Detection of prostate cancer in multiparametric MRI using random forest with instance weighting

Nathan Lay
Yohannes Tsehay
Matthew D. Greer
Baris Turkbey
Jin Tae Kwak
Peter L. Choyke
Peter Pinto
Bradford J. Wood
Ronald M. Summers

Detection of prostate cancer in multiparametric MRI using random forest with instance weighting

Nathan Lay,^a Yohannes Tsehay,^a Matthew D. Greer,^b Baris Turkbey,^b Jin Tae Kwak,^c Peter L. Choyke,^b Peter Pinto,^b Bradford J. Wood,^c and Ronald M. Summers^{a,*}

^aNational Institutes of Health, Clinical Center, Imaging Biomarkers and Computer Aided Diagnosis Laboratory, Bethesda, Maryland, United States

^bNational Institutes of Health, National Cancer Institute, Urologic Oncology Branch and Molecular Imaging Program, Bethesda, Maryland, United States

^cNational Institutes of Health, Clinical Center, Center for Interventional Oncology, Bethesda, Maryland, United States

Abstract. A prostate computer-aided diagnosis (CAD) based on random forest to detect prostate cancer using a combination of spatial, intensity, and texture features extracted from three sequences, T2W, ADC, and B2000 images, is proposed. The random forest training considers instance-level weighting for equal treatment of small and large cancerous lesions as well as small and large prostate backgrounds. Two other approaches, based on an AutoContext pipeline intended to make better use of sequence-specific patterns, were considered. One pipeline uses random forest on individual sequences while the other uses an image filter described to produce probability map-like images. These were compared to a previously published CAD approach based on support vector machine (SVM) evaluated on the same data. The random forest, features, sampling strategy, and instance-level weighting improve prostate cancer detection performance [area under the curve (AUC) 0.93] in comparison to SVM (AUC 0.86) on the same test data. Using a simple image filtering technique as a first-stage detector to highlight likely regions of prostate cancer helps with learning stability over using a learning-based approach owing to visibility and ambiguity of annotations in each sequence. © 2017 Society of Photo-Optical Instrumentation Engineers (SPIE) [DOI: [10.1117/1.JMI.4.2.024506](https://doi.org/10.1117/1.JMI.4.2.024506)]

Keywords: computer-aided diagnosis; prostate; multiparametric MRI.

Paper 17054RR received Mar. 4, 2017; accepted for publication May 12, 2017; published online Jun. 12, 2017.

1 Introduction

Prostate cancer is the sixth most common cancer worldwide.¹ A widely used method to diagnose prostate cancer involves randomly biopsying the prostate in 10 to 12 locations throughout the prostate.² Due to their random nature, biopsies can miss clinically significant lesions and result in either false negatives or inappropriately downgrading the cancer. Incidental findings of prostate cancer are also frequent and may result in overdiagnosis and overtreatment. More recently, multiparametric MRI (mpMRI) has been demonstrated to be the most accurate imaging technique to detect prostate cancer. mpMRI-guided biopsies improve sensitivity for clinically significant prostate cancer while limiting overdiagnosis.³

Computer-aided diagnosis (CAD) of prostate cancer is expected to increase sensitivity further. It can quickly interpret the plethora of images available in mpMRI and assist radiologists with less experience. Contemporary literature demonstrates sensitivity from mpMRI alone to be as much as 81%.² This work proposes a prostate cancer CAD system for the purpose of recommending biopsy sites. The CAD in this work is developed using learning-based methods, namely random forest,⁴ and produces a probability image from T2W, ADC, and B2000 MRI images. The probability image can then be visualized by a prostate reader. The originality of this application includes the use of an AutoContext⁵ detection pipeline, a sample weighting strategy, and the use of various kinds of ground truth information of varying quality. We validate the proposed CAD on a dataset

of 224 patient scans and compare the performance with a previously published CAD operating on the same dataset.

1.1 Related Works

There are several examples of prostate cancer CAD systems in the literature. Some recent notable works include^{6–8} the CAD system in Ref. 6, which uses support vector machine (SVM) with local binary pattern features applied to T2W and B2000 images to predict whether pixels were likely to be cancer or not. This SVM was trained against a cohort of 264 patients with recorded biopsy results. The prostate CAD in Ref. 7 developed a voxel-wise classification approach and a lesion segmentation algorithm. The classifiers used were a linear discriminant classifier, GentleBoost, and random forest and used T2W, ADC, B800, and dynamic contrast enhanced (DCE)-images. The features considered were pixel intensity, relative location in the prostate, texture, peripheral zone likelihood, blobness, and pharmacokinetic heuristics. In Ref. 8, prostate cancer candidates were extracted from ADC using a multiscale Hessian-based blobness filter, and then these candidates were further classified using an linear discriminant analysis classifier with statistics-based features computed from T1W, T2W, DCE, ADC, and Ktrans. While the learning task is slightly different, the work in Ref. 9 employed intensity statistics and Haralick texture features to classify Gleason scores from MR images using a variety of classifiers. A noteworthy contribution was the use of SMOTE to deal with sample imbalance.

*Address all correspondence to: Ronald M. Summers, E-mail: rms@nih.gov

This work differs with previous CAD systems by using instance-level weighting to deal with label imbalance as well as for fair treatment between small and large cancerous lesions and for fair treatment between small and large prostates. This is one approach to dealing with sample imbalance with oversampling techniques, such as SMOTE, being an alternative approach. Furthermore, the CAD uses three kinds of annotations: hand-drawn contours, targeted biopsies, and normal cases, as well as develops sampling strategies for each of these types of annotations. Last, the CAD uses a transition zone distance as a feature to help the model specialize on peripheral and transition zones. This is slightly more powerful than the distance features considered in Ref. 7, which only consider distances to the boundary of the prostate but should be comparable to the peripheral zone likelihood feature used in Ref. 7.

2 Materials and Methods

This study was health insurance portability and accountability act-compliant, approved by our internal review board, and the need for signed informed consent was waived. This study incorporates 224 patient cases involving three types of annotations. MRIs were taken of patients prior to biopsy to establish the presence of cancer. These were then used to acquire MRI-TRUS fusion biopsies where the MRIs were used to guide the biopsy probe. Of the three types of annotations, there were 53 cases with hand-drawn cancerous lesion contours, 24 normal cases, and 147 MRI-TRUS fusion targeted biopsy cases. All cases additionally include an automatic segmentation of the prostate and transition zone produced by a commercial algorithm with some manual correction by an experienced radiologist. The cancer lesion contours were drawn on T2W images by an experienced radiologist examining mpMRI and biopsy results for these cases. The expertise and time required to manually contour cancerous lesions are generally prohibitive and motivate the use of other available information. Biopsy results are numerous and readily available. They are assumed to be reliable owing to the use of MRI-TRUS fusion to target suspicious regions in the prostate. Owing to factors such as human error controlling the biopsy needle and image registration problems, MRI-TRUS fusion biopsies can have inaccuracies, however little.¹⁰ There were a total of 200 patients with recorded biopsy results, including the 53 patients with hand-drawn contours. In each of these patients, biopsy sites were strategically picked from mpMRI by an experienced radiologist prior to the biopsy procedure. Of those biopsy sites, 287 were benign and 123 were Gleason 6 or higher. A table of the Gleason grade and zone breakdown is provided in Table 1. The remaining 24 normal cases were inspected in mpMRI by an experienced radiologist and were considered to have no visible suspicious lesions. This study additionally considers six test cases selected from consecutive prostate cancer patients (April 2012 to June 2015) who underwent mpMRI and prostatectomy (179 total) for the

purpose of comparing with wholemount histopathology in Figs. 1 and 2.

2.1 MRI Protocol

The mpMRI images were acquired using a 3T MR scanner (Achieva-TX, Philips Medical Systems, Best, the Netherlands) using the anterior half of a 32 channel SENSE cardiac coil (In vivo, Philips Healthcare, Gainesville, Florida) and an endorectal coil (ERC) (BPX-30, Medrad, Indianola, Pennsylvania). No pre-examination bowel preparation was required. The balloon of each ERC was distended with ~45 ml of perfluorocarbon (Fluorinert FC-770, 3M, St. Paul, Minnesota) to reduce imaging artifacts related to air-induced susceptibility. The T2W image was acquired with an echo time, repetition time, and section thickness of 4434 ms, 120 ms, and 3 mm, respectively. The standard diffusion weighted imaging (DWI) was acquired with five evenly spaced b -values (0–750), and high b -value DWI was acquired with $b = 2000$.

2.2 Prostate CAD

The proposed prostate CAD is based on a pixel-wise random forest trained against the three types of annotations. The CAD makes use of 13 types of Haralick texture features, four types of intensity features applied to T2W, ADC, and B2000 images, and the signed Euclidean distance to the transition zone. Two other configurations of the CAD based on AutoContext are considered. These CAD systems additionally use four types of intensity features on probability maps from each sequence. Table 2 provides a list of features used in the three systems. Figure 3 shows an example of the MRI sequences and annotations used by the proposed CAD and Fig. 4 illustrates the workflow used by the CAD.

Haralick texture features¹¹ are statistics computed on co-occurrence matrices in an attempt to characterize texture. They have found use in characterizing lesions in various structures throughout the body.^{9,12–14} Haralick features have recently been shown to discriminate not only between cancerous and noncancerous prostate lesions in T2W, ADC, and DCE images but also between different Gleason grades of prostate cancer in the same sequences.^{9,12} The ability to characterize texture and discriminate prostate tissue has motivated their inclusion in this work.

Random forest is an ensemble of random decision trees first described by Breiman.⁴ The predictions by random forest are an average of the predictions of its constituent random trees. A random forest is trained by individually training random trees on random subsets of data. Each random tree is trained in a similar fashion as a decision tree with the exception that a randomized subset of the features is used to determine an optimal

Table 1 Breakdown of severity of biopsied lesions used in the data set.

Zone	Benign	Gleason 6	Gleason 7	Gleason 8	Gleason 9	Gleason 10	Total
Peripheral	183	23	26	12	10	3	257
Transition	104	14	18	11	6	0	153
Overall	287	37	44	23	16	3	410

Fig. 1 Three cases with prominent false positives. The top row shows the prostate of a 67 year old with numerous false positives in the peripheral zone that renders the CAD more difficult to interpret (index lesion is marked 1 with Gleason score 7). The second row shows the prostate of another 67 year old with false positives along the midline base and underestimates the tumor burden of lesion 3 and 4 (index lesion is marked 3 with Gleason score 7). The third row shows the prostate of a 65 year old with several false positives in the right side and transition zone of the prostate (index lesion is marked 1 with Gleason score 7).

Fig. 2 Three cases with false negatives or underestimated tumor burden. The top row shows the prostate of a 50 year old with an underestimated lesion 1 in the transition zone (index lesion is marked 1 with Gleason score 7). The second row shows the prostate of a 64 year old with lesions 2 and 4 completely missed (index lesion is marked 2 with Gleason score 7). The third row shows the prostate of a 65 year old with a completely missed transition zone lesion (index lesion is marked 1 with Gleason score 7). The B2000 image in the third row shows no indication of cancer while ADC and T2W images show low-intensity regions that indicate a potential site of cancer.

decision. Optimality of a decision is determined by the information gain.

First introduced in Ref. 5, AutoContext uses a sequence of classifiers to improve classification accuracy in computer vision tasks. Consecutive classifiers use probability maps produced by

the previous classifier as a supplemental feature to the image features. The probability maps can highlight relatively prominent structure in an image and make the learning task relatively simpler. This work uses probability maps independently produced on T2W, ADC, and B2000 images as a feature to a

Table 2 Features used by the CAD systems. 2-D dimensions indicate that the CAD uses the feature and the patch window dimensions used to compute the feature. The probability map features are only used in the AutoContext-based methods.

	Type	T2W	ADC	B2000	Distance map	Probability map
Intensity	Pixel intensity	1 × 1 pixel				
	Mean					2 mm × 2 mm
	Median					
	Variance					
Haralick	Angular second Moment					Not used
	Contrast					
	Correlation					
	Sum of squares					
	Inverse difference moment					
	Sum average					
	Sum variance					
	Sum entropy					
	Entropy					
	Difference variance					
	Difference entropy					
	Information measure of correlation 1					
	Information measure of correlation 2

Fig. 3 An example of the (b) three sequences and (a) contour annotations. Here, blue is the whole prostate segmentation, green is the transition zone segmentation, and red is the cancerous lesions.

classifier that then produces the final probability map. This helps identify any sequence-specific patterns without correlating any coincidental patterns that might be present on other sequences. Admittedly, the final classifier can still overfit coincidental cross-sequence patterns.

Two AutoContext CAD pipelines were examined: the first applies random forest on each individual sequence using Haralick, intensity, and signed distance features to produce a sequence-specific probability map, and the second uses simple image filtering to produce pseudoprobability maps for each

Fig. 4 Visualization of the prostate CAD pipeline.

sequence. These probability maps are then used as supplemental features to random forest along with image features to produce a final probability map. The image filter in the second configuration subtracts the prostate intensity mean from the image and then sets the resulting pixels to 0 when they have an intensity less than or larger than 0, depending on the sequence (T2W, ADC larger than 0 and B2000 less than 0). This follows from the observation that lesions visible in T2W and ADC images have relatively lower intensity than normal prostate tissue and that lesions visible in B2000 tend to have relatively higher intensity. This filter rule thus behaves like a simple cancer classifier in each sequence.

2.3 Training

Three kinds of training annotations are used to train the random forest: hand-drawn contours, targeted biopsy results, and normal cases. Each tree of the random forest is trained on 20 uniform randomly sampled patient cases, regardless of the annotation type. The positives and negatives used to train the tree are sampled as follows:

1. Contours

Pixels occurring inside a hand-drawn cancerous lesion are taken as positive examples. Pixels 3 mm or further away from any hand-drawn cancerous lesion are taken as negative examples.

2. Targeted biopsy results

Pixels within 5 mm of any biopsy point (regardless of whether they are benign or not) are taken to be positive. Since these biopsies were targeted with the MRI-TRUS fusion approach, the lesions must have appeared abnormal to the radiologist and using them as negative examples were observed to harm detection performance of clinically significant prostate cancer. Furthermore, the practical task of the CAD is to mimic the radiologist

in choosing biopsy sites although this is a different problem than detection of prostate cancer. The chosen radius follows from the observation that lesions have an average diameter of about 10 mm.¹⁵ It is important to note that no negatives are sampled as the extent of the targeted lesion is not known from the biopsy itself. It may also be possible that other suspicious lesions were not biopsied, rendering sampling negatives risky.

3. Normal prostate cases

Every pixel inside a normal prostate is taken to be a negative example.

The three sampling strategies are shown in Fig. 5. This sampling scheme is unbalanced in a number of ways, which could bias a learning-based method. There are a differing number of positives and negatives, and the volumes of prostates and lesions vary by patient and can underrepresent or overrepresent specific lesions and prostates. The random forest in this work is trained with weighted training examples to counter these potential sources of bias. To work with weighted training examples, the decision tree information gain function was modified and is defined as

$$IG(X, f, t) = H[p(X)] - \frac{W(X_{f \leq t})}{W(X)} H[p(X_{f \leq t})] - \frac{W(X_{f > t})}{W(X)} H[p(X_{f > t})],$$

where X is a set of training examples (feature vector x , weight w , and class label y) to partition, f is the feature, t is a threshold to partition the training example set, $H(p)$ is a purity function, p is a probability vector, and $W(X)$ is a function that sums the weights of the given training examples. These are all given as

Fig. 5 Sampling strategy for the three types of annotations. The red regions correspond to positive (cancer) examples, while the green regions correspond to negative (normal) examples. The blue region in the “contours” picture represents a 3-mm band where neither positive nor negatives examples are sampled. No negative examples are sampled when working with biopsy points.

$$W(X) = \sum_{(x,y,w) \in X} w,$$

$$p_k(X) = \frac{W(X_{y=k})}{W(X)},$$

$$H(p) = \sum_{k=1}^K p_k(1 - p_k).$$

The purity function used in this random forest is the Gini index.¹⁶ This weighted gain is just a generalization of the conventional gain function used in classification and regression trees (CART) where a hard count is replaced by a sum of weights.¹⁶ If the weights are all 1, then this reduces to the CART gain function.

The weights for cancerous lesions in hand-drawn contour cases are taken to be 1/lesion volume while biopsies are taken to be $3/(500 \times \pi)$, the inverse of which corresponds to the volume of a sphere of radius 5 mm. The negative examples for hand-drawn contour cases are taken to be 1/volume of normal prostate tissue and 1/volume of prostate in the case of normal cases. The negative weights are then scaled to have the same sum as the positive weights. This ensures that lesions of any size are treated equally and that positives and negatives are treated equally.

3 Statistical Analysis

The random forest CAD and the two AutoContext methods were trained and evaluated on five permutations of twofold cross-validation on 224 patients (112 training and 112 testing for each permutation). The data were augmented by flipping the images, contours, and biopsy points about the y-axis since prostates are symmetric. This effectively doubles the training set size. The cases with contours, targeted biopsy results, and normal cases were all independently split into twofolds and then concatenated to form the twofolds over all patients. This ensures that roughly the same quantities of the three types of annotations are used in each fold. The three systems were additionally compared to the CAD system described in Ref. 6. This additional CAD system was evaluated on the same test data but not trained. Cases from the data set used in Ref. 6 are different and are not included in the data set used in this work.

Performance was measured in the form of an averaged receiver operating characteristic (ROC) curve and a Gleason grade-specific free-response ROC curve (FROC). Significance

was measured between the random forest CAD and the SVM CAD FROC curves using JAFROC2.¹⁷

3.1 Permutations of Twofold Cross-Validation

Owing to the limited number of higher Gleason grade lesions, we consider a twofold cross-validation to maximize the number of lesions of each Gleason grade in both training and testing. However, twofold cross-validation may not adequately demonstrate generalization of CAD systems. Prostate literature often considers more folds, such as 10-fold cross-validation. To compensate for these limitations, we formed five twofold cross-validation sets, all resulting from random permutations of the original data set. This helps better demonstrate generalization while leaving a large number of test lesions of various Gleason grades for analysis.

3.2 ROC Analysis

The ROC analysis was performed in three-dimensional (3-D) images and is a lesion-level analysis. The probability images were taken as a 3-D volume, and the hand-drawn two-dimensional (2-D) contours were stacked to define 3-D cancer volumes. If the 90th percentile of the CAD probability scores in a cancer volume exceeds a probability threshold, then this is taken to be a true positive. In other words, if at least 10% of the cancer volume has a relatively high probability, then the CAD is said to have detected the cancerous lesion. The CAD needs to only predict part of a lesion as suspicious to alert a radiologist, and the use of hand-drawn contours in the ROC analysis reflects this aspect. However, we stress that hand-drawn contours are not needed by users of the CAD as they simply interpret the raw probability map output from the CAD. False positives were computed on normal cases and hand-drawn contour cases by dividing the prostate into $3 \text{ mm} \times 3 \text{ mm} \times 3 \text{ mm}$ cubic cells and evaluating the 90th percentile of the CAD probability scores in each cell that do not coincide with a true positive. This coarse treatment of the background eliminates small detections that a prostate reader would effectively ignore. Additionally, a cell must be at least 3 mm away from a cancer volume to be counted as a false positive. This is meant to account for human error in the annotations.

3.3 FROC Methodology

Like the ROC analysis, the FROC analysis was also performed in 3-D. The FROC analysis is only evaluated on cases with biopsy results, which also includes the contour cases. The objective is to quantify performance with respect to Gleason grade.

While Gleason grade is normally expressed in two numbers, indicating the two most dominant patterns in the pathology, this work uses the sum of those two numbers, ranging from 6 to 10. Three FROC curves were generated to quantify CAD performance: actionable, Gleason ≥ 7 , and Gleason-specific curves. These can be seen in Figs. 6, 7, and 8.

All biopsies in this work are targeted and treated as positive even if they are benign. Every targeted biopsy is said to be “actionable” since a radiologist deemed these biopsy sites

suspicious based on mpMRI. While this treatment of benign examples appears to contradict the task of detection of clinically significant cancer, the practical purpose of this CAD is to imitate the best judgment of expert radiologists. Furthermore, treating benign targeted biopsies as negative examples proved to be confusing to the CAD and was observed to reduce detection performance of clinically significant cancers since targeted benign biopsy sites tended to have a similar appearance to prostate cancer.

The FROC is generated in three steps. First, 3-D nonmaximum suppression (NMS) is computed on the raw 3-D CAD probability map. The NMS uses a window of $10\text{ mm} \times 10\text{ mm} \times 10\text{ mm}$ and produces a small set of points and corresponding probabilities. These roughly reflect local maxima in the probability map. Next, NMS detections are paired with ground truths to count detections, false positives, and false negatives. However, there can be numerous solutions to this problem owing to the potential proximity of multiple NMS detections to multiple ground truths. A simple example of this problem is the presence of a single NMS detection that is equidistant to two ground truths. To help reduce ambiguous solutions and optimally pair NMS detections with ground truths, a bipartite graph is constructed with one group of vertices being the ground truth biopsy locations and the other group being the NMS detections. An edge is placed between a detection and a ground truth if they are 10 mm or less apart. A weighted maximum matching¹⁸ is computed on the graph to find optimal pairs of detections and ground truths. The weight for each edge is computed based on Gleason grade and probability using the following function

Fig. 6 Average FROC curve of detections of cancerous lesions Gleason 3 + 4, 4 + 3, or higher. The curve labeled “Kwak et al.” refers to the method described in Ref. 6 and was run on identical test data but not trained on any of the folds. The corresponding shaded regions are the 95% confidence interval of the detection rate over all the test folds.

Fig. 7 Average test FROC curve over five permutations of twofold cross-validation and plotted with respect to Gleason grade for random forest (forest), filter-based AutoContext (filter), and random forest-based AutoContext (AutoContext). Random forest struggles with Gleason 7 lesions and Gleason 10 while filter and AutoContext approaches struggle with Gleason 7.

Fig. 8 Three examples of images, annotations, and prostate CAD probability maps from this work and Kwak et al.⁶ The red region in the “annotation” column is the hand-drawn cancerous lesion annotation. The red regions in the “probability map” column denote higher predicted probability of cancer, while green and blue denote low predicted probability of cancer. Row 1 is an example of a 74-year-old patient with a cancerous lesion in the left apical midperipheral zone (Gleason 9). Row 2 shows a 63-year-old patient with a cancerous lesion in the right midbase anterior transition zone (Gleason 7). Last, row 3 shows a 49-year-old patient with a cancerous lesion in the left apical midanterior transition zone (Gleason 7).

$$w(g, p) = p \times 1.5^{g-6},$$

where g and p are Gleason grade and detection probability, respectively. Benign biopsies use the Gleason grade $g = 0$ in this analysis. This is intended to further resolve ambiguities and better represent CAD performance in the case where a high probability detection can be paired with a low grade or benign lesion when a high grade lesion is also nearby. This weighting function prefers that higher probability detections be paired with higher Gleason grade biopsy sites. At this stage, singleton detection vertices are counted as false positives while singleton ground truth vertices are counted as false negatives. These vertices represent detections or ground truths further than 10 mm away from any ground truth or detection, respectively. Last, a multiscale blob-similarity is computed along the axial view of the detection. The blob is measured by convolving a Gaussian kernel on a local patch on the axial view of the probability map centered at the detection. The patch is the same size as the Gaussian kernel, so this operation results in just a single value. Three scales of Gaussian kernel are considered, and the maximum convolution is then used as the final probability for the detection. Detections that look locally more like blobs thus have higher probabilities than those that look less like blobs. These remaining detections and probabilities are then used to determine the detection rate.

4 Results

The pipeline using the random forest detector in the first stage achieved an area under the curve (AUC) of 0.93 with a detection rate of 89% and a false positive rate of 20% for the ROC

analysis. This corresponds to a 92% detection rate and an average of 13 false positives per patient in the Gleason ≥ 7 FROC analysis. The image filter-based detector achieved a similar AUC of 0.93 with a detection rate of 91% at the same false positive rate on the ROC curve while achieving a corresponding detection rate of 91% and an average number of false positives per patient of 13 at the same threshold in the Gleason ≥ 7 FROC analysis. Both methods are comparable to the methods that do not use probability map features. The random forest lacking probability map features achieved an AUC of 0.93 with a detection rate 91% at the same false positive rate with a corresponding detection rate of 90% and an average number of false positives per patient of 14 in the Gleason ≥ 7 FROC analysis. The SVM approach had an AUC of 0.86 with a detection rate 81% in the ROC analysis at a false positive rate of 20% and a corresponding detection rate of 92% with an average of false positives per patient of 19 for the Gleason ≥ 7 FROC analysis. All four ROC curves are shown in Fig. 9. Figures 6, 7, and 10 show FROC curves for actionable, Gleason ≥ 7 , and fixed Gleason grade detections. Only the filter-based method is considered as it performs the best of all methods in the ROC analysis. Significance was established for the Gleason ≥ 7 FROC curves using JAFROC2 for the filter-based CAD versus the SVM method. The random forest performance was statistically significantly different in 9 out of 10 experiments.

A histogram of the frequency of features used in the CAD selected by random forest was computed by counting the frequency of features used in decisions from all trees in the random forest. The top five most selected features by random forest were the signed distance to the transition zone, the T2W mean and

Fig. 9 Average test ROC curve of three CAD systems run on five permutations of twofold cross-validation. The curve labeled “Kwak et al.” refers to the method described in Ref. 6 and was run on identical test data but not trained on any of the folds. The corresponding shaded regions are the 95% confidence interval of the detection rate over all the test folds.

Fig. 10 Average test FROC curve of three CAD systems run on five permutations of twofold cross-validation. The curve labeled “Kwak et al.” refers to the method described in Ref. 6 and was run on identical test data but not trained on any of the folds. The corresponding shaded regions are the 95% confidence interval of the detection rate over all the test folds.

median, the B2000 median, and the ADC median. The distance map feature was twice as likely to be picked as the T2W mean and is especially useful for differentiating between the transition zone and peripheral zone. Haralick features are less frequently selected, but all 39 are consistently picked by the random forest.

Figure 8 shows the various sequences, annotations, and CAD probability maps from random forest and the CAD from Ref. 6. Figures 1 and 2 show failure modes of the random forest CAD where Fig. 1 shows prominent false positives and Fig. 2 shows missed cancerous lesions. The cases shown in these figures are test cases with whole mount histopathology and were not included in the training or testing sets considered in experiments in this work.

Figure 5 shows three sampling strategies employed by the CAD during training. These are described in Sec. 2.

5 Discussion

The three random forest CAD systems performed similarly. The filter-based CAD had a slightly narrower confidence region.

It was also more consistent than the other CAD systems since its performance varied less with different training sets. The AutoContext and random forest CAD systems varied more, although AutoContext was worse than random forest with respect to stability. The performance stability of the filter-based method could be attributed to better detection of lesions in the first stage. The filter-based approach attempts to highlight relatively low- or high-intensity regions in each sequence directly without the use of annotations. Annotations can be problematic in mpMRI as not all lesions are visible in all sequences. Thus, the learning task for specific sequences is more challenging as an annotation may refer to a lesion that is not visible.

All three CADs outperform the CAD from Ref. 6 even though the CAD from Ref. 6 was trained on different mpMRI data acquired with the same exact MRI protocol and with biopsies acquired with the same exact MRI-TRUS fusion biopsy procedure. This may be due to a combination of the use of different features, different sequences (Ref. 6 does not use ADC), a different classifier, and different annotations. The work in Ref. 6 uses only biopsies for training data. The data set was selected using inclusion and exclusion criteria to meet specific criteria to reduce ambiguities. This selection could bias the resulting model to specific kinds of prostate cancer configurations. In contrast, this work uses all available training data in three different forms to improve generalizability. Consequently, this work generally outperforms the CAD of Ref. 6. There are other algorithmic differences that can account for such performance differences. Aside from using a different classifier and different features, one noteworthy difference is the lack of transition zone information in the method of Ref. 6. This feature is especially useful for learning specialized rules for transition and peripheral zones.

Due to differences in data sets and evaluation methodology, it is difficult to compare other prostate CAD systems in the literature. The work of Ref. 6 reports an AUC of 0.83 in a cancer versus MR-positive benign experiment. The work of Ref. 7 reports AUCs of 0.72 and 0.82. The first and second AUCs correspond to analysis on the raw probability map and a post-processing of the probability map where the lesion is segmented. The reported FROC shows a sensitivity of 80% detection rate with 1.5 false positive per patient. The work of Ref. 8 reports sensitivities of 0.41, 0.65, and 0.74 at 1, 3, and 5 false positives per patient, respectively.

The three CADs in this work outperform⁶ with respect to prostate cancer detection and recommending actionable biopsy sites. The performance for recommending suspicious actionable sites is worse than for detection of Gleason ≥ 7 lesions. Confusingly, Fig. 7 indicates that all CADs are slightly more performant for Gleason 6 than for Gleason 7. The relatively good performance on Gleason 6 also means that the suspicious lesions found to be benign account for the relatively poorer performance in Fig. 10.

Gleason 7 lesions are problematic for the random forest CAD systems. This can be seen in Fig. 8 where all three CAD systems perform comparable to Ref. 6. This is mainly due to the CAD missing Gleason 7 transition zone lesions. One possible explanation is that some transition zone lesions could be mistaken for fibromuscular stroma.¹⁹ Both structures share similarities, such as relatively low intensity. Another possible explanation is the sensitivity of the CAD to the geometric features. Poor transition zone segmentations may cause the CAD to miss transition zone

lesions. Transition zone distance features are also in mm units. This could be problematic for unusually large or small prostates. A possible solution to this particular problem is to normalize the distance features. The CAD systems described in this work are still at risk of overfitting cross-sequence patterns. Some false negative cases have been observed to show little or no indication of cancer in high b -value images while having prominent cancerous appearance in one or both of the T2W and ADC images. A larger training set with these kinds of images may help CAD systems generalize better. In the case of the AutoContext-based CADs, using a predetermined decision fusion rule may be more robust as these CADs are still liable to learn coincidental patterns in the training set.

The frequency at which a feature is selected signifies its relative importance. The frequent selection of distance map features demonstrates the added value of anatomical knowledge. Generally, lesions found in either one of the zones have similar intensity characteristics, but the backgrounds surrounding the two different regions are visually distinct. The low- or high-intensity appearance of cancerous lesions in T2W, ADC, and B2000 images is consistent with the frequent selection of intensity features from the image sequences. Although individual Haralick features are relatively less frequently selected, there is consistency in the overall selection pattern of all the texture features, which indicates that they are all equally relevant for prostate CAD. This is admittedly a surprising finding since Haralick texture features have been demonstrated to discriminate between cancerous and noncancerous prostate tissue as well as between cancerous tissues of differing Gleason grades.

This study is limited in a number of ways. For one, it examines mpMRI images coming from one institution and using only one kind of scanner. All images also use ERC when ERC is not used in all institutions. The use of high b -value images, such as B2000, excludes many institutions that may not possess the equipment capable of acquiring such images, although b -value images may be predicted from other available b -value images.²⁰ The use of MRI–TRUS fusion biopsies may be seen to be a limitation since many prostate CAD works use wholemount histopathology as the ground truth. However, exclusive use of patients with wholemount histopathology may bias the CAD to work well only on patients who go on to have prostatectomy. The FROC methodology described in this work can produce different results depending on the matching criteria and weight function. The matchings are also not necessarily unique. The matching criteria and weight function used in this work were chosen to help reduce ambiguous matches and to best reflect clinical intuition. The use of probability maps also poses problems for quantitative analysis, and this results in differing evaluation methodologies across the prostate CAD literature. The probability maps are intended to be interpreted by radiologists, and this consequently introduces subjectivity into the detection process. Thus, it is technically impossible to automatically characterize the performance of a CAD that produces probability maps intended for interpretation by radiologists. Only a reader study can completely characterize such a CAD system.

6 Conclusion

This work introduced three forms of prostate CAD, the first of which uses the transition zone for features, AutoContext, and three types of annotations of various qualities. It is the second such CAD that employs a strategy for sample imbalance but uses instance-level weighting for fair treatment of all cancerous

lesions and normal prostate tissue regardless of volume. The three prostate CAD systems presented here all performed similarly, with the filter-based CAD exhibiting slightly more consistent performance. All three CAD systems significantly outperform an SVM-based method on the same data set under the same evaluation methodology. This direct comparison with another algorithm presented in a published work with the same data is unprecedented in the prostate CAD literature.

Disclosures

Dr. Ronald M. Summers receives patent royalties from the iCAD Medical and license royalties from Imbio and Zebra Medical. Furthermore, Dr. Summers' lab receives research support from Ping An. Dr. Bradford J. Wood holds multiple patents related to prostate biopsy. No conflicts of interest were reported by the other authors.

Acknowledgments

We thank Andrew Dwyer, MD, for critical review of the paper. This research was funded by the Intramural Research Program of the National Institutes of Health, Clinical Center.

References

1. H. Grönberg, "Prostate cancer epidemiology," *Lancet* **361**(9360), 859–864 (2003).
2. D. N. Costa et al., "MR imaging–transrectal US fusion for targeted prostate biopsies: implications for diagnosis and clinical management," *Radiographics* **35**(3), 696–708 (2015).
3. M. Valerio et al., "Detection of clinically significant prostate cancer using magnetic resonance imaging–ultrasound fusion targeted biopsy: a systematic review," *Eur. Urol.* **68**(1), 8–19 (2015).
4. L. Breiman, "Random forests," *Mach. Learn.* **45**(1), 5–32 (2001).
5. Z. Tu and X. Bai, "Auto-context and its application to high-level vision tasks and 3D brain image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.* **32**(10), 1744–1757 (2010).
6. J. T. Kwak et al., "Automated prostate cancer detection using T2-weighted and high- b -value diffusion-weighted magnetic resonance imaging," *Med. Phys.* **42**(5), 2368–2378 (2015).
7. G. Litjens et al., "Computer-aided detection of prostate cancer in MRI," *IEEE Trans. Med. Imaging* **33**(5), 1083–1092 (2014).
8. P. Vos et al., "Automatic computer-aided detection of prostate cancer based on multiparametric magnetic resonance image analysis," *Phys. Med. Biol.* **57**(6), 1527–1542 (2012).
9. D. Fehr et al., "Automatic classification of prostate cancer Gleason scores from multiparametric magnetic resonance images," *Proc. Natl. Acad. Sci. U. S. A.* **112**(46), E6265 (2015).
10. A. Muthigi et al., "Missing the mark: prostate cancer upgrading by systematic biopsy over magnetic resonance imaging/transrectal ultrasound fusion biopsy," *J. Urol.* **197**(2), 327–334 (2017).
11. R. M. Haralick, K. Shanmugam, and I. H. Dinstein, "Textural features for image classification," *IEEE Trans. Syst. Man Cybern.* **1973**(6), 610–621.
12. A. Wibmer et al., "Haralick texture analysis of prostate MRI: utility for differentiating non-cancerous prostate from prostate cancer and differentiating prostate cancers with different Gleason scores," *Eur. Radiol.* **25**(10), 2840–2850 (2015).
13. R. M. Rangayyan, F. J. Ayres, and J. L. Desautels, "A review of computer-aided diagnosis of breast cancer: toward the detection of subtle signs," *J. Franklin Inst.* **344**(3), 312–348 (2007).
14. M. Gletsos et al., "A computer-aided diagnostic system to characterize CT focal liver lesions: design and optimization of a neural network classifier," *IEEE Trans. Inf. Technol. Biomed.* **7**(3), 153–162 (2003).
15. T. Wolters et al., "A critical analysis of the tumor volume threshold for clinically insignificant prostate cancer using a data set of a randomized screening trial," *J. Urol.* **185**(1), 121–125 (2011).
16. L. Breiman et al., *Classification and Regression Trees*, CRC Press, Boca Raton, Florida (1984).

17. D. P. Chakraborty and K. S. Berbaum, "Observer studies involving detection and localization: modeling, analysis, and validation," *Med. Phys.* **31**(8) 2313–2330 (2004).
18. J. L. Gross and J. Yellen, *Graph Theory and Its Applications*, CRC Press, Boca Raton, Florida (2005).
19. Y. X. Kitzing et al., "Benign conditions that mimic prostate carcinoma: MR imaging features with histopathologic correlation," *Radiographics* **36**(1) 162–175 (2016).
20. K. B. Grant et al., "Comparison of calculated and acquired high b value diffusion-weighted imaging in prostate cancer," *Abdom. Imaging* **40**(3), 578–586 (2015).

Biographies for the authors are not available.