100 trees was trained using the same set of 160,000 training ROIs as described above for training the DL-CNN. The bladder likelihood map was generated by extracting the 59 Haar features values from each ROI. The feature values were input into the trained random forest classifier that output the likelihood that the input ROI was inside the bladder. The likelihood score for the ROI was assigned to the center pixel of the ROI. The collection of likelihood scores over the voxels in the VOI formed the bladder likelihood map.

The distribution of the Haar-feature-based bladder likelihood scores was different than that from the DL-CNN scores, thus a different threshold of 0.56 was chosen experimentally using the training cases and used to generate the binary bladder mask for initialization of the level sets. After the Haar-feature-based bladder binary mask was generated, the bladder segmentation process was identical to that described in Section 3.3.3.

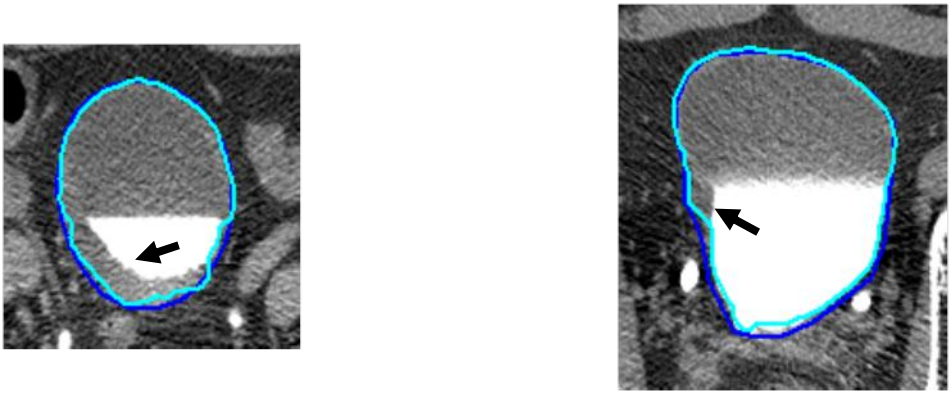
3.3.5 Evaluation Methods

Segmentation performance was evaluated by comparing the automatic segmentation results to the 3D hand-segmented contours. The volume intersection ratio, the volume error, the average minimum distance, and the Jaccard index[29] between the hand-segmented contours and computer-segmented contours were calculated. The performance metrics are described in Chapter II, section 2.3.5.

3.4 Results

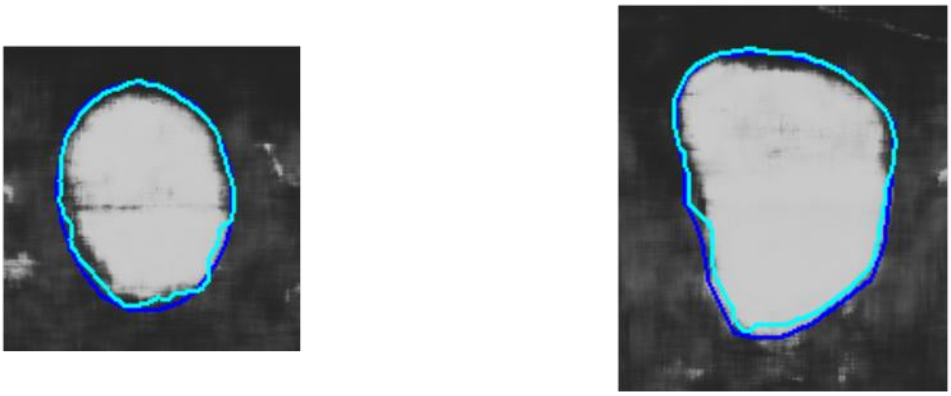
3.4.1 Segmentation performance using DL-CNN bladder likelihood map with level sets

The trained DL-CNN obtained a classification error rate of 0.054 for the training set. The error rate for the classification of the ROIs was not measured, as the classification of the ROI is not the final goal of this study. Examples of the segmentation from cases in the test set are shown in Figure 3.10. Figure 3.11 shows the bladder likelihood maps used to generate the bladder boundaries in Figure 3.10. The segmentation performance measures averaged over the cases in the training and test sets are presented in Table 3.3.



(a) (b)

Figure 3.10: Examples of bladder segmentations using DL-CNN with level sets for two cases in the test set. (a) Malignant bladder wall thickening was fully enclosed within the segmentation. (b) The bladder segmentation enclosed the lesion present in the bladder; however, the bottom of the contrast-enhanced region was slightly under-segmented. Arrows point to the wall thickening and lesion in (a) and (b), respectively. The light blue contour represents segmentation result from DL-CNN with level sets. The dark blue contour represents the radiologist's hand outline.



(a) (b)

Figure 3.11: Bladder likelihood maps and the corresponding bladder segmentation for cases shown in Figure 3.10. (a) Refining the initial contour generated from the likelihood map by level sets results in accurate bladder segmentation. (b) Regions within the non-contrast region of the bladder had low likelihood of being within the bladder. The level sets propagated the initial contour to enclose the lesion and the non-contrast region. The light blue contour represents segmentation result from DL-CNN with level sets. The dark blue contour represents the radiologist's hand outline.

Table 3.3: Segmentation evaluation results using DL-CNN-based likelihood map with level sets averaged over the 81 training cases and 92 test cases.

	Volume intersection ratio R^{3D}	Volume error E^{3D}	Absolute volume error $\lvert E^{3D} \rvert$	Average minimum distance *AVDIST*	Jaccard index $JACCARD^{3D}$
Training Set	87.2±6.1%	6.0±9.1%	8.8±6.4%	3.0±1.2 mm	81.9±7.6%
Test Set	81.9±12.1%	10.2±16.2%	14.0±13.0%	3.6±2.0 mm	76.2±11.8%

The histograms for volume intersection ratio, volume error, average distance, and the Jaccard index for both the training set and the test set are shown in Figures 3.12, 3.13, 3.14, and 3.15, respectively.

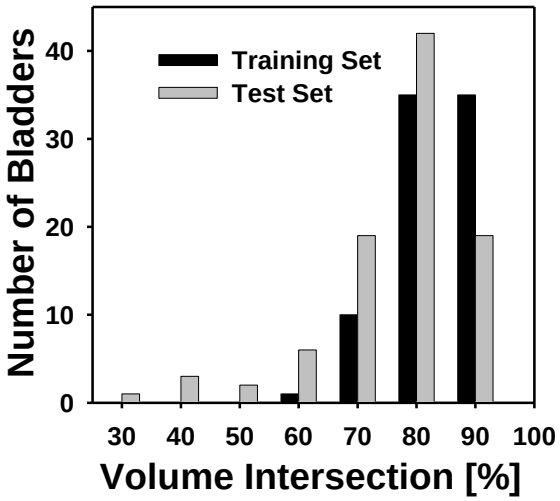


Figure 3.12: Histogram of the percent volume intersection ratio for the training and test sets. The mean volume intersection was 87.2% for the 81 training cases, and 81.9% for the 92 test cases.

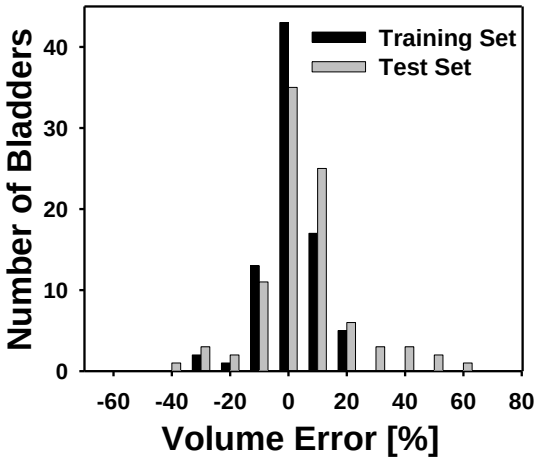


Figure 3.13: Histogram of the volume error for the training and test sets. The mean volume error was 6.0% for the 81 training cases, and 10.2% for the 92 test cases.

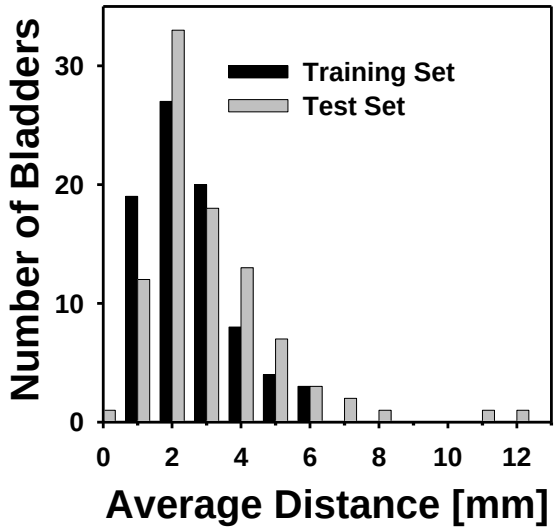


Figure 3.14: Histogram of the average distance for the training and test sets. The mean average distance was 3.0 mm for the 81 training cases, and 3.6 mm for the 92 test cases.

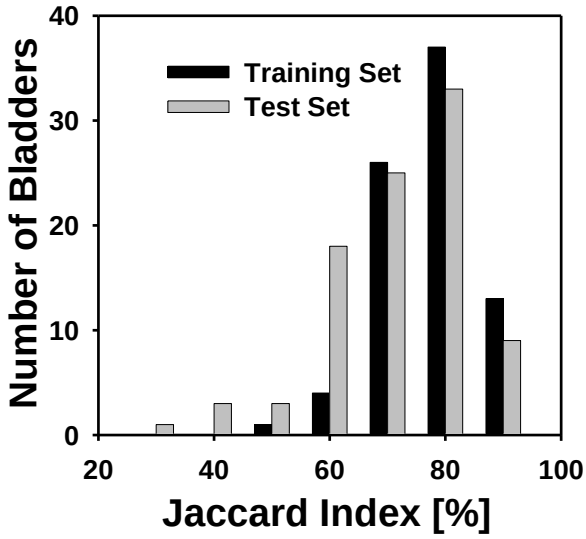


Figure 3.15: Histogram of the Jaccard index for the training and test sets. The mean Jaccard index was 81.9% for the 81 training cases, and 76.2% for the 92 test cases.

Of the 81 cases in the training set, 70 bladders (86.4%) had a volume intersection ratio greater than 80% (Figure 3.12). There were 79 bladders (97.5%) whose absolute volume error for the training set was less than 20% (Figure 3.13). Forty-six bladders (56.8%) in the training

set had an average distance less than 3 mm (Figure 3.14), and 50 bladders (61.7%) had Jaccard indices of over 80% (Figure 3.15).

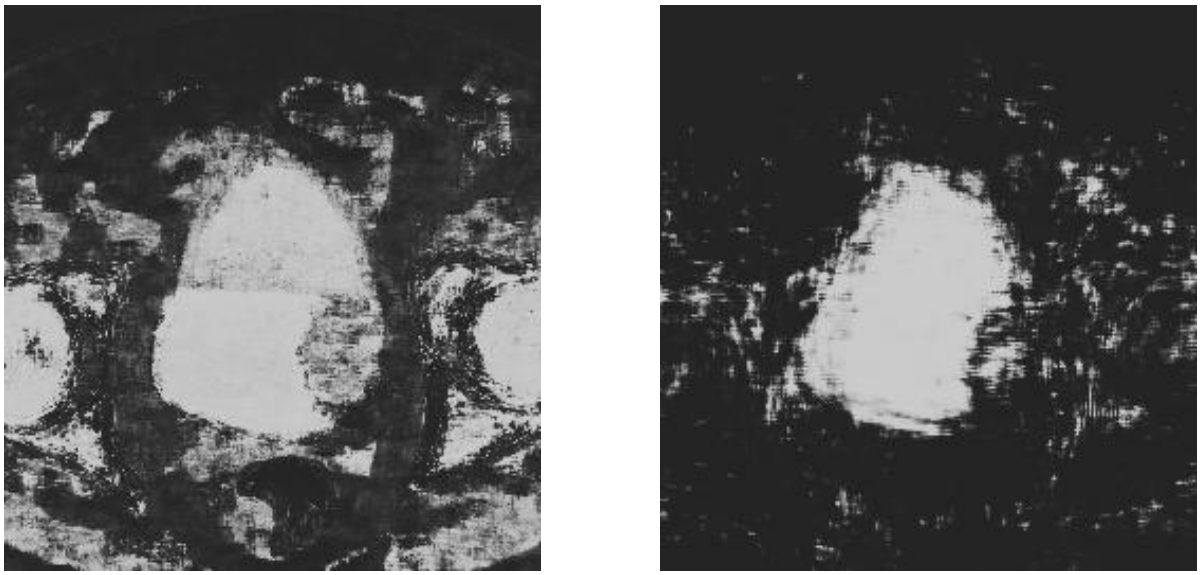
Of the 92 test cases, 61 bladders (66.3%) had a volume intersection ratio greater than 80% (Figure 3.12). There were 73 bladders (79.3%) whose absolute volume error for the test set was less than 20% (Figure 3.13). Forty-six bladders (50.0%) in the test set had an average distance less than 3 mm (Figure 3.14), and 42 bladders (45.7%) had Jaccard indices of over 80% (Figure 3.15).

3.4.2 Dependence of segmentation performance on input ROI size and DL-CNN pooling

Table 3.4 summarizes the segmentation performance on the test cases for the conditions: (1) the maximum pooling layers were replaced by average pooling layers while keeping the input ROI size at 32 x 32 pixels and other parameters the same as those in Section 3.1, (2) the input ROI size was changed to 16 x 16 pixels and 64 x 64 pixels while all other parameters the same as those in Section 3.1. The training set results showed similar trends. Figure 3.16 shows examples of the bladder likelihood map for 16x16-pixel ROI and the 64x64-pixel ROI for the CTU slice shown in Figure 3.4.

Table 3.4: Segmentation evaluation results for DL-CNN with level sets using average pooling with 32x32-pixel ROI, and maximum pooling using 16x16-pixel ROI, and 64x64-pixel ROI averaged over the 92 test cases. Training set results showed similar trends.

	Volume intersection ratio	Volume error	Absolute volume error	Average minimum distance	Jaccard index
	R^{3D}	E^{3D}	$\lvert E^{3D} \rvert$	*AVDIST*	$JACCARD^{3D}$
Average Pooling 32x32-pixel ROI	81.0±12.1%	5.3±21.5%	16.2±14.9%	4.5±2.9 mm	72.1±13.3%
Maximum Pooling 16x16-pixel ROI	79.2±14.2%	11.0±20.1%	17.4±14.8%	4.4±2.5 mm	72.6±14.0%
Maximum Pooling 64x64-pizel ROI	67.1±12.7%	24.9±19.8%	27.9±15.1%	6.4±2.8 mm	62.8±13.1%



(a) (b)

Figure 3.16: Bladder likelihood map of the CTU slice shown in Figure 3.4 using different ROI sizes. (a) Likelihood map generated using 16x16-pixel ROIs. (b) Likelihood map generated using 64x64-pixel ROIs.

3.4.3 Variability of reference standards

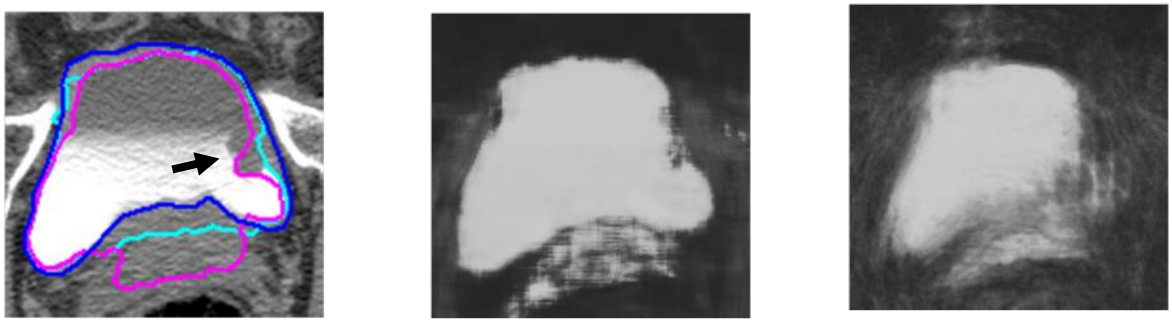
Table 3.5 shows the segmentation results using DL-CNN likelihood map with level sets compared against the two reference standards, as well as the results comparing the two hand-outlines with each other.

Table 3.5: Segmentation evaluation results in a subset of test cases with lesions (41 training cases, 50 test cases) between hand-segmented reference standards (RS1, RS2) by two different readers and DL-CNN with level sets. Segmentation evaluation of RS2 using RS1 as the reference is included to show inter-observer variations.

		Volume intersection ratio	Volume error	Absolute volume error	Average minimum distance	Jaccard index
		R^{3D}	E^{3D}	$\|E^{3D}\|$	*AVDIST*	$JACCARD^{3D}$
DL-CNN vs RS1	Training Set	85.9±6.6%	6.9±9.6%	9.3±7.1%	3.2±1.3 mm	80.4±8.4%
	Test Set	81.2±11.5%	12.5±13.5%	13.4±12.5%	3.6±1.9 mm	76.4±11.5%
DL-CNN vs RS2	Training Set	84.3±7.1%	9.7±10.0%	11.4±7.9%	3.4±1.3 mm	79.8±8.2%
	Test Set	78.2±10.9%	17.5±12.0%	17.7±11.6%	4.0±2.1 mm	75.1±11.0%
RS2 vs RS1	Training Set	96.2±2.8%	-3.0±4.8%	4.2±3.8%	1.4±0.5 mm	90.2±4.8%
	Test Set	95.0±8.1%	-6.2±15.3%	10.3±12.8%	1.7±1.0 mm	86.1±9.5%

3.4.4 Comparison of segmentation performance using DL-CNN-based and Haar-feature-based bladder likelihood maps

Table 3.6 summarizes the segmentation performance measures using the Haar-feature-based likelihood map to guide the level sets, averaged over the cases in the training and test sets. An example comparing the segmented bladder using the Haar-feature-based likelihood map with that using the DL-CNN-based likelihood map is shown in Figure 3.17.



(a) (b) (c)

Figure 3.17: Comparison of bladder segmentations using DL-CNN-based likelihood map and Haar-feature-based likelihood map. (a) DL-CNN-based segmentation (light blue contour) encloses the bladder lesion within the segmentation, while the Haar-feature-based segmentation (pink contour) does not fully enclose the lesion and leaks into the prostate. The arrow points to the lesion. The dark blue contour represents the radiologist's hand outline. (b) Bladder likelihood map generated using DL-CNN. (c) Bladder likelihood map generated using Haar features and random forest classifier.

Table 3.6: Segmentation evaluation results using Haar-feature-based likelihood map with level sets averaged over 81 training cases and 92 test cases.

	Volume intersection ratio	Volume error	Absolute volume error	Average minimum distance	Jaccard index
	R^{3D}	E^{3D}	$\|E^{3D}\|$	*AVDIST*	$JACCARD^{3D}$
Training Set	76.2±10.4%	15.5±15.0%	18.1±11.6%	5.2±1.7 mm	70.7±10.0%
Test Set	74.3±12.7%	13.0±22.3%	20.5±15.7%	5.7±2.6 mm	66.7±12.6%

Table 3.7 shows the initial segmentation surface ($\phi_0(x)$) generated from the DL-CNN-based and Haar feature-based bladder likelihood maps in comparison to the hand outlines (RS1). The results show the segmentation performance without the refinement by the level sets and the differences between the DL-CNN-based likelihood maps and the Haar-feature-based likelihood maps. We observer better segmentation performance when DL-CNN is used.

Table 3.7: Segmentation evaluation results using initial contours (no level sets) generated using bladder likelihood maps with DL-CNN and Haar Features averaged over the 92 test cases. Training cases showed similar trends.

	Volume intersection ratio	Volume error	Absolute volume error	Average minimum distance	Jaccard index
	R^{3D}	E^{3D}	$\|E^{3D}\|$	*AVDIST*	$JACCARD^{3D}$
DL-CNN	68.7±12.0%	27.3±13.7%	27.4±13.6%	5.7±2.2 mm	66.2±11.8%
Haar Features	59.8±12.1%	32.3±18.6%	34.0±15.2%	8.1±2.6 mm	55.6±11.4%

3.4.5 Comparison of segmentation performance using DL-CNN bladder likelihood map with level sets and CLASS with LCR

Segmentation results of several test cases for both CLASS with LCR and DL-CNN with level sets are shown in Figure 3.18. The segmentation performance measures for CLASS with LCR method are shown in Table 3.8. These results should be compared to those in Table 3.3 that were obtained with the DL-CNN with level sets method for the same training and test sets and the reference standards.

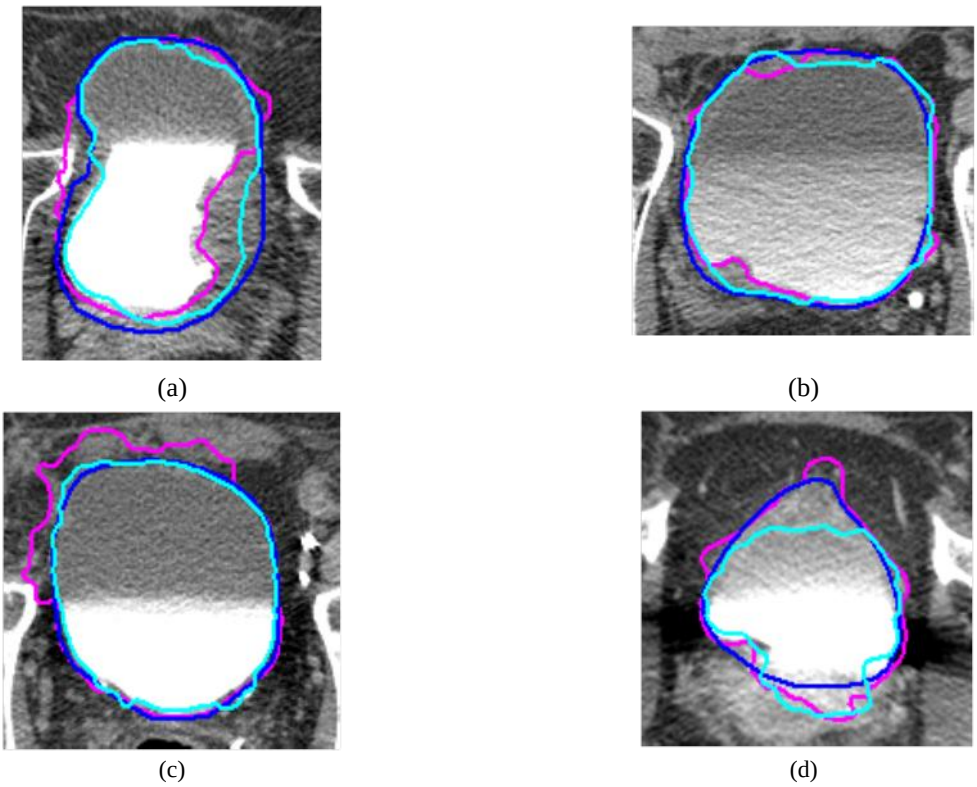


Figure 3.18: Comparison of bladder segmentation using DL-CNN with level sets and CLASS with LCR. The pink contour represents segmentation using CLASS with LCR. The dark blue contour represents the radiologist's hand outline. (a) DL-CNN slightly under-segments the upper region of the non-contrast region, but encloses more of the large, malignant lesion and does not leak towards the bones. (b) The two segmentation methods perform similarly, but DL-CNN with level sets encloses the lesion, whereas CLASS does not. (c) DL-CNN with level sets does not leak into the surrounding organs in the non-contrast region, unlike CLASS. (d) CLASS performs better than DL-CNN with level sets in the non-contrast enhanced region. Both methods over-segment the contrast-enhanced region. The light blue contour represents segmentation using DL-CNN with level sets.

Table 3.8: CLASS with LCR segmentation evaluation results averaged over the 81 training cases and 92 test cases.

	Volume intersection ratio R^{3D}	Volume error E^{3D}	Absolute volume error $\|E^{3D}\|$	Average minimum distance *AVDIST*	Jaccard index *JACCARD*^{3D}
Training Set	84.2±11.4%	8.2±17.4%	13.0±14.1%	3.5±1.9 mm	78.8±11.6%
Test Set	78.0±14.7%	16.5±16.8%	18.2±15.0%	3.8±2.3 mm	73.9±13.5%

3.5. Discussion

In this study, a new segmentation method that combines a likelihood map generated by DL-CNN with cascading level sets was developed and applied to a data set containing bladders in CTUs having a wide range of image quality. Most of the bladders were partially filled with excreted contrast material; however, some bladders were entirely filled with excreted contrast material and others did not contain any contrast-enhanced urine due to variation in timing for

image acquisition. The presence of the two distinct areas that have very different attenuation values: an area filled with contrast material and an area without contrast material, poses a challenge for segmentation that needs to go across the strong boundary between the two areas. To alleviate this problem, we previously required two manually input VOIs: one for the non-contrast region and the other for the contrast-enhanced region using our CLASS segmentation method[1, 27] and an LCR method was needed to refine and connect the two contours. However, by combining the DL-CNN bladder likelihood maps with the level set methods, we no longer needed the separate input user inputs for the two different regions. A major contribution of this work is that it demonstrates the DL-CNN can overcome the strong boundary between two regions that have large difference in gray levels and provides a seamless mask to guide level set segmentation. This has been a problem for many gradient-based segmentation methods. As a result, this new method requires only one user input bounding box for the entire bladder to start the segmentation procedure compared to the two user input bounding boxes for the previous method.

Compared to our CLASS with LCR method using the same data set, segmentation using DL-CNN performed better. All performance measures were improved using DL-CNN with level sets (Table 3.3) compared to CLASS with LCR (Table 3.8) for both the training and test sets. The differences in the volume intersection ratio, absolute volume error, average minimum distance, and the Jaccard index for the training set were statistically significant, with *p*-values of 0.01, 0.007, 0.01, and 0.002, respectively, by two-tailed paired t-test at an alpha level of 0.01 after the Bonferroni correction for the 5 comparisons. For the test set, the differences in the volume intersection ratio, volume error, and the absolute volume error were statistically significant with *p*-values of 0.004, 0.001, and 0.005 by two-tailed paired t-test at the alpha level of 0.01. For the training set, the percentages of cases obtaining improvements in the 5 performance measures ranged from 54% to 64%. For the test set, the improvements ranged from 54% to 67%. More importantly, DL-CNN with level sets better included lesions within its segmented region; 50 out of 59 (84.7%) lesions in the training set, and 64 out of 78 (82.1%) lesions in the test set were included better than or similar to the bladder segmented with CLASS with LCR. These improvements were obtained while reducing the number of user inputs (one box vs. two boxes).

DL-CNN with level sets generally enclosed more of the lesions within the segmented regions compared to CLASS with LCR (Figure 3.18(a, b)), which is important because further steps of the CAD system for lesion detection and characterization will be performed within the segmented bladder. The non-contrast enhanced region was segmented more accurately, without leaking into the adjacent organs using the DL-CNN (Figure 3.18(a, c)). However, there were cases that performed worse than our previous method (Figure 3.18(d)). These were caused by either the network giving low likelihood scores for portions of the bladder, causing under-segmentation, or the network giving relatively high likelihood scores for other organs, such as the bone, causing over-segmentation. Organs that were given relatively high bladder likelihood scores, such as the femoral heads, can be seen in the regions outside the VOI in Figure 3.7.

For a few test set cases, DL-CNN with level sets performed well below the average performance of the test data set. Some of these cases had poor image quality due to noise caused by the large patient size or the presence of hip prosthesis. Other large segmentation mistakes were due to the patient having advanced bladder cancer spreading into the neighboring organs and causing the segmentation to leak into those areas. In the future, we will improve our method to reduce the errors caused by these types of cases.

When average pooling was used instead of maximum pooling in the network structure, the segmentation performance measures deteriorated in general. The differences between the two methods were statistically significant for the volume error, average minimum distance, and the Jaccard index for the test set.

The graphics processing units (GPUs) used for DL-CNN are designed such that they perform efficiently when the inputs are multiples of 16. Therefore, we studied 16x16, 32x32, and 64x64-pixel ROI sizes as input to the network. Using the 16x16-pixel ROIs as input to the network resulted in bladder likelihood maps with finer details, such as lesion boundaries and the boundary between the non-contrast and the contrast-enhanced regions of the bladder. However, these maps did not lead to better segmentation than the likelihood maps obtained from 32x32-pixel ROIs likely due to the fine details hindering the generation of the initial contour for the entire bladder. On the other hand, the bladder likelihood maps obtained from 64x64-pixel ROIs contained fewer details from the structures surrounding the bladder. However, the shapes of the bladder might have lost too much details compared to those in the

likelihood maps generated using 32x32-pixel ROIs. It also had the tendency of misclassifying large lesions as outside of the bladder. As shown in Table 3.4, both the smaller 16x16-pixel ROI and the larger 64x64-pixel ROI were inferior to the 32x32-pixel ROI for generating the bladder likelihood maps to guide bladder segmentation.

The bladder segmentation using DL-CNN with level sets performed comparably regardless of which of the two hand outlines was used as the reference standard. The agreement between the computer and the hand outlines are slightly lower than the agreement between the two observers (approximately 10% for the volume intersection ratio and the Jaccard index), but the computer segmentation in this range of accuracy is acceptable and still useful for defining the search region for bladder lesion detection, as shown in our previous work on bladder lesion detection[3].

Comparing Table 3.3 and Table 3.6, it is seen that the bladder likelihood maps obtained from the Haar features and the random forest classifier were not as effective as those from the DL-CNN, resulting in lower bladder segmentation performance. The differences in all performance measures but the volume error for the test set were statistically significant.

The comparison of the initial segmentation surfaces generated from the bladder likelihood maps with the reference standards show that the DL-CNN-based maps are closer to the hand outlines than the Haar-feature-based maps (Table 3.7). The result also shows that segmentation using the DL-CNN alone cannot reach the high performance level achieved by DL-CNN with refinement by level sets. The DL-CNN bladder likelihood maps are generally under-segmenting the bladder, often catching the edge of the inner bladder wall for cases with circumferential bladder wall thickening while lowering the threshold for the DL-CNN bladder likelihood map would lead to leaking. Applying the level sets to the slightly under-segmented contours allows better control of the balance between under- and over-segmentation.

We chose the network structure size and level sets parameters by experimentation where each parameter was varied over a reasonable range, and the best parameter within the studied range was chosen based on the evaluation of the training set results. Our sensitivity analysis of the level sets can be found in the literature[28]. We have performed a sensitivity analysis of the network structure size. The number of kernels within the first two convolution layers was varied between 32, 64, and 96. The network was trained on the training set, and the bladders were segmented using DL-CNN likelihood maps with level sets. The change in the volume

intersection ratio was in the range of 0.5-1.9%, absolute volume error 0.2-9.3%, average minimum distance 0.6-10.1%, and the Jaccard index 0.1-2.2%. These results demonstrate that our DL-CNN based segmentation system is robust within a reasonable range of parameters.

A limitation to the new method is the long training time for the DL-CNN. The DL-CNN requires training, which takes approximately 5.5 hours for 160,000 ROIs and 1500 iterations. However, the processes involving the DL-CNN have not been optimized, and a slower GPU was used for compatibility reasons for this study. Optimizing the process and using faster hardware will reduce the runtime for training the DL-CNN. Once the DL-CNN has been trained, it takes approximately 4 minutes to generate the bladder likelihood maps within the VOI of a case. It takes 2 – 5 minutes to mark the VOI and run the level set segmentation, depending on the bladder size. On the other hand, CLASS with LCR takes approximately 4 – 10 minutes per case to mark the VOI and run the segmentation. Therefore, for an unknown case, it may require up to 10 minutes for DL-CNN with level sets, which is comparable to the CLASS with LCR method.

It is difficult to perform direct comparison of segmentation performance to the previous methods used by other investigators, as summarized in the Introduction due to the differences in the data sets and in their degrees of difficulty. A rough comparison can be made to only one of the studies[24] that reported quantitative results. Chai et al.[24] achieved Jaccard indices of 70.5% and 77.7% for their automatic and semi-automatic methods, respectively, using 95 scans of 8 patients for training, and 233 scans of 22 patients for testing. Our segmentation method using DL-CNN achieved higher accuracy than the automatic method from Chai et al.[24], and achieved comparable results to their semi-automatic method, while using a larger independent test set.

3.6 Conclusion

Our results show that the proposed segmentation method using DL-CNN can accurately segment the bladders on CTU scans. While only using a single bounding box for the entire bladder as the input to the system, the new method performed comparably to or better than our previous CLASS with LCR method for all performance measures that required two bounding boxes as input. However, the cost of this improvement is the increased runtime for training the DL-CNN. Once the DL-CNN is trained and implemented as a part of the segmentation package, the runtime for an unknown case becomes comparable. The agreement of the segmentations

between DL-CNN and the two hand outlines were comparable, but the agreement was lower than the agreement between the two hand outlines. We observed that DL-CNN can differentiate the inside and outside of the bladder regions better than the Haar features with random forest classifier, resulting in a more accurate bladder likelihood map and segmentation after refinement by level sets. Further work is underway to optimize the segmentation process and to improve the segmentation accuracy. It is especially important to include bladder lesions inside the segmented bladder boundaries. This study is a step toward the development of a reliable system for segmentation of bladders, which is a critical component of a CAD system for detection of urothelial lesions imaged with CT urography.

Chapter IV

Detection of Urinary Bladder Mass in CT Urography (CTU) with SPAN

4.1 Abstract

In this chapter, we focused on developing a system for detecting masses fully or partially within the contrast-enhanced (C) region of the bladder. With IRB approval, a data set of 70 patients with 98 biopsy-proven bladder lesions fully or partially immersed within the contrast-enhanced region (C region) of the bladder was collected for this study: 35 patients for the training set (38 malignant, 7 benign lesions), and 35 patients for the test set (49 malignant, 4 benign lesions). The bladder in the CTU images was automatically segmented using our Conjoint Level set Analysis and Segmentation System (CLASS) that we developed specifically to segment the bladder. A closed contour of the C region of the bladder was generated by maximum intensity projection using the property that the dependently layering contrast material in the bladder will be filled consistently to the same level along all CTU slices due to gravity. Potential lesion candidates within the C region contour were found using our Straightened Periphery ANalysis (SPAN) method. SPAN transforms a bladder wall to a straightened thickness profile, marks suspicious pixels on the profile, and clusters them into regions of interest to identify potential lesion candidates. The candidate regions were automatically segmented using our Auto-Initialized Cascaded Level Set (AI-CALS) segmentation method. Twenty-three morphological features were automatically extracted from the segmented lesions. The training set was used to determine the best subset of these features using simplex optimization with the leave-one-out case method. A linear discriminant classifier (LDA) was designed for the classification of bladder lesions and false positives. The detection performance was evaluated on the independent test set by free-response receiver operating characteristic (FROC) analysis. At the prescreening step, our system achieved 84.4% sensitivity with an average of 4.3 false positives per case (FPs/case) for the training set, and 84.9% sensitivity with 5.4 FPs/case for the test set. After LDA classification with the selected features, the FP rate improved to 2.5 FPs/case for the training set, and 4.3 FPs/case for the test set without

missing additional true lesions. By varying the threshold for the LDA scores, at 2.5 FPs/case, the sensitivities were 84.4% and 81.1% for the training and test sets, respectively. At 1.7 FPs/case, the sensitivities decreased to 77.8% and 75.5%, respectively. The results demonstrate the feasibility of our method for detection of bladder lesions fully or partially immersed in the contrast-enhanced region of CTU. The results presented in this chapter have been published[3], before the methods for bladder segmentation presented in Chapter III were published.

4.2 Introduction

Interpretation of a CTU study requires thorough image analysis, often requiring extensive time. The radiologist interpreting the study must visually determine the presence of lesions within the urinary tracts on a display workstation, frequently adjusting the brightness, contrast, and zoom levels. The radiologist must pay close attention throughout the entire urinary tract, as multiple lesions may be present. In addition, many different urinary anomalies may be found in a single CTU study. Not only do the radiologists have to identify these anomalies, they must also determine their likelihood each of these of being a urothelial neoplasm. The challenges of interpreting a CTU leads to a substantial variability among radiologists in detection of bladder cancer, with reported sensitivities ranging from 59% to 92%[18, 19]. The chance that the radiologist misses a subtle lesion may not be negligible due to overall workload, and the need to maintain rapid case throughput. Thus, any technique that helps radiologists identify urothelial neoplasms will be useful. Computer-aided detection (CAD), used as an adjunct, may reduce the chance of oversight by the radiologists. We are developing a CAD system that detects bladder cancer in CTU to be used for such purposes.

Few investigators have worked on CAD for bladder cancer to-date. Duan et al.[22] used an adaptive window setting to segment bladder tumor surfaces for magnetic resonance (MR) cystography. They generated multiple windows that covered the inner wall of the bladder and from which features were extracted. Using quadratic discriminant analysis, they determined if a window contained a true positive or false positive in a data set of 10 patients. Jaume et al.[46] detected bladder tumors in abdominal CT images by estimating the bladder wall thickness using inner and outer bladder surface meshes generated using the Marching Cubes algorithm. They separated each of the 26 bladders in their data set into 6 regions and created an atlas to distinguish between normal and diseased regions. Hadjiiski et al.[25] performed a pilot study in

detecting lesions within the contrast-enhanced region of the bladder in 15 patients in CTU. They used a rule-based system based on shape measures and uniformity measures to identify lesion candidates.

Automatic detection of bladder lesions in CTU is challenging. The imaged bladders, along with their lesions, can assume a variety of shapes and sizes. Bladders may also be partially or fully filled and may or may not contain excreted intravenous (IV) contrast material, resulting in variations in the degree of bladder opacification. The lesions in the bladder region that is filled with contrast material have much different contrast relative to those in the bladder region not filled with contrast material. This requires different strategies for the detection of lesions in the contrast enhanced (C) and non-contrast enhanced (NC) regions of the bladder.

In this study, we focused on developing a system for detection of bladder lesions fully or partially immersed in the C region of the bladder. We designed the system and evaluated its performance using free-response receiver operating characteristic (FROC) analysis using a data set of 70 cases. Although the data set was still small, to our knowledge it was the largest data set compared to those used in other reported studies.

4.3 Materials and Methods

The bladder as seen on the CTU images was automatically segmented within a manually marked bounding box, then as part of the prescreening step, the C region was delineated from the segmented bladders. The bladder wall of the C region was transformed into a wall thickness profile that was analyzed to determine lesion candidates. These candidates were automatically segmented, and morphological features were extracted. The best subset of these features was determined and a linear discriminant classifier was designed with a training set for classification of the bladder lesions and false positives. The block diagram of the detection system is presented in Figure 4.1. The detection performance was evaluated in an independent test set by FROC analysis.

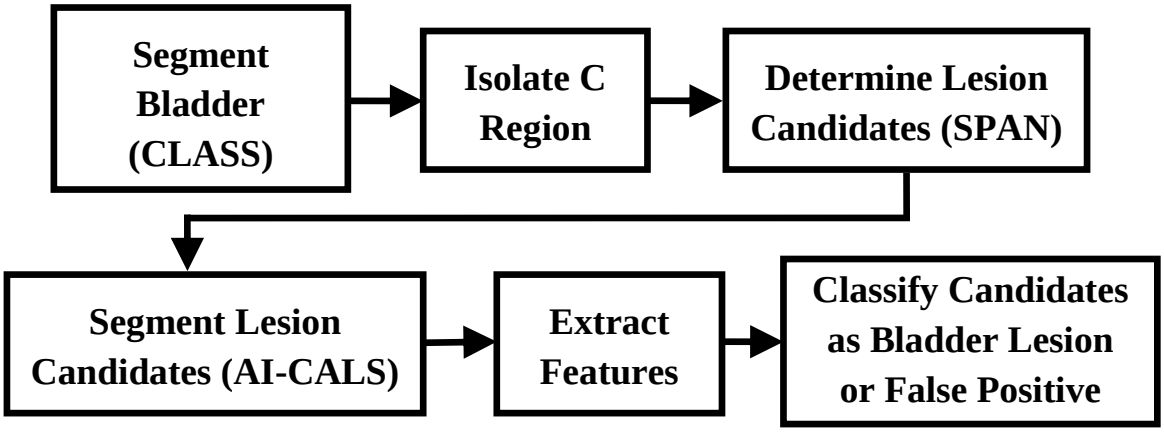


Figure 4.1: Block diagram of the detection system.

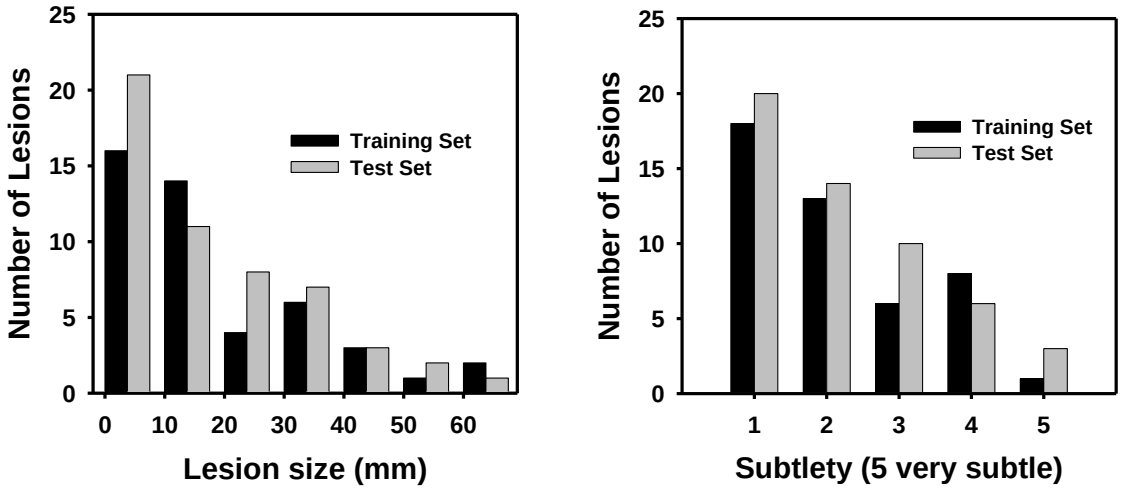
4.3.1 Data set

With approval of the Institutional Review Board, a data set of 70 patients undergoing CTU who subsequently underwent cystoscopy and biopsy was collected retrospectively from the Department of Radiology at the University of Michigan. The CTU scans were acquired with GE Healthcare LightSpeed MDCT scanners at a slice interval of 1.25 mm or 0.625 mm using 120 kVp and 160-560 mAs, and reconstructed with filtered back projection using the standard reconstruction kernel. The excretory phase images used were obtained 12 minutes after the initiation of intravenous injection of 125-175 mL of iodinated contrast material (at a concentration of 300 mg iodine per ml). Since patients were not moved prior to image acquisition, dependently layering IV contrast material that had been excreted into the renal collecting systems partially or fully filled the bladder on the excretory phase CTU images.

All lesions were marked by experienced radiologists on the CTU volumes as reference standard. Two radiologists (26 years of experience, 16 years of experience) marked the lesion by placing an ROI over the lesion, indicating the starting and ending slice of the lesion, measuring the longest diameter of the lesion, and giving a subtlety rating. Consensus was obtained if the lesion locations were different, and the final radiologist impression was correlated with radiology and pathology reports. The size and subtlety of the lesions given by the single, more experienced radiologist were reported to illustrate the detection performance of the system for lesions of different degrees of difficulty as seen by an experienced radiologist.

A total of 98 biopsy-proven bladder lesions was identified in the fully or partially contrast-enhanced region of the bladder. The cases were split evenly into independent training and test

sets. The training set contained 35 subjects having 38 malignant and 7 benign lesions with an average size of 20.1 mm (range: 4.2–61.7 mm), measured as the longest diameter on an axial slice (Figure 4.2(a)). The test set contained 35 subjects having 49 malignant and 4 benign lesions with an average size of 18.8 mm (range: 1.4–61.1 mm) (Figure 4.2(a)). The average lesion subtlety ratings in both sets were 2.2 (scale 1 to 5, 5 very subtle) (Figure 4.2(b)).



(a) (b)

Figure 4.2: Histograms of lesion size (a) and lesion subtlety (b) for lesions in the training and test set. The average lesion size was 20.1 mm (range: 4.2–61.7 mm) for the training set, and 18.8 mm (range: 1.4–61.1 mm) for the test set. The average lesion subtlety ratings in both sets were 2.2 (scale 1 to 5, 5 very subtle).

For image processing purposes, all CT voxel values in terms of Hounsfield Units (HU) are linearly shifted into gray level, where gray level = HU + 1024 so that all image voxel values are positive.

4.3.2 Bladder segmentation using CLASS

A critical component of CAD system that detects bladder cancer is accurate bladder segmentation that isolates the bladder from the surrounding anatomical structures. An axial CTU scan of the bladder is shown in Figure 4.3. The bladder shown is partially filled with IV contrast material and a malignant lesion can be identified in the posterior, contrast-enhanced portion of the bladder. The presence of the two distinct regions in the bladder lumen that have very different attenuation values: a region filled with IV contrast material and a region without contrast material, poses a challenge for segmentation algorithm that needs to go across the strong boundary.

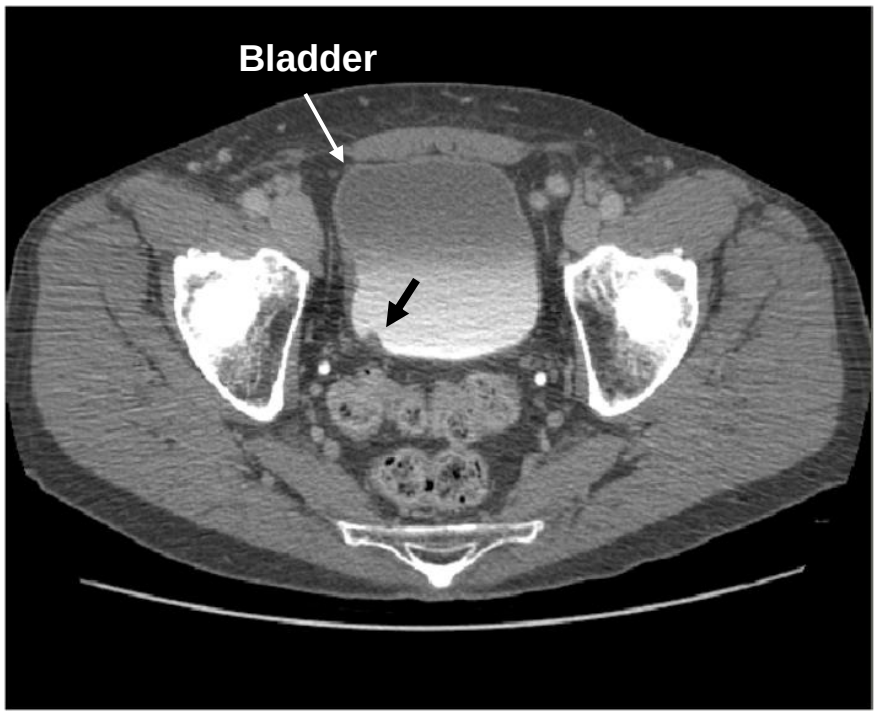


Figure 4.3: An axial slice of a CTU scan in which the bladder is partially filled with IV contrast material. A malignant lesion is present in the contrast-enhanced region of the bladder, indicated by the bold arrow.

We have applied our methods described in Chapter II, called Conjoint Level set Analysis and Segmentation System (CLASS)[1, 11] in order to segment the bladder in CTU. An example of a bladder and its segmentation is shown in Figures 4.4(a) and 4.4(b), respectively.

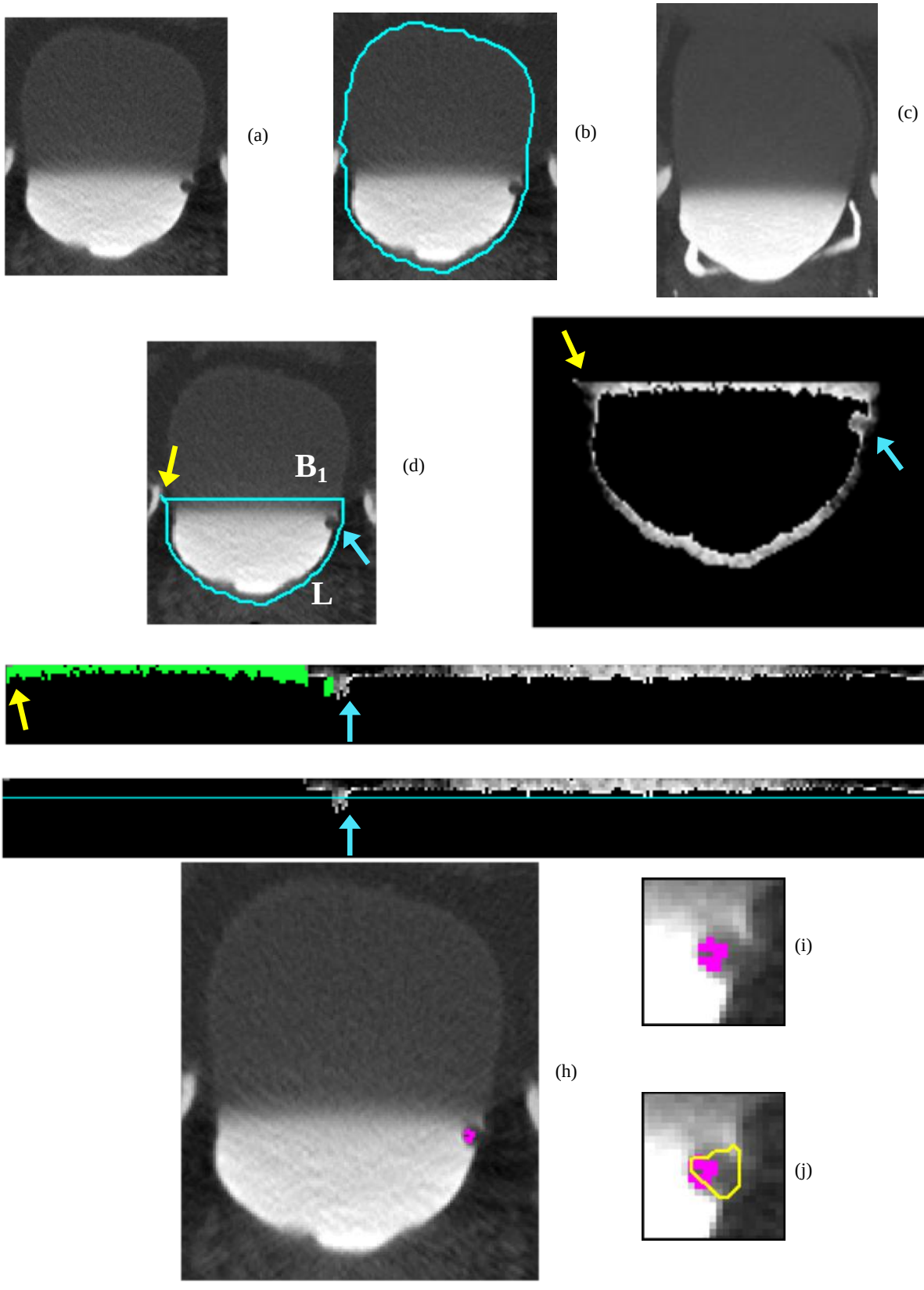

(a)
(b)
(c)
B1
L
(d)
(e)
(f)
(g)
(h)
(i)
(j)

Figure 4.4: Bladder lesion candidate prescreening and segmentation – example of true positive. (a) ROI of the CTU slice that includes the lesion. (b) Bladder segmentation of the slice encompassing the bladder wall. (c) Maximum Intensity Projection (MIP) of the bladder used to determine the boundary between the contrast-enhanced and non-contrast regions. (d) Segmentation of the contrast-enhanced region of the bladder (L contour). The horizontal line on top of the contour is the boundary between the NC and C regions of the bladder, B_1. The arrow on the left side of the bladder indicates the starting point for the wall thickness profile. The arrow on the right points to a true lesion. (e) Magnified C region image after adaptive thresholding. (f) Bladder wall profile. The pixels marked in green were removed during the false positive reduction of voxel candidate. The left section of the profile was removed using location-based rules as described in section 2.3.3. (g) Bladder wall profile used for candidate detection. The line is the threshold used to determine lesion candidates. The arrow points to a lesion candidate that is mapped onto the bladder in (h). (i) Magnified image of the region around the lesion candidate. The windowing of the image was adjusted to better visualize the bladder wall. (j) Lesion candidate segmentation. The segmentation refines the initial region (in pink), resulting in a better representation of the lesion.

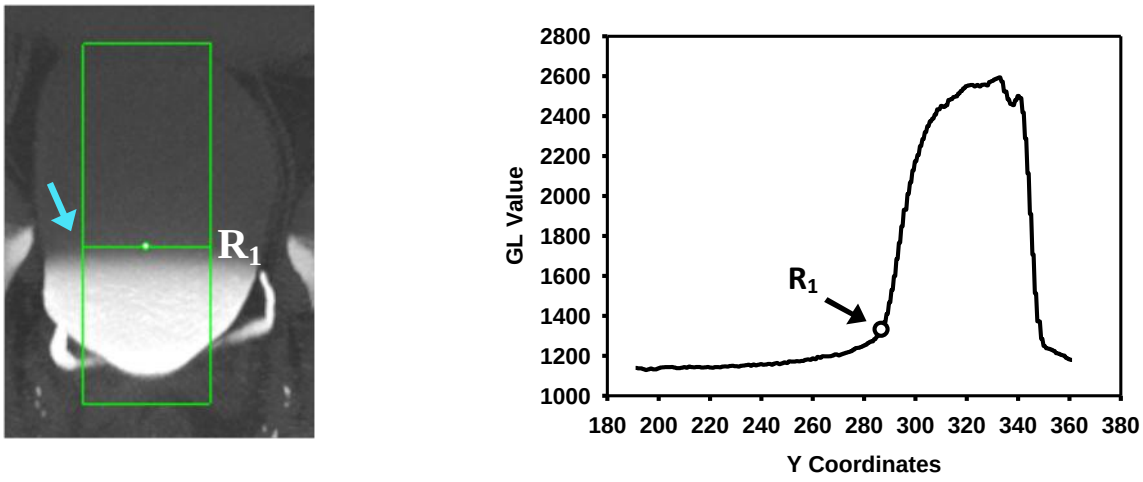
4.3.3 Bladder wall profile generation and lesion candidate identification with SPAN

Bladder lesion candidates are identified by first isolating the contrast-enhanced region of the bladder, and then by using our newly developed method, referred to as Straightened Periphery ANalysis (SPAN). SPAN consists of three stages: (1) wall thickness profile generation, (2) false positive reduction of voxel candidate, and (3) lesion candidate identification.

4.3.3.1 Isolating the contrast-enhanced region of the bladder

The contrast-enhanced region of the bladder is separated from the non-contrast region using the property that the dependently layering excreted contrast material in the bladder will be filled to the same level consistently along all CTU slices due to gravity. We use maximum intensity projection (MIP) along the slices of the bladder to estimate the upper boundary of the contrast enhanced region. The ROIs initializing the segmentation of the C and NC regions of the bladder are used to determine the range of the CTU slices for the MIP. As the bladder is located anterior to the larger pelvic bones when the patient lies in a supine position, it is common for the bones intrude into the bladder's ROI towards the bladder neck. If the high attenuation bones show up on the MIP image, their interference makes it difficult to accurately determine the upper level of the contrast material. Therefore, only a portion of the slices included in the ROI is used. We experimentally determined that 30% of the slices towards caudal aspect of the bladder (towards the patient's feet) from the best slice (to avoid the pelvic bones) and 90% of the slices towards the cephalic aspect of the bladder (towards the patient's head) from the best slice (to avoid other organs above the bladder) worked the best. The best slice is the slice that best represents the bladder region, e.g., where the bladder to be the largest. It was selected manually when the ROIs were defined. An MIP image is shown in Figure 4.4(c).

From the MIP image, a gray level profile is generated as described below. The two ROI boxes marked for the NC and C region segmentations are combined to create a rectangular box such that the width of the box is the width of intersection of the NC and C boxes and the height is the union of the heights of the NC and C boxes. Then the box's width is reduced by 50% while keeping the same center (Figure 4.5(a)) to minimize the negative effects of the peripheries of the irregularly shaped bladders on the estimation of the transition point between the contrast and non-contrast regions. For every row of the box, the gray levels of the pixels belonging to the row are averaged and recorded into a profile (Figure 4.5(b)). The profile is analyzed to find the first row R_1 whose average gray level is greater than a gray level threshold Th_p. By using the training data set the Th_p was determined experimentally as 1330, which provided adequate separation of the NC and C regions for the training cases. A horizontal line B_1 at row R_1 that intersects with the bladder boundary is determined as the boundary between the NC and C regions of the bladder (Figure 4.4(d)). The portion of the C region of the bladder contour and the B_1 boundary then form a new closed contour, referred to as the L contour, that encloses the C region (Figure 4.4(d)). The image pixels that are within the L contour, i.e., the C region of the bladder, are analyzed in the subsequent steps.



(a) (b)

Figure 4.5: Estimation of boundary (row R_1) between the NC and C regions. (a) The box used to calculate the average bladder GL profile shown in (b). The arrow points to the row that the y-coordinate was determined to be that of R_1. (b) Profile of the average GL values for each row of the box in (a). The arrow indicates the average GL of the first row that has value above the 1330 threshold and therefore identified as R_1.

4.3.3.2 Wall thickness profile generation with SPAN

Generating the bladder wall thickness profile using the previously obtained C region image and L contour involves two steps: (1) adaptive thresholding of the contrast material, and (2) transformation of the L contour to a straightened wall profile.

Adaptive thresholding. The contrast material within the C region is removed by adaptive thresholding. Using a constant threshold for the contrast material may result in missing lesions that have relatively high gray levels in some cases, and may also lead to portions of the contrast material not being eliminated that may become false positives in other cases. Using adaptive thresholding method resolves these potential issues.

For each slice within the C region, the mean and the standard deviation of the pixel gray levels are calculated. For pixels whose gray level was greater than 1800, their gray level was set to 1800 for calculating the average. The average gray level, GL_{Avg}, is stratified into four different groups that are used to determine the initial threshold, Th_{init}, using the following decision rules, which were determined using the training set:

$$Th_{init} = \begin{cases} Th_I^H & GL_{Avg} \geq Th_I^H \\ Th_I^{MH} & Th_I^{MH} \leq GL_{Avg} < Th_I^H \\ Th_I^{ML} & Th_I^{ML} \leq GL_{Avg} < Th_I^{MH} \\ Th_I^L & GL_{Avg} < Th_I^{ML} \end{cases} \quad (4.1)$$

where Th_I^H , Th_I^{MH}, Th_I^{ML}, and Th_I^L are the high, medium high, medium low, and low gray level thresholds, respectively. The thresholds were determined after analyzing the histogram of pixel gray levels within the C region using the training set. The histograms of pixel gray levels within the C region for both the training and test sets are shown in Figure 4.6. Assuming a Gaussian mixture distribution, multiple Gaussians were fit to the histogram for the training set (Figure 4.6(a)), and the peaks of the Gaussians, (Th_I^{MH}, Th_I^L) along with intersections of the Gaussians that corresponded to a sharp drop in the histogram (Th_I^H, Th_I^{ML}) were used as the thresholds after adjusting for outlier cases (Table 4.1). The peak of the Gaussian whose gray level was above 1330 was not used as a threshold, as our analysis during the isolation of the C region showed that these pixels are generally part of the contrast material and not the bladder wall. For slices with high GL_{Avg}, Th_{init} is set to be Th_I^H to ensure most of the contrast material is removed after thresholding. For slices with average contrast enhancement between Th_I^H and Th_I^{MH} , setting Th_{init} to be Th_I^{MH} will remove the majority of the contrast material in the slice (Table 4.1 and Figure 4.6). Th_{init} is set to be Th_I^{ML} for slices with GL_{Avg} within the range of Th_I^{MH} to

Th_I^{ML}, which are usually near the end of the bladder. The amount of contrast material in these slices is generally less than the amount in slices near the center of the bladder, resulting in lower intensity of the contrast region. The slices whose GL_{Avg} is less than Th_I^{ML} typically are those with C regions that were not well-enhanced by the contrast material. These slices require a smaller gray level threshold or the contrast material would not be removed, thus their Th_{init} is set to be Th_I^L.

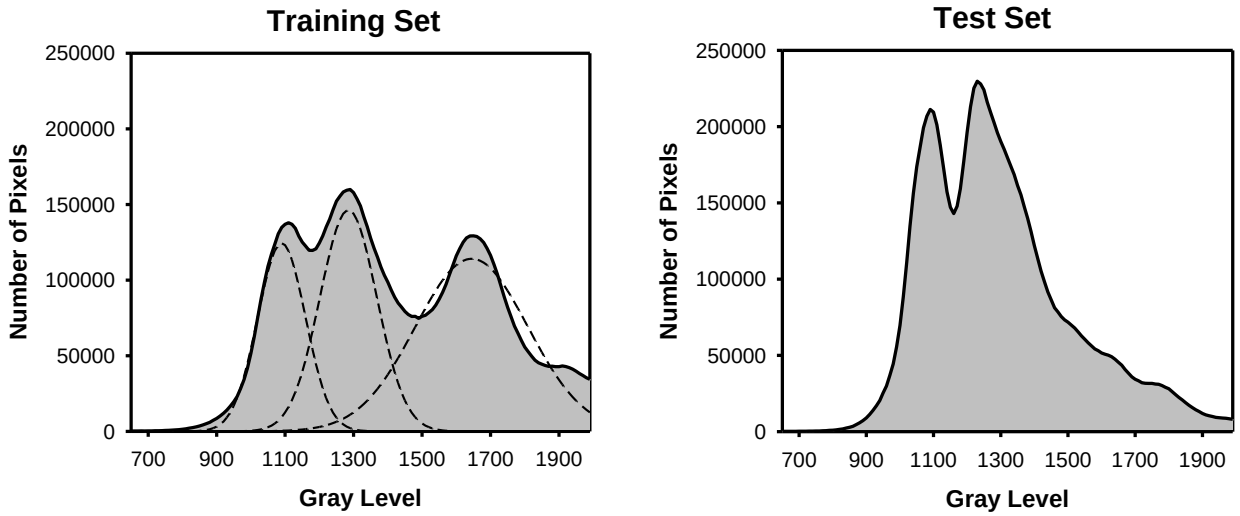


Figure 4.6: Histogram of gray level values of pixels within the C region of the (a) training set and (b) test set. The multiple Gaussians that were fitted to the training set to determine threshold values are shown with dotted lines in (a).

Table 4.1: Parameter values for adaptive thresholding in wall thickness profile generation.

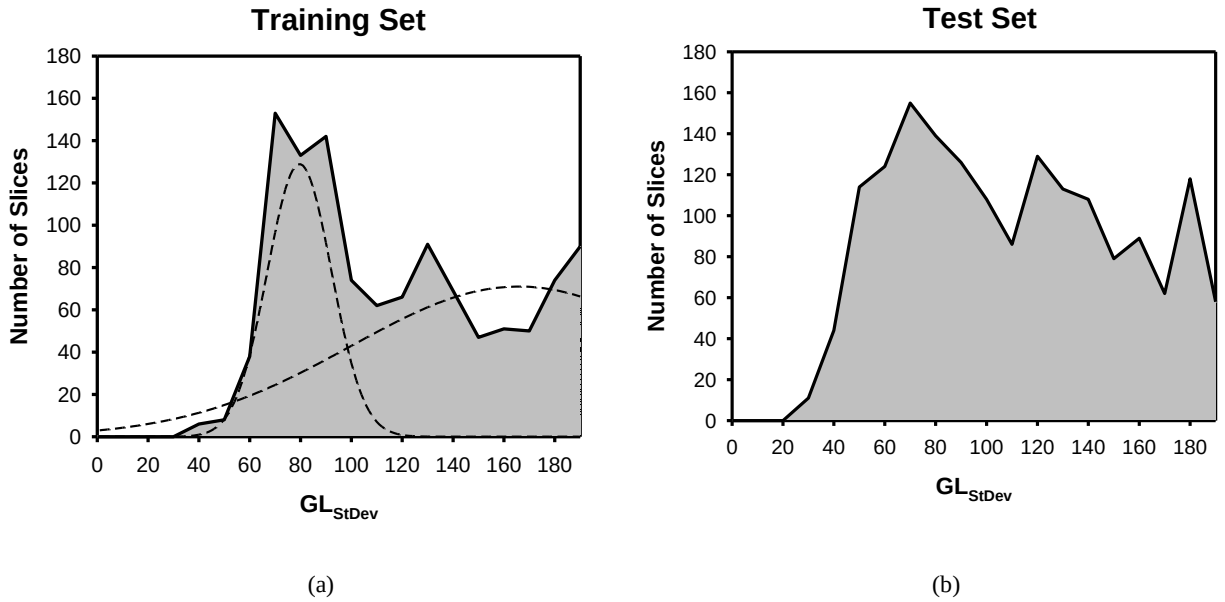
Th_I^H	Th_I^{MH}	Th_I^{ML}	Th_I^L	ε	λ	τ	Th_{GL}^H	Th_{GL}^L
1400	1300	1200	1150	15	-60	20	100	80

The gray level threshold for the contrast material, Th_C, is then refined adaptively by both Th_{init} and the standard deviation of the pixel gray levels, GL_{StDev}, within the C region as follows:

$$Th_C = \begin{cases} Th_{init} - \dfrac{GL_{StDev}}{\varepsilon} & GL_{StDev} \geq Th_{GL}^H \\ Th_{init} - \lambda & Th_{GL}^L \leq GL_{StDev} < Th_{GL}^H \\ Th_{init} + \tau & GL_{StDev} < Th_{GL}^L \end{cases} \tag{4.2}$$

where ε, λ, and τ are constants, and Th_{GL}^H, and Th_{GL}^L are the high and low thresholds for GL_{StDev}, respectively. Using the training set, the constants were determined experimentally, while the thresholds were determined by analyzing the histogram of the GL_{StDev}. The histogram of GL_{StDev}

for both the training and test set are shown in Figure 4.7. After fitting two Gaussians to the histogram of the training set GL_{StDev} (Figure 4.7(a)), the peak of one of the Gaussians, Th_{GL}^{L}, and the intersection of the Gaussians corresponding to a sharp drop, Th_{GL}^{H}, on the histogram were used as the thresholds, leading to three different categories. For slices with GL_{StDev} greater than Th_{GL}^{H}, the C region is usually very inhomogeneous and requires a lower threshold to ensure most of the contrast material is removed. For slices with GL_{StDev} values in the range of Th_{GL}^{L} to h_{GL}^{H}, they generally also contain pixels with high gray levels so that the Th_{init} is also lowered to ensure that sufficient contrast material is removed for the subsequent stages. Slices with GL_{StDev} values lower than Th_{GL}^{L} usually contain fairly homogenous C region. The dependent layering of the contrast material on these slices is not as prevalent as other slices with higher GL_{StDev}, thus their gray level are lower. These cases need a higher threshold to ensure that lesions are not removed along with the contrast material. The values of the chosen constants and thresholds are shown in Table 4.1.



(a) (b)

Figure 4.7: Histogram of standard deviation values of pixel gray level within the C region of a CTU slice, GL_{StDev} for the (a) training set and (b) test set. The multiple Gaussians that were fitted to the training set to determine threshold values are shown with dotted lines in (a).

Once Th_C is determined, the contrast material is eliminated from within the C region by setting the gray level to 0 for all pixels whose gray level is greater than Th_C. Examples of C region images after adaptive thresholding are shown in Figures 4.4(e), 4.8(b), and 4.9(b).

Comparison of the C region with and without the adaptive thresholding for cases in the three different GL_{StDev} categories is shown in Figure 4.10. Notice that without the adaptive thresholding (Figure 4.10(d-f)), much of the contrast material remains within the C region image, which would lead to incorrect wall thickness profiles and thus missing lesion candidates and false positive detections. With adaptive thresholding, well-defined bladder wall can be obtained (Figure 4.10(g-i)).

Wall profile generation. Once the contrast material is removed from the slice, a straightened profile of wall thickness is generated by mapping all of the points along the L contour, L_i, $i=1, \ldots, n$, sequentially to the X-axis of a new coordinate system such that $L_i(x_i, y_i)$ has the coordinate $(X_i,0)$. The origin of this new coordinate system is defined at the top left of the profile, with Y-values increasing in the downward direction. For a given pixel $L_i(x_i, y_i)$, the path normal to the point towards the interior of the L contour is calculated using the normal angle θ defined as:

$$\theta = 90^{\circ} + \frac{1}{2}\left(\tan^{-1}\left(\frac{y_i - y_{i-1}}{x_i - x_{i-1}}\right) + \tan^{-1}\left(\frac{y_{i+1} - y_i}{x_{i+1} - x_i}\right)\right), \quad (4.3)$$

where (x_{i+1},y_{i+1}) and (x_{i-1},y_{i-1}), respectively, are the coordinates of the next L_{i+1} and previous L_{i-1} neighboring points of L_i,. The pixels along the normal path are sequentially mapped onto the profile at increasing Y-values such that the new coordinates of the pixels are given by (X_i,Y_j) $j=1,\ldots i_m$, while X_i is fixed. The path along the normal ends when four black pixels are encountered consecutively, indicating that the path reaches the lumen of the C-region where the pixel gray level has been set to 0. The number of pixels along the normal path at L_i is denoted by i_m. Figures 4.4(f), 4.8(c), and 4.9(c) show examples of the transformation. Pixels that will be removed by the following false reduction step are also marked on the figures.

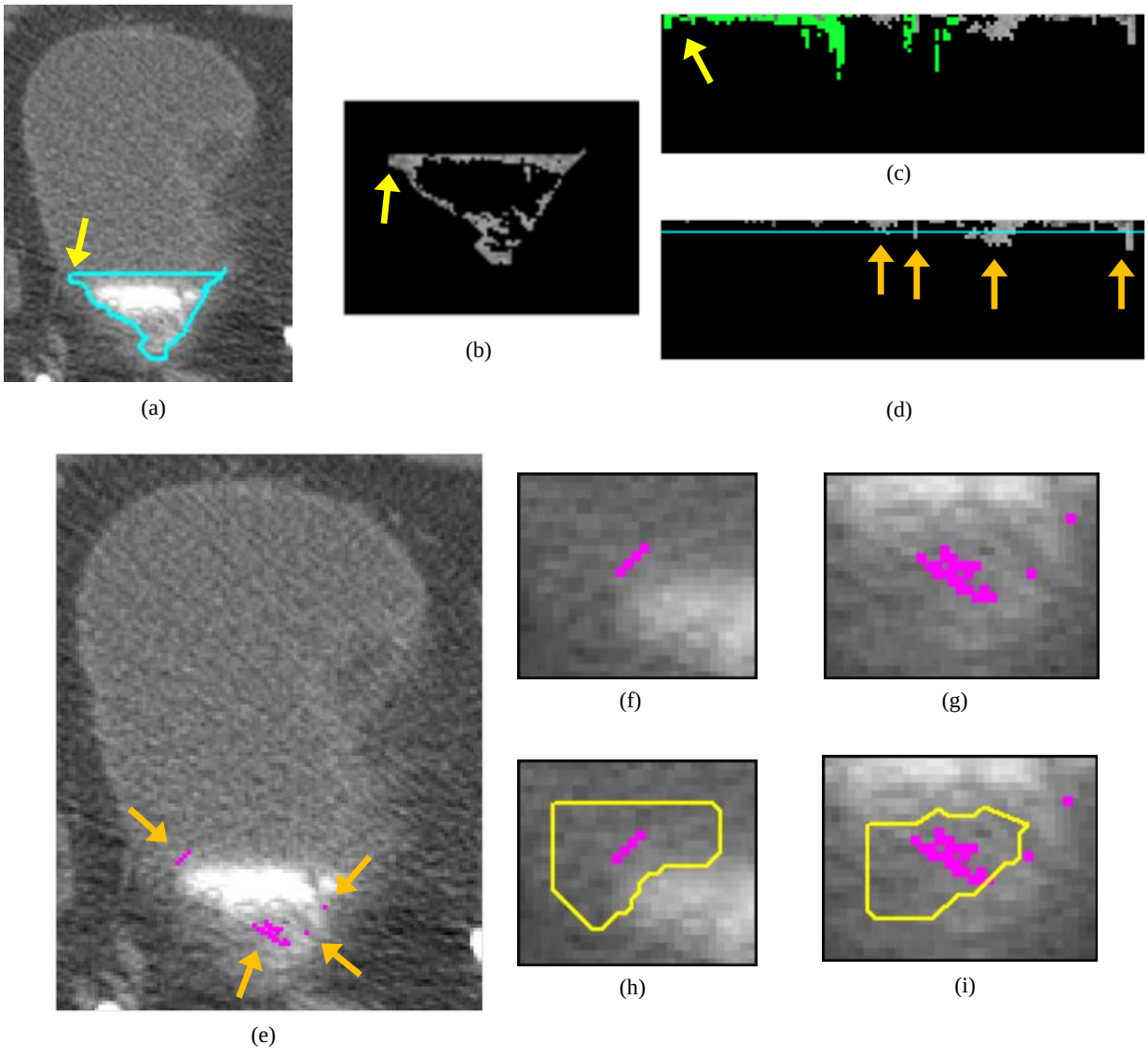


Figure 4.8: Bladder lesion candidate prescreening and segmentation at a slice near the end of the bladder – example of false positives. (a) Segmentation of the C region of the bladder (L contour). (b) C region image after adaptive thresholding. (c) Bladder wall profile. The pixels marked in green were removed during the false positive reduction of voxel candidate. (d) Bladder wall profile used for candidate detection. The line is the threshold used to determine lesion candidates. The arrows point to lesion candidates. (e) Lesion candidates projected onto the bladder. Arrows point to lesion candidates. (f,g) Magnified image of the region around the lesion candidate. (h,i) Lesion candidate segmentation. Two single pixel lesion candidates shown in (e) and (g) were discarded during the lesion candidate determining stage using the size criteria. The two remaining candidates were both false positive lesions and were removed by the LDA classifier.

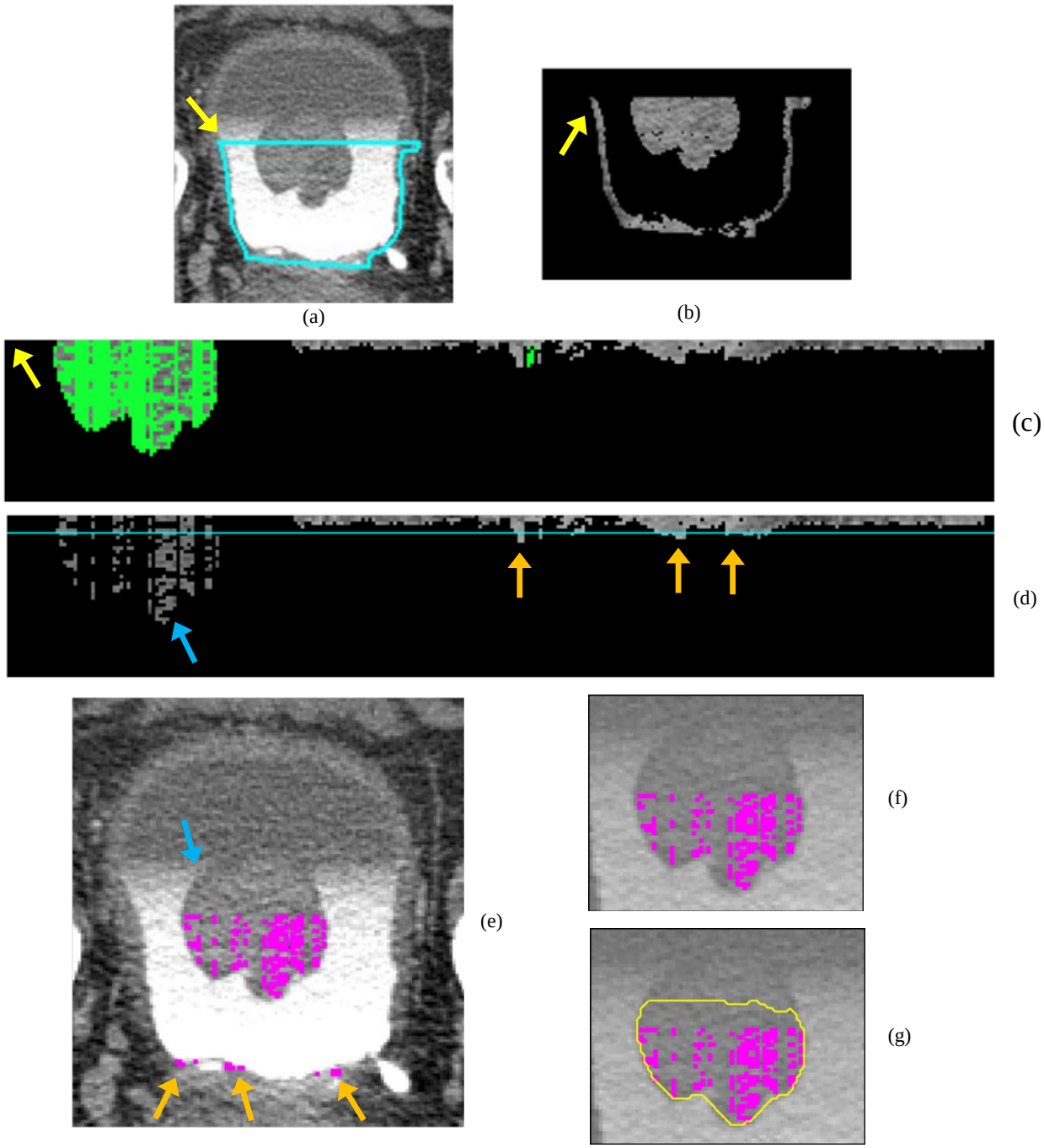


Figure 4.9: Bladder lesion candidate prescreening and segmentation for a lesion along B_1 – example of true positive. (a) Segmentation of the C region of the bladder (L contour). (b) C region image after adaptive thresholding. (c) Bladder wall profile. The pixels marked in green were removed during the false positive reduction of voxel candidate. (d) Bladder wall profile used for candidate detection. The line is the threshold used to determine lesion candidates. The arrows point to lesion candidates. (e) Lesion candidates projected onto the bladder. Arrows point to lesion candidates. (f) Magnified image of the region around the lesion candidate. The windowing of the image was adjusted to better visualize the lesion. (g) Lesion candidate segmentation. The three candidate pixel regions at the bottom of the bladder were discarded during the lesion candidate determining stage using the size criteria.

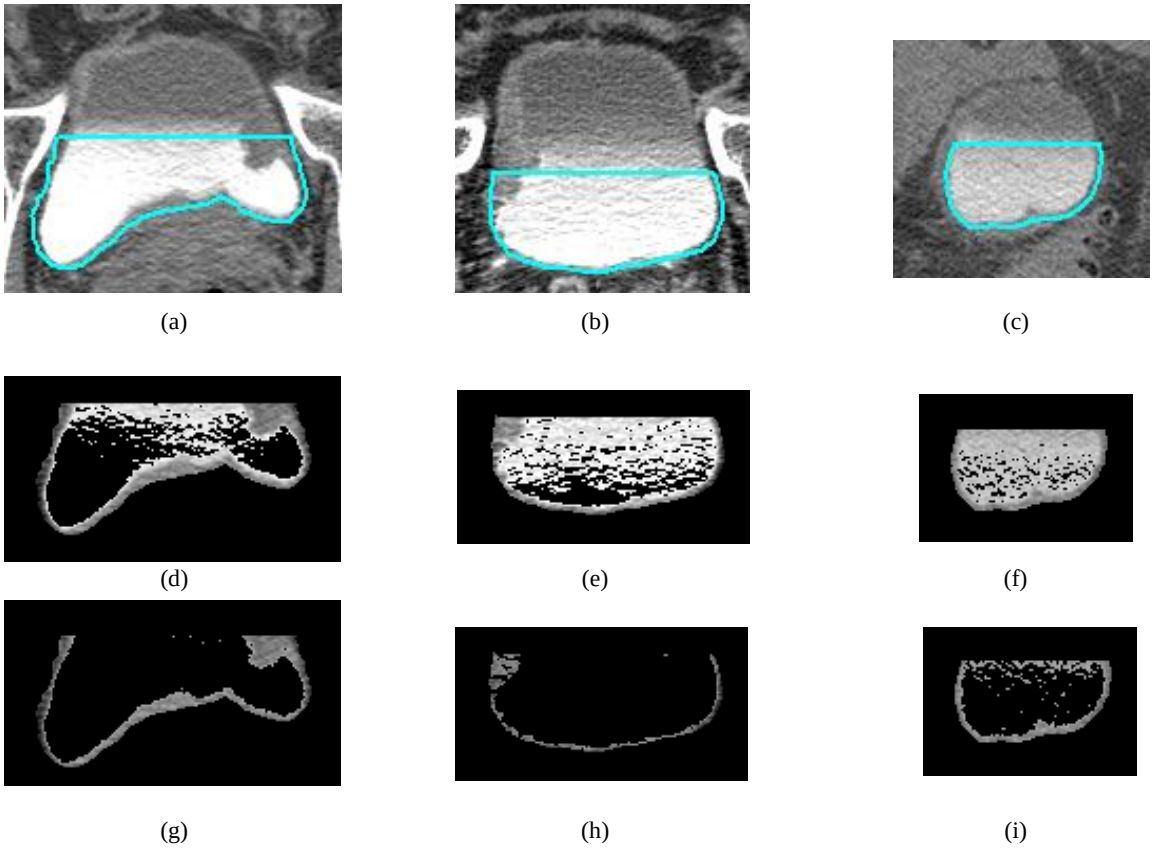


Figure 4.10: C region images with and without adaptive thresholding for cases that fall within the different categories of GL_{StDev}. (a-c) Bladder slices with L contour with different standard deviations, GL_{StDev}: (a) GL_{StDev} = 133, (b) GL_{StDev} = 89, (c) GL_{StDev} = 77. (d-f) C region after hard thresholding slices in (a-c) using Th_C of 1330 without adaptive thresholding. (g-i) C region after adaptive thresholding with rules in Equation (4.3) for slices in (a-c). (g) Th_C = 1191, ($GL_{StDev} \geq Th_{GL}^{H}$), (h) Th_C = 1140, ($Th_{GL}^{L} \leq GL_{StDev} < Th_{GL}^{H}$), (i) Th_C = 1220, ($GL_{StDev} < Th_{GL}^{L}$).

4.3.3.3 False positive reduction of voxel candidate with SPAN

False positive reduction is applied to the straightened bladder wall profile. One step for reducing false positives is to remove pixels with gray level values above 1100 on the profile in the section corresponding to B_1 of the L contour. Columns on the profile in the section whose starting points, (X_i,0), has a gray level value above 1100 are also removed. Because the B_1 section is the border separating the C region and the NC region of the bladder, the gray level of the region along B_1 is often lower than other regions within the C region of the bladder. This may cause many false positive findings, as the darker regions transformed onto the profile may be marked as lesion candidates. After studying lesions in the training set that are located near B_1, we found that the pixels of lesions within the B_1 region typically have gray level values less than

1100. By applying this criterion, we can reduce false positive findings that may be caused by the lower pixel gray levels of the profile along the B_1 boundary neighboring the non-contrast region.

Sharp peaks along the bladder wall profile that are likely caused by noise are also removed. Since noise is random and forms small peaks, we set a criterion to exclude peaks less than three pixels in width and greater than five pixels in height relative to its surroundings. Examples of SPAN for a true lesion, false positive lesions, and a lesion located along B_1 are shown in Figures 4.4(f), 4.8(c) and 4.9(c), respectively.

4.3.3.4 Lesion candidate identification with SPAN

Lesion candidates are found by analyzing the bladder wall profile. For a given bladder profile, the average height of the profile (μ) is calculated, excluding the pixels from the B_1 section (the border between the NC and C regions) and locations whose height is greater than 10 pixels. The standard deviation (σ) of the pixels used is also calculated, and a height threshold (H), in pixels, is set using the following equation that was determined experimentally to maximize the number of true positives found while keeping the false positive findings low:

$$H = floor\left(\mu - \frac{\sigma}{2} + 0.8\right) + 2 \tag{4.4}$$

where the *floor* function represents the rounding down operation. On the profile, pixels with heights larger than H are considered to be lesion candidate pixels (Figure 4.4(g), Figure 4.8(d), Figure 4.9(d)). After the lesion candidate pixels are identified, they are mapped back to the original bladder slices as lesion candidate voxels (Figure 4.4(h, i), Figure 4.8(e-g), Figure 4.9(e-f)).

The lesion candidate voxels are then grouped into regions. Candidate voxels that are one slice apart, and within five voxels apart in 3D space are clustered into the same region. A candidate voxel is ignored if there are no other candidate voxels that are within five voxels on the same slice. Regions that contain less than five candidate voxels or greater than 50,000 candidate voxels are ignored. This size range was determined by analyzing the training cases. Each of the regions retained is enclosed with a 3D bounding box to indicate an ROI for a lesion candidate.

4.3.4 Lesion candidate segmentation, feature extraction, and classification

4.3.4.1 Lesion segmentation with Auto-Initialized Cascaded Level Set (AI-CALS)

Using the ROI obtained for a lesion candidate, the lesion is segmented from the surrounding tissue using the AI-CALS segmentation system[47]. The AI-CALS system consists of three stages: preprocessing, initial segmentation, and level set segmentation. While this process seems similar to the CLASS system used for bladder segmentation, AI-CALS uses a different method for initial segmentation, and different sets of parameters that were specifically developed for segmentation of bladder lesions. In the first stage, preprocessing techniques are applied in 3D to the ROI obtained from the candidate prescreening process above. Smoothing, anisotropic diffusion, gradient filters and the rank transform of the gradient magnitude are applied to the slices within the ROI to obtain a set of smoothed images, a set of gradient magnitude images, and a set of gradient vector images. The set of smoothed images is used in the second stage, while the other two sets are used during level set propagation in the third stage.

We modified the AI-CALS method in the second stage to improve the lesion segmentation performance for this study. In the second stage, the system automatically labels a subset of voxels in the ROI for analysis based on the attenuation, gradient, and location of the voxels. First, voxels with gray level value below 600 are removed. The voxels with gradient values in the top 50 percentile of all voxels in the ROI are identified using the gradient magnitude image and removed from the ROI. In order to distinguish between the contrast material and the lesion candidate that may be present within the ROI, a step that uses Otsu's method[48] is added that is not part of the original AI-CALS. The region with voxel gray level values between 1024 and the threshold gray level determined by the Otsu's method is marked by a binary mask. An elliptical cylinder whose radius is 0.8 of the width and height of the ROI, centered at the centroid of the binary mask, is placed within the binary mask. The intersection of the binary mask and the elliptical cylinder is labeled as the object region. In the original AI-CALS, an ellipsoid centered at the ROI is to determine the object region; however, we found that using an ellipsoid causes the lesions to be under-segmented towards the first and last slices of the ROI, while using an elliptical cylinder, as described above, alleviate this problem. A morphological dilation filter with a spherical structuring element of 2 voxels in radius, 3D flood fill algorithm, and a morphological erosion filter with a spherical structuring element of 2 voxels

in radius are applied to the object region to connect neighboring components and extract an initial segmentation surface.

In the third stage, the initial segmentation surface is propagated towards the lesion boundary using cascading level sets. Our chosen level set implementation evolves according to the equation:

$$\frac{\partial}{\partial t}\Psi(x) = -\alpha A(x)\cdot\nabla\Psi(x) - \beta P(x)|\nabla\Psi(x)| + \gamma\kappa(x)|\nabla\Psi(x)| \tag{4.5}$$

where α, β, and γ are the coefficients for the advection, propagation, and curvature terms, respectively, *A(x)* is a vector field image (assigning a vector to each voxel in the image) that drives the contour to move towards regions of high gradient, *P(x)* is a scalar speed term between 0 and 1 causing the contour to expand at the local rate, and $\kappa(x) = div\left(\frac{\nabla\Psi(x)}{|\nabla\Psi(x)|}\right)$ is the mean curvature of the level set at point x. The symbol ∇ denotes the gradient operator and *div* is the divergence operator[28].

Three 3D level sets with predefined sets of parameters are applied in series to the initial segmentation surface. The corresponding parameters of the 3 level sets are presented in Table 4.2.

Table 4.2: Parameters for the AI-CALS level sets

Level set:	α	β	γ	n
First	1	2	1	10
Second	1	0.4	q	100
Third	0	1.0	0	20
2D slices	4.0	0.3	0.5	100

The first 3D level set slightly expands and smoothes the initial contour. The second 3D level set brings the contour towards the sharp edges, but also expands it slightly in regions of low gradient. The parameter "q" in Table 4.2 is defined to be a linear function $\sigma M + \phi$ of the 2D diagonal distance *M* of the ROI box in millimeters (mm), where $\sigma = 0.06$, $\phi = -0.11$ as shown previously[28]. The third 3D level set further draws the contour towards sharp edges. As a final step, a 2D level set is applied to every slice of the segmented object to refine the 3D contours using the 3D level set contours as the initial contour. Further details on the AI-CALS method can be found in the literature[47]. Examples of true and false segmented bladder lesion candidates are shown in Figures 4.4(j), 4.8(h, i), and 4.9(g).

4.3.4.2 Feature extraction and classification

For each segmented lesion candidate, 23 morphological features are automatically extracted from the central slice of the segmented lesion. Five of the morphological features are based on the normalized radial length that is defined as the Euclidean distance from the object's centroid to each of its edge pixels, i.e., the radial length, normalized relative to the maximum radial length of the object[49]. Table 4.3 lists the features that were used. The definitions of these features can be found in the literature[50]. These features are studied because we found that these features are useful for lesion classification from our previous experience with breast masses.

Table 4.3: Table of morphological features used.

NRL	Shape-based	Gray level-based
NRL Mean	Perimeter	10 Contrast Features
NRL Standard Deviation	Area	Gray Level Average
NRL Entropy	Circularity	Gray Level Standard Deviation
NRL Area Ratio	Rectangularity	
NRL Zero Crossing Count	Perimeter-to-area Ratio	
	Fourier Descriptor	

Using the training set, stepwise feature selection is used to select the best feature subset. Using simplex optimization with leave-one-out case method, the best combination of values for the feature selection parameters, F_{in}, F_{out}, and tolerance, is determined from the training set. Features for classification are then selected from the entire training set with the best thresholds. The six features selected were normalized radial length area, rectangularity, area, average gray level, and two contrast features. More details about these features and the feature selection can be found in the literature[51]. A linear discriminant (LDA) classifier was then designed with the training set for classification of the bladder lesions and false positives using the selected features as input predictor variables. The trained LDA classifier was applied to the test set for independent testing.

4.3.5 Evaluation Methods

The performance of the lesion candidate prescreening steps was evaluated by determining the sensitivity and specificity. The overall performance of the bladder lesion detection CAD system with feature extraction and FP reduction by the LDA classifier was evaluated using free-response receiver operator characteristics (FROC) analysis using our in-house developed package that calculated the sensitivity and specificity at specific operating

points. The FROC curve was generated by varying the decision threshold for the LDA discriminant scores.

4.4 Results

At the prescreening step, our system achieved 84.4% (38/45) sensitivity with an average of 4.3 false positives per case (FPs/case) for the training set, and 84.9% (45/53) sensitivity with 5.4 FPs/case for the test set. The prescreening step generated 215 lesion candidates for the training set, (66 true lesion candidates, 149 false positive lesion candidates) that were used to train the LDA classifier.

Tables 4.4 and 4.5 summarize the detected lesions during prescreening by size and subtlety, respectively. For both the training and test sets, most of the lesions that were missed had subtlety ratings of greater than 3 and were smaller than 10 mm; however, the detection system was able to find the majority of the lesions that fit into these categories. For lesions smaller than 10 mm, 66.7% (10/15) and 71.4% (15/21) were found by the system in the training and test sets, respectively (Table 4.4). For lesions with subtlety ratings greater than 2, 64.3% (9/14) of them were detected in the training set, while 78.9% (15/19) of them were detected in the test set (Table 4.5). The system detected 86.8% (33/38) and 85.7% (42/49) of the malignant lesions in the training and test sets.

Table 4.4: Detected lesions at the prescreening stage for lesions of different sizes.

	Lesion Size (mm)						
	0-10	10-20	20-30	30-40	40-50	50-60	60-70
Training set	66.7% (10/15)	100% (14/14)	50.0% (2/4)	100% (6/6)	100% (3/3)	100% (1/1)	100% (2/2)
Test set	71.4% (15/21)	100% (11/11)	85.7% (7/8)	85.7% (6/7)	100% (3/3)	100% (2/2)	100% (1/1)

Table 4.5: Detected lesions at the prescreening stage for lesions of different subtleties.

	Lesion Subtlety (1-5, 5 very subtle)				
	1	2	3	4	5
Training set	100% (18/18)	84.6% (11/13)	100% (5/5)	50.0% (4/8)	0% (0/1)
Test set	90.0% (18/20)	85.7% (12/14)	90.0% (9/10)	83.3% (5/6)	33.3% (1/3)

Using feature extraction and LDA classifier, the false positive (FP) rate improved to 2.5 FPs/case for the training set and 4.3 FPs/case for the test set without missing additional true lesions. By varying the threshold for the LDA scores, the FROC curve was generated as shown in Figure 4.11. At 2.5 FPs/case, the training set achieved a sensitivity of 84.4%, while the test set achieved 81.1%. At 1.7 FPs/case, the sensitivities were 77.8% and 75.5% for the training and test sets, respectively. Table 4.6 shows the sensitivities of the system at different false positive rates for the training and test sets.

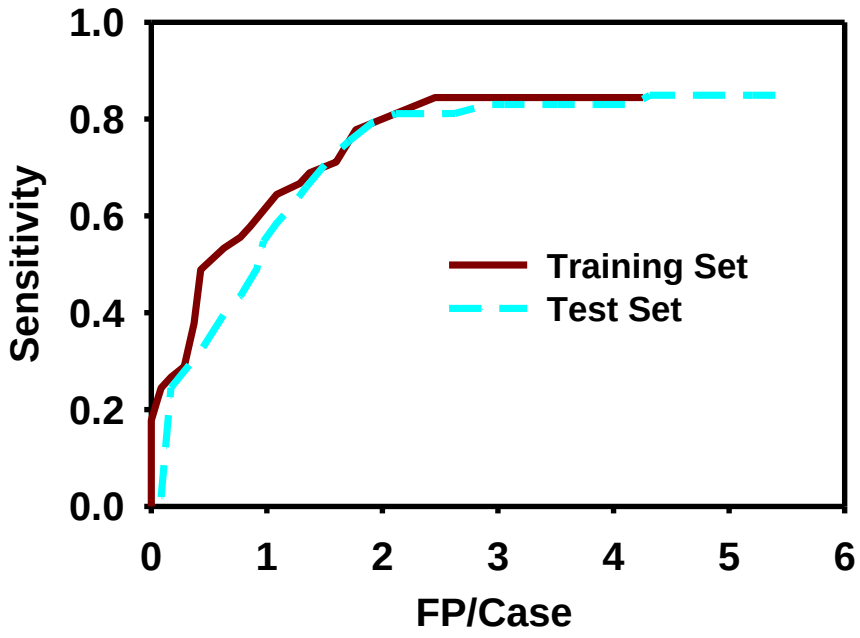


Figure 4.11: FROC curves for automatic computer detection after feature classification with LDA. After prescreening, the system achieved 84.4% sensitivity with 4.3 FPs/case for the training set, and 84.9% sensitivity with 5.4 FPs/case for the test set. After LDA classification, at 1.7 FPs/case the sensitivities were 77.8% and 75.5% for the training and test sets, respectively.

Table 4.6: Sensitivity at a given FP rate after using LDA classifier.

	False positive rate (FPs/case)					
	0.5	1.0	1.5	2.0	2.5	3.0
Training set	48.9%	64.4%	70.0%	77.8%	84.4%	84.4%
Test set	36.6%	54.7%	70.6%	80.0%	81.1%	83.0%

Examples of true positive lesions detected are shown in Figure 4.12. Figure 4.13 shows examples of false positive detections, while Figure 4.14 shows lesions that were missed.

(a) (b) (c)

Figure 4.12: Examples of detected bladder lesion. Lesions of varying sizes and shapes were correctly identified by the CAD system. (a) Small lesion located along the posterior aspect of the bladder. (b) Large lesion partially obstructing the left ureterovesical junction. (c) Lesion covering large amount of the bladder wall. All three lesions were malignant.

(a) (b)

Figure 4.13: Examples of false positives. (a) Prostate protruding onto the bladder was detected as a lesion candidate. (b) Ureterovesical junction detected as a lesion candidate. Neither was removed by the LDA classifier.

(a) (b)

Figure 4.14: Examples of lesions missed by prescreening. The inhomogeneous contrast material in both (a) and (b) prevented the prescreening steps from identifying these lesion candidates. Both were malignant lesions.

We have performed a sensitivity analysis of our CAD system by changing the ROI box size, ROI centroid and best slice locations. The size of the ROI box was changed in the range from -5% up to 20% in the X-Y direction. The centroid of the ROI box was randomly shifted by 5% in the X or Y direction. The best slice of the ROI box was randomly moved by -5% or 5%.

The sensitivity of each ROI box change was estimated at the operating points used for the original box for the training (2.5 FP/case) and test (4.3 FP/case) sets, and shown in Tables 4.7 and 8. Overall, we observed stable performance for lesion detection within the ranges of ROI box sizes and locations studied.

Table 4.7: Detection sensitivity at 2.5 FP/case for training set.

ROI Change	Original	XY -5%	XY +5%	XY +10%	XY +15%	XY +20%	Best Slice ±5%	XY Centroid ±5%
Sensitivity	84.4%	77.8%	82.2%	82.2%	86.7%	86.7%	82.2%	84.4%

Table 4.8: Detection sensitivity at 4.3 FP/case for test set.

ROI Change	Original	XY -5%	XY +5%	XY +10%	XY +15%	XY +20%	Best Slice ±5%	XY Centroid ±5%
Sensitivity	84.9%	84.9%	83.0%	81.1%	84.9%	73.6%	86.8%	84.9%

4.5 Discussion

In this study, we developed a system that can automatically detect bladder lesions of a wide range of sizes and subtleties within the contrast-enhanced region of the bladders in CTU that only requires two ROI boxes as inputs. At the pre-processing stage, the system detected over 80% of the lesions in both the training and test sets, while having about 5 false positives per case.

Our system was able to detect lesions of various shapes and sizes at different locations within the contrast-enhanced region of the bladder. Small lesions protruding out from the bladder wall, large lesions that occupy a large portion of the contrast-enhanced region, and bladder masses that look like an extension of the bladder wall were all detected by the system (Figure 4.12).

During the prescreening stage, a relative large number of false positive lesions appeared close to the cephalic or caudal most aspects of the bladder. This may be caused by the fact that the CT slices at these locations are close to or even intersect the bladder wall in some regions; the partial volume effects may contribute to the inhomogeneous and low contrast appearance of the bladder region (Figure 4.7(e)). However, the LDA classifier was able to correctly remove the two false positives shown in that example. A common false positive finding that the LDA classifier has difficulty differentiating from true lesions is the prostate in male patients. For some cases, the prostate protruding into the bladder has a similar appearance to an intrinsic bladder lesion (Figure 4.13(a)). The bladder segmentation may also leak into the prostate due to the

interface between the bladder and the prostate being difficult to distinguish. The detection system, therefore, often identifies the portions of the prostate that are segmented as lesion candidates. Another common cause of false positive findings not removed by the LDA classifier is the ureterovesical junction (UVJ). When the location near the UVJ is imaged, the ureter wall at the junction between the ureter and the bladder may appear as a protrusion of tissue from the bladder wall, similar to a lesion (Figure 4.13(b)) that the system detects as a lesion candidate. Other causes of false positives include the inhomogeneous contrast material in the urine that did not get fully removed by the adaptive thresholding and folding of the bladder wall for cases with bladder diverticula. A common cause for false negatives is the non-uniformity of the contrast material that camouflages the lesions as a part of the bladder wall (Figure 4.14).

Our system detected the majority of both malignant and benign lesions. For both the training and the test sets, more malignant lesions were detected than benign lesions. However, the number of benign lesions was much smaller than the number of malignant lesions in our data set. Therefore, we cannot draw a conclusion on whether the system detection performance is related to lesion malignancy.

With feature extraction and FP reduction by LDA, our system achieved 84.4% sensitivity with an average of 2.5 FPs/case for the training set, and 84.9% sensitivity with 4.3 FPs/case for the test set. Our radiologist and urologist co-investigators expect that a CAD system with 85-90% sensitivity with 2 FP/case for the entire bladder would be useful in their practice. We were close to reaching this goal for our training set, but the system needs to be improved to meet our performance goals for unknown cases. Further improvement of the system is reported in Chapter V.

Our system takes approximately 2 to 5 minutes to run for a case on a system with an Intel Xeon 5160 processor at 3 GHz, depending on the bladder size. It takes approximately 2 minutes per case for manual input, which includes loading the case and marking the two ROIs.

It is difficult to make a direct comparison with the previous methods by other investigators described in the Introduction due to the differences in the data sets, lesion sizes and subtlety, and the performance evaluation methods. In the study by Jaume et. al[46], they delineated each bladder in their data set into 6 different zones, and measured the performance by determining whether or not a zone was diseased or not. Duan et. al[22] measured their performance by determining the percentage in which the lesions in their data set was covered by windows

marking a region representing a part of a lesion. In comparison to our pilot study[25], in which 83% of bladder lesions were detected with 1.4 FPs/case using a data set of 15 cases, the current system had sensitivities of 84.4% and 84.9% at an FP rate of 2.5 FPs/case and 4.3 FPs/case for the training and test sets, respectively. When the system from our pilot study was applied to the larger data set used in this study, it had sensitivities of 77.8% and 50.9% at an FP rate of 2.6 FPs/case and 2.7 FPs/case for the training and test sets, respectively. Our current system achieved a higher sensitivity when using a data set of 70 patients compared to the system in our pilot study.

This study has several limitations. First, the detection method was designed for detection in the contrast-enhanced region of the bladder. As the contrast between the lesion and its surroundings is much smaller in the non-contrast region, a different method will have to be developed for detection in the non-contrast region. Second, this study was directed towards detection of bladder masses, but not bladder wall thickening. Wall thickening that was found by the system (153 candidates in the training set, and 197 candidates in the test set that involve any portion of a thickening bladder wall) was not considered to be a true lesion and was excluded as a false positive in this study. A different method may need to be developed to detect wall thickenings accurately, as wall thickenings possess different characteristics than masses. We will present methods to detect bladder lesions in the entire bladder, and bladder wall thickenings in Chapters V and VI, respectively.

4.6 Conclusion

This study demonstrates the feasibility of our method for detection of bladder lesions located fully or partially in the contrast-enhanced region of the CTU scans for lesions of a variety of shapes and sizes. The prescreening stage detected most of the true lesions, but also many false positive lesions. Using feature extraction and a trained classifier, the number of false positive findings was reduced while keeping the sensitivity high. The results indicate the usefulness of the methods for bladder lesion detection for lesions partially or fully within the contrast-enhanced region of CTU. Further work is underway to increase the sensitivity, detect lesions within the non-contrast enhanced region and detect lesions manifested as bladder wall thickening. This study is a step towards the development of a CAD system for detection of urothelial lesions in the bladder that are imaged with CT urography.

Chapter V
Automatic Detection of Urinary Bladder Mass on CT Urography within the Whole Bladder

5.1 Abstract

We have continued the development of the computer-aided detection system for bladder cancer in CT urography (CTU). We have previously developed methods for detection of bladder masses within the contrast-enhanced regions of the bladder individually as reported in Chapter IV. In this study, we investigated methods for detection of bladder masses within the entire bladder. The bladder was segmented using our method that was described in Chapter III that combined deep-learning convolutional neural network with level sets. The non-contrast-enhanced region was separated from the contrast-enhanced region with a maximum-intensity-projection-based method. The non-contrast region was smoothed and gray level threshold was applied to the contrast and non-contrast regions separately to extract the bladder wall and potential masses. The bladder wall was transformed into a straightened thickness profile that was analyzed to identify lesion candidates as a prescreening step. The candidates were segmented using our auto-initialized cascaded level set (AI-CALS) segmentation method, and 91 features were extracted for each candidate. A data set of 87 patients with 114 biopsy-proven bladder lesions was used, which was split into independent training and test sets: 47 training cases with 61 lesions, and 40 test cases with 53 lesions. Using the training set, feature selection was performed and a linear discriminant (LDA) classifier was designed to merge the selected features for classification of bladder lesions and false positives. The trained classifier was evaluated with the test set. FROC analysis showed that the system achieved a sensitivity of 96.7% at 4.4 FPs/case for the training set, and 90.6% at 4.9 FPs/case for the test set. At 2.0 FPs/case, the sensitivities were 88.5% and 84.9% for the training and test sets, respectively. The preliminary results using methods presented in this chapter have been published as conference proceedings[6]. Preparation for submission as a journal article is underway.

5.2 Introduction

This chapter describes a system for detection of bladder lesions within the whole bladder on CT urogram. When we started looking into designing a system for detection of lesions within the bladder, we observed that the segmentation of the non-contrast enhanced (NC) region using the CLASS with LCR method had many errors. We noticed that the NC segmentation would either over-segment the bladder, enclosing non-bladder structures, or under-segment the bladder, resulting in missed bladder lesions within the NC region. Therefore, we first focused on performing bladder lesion detection within the contrast-enhanced (C) region of the bladder, as demonstrated in Chapter IV. Afterwards, we developed a new method for bladder segmentation using DL-CNN with level sets, as described in Chapter III that improved the performance of the segmentations.

In this study, we developed a system for detection of bladder lesions within the entire bladder, using the bladder segmentation method described in Chapter III. We designed the system and evaluated its performance using free-response receiver operating characteristic (FROC) analysis using a data set of 87 cases.

5.3 *Materials and Methods*

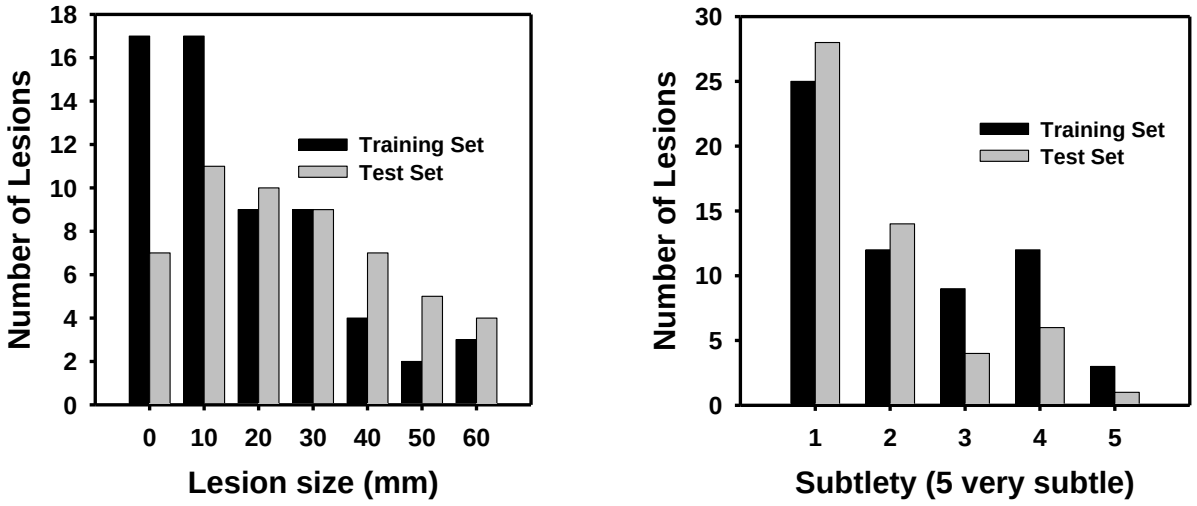
5.3.1 Data set

The data set used in this retrospective study was collected using protocols approved by the Institutional Review Board. A total of 87 patients who had undergone CTU with subsequent cystoscopy and biopsy was collected from our institution. The CTU scans were acquired with the parameters described in Chapter II, section 2.3.5.

All locations of lesions were identified by experienced radiologists in CTU volume as reference standard. Two experienced radiologists marked the lesion location by drawing a bounding box around the lesion, including the starting (most cephalad) and the ending (most caudal) slice of the lesion. They also measured the longest diameter of the lesion and gave a subtlety rating. Consensus was obtained if the lesion locations were different between the radiologists, and findings were correlated with radiology, pathology, and biopsy reports. The size and subtlety of the lesions given by the more experienced radiologist were reported to illustrate the detection performance of the system for lesions of different degrees of difficulty as seen by

an experienced radiologist. Manual hand outlines of the lesions were performed by the more experienced radiologist as a reference standard to evaluate the automatic detection system's performance.

Within the data set, 114 biopsy-proven bladder lesions were identified from 87 cases. The cases were split into independent training and test sets. All bladder lesions were found to be malignant according to biopsy reports. The training set consisted of 61 lesions from 47 cases with an average size of 22.4 mm (range: 1.4-61.7 mm) (Figure 5.1(a)), while the test set contained 53 lesions from 40 cases with an average size of 31.4 mm (range: 5.7-96.4 mm) (Figure 5.1(a)). The average lesion subtlety ratings were 2.3 for the training set, and 1.8 for the test set (scale 1 to 5, 5 very subtle) (Figure 5.1(b)).



(a) (b)

Figure 5.1: Histograms of lesion size (a) and lesion subtlety (b) for lesions in the training and test set. The average lesion size was 22.4 mm (range: 1.4–61.7 mm) for the training set, and 31.4 mm (range: 5.7–96.4 mm) for the test set. The average lesion subtlety ratings in both sets were 2.3 for the training set and 1.8 for the test set (scale 1 to 5, 5 very subtle).

5.3.2 Bladder segmentation using DL-CLASS

A critical component of CAD system that detects bladder cancer is accurate bladder segmentation that isolates the bladder from the surrounding anatomical structures, as it determines the search region for the detection process. Bladder segmentation determines the search region for the detection process. Accurate bladder segmentation will help ensure high accuracy with low number of false positive findings. We have previously developed computerized methods and a software package for segmenting the bladder in CTU that we

refer to as the deep-learning cascaded level set analysis and segmentation system (DL-CLASS) that was described in Chapter III[2]. Briefly, the DL-CLASS method uses a deep-learning convolutional neural network (DL-CNN) trained with 160,000 regions of interests (ROI) to distinguish between the area inside the bladder and the surrounding area containing other organs within the CTU slice. The trained DL-CNN is applied to each voxel of a volume of interest containing the bladder in the CTU to generate a bladder likelihood score, resulting in a bladder likelihood map for each slice. A threshold is applied to the likelihood maps that are smoothed using a morphological dilation filter and a hole-filling algorithm to generate an initial contour. The CTU slices are then pre-processed, which apply smoothing, anisotropic diffusion, gradient filters, and the rank transform of the gradient magnitude. 3D level sets, followed by 2D level sets are applied to the initial contour to obtain the final bladder segmentation. Example of a segmented bladder can be seen in Figures 5.2(a) and 5.3(a).

5.3.3 Bladder wall profile generation and lesion candidate identification with A-SPAN

Bladder lesion candidates are identified by first isolating the contrast-enhanced region of the bladder, and then by using our newly developed method, referred to as Adaptive-Straightened Periphery ANalysis (SPAN). A-SPAN consists of three stages: (1) wall thickness profile generation, (2) false positive reduction of voxel candidate, and (3) lesion candidate identification.

5.3.3.1 Finding the boundary line between the contrast-enhanced and the no-contrast regions of the bladder

The steps for distinguishing the contrast-enhanced (C) and the non-contrast enhanced (NC) regions of the bladder were previously described in Chapter IV, section 4.3.3.1 which described our work of detecting lesions within the C region of the bladder[3], and adapted to use with the DL-CLASS segmentations. Briefly, the contrast-enhanced region of the bladder is separated from the non-contrast region using the property that the dependently layering IV contrast material in the bladder will be filled to the same level consistently along all CTU slices due to gravity. We use maximum intensity projection (MIP) along the slices of the bladder to estimate the upper boundary of the contrast enhanced region. The ROIs initializing the segmentation of the C and NC regions of the bladder are used to determine the range of the CTU slices for the MIP. As the bladder is located on top of the pelvic bones when the patient

lies in a supine position, it is common that the bones intrude into the bladder's ROI towards the bottom of the bladder. If the bright bones show up on the MIP image, its interference makes it difficult to accurately determine the upper level of the contrast material. Therefore, only a portion of the slices included in the ROI is used. We experimentally determined that 10% of the slices towards the bottom of the bladder from the best slice (to avoid the pelvic bones) and 90% of the slices towards the top of the bladder from the best slice (to avoid other organs above the bladder) worked the best. The best slice is the slice that best represents the bladder region, e.g., where the bladder is seen the largest. It was selected manually when the ROIs were defined.

From the MIP image, a gray level profile is generated. The width of the input bounding box of the bladder used for the DL-CLASS segmentations is reduced by 50% while keeping the same center to minimize the negative effects of the peripheries of the irregularly shaped bladders on the estimation of the transition point between the contrast and non-contrast regions. For every row of the box, the gray levels of the pixels belonging to the row are averaged and recorded into a profile. The profile is analyzed to find the first row R_1 whose average gray level is greater than a gray level threshold Th_p. By using the training data set the Th_p was determined experimentally as 1330, which provided adequate separation of the NC and C regions for the training cases. The horizontal value B_1 at row R_1 that intersects with the bladder boundary is determined as the boundary between the NC and C regions of the bladder. The region above the horizontal value B_1 is assigned as the NC region, and the region above the line is assigned as the C region. The two regions will be analyzed with different techniques as their properties vastly differ.

5.3.3.2 Wall thickness profile generation with A-SPAN

Generating the bladder wall thickness profile involves two steps: (1) refining the bladder segmentation in the C region of the bladder, (2) adaptive thresholding of the urine to isolate the bladder wall and masses, and (3) transformation of the bladder segmentation to a straightened wall profile.

Refining the C region contour. The DL-CNN contour may under-segment the bladder in the C region, resulting in portions of lesions being outside of the bladder segmentation. To alleviate this issue, our previously developed Model-Guided Refinement (MGR) method[1] that was described in Chapter II, section 2.3.2.1, was applied to the segmentation contour points

whose Y-value is greater than the horizontal value B_1. The new, refined resulting contour will be referred as the L contour.

Adaptive thresholding of urine. Different strategies are used for the adaptive thresholding of the NC and C regions of the bladder to isolate the bladder wall from the urine within the bladder. The methods used for the C region adaptive thresholding is described in chapter IV, section 4.3.3.2[3]. For the NC region, the region is first blurred using a 7-pixel by 7-pixel average filter. As the contrast material is denser towards the bottom of the bladder, a threshold that is based on the horizontal location is implemented. The threshold value, Th_{NC}^{j} is given by the following equation:

$$Th_{NC}^{j} = \varepsilon - \tau * \left(\frac{j - Y_{NC}}{B_1 - Y_{NC}}\right) \tag{5.2}$$

where j is the Y-value of a given pixel within the NC region, B_1 is the horizontal value mentioned in section 2.3.1, and Y_{NC} is the Y-value of the upper most point (point with the smallest Y-value) of the bladder segmentation. ε and τ are constants experimentally determined to be 1210 and 410, respectively. Once Th_{NC}^{j} is determined, the urine is eliminated from the NC region by setting the gray level to 0 for all pixels whose gray level is greater than Th_{NC}^{j}.

Wall profile generation. Once the urine is removed from the CTU slice, a straightened profile of wall thickness is generated by mapping all of the points along the L contour, L_i, $i=1, \ldots, n$, sequentially to the X-axis of a new coordinate system such that $L_i(x_i, y_i)$ has the coordinate $(X_i,0)$. The origin of this new coordinate system is defined at the top left of the profile, with Y-values increasing in the downward direction. For a given pixel $L_i(x_i, y_i)$, the path normal to the point towards the interior of the L contour is calculated using the normal angle θ defined as:

$$\theta = 90^{\circ} + \frac{1}{2}\left(\tan^{-1}\left(\frac{y_i - y_{i-1}}{x_i - x_{i-1}}\right) + \tan^{-1}\left(\frac{y_{i+1} - y_i}{x_{i+1} - x_i}\right)\right), \tag{5.3}$$

where (x_{i+1}, y_{i+1}) and (x_{i-1}, y_{i-1}), respectively, are the coordinates of the next L_{i+1} and previous L_{i-1} neighboring points of L_i,. The pixels along the normal path are sequentially mapped onto the profile at increasing Y-values such that the new coordinates of the pixels are given by (X_i, Y_j) $j=1, \ldots i_m$, while X_i is fixed. The path along the normal ends when 16 black pixels are encountered consecutively, indicating that the path reaches the lumen of the bladder where the pixel gray level has been set to 0. The number of pixels along the normal path at L_i is denoted by i_m.

(a) (b) (c)

Figure 5.2: Detection of bladder mass: example of mass in NC region. (a) Bladder segmentation. (b) Wall profile after thresholding to remove the urine within the entire bladder. The arrows indicate bladder lesion candidates. (c) Detected bladder mass candidates mapped back to CTU image.

(a) (b) (c)

Figure 5.3: Detection of bladder mass: example of mass in C region. (a) Bladder segmentation. (b) Wall profile after thresholding to remove the urine within the entire bladder. The arrows indicate bladder lesion candidates. The Orange arrow points to false positives, while the blue arrow points to a bladder lesion. (c) Detected bladder mass candidates mapped back to CTU image.

5.3.3.3 False positive reduction of voxel candidate with A-SPAN

Sharp peaks along the bladder wall profile that are likely caused by noise are removed. Since noise is random and forms small peaks, we set a criterion to exclude peaks less than three pixels in width and greater than five pixels in height relative to its surroundings.

One of the main false positive (FP) findings is organs situated just below the bladder, which include the prostate and the uterus. These possible false positive candidate pixels are removed based on their location mapped back to the original CTU slices, and the presence of other candidates near their locations. As the prostate and the uterus is generally located near the horizontal center of the CTU slice, candidate pixels whose x-value on the original CTU slice is within 30 pixels of the center of the CTU slice is studied as possible FP. If there are at least 15 other candidate pixels within a 3-pixel by 3-pixel region centered on the studied candidate pixel location from all the CTU slices below in the axial direction, towards the caudal aspect of the bladder, the studied candidate pixel is considered as an FP, and is removed. As true positive lesions can be present on the center of the bladder as well, and these parameters were selected using the training set to remove the false positive findings while retaining the true positive pixel

candidates. Examples of A-SPAN for a true lesion in the NC region, and a lesion in the C region are shown in Figures 5.2 and 5.3, respectively.

5.3.3.4 Lesion candidate identification with SPAN

Lesion candidates are found by analyzing the bladder wall profile. After the FP removal step, all remaining pixels for a given bladder profile are considered to be lesion candidate pixels (Figure 5.2(b), 5.3(b)). After the lesion candidate pixels are identified, they are mapped back to the original bladder slices as lesion candidate voxels (Figure 5.2(c), 5.3(c)).

The lesion candidate voxels are then grouped into regions. Candidate voxels that are one slice apart, and within two voxels apart in 3D space are clustered into the same region. A candidate voxel is ignored if there are no other candidate voxels within two voxels on the same slice. Regions that contain less than five candidate voxels or greater than 50,000 candidate voxels are ignored. This size range was determined by analyzing the training cases. Each of the regions retained is enclosed with a 3D bounding box to indicate an ROI for a lesion candidate. Additional possible FP candidate regions are removed based on size and gray level of the regions, determined using the training set.

5.3.4 Lesion candidate segmentation, feature extraction, and classification

5.3.4.1 Lesion segmentation with Auto-Initialized Cascaded Level Set (AI-CALS)

Using the ROI obtained for a lesion candidate, the lesion is segmented from the surrounding tissue using the AI-CALS segmentation system, described in Chapter IV, section 4.3.4.1[3]. The AI-CALS segmentation system is briefly described in the following. The AI-CALS consists of three stages: preprocessing, initial segmentation, and level set segmentation. In the first stage, preprocessing techniques are applied in 3D to the ROI obtained from the candidate prescreening process above. In the second stage, the system generates an initial segmentation based on the attenuation, gradient, and location of the voxels. In the third stage, the initial segmentation surface is propagated towards the lesion boundary using cascading level sets. Further details on the AI-CALS method can be found in the literature[3, 47].

5.3.4.2 Feature extraction and classification

For each segmented lesion candidate, 91 features are automatically extracted from the central slice of the segmented lesion. The definitions of these features can be found in the literature[30, 50, 51]. These features, which include: morphological features, such as volume, circularity, rectangularity, and Fourier descriptor; (2) gray level features, such as the average gray level and contrast features; (3) texture features, such as run length statistics, and (4) gradient field features, such as the gradient magnitudes statistics for all voxels on the surface of the segmented lesion; are studied because we found that these features are useful for lesion classification from our previous experience with breast masses and lung nodules.

Using the training set, stepwise feature selection is used to select the best feature subset. Using leave-one-out case method on the training set, a combination of four features is selected. The four features consist of two run-length statistics feature (texture feature), an average radial length feature (morphological feature), and a contrast feature (gray level feature). A linear discriminant (LDA) classifier was then designed with the training set for classification of the bladder lesions and false positives using the selected features as input predictor variables. The trained LDA classifier was applied to the test set for independent testing.

5.3.5 Evaluation Methods

The performance of the lesion candidate prescreening steps was evaluated by determining the sensitivity and specificity. The overall performance of the bladder lesion detection CAD system with feature extraction and FP reduction by the LDA classifier was evaluated using free-response receiver operator characteristics (FROC) analysis. The FROC curve was generated by varying the decision threshold for the LDA discriminant scores.

5.4 Results

At the prescreening step, our system achieved 96.7% (59/61) sensitivity with an average of 4.4 false positives per case (FPs/case) for the training set, and 90.6% (48/53) sensitivity with 4.9 FPs/case for the test set. The prescreening step generated 325 lesion candidates for the training set (118 true lesion candidates, 207 false positive lesion candidates) that were used to train the LDA classifier.

Tables 5.1 and 5.2 summarize the detected lesions during prescreening by size and subtlety, respectively. For lesions smaller than 10 mm, 94.4% (17/18) and 85.7% (6/7) were

found by the system in the training and test sets, respectively (Table 5.1). For lesions with subtlety ratings greater than 2, 94.4% (17/18) of them were detected in the training set, while 90.9% (10/11) of them were detected in the test set (Table 5.2).

Table 5.1: Detected lesions at the prescreening stage for lesions of different sizes.

	Lesion Size (mm)						
	0-10	10-20	20-30	30-40	40-50	50-60	60-70
Training set	94.4% (17/18)	100% (11/11)	87.5% (7/8)	100% (6/6)	100% (4/4)	100% (2/2)	100% (3/3)
Test set	85.7% (6/7)	91.7% (11/12)	77.8% (7/9)	100% (6/6)	50% (1/2)	100% (2/2)	100% (1/1)

Table 5.2: Detected lesions at the prescreening stage for lesions of different subtleties.

	Lesion Subtlety (1-5, 5 very subtle)				
	1	2	3	4	5
Training set	95% (19/20)	100% (14/14)	85.7% (6/7)	100% (8/8)	100% (3/3)
Test set	88.9% (16/18)	80% (8/10)	100% (4/4)	83.3% (5/6)	100% (1/1)

Using feature extraction, LDA classifier, and by varying the threshold for the LDA scores, the FROC curve was generated as shown in Figure 5.4. At 4 FPs/case, the training set achieved a sensitivity of 95.0%, while the test set achieved 86.8%. At 2 FPs/case, the sensitivities were 88.5% and 84.9% for the training and test sets, respectively. Table 5.3 shows the sensitivities of the system at different false positive rates for the training and test sets.

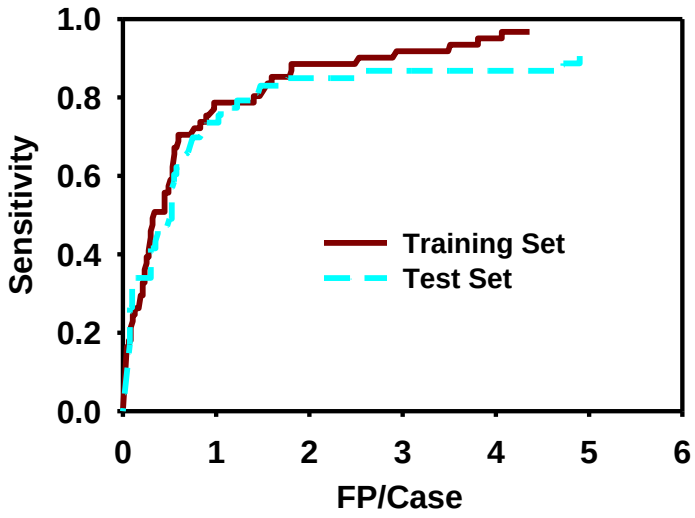
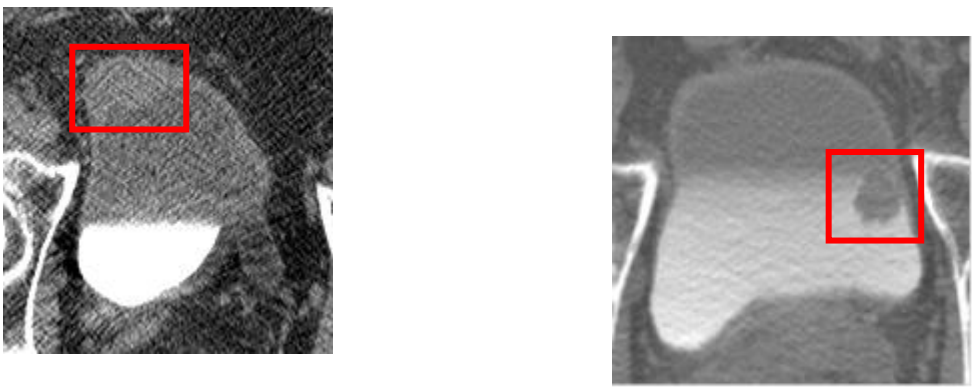


Figure 5.4: FROC curves for automatic computer detection after feature classification with LDA. After prescreening, the system achieved 91.8% sensitivity with 4.4 FPs/case for the training set, and 90.6% sensitivity with 4.9 FPs/case for the test set. After LDA classification, at 2 FPs/case the sensitivities were 90.2% and 84.9% for the training and test sets, respectively.

Table 5.3: Sensitivity at a given FP rate after using LDA classifier.

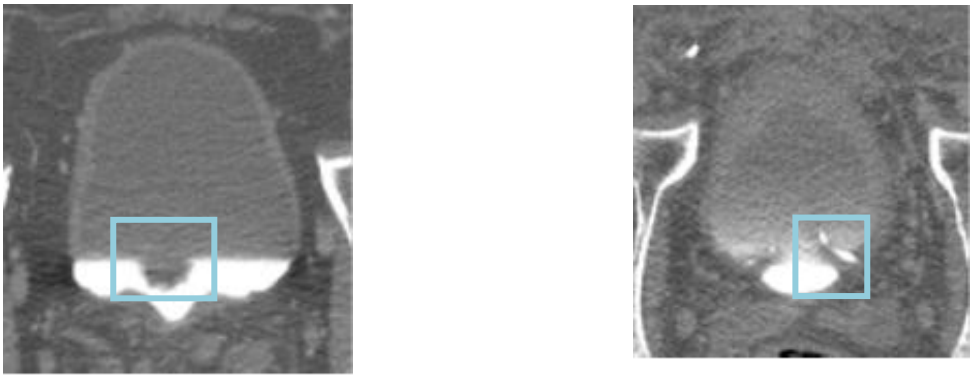
	False positive rate (FPs/case)					
	0.5	1.0	1.5	2.0	2.5	3.0
Training set	57.4%	78.7%	82.0%	88.5%	90.2%	91.8%
Test set	49.1%	73.6%	83.0%	84.9%	84.9%	86.8%

Examples of true positive lesions detected are shown in Figure 5.5. Figure 5.6 shows examples of false positive detections, while Figure 5.7 shows a lesion that was missed.



(a) (b)

Figure 5.5: Examples of detected bladder lesions (true positives). Lesions within both the contrast-enhanced and non-contrast regions of the bladder were correctly identified by the CAD system. (a) Lesion located in the non-contrast region. (b) Detected lesion within the contrast-enhanced bladder.



(a) (b)

Figure 5.6: Examples of false positives. (a) Prostate protruding onto the bladder was detected as a lesion candidate. (b) Ureterovesical junction detected as a lesion candidate. Neither was removed by the LDA classifier.

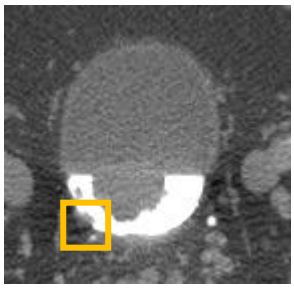


Figure 5.7: Example of lesion missed by prescreening (false negative). The relatively flat lesion near other focal lesions prevented the prescreening steps from identifying this lesion candidate, which was a malignant lesion.

5.5 Discussion

In this study, we developed a system that can automatically detect bladder lesions of a wide range of sizes and subtleties in CTU that only uses an ROI box as manual input. At the pre-processing stage, the system detected over 90% of the lesions in both the training and test sets, while having about 4 to 5 false positives per case. Our radiologist and urologist co-investigators expect that a CAD system with 85-90% sensitivity with 2 FP/case for the entire bladder would be useful in their practice. We were able to reach this goal with the methods proposed in this study.

A common false positive finding that the LDA classifier has difficulty to differentiate from true lesions is the prostate in male patients. For some cases, the portion of an enlarged prostate that protrudes into the bladder has a similar appearance to an intrinsic bladder lesion (Figure 5.6(a)). The bladder segmentation may also leak into the prostate due to the interface between the bladder and the prostate being difficult to distinguish. The detection system, therefore, identifies the portions of the prostate that are segmented as lesion candidates. We attempted to remove possible candidate pixels to try and remove some of the FPs, but we were not able to remove all of them. Another common cause of false positive findings not removed by the LDA classifier is the ureterovesical junction (UVJ). When the location near the UVJ is

imaged, the ureteral wall at the junction between the ureter and the bladder may appear as a protrusion from the bladder wall, similar to a lesion (Figure 5.6(b)) that the system detects as a lesion candidate. Other causes of false positives include the inhomogeneous contrast material in the urine that did not get fully removed by the adaptive thresholding and folding of the bladder wall for cases with bladder diverticula.

Compared to the methods used in Chapter IV for the detection of bladder masses only within the contrast-enhanced region of the bladder, we expect that the new method performs better, as the bladder segmentation using DL-CNN performs better than the segmentation using CLASS with LCR. We observe this trend when looking at the test set pre-screening performances (90.6% at 4.9 FPs/case for whole bladder, and 84.9% at 5.4 FPs/case for C region only); however, it is difficult to make a direct comparison, as the data set between the two studies are slightly different.

Using other machine-learning classifiers with the selected features performed similarly as the LDA classifier in the removal of false positive detections. The LDA, support vector machine, back-propagation neural network, and the random forest classifiers all performed similarly, with minor differences at some operating points, ranging from 87% to 89% sensitivity at 4 FPs/case, and 85% to 83% at 2 FPs/case.

This study has several limitations. First, the data set consists of 87 cases and it is relatively small, with 40 cases being used as the test set. Our method should be evaluated on a much larger data set to study the feasibility of our method. Second, this study was directed towards detection of bladder masses, but not bladder wall thickening. Wall thickening that was found by the system (290 candidates in the training set, and 221 candidates in the test set that involve any portion of a thickening bladder wall) was not considered to be a true lesion and was excluded as a false positive in this study. A different method may need to be developed to detect wall thickening accurately, as wall thickening possesses different characteristics than masses. We develop methods to detect bladder wall thickenings that will be introduced in the next chapter, Chapter VI.

5.6 Conclusion

This study demonstrates the feasibility of our method for detection of bladder lesions on CTU for bladder lesions of a variety of shapes and sizes, in both the contrast-enhanced and non-

contrast-enhanced regions of the bladder. The prescreening stage detected most of the true lesions, but also many false positive lesions. Using feature extraction and a trained classifier, the false positives were reduced while keeping the sensitivity high. The results indicate the usefulness of the methods for bladder lesion detection in CTU. Further work is underway to increase the sensitivity. This study is a step towards the development of a CAD system for detection of urothelial lesions imaged with CT urography.

Chapter VI
Bladder Wall Thickening Detection in CTU

6.1 Abstract

We are developing a computer-aided detection system for bladder cancer in CT urography (CTU). We have described developed methods for detection of focal bladder masses in Chapters IV and V. Bladder wall thickening is another manifestation of bladder cancer. In this study, we investigate methods for detection of bladder wall thickenings. The inner and outer bladder walls were segmented using our method that combined deep-learning convolutional neural network with level sets. The non-contrast-enhanced region was separated from the contrast-enhanced region with a maximum-intensity-projection-based method. The non-contrast region was smoothed and gray level threshold was applied to the contrast and non-contrast regions separately to extract the bladder wall and potential masses. The bladder wall was transformed into a straightened thickness profile that was analyzed to identify regions of wall thickenings. A data set of 112 patients, 87 with wall thickening, and 25 with normal bladders was used that was split into independent training and test sets: 57 training cases, of which 44 had bladder wall thickening, and 13 of which had normal bladders, and 55 test cases of which 43 had wall thickening and 12 had normal bladders. The volume of the wall thickening candidate was used as a feature to build an LDA classifier for the classification of bladder wall thickenings and false positives. The trained classifier was evaluated with the test set. FROC analysis showed that the system achieved a sensitivity of 93.2% at 2 FPs/case for the training set, and 88.4% at 2 FPs/case for the test set. Preparation for submission as a journal article is underway at the time of this dissertation.

6.2 Introduction

CT urography can detect cancers that produce only thickening of the urothelium without any associated abnormality of the urinary tract lumen. This phenomenon was not widely known to exist until CT urography began to appear. CT urography may also detect bladder lesions that

were missed by cystoscopy. We have described developed methods for detection of bladder masses in Chapters IV and V. Bladder wall thickening is another indication for the presence of bladder cancer. In this study, we investigated methods for detection of bladder wall thickenings. We designed the system and evaluated its performance using free-response receiver operating characteristic (FROC) analysis.

6.3 Materials and Methods

6.3.1 Data set

The data set used in this retrospective study was collected using protocols approved by the Institutional Review Board. A total of 112 patients who had undergone CTU with subsequent cystoscopy and biopsy was collected from our institution. The CTU scans were acquired the protocol described in Chapter II, section 2.3.4.

All locations of the wall thickening were identified by experienced radiologists in a CTU volume and this was used as a reference standard. Two experienced radiologists marked the thickening location by drawing a bounding box around the thickening, including the starting (most cephalad) and the ending (most caudal) slice of the lesion. Manual hand outlines of inner and outer bladder walls were performed by the more experienced radiologist as a reference standard to evaluate the automatic detection system's performance.

Within the data set, 87 patients who had bladder wall thickenings were identified from 112 a total of cases, which also included 25 patients with normal bladders. The cases were split into independent training and test sets. The training set consisted of 44 cases with bladder wall thickenings and 13 normal bladder cases, while the test set contained 43 cases with bladder wall thickenings and 12 normal bladder cases.

6.3.2 Bladder inner and outer wall segmentation using DL-CNN

We have previously developed computerized methods and a software package for segmenting the inner and outer bladder wall in CTU[52]. The method uses a deep-learning convolutional neural network (DL-CNN) trained with 240,000 regions of interests (ROI) to distinguish between the area inside the bladder wall and the surrounding areas including the bladder lumen and other organs within the CTU slice. The trained DL-CNN is applied to each voxel of a volume of interest containing the bladder in the CTU to generate a bladder

likelihood score, resulting in a bladder wall likelihood map for each slice. A threshold is applied to the likelihood maps, which are smoothed using a morphological dilation filter and a hole-filling algorithm to generate an initial contour. The CTU slices are then pre-processed, which apply smoothing, anisotropic diffusion, gradient filters, and the rank transform of the gradient magnitude. 3D level sets, followed by 2D level sets, with two separate parameters for the inner and outer wall segmentations, are applied to the initial contour to obtain the final bladder inner and outer wall segmentation. Example of a segmented bladder wall can be seen in Figure 6.1(a). More details about our method can be found in literature[52].

6.3.3 Bladder wall profile generation and wall thickening candidate identification

Bladder wall thickening candidates are identified by first isolating the contrast-enhanced region of the bladder, and analyzing the bladder wall, which consists of the following steps: (1) wall thickness profile generation, (2) false positive reduction of voxel candidate, and (3) wall thickening candidate identification.

6.3.3.1 Finding the boundary line between the contrast-enhanced and the no-contrast regions of the bladder

The contrast-enhanced (C) or opacified, and the non-contrast-enhanced (NC) regions, or unopacified, of the bladder were separated using methods described in Chapter V, section 5.3.3.1 to obtain the horizontal value B_1 that is the boundary between the NC and C regions of the bladder. The region above the horizontal value B_1 is assigned as the NC region, and the region above the line is assigned as the C region. The two regions will be analyzed with different techniques as their properties vastly differ.

6.3.3.2 Wall thickness profile generation

Generating the bladder wall thickness profile involves two steps: (1) refining the bladder segmentation in the C region of the bladder, (2) adaptive thresholding of the urine to isolate the bladder wall and masses, and (3) transformation of the bladder segmentation to a straightened wall profile.

Refining the C region contour. The DL-CNN contour may under-segment the bladder in the C region, resulting in portions of lesions being outside of the bladder segmentation. To

alleviate this issue, our previously developed Model-Guided Refinement (MGR) methods that is described in Chapter II, section 2.3.2.1[1], was applied to the segmentation contour points whose Y-value is greater than the horizontal value B_1. The new, refined resulting contour will be referred as the L contour.

Adaptive thresholding of urine. Different strategies are used for the adaptive thresholding of the NC and C regions of the bladder to isolate the bladder wall from the urine within the bladder. The methods were previously described in Chapter V, section 5.3.3.2. In brief, the C region uses the gray level information for setting the threshold, while the NC region uses a location-based method to set the threshold for removing the urine within the bladder. Once the threshold is determined, the urine is eliminated from the NC and C regions by setting the gray level to 0 for all pixels whose gray level is greater than Th_{NC}^{j}. Additionally, any pixels that are contained within the inner bladder wall segmentation also had their gray level set to 0.

Wall profile generation. Once the urine is removed from the CTU slice, a straightened profile of wall thickness is generated by mapping all of the points along the L contour, L_i, $i=1, \ldots, n$, sequentially to the X-axis of a new coordinate system such that $L_i(x_i, y_i)$ has the coordinate $(X_i,0)$. The origin of this new coordinate system is defined at the top left of the profile, with Y-values increasing in the downward direction. For a given pixel $L_i(x_i, y_i)$, the path normal to the point towards the interior of the L contour is calculated using the normal angle θ defined as:

$$\theta = 90^{\circ} + \frac{1}{2}\left(\tan^{-1}\left(\frac{y_i - y_{i-1}}{x_i - x_{i-1}}\right) + \tan^{-1}\left(\frac{y_{i+1} - y_i}{x_{i+1} - x_i}\right)\right), \quad (6.3)$$

where (x_{i+1}, y_{i+1}) and (x_{i-1}, y_{i-1}), respectively, are the coordinates of the next L_{i+1} and previous L_{i-1} neighboring points of L_i,. The pixels along the normal path are sequentially mapped onto the profile at increasing Y-values such that the new coordinates of the pixels are given by (X_i, Y_j) $j=1, \ldots i_m$, while X_i is fixed. The path along the normal ends when 16 black pixels are encountered consecutively, indicating that the path reaches the lumen of the bladder where the pixel gray level has been set to 0. The number of pixels along the normal path at L_i is denoted by i_m. Pixels that will be removed by the following false reduction step are also marked on the figures.

(a) (b) (c)

Figure 6.1: Detection of bladder wall thickening. (a) Bladder wall segmentation. The light blue contour represents the outer bladder wall segmentation, while the dark blue contour represents the inner bladder wall segmentation. (b) Wall profile after thresholding to remove the urine within the entire bladder. The arrows indicate location where candidate points are mapped back to the CTU image. (c) Detected bladder wall thickening candidates mapped back to CTU image after region growing.

6.3.3.3 False positive reduction of voxel candidate

Sharp peaks along the bladder wall profile that are likely caused by noise are removed. Since noise is random and forms small peaks, we set a criterion to exclude peaks less than three pixels in width and greater than five pixels in height relative to its surroundings.

6.3.3.4 Thickening candidate identification

Wall thickening candidates are found by analyzing the bladder wall profile. For a given bladder profile, the histogram of the height of the profile is calculated, using a binning of 8 pixels for the height. A height threshold in pixels is set by taking the binning value of the two bins greater than the peak of the histogram. Figure 6.2 shows the histogram for the wall profile shown in Figure 6.1(b).

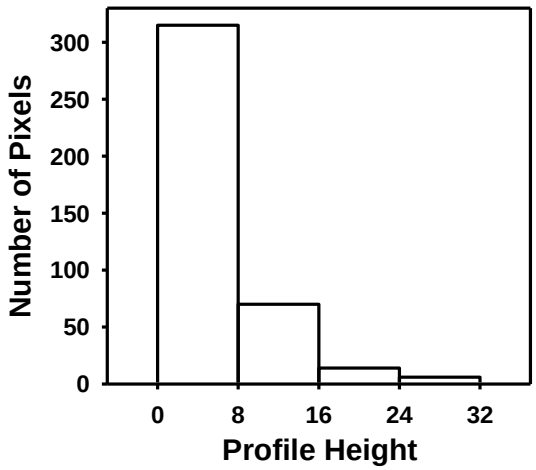


Figure 6.2: Histogram of the wall profile shown in Figure 6.1(a). The peak of the histogram is located at a pixel height of 0 for this slice; therefore, the height threshold was set to 16 pixels, which is two bins greater than the peak.

After the wall thickening candidate pixels are identified, they are mapped back to the original bladder slices as lesion candidate voxels. Region growing is performed on the candidates such that they will grow into a target touching region within the image if the target region is within the bladder wall segmentation. Figure 6.1(c) shows an example of candidates that were found and were grown.

The thickening candidate voxels are then grouped into regions. Candidate voxels that are one slice apart, and within two voxels apart in 3D space are clustered into the same region. A candidate voxel is ignored if there are no other candidate voxels that are within two voxels on the same slice. Regions that contain less than five candidate voxels or greater than 50,000 candidate voxels are ignored. This size range was determined by analyzing the training cases.

6.3.4 Thickening feature extraction and classification

The volumes of each of the regions were calculated in cubic millimeters to use as features for classification purposes. A linear discriminant analysis (LDA) classifier was then designed with the training set for classification of the bladder lesions and false positives using the volume features as input predictor variables. The trained LDA classifier was applied to the test set for independent testing.

6.3.5 Evaluation methods

The performance of the lesion candidate prescreening steps was evaluated by determining the sensitivity and specificity. The overall performance of the bladder wall thickening detection CAD system with feature extraction and FP reduction by the LDA classifier was evaluated using free-response receiver operator characteristics (FROC) analysis. The FROC curve was generated by varying the decision threshold for the LDA discriminant scores.

6.4 Results

At the prescreening step, our system achieved 93.2% (41/44) sensitivity with an average of 2.6 false positives per case (FPs/case) for the training set, and 88.4% (38/43) sensitivity with 3.4 FPs/case for the test set. Using feature extraction, LDA classifier and by varying the threshold for the LDA scores, the FROC curve was generated as shown in Figure 6.3. At 2 FPs/case, the sensitivity of both training and testing set did not change. At 1 FPs/case, the

sensitivities were 93.2% and 88.4% for the training and test sets, respectively. Table 6.6 shows the sensitivities of the system at different false positive rates for the training and test sets.

Table 6.1: Sensitivity at a given FP rate after using LDA classifier.

	False positive rate (FPs/case)			
	0.5	1.0	1.5	2.0
Training set	90.9%	93.2%	93.2%	93.2%
Test set	88.4%	88.4%	88.4%	88.4%

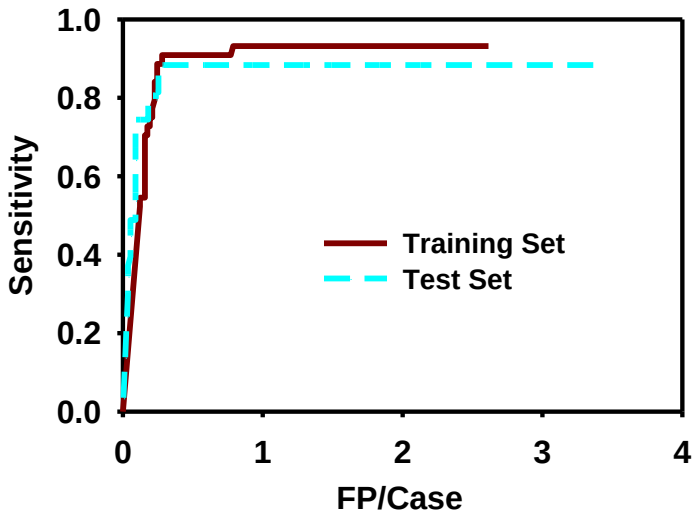
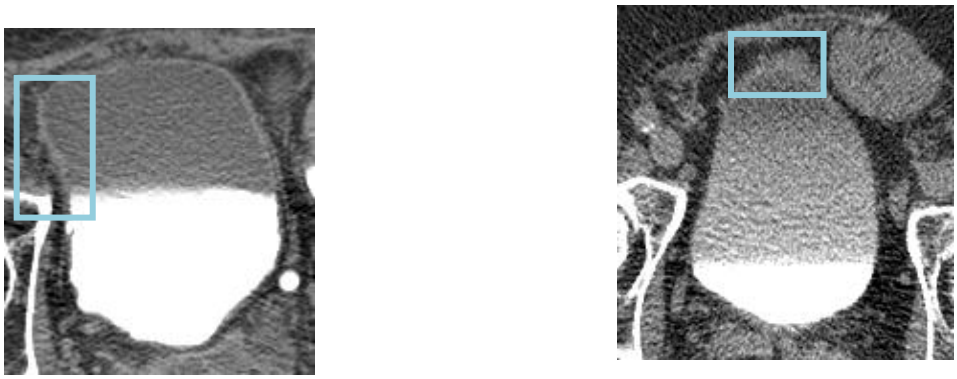


Figure 6.3: FROC curves for automatic computer detection after feature classification with LDA. After prescreening, the system achieved 93.2% sensitivity with 2.6 FPs/case for the training set, and 88.4% sensitivity with 3.4 FPs/case for the test set. After LDA classification, at 1 FPs/case the sensitivities were 93.2% and 88.4% for the training and test sets, respectively.

Examples of true positive lesions detected are shown in Figure 6.4. Figure 6.5 shows examples of false positive detections, while Figure 6.6 shows a thickening that was missed.

(a) (b)

Figure 6.4: Examples of detected areas of bladder wall thickening (true positives). (a) Wall thickening located in the non-contrast region. (b) Detected wall thickening partially within the contrast-enhanced region of the bladder.



(a) (b)

Figure 6.5: Examples of false positives. (a) Slightly thicker wall compared to the remainder of the bladder wall is present in a normal bladder. (b) Bladder wall looks thicker due to volume averaging. Neither was removed by the LDA classifier.

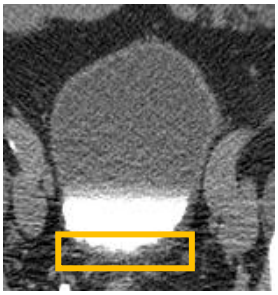


Figure 6.6: Example of bladder wall thickening missed by prescreening (false negative). The relatively flat wall thickening on the bottom of the bladder prevented the prescreening steps from identifying this lesion candidate, which was malignant.

6.5 Discussion

In this study, we developed a system that can automatically detect bladder wall thickenings using only an ROI box as input. At the pre-processing stage, the system detected most of the thickenings in both the training and test sets, while having about 3 false positives per case.

Common false positive findings were due to the slightly thicker-looking bladder walls that the radiologists did not marked as an abnormality, which was sometimes produced by volume averaging with adjacent structures (Figure 6.5). A common cause of false negative finding was due to only minimally increased wall thickening along the posterior aspect of the bladder (Figure 6.6).

This study has several limitations. First, the reference standard used for this study was the radiologists marking the locations of bladder wall thickening, which may include regions that do not represent true bladder cancers. In the future, the system should be tuned to detect only bladder wall thickening with true pathology. Second, while the data set consists of 112 cases, it is relatively small. The method should be evaluated on a much larger data set to study the feasibility of our method. Third, this study was directed towards detection of bladder wall thickenings, but not focal bladder masses. As we have developed a system for detection of bladder masses, we plan on combining our two methods and developing a system for full detection of bladder abnormalities as components of a complete CAD system for detection of bladder cancer.

6.6 Conclusion

This study demonstrates the feasibility of our method for detection of areas of bladder wall thickenings in CT urogram studies. The prescreening stage detected most of the suspicious areas of thickening, but there were some false positive results. Using a trained classifier with volume of the detected region as the input feature, the false positives were reduced while keeping the sensitivity high. The results indicate the usefulness of the methods for bladder wall thickening detection in CTU. This study is a step towards the development of a CAD system for detection of urothelial wall thickenings imaged with CT urography.

Chapter VII
Urinary Bladder Cancer Staging in CT Urography using Machine Learning

7.1 Abstract

We are evaluating the feasibility of using an objective computer aided system to assess bladder cancer stage in CT Urography (CTU). A data set consisting of 84 bladder cancer lesions from 76 CTU cases was used to develop the computerized system for bladder cancer staging based on machine learning approaches. The cases were grouped into two classes based on pathological stage ≥T2 or below T2, which is the decision threshold for neoadjuvant chemotherapy treatment clinically. There were 43 cancers below stage T2 and 41 cancers at stage T2 or above. All 84 lesions were automatically segmented using our previously developed auto-initialized cascaded level sets (AI-CALS) method. Morphological and texture features were extracted. The features were divided into subspaces of morphological features only, texture features only, and a combined set of both morphological and texture features. The data set was split into Set 1 and Set 2 for two-fold cross validation. Stepwise feature selection was used to select the most effective features. A linear discriminant analysis (LDA), a neural network (NN), a support vector machine (SVM), and a random forest (RAF) classifier were used to combine the features into a single score. The classification accuracy of the four classifiers was compared using the area under the receiver operating characteristic (ROC) curve (A_z). Based on the texture features only, the LDA classifier achieved a test A_z of 0.91 on Set 1 and a test A_z of 0.88 on Set 2. The test A_z of the NN classifier for Set 1 and Set 2 were 0.89 and 0.92, respectively. The SVM classifier achieved test A_z of 0.91 on Set 1 and test A_z of 0.89 on Set 2. The test A_z of the RAF classifier for Set 1 and Set 2 was 0.89 and 0.97, respectively. The morphological features alone, the texture features alone, and the combined feature set achieved comparable classification performance. The predictive model developed in this study shows promise as a classification tool for stratifying bladder cancer into two staging categories: greater than or equal to stage T2 and below stage T2. The preliminary results using methods presented in this chapter have been

published as conference proceedings[7]. A journal article has been submitted and is under revision at the time of this dissertation.

7.2 Introduction

The initial treatment for bladder cancer is transurethral resection of the bladder tumor (TURBT) that removes the tumor from the bladder and also helps provide information regarding the stage of the cancer[53-55]. Bladder cancer is staged in order to determine treatment options and estimate a prognosis for the patient. Accurate staging provides the physician with information about the extent of the cancer. The tumor stages T refer to the depth of the penetration of the tumor into the layers of the bladder. T0 indicates no primary tumor, T1 indicates that the tumor has invaded the connective tissue under the epithelium, T2 indicates that the tumor has invaded the bladder muscle, T3 indicates that the tumor has invaded the fatty tissue around the bladder, and T4 indicates that the tumor has spread beyond the fatty tissue into other areas such as the pelvic wall, uterus, prostate or abdominal wall[12] (Figure 7.1). An example of bladder cancer stage T2 is presented in Figure 7.2.

Accurate staging of bladder cancer is crucial to providing proper treatment to the patient. Superficial diseases (under stage T2) can be managed with less aggressive treatment than invasive diseases (stage T2 and above)[53-55]. There are two types of staging for bladder cancer - clinical and pathological. The clinical stage is the physicians' best estimate for the extent of the cancer based on physical exams and imaging. The pathological stage is determined by analysis of the tissue collected from the cancer after biopsy, tumor resection or bladder cystectomy. The accuracy of the staging depends on the complete resection of the tumor. Incomplete resection of the tumor may reduce the reliability of the staging at the beginning of the tumor management process[56]. Cystectomy ensures that the entire bladder tumor is present for pathological review; therefore, the pathological staging is based on the histological review of the cystectomy specimen[57]. Adjuvant chemotherapy is used in patients with locally advanced bladder cancer in order to reduce the chances of cancer recurrence following radical cystectomy[58]. Neoadjuvant chemotherapy is used prior to radical cystectomy in order to reduce the tumor size before surgical removal; for example, a cisplatin-based regimen has been shown to decrease the probability of finding extravesical disease and improves survival when compared to radical cystectomy alone[58-60].

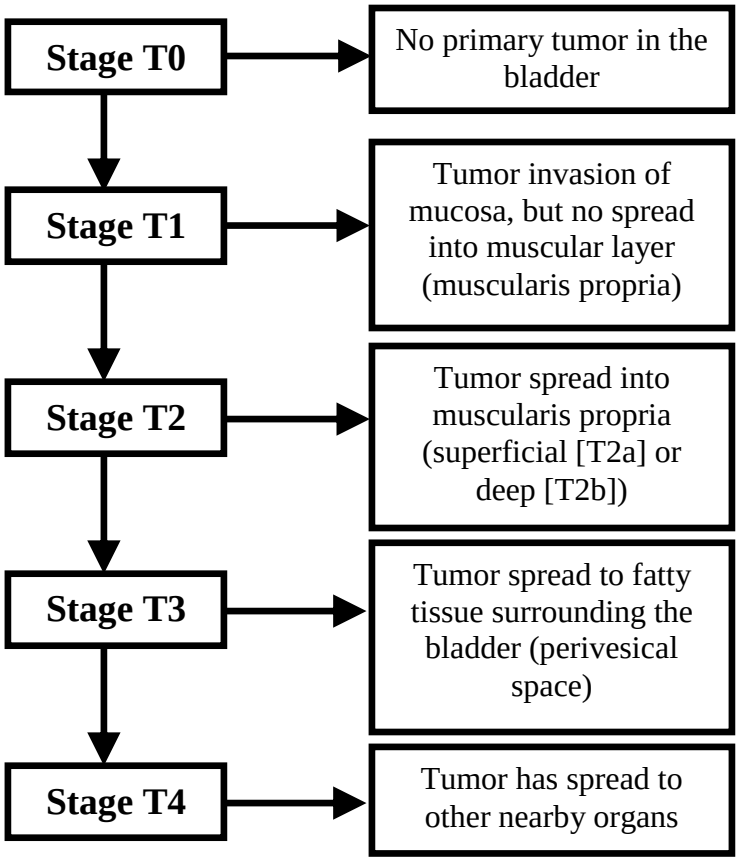


Figure 7.1: Bladder cancer stage grading scale definition.

Correct staging of bladder cancer is crucial for the decision of neoadjuvant chemotherapy treatment and minimizing the risk of under-treatment or over-treatment. Many patients with stage T2 to T4 carcinomas of the bladder are often referred for neoadjuvant chemotherapy.

Studies found that up to 50% of the patients who are estimated to have a T1 disease at clinical staging are under-staged and later upstaged after radical cystectomy[61-64]. This inaccuracy in staging can partly be attributed, in part, to the possibility that invasive portions of a bladder tumor may not be sampled during TURBT. Additionally, limitations in the ability of CT and MRI in detecting perivescial spread of tumor are well known.

The purpose of this study is to develop an objective decision support system that can potentially reduce the risk of under-treatment or over-treatment by merging radiomic information in a predictive model using statistical outcomes and machine learning.

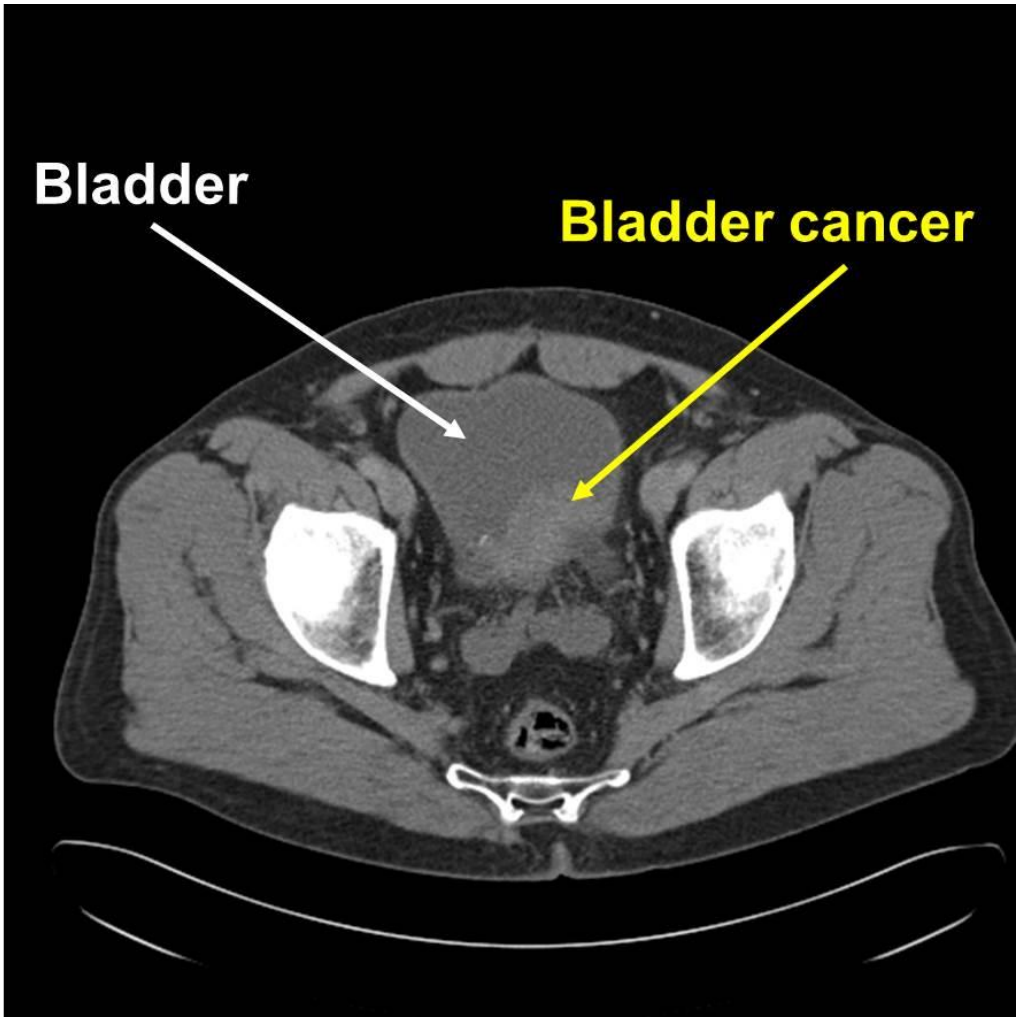


Figure 7.2: Urinary Bladder CT. The bladder cancer is marked and clearly visible. The cancer stage is T2.

7.3 Materials and Methods

7.3.1 Data set

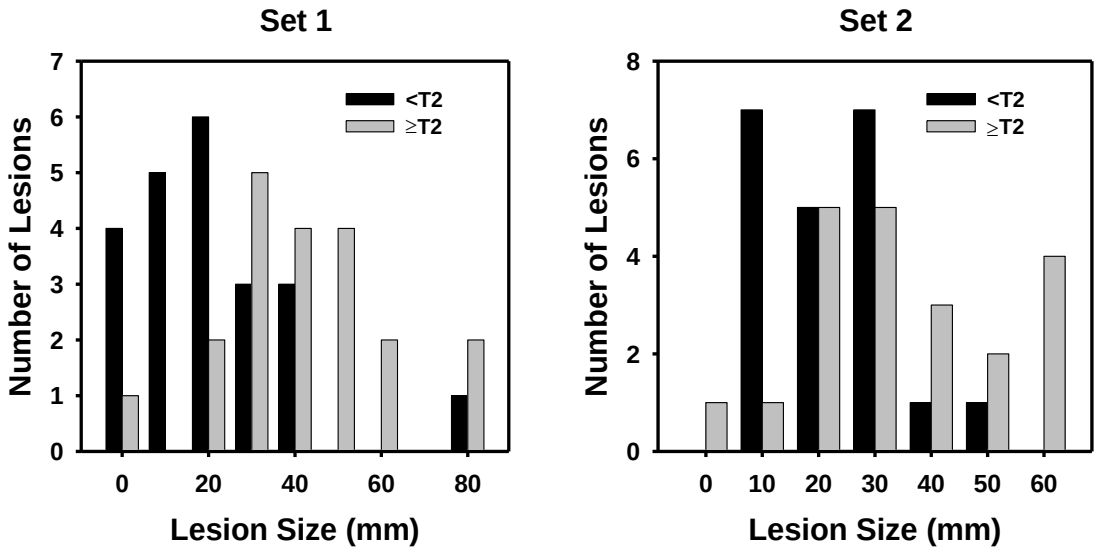
The data collection protocol was approved by our institutional review board and is HIPAA compliant. Patient informed consent was waived for this retrospective study. Our data set consisted of 84 bladder cancer lesions from 76 bladder cancer CTU cases collected from patient files without additional imaging for research purpose. The CTU scans in this data set were acquired using the protocol described in Chapter II, section 2.3.4. The data set consisted of 22 non-contrast cases (22 lesions), 22 early phase contrast-enhanced cases (22 lesions), and 32 delayed-phase contrast-enhanced cases (40 lesions). Per imaging protocol, the early phase

contrast-enhanced images are obtained 60-70 seconds following the initiation of a contrast injection. The delayed-phase contrast-enhanced images are obtained 12 min after the initiation of contrast injection. The type of scan a patient receives is determined by the protocol of the hospital performing the scan. Our data set includes patients referred to our hospital for treatment so that some scans were performed at outside hospitals and followed different scanning protocols, resulting in scans with inconsistent contrast-enhancement phase. A patient may also get a non-contrast scan due to risk factors, such as allergy to the contrast media, asthma or renal insufficiency[65].

For all cases, clinical and pathological staging were performed during the patient's clinical care. Cystectomy was performed after completing the course of neoadjuvant chemotherapy. The primary chemotherapy regimen used for the patients in our data set was MVAC, which is a combination of four medications: Methotrexate, Vinblastine, Doxorubicin, and Cisplatin. Stage T2 is identified to be clinically important as a decision threshold for neoadjuvant chemotherapy treatment. The stage at the beginning of the tumor management process, based on the clinical staging and pathological staging was used as a reference standard of the tumor stage for our study.

In addition, for all bladder cancer lesions, a radiologist measured the longest diameter on the pre-treatment scans by using an electronic caliper provided by an in-house developed graphical user interface.

The 84 bladder cancer lesions were separated into two classes. The first class consisted of 41 cancers that were stage T2 or above and the patients were treated with neoadjuvant chemotherapy. The second class consisted of 43 cancers that were below stage T2 and patients were not referred to neoadjuvant chemotherapy treatment. The data set was then split randomly by case into two sets with 42 cancers each while keeping the proportion of cancers between the two classes similar. The first set (Set 1) consisted of 22 cancers below stage T2 and 20 cancers stage T2 or above. The second set (Set 2) consisted of 21 cancers below stage T2 and 21 cancers stage T2 or above.



(a) (b)

Figure 7.3: Distribution of tumor sizes (the longest diameters) for Set 1 and Set 2. (a) Set 1: The average tumor sizes of stage < T2 and ≥ T2 were 26.4±17.3 mm and 45.6±19.1 mm respectively. (b) Set 2: The average tumor sizes of stage < T2 and ≥ T2 were 27.3±10.8 mm and 40.6±17.3 mm respectively.

In Set 1, two patients had two lesions and one patient had three lesions. In Set 2, three patients had two lesions. In Set 1, the average tumor sizes (the longest diameters) of stage < T2 and ≥ T2 were 26.4±17.3 and 45.6±19.1 mm, respectively (Figure 7.3(a)). In Set 2, the average tumor sizes (the longest diameters) of stage <T2 and ≥T2 were 27.3±10.8 mm and 40.6±17.3 mm, respectively (Figure 7.3(b)).

7.3.2 Segmentation of bladder lesions on CT urography

Our previously developed method for bladder lesion segmentation using an auto-initialized cascaded level set (AI-CALS) was used[47], which was previously described in Chapter IV, section 4.3.4.1. The details of the AI-CALS method can be found in our previous paper[47].

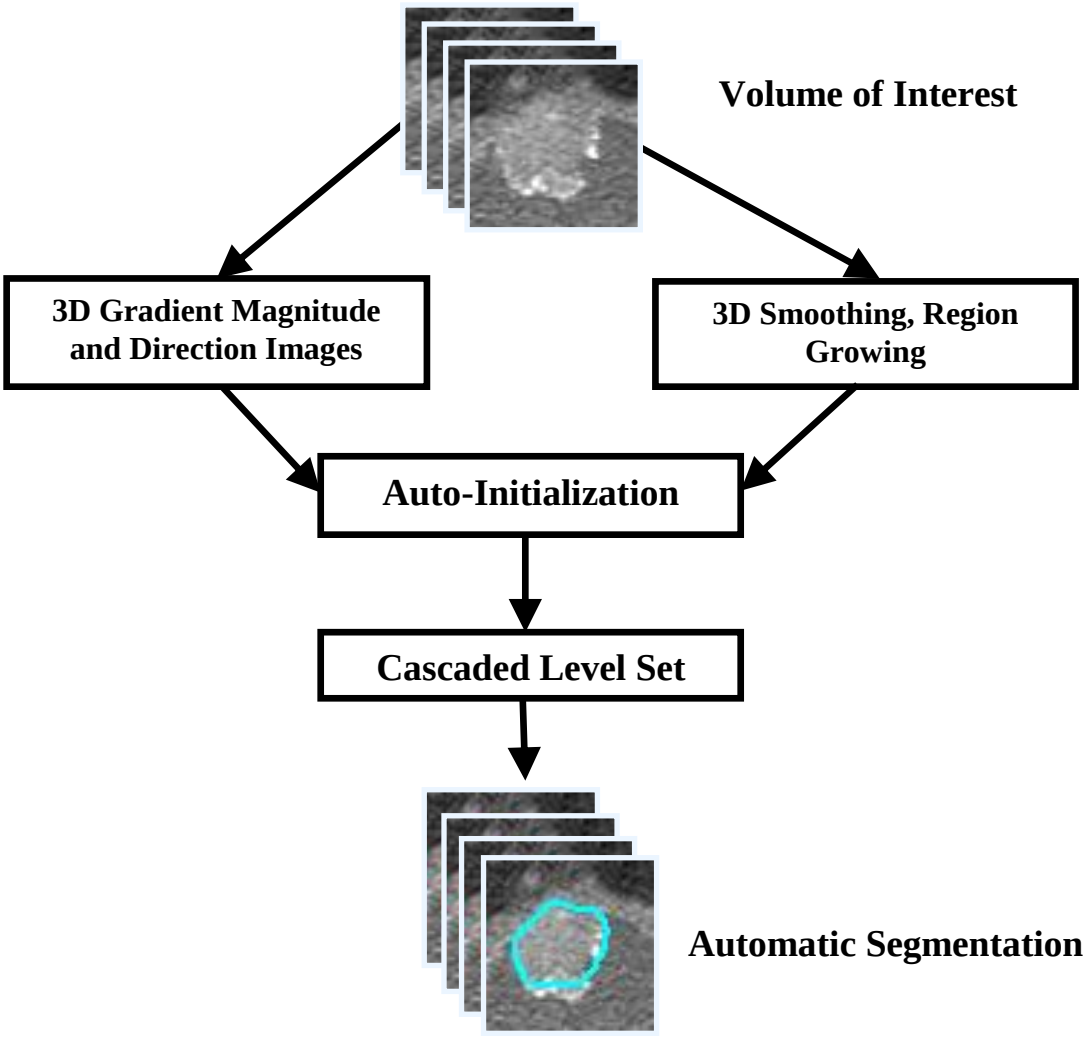


Figure 7.4: Block diagram of the auto-initialized cascaded level sets (AI-CALS) method.

7.4 Classification

7.4.1 Feature extraction

Following automated computer segmentation, texture features and morphological features were extracted to characterize the lesion. The mass size was measured as its 3D volume. Five morphological features were extracted based on the normalized radial length (NRL). NRL is defined as the radial length normalized relative to the maximum radial length for the segmented object[51]. The NRL features extracted include zero crossing count, area ratio, standard deviation, mean, and entropy. In addition, ten contrast features and a number of features including circularity, rectangularity, perimeter-to-area ratio, Fourier descriptor, gray level average,

standard deviation of gray level, mean density, eccentricity, moment ratio, and axis ratio were extracted as shape descriptors.

The texture of the tumor margin can provide important information about its characteristics. We calculated texture features from the rubber band straightening transform (RBST) images[66] of the tumor margin including those from the run-length statistics matrices, filtered Dasarathy east-west direction and filtered Dasarathy horizontal direction[30, 67]. The texture feature set also included the gray level radial gradient direction features.

In total, 91 features were extracted to form the feature space, including 26 morphological features and 65 texture features.

7.4.2 Feature selection/classification

A block diagram of the machine learning based bladder cancer staging system is shown in Figure 7.5. Stepwise feature selection was used to select the best subset of features to create an effective classifier[68]. A number of different classification experiments were performed to determine the best collection of input features. The classification performance was compared in three feature spaces: (1) morphological features only, (2) texture features only, and (3) morphological and texture features combined. A two-fold cross validation was conducted by partitioning the data set into Set1 and Set 2. In the first fold, Set 1 was used for feature selection and classifier training. The trained classifier was then tested on Set 2. In the second fold, feature selection and classifier training were performed on Set 2 and then tested on Set 1.

When training on a given fold (for example, Set 1) a leave-one-case-out resampling scheme with stepwise feature selection was used to reduce the dimensionality of the feature space. In stepwise feature selection, one feature is entered or removed in alternate steps while their effect is analyzed using the Wilks' lambda criterion[68]. The significance of the change in the Wilks' lambda when a feature is included or removed was estimated by F statistics. F_{in}, F_{out}, and tolerance are the parameters of the stepwise feature selection that define the thresholds for inclusion or exclusion of a given feature. A range of F_{in}, F_{out}, and tolerance values is evaluated by using an automated simplex optimization method. The set of F_{in}, F_{out}, and tolerance values that lead to the highest classification result with the lowest number of features based on the training set are selected. A smaller number of features are preferred in order to reduce the chance of overfitting. Once the set of F_{in}, F_{out}, and tolerance is selected, the stepwise feature selection

with the selected parameter set is applied to the entire training fold to select a single set of features and train a single classifier. After the classifier is fixed it is applied to the test fold (for example, Set 2) for performance evaluation.

Four different classifiers were evaluated in this study. The same partitioning of Set 1 and Set 2 was used for all classifiers. We compared the four classifiers for this classification task. The first classifier was linear discriminant analysis (LDA)[69, 70]. The LDA with the stepwise feature selection was used to determine the most effective features using the training set in each fold, as described above. The second classifier was a back-propagation neural network (NN) [71] with a single hidden layer and a single output node. The selected features from LDA were used for this classifier and they determined the number of input nodes to the NN. The parameters for the NN were adjusted using the training set, and the best performing network was applied to the test set. The third classifier was a support vector machine (SVM)[72, 73] with a radial basis kernel. Using training data, a SVM determines a decision hyperplane to separate the two classes by maximizing the distance, or the margin, between the training samples of both classes and the hyperplane. The width of the SVM radial basis kernels γ was varied between 0.02 and 0.14 for the experiments. The best parameters for the SVM kernels for a specific experiment were selected using the training set, which were then applied to the test set. The LDA selected features were also used as the input to the SVM. The fourth one is the Random Forest (RAF) classifier[74]. We used the WEKA[75] implementation and selected 50 to 100 trees and 5 to 7 features per tree for our classification task using the training set in each fold. The parameters for the random forest classifier were determined experimentally using the training sets. All 91 features were used as an input to the RAF.

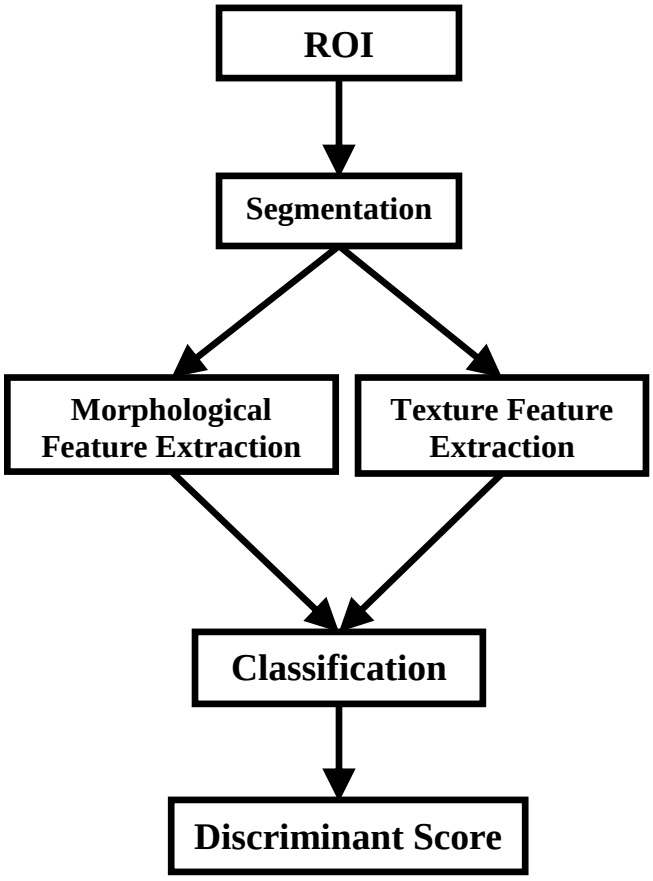


Figure 7.5: Block diagram of our machine learning based staging system. We compared the linear discriminant analysis (LDA), back-propagation neural network (NN), Support vector machine (SVM), and Random forest classifiers (RAF) in the classification stage for this study.

7.4.3 Evaluation methods

Lesion segmentation performance was evaluated using radiologists' 3D hand-segmented contours as reference standards. The hand outlines of all 84 lesions were obtained from an experienced abdominal radiologist (RAD1). Hand outlines for a subset of 12 lesions were obtained from a second experienced abdominal radiologist (RAD2). The average distance and the Jaccard index[29] were calculated between the computer outlines and the hand outlines, using methods described in Chapter II, section 2.3.5.

To evaluate the classifier performance, the training and test scores output from the classifier were analyzed using the receiver operating characteristic (ROC) methodology[76]. The classification accuracy was evaluated using the area under the ROC curve, A_z. The statistical

significance of the differences between the different classifiers and feature spaces were estimated by the CLABROC program using ROC software by Metz et al.[77, 78].

7.5 Results

The lesion segmentation performance of the AI-CALS compared to the radiologist hand outlines for the 84 lesions are shown in Table 7.1. Table 2 shows the computer segmentation performance compared to two different radiologists' hand outlines for a subset of 12 lesions.

Table 7.1: Segmentation performance of the 84 lesions compared to hand-outlines performed by radiologist 1 (RAD1).

	AI-CALS vs. RAD1
Average distance ***AVDIST***	4.9 ± 2.7 mm
Jaccard index ***JACCARD*^{3D}**	43.5 ± 14.0%

Table 7.2: Segmentation performance for a subset of 12 lesions compared to hand-outlines performed by two different radiologists (RAD1, RAD2).

	AI-CALS vs. RAD1	***AI-CALS vs. RAD2***	***RAD1 vs. RAD2***
Average distance ***AVDIST***	5.2 ± 2.5 mm	4.1 ± 1.5 mm	2.9 ± 1.1 mm
Jaccard index ***JACCARD*^{3D}**	43.2 ± 13.2%	50.1 ± 14.7%	58.7 ± 11.1%

The performance of the classifiers based on different machine learning techniques, the LDA, NN, SVM, and RAF, is summarized in Table 7.3. Different feature spaces containing the morphological features, the texture features, and the combined set of both morphological and texture features were used for classification. The features selected with LDA were used in the SVM and NN classifiers. The LDA classifier with morphological features achieved a training A_z of 0.91 on Set 1 and a test A_z of 0.81 on Set 2. For training on Set 2 it achieved a A_z of 0.97 and a test A_z of 0.90 on Set 1. The selected features on the training sets included volume, a contrast feature, and gray level feature. The test A_z of the NN for Set 1 and Set 2 was 0.88 and 0.91 respectively. The SVM achieved test A_z of 0.88 on Set 1 and test A_z of 0.90 on Set 2. The test A_z of the RAF for Set 1 and Set 2 was 0.83 and 0.88 respectively. The distribution of the

discriminant scores from the four classifiers for testing on Set 1 and Set 2 in two fold cross-validation in the morphological feature space is presented in Figure 7.6. It can be observed that most of the classifiers were able to provide a relatively good separation between the two classes.

By using the texture features the LDA classifier achieved a test A_z of 0.91 on Set 1 and a test A_z of 0.88 on Set 2. When trained on Set 1 or Set 2 the stepwise feature selection procedure selected subsets of the filtered Dasarathy east-west direction features, the filtered Dasarathy horizontal direction features and the gray level radial gradient direction features. The test A_z of the NN classifier for Set 1 and Set 2 was 0.89 and 0.92, respectively. The SVM classifier achieved test A_z of 0.91 on Set 1 and test A_z of 0.89 on Set 2. The test A_z of the RAF classifier for Set 1 and Set 2 was 0.89 and 0.97, respectively.

When the morphological and the texture features were combined, the LDA classifier achieved a test A_z of 0.89 on Set 1 and a test A_z of 0.90 on Set 2. When trained on Set 1 or Set 2 the stepwise feature selection procedure selected a contrast feature, subsets of the filtered Dasarathy horizontal direction features, and subsets of the gray level radial gradient direction features. The test A_z of the NN classifier for Set 1 and Set 2 was 0.91 and 0.95, respectively. The SVM classifier achieved test A_z of 0.92 on Set 1 and test A_z of 0.89 on Set 2. The test A_z of the RAF classifier for Set 1 and Set 2 was 0.86 and 0.96, respectively. The test ROC curves for all of the classifiers when tested on Set 1 and Set 2 in the two fold cross-validation in the different feature spaces are shown in Figure 7.7.

The differences in the A_z values between pairs of classifiers did not achieve statistical significance. The classifiers achieved slightly higher A_z values in the texture and combined feature spaces than in the morphological feature space; however, the differences did not achieve statistical significance after Bonferroni correction for the multiple comparisons (p-value < 0.05/18=0.0028 to be considered significant).

Table 7.3: Summary results for LDA, NN, SVM and RAF classifiers in morphological, texture, and combined feature spaces. The column "Number of Features" did not apply to the RAF classifier. All features were used for the RAF classifier. The differences in the A_z values in pair-wise comparison of the different classifiers did not achieve statistical significance after performing Bonferroni correction for the 18 comparisons ($p>0.0028$).

		LDA		**NN**		**SVM**		**RAF**	
Feature Type	***Number of Features***	***Training***	***Testing***	***Training***	***Testing***	***Training***	***Testing***	***Training***	***Testing***
Morphological Features									
Training (Set 1) Testing (Set 2)	4	0.91	0.81	0.96	0.91	0.95	0.90	1	0.88
Training (Set 2) Testing (Set 1)	4	0.97	0.90	0.98	0.88	0.97	0.88	1	0.83
Texture Features									
Training (Set 1) Testing (Set 2)	2	0.91	0.88	0.95	0.92	0.92	0.89	1	0.97
Training (Set 2) Testing (Set 1)	7	1	0.91	1	0.89	1	0.91	1	0.89
Combined Features									
Training (Set 1) Testing (Set 2)	3	0.92	0.90	0.97	0.95	0.92	0.89	1	0.96
Training (Set 2) Testing (Set 1)	7	1	0.89	1	0.91	1	0.92	1	0.86

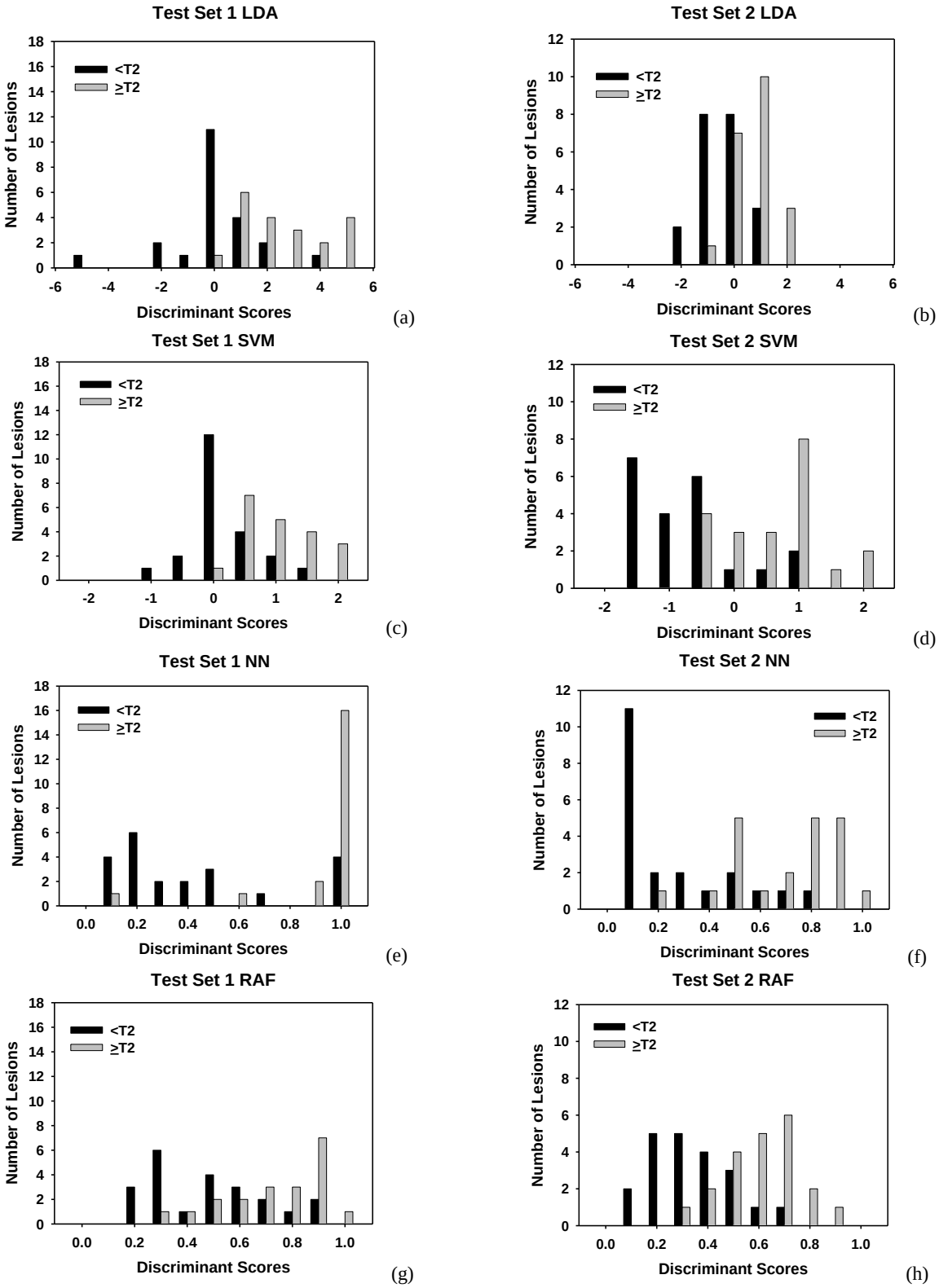
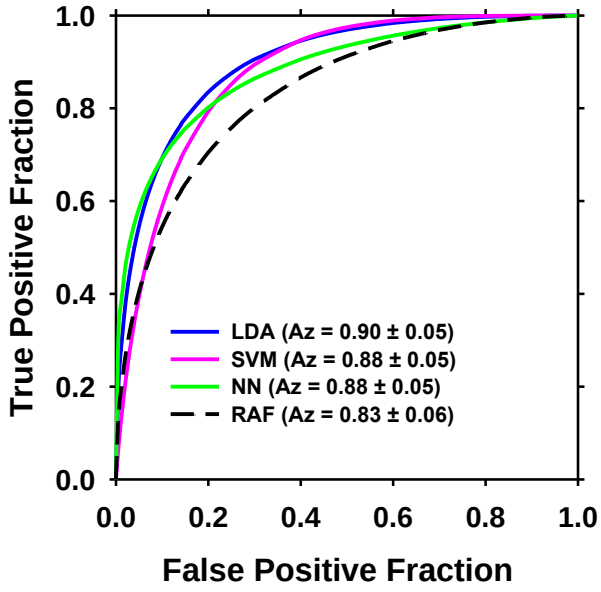
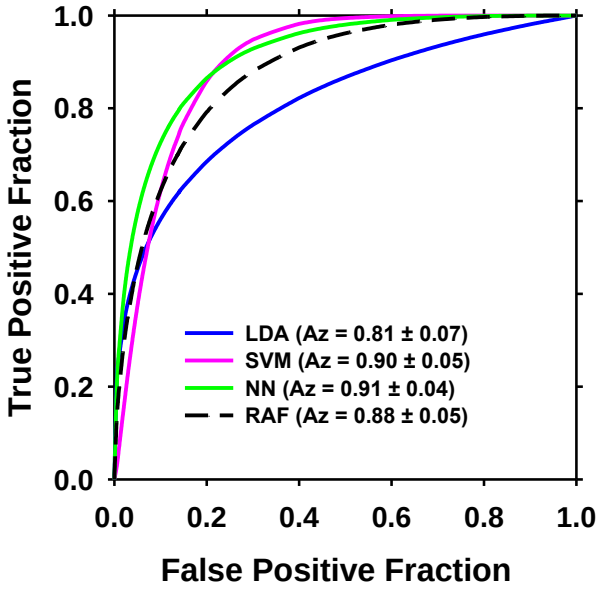


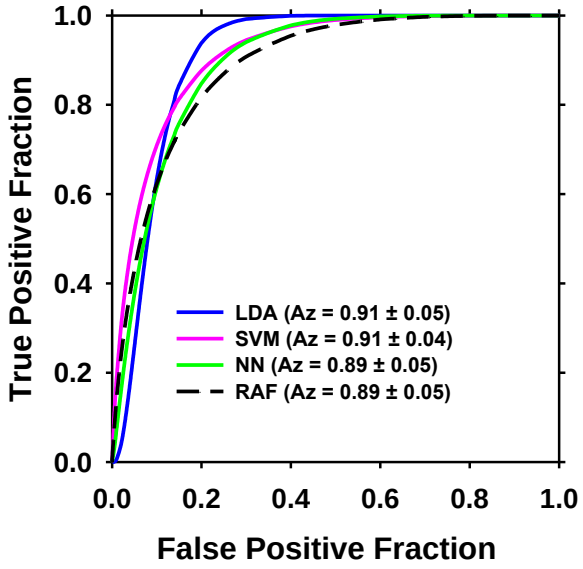
Figure 7.6: Distribution of the classifiers discriminant scores for testing on Set 1 and Set 2 in two-fold cross validation using the morphological features. (a) LDA (Set 1) $A_z = 0.90$, (b) LDA (Set 2) $A_z = 0.81$, (c) SVM (Set 1) $A_z = 0.88$, (d) SVM (Set 2) $A_z = 0.90$, (e) NN (Set 1) $A_z = 0.88$, (f) NN (Set 2) $A_z = 0.91$, (g) RAF (Set 1) $A_z = 0.83$, (h) RAF (Set 2) $A_z = 0.88$.


True Positive Fraction
False Positive Fraction
LDA (Az = 0.90 ± 0.05)
SVM (Az = 0.88 ± 0.05)
NN (Az = 0.88 ± 0.05)
RAF (Az = 0.83 ± 0.06)

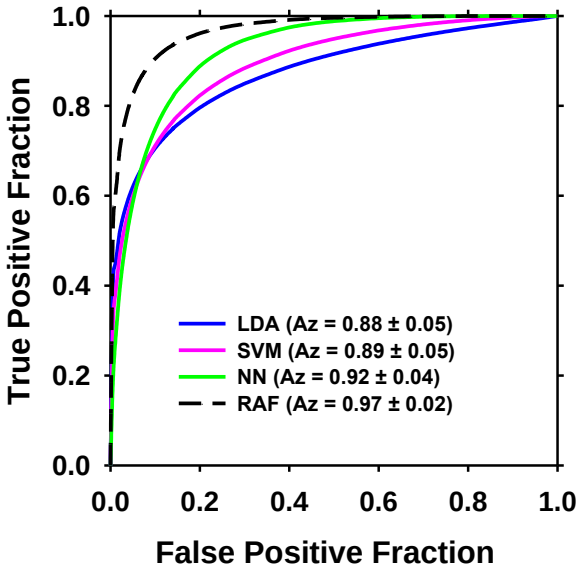
(a)


True Positive Fraction
False Positive Fraction
LDA (Az = 0.81 ± 0.07)
SVM (Az = 0.90 ± 0.05)
NN (Az = 0.91 ± 0.04)
RAF (Az = 0.88 ± 0.05)

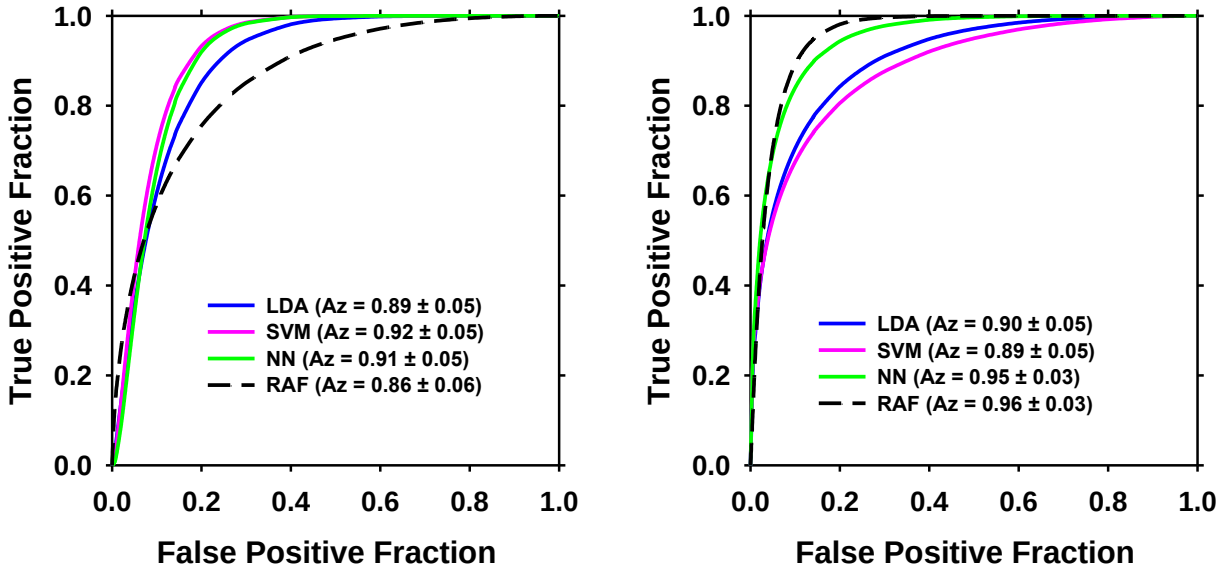
(b)


True Positive Fraction
False Positive Fraction
LDA (Az = 0.91 ± 0.05)
SVM (Az = 0.91 ± 0.04)
NN (Az = 0.89 ± 0.05)
RAF (Az = 0.89 ± 0.05)

(c)


True Positive Fraction
False Positive Fraction
LDA (Az = 0.88 ± 0.05)
SVM (Az = 0.89 ± 0.05)
NN (Az = 0.92 ± 0.04)
RAF (Az = 0.97 ± 0.02)

(d)



(e) (f)

Figure 7.7: ROC curves for testing on Set 1 and Set 2 in two-fold cross validation for LDA, SVM, NN, and RAF classifiers: Left column: testing on Set 1, right column: testing on Set 2. (a) and (b) morphological features; (c) and (d) texture features; (e) and (f) combined features.

7.6 Discussion

The agreement between the AI-CALS lesion segmentation and the radiologists' manual segmentation was slightly lower than the agreement between two radiologists' hand outlines, indicating that the computer segmentation will need to be further improved. An improved method for the bladder lesion segmentation is presented in Chapter VIII. Both the morphological and the texture features were important for classifying the bladder cancer stage. When only morphological features were used in the classifier, volume and contrast features were always selected. Volume was the primary feature used to describe lesion size. When the classifier used only the texture features, the features from the 3 main groups, the filtered Dasarathy east-west direction features, the filtered Dasarathy horizontal direction features, and the gray level radial gradient direction features were consistently selected. There was essentially no change in classification accuracy when the morphological features were added to the texture features in the combined set.

The LDA, SVM, and NN classifiers all led to relatively consistent results. There was no statistically significant difference in the performances between pairs of the classifiers. The best overall results for the two-fold cross validation were obtained when a combined feature set was

used with an NN classifier. Using Set 1 for training, the training A_z was 0.97 and the test A_z was 0.95. Using Set 2 for training, the training A_z was 1.00 and the test A_z was 0.91.

The RAF classifier showed greater imbalance between Set 1 and Set 2 than the other classifiers. When training was done on Set 2 and testing on Set 1, the A_z were substantially lower than the A_z values when training was done on Set 1 and testing on Set 2. For example, the test A_z decreased from 0.88 to 0.83 for morphological features, from 0.97 to 0.89 for texture features only, and from 0.96 to 0.86 for the combined features. This imbalance between the two sets could be due to the fact that RAF utilized all the features in the subspace whereas the other three classifiers involved feature selection.

Examples of bladder cancers with stages $\geq$ T2 or $<$ T2 and the corresponding classifier scores are shown in Figure 7.8. The reported scores are test scores for the LDA, SVM, NN, and RAF classifiers based on the morphological features. In Figure 7.8(a), (b) and Figure 7.8(c), (d) are shown T1 stage cancers of different sizes that were correctly classified with low scores by all classifiers. Note that the output score ranges are different for different classifiers so that the score values should not be compared across classifiers. T3 stage and T2 stage cancers that were correctly classified with high scores from all classifiers are presented in Figure 7.8(e), (f) and Figure 7.8(g), (h), respectively. A case that was clinically identified as T1 stage pre-surgery but later was identified as a T2 stage cancer post-surgery is shown in Figure 7.8(k), (l). The classifiers classified the cancer as $\geq$ T2 with high scores. Figure 7.8(m), (n) show a T2 stage cancer that was incorrectly identified by the LDA, SVM, and NN classifiers with low scores, but correctly identified by the RAF with a high score.

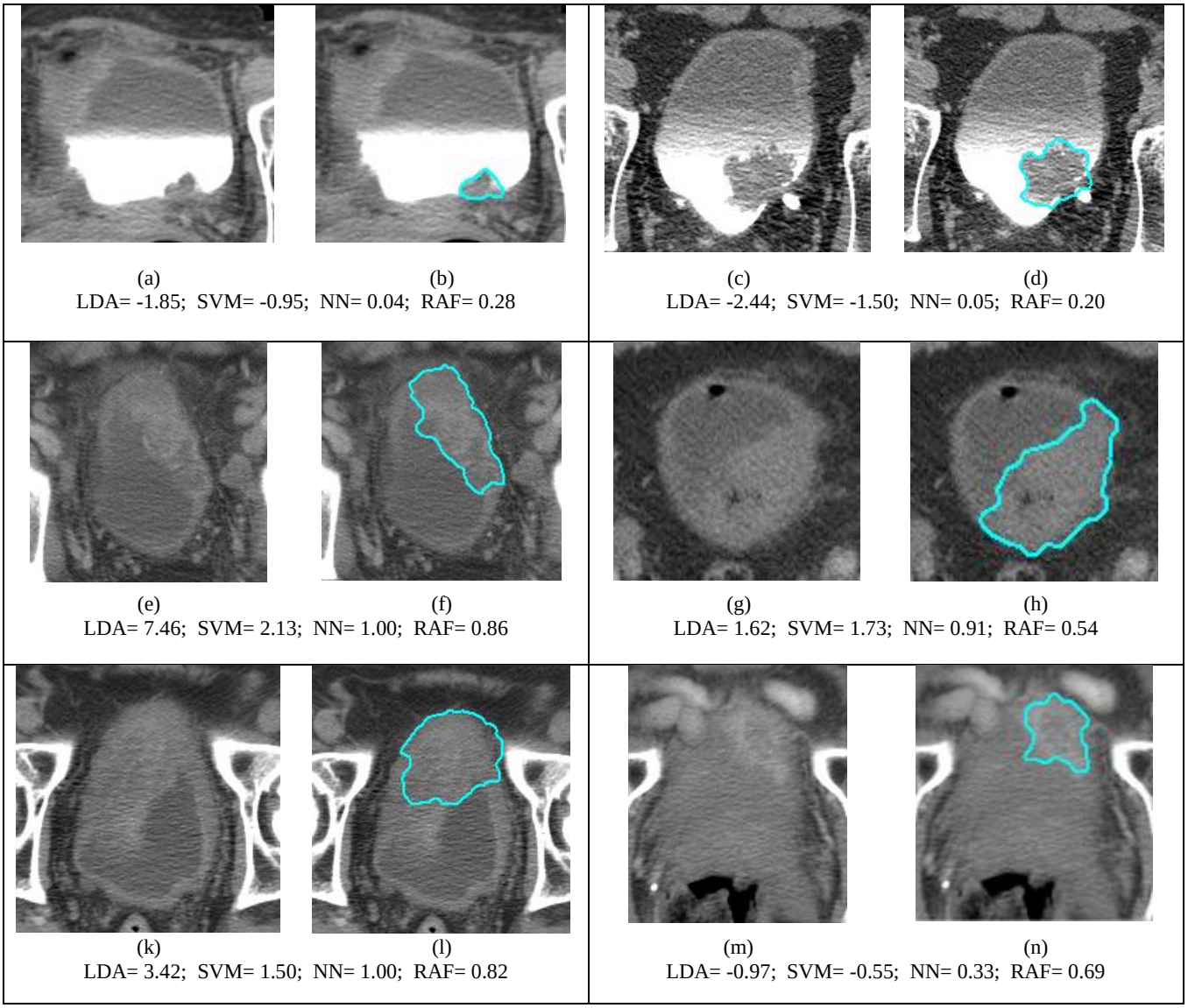


Figure 7.8: Examples of bladder cancers with stages ≥ T2 or < T2. The blue outlines represent the AI-CALS segmentation. The reported scores are test scores for the LDA, SVM, NN, and RAF classifiers based on the morphological features. Note that the output score ranges are different for different classifiers so that the score values should not be compared across classifiers. The two cases in (a)(b) and (c)(d) both contained was a T1 stage cancer that was properly classified with low scores from all classifiers. (e)(f) was a T3 stage case that was properly classified with high scores from all classifiers. (g)(h) was a T2 stage case that was properly classified with high scores from all classifiers. (k)(l) was a case that was clinically identified as T1 pre-surgery but was identified as a T2 stage cancer post-surgery. The classifiers classified the cancer as ≥T2 with high scores. (m)(n) was T2 stage cancer that was incorrectly identified by the LDA, SVM, and NN classifiers with low scores and correctly identified by the RAF with a high score.

We also have extracted features from the manually segmented bladder lesions and applied the 4 different types of classifiers with the different feature sets to the cancer stage prediction. The classifiers using features extracted from the manually segmented lesions

performed similarly to the classifiers using features extracted from the AI-CALS segmented lesions. The test A_z values ranged from 0.77 to 0.95. For 6 out of the 24 experiments the classifiers using features extracted from the manually segmented lesions performed better than classifiers using features extracted from the AI-CALS segmentations. However, the differences did not reach statistical significance. Therefore, although the performance of the AI-CALS lesion segmentation was slightly lower than the radiologists' hand outlines the final classification results were similar.

The main limitation of the study is the small data set. Another limitation is that we have not applied the deep learning convolution neural network (DLCNN) to this bladder cancer staging task. DLCNN has been shown to be superior to conventional classifiers in many classification tasks, especially the classification of natural scene images with millions of training samples. It also shows promise in number of medical imaging applications[79, 80] including bladder segmentation[2] and bladder cancer treatment response monitoring[10]. However, our experience with DLCNN also indicates that it is not always the best, perhaps limited by the relatively small annotated training set in medical imaging, even with transfer learning. As the performances of the four conventional classifiers used in this study were quite high, it would not be a fair comparison for DLCNN if we do not have adequate training for it. We will continue to collect additional cases and compare the conventional classifiers with DLCNN for bladder cancer staging in a future study.

7.7 Conclusion

In this preliminary study we proposed machine learning methods for prediction of bladder cancer stage. It was found that the morphological features and texture features were useful for assessing the stage of bladder lesions. The LDA, SVM, and NN classifiers all led to relatively consistent results. There was a trend that the SVM and NN classifier slightly outperformed the LDA classifier. The best overall results for the two-fold cross validation were obtained when a combined feature subspace was used with the NN classifier. Further studies are under way to improve the staging of bladder cancer and test the classifier on a larger data set, and to investigate the potential of improving the predictive model by combining imaging biomarkers with non-imaging biomarkers.

Chapter VIII

Bladder Cancer Segmentation in CT for Treatment Response Assessment: Application of Deep-Learning Convolution Neural Network - A Pilot Study

8.1 Abstract

Assessing the response of bladder cancer to neoadjuvant chemotherapy is crucial for reducing morbidity and increasing quality of life of patients. Changes in tumor volume during treatment are generally used to predict treatment outcome. We are developing a method for bladder cancer segmentation in CT using a pilot data set of 62 cases. 65,000 regions of interests were extracted from pre-treatment CT images to train a deep-learning convolution neural network (DL-CNN) for tumor boundary detection using leave-one-case-out cross-validation. The results were compared to our previous AI-CALS method. For all lesions in the data set, the longest diameter and its perpendicular were measured by two radiologists, and 3D manual segmentation was obtained from one radiologist. The World Health Organization (WHO) criteria and the Response Evaluation Criteria In Solid Tumors (RECIST) were calculated, and the prediction accuracy of complete response to chemotherapy was estimated by the area under the receiver operating characteristic curve (AUC). The AUCs were 0.73 ± 0.06, 0.70 ± 0.07 and 0.70 ± 0.06, respectively, for the volume change calculated using DL-CNN segmentation, the AI-CALS and the manual contours. The differences did not achieve statistical significance. The AUCs using the WHO criteria were 0.63 ± 0.07 and 0.61 ± 0.06, while the AUCs using RECIST were 0.65 ± 007 and 0.63 ± 0.06 for the two radiologists, respectively. Our results indicate that DL-CNN can produce accurate bladder cancer segmentation for calculation of tumor size change in response to treatment. The volume change performed better than the estimations from the WHO criteria and RECIST for the prediction of complete response. The results presented in this chapter have been published[4].

8.2 Introduction

The standard treatment method for bladder cancer involves radical cystectomy of the bladder; however, approximately 50% of the patients who have under gone cystectomy and were thought to have only locally invasive cancer at the time of the surgery develop metastatic disease within 2 years and subsequently die of the disease[81]. This may be due to the presence of micro-metastatic disease, or presence of neoplasms that had spread to perivascular tissue, with the spread not detected at the time of the treatment. Neoadjuvant chemotherapy has been shown to improve resectability of large tumors, as well as being beneficial for the treatment of micrometastases before radical cystectomy[58-60]. The methotrexate, vinblastine, doxorubicin, and cisplatin (MVAC) treatment regimen has been shown to decrease the chance for finding residual cancer after cystectomy than cystectomy alone, and increase the survival of patients with locally-advanced bladder cancers[82, 83]. The side effects of this treatment are severe, however, which include neutropenic fever, sepsis, mucositis, nausea, vomiting, malaise, and alopecia[84]. As there are no reliable methods at present to predict a patient's response to chemotherapy, it is possible that patients suffer through these conditions while having to endure treatment that may or may not achieve the desirable benefits. Therefore, early assessment of the bladder cancer treatment response is important, as it may allow the clinician to stop the treatment early if the treatment does not work well. This will help reduce morbidity of the patient and increase their quality of life. It may also preserve the patient's physical conditions and allow the patient to pursue alternative treatment that may be more beneficial.

The response to treatment can be measured via pathological information from the removed bladder after cystectomy or other surgical procedures such as transurethral resection of bladder tumor (TURBT). However, as these patients are receiving chemotherapy, surgery may not be the ideal method for assessing treatment response. The patients are weak from going through the chemotherapy and may have difficulties with the additional burden of surgery in the middle of their chemotherapy treatment; therefore, noninvasive evaluation would be preferable. Image-based evaluation, using CT or MRI images, can noninvasively visualize the tumor during the chemotherapy treatment. Specific features from these images, known as radiomics features, may be extracted and analyzed to determine the properties of the tumors.

The clinical estimation of the tumor size and its response to treatment is based on the World Health Organization (WHO) criteria[85] and the Response Evaluation Criteria in Solid

Tumors (RECIST)[86]. In the WHO criteria, the longest diameter of a tumor and its perpendicular diameter are measured. The response to treatment is defined as the percentage reduction of the products of the two diameters between the pre- and post-treatment measurements. The RECIST criteria use the percentage reduction of the longest diameter between the pre- and post-treatment measurements. Both of these methods can be inaccurate and can have large inter- and intra-observer variations, especially for tumors with irregular and complex shapes[87]. As RECIST criteria and the WHO criteria involve only one-dimensional (1D) and 2D measurements, respectively, the volumetric (3D) information from a CT scan is not fully utilized, and it is possible that 3D information may provide better response evaluation.

The gross tumor volume (GTV), can be effectively measured in CT images, and has been shown previously to predict outcomes of bladder cancers[88]. For an accurate GTV measure from the CT images, the bladder tumor in the images needs to be delineated slice by slice; however, this is a time and labor intensive procedure, thus the burden of the additional workload on the radiologists cannot be ignored. Computerized segmentation tools that can automatically or semi-automatically delineate the tumors from its surroundings would greatly reduce the additional workload. In our previous preliminary study, we have shown that our method for computerized segmentation of bladder tumors, Auto-Initialized Cascaded Level Sets (AI-CALS) can reliably produce 3D segmentation for a variety of bladder tumors[47], and that the volume estimates more accurately predict the complete response to treatment compared to the WHO and the RECIST criteria on a small data set[89].

Although the 3D volumetric measurement by AI-CALS provides better estimates than the 1D and 2D estimates in tumor size changes, the segmentation method still needs improvement in many cases given the wide variety of the tumor shapes and characteristics in the patient population. In this study, we explored the application of deep-learning convolution neural network (DL-CNN) to the segmentation of bladder tumors. Convolution neural networks (CNNs) have been used previously to classify patterns in medical images for use with computer-aided detection, specifically for microcalcification and mass detection in mammograms[31, 32, 34-38, 90, 91]. The training set used in these applications were typically small, generally less than 500 samples. With the advances in computation power, it becomes practical to design CNNs with very large and complex architectures that require massive number of training samples to solve more challenging pattern recognition problems. The deep-learning CNN (DL-CNN) using graphics

processing units (GPU) has been shown to be successful in classifying natural scene images using a large training set. Krizhevsky et al.[39, 40] has demonstrated that high classification accuracy can be achieved using DL-CNN on the ImageNet ILSVRC-2010 and ILSVRC-2012 data sets[41], and the CIFAR-10 data set[42]. DL-CNN has also been successfully used for computer-aided detection in medical imaging[92]. We have previously applied DL-CNN to the segmentation of whole bladders in CT images[2]; however, the segmentation of the tumors are more difficult because contrast material is generally not used in CT for patients undergoing chemotherapy, resulting in low contrast between the tumor and the inside of the bladder.

In this pilot study, we applied DL-CNN to bladder lesion segmentation. For this task, the DL-CNN was trained to recognize the patterns in the regions that were inside and outside of the bladder lesion and generate a lesion likelihood map. Minor refinement on the likelihood map was performed by level sets to obtain the segmented boundaries of the bladder cancer.

The bladder cancer segmentation performance of the DL-CNN and the AI-CALS were compared quantitatively with the radiologists' manual outlines. The cancer volumes were calculated from the segmented tumor boundaries, and the GTV change in response to neoadjuvant chemotherapy was calculated. The results obtained from the DL-CNN were compared to the results obtained using our previous AI-CALS segmentation method, a radiologist's manual outlines, as well as the response estimation using the WHO and the RECIST criteria.

8.3 Materials and Methods

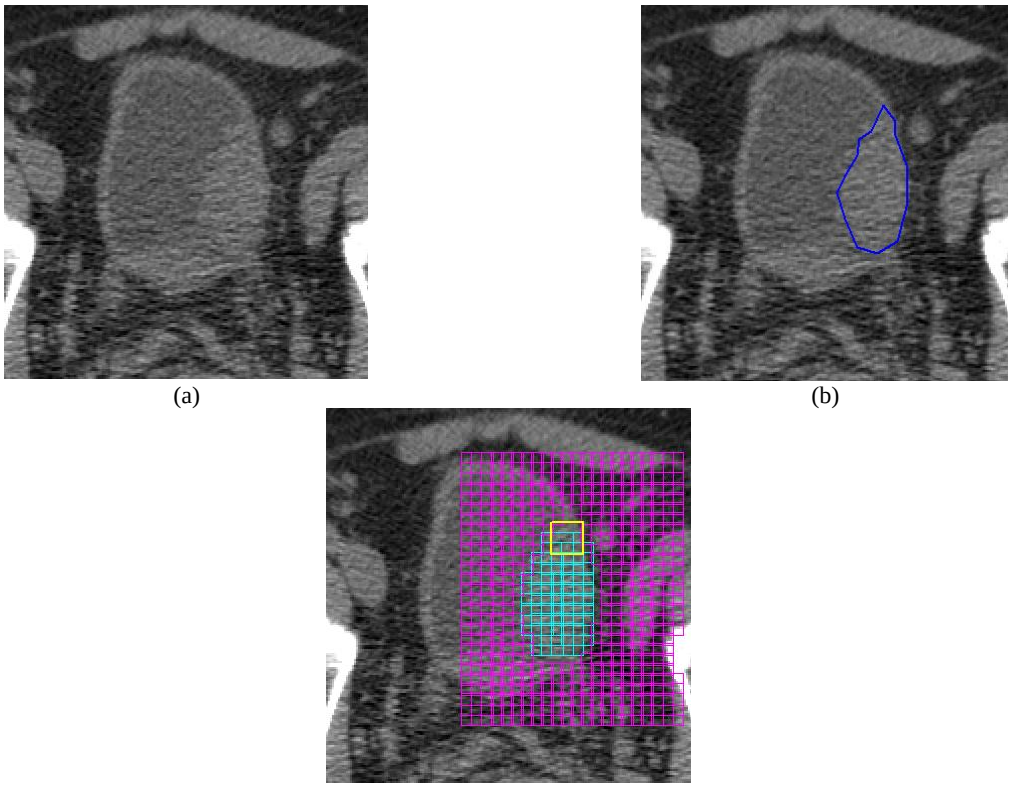
8.3.1 Data set

A data set of 62 cases was collected retrospectively from the Abdominal Imaging Division of the Department of Radiology at the University of Michigan with Institutional Review Board approval for this pilot study. All of the patients in the data set had undergone CT examination before and after chemotherapy, and subsequently underwent cystoscopy, biopsy, or radical cystectomy. The CT scans used in this study were acquired in our clinic with GE Healthcare LightSpeed MDCT scanners. The images were acquired using 120 kVp and 120 – 280 mA and reconstructed at a slice interval of 0.625, 1.25, 2.5, or 5 mm with pixel sizes ranging from 0.586 to 0.977. The data set contained 64 tumors forming 74 temporal pairs, and 27% (17/62) of the cases had stage pT0 after surgery, indicating a complete response to treatment.

A reference standard for the computerized segmentation was obtained via 3D hand-segmented contours of the bladder tumors in the pre- and post-treatment CT of the 62 cases. An experienced radiologist with 27 years of experience reading CT bladder cases identified and marked focal tumor locations within the CT scans. The radiologist also manually outlined the full 3D contour for all cases (reference standard 1), and measured the longest diameter and its perpendicular using a graphical user interface (GUI) that we have developed. A second radiologist with 17 years of experience in CT bladder cases also manually outlined the bladder tumors in the pre- and post-treatment CT for a subset of 29 cases (reference standard 2), and measured the longest diameter and its perpendicular independently following the WHO criteria and RECIST criteria for all 62 cases.

8.3.2 DL-CNN training

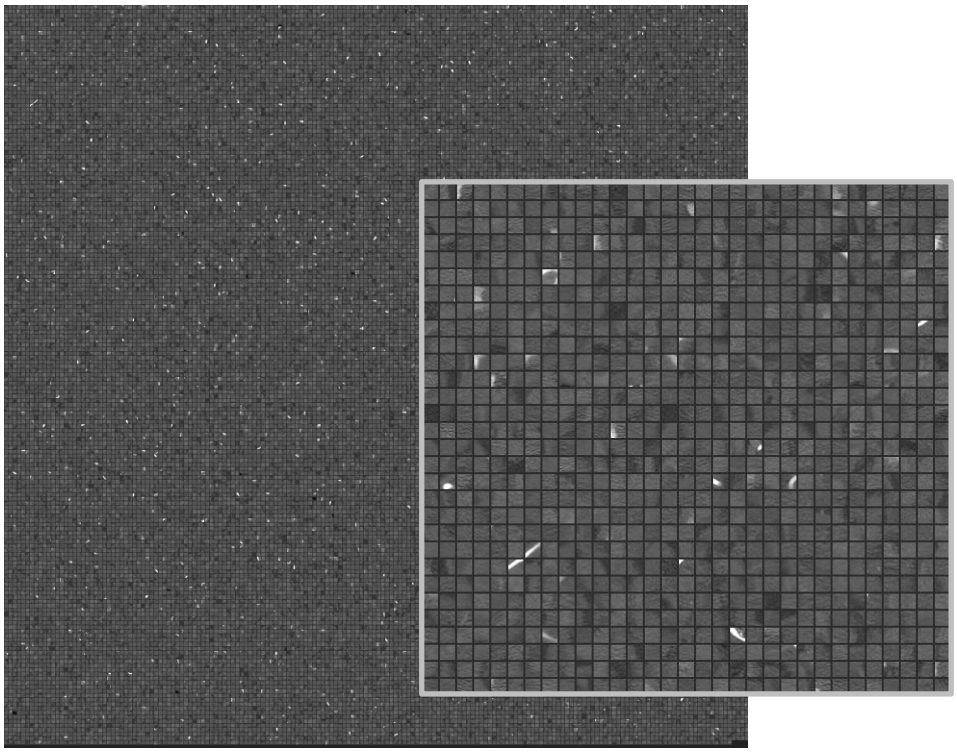
Our research work on whole bladder segmentation using DL-CNN was expanded further to bladder tumor segmentation. The DL-CNN by Krizhevsky et al. called cuda-convnet[39, 40] was used. The neural network was trained to classify regions of interests (ROIs) on 2D slices as being inside or outside of the bladder cancer. Details on the DL-CNN can be found in the literature[2] and in Chapter III. The DL-CNN was trained with the pre-treatment scans of the cases. For each axial slice of the cases, a large number of overlapping 16 x 16-pixel ROIs were extracted from the region including the cancer marked by the radiologist. If more than 80% of an ROI was within the hand-outlined bladder cancer, the ROI was labeled as being inside of the cancer, whereas the ROI had to be completely outside of the cancer in order for it to be classified as being outside the cancer. ROIs not labeled as either inside or outside of the cancer were excluded. The ROI size of 16 x 16-pixels was used to ensure sufficiently accurate coverage of both small and large cancers. This size also allows for the ROIs to contain mostly regions that are either inside of the cancer or outside of the cancer. Figure 8.1 shows an example of ROIs obtained from a CT slice. The number of ROIs within the two classes was balanced, resulting in approximately 65,000 ROIs. Figures 8.2(a) and (b) show examples of ROIs inside and outside of the bladder cancer, respectively, used to train the DL-CNN.



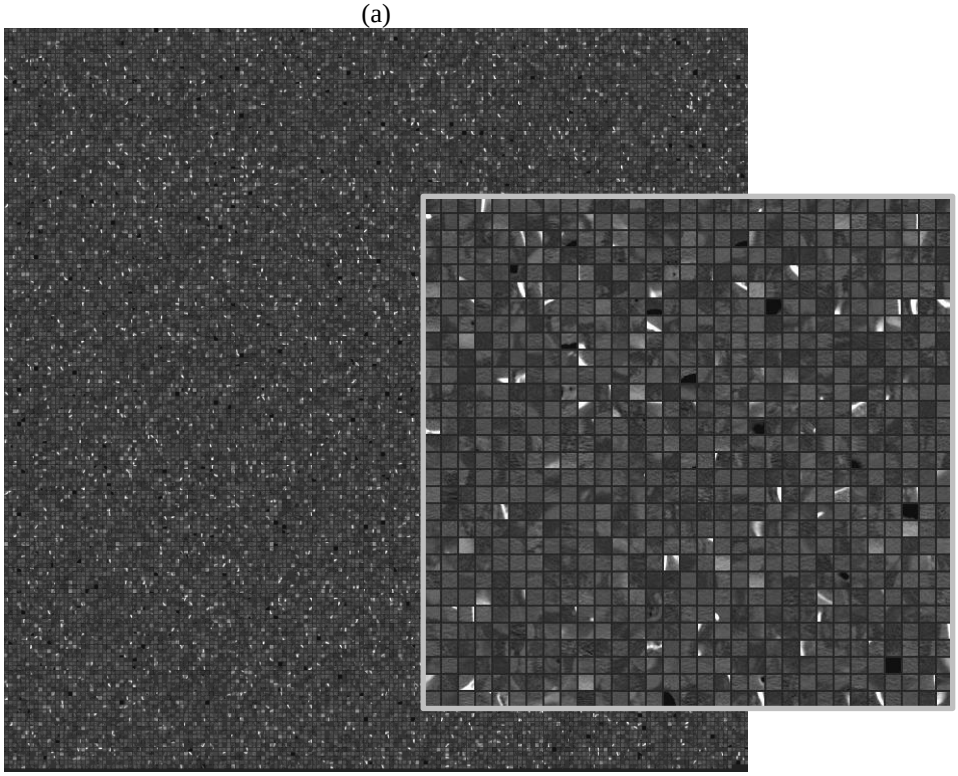
(a) (b)

(c)

Figure 8.1: An axial slice of a pre-treatment CT scan from a training case. (a) Cropped CT slice centered at the bladder. (b) Radiologist's hand-outline of the cancer overlaid on the CT slice. (c) ROIs extracted from this slice. The yellow ROI shows the size of a 16 x 16-pixel ROI. The ROIs are partially overlapping. The blue ROIs are labeled as inside the bladder cancer. The pink ROIs are labeled as outside the bladder cancer for training the DL-CNN.



(a)



(b)

Figure 8.2: Composite images of the 65,000 ROIs from the training set used to train the DL-CNN. Each ROI is 16 x 16 pixels. (a) ROIs labeled as being inside bladder cancers. (b) ROIs labeled as being outside bladder cancers. A portion of each composite image is enlarged to show the typical ROIs in each class.

The network architecture used in this study consists of five main layers: two convolution layers, two locally connected layers, and one fully connected layer. The first convolution layer consists of 64 kernels with a size of 5 x 5 pixels. The output of this layer is pooled, normalized and input into the second convolution layer that also consists of 64 kernels with a size of 5 x 5 pixels. The output of this layer is also pooled, normalized and input to the first locally connected layer that has 64 kernels with a size of 3 x 3 pixels. The second locally connected layer has 32 kernels with a size of 3 x 3 pixels and output to the fully connected layer. The fully connected layer outputs two values to a Softmax layer that converts the values to a range from 0 to 1. The output of the DL-CNN can be interpreted as the likelihood of an input ROI being classified into one of the two categories. The neural network was trained for 1500 iterations, which was sufficient for the classification error rate to converge to a minimum and remained stable. Leave-one-case-out cross-validation was employed for this study. In each of the leave-one-case-out partitions, all ROIs associated with a case were removed and the DL-CNN was trained using the remaining ROIs. The training of the DL-CNN for one partition took approximately 1.5 hours on average using an Nvidia Tesla K20 GPU.

8.3.3 Bladder cancer likelihood map generation using DL-CNN

For each leave-one-case-out partition, the trained DL-CNN network was applied to the removed case to generate the bladder cancer segmentation likelihood map. A bladder cancer likelihood map was generated by applying the trained DL-CNN to a volume of interest (VOI) of a CT scan that contained the bladder tumor to be segmented. In this study, a VOI that approximately enclosed the bladder cancer was manually marked in each CT scan. For every voxel within the VOI, an ROI of 16 x 16 pixels in size centered at the voxel was automatically extracted from the corresponding axial slice and input into the DL-CNN that estimated a likelihood score that the voxel was inside the tumor. After the likelihood values were estimated for all voxels in the VOI, the likelihood values on each slice constituted a 2D likelihood map on the axial slice, and the stack of 2D likelihood maps provided the 3D map for the VOI. Figure 8.3 shows the bladder cancer likelihood map for the CT slice shown in Figure 8.1. The DL-CNN was applied to the CT scan for both the pre- and post-treatment scans for each bladder cancer case.

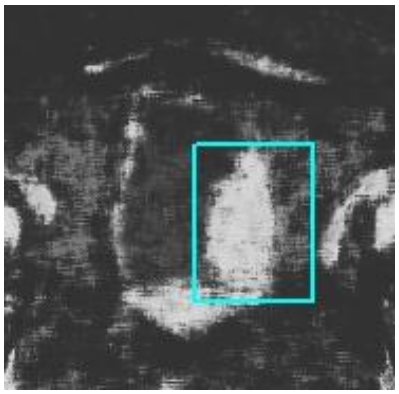


Figure 8.3: Bladder Cancer likelihood map of the CT slice shown in Figure 8.1. Regions that are highly likely to be bladder cancer have higher intensity values. The VOI that was used for this lesion is shown in blue. For demonstration purposes, the bladder cancer likelihood map was generated in the region around the entire bladder.

8.3.4 Bladder cancer segmentation from likelihood map

As seen in the example of Figure 8.3, the likelihood map identifies the bladder tumor region very well but the tumor boundary is not sharply demarcated. The level sets, therefore, are used to perform minor refinements to the contour. First, a binary cancer mask, DL_{Mask} is generated by applying the following equation to every pixel of every slice of the likelihood map:

$$DL_{Mask}(x,y) = \begin{cases} 1, & DL_{Score}(x,y) \geq \rho \\ 0, & DL_{Score}(x,y) < \rho \end{cases} \tag{8.1}$$

where $DL_{Mask}(x,y)$ is the pixel value on the cancer mask at the coordinates (x,y) of a slice, $DL_{Score}(x,y)$ is the bladder cancer likelihood score at the coordinates (x,y), and ρ is the likelihood score threshold. The value of ρ was experimentally determined to be 0.60 to generate reasonable binary masks in comparison to radiologist's manual segmentation. A morphological dilation filter with a spherical structuring element of two voxels in radius, 3D flood fill algorithm, and a morphological erosion filter with a spherical structuring element of two voxels in radius are used to smooth the cancer mask and connect neighboring components in order to extract an initial segmentation surface, $\phi_0(x)$.

The initial segmentation surface is refined using level sets. For this study, the level set uses the following equation:

$$\begin{cases} \frac{\partial}{\partial t}\Psi(x) = -\alpha A(x)\nabla\Psi(x) - \beta P(x)|\nabla\Psi(x)| + \gamma\kappa(x)|\nabla\Psi(x)| \\ \Psi(x, n=0) = \phi_0(x) \end{cases} \tag{8.2}$$

where $\alpha, \beta,$ and γ are the coefficients for the advection, propagation, and curvature terms, respectively, $A(x)$ is a vector field image that drives the contour towards regions of high gradient,

$P(x)$ is a scalar speed term between 0 and 1 causing the contour to expand at the local rate, and $\kappa(x) = div\left(\frac{\nabla\Psi(x)}{|\nabla\Psi(x)|}\right)$ is the mean curvature of the level set at point x. The symbol ∇ denotes the gradient operator and div is the divergence operator[28]. n is the number of iterations.

A 3D level set with a predefined set of parameters is applied to the initial segmentation surface, and the segmentation on each slice is further refined by a 2D level set. The parameters of the 3D and 2D level sets are presented in Table 8.1.

Table 8.1: Parameters for the level sets.

Level set:	α	β	γ	n
3D	1	0.4	q	20
2D	4.0	0.2	0.5	10

The 3D level set brings the contour towards the sharp edges, but also expands it slightly in regions of low gradient. The parameter "q" in Table 8.1 is defined to be a linear function $\sigma M + \lambda$ of the 2D diagonal distance M of the VOI box in millimeters (mm), where $\sigma = 0.06, \lambda = -0.11$ as shown previously [28]. After the 3D level set refinement, a 2D level set is applied to every slice of the 3D contours to further refine the contours. Details on the level sets used can be found in the literature[28]. Figure 8.4 shows the final contour of the bladder cancer on the CT slice from Figure 1 using the likelihood map shown in Figure 8.3.

Figure 8.4: Bladder cancer segmentation on the CT slice shown in Figure 8.1 using the bladder likelihood map shown in Figure 8.3.

8.3.5 Evaluation methods

Segmentation performance was evaluated by comparing the automatic segmentation results to the 3D hand-segmented contours. The average minimum distance, and the Jaccard index[29] between the hand-segmented contours and computer segmented contours were calculated. The performance metrics are described in Chapter II, section 2.3.5.

Receiver operating characteristics (ROC) analysis and area under the curve (AUC) was used to estimate the accuracy for the prediction of T0 disease (complete response) after surgery based on the calculated change in GTV between pre- and post-treatment CT scans using the DL-CNN, AI-CALS, and the manual segmentation methods. The AUCs from the radiologists' WHO criteria and RECIST estimates were also calculated.

8.4 Results

Examples of DL-CNN segmented bladder cancer on pre- and post-treatment CT scans, along with the AI-CALS segmentation, are shown in Figure 8.5. The segmentation performance measures of both the DL-CNN and AI-CALS methods compared with reference standard 1 averaged over the pre-treatment lesions, post-treatment lesions, and all lesions, along with the p-values from Student's two-tailed paired t-test for the differences between the methods, are presented in Table 8.2. For all lesions, the difference in the average minimum distance was statistically significant with a p-value of 0.001, while the difference in the Jaccard index approached significance with a p-value of 0.058. The differences in the pre-treatment lesion segmentation performances were statistically significant with p-values of less than 0.001 and 0.015 for the average minimum distance and the average Jaccard Index, respectively. The differences in the post-treatment lesion segmentation performances did not reach statistical significance.

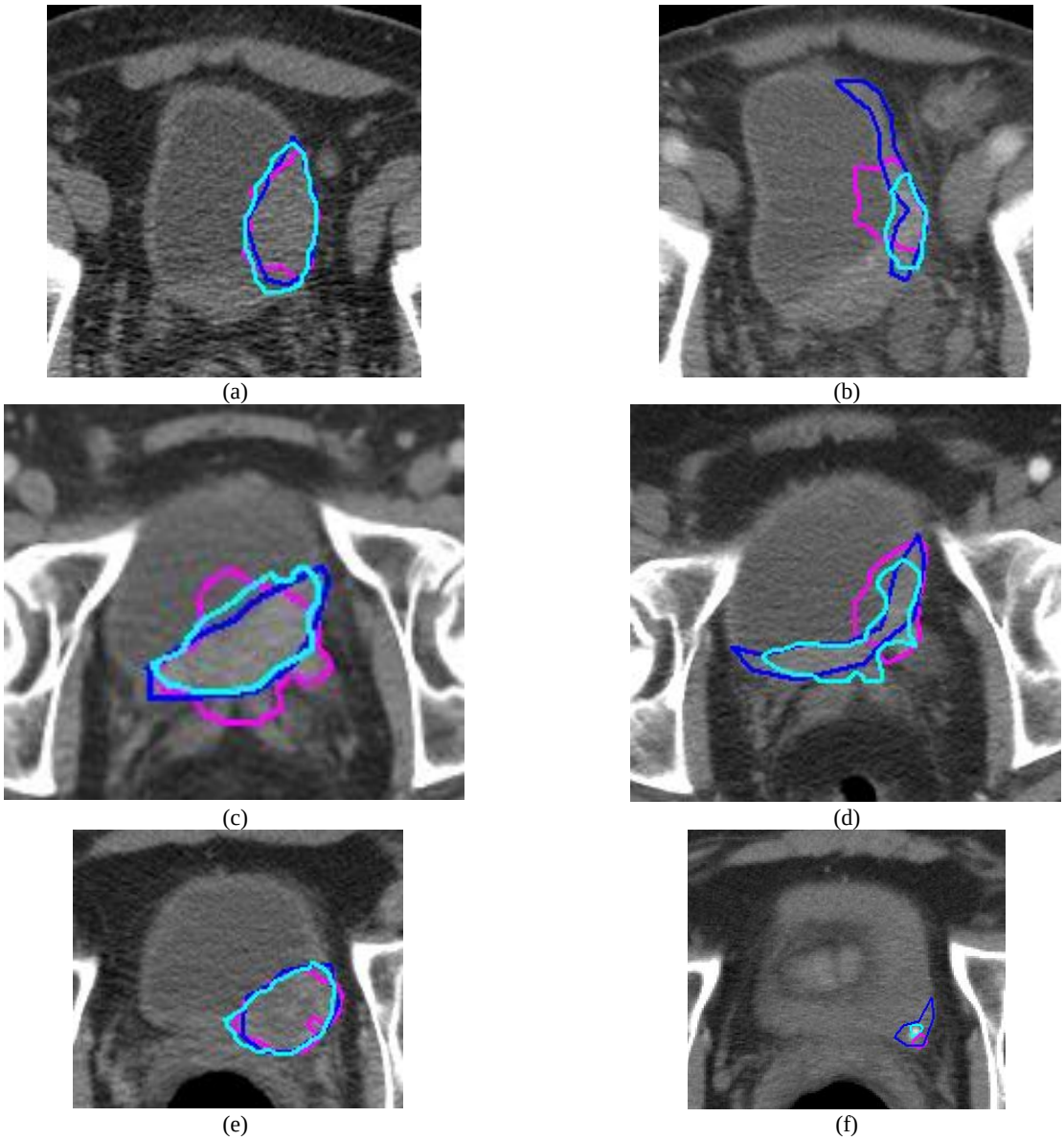


Figure 8.5: Examples of segmentations of bladder tumors in pre-treatment (a, c, e) and post-treatment (b, d, f) CT scans. The DL-CNN segmentation is shown in light blue. The AI-CALS segmentation is shown in pink. The hand outline is shown in dark blue. (a) DL-CNN segmentation with AI-CALS segmentation and hand outline for the cancer shown in Figure 1. Both computer methods segmented the lesion reasonably. (b) The cancer shrunk due to treatment, and became a part of the bladder wall. The DL-CNN under-segmented the cancer, not extending enough into the bladder wall. AI-CALS over-segmented the lesion, leaking into the bladder. (c) The DL-CNN segmentation outlined the cancer relatively accurately, while the AI-CALS segmentation leaked. (d) In this post-treatment scan, the cancer along the bladder wall was reasonably segmented by DL-CNN, while the AI-CALS was unable to follow the shape and leaked into the bladder. (e) Both DL-CNN and AI-CALS segmented the bladder cancer reasonably well, but the AI-CALS slightly under-segmented the cancer. (f) The bladder cancer responded to treatment, thus had shrunk considerably, making the segmentation difficult. Both the DL-CNN and the AI-CALS under-segmented the lesion.

Table 8.2: Segmentation evaluation using reference standard 1 (RS1). The results are shown in groups of pre-treatment, post-treatment, and both pre- and post-treatment lesions (126 lesions). The p-values from Student's two-tailed paired t-test for the differences between the DL-CNN and the AI-CALS segmentation methods are also shown. Some post-treatment lesions were determined to have shrunk completely by radiologist, thus no segmentation was performed.

		DL-CNN vs. RS1	AI-CALS vs. RS1	p-value
Average minimum distance *AVDIST*	Pre-treatment	4.8±2.3 mm	6.1±3.6 mm	0.001*
	Post-treatment	4.6±1.8 mm	4.9±2.6 mm	0.389
	Both	4.7±2.1 mm	5.5±3.2 mm	0.001*
Jaccard index *JACCARD*^{3D}	Pre-treatment	39.5±17.1%	34.7±15.8%	0.015*
	Post-treatment	32.6±17.8%	32.7±14.4%	0.936
	Both	36.3±17.7%	33.8±15.1%	0.058

* Statistically significant at $p < 0.05$

The segmentation performance measures of the DL-CNN and AI-CALS methods compared with the two reference standards averaged over the pre-treatment lesions, post-treatment lesions, and both pre- and post-treatment lesions for a subset of 29 cases are presented in Table 8.3. None of the differences reached statistical significance for this subset of cases.

Table 8.3: Segmentation evaluation results for a subset of 29 cases divided into pre-treatment, post-treatment, and both pre- and post-treatment lesions (58 lesions) between hand-segmented reference standards (RS1, RS2) by two different readers for DL-CNN and the AI-CALS segmentation methods. None of the paired differences between the two methods reached statistical significance for this subset, probably due to the small sample size.

		DL-CNN vs. RS1	AI-CALS vs. RS1	DL-CNN vs. RS2	AI-CALS vs. RS2
Average minimum distance *AVDIST*	Pre-treatment	4.8±1.8 mm	5.3±2.7 mm	4.9±3.4 mm	4.5±1.9 mm
	Post-treatment	4.3±1.7 mm	4.4±1.8 mm	4.7±3.1 mm	4.9±3.7 mm
	Both	4.6±1.8 mm	4.8±2.3 mm	4.8±3.2 mm	4.7±2.9 mm
Jaccard index *JACCARD*^{3D}	Pre-treatment	45.3±8.5%	42.5±14.1%	46.8±9.3%	42.8±12.5%
	Post-treatment	29.8±17.7%	32.9±14.8%	28.8±19.7%	28.6±18.2%
	Both	37.5±15.8%	37.7±15.2%	37.8±17.8%	35.7±17.1%

Table 8.4 shows the AUC values for the different methods. The AUC for the prediction of complete response using GTV calculated using the DL-CNN segmentation achieved 0.73 ±

0.06, while the AI-CALS segmentation achieved 0.70 ± 0.07, compared to 0.70 ± 0.06 for the radiologist's hand outline based GTV. The differences between the three methods did not reach statistical significance. For the WHO criteria, the AUCs were 0.63 ± 0.07 and 0.61 ± 0.06 for the two radiologists. For the RECIST estimates, the AUCs were 0.65 ± 0.07 and 0.63 ± 0.06 for the two radiologists. Figure 8.6 compares the ROC curves for the different methods.

Table 8.4: AUC values for prediction of cancer stage pT0 after surgery.

Method	AUC
Volume	
DL-CNN	0.73 ± 0.06
AI-CALS	0.70 ± 0.07
Hand Outline	0.70 ± 0.06
WHO criteria	
Radiologist 1	0.63 ± 0.07
Radiologist 2	0.61 ± 0.06
RECIST	
Radiologist 1	0.65 ± 0.07
Radiologist 2	0.63 ± 0.06

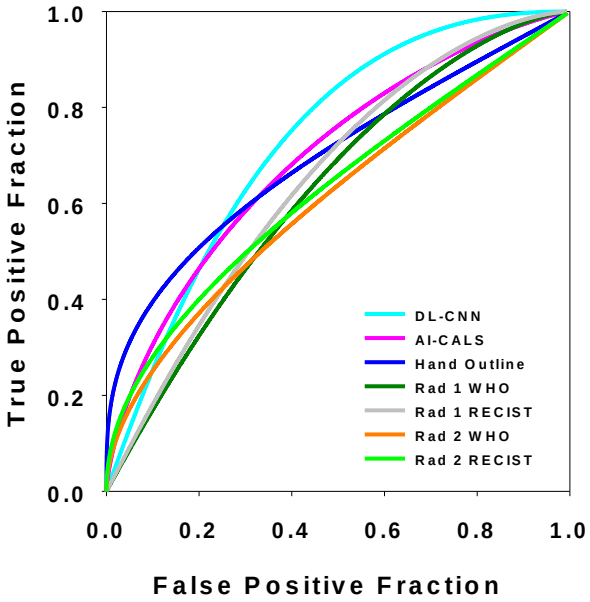


Figure 8.6: ROC curves for the prediction of complete response to chemotherapy. The AUCs for GTV-based estimates were 0.73 ± 0.06 for DL-CNN, 0.70 ± 0.07 for AI-CALS, 0.70 ± 0.06 for the radiologist's hand outlines; The AUCs for the WHO criteria based estimates were 0.63 ± 0.07 for radiologist 1 (Rad 1) and 0.61 ± 0.06 for radiologist 2 (Rad 2); and the AUCs for the RECIST based estimates were 0.65 ± 0.07 for radiologist 1 and 0.63 ± 0.06 for radiologist 2.

8.5 Discussion

The results of this pilot study show that DL-CNN can be trained to segment bladder cancers in CT. The trained DL-CNN generates a likelihood map that identifies regions in the CT scans that are likely to be bladder cancers. By thresholding this map and performing minor refinement with level sets, we are able to obtain reasonable bladder cancer segmentations.

In our previous work on bladder segmentation with DL-CNN, level sets were also used to refine the output of the DL-CNN; as the contrast between the bladder and its surroundings is relatively high, cascaded level sets can refine the bladder boundary with high accuracy. On the other hand, with bladder cancer segmentation, the contrast between the tumor and the inside of the bladder can be much lower because contrast material is often not used for the pre- and post-treatment CT. The level sets did not perform as well under these conditions, causing the segmentation to leak or shrink incorrectly. Therefore, only a few iterations of the level sets were applied to the bladder cancer likelihood map to fill holes and smooth the segmentation.

The segmentation performance of the DL-CNN was better than the performance of AI-CALS in comparison with a radiologist's reference standard using the entire data set. The differences in the average minimum distance were statistically significant, while the difference in the Jaccard index approached significance (Table 8.2). When the data set is divided into pre-treatment and post-treatment lesions, the DL-CNN performed significantly better than the AI-CALS for the pre-treatment lesions, while the two methods performed comparably on the post-treatment lesions. The pre-treatment lesions are generally better defined than the post-treatment lesions. As the lesions change due to the treatment, the lesion generally shrink and the boundaries become less distinct, making the post-treatment lesions more difficult to segment. Nevertheless, the changes in GTV estimated by the two computer methods were comparable to the estimates using radiologist's hand outlines at prediction of complete response to treatment (Table 8.4).

The segmentation performances for the DL-CNN and the AI-CALS were compared with two reference standards in the subset that had both radiologists' manual outlines (Table 8.3). The results indicate that the performances of the two methods were consistent regardless of which reference standard was used.

The change in GTV estimated by the new DL-CNN segmentation method performed better than that by our previous AI-CALS system at prediction of complete response to treatment.

While the difference in the AUC for prediction of complete response to treatment did not reach statistical significance, probably due to the small data set, the segmentation performance results show that the segmentation by DL-CNN is better than that by AI-CALS.

Comparisons of the volume measurements with the WHO criteria and the RECIST show that 3D measure of the bladder cancer (GTV) performs better than the 2D (WHO criteria) and the 1D (RECIST) measurements. The WHO criteria and the RECIST measurements performed worse compared to the GTV measures. This indicates that the 3D information (GTV) may be more reliable for assessment of bladder cancer treatment response.

There are limitations in this study. While the data set was expanded compared to our previous study, the number of cases was still relatively small. Testing the method on a larger data set with wider ranges of sizes and types of bladder cancers would allow us further validate the generalizability of the method. We will continue to enlarge the data set. Another limitation is that we have only one set of radiologist's hand segmentations as reference standard for the entire data set. In order to study the inter- and intra-observer variability in the hand segmentations of the bladder cancer, additional independent hand segmentations by different radiologists would be needed.

Bladder cancer segmentation is important as it defines the regions to be analyzed for the characterization of the lesion. We plan to expand our work to investigate if radiomics features extracted from the segmented bladder cancers, in conjunction with the GTV change, can improve the assessment of response to chemotherapy or other treatments. Although the DL-CNN method shows more promising results than the AI-CALS method at present, there is still more room for improvement on the segmentation performance of both methods, especially for the post-treatment tumors. Further development and validation with a larger data set are also needed to confirm the relative performance of the two approaches.

8.6 Conclusion

Our results demonstrate that DL-CNN is useful for 3D segmentation of bladder cancers for a variety of bladder cancer shapes and sizes. The DL-CNN and the AI-CALS methods were able to automatically segment the cancers, with results similar to those of the radiologists. The 3D information from CT may provides more accurate information on the changes in the tumor size in response to treatment compared to the 2D (WHO criteria) and 1D (RECIST) estimations

used in current clinical practice. This study suggests that computerized segmentation of bladder cancers using DL-CNN has the potential to assist in the assessment of tumor volume and treatment response of bladder cancer by providing the more accurate 3D information without the extensive effort of manual segmentation.

Chapter IX

Bladder Cancer Treatment Response Assessment in CT using Radiomics with Deep-Learning

9.1 Abstract

Cross-sectional X-ray imaging has become the standard for staging most solid organ malignancies. However, for some malignancies such as urinary bladder cancer, the ability to accurately assess local extent of the disease and understand response to systemic chemotherapy is limited with current imaging approaches. In this study, we explored the feasibility that radiomics-based predictive models using pre- and post-treatment computed tomography (CT) images might be able to distinguish between bladder cancers with and without complete chemotherapy responses. We assessed three unique radiomics-based predictive models, each of which employed different fundamental design principles ranging from a pattern recognition method via deep-learning convolution neural network (DL-CNN), to a more deterministic radiomics feature-based approach and then a bridging method between the two, utilizing a system that extracts radiomics features from the image patterns. Our study indicates that the computerized assessment using radiomics information from the pre- and post-treatment CT of bladder cancer patients has the potential to assist in assessment of treatment response. The results presented in this chapter have been published[5].

9.2 Introduction

Radical cystectomy provides the best local control for patients with localized muscle invasive or recurrent non-muscle invasive bladder cancer. Despite adequate local cancer control, approximately 50% of patients who have undergone cystectomy develop metastases within two years of cystectomy and subsequently die of the disease. This is likely due to the presence of regional or distant microscopic metastatic disease at the time of surgery.

Neoadjuvant chemotherapy prior to cystectomy utilizing a cisplatin-based regimen has been shown to decrease the probability of finding extravesical disease when compared to radical cystectomy alone and also improve survival[58-60]. Patients with a complete local response within

their bladder following neoadjuvant chemotherapy (approximately 30% of the patients) have 5-year recurrence free survival, equivalent to patients who undergo cystectomy for non-muscle invasive disease (85-90%). Currently, there is no reliable method for predicting the response of an individual case to neoadjuvant chemotherapy prior to or during its administration. As a result, some patients may suffer from adverse reactions to the drugs without achieving beneficial effects. Furthermore, these patients may miss the opportunity to receive alternative therapy as a result of the deterioration of their physical condition from the initial chemotherapeutic regimen.

Early assessment of therapeutic efficacy and prediction of treatment failure would help clinicians decide whether to discontinue chemotherapy at an early phase before additional toxicity develops, thus improve the quality of life of a patient and reduce unnecessary morbidity and cost. The ultimate goal is to improve survival for those with a high risk of recurrence while minimizing toxicity to those who will have minimal benefit. Therefore, development of an accurate and early predictive model of the effectiveness of neoadjuvant chemotherapy is important for patients with bladder cancer. Many bladder cancers are well visualized with CT[93-95]. It might be possible to characterize the responsiveness of a cancer to chemotherapy by changes in its imaging characteristics.

Radiomics is the study of using radiological images to analyze anatomic or physiologic abnormalities based upon the imaging characteristics[96-98]. Previous studies decoded tumor phenotypes[99] using radiomics features, or evaluated the correlation between radiomics features with radiation therapy dose and radiation pneumonitis[100] in lung CT. A review paper on radiomics for lung cancer concluded that radiomics have the potential to improve lung cancer diagnosis[101].

One promising method that may be useful for extracting image information for radiomics application is convolution neural network (CNN). A CNN learns the underlying patterns and features in the input training images using many small convolution kernels that contain the weights that the network incorporates the knowledge learned from the training images and generalizes it to recognize unknown images outside the training set. CNN has been used for many years and has been implemented successfully to classify different types of medical image patterns[31, 34-38, 90, 102, 103]. With the recent increase in the power for parallel computing, primarily resulting from the development of powerful graphics processor units (GPUs), training of a very large CNN with many layers and connections in the network has become feasible. Deep-learning

convolutional neural networks (DL-CNN) are CNNs capable of learning complex patterns using large numbers of input images via a large neural network with many connections. The application of DL-CNN to various tasks in medical image analysis has been examined recently[2, 104-106].

We have shown in Chapter VIII that the change in gross tumor volume can be used to distinguish between bladder cancers that have fully responded to chemotherapy and those that have not. In this study, we explored the possibility that radiomics-based predictive models might be able to distinguish between bladder cancers that have fully responded to chemotherapy and those that have not, based upon analysis of pre- and post-treatment CT images. We evaluated three unique radiomics predictive models that employ different fundamental design principles: 1) a pattern recognition method (DL-CNN), 2) a more deterministic radiomics feature based approach (RF-SL), and 3) a bridging method between the two that extracts radiomics features from image patterns (RF-ROI). We studied both the properties of the different predictive models and the relationship between these different radiomics approaches. We also compared the performance of the models in predicting a complete response of bladder cancer to neoadjuvant chemotherapy with that of expert physicians.

9.3 Materials and Methods

9.3.1 Data set

The patient images were collected and de-identified using methods approved by our Institutional Review Board (IRB) and are HIPAA compliant. All methods were performed in accordance to the guidelines and regulations of the IRB. A training data set of 82 patients (67 males, 15 females, age 64.0 ± 10.6, age range 37-84) with 87 bladder cancers was collected. The CT scans used in this study were acquired with the process described in Chapter II, section 2.3.4. A total of 172 CT scans (pixel size range 0.586 – 0.977 mm, slice thickness range 0.625 – 7 mm) were obtained for the training set, of which 28 scans were performed without contrast material, while the remaining 144 scans were contrast-enhanced CT scans. Using the 87 lesions, 104 temporal lesion pairs were generated. Twenty-seven percent (28/104) of the lesions pairs had T0 cancer after neoadjuvant chemotherapy that corresponds to a complete response to treatment, using the clinical information available from the patient files as reference truth. These temporal

lesion pairs were used to generate 6,700 pre-post-treatment paired ROIs for use with DL-CNN and RF-ROI.

In addition, a test data set of 41 patients (33 males, 8 females, age 60.9 ± 9.2, age range 42-84) with 42 lesions was collected. A total of 88 CT scans (pixel size range 0.586 – 0.977 mm, slice thickness range 0.5 – 7.5 mm) were obtained for the test set, of which 16 scans were performed without contrast material, while the remaining 72 scans were contrast-enhanced CT scans. Fifty-four temporal pairs were generated from the 42 lesions. Twenty-two percent (12/54) of the lesion pairs had T0 cancer after neoadjuvant chemotherapy. Chemotherapy regimens used for the majority of these patients were MVAC (methotrexate, vinblastine, doxorubicin, and cisplatin treatment), while other patients were treated with regimens including carboplatin, paclitaxel, gemcitabine, and etoposide. The pre-treatment scans were acquired approximately 1 month (max 3 months) before the first cycle of chemotherapy. The post-treatment images were acquired after the completion of three cycles of chemotherapy, generally within 1 month (max 2 months) of cessation of the therapy. The pre-post-treatment scans were on average acquired 4 months apart. Each patient underwent cystectomy at the end of his or her chemotherapy, usually 4-6 weeks after completion of neoadjuvant chemotherapy, which is generally within 4 months after the post-treatment CT scan. Pathology obtained from the bladder at the time of surgery was used to determine the final cancer stage after chemotherapy and was used as the reference standard to determine whether or not the patient had responded completely to treatment.

9.3.2 Lesion segmentation

The AI-CALS system was used for the segmentation of the bladder lesions. Additional information regarding the AI-CALS system can be found in literature[47]. Figure 9.1 shows examples of segmented lesions.

9.3.3 DL-CNN predictive model

To train the DL-CNN with both the pre- and post-treatment bladder lesion information, we generated a single image containing information from the bladder lesion at the two time points. From every slice of a segmented bladder lesion, ROIs with dimension 16 x 32 pixels that were shifted with respect to one another but stayed within the segmented bladder lesion were

extracted. If the bladder lesion was smaller than the ROI size, a single ROI centered at the centroid of the bladder lesion was extracted. Once these ROIs were extracted from both the pre- and post-treatment CT of a temporal lesion pair, the ROIs were combined by pasting them side-by-side, with a pre-treatment lesion ROI located on the left half and a post-treatment lesion ROI located on the right half of the combined ROI. The resulting image was a paired 32 x 32-pixel ROI containing both pre-and post-treatment information (Figure 9.2). Different combinations of the pre- and post-treatment lesion ROIs from the same case were used to form multiple paired ROIs. Each paired ROI was labeled as having or not having responded completely to treatment based on whether the cancer went down to stage T0 or not after treatment as determined by clinical information from the patient files. If the post-treatment ROI of a paired ROI was stage T0, the paired ROI was labeled as having complete response, but if the stage was greater than T0, the paired ROI was labeled as not having complete response. Figure 9.2 shows an example of how a paired ROI was formed and examples of paired ROIs and their labels. A total of 6,700 pre-post-treatment paired ROIs were generated (Figure 9.3).

A DL-CNN learns patterns of different classes from a set of labeled training images that are representative of the population of patterns being classified. After the user specifies a few parameters that control the training process, the DL-CNN reads the labeled images and learns the differences between the classes without any additional inputs. We applied the DL-CNN developed by Krizhevski et al., called cuda-convnet[39, 40], to the classification of the paired ROIs with complete response, and those without. The parameters of the original network were modified, and a set of parameters was found that worked best for the task of identifying cases that completely responded to treatment using the training set. The network consists of five main layers: two convolution layers, two locally-connected layers, and one fully-connected layer. The locally-connected layers perform the same operation as the convolution layer except that, instead of applying a single convolution kernel to every location of the input image to obtain a kernel map, different convolution kernels are applied to every location of the input image and the resulting values are collected into the corresponding neurons within the corresponding kernel map. The fully-connected layer uses every kernel map element multiplied by a weight as input. Pooling and normalization are performed after each of the two convolution layers. All convolution layers consist of 16 kernels with a kernel filter size of 3x3. The “per-lesion” score was obtained by using the average value among the ROIs associated with the lesion. We trained

a DL-CNN to distinguish between bladder lesions that were diagnosed as stage T0 post-treatment (no residual tumor) and those that were greater than stage T0 (any residual tumor) (Figure 9.4).

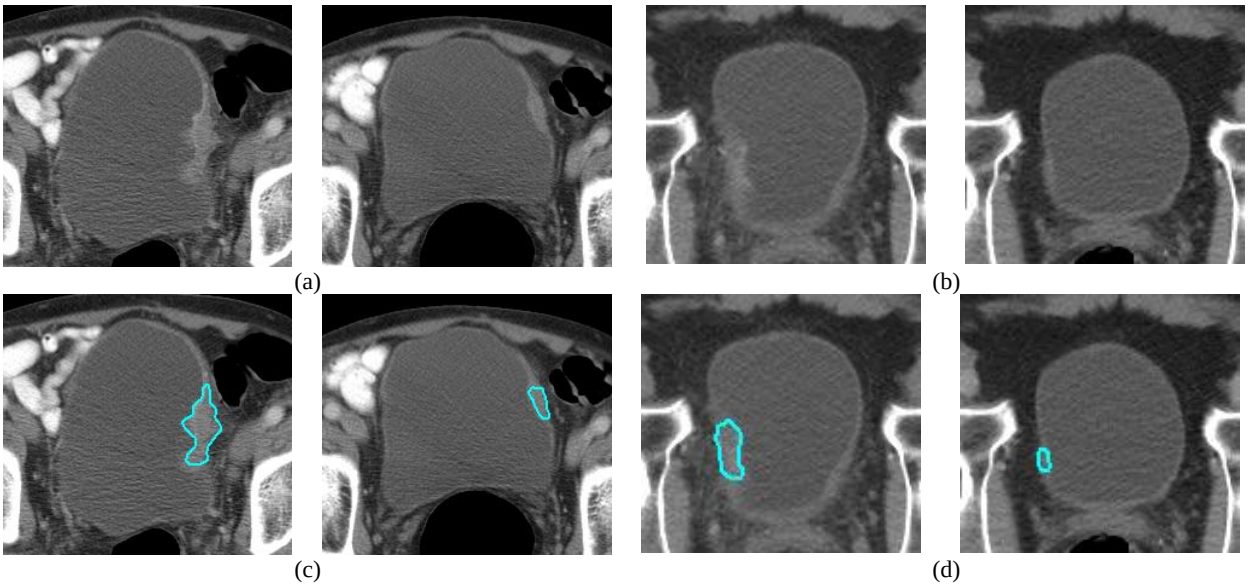


Figure 9.1: Bladder lesion segmentations. Two segmented bladder cancers are illustrated. The lesions in the pre- and post-treatment scan pairs shown in (a) and (b) are segmented using AI-CALS, as shown in (c) and (d), respectively. The pre-treatment scan is on the left and the post-treatment scan is located on the right of each pair.

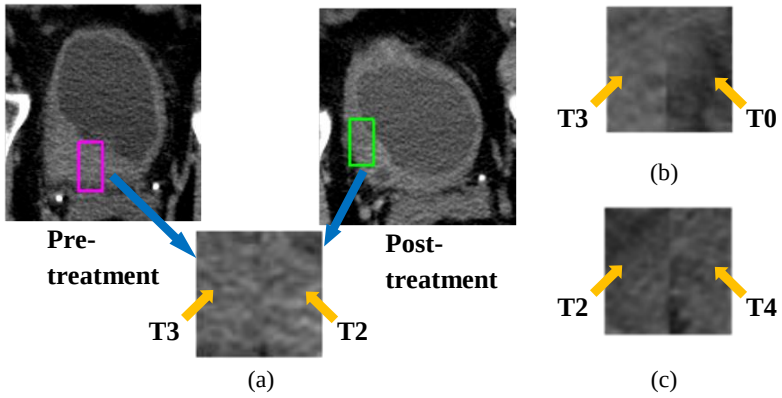
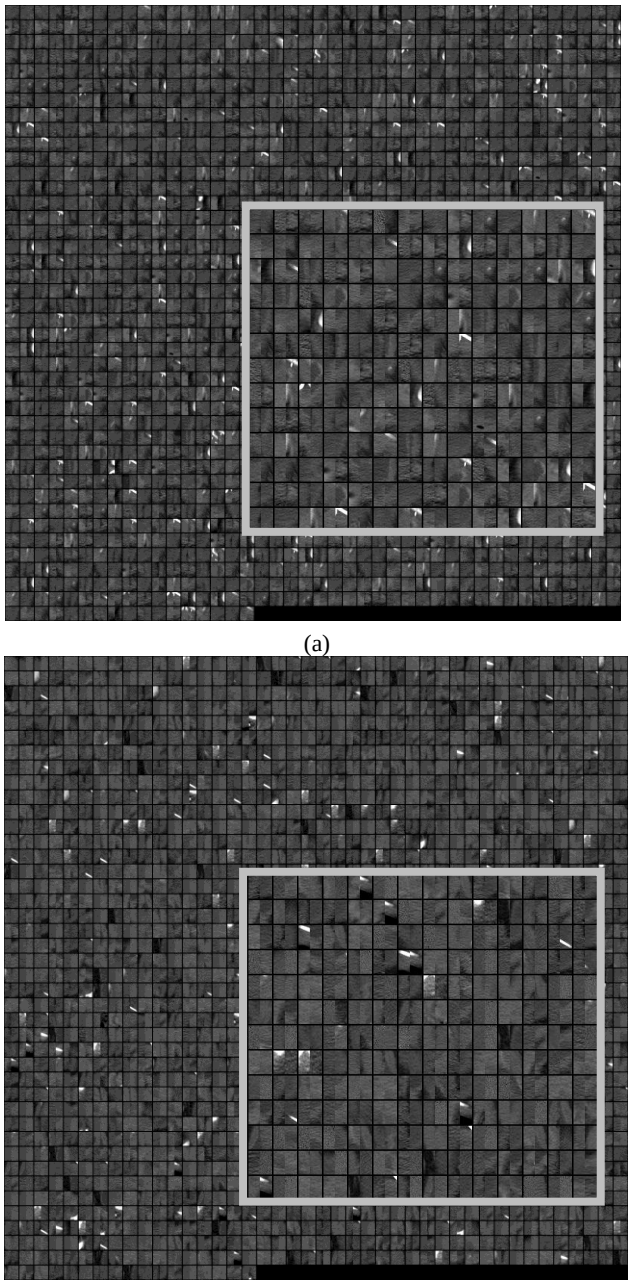


Figure 9.2: Creating ROIs to train the DL-CNN. (a) ROIs were generated by combining regions from the pre- and post-treatment scan lesions. In this example, the pre-treatment stage was T3, and the post-treatment stage was T2. Therefore, the ROI was labeled as being greater than stage T0 after treatment. (b) ROI of a case that was stage T3 pre-treatment and stage T0 after treatment. (c) ROI of a case that was stage T2 pre-treatment and stage T4 post-treatment. Therefore the ROI was labeled as greater than stage T0 after treatment.



(a)

(b)

Figure 9.3: Subset of Paired ROIs used to train the DL-CNN. Each ROI is 32x32 pixels. (a) ROIs that were labeled as being stage T0 after treatment. (b) ROIs that were labeled as being greater than stage T0 after treatment. A portion of ROIs in each class is zoomed in to illustrate the content of typical ROIs.

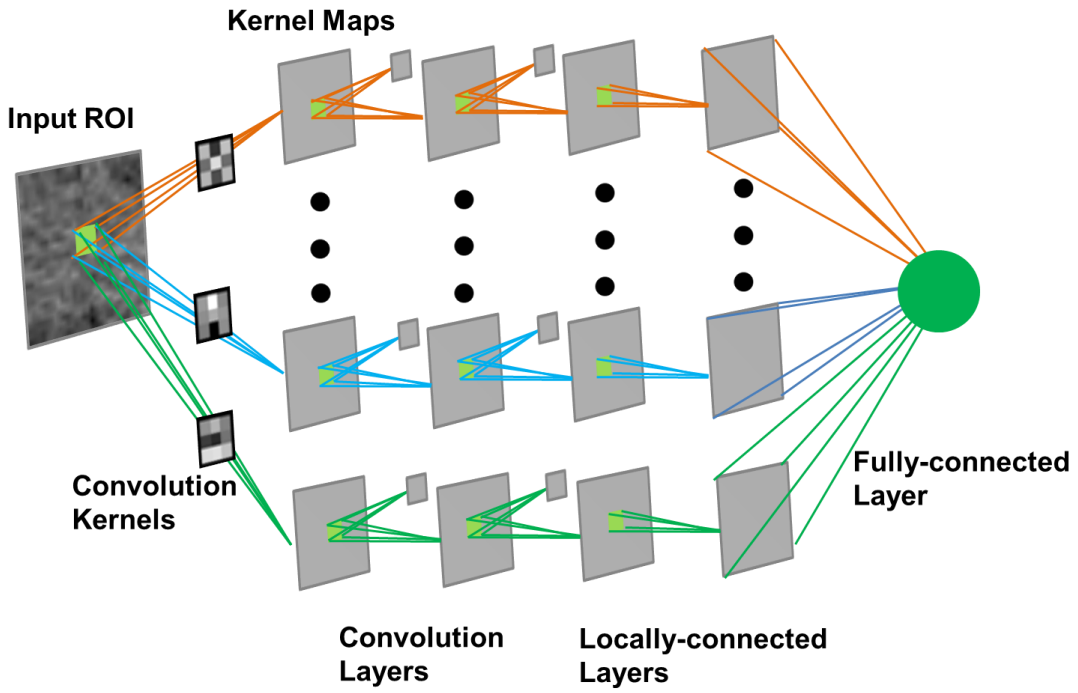


Figure 9.4: DL-CNN Structure. An input ROI is convolved with multiple convolution kernels, and the resulting values are collected into the corresponding kernel maps. This process repeats for several layers, giving the "deep" convolutional neural network. The network used in this study contains two convolution layers and two locally-connected layers, each of which contains 16 kernels.

9.3.4 RF-SL predictive model

For this predictive model, a radiomics-feature-based analysis was applied to the segmented lesions (RF-SL). We extracted 91 radiomics features from every segmented lesion. Four types of features were extracted: (1) morphological features, such as volume, circularity, rectangularity, and Fourier descriptor, (2) gray level features, such as the average gray level and contrast features, (3) texture features, such as run length statistics, and (4) gradient field features, such as the gradient magnitudes statistics for all voxels on the surface of the segmented lesion. Shape-based features and the contrast features were extracted from the central slice of the lesion. Other features, such as the gray-level features, texture features, and the gradient field features, were extracted from the segmented lesion in 3D after performing interpolation to obtain isotropic images. For every temporal lesion pair, the percent change between each radiomics feature extracted from the pre- and post-treatment lesion was calculated. These features were found to be potentially useful for lesion classification from our previous experience with breast masses and lung nodules[30, 50]. Features that were useful specifically for the bladder cancer treatment response assessment were further identified during the training of the predictive model. The

percent change of each of the feature values before and after the treatment was calculated. Feature selection was performed and a random forest classifier using 6 trees and the minimum number of observations per tree leaf set of 13 was trained to use the selected radiomics features to predict the likelihood of the post-treatment lesion being T0 stage. The parameters were selected experimentally using the training set.

9.3.5 RF-ROI predictive model

In this model, the radiomics-feature-based analysis was applied to the paired ROIs (RF-ROI). Gray-level and texture features were extracted from the paired ROIs used for the DL-CNN. Thirty-eight features, including gray-level histogram statistics, and run length statistics features, were calculated for every ROI. The "per-lesion" features were generated by averaging the feature values among the ROIs associated with the lesion. Similar to the RF-SL model, feature selection was performed and a random forest classifier using 2 trees and the minimum number of observations per tree leaf set of 29 was trained to use the selected radiomics features to predict the likelihood of the post-treatment lesion being T0 stage. The parameters were selected experimentally using the training set.

9.3.6 Expert physician performance

An observer study with two experienced fellowship-trained abdominal radiologists, one with 17 years of experience and the other with 27 years of experience, was performed as references for comparison with the three predictive models. Each radiologist independently read the pre- and post-treatment images of patients that were loaded side-by-side, and estimated the likelihood of the patient having stage T0 cancer post-treatment. The images were presented in a randomized order to reduce potential observer bias.

9.3.7 Performance evaluation

For the three methods, the trained model was applied to the test set and the area under the curve (AUC) was calculated using the scores from the test cases. For the radiologists, the AUC was calculated for the test cases using their given likelihood of complete response to treatment.

9.4 Results

Receiver operating characteristic (ROC) analysis was performed and the area under the curve (AUC) was calculated as a measure of performance. Figure 9.5 shows the ROC curves for the DL-CNN, RF-SL, and RF-ROI methods and the radiologists for the test set. Table 9.1 shows the performances for the DL-CNN, RF-SL, and RF-ROI methods, along with the radiologists' results for the test set.

Table 9.1: Performances of bladder cancer treatment response assessment in the test set.

	DL-CNN	RF-SL	RF-ROI	Radiologist 1	Radiologist 2
AUC	0.73 ± 0.08	0.77 ± 0.08	0.69 ± 0.08	0.76 ± 0.08	0.77 ± 0.07

DL-CNN: Deep-learning convolution neural network
RF-SL: Radiomics features extracted from segmented lesions
RF-ROI: Radiomics features extracted from pre- and post-treatment paired ROIs
The area under the curve (AUC) is shown with the standard deviations

The areas under the curve (AUC) for prediction of T0 disease after treatment were similar. The performances of all three methods are comparable to those of the radiologists. The differences between any two AUCs did not reach statistical significance. Examples of the treatment response prediction of pre- and post-treatment case pairs are shown in Figure 9.6. Figure 9.7 shows examples of cancers that responded fully to treatment and the differences in the predictions by the computer models and the radiologists.

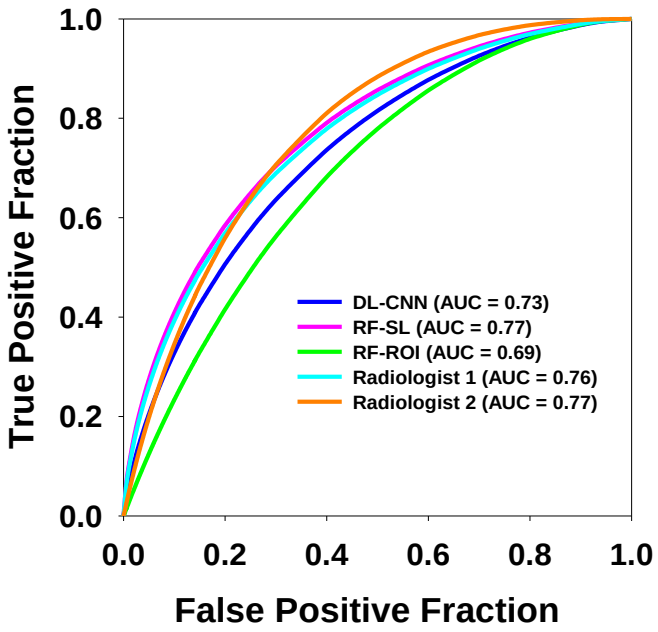


Figure 9.5: Test set ROC curves for the three models and two expert radiologists. The results from the test set for prediction of T0 stage after neoadjuvant chemotherapy for the three models. The differences between any pairs of AUCs did not reach statistical significance.

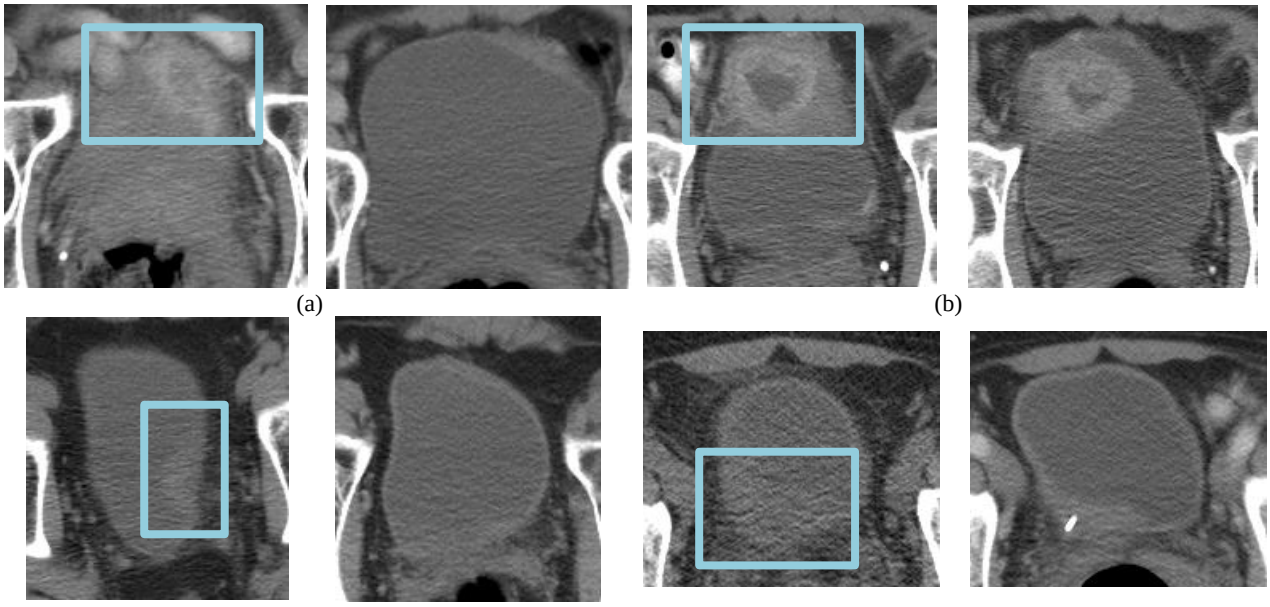


Figure 9.6: Examples of pre- and post-treatment bladders and their predictions. (a) The computer methods and the radiologists correctly predicted the treatment outcome for this case, which was a non-responding, progressive disease that went from stage T2 before treatment to T3a after treatment. (b) In this stable disease case (stage T3), the computer methods and the radiologists correctly identified the case as non-responding. (c) This case fully responded, going from stage T2 to T0, and the computer methods and the radiologists correctly predicted the treatment response. (d) A full-responding case, going from stage T3 to T0. The computers correctly predicted the response, while the radiologists did not. The region around the right ureterovesical junction was asymmetrically thickened, which might have misled the radiologist to assess that cancer was present. The pre-treatment scan is on the left and the post-treatment scan is located on the right of each pair. The box on the pre-treatment scan represents the location of the lesion as marked by one of the radiologists.

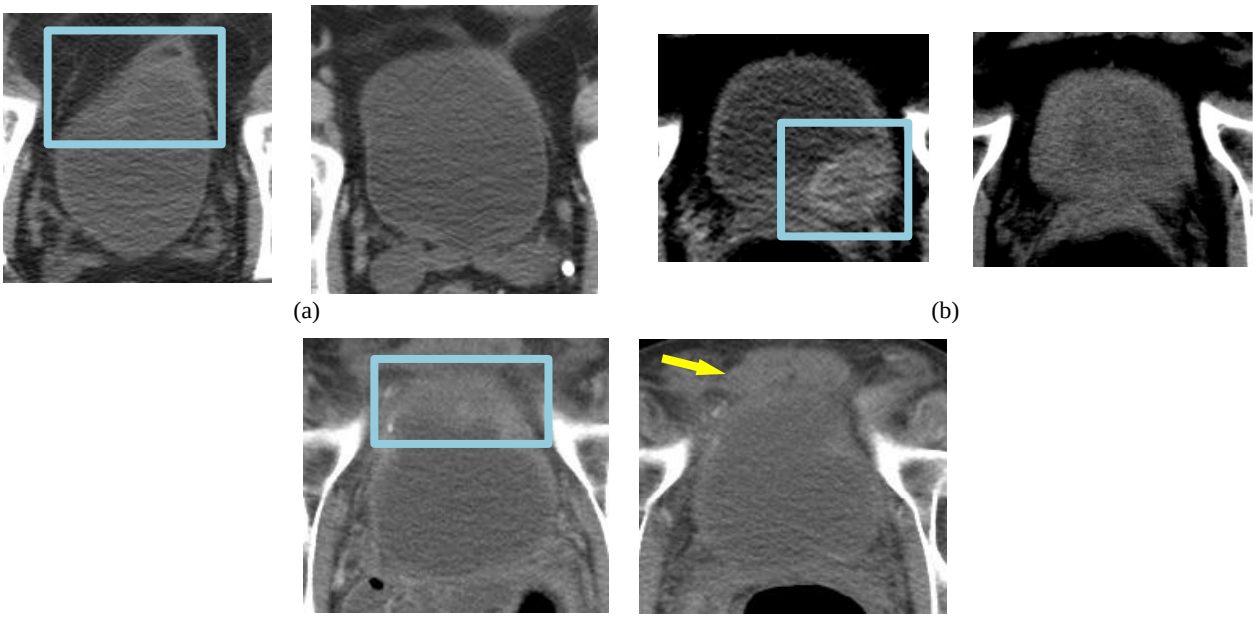


Figure 9.7: Examples of pre- (on the left side of each image pair) and post- (on the right side of each image pair) treatment bladders that responded fully to treatment, and the differences in the predictions by the computer models and radiologists. (a) All three computer methods and the radiologists correctly predicted the outcome of treatment for this case. (b) The three computer methods correctly identified the case as becoming T0 tumor, while the radiologists did not. There was residual bladder wall thickening, presumably related to the treatment, causing the radiologists to falsely conclude that there was persistent tumor. (c) The radiologist correctly identified that there was no residual tumor on post-treatment images, while the three computer methods failed to classify this case correctly. This was likely due to misidentification of perivesical tissue (arrow) as residual tumor by the computer models. The box on the pre-treatment scan represents the location of the lesion as marked by one of the radiologists.

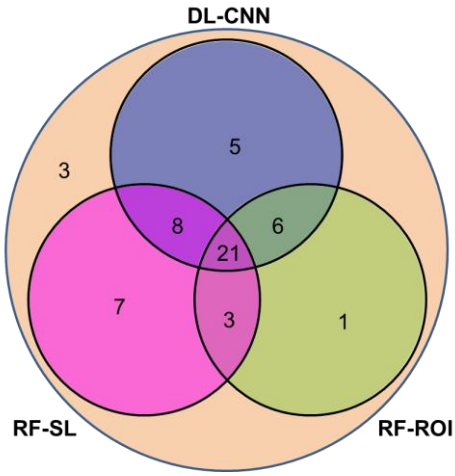
9.5 Discussion

In this study, we evaluated a system that can distinguish between bladder cancers that have completely responded to neoadjuvant chemotherapy from those that have not, based upon computer analysis of pre- and post-treatment CT images. Although we used CT images before and after completion of three cycles of chemotherapy, we expect that the trained predictive models will be applicable to any treatment time points because they are trained to recognize the change in the image patterns to T0 stage regardless of when it occurs. If the models are validated, the models may be used to assess treatment response at any clinically relevant time point, and the treatment may be stopped or changed prior to the appearance of other toxic effects if the cancer is resistant to the treatment. The performance of the DL-CNN was compared against a radiomics feature-based method, where the percent change in the features extracted from the segmented lesions pre- and post-treatment was used (RF-SL), and against a third method extracting radiomics features from the paired ROIs used by the DL-CNN (RF-ROI).

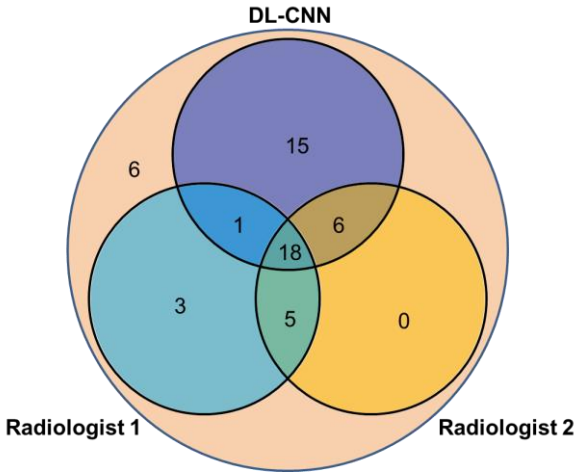
For the RF-SL, five features were consistently selected which included a contrast feature and four run length statistics texture features. We have also performed an ordering of importance on the radiomics features based on the AUC of the individual features, and found that the selected features were highly ranked. For the RF-ROI, the gray-level average, the skewness of the gray-level histogram, and two run length statistics texture features were consistently selected. These results show that the texture, which characterizes the heterogeneity, of the bladder lesions is an important indicator for the estimation of full responders to chemotherapy.

All three methods performed comparably to the two expert radiologists. The RF-SL performed slightly better than the DL-CNN; however, the RF-ROI method resulted in worse performance compared to the DL-CNN, indicating that the DL-CNN is able to better characterize the paired ROIs to identify full responders compared to extracting features from the ROIs and using the random forest classifier. The absence of contrast material in some of the CT scans did not affect the results of the computer system or the radiologists. The pre- or post-treatment

cancer stage did not have an observable effect on the performances of the system or the radiologists either. Although the overall performance was similar across the methods, there are variations in the prediction performance in individual cases. In the case shown in Figure 9.7(a), all methods and radiologists correctly identified that the detected cancer had responded fully to treatment. However, in the case shown in Figure 9.7(b), the computer methods correctly identified the cancer as having a complete response, while the radiologists did not. Upon further review, the radiologists commented that the uncertainty was due to the presence of residual/persistent bladder wall thickening in the region of the tumor (which turned out to be benign at surgery). In the case shown in Figure 9.7(c), the radiologists correctly identified the case as having responded fully to treatment, whereas the computer indicated that the case had not. The lesion in this case was located at the top of the bladder, and the automated computer segmented region incorrectly extended into the perivesical tissue outside the bladder. The analysis of this extravesical tissue by the computer models might have caused the incorrect decision that residual tumor was present.



(a)



(b)

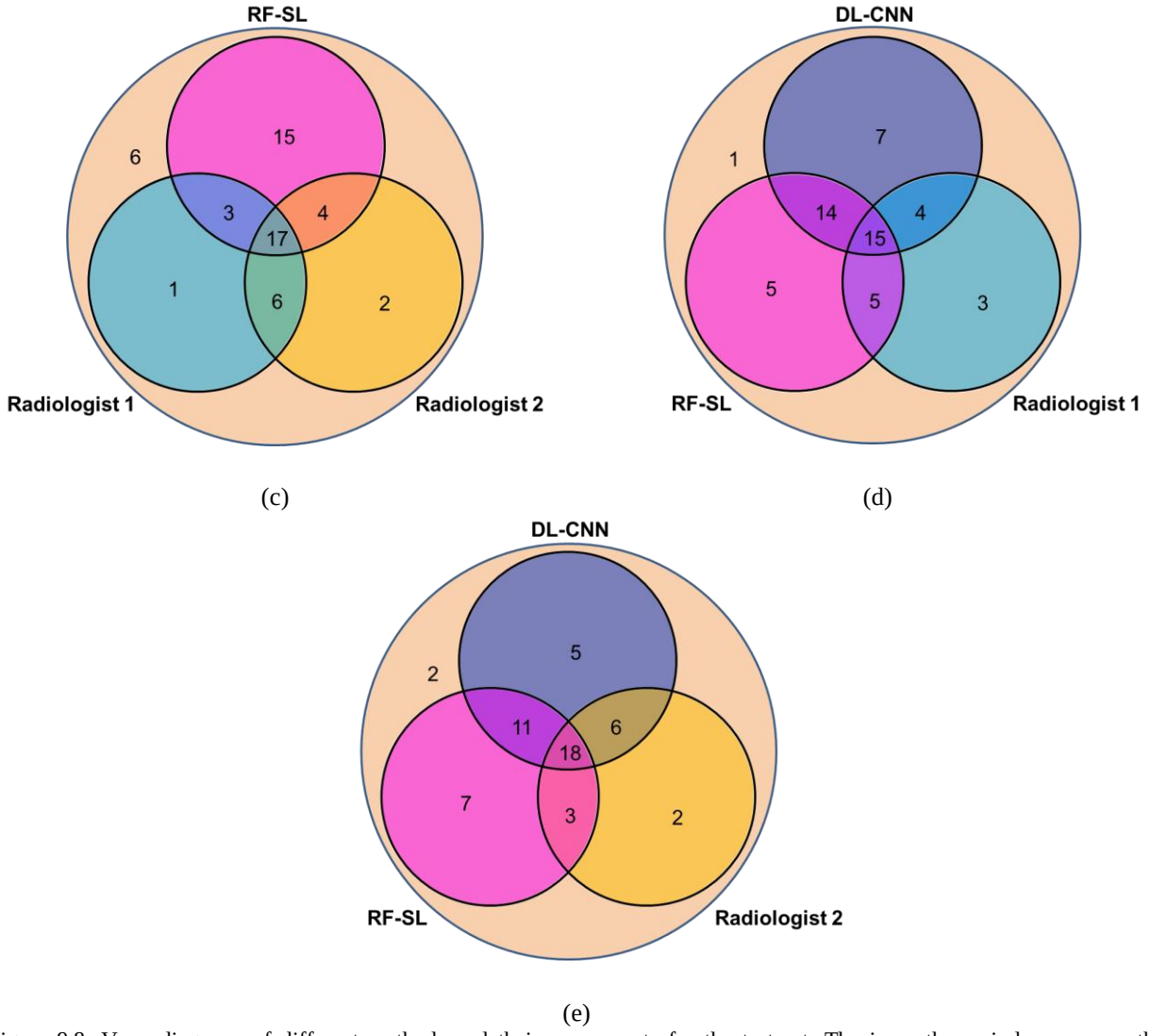


Figure 9.8: Venn diagrams of different methods and their assessments for the test set. The inner three circles compare the methods when at least one method correctly predicted the treatment outcome for a pre- and post-treatment pair. The outer circle contains the pre- and post-treatment pairs for which all three methods incorrectly predicted the treatment outcome. (a) The three computer methods correctly predicted the same outcome of the patients for 39% (21/54) of the pre- and post-treatment pairs. (b-c) The two radiologists correctly agreed on the outcome of 43% of the cases ((18+5)/54 for (b) and (17+6)/54 for (c)). (d-e) Radiologists 1 correctly agreed with the DL-CNN and the RF-SL methods for 19 and 20 cases, respectively, while Radiologist 2 correctly agreed with the DL-CNN and the RF-SL methods for 24 and 21 cases, respectively.

Given the fact that in some instances the computer models were correct about complete tumor responses and the radiologists were incorrect, we speculate that use of one or more of these models alongside a radiologist might improve the radiologist's ability to identify patients who responds fully to chemotherapy. Venn diagrams depicting the agreements of the assessments performed by the computer methods and the radiologists are shown in Figure 9.8. In cases like that in Figure 9.6(d), radiologists will generally decide that the case is a non-responder because they see residual bladder wall thickening, which is an indicator of cancer. If the

computer models suggested that there was a high likelihood of T0 after treatment in this case, it might lead the radiologists to re-evaluate their decision, and, possibly come to a different (and correct) conclusion. In the case shown in Figure 9.1(a), the tumor stage was found to be T2 at the time of subsequent cystectomy. Even though the computer indicated that the case had a high likelihood of having fully responded, the radiologist might have concluded that cancer was still present on the post-treatment images and decided not to be influenced by the computer's opposite (and incorrect) prediction. Such decisions could lead to a higher accuracy in determining whether a patient had responded completely to neoadjuvant chemotherapy. Of course, it is also possible for the computer models to sway a radiologist's decision in the wrong direction.

All currently approved Food and Drug Administration (FDA) computer-aided detection (CAD) devices are labeled for use as "second readers", i.e., the workflow is to use the CAD system as a second opinion rather than a first reader or concurrent reader. For future decision support systems for treatment response assessment or other applications, it is unknown what the best approach is at this point. One possibility is to follow the second opinion approach: the clinician will first make his/her assessment without being influenced by the computer. The computer then shows the clinician its prediction. The clinician may reconsider his/her assessment using the computer's prediction as additional information. Alternatively, a concurrent approach could be used, where prediction by the decision support system may be displayed from the beginning and it is up to the clinician to use the prediction as additional information in his/her decision making process. Generally, when clinicians' assessments differ from each other, they may discuss and try to reach consensus or bring in additional expert opinion. If the clinician disagrees with the computer, he/she can override it because the computer models, in the current state of development, does not consider many other factors, such as clinical, genetic, or demographic factors, which a clinician may figure into their decision. The clinician may also bring in additional expert opinion if he/she deems it necessary.

Further study of the accuracy of the computer models in tandem with radiologist assessment is needed to determine whether or not such decision support systems will improve radiologist performance in treatment response assessments for bladder cancers. Previous observer performance studies showed that when radiologists and the computer systems had comparable performances individually, using the computer system to assist the radiologists

improved their performance for breast cancer detection[107], breast mass classification[108, 109], and colon polyp detection[110]. We plan to perform an observer performance study using either the second opinion approach or the concurrent approach to evaluate if a similar trend will be observed for bladder cancer treatment response assessment.

There are several limitations to this study. First, the results of this study need to be improved. An AUC values in the range of 0.69 to 0.77 is not optimal for distinguishing between fully and partially or non-responding bladder cancers. Nevertheless, our feasibility study is the first to demonstrate the promise of using radiomics methods, as well as the DL-CNN to distinguish patients who respond to bladder cancer treatment from those who do not. Second, our pilot training data set was relatively small, consisting of only 82 patients with bladder cancer and subsequent generation of 6,700 ROIs. This is a very small number of ROIs for training a DL-CNN. The limited training set size could be a major factor that impacted the performance of a DL-CNN, compared to other pattern recognition tasks such as natural scene classification that can collect millions of training samples much more easily than collecting medical images. Table 9.2 shows the number of correctly identified completely responding pairs and non-completely responding pairs at the operating point used to generate the Venn diagrams in Figure 9.8.

Table 9.2: Number of correctly predicted bladder cancer treatment response assessment of the test set at an operating point determined using the training set.

	DL-CNN	**RF-SL**	**RF-ROI**	**Radiologist 1**	**Radiologist 2**
Complete Response	6/12	6/12	8/12	11/12	11/12
Non-complete Response	34/42	33/42	23/42	18/42	16/42

DL-CNN: Deep-learning convolution neural network
RF-SL: Radiomics features extracted from segmented lesions
RF-ROI: Radiomics features extracted from pre- and post-treatment paired ROIs

We can see from the table that the accuracy of correctly identified completely responding lesion pairs for the DL-CNN and the RF-SL are lower than the RF-ROI and the two Radiologists. We also see that for those two methods, the accuracy of correctly identifying the non-completely responding lesions is higher than the accuracy for the completely responding lesion pairs. This is a limitation of our small and unbalanced data set. The operating points that each of the methods performed similarly to the others on the training set were chosen. While the performances on the training set were similar at this point, we see that the performances on the test set vary. A larger data set is needed so that hopefully, the populations in the training and testing data sets are

similar. We want a representative training population, which can be achieved with large number of training cases. If a bigger training data set was used, we hope that the methods would be more generalizable, thus this issue would not occur. If both the training and testing are unbalanced similarly, we will likely not see as much of a difference between the two classes. This pilot study indicates the potential of using DL-CNN and radiomics methods for treatment response prediction. We still need to study in greater detail the generalizability of the methods using larger data sets in future studies. A third limitation is that this study is a retrospective study. In the future, after the development is completed and the performance of the system is further improved, a prospective clinical trial should be conducted to assess its robustness in the population.

9.6 Conclusion

This study indicates that the computerized assessment using radiomics information from the pre- and post-treatment CT of patients who have undergone neoadjuvant chemotherapy for bladder cancer has the potential to assist in assessment of treatment response. This topic is further studied in Chapter X.

Chapter X

Observer Performance Study for Bladder Cancer Treatment Response Assessment in CT Urography with and without Computerized Decision Support

10.1 Abstract

We evaluated whether a computerized decision support system for bladder cancer treatment response assessment (CDSS-T) can assist radiologists in identifying patients who have complete response after neoadjuvant chemotherapy. With IRB approval, pre- and post-chemotherapy CTU scans of 123 patients with 158 bladder cancers were collected retrospectively, resulting in 158 pre- and post-treatment lesion pairs. The pathological cancer stage after treatment, as determined by cystectomy, was collected as the reference standard in order to determine whether a patient fully responded to treatment. Twenty-five percent of the lesion pairs (40/158) had a complete response, or T0 disease, after chemotherapy. We have developed a CDSS-T system that uses a combination of DL-CNN and radiomics features to distinguish between cases that have fully responded to treatment (T0 disease) and those that have not (>T0 disease). Six abdominal radiologists, four residents trained in abdominal radiology, and one urologist estimated the likelihood of stage T0 disease (complete response) after treatment by viewing each pre-post-treatment CTU pair displayed side by side on a graphic user interface designed for CDSS-T. Each observer provided an estimate without CDSS-T first and then revised the estimate, if desired, after taking into consideration the CDSS-T score. The cases were randomized differently for each observer. The observers' estimates with and without CDSS-T were analyzed with multi-reader, multi-case (MRMC) receiver operating characteristic (ROC) methodology. The area under the curve (AUC) and the statistical significance of the difference were calculated. The AUC for prediction of T0 disease after treatment was 0.80±0.04 for the CDSS-T alone. Ten out of the 11 observers' performance improved with the aid of CDSS-T. The average AUC for the observers was 0.74 (range: 0.72-0.77) without CDSS-T, and 0.77 (range: 0.73-0.81) with CDSS-T. The differences in the average AUC values between without CDSS-T

and with CDSS-T were statistically significant ($p < 0.01$). The study demonstrated that our CDSS-T system for bladder cancer treatment response assessment in CTU can improve radiologists' performance in identifying patients who fully responds to treatment. The preliminary results using methods presented in this chapter have been accepted for presentation at the RSNA 2017 conference. Preparation for submission as a journal article is underway at the time of this dissertation.

10.2 Introduction

Early assessment of therapeutic efficacy and prediction of neoadjuvant chemotherapy failure would help clinicians decide whether to discontinue treatment at an early phase before additional toxicity develops, thus improve the quality of life of a patient and reduce unnecessary morbidity and cost. The ultimate goal is to improve survival for those with a high risk of recurrence while minimizing toxicity to those who will have minimal benefit.

In Chapter IX, we have introduced methods that can estimate the likelihood that a patient responds completely to treatment after neoadjuvant chemotherapy. In this study, we explored the potential that a computerized decision support system for bladder cancer treatment response assessment (CDSS-T) can assist radiologists in better identifying patients who have complete response after neoadjuvant chemotherapy. We compared the performance of three different methods: 1) the CDSS-T alone, 2) the radiologist alone, and 3) radiologists with the CDSS-T.

10.3 Materials and Methods

10.3.1 Data set

With the approval of the Institutional Review Board (IRB), a data set of 123 patients with 158 bladder cancers who were evaluated with CTU before and after the administration of neoadjuvant chemotherapy was collected retrospectively, resulting in 158 lesion pairs. Chemotherapy regimens used for the majority of these patients were MVAC (methotrexate, vinblastine, doxorubicin, and cisplatin treatment), while other patients were treated with regimens including carboplatin, paclitaxel, gemcitabine, and etoposide. The pre-treatment scans were acquired approximately 1 month (max 3 months) before the first cycle of chemotherapy. The post-treatment images were acquired after the completion of three cycles of chemotherapy, generally within 1 month (max 2 months) of cessation of the therapy. The pre-post-treatment

scans were on average acquired 4 months apart. Each patient underwent cystectomy at the end of his or her chemotherapy, usually 4-6 weeks after completion of neoadjuvant chemotherapy, which was generally within 4 months after the post-treatment CT scan. Pathology obtained from the bladder at the time of surgery was used to determine the final cancer stage after chemotherapy and was used as the reference standard to determine whether or not the patient had responded completely to neoadjuvant chemotherapy.

10.3.2 Computerized decision support system for treatment response assessment (CDSS-T)

We have developed a CDSS-T system that uses a combination of DL-CNN and radiomics features to distinguish between cases that have fully responded to treatment and those that have not. The detailed description of the CDSS-T system can be found in Chapter IX. A brief summary is given below.

10.3.2.1 Bladder Lesion Segmentation

The bladder lesions were segmented using our previously developed method known as auto-initialized cascaded level set (AI-CALS)[47]. The details of the AI-CALS method can be found in our previous paper[47].

10.3.2.2 DL-CNN predictive model

Regions of interest (ROIs) were extracted from within the segmented lesions from corresponding pre- and post-treatment scans of a patient and were paired together in multiple combinations to generate pre-post-treatment paired ROIs, as shown in Chapter IX, section 9.3.3. We trained a DL-CNN to distinguish between bladder lesions that were diagnosed as stage T0 post-treatment (no residual tumor) and those that were greater than stage T0 (any residual tumor). For training and testing of the predictive model, a leave-one-case-out cross-validation scheme was used. For each leave-one-case-out partition, the trained DL-CNN outputted a likelihood of stage T0 score for the left-out test case ROIs. The "per-lesion" score was obtained by using the average value among the ROIs associated with the lesion.

10.3.2.3 Radiomics predictive model

For this predictive model, a radiomics-feature-based analysis was applied to the segmented lesions. We extracted 91 features for every segmented lesion, subdivided into four different types: (1) morphological features, such as volume, circularity, rectangularity, and Fourier descriptor, (2) gray level features, such as the average gray level and contrast features, (3) texture features, such as run length statistics, and (4) gradient field features, such as the gradient magnitudes statistics for all voxels on the surface of the segmented lesion. For every temporal lesion pair, the percent differences of each radiomics features between the pre- and post-treatment lesions were calculated. A two-loop leave-one-case-out cross-validation scheme[111] was used for this predictive model. The two-loop leave-one-case-out involves two loops: The inner loop that is used to select the subset of features using a leave-one-case-out scheme on the training partition, and the outer loop that generates the scores for the left-out test case. An average of four features was selected, including two run-length statistics features and two contrast features.

10.3.2.4 Score generation

A combined score using the test scores from both the DL-CNN and the radiomics predictive model were generated by taking the maximum of the two scores. A relative computer-aided diagnosis (CAD) score was obtained by linearly scaling the combined score within the interval between 1 and 10, rounding to the nearest integer. A score of 1 corresponded to the lowest likelihood that the lesion pair completely responded (greater than stage T0). A score of 10 corresponded to the highest likelihood that the lesion pair completely responded (stage T0). This transformation is more intuitive than presenting the raw scores to the observers. The ROC analysis was performed on the combined scores. Curves were fitted to the linearly-transformed distribution of the non-responders and the complete responders, and the area under both of the distribution curves were normalized to a value of one, to obtain fitted distribution for both categories. The distribution of the fitted scores was displayed to the observers as reference, which is shown in Figure 10.1.

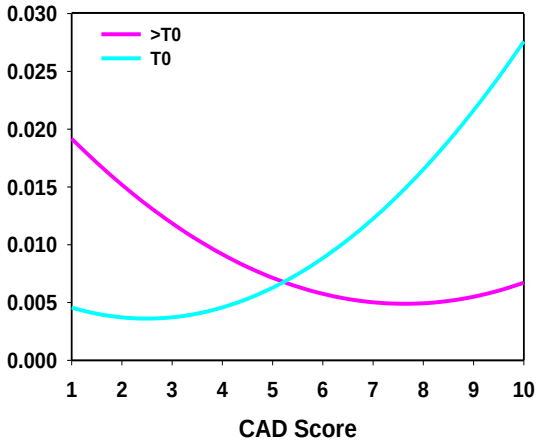


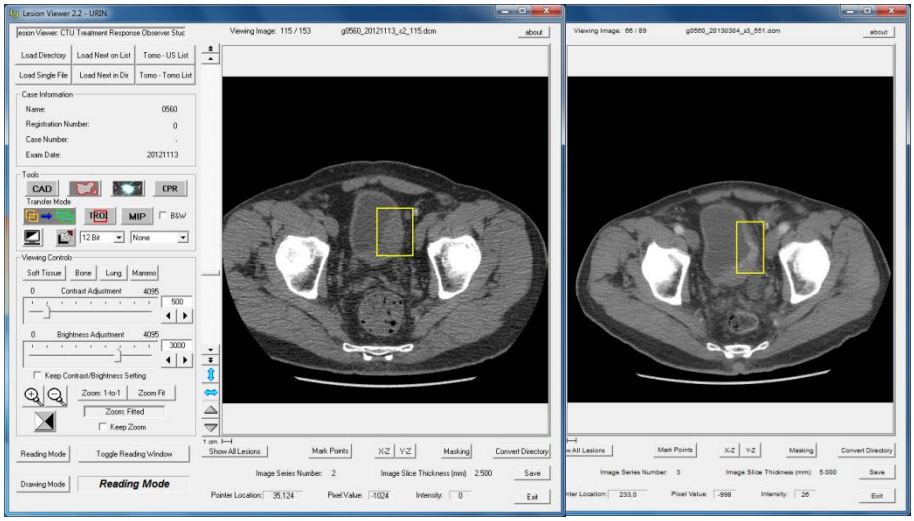
Figure 10.1: The fitted normalized distribution of the scores generated by the combined DL-CNN and radiomics predictive models.

10.3.3 Observer performance study

Six abdominal radiologists, four residents trained in abdominal radiology, and one urologist estimated the likelihood of stage T0 disease (complete response) after treatment by viewing each pre-post-treatment CTU pair displayed side by side on a specialized graphic user interface designed for CDSS-T. The observers were instructed to look within the VOI used for the AI-CALS segmentation, as some cases had multiple lesions, and the CAD score was on a per-lesion basis.

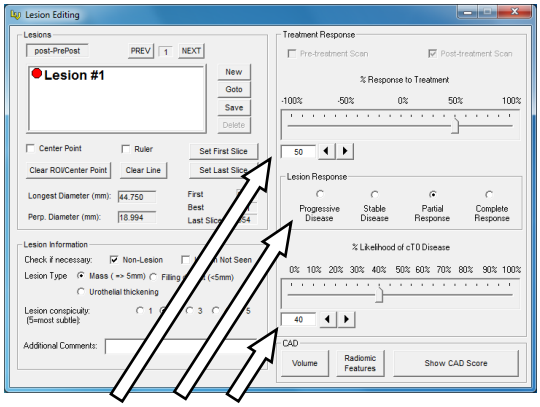
The observer provided an estimate of likelihood of stage T0 disease after treatment, first without CAD, on a scale of 0% to 100%, where 0% indicates that it is very unlikely that the cancer has responded completely to treatment (not T0), while a 100% indicates that it is very likely that there is no residual disease (T0) after treatment. Once the observer had given their likelihood, the CAD score, along with the fitted CAD score distribution was displayed. The observer then may revise their estimate of the likelihood of stage T0, if he or she so desired. The observer was not allowed to change his or her likelihood estimate for without CAD assessment once the CAD score had been displayed.

All observers read the CTU pairs independently, and were allowed unlimited time for the evaluation. The cases were randomized differently for each observer. Figure 10.2 shows the graphical user interface for the reading without, then with the CDSS-T system.



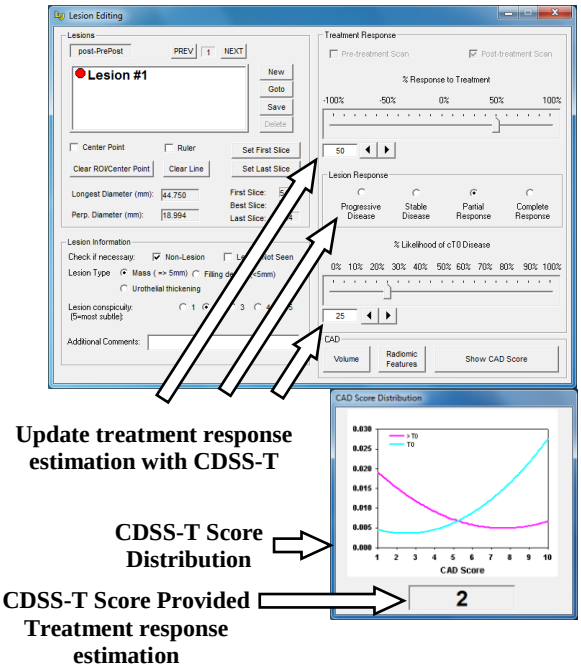
Pre-treatment **Post-treatment**

(a)



Treatment response estimation

(b)



(c)

Figure 10.2: The graphical user interface for reading with and without our computer-aided diagnosis (CAD) system designed for supporting treatment response assessment (CDSS-T). (a) The pre- and post-treatment scans are shown side-by-side, and (b) the observer estimates the treatment response. (c) The observer is shown the CAD score. The score distribution of the two classes is displayed for reference. The observer may revise their treatment response assessment after considering the CAD score.

10.3.4 Evaluation

The observers' estimates with and without CAD were analyzed with multi-reader, multi-case (MRMC) receiver operating characteristic (ROC) methodology[112]. The area under the curve (AUC) and the statistical significance of the difference read with and without CAD were calculated.

The average standard deviation of the likelihood estimates given by the observers per treatment pair was analyzed to study the effects of CAD on the inter-observer variability. The effects of CAD based on the difficulty of the treatment pairs were also studied. The difficulty of a case was estimated by the standard deviation of the observers' likelihood estimates, assuming that the inter-observer variability would be smaller for easier cases. Using a threshold of the standard deviation value of 25, the treatment pairs were categorized into easy (value $\leq$ 25) or difficult treatment pairs (value > 25). The threshold was determined by taking into consideration the data set balance, including the number of T0 treatment pairs. This resulted in 93 treatment pairs categorized as easy with 17.2% (16/93) of them being complete responders, and 65 treatment pairs categorized as difficult, with 36.9% (24/65) of them being complete responders.

10.4 Results

The AUC for prediction of T0 disease after treatment was 0.80 ± 0.04 for the CDSS-T alone. The AUC values of the observers are shown in Table 10.1. A graph of the results for individual observers is shown in Figure 10.3. The average ROC curves for the observers with and without the aid of CDSS-T are shown in Figure 10.4.

Table 10.1: AUC of observers with and without CAD.

Observer Number	Without CAD	With CAD
1	0.76 ± 0.04	0.79 ± 0.04
2	0.74 ± 0.04	0.76 ± 0.04
3	0.74 ± 0.04	0.77 ± 0.04
4	0.74 ± 0.04	0.76 ± 0.04
5	0.72 ± 0.04	0.76 ± 0.04
6	0.75 ± 0.04	0.80 ± 0.04
7	0.73 ± 0.04	0.73 ± 0.04
8	0.75 ± 0.04	0.77 ± 0.04
9	0.73 ± 0.04	0.75 ± 0.04
10	0.77 ± 0.04	0.81 ± 0.04
11	0.73 ± 0.04	0.76 ± 0.04

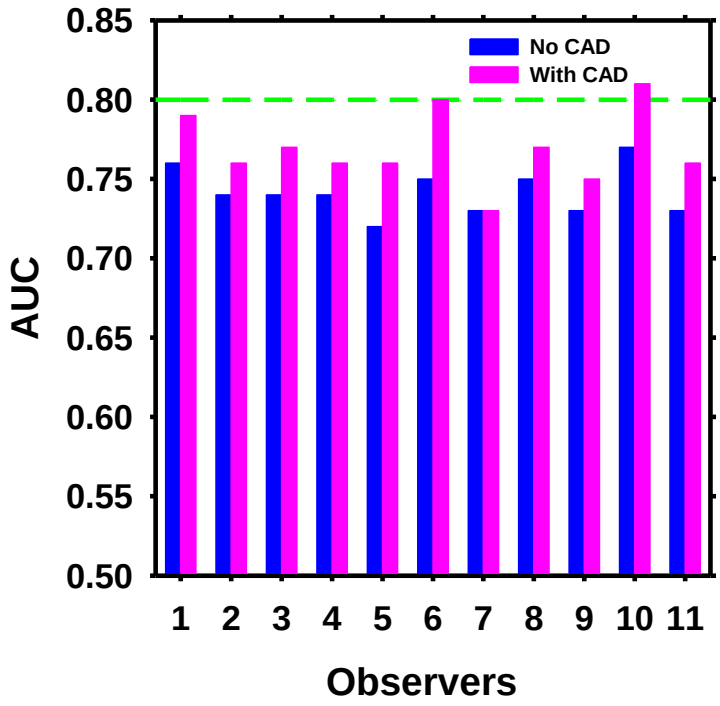


Figure 10.3: AUC values for the 11 observers with and without CDSS-T. The performance of the CDSS-T is shown with the dashed line. The performance all but one of the observers increased with using CDSS-T.

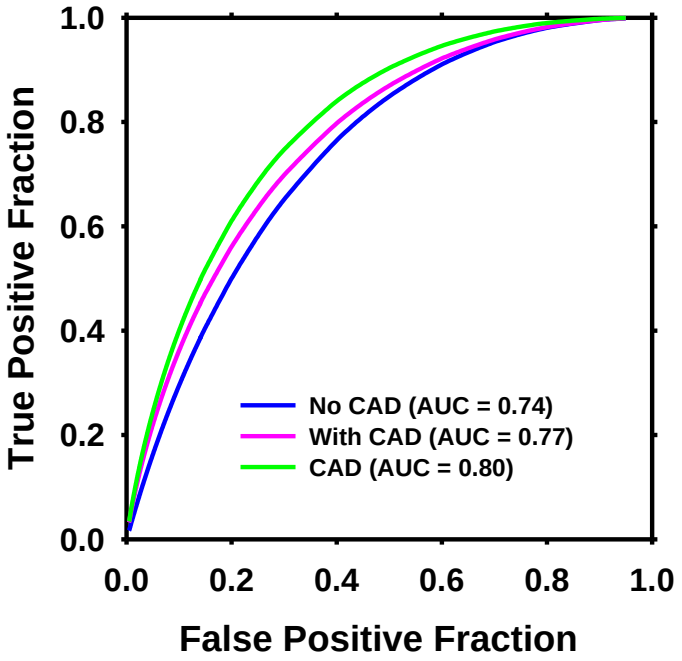


Figure 10.4: Average ROC curves for prediction of T0 stage after neoadjuvant chemotherapy for the 11 observers viewing the pre- and post-treatment CTU pairs without and then with CAD. The average AUC without CAD was 0.74, while it was 0.77 with CAD.

In general, the observers' performances increased with the aid of CDSS-T. The average AUC for the observers were 0.74 (range: 0.72-0.77) without CDSS-T, and increased to 0.77 (range: 0.73-0.81) with CDSS-T. The differences in the average AUC values between without CDSS-T and with CDSS-T were statistically significant ($p = 0.01$). The average standard deviations of the likelihood estimates given by the observers per treatment pair were 20.0 without CAD and 17.4 with CAD with the difference being statistically significant ($p < 0.001$).

The average AUC of the observers increased from 0.83 without CAD to 0.85 with CAD for the easy treatment pairs. The average standard deviation of the observers' likelihood estimates decreased from 13.5 without CAD to 12.0 with CAD. The decrease was statistically significant ($p = 0.005$). For the difficult treatment pairs, the average AUC of the observers was 0.56 without CAD and 0.60 with CAD, and the average standard deviation of the observers' likelihood estimates decreased from 29.2 without CAD to 25.1 with CAD. The decrease was also statistically significant ($p < 0.001$) (Table 10.2).

Table 10.2: Average AUC and average standard deviation of the observers' likelihood estimates with and without CAD for the entire set of cases, and the subsets of easy, and difficult treatment pairs.

	Without CAD		**With CAD**			
	Average AUC	Average standard deviation of observers' likelihood estimates	Average AUC	Average standard deviation of observers' likelihood estimates	Difference in standard deviation (Without CAD – With CAD)	p-value
Entire Set	0.74	20.0	0.77	17.4	2.6	< 0.001*
Easy Subset	0.83	13.5	0.85	12.0	1.5	0.005*
Difficult Subset	0.56	29.2	0.60	25.1	4.1	< 0.001*

* $p < 0.05$ is considered statistically significant

10.5 Discussion

In this study, we studied the effect of a decision support system for treatment response assessment on clinician performance. As far as we are aware, this is the first study of this type. We observed an increase in performance for the estimation of T0 disease by the observers when they read with CDSS-T. On average, we saw an increase of 0.03 in AUC between all of the observers. While the difference in the AUC with and without CAD for the individual observers did not reach statistical significance for 6 out of the 11 observers, performing the MRMC analysis with all of the observers showed a statically significant difference between reading with and without CAD. The AUC of only one observer reading with CDSS-T exceeded the AUC of the CDSS-T by itself.

The abdominal radiologists performed slightly better on average without CAD (AUC = 0.75) compared to the residents (AUC = 0.74), while they had the same performance with CAD (AUC = 0.77). The urologist performed similarly to the average performance of all observers.

The average standard deviation of the likelihood estimates given by the observers per treatment pairs decreased with CAD, which shows that for the task of identifying complete responders vs non-complete responders, CAD reduces the inter-observer variability. Upon further review of the observers' likelihood estimates, it was noted that the observers were resistant to call a case complete response.

The performance with CAD increased with both easy and difficult treatment pairs. We observed a larger increase in the performance with CAD on the treatment pairs categorized as difficult, which indicates that the observers tend to consider more the second opinion by the

computer assessment for the difficult treatment pairs. The larger decrease in the standard deviation for the difficult treatment pairs with CAD indicates that the inter-observer variability is reduced more for the pairs when the observers have less confidence in their own readings without CAD.

There are several limitations to this study. First, due to the lack of a large data set, the CAD scores were obtained through leave-one-case-out cross-validation. Ideally, we should evaluate the system on an independent test set[113]. On the other hand, the leave-one-case-out cross-validation is well established in the pattern recognition literature as a statistically valid technique for estimation of the classifier performance in an unknown population.

Second, we were unable to train the observers to familiarize them with the performance of the CAD system and build up their confidence in using CAD, again due to the lack of a large data set. Although the score distributions of complete responders to non-complete responders were shown to the observers, they were not familiar with the characteristics of the CTU scans that may influence the CDSS-T to give a higher or lower score. In the future, when we collect a larger data set, we will evaluate our system on an independent test set after giving the observers a training session to get acquainted with the performance of the CDSS-T.

10.6 Conclusions

Our study demonstrated the potential that our CDSS-T system for bladder cancer treatment response assessment in CTU can improve clinicians' performance in identifying patients who fully responds to treatment.

However, further improvement in the performance of the CDSS-T will be needed and a large scale observer study with independent test cases should be conducted to validate the impact of the CDSS-T system on clinicians for treatment response assessment of bladder cancer patients.

Chapter XI
Simulation of Incomplete Data for Bladder Cancer Treatment Response Assessment

11.1 Abstract

Predictive models using multidimensional input from different sources are useful but may have higher risk of having incomplete input data for a given case. It is important for a decision support system to be able to handle cases with incomplete data, as this situation may not be uncommon clinically. We studied the effect of incomplete data on the classification of pre- and post-neoadjuvant chemotherapy treatment into complete responders and non-complete responders by a simulation experiment. Using a subset of data set studied in Chapter VIII that had additional clinical information, a linear discriminant (LDA) classifier was built with the new data set. The incomplete data was simulated by randomly replacing a specific feature value with the average feature value estimated from the training set for a fraction of the test cases, and the area under the receiver operating characteristic curve (AUC) was calculated for the entire test set. The experiment was performed with different features used in the LDA classifier, missing one at a time, and repeated multiple times to estimate the average AUC for a given condition. We observed that the impact of missing different features affected the models to a different degree, likely depended on the importance of the specific feature. The test AUC decreased on average compared to without missing data, but the average performance was higher than that obtained with a classifier trained without the missing feature when the proportion of test cases with missing data was small.

11.2 Introduction

Predictive models using multidimensional input from different sources are useful but may have higher risk of having incomplete input data for a given case. It is important for a decision support system to be able to handle cases with incomplete data, as this situation may not be uncommon clinically. For bladder cancer, other clinical information in addition to CT radiomic features may be useful for decision support. For example, urologists perform examination under anesthesia (EUA) by pushing on the bladder with both hands to see whether the bladder is

mobile, or whether they can feel a palpable mass or a fixed bladder. Other clinical information, such as the urine cytology results may also be useful information, although urine cytology results may not be readily available. However, not all patients undergo the EUA or the clinical examinations such that these patients may have incomplete data for the predictive model. We conducted a study to investigate the effect of incomplete data on the classification of pre- and post-neoadjuvant chemotherapy treatment into complete responders and non-complete responders. We simulated missing clinical data by removing feature data from random subset of cases. We used the EUA as a clinical feature that was removed to simulate incomplete data for the study. We also simulated additional incomplete clinical data other than EUA by removing radiomic features as substitutes of the unavailable clinical information.

11.3 Materials and Methods

11.3.1 Data set

The data set used in this study is a subset of the data set described in the study in Chapter VIII. In addition to the imaging data, we obtained the results of the patients who had undergone EUA and were categorized as positive EUA (mass/cancer likely present) or negative EUA (abnormality not observed). The training set consisted of 68 patients with 84 bladder cancers, resulting in 84 pre- and post-treatment lesion pairs, 30% (25/84) of which had completely responded (stage T0) to neoadjuvant chemotherapy. The test data set consisted of 30 cases with 38 temporal lesion pairs, with 26% (10/38) of cancers had complete response to treatment (stage T0). Pathology obtained from the bladder at the time of surgery was used to determine the final cancer stage after chemotherapy.

11.3.2 Radiomic features – segmented lesions (RF-SL) predictive model

Using the current training set, two different Radiomic Features – Segmented Lesions (RF-SL) predictive models described in Chapter IX, section 9.3.4 were built: one incorporating the EUA information as a feature, and the other without. Briefly, we extracted 91 radiomics features from every segmented bladder lesion. For every temporal lesion pair, the percent change between each radiomics feature extracted from the pre- and post-treatment lesion was calculated. Feature selection was performed, and a linear discriminant analysis (LDA) classifier was built using the training set, and applied to the test set.

11.3.3 Simulating incomplete data

To simulate the situation in which a radiomics feature used in the RF-SL model is missing, the value of the selected feature of the RF-SL model on a test case was randomly replaced with the average of the feature values estimated from the training set. A series of experiments was performed where the fraction of the cases with missing data was set to a range between 10% and 100% of the test set for each feature in the predictive model. For each feature and each fraction, the experiment was repeated 100 times for 3 different random seeds.

Two series of experiments were performed: one when the clinical feature (EUA) was not selected for the RF-SL model, and the other when the EUA was selected for the RF-SL model. LDA classifiers were also built on the training set without the chosen studied feature to obtain the performance of the classifier without the feature as a reference for comparison.

11.3.4 Performance evaluation

For every experiment, receiver operating characteristic analysis was performed, and the average area under the curve (AUC) of the test set was calculated for the different levels of missing data.

11.4 Results

From the training set with all the available features, the features selected for the RF-SL model include two contrast features (CONT33, CONT35), and three run-length-statistics texture features (FDH_4, FDH_5, FDH_12), as well as the EUA feature. Without simulating the incomplete data, the test AUC values were 0.76 for RF-SL without EUA, and 0.78 for RF-SL with EUA. Tables 11.1 and 11.2 shows, without EUA and with EUA, respectively, the change in the test AUC as a larger fraction of the test set was missing a feature that was then replaced with the average feature value of the training set. Figures 11.1 and 11.2 show the corresponding graphs.

Table 11.1: Test AUC values of the RF-SL model without EUA with simulated incomplete data on the test set. The last row shows the test AUC of the model trained without the given feature.

Missing Data %	Feature				
	CONT33	FDH_4	FDH_5	CONT35	FDH_12
0%	0.76	0.76	0.76	0.76	0.76
10%	0.73	0.70	0.74	0.73	0.74
20%	0.72	0.67	0.73	0.72	0.73
30%	0.72	0.64	0.72	0.71	0.73
40%	0.71	0.62	0.72	0.70	0.72
50%	0.70	0.59	0.71	0.69	0.72
60%	0.69	0.57	0.70	0.69	0.71
70%	0.69	0.54	0.70	0.68	0.71
80%	0.68	0.52	0.70	0.67	0.70
90%	0.67	0.50	0.70	0.66	0.70
100%	0.67	0.49	0.71	0.66	0.70
RF-SL w/o Feature	0.66	0.64	0.69	0.68	0.71

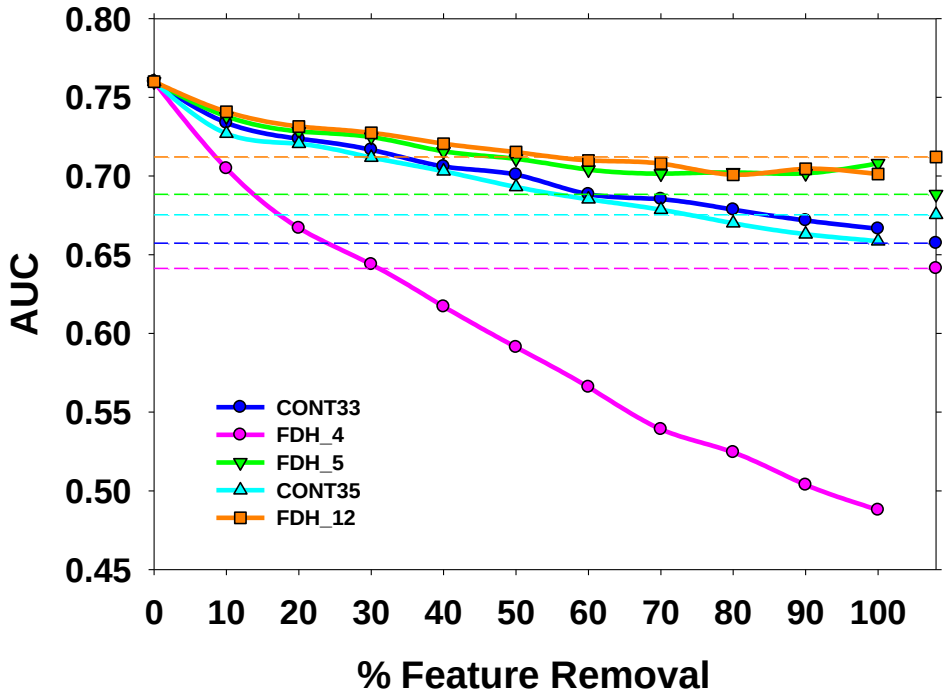


Figure 11.1: Each curve shows the change in AUC of the predictive model as a single feature for different fractions of samples from the test set is missing. The missing feature data was replaced with the average of the feature values in the training set for the RF-SL model without EUA. The dashed lines show the test AUC if the model was trained without the given feature.

Table 11.2: Test AUC values of the RF-SL model with EUA with simulated incomplete data on the test set. The last row shows the test AUC of the model trained without the given feature.

Missing Data %	Feature					
	CONT33	FDH_4	FDH_5	CONT35	FDH_12	EUA
0%	0.78	0.78	0.78	0.78	0.78	0.78
10%	0.77	0.75	0.77	0.77	0.77	0.77
20%	0.77	0.72	0.77	0.76	0.77	0.77
30%	0.76	0.70	0.76	0.75	0.77	0.76
40%	0.76	0.68	0.76	0.75	0.76	0.76
50%	0.76	0.66	0.76	0.74	0.76	0.75
60%	0.75	0.64	0.76	0.73	0.76	0.75
70%	0.75	0.61	0.75	0.73	0.76	0.74
80%	0.74	0.60	0.75	0.72	0.75	0.74
90%	0.73	0.59	0.74	0.72	0.75	0.73
100%	0.72	0.60	0.73	0.72	0.75	0.72
RF-SL w/o Feature	0.75	0.72	0.75	0.71	0.77	0.76

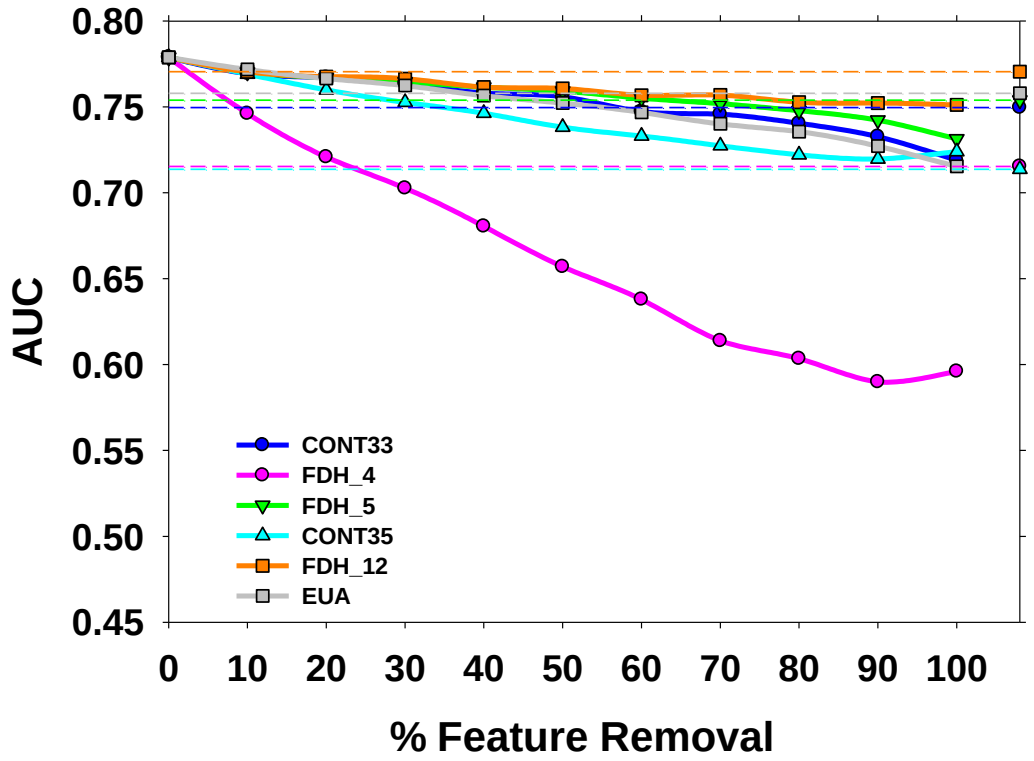


Figure 11.2: Each curve shows the change in AUC of the predictive model as a single feature for different fractions of samples from the test set is missing. The missing feature data was replaced with the average of the feature values in the training set for the RF-SL model with EUA. The dashed lines show the test AUC if the model was trained without the given feature.

11.5 Discussion

Removing any of the features in even 10% of the cases for the two model resulted in worsening of the test performance. As a feature is missing from more cases, the test AUC continues to drop. The FDH_4 feature had the most impact on the classification performance. The general trend between the model with and without the EUA feature was similar.

This simulation study is limited by the number of available cases and the number of features in the predictive model. A more thorough simulation study may be performed in the future by (1) replacing real patient cases with a simulated population with multi-dimensional and multi-domain features so that a large number of samples can be available, (2) studying the effect of feature correlation on the missing feature, (3) study various ways to estimate the missing data, (4) missing multiple features at once, (4) studying the sensitivity to missing data for different classifiers such as neural network, support vector machine, and random forest, and (6) studying the effects of missing data in the training set as well as in test cases.

In the future, additional studies using more clinical features will be conducted when different clinical features capable of distinguishing between completely responding cancers and non-completely responding cancers are utilized in the predictive model. Another planned future effort is to implement a system capable of switching between two different classifiers: one trained with all the features, and one trained without the features that are likely to be missed in the test cases. As the results show, the performance of the classifier built with all the features drops when it is used to evaluate a case with a missing feature. The average performance in terms of AUC therefore decreases as more cases miss the feature. However, it is important to note that the prediction accuracy on individual test cases is independent such that the prediction for an individual case without missing data will not be affected by the prediction for the cases with missing data although the average performance over the entire test set (or "population") drops. Comparing the classifier built with all selected features (the AUC at 0% cases with missing feature) to the classifiers built without an individual feature one at a time (the AUCs plotted on the right vertical axis), it can be seen that the AUC of the classifier built with all selected features is consistently higher. On the other hand, the performances of most of the classifiers built without an individual feature are still higher than applying the classifier built with all features to the test cases missing the feature. The results indicate that a better strategy will be building more than one predictive models, one with the complete set of features and the others with missing a certain feature or features in anticipation of cases with missing data. For a new patient case, such a system will first determine whether all data are available or missing certain data and then triage the case to the appropriate predictive model for assessment. We plan to further study the effectiveness of various clinical data in combination with radiomics features

and compare the different strategies of building and applying the predictive models when a large data set with the variety of clinical tests used in bladder cancer management is available.

11.6 Conclusion

Our study demonstrates that for the task of using the RF-SL model to predict complete response to chemotherapy, the average performance of the predictive model on a test set with missing data replaced with the average value of the training set decreased on average compared to that without missing data, likely due to the prediction on the test cases with missing data being poor. The average performance was higher than retraining the classifier without the missing feature when the proportion of test cases with missing data was small. However, the higher performance was most likely contributed by the cases with complete data such that prediction for these cases was better than the model trained without the feature. As the fraction of test cases with missing feature increased, the AUC would mainly be determined by cases with missing data, and the average performance could drop below the model trained without the feature due to the mismatching between the model and the input data for the majority of the cases. This preliminary study indicates that applying a model built with a certain set of input data to a case missing some of the input data increases the likelihood of obtaining erroneous prediction result. Further study is needed to investigate the best strategy to handle missing clinical test data due to physician's preference or case characteristics, which can often happen in clinical practice.

Chapter XII
Summary and Future Work

12.1 Summary

We have investigated the feasibility of developing a computer-aided image processing and decision support system for bladder cancer. For this dissertation, we focused on CT urography (CTU) images of the bladder. We automatically segmented the bladder using different methods including level sets in combination with other techniques, and achieved reasonable results when comparing the segmentations to the manual hand outlines of radiologists. We were able to reduce the number of inputs from two manually defined region of interest to one and improve the segmentation performance by applying deep learning convolutional neural network (DL-CNN) to the task of bladder segmentation. By training the network to distinguish between regions inside and outside of the bladder, a bladder likelihood map was generated which was used as an initial contour for the level set segmentation. We saw that the DL-CNN by itself is not guaranteed to perform better than the conventional methods previously developed. We did see, however, that combining DL-CNN with the conventional methods can increase the performance for a given task.

Using the segmentations, we were able to automatically detect bladder cancers on CTU images. We first detected masses that were present within the contrast-filled region of the bladder where the contrast between the masses and the urine within the bladder is generally high. We then extended the methods to work with detecting masses within the region that is not filled with contrast material, which makes the masses much more subtle. We adapted the method again to work with finding areas of bladder wall thickening.

Accurate staging of bladder cancer is crucial in providing proper treatment to the patient. Superficial bladder cancers (stage < T2) can be managed with less aggressive treatment than invasive diseases (stage $\geq$ T2), whereas patients with stage T2 to T4 carcinomas of the bladder are referred for neoadjuvant chemotherapy. We extracted radiomics features from the segmented

lesions, and built classifiers to distinguish between lesions whose stage is $\geq$ T2 or < T2, using different machine learning classifiers.

Early assessment of therapeutic efficacy and prediction of neoadjuvant chemotherapy treatment failure for bladder cancer would help clinicians decide whether to discontinue chemotherapy at an early phase before additional toxicity develops, thus improving the quality of life of a patient and reducing unnecessary morbidity and cost. The ultimate goal is to improve survival for those with a high risk of recurrence while minimizing toxicity to those who will have minimal benefit. Therefore, development of an accurate and early predictive model of the effectiveness of neoadjuvant chemotherapy is important for patients with bladder cancer. We applied DL-CNN for the task of bladder lesion segmentation, and attempted to predict complete response to treatment by measuring the change in gross tumor volume calculated from the lesion segmentations. We developed multiple predictive models to estimate the likelihood that a pre- and post-treatment lesion pair has completely responded to treatment. We built different systems using DL-CNN and radiomic features, and found that they performed comparably to radiologists in distinguishing complete responders to non-complete responders. We built a decision support system that combines the DL-CNN and the radiomics features, tested to see whether such a system can help observers identify lesions that fully respond to treatment and found that our system can improve the clinicians' performances.

Working with multi-modality data could mean that the data for certain patients may be incomplete. We simulated what would happen to a predictive model for bladder cancer treatment response assessment if we estimated missing data on the test set using the average data from the training set.

12.2 Future Work

The computer-aided image processing and decision support system for bladder cancer developed and presented in this dissertation demonstrated the feasibility of several systems that may help clinicians diagnose and manage bladder cancer patients. Further development is needed to bring the systems demonstrated in this dissertation to clinical practice. These include the following:

- Develop a deep learning system for tasks other than classification, such as directly giving segmentation contours, and detection of bladder cancers.

- Optimize the parameters for DL-CNN for a given task to obtain high performance while reducing over-fitting. As there are many parameters involved in training of a DL-CNN, the different combinations of variables such as random seed, initial weights, learning rate, and network structure, need to be optimized for a task. With the accumulating experience of designing DL-CNN, we will be able to decide on many of these parameters faster, and in a more systematic way for future tasks.
- Improve bladder segmentation performance by reducing errors caused by poor image quality due to patient size or presence of hip prosthesis. Segmentation errors due to bladder cancer spreading into neighboring organs would also be need to be addressed.
- Increase performance for the detection of lesions within the bladder (both masses and wall thickening) by training with a much larger data set and testing it on a much larger independent test set to study the robustness of the methods. One possible method for removing additional false positive findings would be to automatically detect the prostate, and remove it before running the algorithm for bladder lesion detection.
- Perform an observer performance study with the bladder cancer detection system to see whether a detection system used as an adjunct can improve clinicians' sensitivity in identifying bladder cancers.
- Find incorrectly staged bladder cancers, and see if our staging system can correctly classify these cases.
- Increase the performance of the bladder cancer treatment response assessment system, as the performance is relatively low compared to other tasks. Although the performance of the current system is already comparable to the radiologists, the performance is low compared with other tasks tested in these studies; increasing the performance of the system would increase its potential benefit to clinicians.
- Develop a system for bladder cancer treatment response assessment that evaluates changes in bladder wall thickening and estimates the likelihood of complete response to treatment, as our currently developed system only applies to bladder masses.
- Apply the bladder cancer treatment response decision support system to a larger independent test set and assess its performance, and perform another observer performance study to check the robustness of the system on a much larger and more diverse data set.

BIBLIOGRAPHY

[1] K.H. Cha, L.M. Hadjiiski, H.-P. Chan, E.M. Caoili, R.H. Cohan, C. Zhou, "CT urography: segmentation of urinary bladder using CLASS with local contour refinement," Phys Med Biol **59**, 2767-2785 (2014).

[2] K.H. Cha, L. Hadjiiski, R.K. Samala, H.P. Chan, E.M. Caoili, R.H. Cohan, "Urinary bladder segmentation in CT urography using deep-learning convolutional neural network and level sets," Medical Physics **43**, 1882-1896 (2016).

[3] K.H. Cha, L. Hadjiiski, H.-P. Chan, R.H. Cohan, E.M. Caoili, C. Zhou, "Detection of urinary bladder mass in CT urography with SPAN," Medical Physics **42**, 4271-4284 (2015).

[4] K.H. Cha, L.M. Hadjiiski, R.K. Samala, H.P. Chan, R.H. Cohan, E.M. Caoili, C. Paramagul, A. Alva, A.Z. Weizer, "Bladder Cancer Segmentation in CT for Treatment Response Assessment: Application of Deep-Learning Convolution Neural Network-A Pilot Study," Tomography **2**, 421-429 (2016).

[5] K.H. Cha, L. Hadjiiski, H.-P. Chan, A.Z. Weizer, A. Alva, R.H. Cohan, E.M. Caoili, C. Paramagul, R.K. Samala, "Bladder Cancer Treatment Response Assessment in CT using Radiomics with Deep-Learning," Scientific Reports **7**, s41598-41017 (2017).

[6] K.H. Cha, L.M. Hadjiiski, H.-P. Chan, E.M. Caoili, R.H. Cohan, A. Weizer, R.K. Samala, "Computer-aided detection of bladder masses in CT Urography (CTU)," Proc SPIE, 1013421-1013426 (2017).

[7] S.S. Garapati, L.M. Hadjiiski, K.H. Cha, H.-P. Chan, E.M. Caoili, R.H. Cohan, A. Weizer, A. Alva, C. Paramagul, J. Wei, C. Zhou, "Automatic staging of bladder cancer on CT urography," Proc SPIE **9785**, 97851G-97851G-97856 (2016).

[8] S.A. Woolen, L. Hadjiiski, K.H. Cha, H.-P. Chan, F.P. Worden, P. Swiecicki, B. Wasserman, A. Srinivasan, "Detecting Tumor Features of Head and Neck Cancers on CT Using Computerized Analysis," Oral presentation at the 102nd Scientific Assembly and Annual Meeting of the Radiological Society of North America (RSNA), Chicago, IL. Nov 27-Dec 2(2016).

[9] L. Hadjiiski, H.-P. Chan, K.H. Cha, A. Srinivasan, J. Wei, C. Zhou, M. Prince, S. Papagerakis, "Radiomics biomarkers for accurate tumor progression prediction of oropharyngeal cancer," Proc SPIE **10134**, 101341Z-101341Z-101347 (2017).

[10] K.H. Cha, L.M. Hadjiiski, H.-P. Chan, R.K. Samala, R.H. Cohan, E.M. Caoili, C. Paramagul, A. Alva, A.Z. Weizer, "Bladder cancer treatment response assessment using deep learning in CT with transfer learning," Proc SPIE **10134**, 1013431-1013436 (2017).

[11] L. Hadjiiski, H.P. Chan, R.H. Cohan, E.M. Caoili, Y. Law, K. Cha, C. Zhou, J. Wei, "Urinary bladder segmentation in CT urography (CTU) using CLASS," Medical Physics **40**, 111906 (2013).

[12] *American Cancer Society. Cancer Facts & Figures 2017.* . (American Cancer Society, Inc., Atlanta, 2017).

[13] S.A. Akbar, K.J. Mortele, K. Baeyens, M. Kekelidze, S.G. Silverman, "Multidetector CT urography: Techniques, clinical applications, and pitfalls," Seminars in Ultrasound CT and MRI **25**, 41-54 (2004).

[14] E.M. Caoili, R.H. Cohan, M. Korobkin, J.F. Platt, I.R. Francis, G.J. Faerber, J.E. Montie, J.H. Ellis, "Urinary tract abnormalities: Initial experience with multi-detector row CT urography " Radiology **222**, 353-360 (2002).

[15] W.C. Liu, K.J. Mortele, S.G. Silverman, "Incidental extraurinary findings at MDCT urography in patients with hematuria: Prevalence and impact on Imaging costs," American Journal of Roentgenology **185**, 1051-1056 (2005).

[16] C.L. McCarthy, N.C. Cowan, "Multidetector CT urography (MD-CTU) for urothelial imaging," Radiology (P) **225**, 237 (2002).

[17] M. Noroozian, R.H. Cohan, E.M. Caoili, N.C. Cowan, J.H. Ellis, "Multislice CT urography: State of the art " British Journal of Radiology **77**, S74-S86 (2004).

[18] S.B. Park, J.K. Kim, H.J. Lee, H.J. Choi, K.-S. Cho, "Hematuria: portal venous phase multi detector row CT of the bladder--a prospective study," Radiology **245**, 798-805 (2007).

[19] G.S. Sudakoff, D.P. Dunn, M.L. Guralnick, R.S. Hellman, D. Eastwood, W.A. See, "Multidetector computerized tomography urography as the primary imaging modality for detecting urinary tract neoplasms in patients with asymptomatic hematuria," Journal of Urology **179**, 862-867 (2008).

[20] L. Li, Z. Wang, X. Li, X. Wei, A.H. L., W. Huang, S. Rizvi, M. H., D.P. Harrington, Z. Liang, "A new partial volume segmentation approach to extract bladder wall for computer aided detection in virtual cystoscopy," Proc. SPIE **5369**, 199-206 (2004).

[21] C. Duan, Z. Liang, S. Bao, H. Zhu, S. Wang, G. Zhang, J.J. Chen, H. Lu, "A coupled level set framework for bladder wall segmentation with application to MR cystography," IEEE Trans Med Imaging **29**, 903-915 (2010).

[22] C.J. Duan, K.H. Yuan, F.H. Liu, P. Xiao, G.Q. Lv, Z.R. Liang, "An Adaptive Window-Setting Scheme for Segmentation of Bladder Tumor Surface via MR Cystography," IEEE Transactions on Information Technology in Biomedicine **16**, 720-729 (2012).

[23] H. Han, L. Li, C. Duan, H. Zhang, Y. Zhao, Z. Liang, "A unified EM approach to bladder wall segmentation with coupled level-set constraints," Medical Image Analysis **17**, 1192-1205 (2013).

[24] X.F. Chai, M. van Herk, A. Betgen, M. Hulshof, A. Bel, "Automatic bladder segmentation on CBCT for multiple plan ART of bladder cancer using a patient-specific bladder model," Physics in Medicine and Biology **57**, 3945-3962 (2012).

[25] L.M. Hadjiiski, B. Sahiner, H.-P. Chan, E.M. Caoili, R.H. Cohan, C. Zhou, "Automated segmentation of urinary bladder and detection of bladder lesions in multi-detector row CT urography," Proc. SPIE **7260**, 72603R72601- 72603R72607 (2009).

[26] L. Hadjiiski, H.-P. Chan, Y. Law, R.H. Cohan, E.M. Caoili, H.C. Cho, C. Zhou, J. Wei, "Segmentation of Urinary Bladder in CT Urography (CTU) using CLASS," Proc. SPIE **8315**, 83150J83151-83150J83157 (2012).

[27] L.M. Hadjiiski, H.-P. Chan, R.H. Cohan, E.M. Caoili, Y. Law, K. Cha, C. Zhou, J. Wei, "Urinary bladder segmentation in CT urography (CTU) using CLASS," Medical Physics **40**, 11190601-11190610 (2013).

[28] E. Street, L. Hadjiiski, B. Sahiner, S. Gujar, M. Ibrahim, S.K. Mukherji, H.-P. Chan, "Automated Volume Analysis of Head and Neck Lesions on CT Scans Using 3D Level Set Segmentation," Medical Physics **34**, 4399-4408 (2007).

[29] P. Jaccard, "The distribution of the flora in the alpine zone," New phytologist **11**, 37-50 (1912).

[30] T.W. Way, L.M. Hadjiiski, B. Sahiner, H.-P. Chan, P.N. Cascade, E.A. Kazerooni, N. Bogot, C. Zhou, "Computer-aided diagnosis of pulmonary nodules on CT scans: segmentation and classification using 3D active contours," Medical Physics **33**, 2323-2337 (2006).

[31] H.-P. Chan, S.C.B. Lo, B. Sahiner, K.L. Lam, M.A. Helvie, "Computer-aided detection of mammographic microcalcifications: Pattern recognition with an artificial neural network," Medical Physics **22**, 1555-1567 (1995).

[32] R.K. Samala, H.-P. Chan, Y. Lu, L.M. Hadjiiski, J. Wei, M.A. Helvie, "Digital breast tomosynthesis: computer-aided detection of clustered microcalcifications on planar projection images," Physics in Medicine and Biology **59**, 7457-7477 (2014).

[33] M.N. Gurcan, B. Sahiner, H.-P. Chan, L.M. Hadjiiski, N. Petrick, "Selection of an optimal neural network architecture for computer-aided diagnosis - comparison of automated optimization techniques," Radiology **217(P)**, 436 (2000).

[34] M.N. Gurcan, B. Sahiner, H.-P. Chan, L.M. Hadjiiski, N. Petrick, "Selection of an optimal neural network architecture for computer-aided detection of microcalcifications - comparison of automated optimization techniques," Medical Physics **28**, 1937-1948 (2001).

[35] J. Ge, B. Sahiner, L.M. Hadjiiski, H.-P. Chan, J. Wei, M.A. Helvie, C. Zhou, "Computer aided detection of clusters of microcalcifications on full field digital mammograms," Medical Physics **33**, 2975-2988 (2006).

[36] J. Ge, L.M. Hadjiiski, B. Sahiner, J. Wei, M.A. Helvie, C. Zhou, H.-P. Chan, "Computer-aided detection system for clustered microcalcifications: comparison of performance on full-field digital mammograms and digitized screen-film mammograms," Physics in Medicine and Biology **52**, 981-1000 (2007).

[37] P. Filev, L. Hadjiiski, H.-P. Chan, B. Sahiner, J. Ge, M.A. Helvie, M. Roubidoux, C.A. Zhou, "Automated regional registration and characterization of corresponding microcalcification clusters on temporal pairs of mammograms for interval change analysis," Medical Physics **35**, 5340-5350 (2008).

[38] R.K. Samala, H.-P. Chan, Y. Lu, L.M. Hadjiiski, J. Wei, M.A. Helvie, "Computer-aided Detection System for Clustered Microcalcifications in Digital Breast Tomosynthesis using Joint Information from Volumetric and Planar Projection Images," Physics in Medicine and Biology **60**, 8457-8479 (2015).

[39] A. Krizhevsky, "cuda-convnet", 2012, (09/12/2016). [http://code.google.com/p/cuda-convnet/]

[40] A. Krizhevsky, I. Sutskever, G.E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *NIPS 2012: Neural Information Processing Systems* (Lake Tahoe, Nevada 2012).

[41] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, "Imagenet large scale visual recognition challenge," International Journal of Computer Vision, 1-42 (2014).

[42] A. Krizhevsky, "Learning Multiple Layers of Features from Tiny Images," University of Toronto, 2009.

[43] P. Viola, M. Jones, "Rapid object detection using a boosted cascade of simple features," in *2001 Ieee Computer Society Conference on Computer Vision and Pattern Recognition, Vol 1, Proceedings*, edited by A. Jacobs, T. Baldwin (2001), pp. 511-518.

[44] R. Lienhart, J. Maydt, Ieee, "An extended set of haar-like features for rapid object detection," in *2002 International Conference on Image Processing, Vol I, Proceedings* (2002), pp. 900-903.

[45] V. Nair, G.E. Hinton, "Rectified linear units improve restricted boltzmann machines," Proceedings of the 27th International Conference on Machine Learning (ICML-10), 807-814 (2010).

[46] S. Jaume, M. Ferrant, B. Macq, L. Hoyte, J.R. Fielding, A. Schreyer, R. Kikinis, S.K. Warfield, "Tumor detection in the bladder wall with a measurement of abnormal thickness in CT scans," IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING **50**, 383-390 (2003).

[47] L.M. Hadjiiski, H.-P. Chan, E.M. Caoili, R.H. Cohan, J. Wei, C. Zhou, "Auto-Initialized Cascaded Level Set (AI-CALS) Segmentation of Bladder Lesions on Multi-Detector Row CT Urography," Academic Radiology **20**, 148-155 (2013).

[48] N. Otsu, "A threshold selection method from gray-level histograms," IEEE Trans. System, Man, Cybernetics **9**, 62-66 (1979).

[49] J. Kilday, F. Palmieri, M.D. Fox, "Classifying mammographic lesions using computer-aided image analysis," IEEE Transactions on Medical Imaging **12**, 664-669 (1993).

[50] B. Sahiner, H.-P. Chan, N. Petrick, M.A. Helvie, L.M. Hadjiiski, "Improvement of mammographic mass characterization using spiculation measures and morphological features," Medical Physics **28**, 1455-1465 (2001).

[51] L.M. Hadjiiski, B. Sahiner, H.-P. Chan, N. Petrick, M.A. Helvie, M.N. Gurcan, "Analysis of Temporal Change of Mammographic Features: Computer-Aided Classification of Malignant and Benign Breast Masses," Medical Physics **28**, 2309-2317 (2001).

[52] M. Gordon, L. Hadjiiski, K. Cha, H.-P. Chan, R. Samala, R.H. Cohan, E.M. Caoili, "Segmentation of inner and outer bladder wall using deep-learning convolutional neural networks in CT urography," Proc SPIE **10134**, 1013411-1013417 (2017).

[53] S.S. Chang, S.A. Boorjian, R. Chou, P.E. Clark, S. Daneshmand, B.R. Konety, R. Pruthi, D.Z. Quale, C.R. Ritch, J.D. Seigne, E.C. Skinner, N.D. Smith, J.M. McKiernan, "Diagnosis and Treatment of Non-Muscle Invasive Bladder Cancer: AUA/SUO Guideline," Journal of Urology **196**, 1021-1029 (2016).

[54] J.A. Witjes, E. Comperat, N.C. Cowan, M. De Santis, G. Gakis, N. James, T. Lebret, A. Sherif, A.G. Van der Heijden, M.J. Ribal, "Guidelines on Muscle-invasive and Metastatic Bladder Cancer," European Association of Urology (2016).

[55] M. Babjuk, A. Bohle, M. Burger, E. Comperat, E. Kaasinen, J. Palou, M. Roupret, B.W.G. Van Rhijn, S. Shariat, R. Sylvester, R. Zigeuner, "Guidelines on Non-muscle-invasive Bladder Cancer (Ta, T1 and CIS)," European Association of Urology (2016).

[56] H.W. Herr, S.M. Donat, "Quality control in transurethral resection of bladder tumours," Bju International **102**, 1242-1246 (2008).

[57] *AJCC Cancer Staging Handbook*, 8th ed. (American Joint Committee on Cancer, Chicago, IL, 2016).

[58] J.J. Meeks, J. Bellmunt, B.H. Bochner, N.W. Clarke, S. Daneshmand, M.D. Galsky, N.M. Hahn, S.P. Lerner, M. Mason, T. Powles, C.N. Sternberg, G. Sonpavde, "A Systematic Review of Neoadjuvant and Adjuvant Chemotherapy for Muscle-invasive Bladder Cancer," European Urology **62**, 523-533 (2012).

[59] S.L. Fagg, P. Dawsonedwards, M.A. Hughes, T.N. Latief, E.B. Rolfe, J.W.L. Fielding, "CIS-Diamminedichloroplatinum (DDP) as initial treatment of invasive bladder cancer," British Journal of Urology **56**, 296-300 (1984).

[60] D. Raghavan, B. Pearson, G. Coorey, W. Woods, D. Arnold, J. Smith, J. Donovan, P. Langdon, "Intravenous CIS-platinum for invasive bladder cancer – safety and feasibility of a new approach," Medical Journal of Australia **140**, 276-278 (1984).

[61] J. Huguet, M. Crego, S. Sabate, J. Salvador, J. Palou, H. Villavicencio, "Cystectomy in patients with high risk superficial bladder tumors who fail intravesical BCG therapy: Pre-cystectomy prostate involvement as a prognostic factor," European Urology **48**, 53-59 (2005).

[62] H.M. Fritsche, M. Burger, R.S. Svatek, C. Jeldres, P.I. Karakiewicz, G. Novara, E. Skinner, S. Denzinger, Y. Fradet, H. Isbarn, P.J. Bastian, B.G. Volkmer, F. Montorsi, W. Kassouf, D. Tilki, W. Otto, U. Capitanio, J.I. Izawa, V. Ficarra, S. Lerner, A.I. Sagalowsky, M. Schoenberg, A. Kamat, C.P. Dinney, Y. Lotan, S.F. Shariat, "Characteristics and Outcomes of Patients with Clinical T1 Grade 3 Urothelial Carcinoma Treated with Radical Cystectomy: Results from an International Cohort," European Urology **57**, 300-309 (2010).

[63] P. Turker, P.J. Bostrom, M.L. Wroclawski, B. van Rhijn, H. Kortekangas, C. Kuk, T. Mirtti, N.E. Fleshner, M.A. Jewett, A. Finelli, T.V. Kwast, A. Evans, J. Sweet, M. Laato, A.R. Zlotta, "Upstaging of urothelial cancer at the time of radical cystectomy: factors associated with upstaging and its effect on outcome," Bju International **110**, 804-811 (2012).

[64] S.F. Shariat, G.S. Palapattu, P.I. Karakiewicz, C.G. Rogers, A. Vazina, P.J. Bastian, M.P. Schoenberg, S.P. Lerner, A.I. Sagalowsky, Y. Lotan, "Discrepancy between clinical and pathologic stage: Impact on prognosis after radical cystectomy," European Urology **51**, 137-151 (2007).

[65] *ACR Manual on Contrast Media*. (ACR Committee on Drugs and Contrast Media, 2016).

[66] B. Sahiner, H.-P. Chan, N. Petrick, M.A. Helvie, M.M. Goodsitt, "Computerized characterization of masses on mammograms: The rubber band straightening transform and texture analysis," Medical Physics **25**, 516-526 (1998).

[67] B.R. Dasarathy, E.B. Holder, "Image characterizations based on joint gray-level run-length distributions," Pattern Recog. Letters **12**, 497-502 (1991).

[68] H.-P. Chan, D. Wei, M.A. Helvie, B. Sahiner, D.D. Adler, M.M. Goodsitt, N. Petrick, "Computer-aided classification of mammographic masses and normal tissue: Linear discriminant analysis in texture feature space," Physics in Medicine and Biology **40**, 857-876 (1995).

[69] P.A. Lachenbruch, *Discriminant Analysis*. (Hafner Press, New York, 1975).

[70] M.M. Tatsuoka, *Multivariate Analysis, Techniques for Educational and Psychological Research*, 2nd ed. (Macmillan, New York, 1988).

[71] D.E. Rumelhart, G.E. Hinton, R.J. Williams, *Learning Internal Representation by Error Propagation*. (MIT Press, Cambridge, MA, 1986).

[72] V.N. Vapnik, *Statistical Learning Theory*. (Wiley, New York, 1998).

[73] C.J.C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," Data Mining and Knowledge Discovery **2**, 121-167 (1998).

[74] T.K. Ho, "The random subspace method for constructing decision forests," Ieee Transactions on Pattern Analysis and Machine Intelligence **20**, 832-844 (1998).

[75] I.H. Witten, E. Frank, M.A. Hall, C.J. Pal, *The WEKA Workbench. Online Appendix for "Data Mining: Practical machine learning tools and techniques"*. (Morgan Kaufmann, 2016).

[76] C.E. Metz, "ROC methodology in radiologic imaging," Investigative Radiology **21**, 720-733 (1986).

[77] C.E. Metz, B.A. Herman, J.H. Shen, "Maximum-likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data," Statistics in Medicine **17**, 1033-1053 (1998).

[78] "Metz ROC Software. University of Chicago Medical Center Department of Radiology, see http://metz-roc.uchicago.edu/MetzROC/software",

[79] G. Litjens, T. Kooi, B.E. Bejnordi, A.A.A. Setio, F. Ciompi, M. Ghafoorian, J.A. van der Laak, B. van Ginneken, C.I. Sánchez, "A Survey on Deep Learning in Medical Image Analysis," arXiv:1702.05747 (2017).

[80] H. Greenspan, B. van Ginneken, R.M. Summers, "Deep Learning in Medical Imaging: Overview and Future Promise of an Exciting New Technique," Ieee Transactions on Medical Imaging **35**, 1153-1159 (2016).

[81] C.N. Sternberg, "The treatment of advanced bladder cancer," Annals of Oncology **6**, 113-126 (1995).

[82] H. Abol-Enein, A.V. Bono, M. Boyer, N.W. Clarke, C.M.L. Coppin, E. Cortesi, P.J. Goebell, S. Groshen, R.R. Hall, A. Horwich, P.U. Malmstrom, J.A. Martinez-Pineiro, M.K.B. Parmar, D. Raghavan, J.T.G. Roberts, L. Sengelov, A. Sherif, L.A. Stewart, M. Stockle, R. Sylvester, J.F. Tierney, F.M. Torti, C.L. Vale, D.M.A. Wallace, A.B.C.M.-A. Collaboration, "Neoadjuvant chemotherapy in invasive bladder cancer: a systematic review and meta-analysis," Lancet **361**, 1927-1934 (2003).

[83] H.B. Grossman, R.B. Natale, C.M. Tangen, V.O. Speights, N.J. Vogelzang, D.L. Trump, R.W.D. White, M.F. Sarosdy, D.P. Wood, D. Raghavan, E.D. Crawford, "Neoadjuvant chemotherapy plus cystectomy compared with cystectomy alone for locally advanced bladder cancer," New England Journal of Medicine **349**, 859-866 (2003).

[84] J.A. Witjes, M. Wullink, G.O.N. Oosterhof, P. deMulder, "Toxicity and results of MVAC (methotrexate, vinblastine, adriamycin and cisplatin) chemotherapy in advanced urothelial carcinoma," European Urology **31**, 414-419 (1997).

[85] *WHO handbook for reporting results of cancer treatment. Geneva (Switzerland): World Health Organization Offset Publication No. 48;*. (1979).

[86] E.A. Eisenhauer, P. Therasse, J. Bogaerts, L.H. Schwartz, D. Sargent, R. Ford, J. Dancey, S. Arbuck, S. Gwyther, M. Mooney, L. Rubinstein, L. Shankar, L. Dodd, R. Kaplan, D. Lacombe, J. Verweij, "New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1)," European Journal of Cancer **45**, 228-247 (2009).

[87] J.E. Husband, L.H. Schwartz, J. Spencer, L. Ollivier, D.M. King, R. Johnson, R. Reznek, S. Int Canc Imaging, "Evaluation of the response to treatment of solid tumours - a consensus statement of the International Cancer Imaging Society," British Journal of Cancer **90**, 2256-2260 (2004).

[88] E.M. Bessell, H.M. Price, P.J. McMillan, "The measurement of the regression of carcinoma of the bladder using ultrasonography and CT scanning during and after radical radiotherapy," Radiotherapy and Oncology **19**, 145-157 (1990).

[89] L. Hadjiiski, A.Z. Weizer, A. Alva, E.M. Caoili, R.H. Cohan, K. Cha, H.P. Chan, "Treatment Response Assessment for Bladder Cancer on CT Based on Computerized Volume Analysis, World Health Organization Criteria, and RECIST," American Journal of Roentgenology **205**, 348-352 (2015).

[90] B. Sahiner, H.-P. Chan, N. Petrick, D. Wei, M.A. Helvie, D.D. Adler, M.M. Goodsitt, "Classification of mass and normal breast tissue: A convolution neural network classifier with spatial domain and texture images," IEEE Transactions on Medical Imaging **15**, 598-610 (1996).

[91] M.N. Gurcan, H.-P. Chan, B. Sahiner, L. Hadjiiski, N. Petrick, M.A. Helvie, "Optimal neural network architecture selection: Improvement in computerized detection of microcalcifications," Academic Radiology **9**, 420-429 (2002).

[92] H.-C. Shin, H.R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, R.M. Summers, "Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning," Ieee Transactions on Medical Imaging **35**, 1285-1298 (2016).

[93] C.A. Sadow, S.G. Silverman, M.P. O'Leary, J.E. Signorovitch, "Bladder cancer detection with CT urography in an academic medical center," Radiology **249**, 195-202 (2008).

[94] R.H. Cohan, E.M. Caoili, N.C. Cowan, A.Z. Weizer, J.H. Ellis, "MDCT Urography: Exploring a New Paradigm for Imaging of Bladder Cancer," American Journal of Roentgenology **192**, 1501-1508 (2009).

[95] M. Jinzaki, A. Tanimoto, H. Shinmoto, Y. Horiguchi, K. Sato, S. Kuribayashi, S.G. Silverman, "Detection of bladder tumors with dynamic contrast-enhanced MDCT," American Journal of Roentgenology **188**, 913-918 (2007).

[96] P. Lambin, E. Rios-Velazquez, R. Leijenaar, S. Carvalho, R. van Stiphout, P. Granton, C.M.L. Zegers, R. Gillies, R. Boellard, A. Dekker, H. Aerts, I.C.C.C. Qu, "Radiomics: Extracting more information from medical images using advanced feature analysis," European Journal of Cancer **48**, 441-446 (2012).

[97] V. Kumar, Y.H. Gu, S. Basu, A. Berglund, S.A. Eschrich, M.B. Schabath, K. Forster, H. Aerts, A. Dekker, D. Fenstermacher, D.B. Goldgof, L.O. Hall, P. Lambin, Y. Balagurunathan, R.A. Gatenby, R.J. Gillies, "Radiomics: the process and the challenges," Magnetic Resonance Imaging **30**, 1234-1248 (2012).

[98] R.J. Gillies, P.E. Kinahan, H. Hricak, "Radiomics: Images Are More than Pictures, They Are Data," Radiology **278**, 563-577 (2016).

[99] H. Aerts, E.R. Velazquez, R.T.H. Leijenaar, C. Parmar, P. Grossmann, S. Cavalho, J. Bussink, R. Monshouwer, B. Haibe-Kains, D. Rietveld, F. Hoebers, M.M. Rietbergen, C.R. Leemans, A. Dekker, J. Quackenbush, R.J. Gillies, P. Lambin, "Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach," Nature Communications **5**(2014).

[100] A. Cunliffe, S.G. Armato, R. Castillo, N. Pham, T. Guerrero, H.A. Al-Hallaq, "Lung Texture in Serial Thoracic Computed Tomography Scans: Correlation of Radiomics-based Features With Radiation Therapy Dose and Radiation Pneumonitis Development," International Journal of Radiation Oncology Biology Physics **91**, 1048-1056 (2015).

[101] M. Scrivener, E.E.C. de Jong, J.E. van Timmeren, T. Pieters, B. Ghaye, X. Geets, "Radiomics applied to lung cancer: a review," Translational Cancer Research **5**, 398-409 (2016).

[102] S.C.B. Lo, H.-P. Chan, J.S. Lin, H. Li, M. Freedman, S.K. Mun, "Artificial Convolution neural network for medical image pattern recognition," Neural Networks **8**, 1201-1214. (1995).

[103] R.K. Samala, H.-P. Chan, Y. Lu, L. Hadjiiski, J. Wei, B. Sahiner, M.A. Helvie, "Computer-aided detection of clustered microcalcifications in multiscale bilateral filtering regularized reconstructed digital breast tomosynthesis volume," Medical Physics **41**, 021901-021901 (021914 pages) (2014).

[104] R.K. Samala, H.-P. Chan, L.M. Hadjiiski, M.A. Helvie, J. Wei, K. Cha, "Mass Detection in Digital Breast Tomosynthesis: Deep Convolutional Neural Network with Transfer Learning from Mammography," Medical Physics **43**, 6654-6666 (2016).

[105] H.C. Shin, H.R. Roth, M.C. Gao, L. Lu, Z.Y. Xu, I. Nogues, J.H. Yao, D. Mollura, R.M. Summers, "Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning," Ieee Transactions on Medical Imaging **35**, 1285-1298 (2016).

[106] M. Kallenberg, K. Petersen, M. Nielsen, A.Y. Ng, P.F. Diao, C. Igel, C.M. Vachon, K. Holland, R.R. Winkel, N. Karssemeijer, M. Lillholm, "Unsupervised Deep Learning Applied to Breast Density

Segmentation and Mammographic Risk Scoring," Ieee Transactions on Medical Imaging **35**, 1322-1331 (2016).

[107] M.A. Helvie, L.M. Hadjiiski, E. Makariou, H.-P. Chan, N. Petrick, B. Sahiner, S.C.B. Lo, M. Freedman, D. Adler, J. Bailey, C. Blane, D. Hoff, K. Hunt, L. Joynt, K. Klein, C. Paramagul, S. Patterson, M.A. Roubidoux, "Sensitivity of noncommercial computer-aided detection system for mammographic breast cancer detection - A pilot clinical trial," Radiology **231**, 208-214 (2004).

[108] H.P. Chan, B. Sahiner, M.A. Helvie, N. Petrick, M.A. Roubidoux, T.E. Wilson, D.D. Adler, C. Paramagul, J.S. Newman, S.S. Gopal, "Improvement of radiologists' characterization of mammographic masses by computer-aided diagnosis: an ROC study," Radiology **212**, 817-827 (1999).

[109] Z.M. Huo, M.L. Giger, C.J. Vyborny, C.E. Metz, "Breast Cancer: Effectiveness of Computer-aided Diagnosis - Observer Study with Independent Database of Mammograms," Radiology **224**, 560-568 (2002).

[110] N. Petrick, M. Haider, R.M. Summers, S.C. Yeshwant, L. Brown, E.M. Iuliano, A. Louie, J.R. Choi, P.J. Pickhardt, "CT colonography with computer-aided detection as a second reader: Observer performance study," Radiology **246**, 148-156 (2008).

[111] T.W. Way, B. Sahiner, H.-P. Chan, L. Hadjiiski, P.N. Cascade, A. Chughtai, N. Bogot, E. Kazerooni, "Computer aided diagnosis of pulmonary nodules on CT scans: improvement of classification performance with nodule surface features," Medical Physics **36**, 3086-3098 (2009).

[112] D.D. Dorfman, K.S. Berbaum, C.E. Metz, N.A. Obuchowski, H. Rockette, "http://perception.radiology.uiowa.edu/Software/ReceiverOperatingCharacteristicROC/MRMCAnalysis/tabid/116/Default.aspx."

[113] N. Petrick, B. Sahiner, S.G. Armato, A. Bert, L. Correale, S. Delsanto, M.T. Freedman, D. Fryd, D. Gur, L. Hadjiiski, Z.M. Huo, Y.L. Jiang, L. Morra, S. Paquerault, V. Raykar, F. Samuelson, R.M. Summers, G. Tourassi, H. Yoshida, B. Zheng, C. Zhou, H.-P. Chan, "Evaluation of computer-aided detection and diagnosis systems," Medical Physics **40**, 087001-087001 (087017 pages) (2013).