Article Type: Research Article

Urinary Bladder Cancer Staging in CT Urography using Machine Learning

Sankeerth S. Garapati
Lubomir Hadjiiski
Kenny H. Cha
Heang-Ping Chan
Elaine M. Caoili
Richard H. Cohan
Alon Weizer*
Ajjai Alva**
Chintana Paramagul
Jun Wei
Chuan Zhou

Department of Radiology, The University of Michigan, Ann Arbor, Michigan 48109

**Department of Urology, Comprehensive Cancer Center, The University of Michigan, Ann Arbor, Michigan 48109*

***Department of Internal Medicine, Hematology-Oncology, The University of Michigan, Ann Arbor, Michigan 48109*

Running Title: Bladder Cancer Staging by Machine Learning

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/mp.12510

Correspondence:

Lubomir Hadjiiski
Department of Radiology
University of Michigan
1500 E. Medical Center Drive
MIB C476
Ann Arbor, MI 48109-5842
Telephone: (734) 647-7428
Fax: (734) 615-5513
E-mail:lhadjisk@umich.edu

ABSTRACT

Purpose: To evaluate the feasibility of using an objective computer aided system to assess bladder cancer stage in CT Urography (CTU).

Materials and Methods: 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).

Results: 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.

Conclusion: 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.

Keywords: Radiomics, Computer-Aided Diagnosis, CT Urography, Bladder Cancer Staging, Segmentation, Feature Extraction, Classification, Machine Learning.

1. INTRODUCTION

Bladder cancer is one of the most common cancers affecting both men and women[1]. It can cause substantial morbidity and mortality among the patients with the disease. In 2017, it is estimated that there will be 79,030 new cases and 16,870 deaths from bladder cancer[1]. One in 42 Americans will be diagnosed with bladder cancer in their lifetime and 9 out of 10 patients with this cancer are over the age of 55[1,2]. The average age of diagnosis is 73[1]. Approximately half of all bladder cancer cases are first found while the cancer is still confined to the inner wall of the bladder and has not invaded into deeper layers or distant parts of the body[1]. Bladder cancer has a recurrence rate of 50-80 percent and requires constant surveillance. This makes it the most expensive cancer to treat, requiring a total of $4.1 billion yearly, on a per patient basis in the United States[2]. Bladder cancer can be divided into three categories that include noninvasive, superficial, and invasive. The initial treatment for bladder cancer is transurethral resection of the bladder tumor (TURBT), which removes the tumor from the bladder and also helps provide information regarding the stage of the cancer[3-5]. 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[6] (Fig. 1). An example of bladder cancer stage T2 is presented in Fig. 2.

The 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)[3-5]. 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[7]. Bladder 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[6]. Adjuvant chemotherapy is used in patients with locally advanced bladder cancer in order to reduce the chances of cancer recurrence following radical cystectomy[8]. 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 improve survival when compared to radical cystectomy alone[8-10].

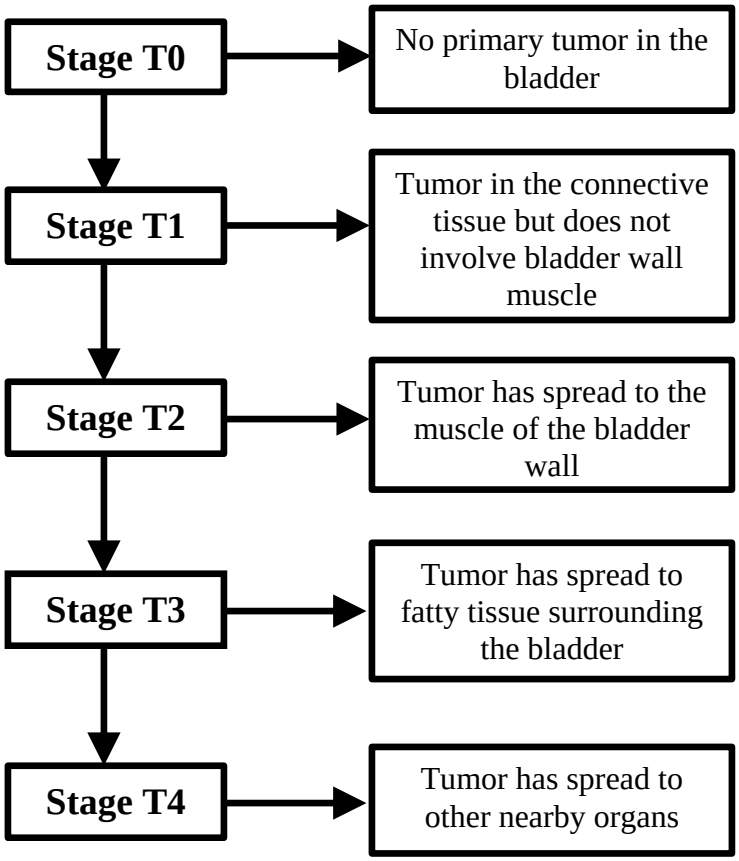


Figure 1. Bladder cancer stage grading scale definition.

Correct staging of bladder cancer is crucial for the decision of neoadjuvant chemotherapy treatment and minimizing the risk of under-treatment or over-treatment. Patients with stage T2 to T4 carcinomas of the bladder are recommended for treatment with 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[11-14]. This inaccuracy in staging can partly be attributed to the subjectivity and variability of clinicians in utilizing various diagnostic information. 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.

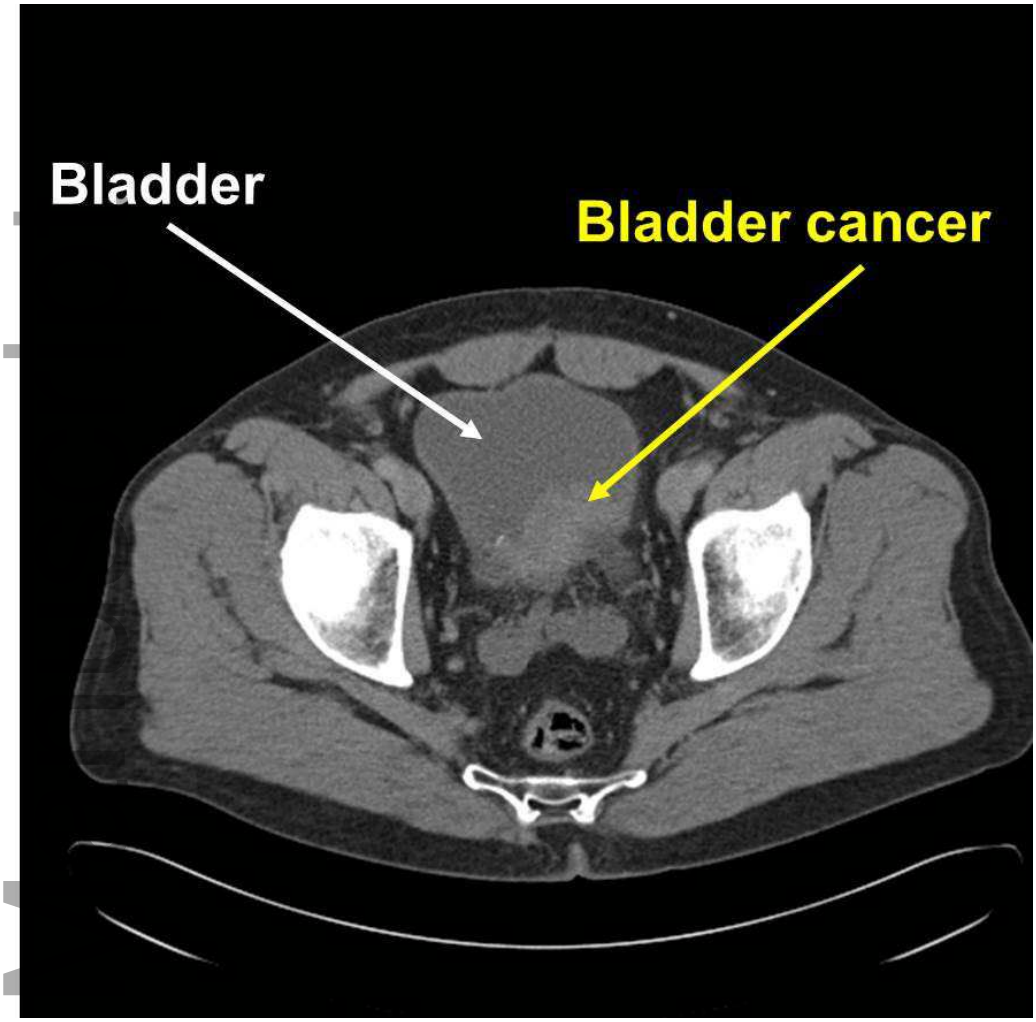


Figure 2. Urinary Bladder CT. The bladder cancer is marked and clearly visible. The cancer stage is T2.

2. MATERIALS AND METHODS

2.1 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 at an image slice interval of 0.625 to 1.25 mm using 120 kVp and 120-280 mA. 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 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, renal insufficiency, significant cardiac disease, or anxiety[15].

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 were 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.

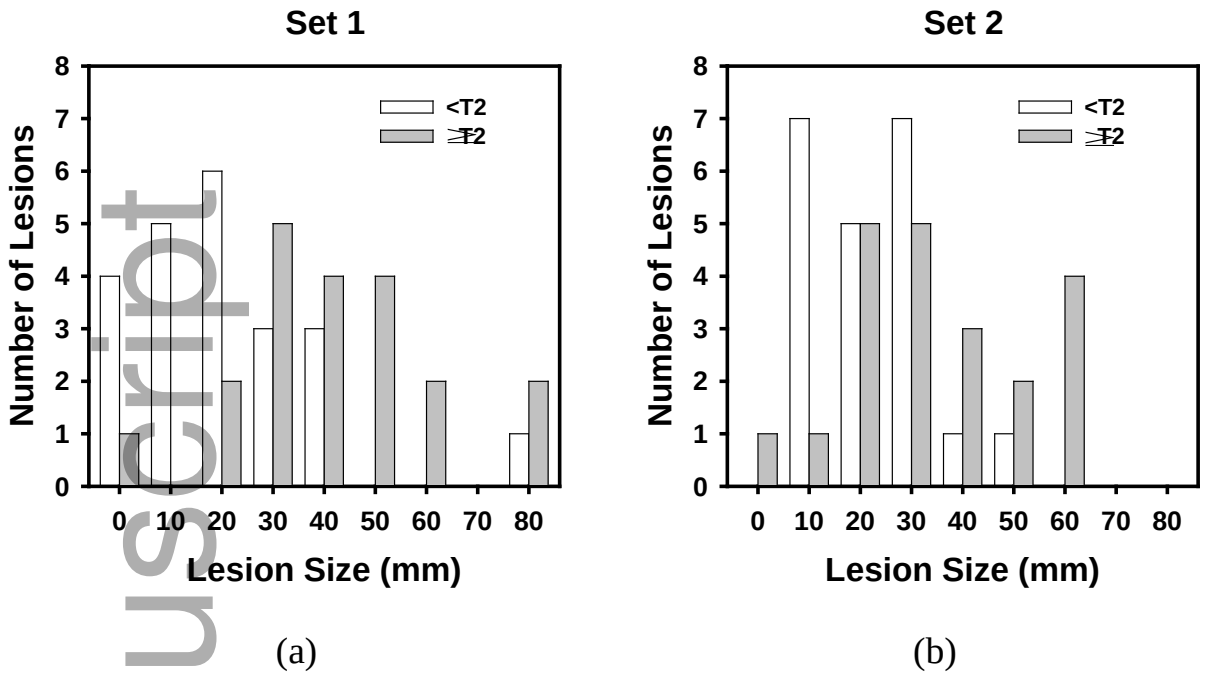


Figure 3. Distribution of tumor sizes (the longest diameters) for Set 1 and Set 2. (a) Set 1: The average tumor sizes of stage < T2 and≥ T2 were 26.4±17.3 mm and 45.6±19.1 mm respectively. (b) Set 2: The average tumor sizes of stage < T2 and ≥ T2 were 27.3±10.8 mm and 40.6±17.3 mm respectively.

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 (Fig. 3a). 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 (Fig. 3b).

2.2 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[16]. Briefly, the system consists of three stages that include preprocessing, initial segmentation, and 3D level set segmentation (Fig. 4). The segmentation of bladder lesions is often difficult as some lesions are located in the non-contrast enhanced region of the bladder such that contrast between the lesion and the surrounding background was low. Additionally, lesions often have irregular boundaries and can be very small and subtle. Each

lesion in the data set was marked by a bounding box as an input volume of interest (VOI). The lateral dimensions of the box were determined by an adjustable rectangle within the image slice that contains the best view of the lesion. The top and bottom slices are marked to completely enclose the lesion. The AI-CALS segmentation is then automatically performed in the VOI. In the pre-processing stage, image processing techniques including smoothing, anisotropic diffusion, gradient filters, and a rank transform of the gradient magnitude are used to generate sets of smoothed images, gradient magnitude images, and gradient vector images. The initial segmentation surface is obtained by combining information from these images. Three dimensional (3D) flood fill algorithm, morphological dilation filter, and morphologic erosion filter are applied to the initial segmentation surface to connect nearby components, which is then used to initialize the level set segmentation. The initial contour is propagated toward the lesion boundary using a bank of cascaded level sets. The level sets help refine the initial contour. The details of the AI-CALS method can be found in our previous paper[16].

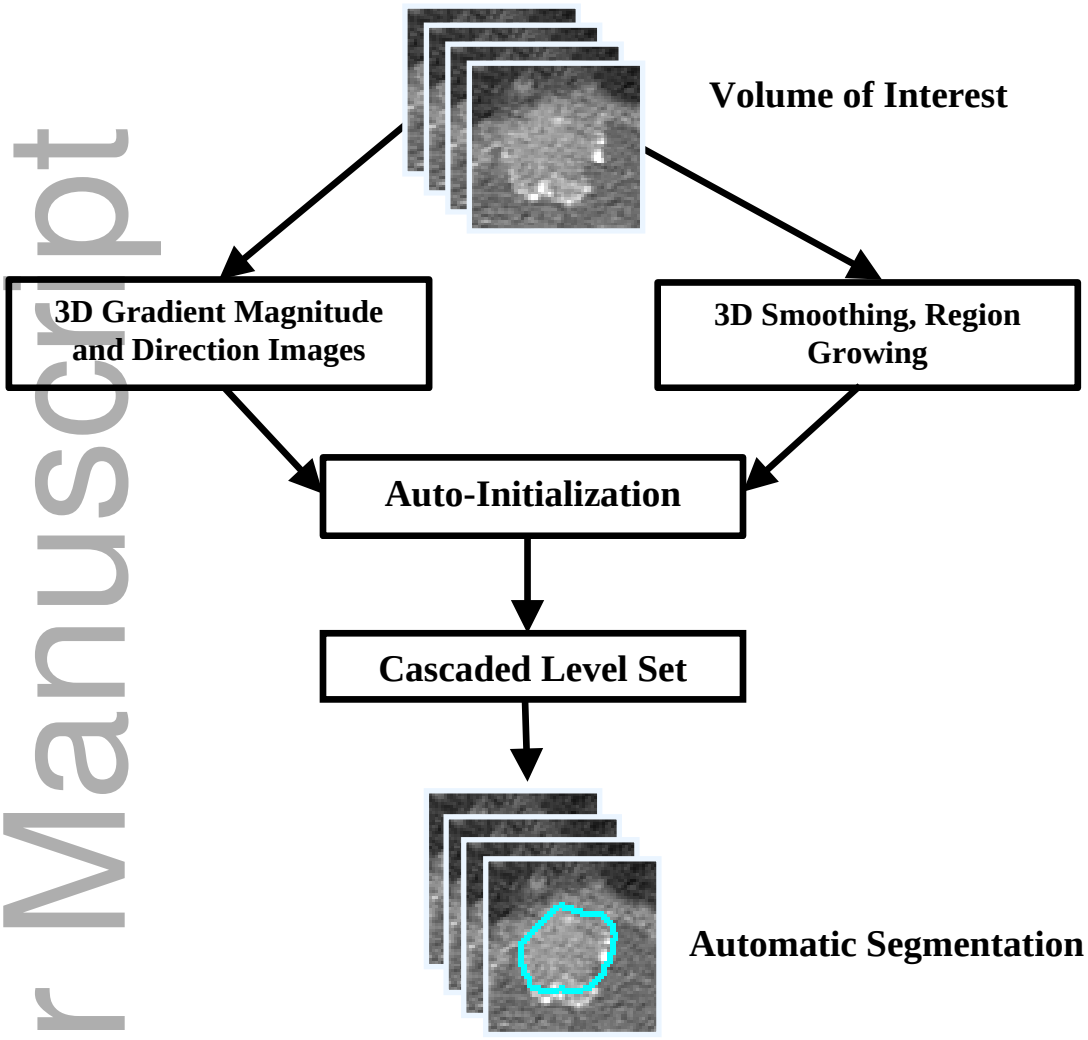


Figure 4. Block diagram of the auto-initialized cascaded level sets (AI-CALS) method.

3. CLASSIFICATION

3.1 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[17]. 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[18] of the tumor margin including those from the run-length statistics matrices, filtered Dasarathy east-west direction and filtered Dasarathy horizontal direction[19,20]. 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.

3.2 Feature Selection/Classification

A block diagram of the machine learning based bladder cancer staging system is shown in Fig. 5. Stepwise feature selection was used to select the best subset of features to create an effective classifier[21]. 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[21]. 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, which 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)[22,23]. 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) [24] 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)[25,26] 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 to 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[27]. We used the WEKA[28] 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.

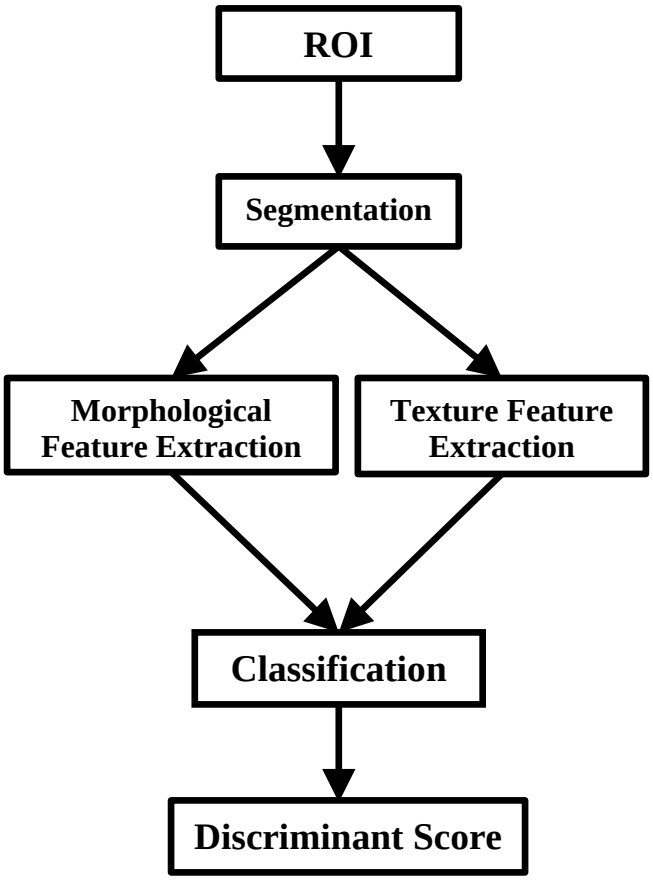


Figure 5. Block diagram of our machine learning based staging system. We compared the linear discriminant analysis (LDA), back-propagation neural network (NN), Support vector machine (SVM), and Random forest classifiers (RAF) in the classification stage for this study.

3.3 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. The average distance, *AVDIST*, is defined as the average of the distances between the closest points of the two contours:

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

where *G* and *U* are two contours being compared. 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.

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:

$$JACCARD^{3D} = \frac{V_G \cap V_U}{V_G \cup V_U}, \quad (2)$$

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.

To evaluate the classifier performance, the training and test scores output from the classifier were analyzed using the receiver operating characteristic (ROC) methodology[30]. 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.[31,32].

4. RESULTS

The lesion segmentation performance of the AI-CALS compared to the radiologist hand outlines for the 84 lesions are shown in Table 1. Table 2 shows the computer segmentation performance compared to two different radiologists' hand outlines for a subset of 12 lesions.

Table 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 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 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 are presented in Fig 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 Fig. 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 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 between 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

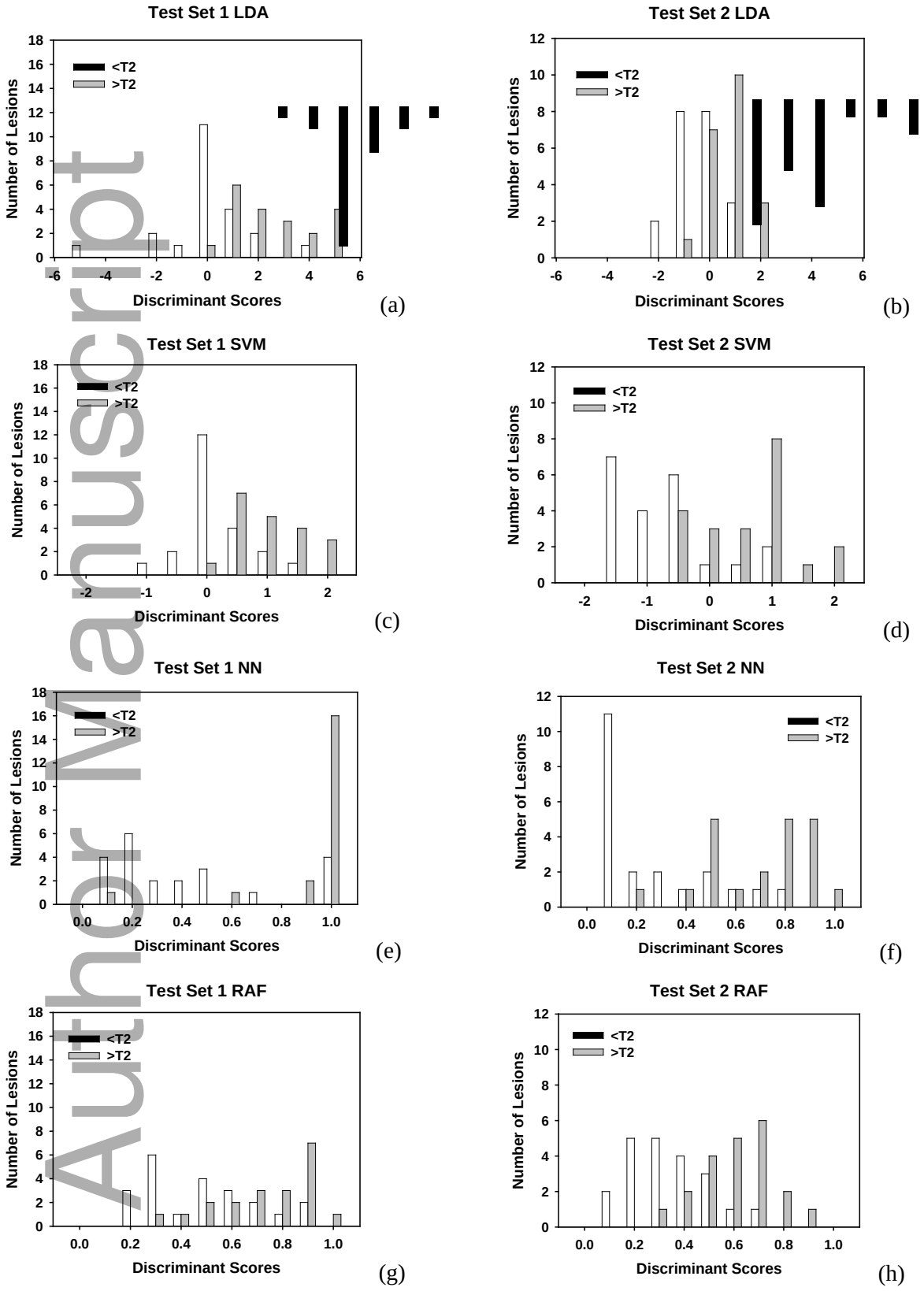


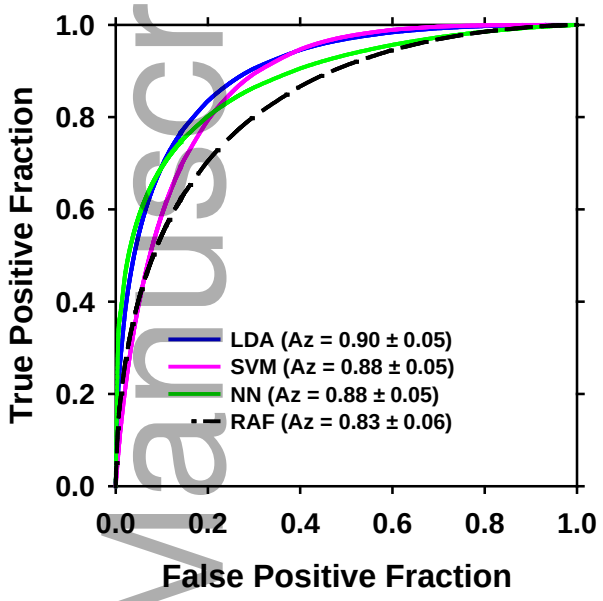
Figure 6. Distribution of the classifiers discriminant scores for testing on Set 1 and Set 2 in two-fold

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.

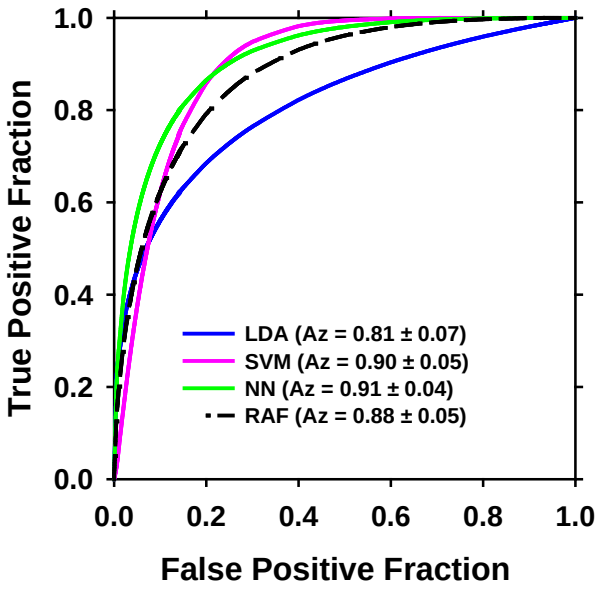
337

338

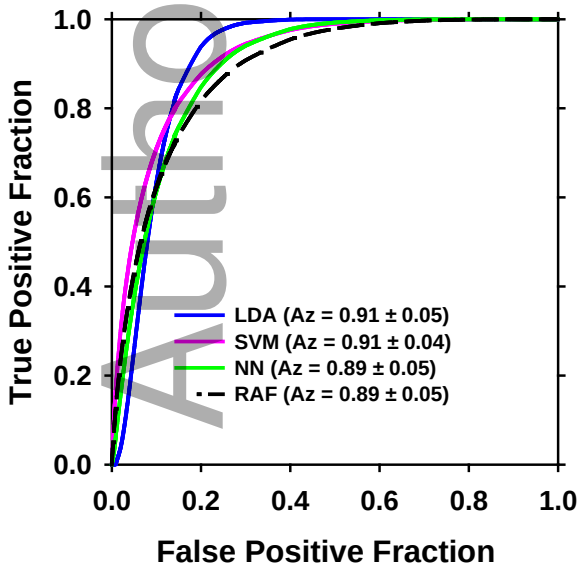
339

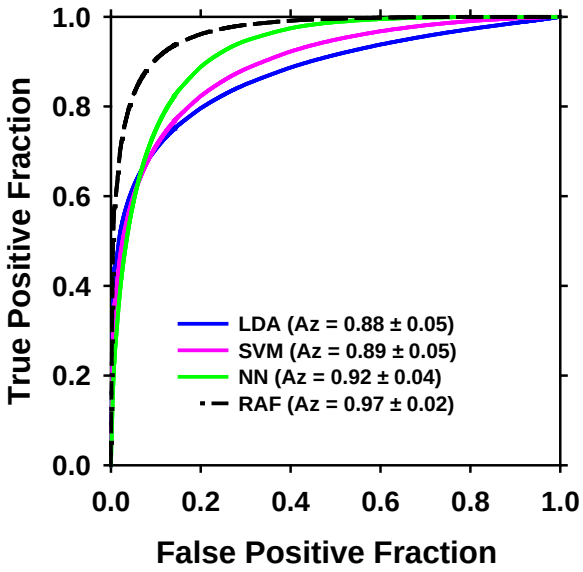


(a)

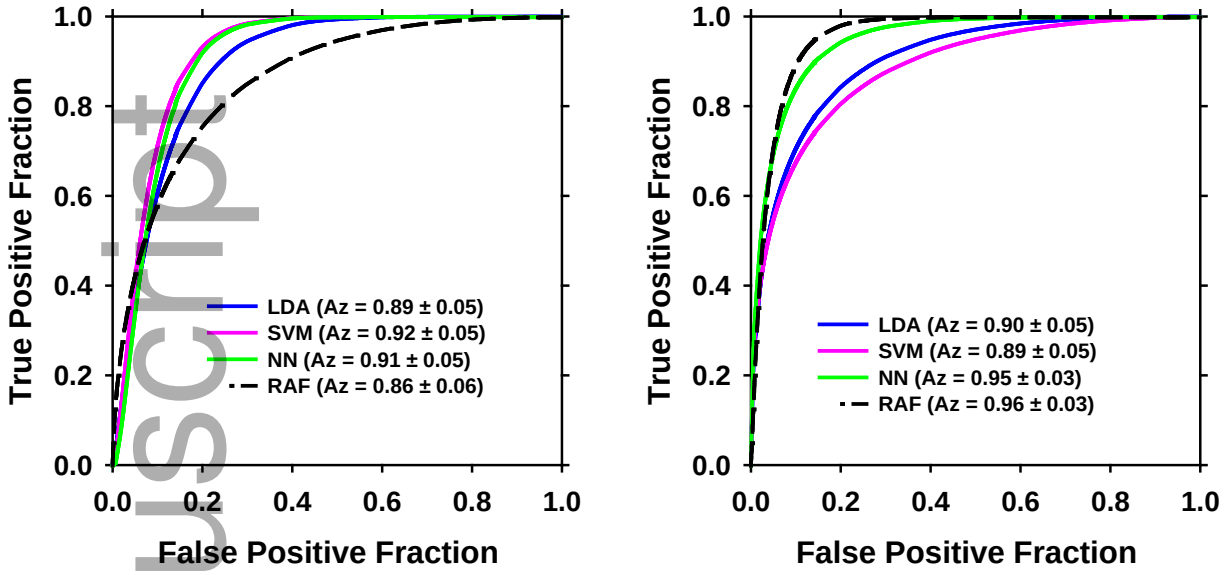


(b)





(c) (d)



(e) (f)

Figure 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.

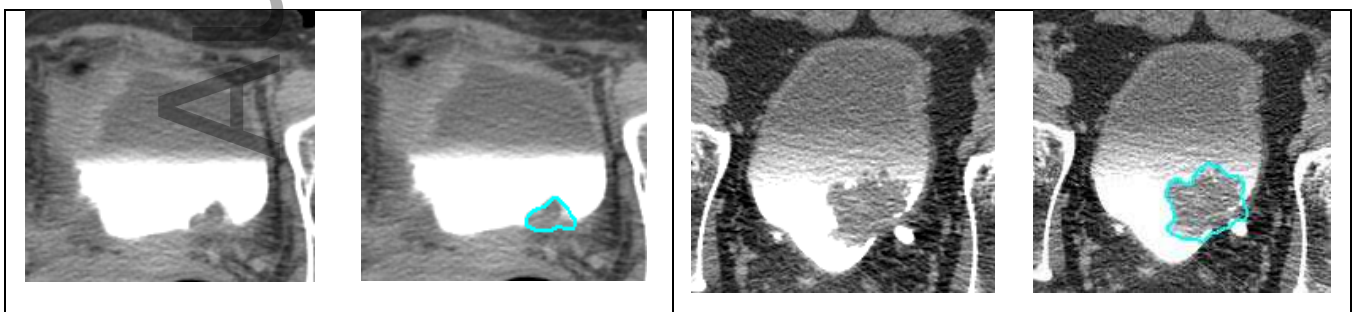
5. DISCUSSION

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. 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 ≥ T2 or < T2 and the corresponding classifier scores are shown in Fig. 8. The reported scores are test scores for the LDA, SVM, NN, and RAF classifiers based on the morphological features. In Fig. 8a, b and Fig. 8c, 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 Fig. 8e, f and Fig 8g, 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 Fig. 8k, l. The classifiers classified the cancer as ≥T2 with high scores. Fig. 8m, 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.



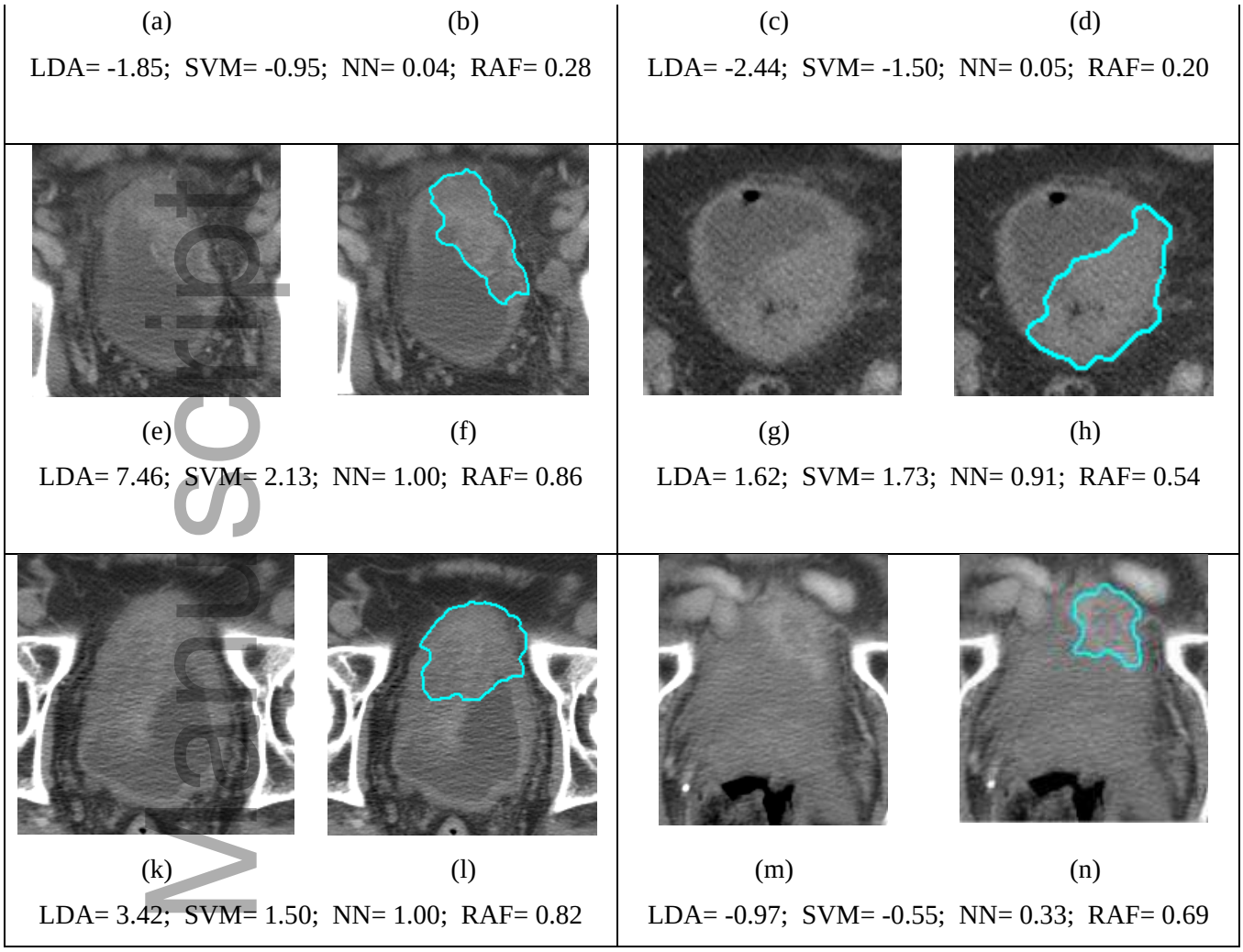


Figure 8. Examples of bladder cancers with stages ≥ T2 or < T2. The blue outlines represent the AI -CALS segmentation. The reported scores are test scores for the LDA, SVM, NN, and RAF classifiers based on the morphological features. Note that the output score ranges are different for different classifiers 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[33,34] including bladder segmentation[35] and bladder cancer treatment response monitoring[36]. 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 the latter. We will continue to collect additional cases and compare the conventional classifiers with DLCNN for bladder cancer staging in a future study.

6. 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.

Acknowledgments

This work is supported by National Institutes of Health grant number U01CA179106.

References

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

[2]"Bladder Cancer Advocacy Network, www.bcan.org/facts 2017, "Bladder Cancer Facts" " (2017).

[3]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, et al., "Diagnosis and Treatment of Non-Muscle Invasive Bladder Cancer: AUA/SUO Guideline," Journal of Urology 196, 1021-1029 (2016).

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

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

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

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

[8]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, et al., "A Systematic Review of Neoadjuvant and Adjuvant Chemotherapy for Muscle-invasive Bladder Cancer," European Urology 62, 523-533 (2012).

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

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

[11]J. Huguet, M. Crego, S. Sabate, J. Salvador, J. Palou, and 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).

[12]H. M. Fritsche, M. Burger, R. S. Svatek, C. Jeldres, P. I. Karakiewicz, G. Novara, E. Skinner, S. Denzinger, Y. Fradet, H. Isbarn, et al., "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).

[13]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, et al., "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).

[14]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, and Y. Lotan, "Discrepancy between clinical and pathologic stage: Impact on prognosis after radical cystectomy," European Urology 51, 137-151 (2007).

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

[16]L. M. Hadjiiski, H.-P. Chan, E. M. Caoili, R. H. Cohan, J. Wei, and 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).

[17]L. M. Hadjiiski, B. Sahiner, H.-P. Chan, N. Petrick, M. A. Helvie, and 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).

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

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

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

[21]H.-P. Chan, D. Wei, M. A. Helvie, B. Sahiner, D. D. Adler, M. M. Goodsitt, and 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).

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

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

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

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

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

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

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

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

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

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

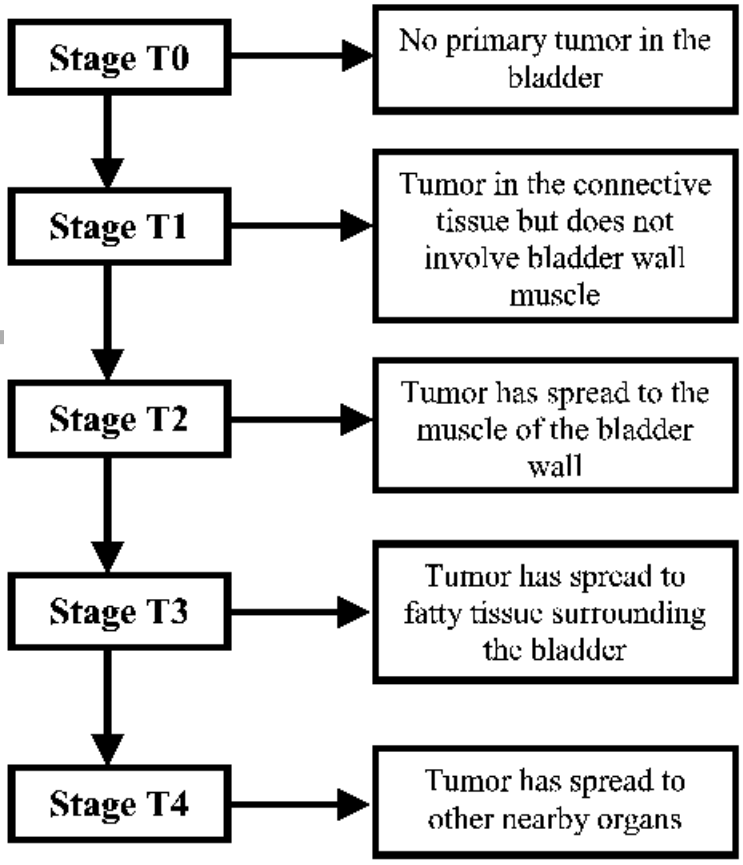
[32]"Metz ROC Software. University of Chicago Medical Center Department of Radiology, see http://metz-roc.uchicago.edu/MetzROC/software,"

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

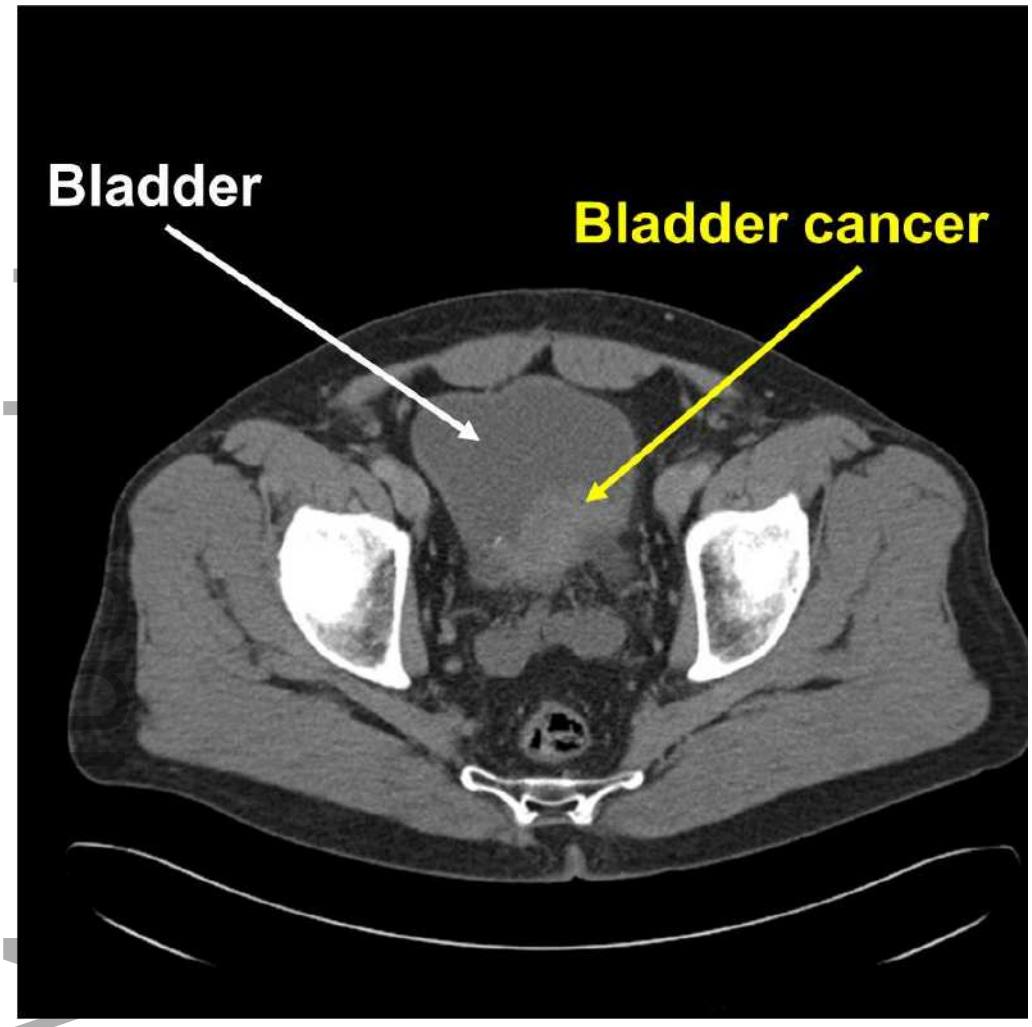
[34]H. Greenspan, B. van Ginneken, and R. M. Summers, "Deep Learning in Medical Imaging: Overview and Future Promise of an Exciting New Technique," Ieee Transactions on Medical Imaging 35, 1153-1159 (2016).

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

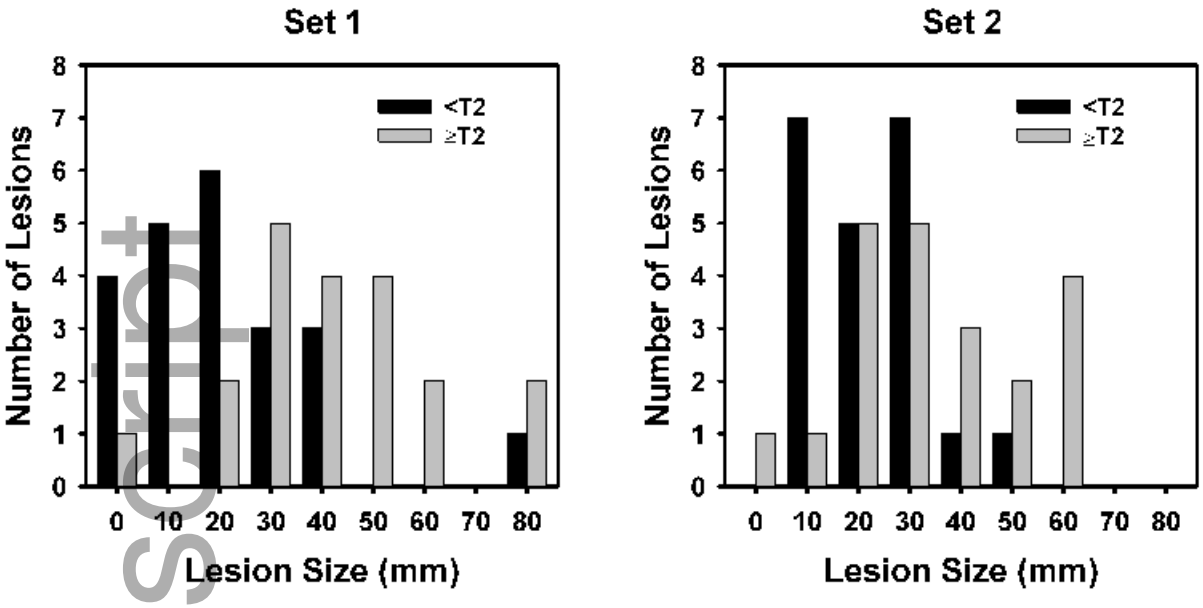
[36]K. H. Cha, L. M. Hadjiiski, H.-P. Chan, R. K. Samala, R. H. Cohan, E. M. Caoili, C. Paramagul, A. Alva, and A. Z. Weizer, "Bladder cancer treatment response assessment using deep learning in CT with transfer learning," Proc SPIE 10134, 101341-6 (2017).



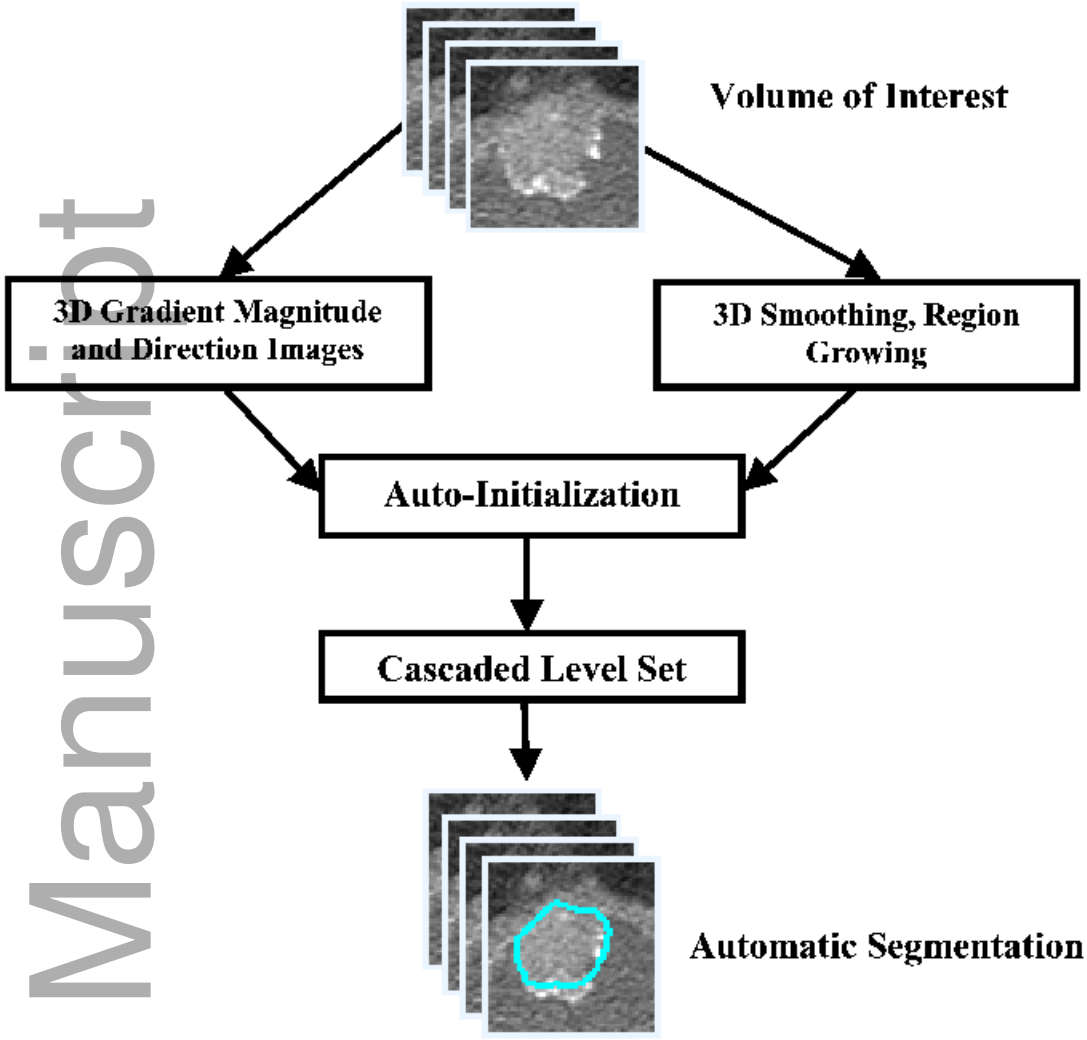
mp_12510_f1.tif



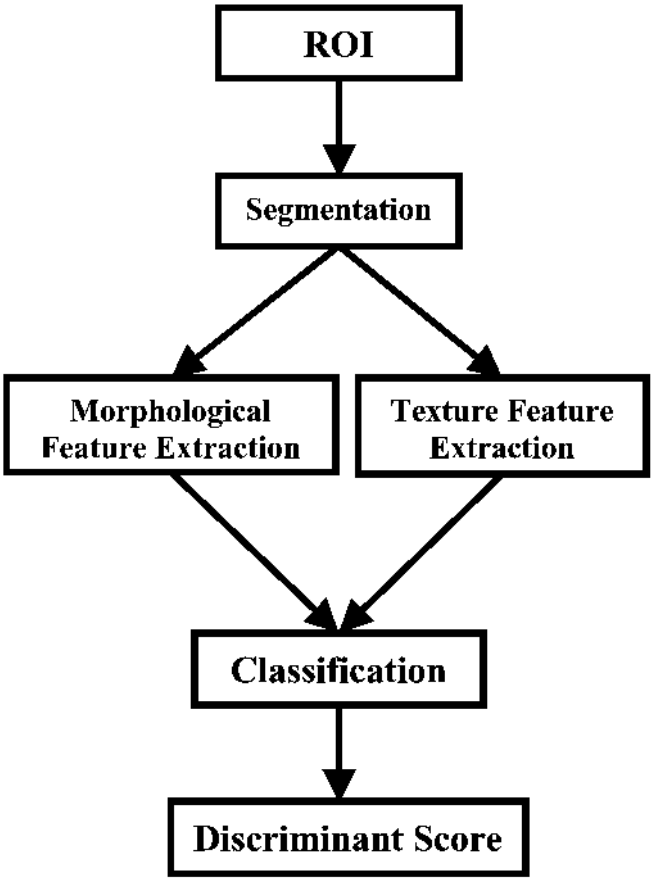
mp_12510_f2.tif



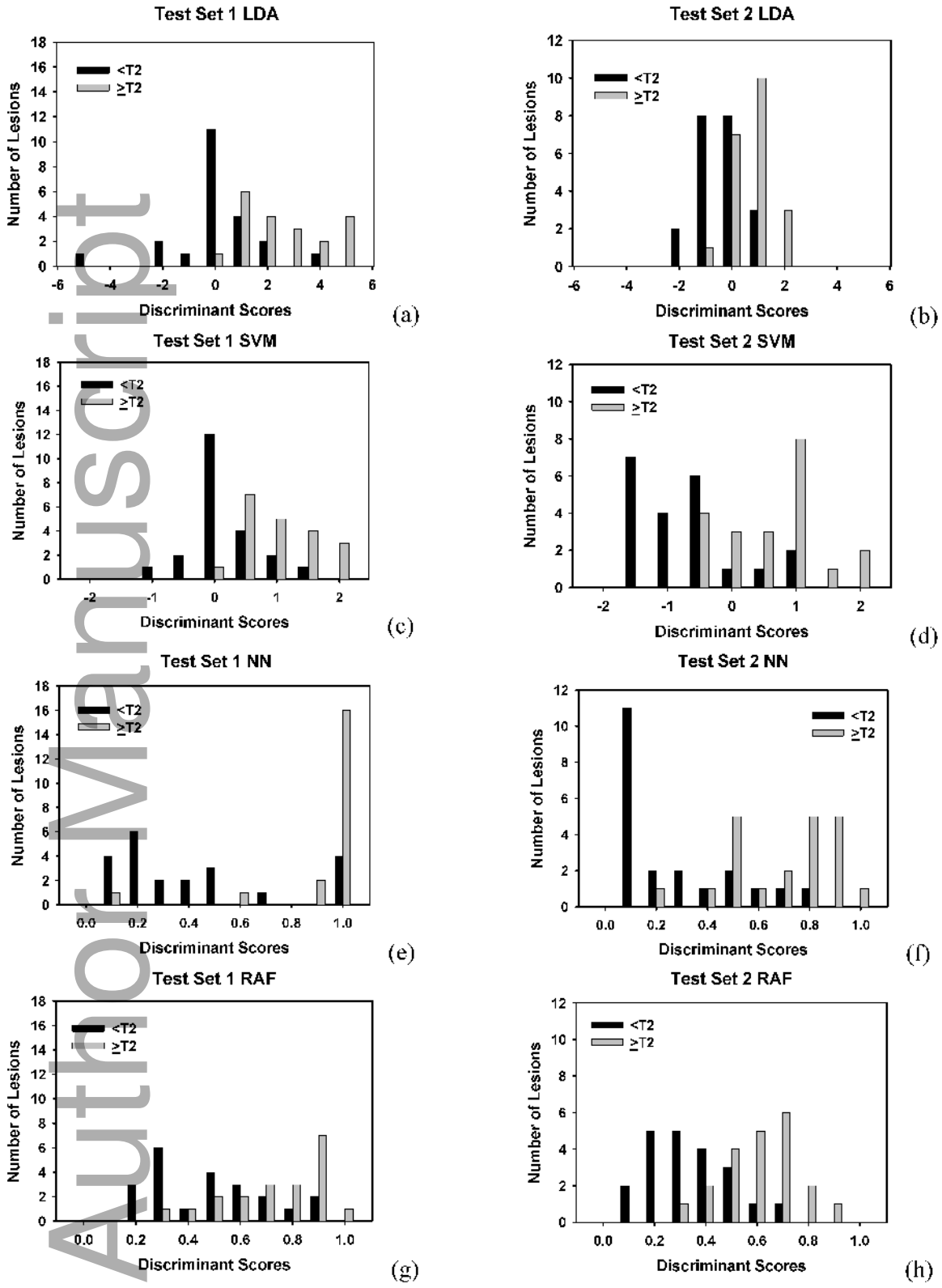
mp_12510_f3.tif



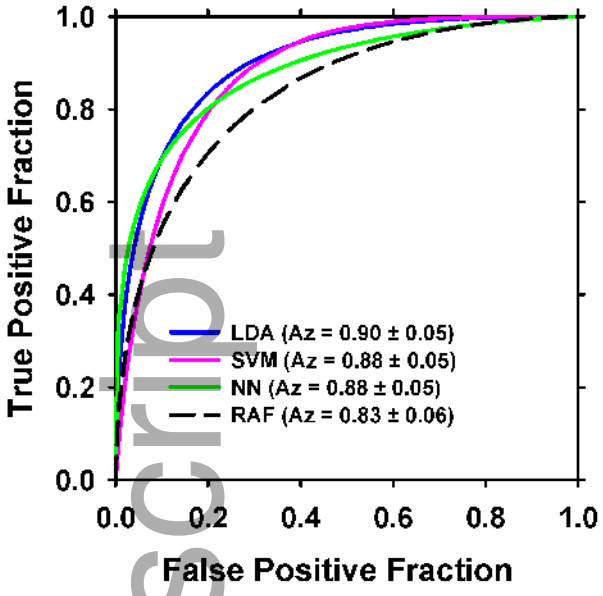
mp_12510_f4.tif



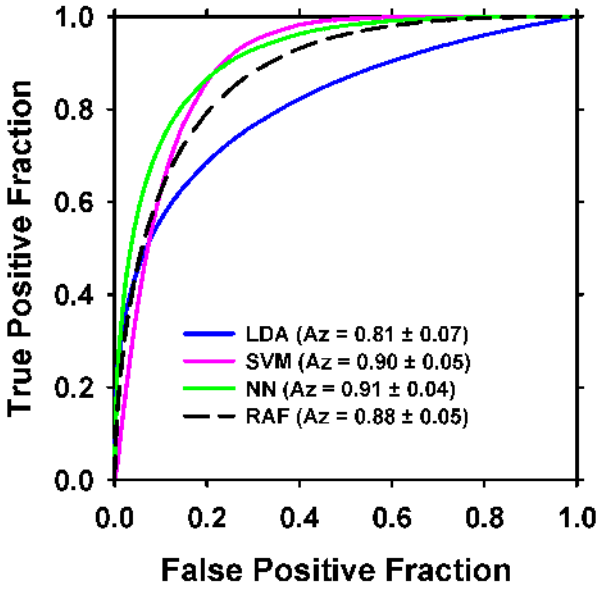
mp_12510_f5.tif



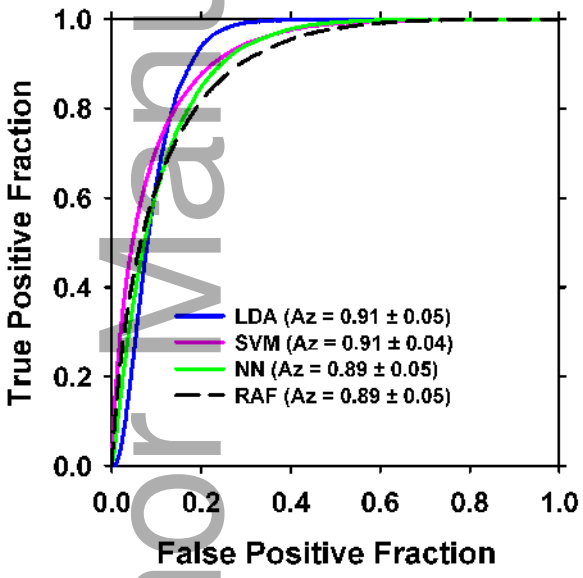
mp_12510_f6.tif



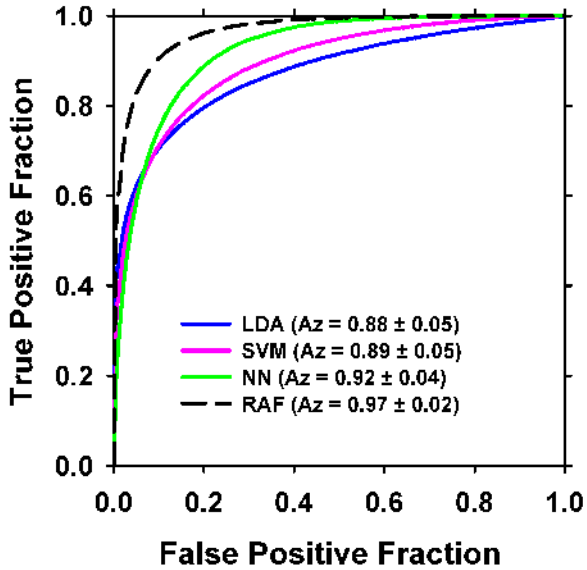
(a)



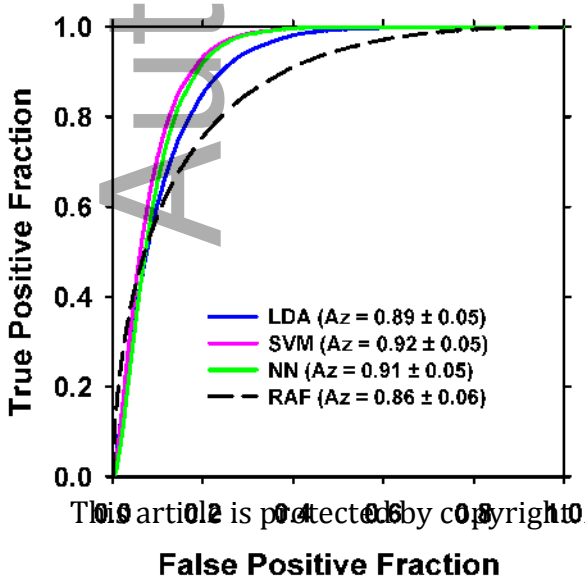
(b)



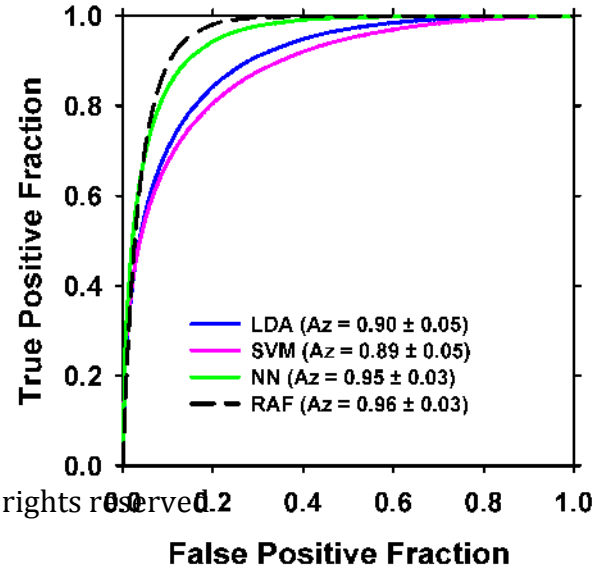
(c)



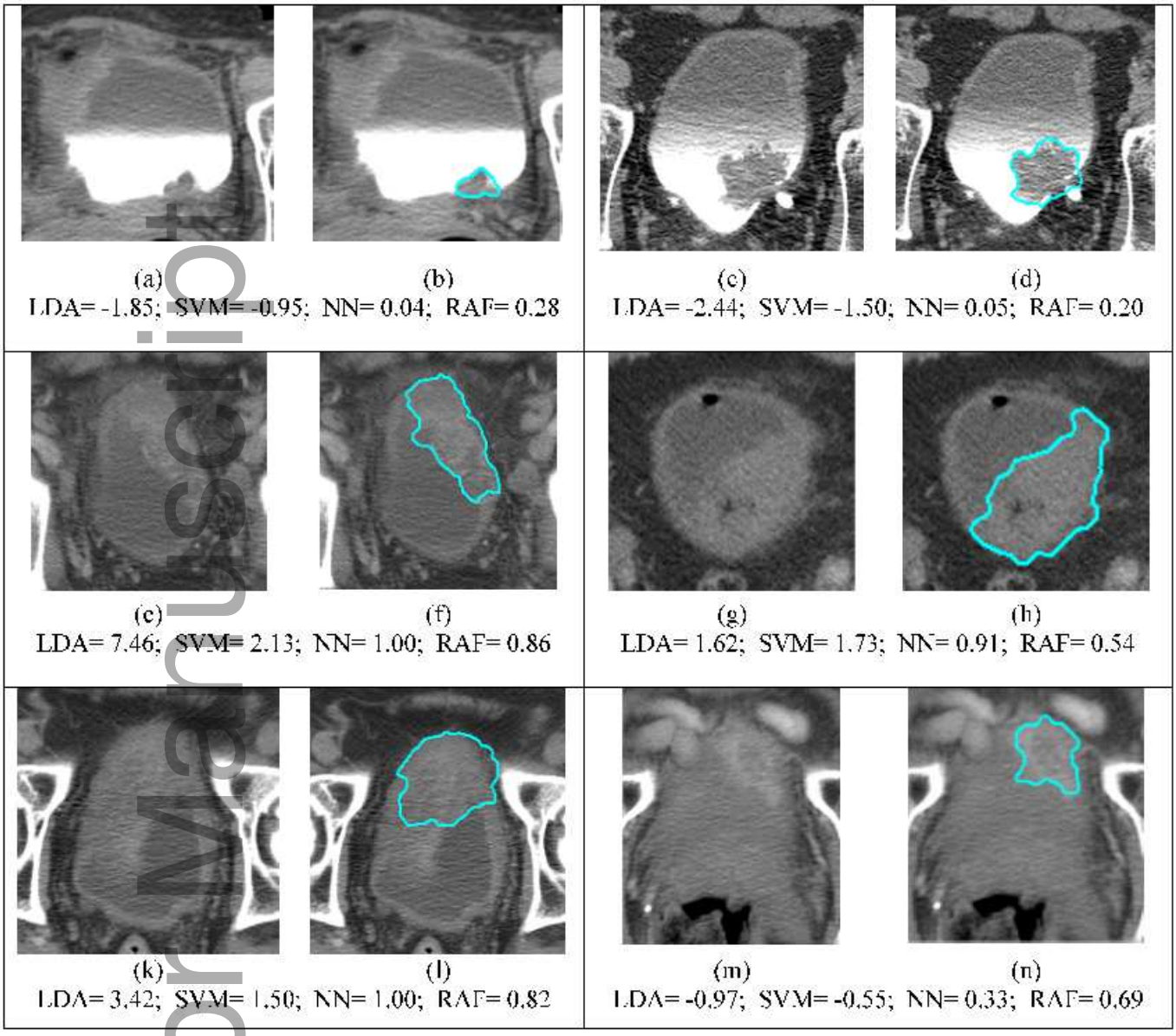
(d)



(e)



(f)



mp_12510_f8.tif