PRIFYSGOL ABERYSTWYTH UNIVERSITY

Aberystwyth University

Computer aided diagnosis of prostate cancer

Rampun, Yambu Andrik; Tiddeman, Bernard; Zwiggelaar, Reyer; Malcolm, Paul

Published in:
Medical Physics

DOI:
10.1118/1.4962031

Publication date:
2016

Citation for published version (APA):
Rampun, Y. A., Tiddeman, B., Zwiggelaar, R., & Malcolm, P. (2016). Computer aided diagnosis of prostate cancer: A texton based approach. *Medical Physics*, *43*(10), [5412]. https://doi.org/10.1118/1.4962031

Document License
CC BY

General rights
Copyright and moral rights for the publications made accessible in the Aberystwyth Research Portal (the Institutional Repository) are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the Aberystwyth Research Portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the Aberystwyth Research Portal

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

tel: +44 1970 62 2400
email: is@aber.ac.uk

Download date: 09. Jul. 2020

Computer Aided Diagnosis of Prostate Cancer: A Texton Based Approach

Andrik Rampun, Bernie Tiddeman and Reyer Zwiggelaar [a)]

Department of Computer Science, Aberystwyth University, SY23 3DB, UK

Paul Malcolm

Department of Radiology, Norfolk & Norwich University Hospital, Norwich NR4 7UY, UK

(Dated: 19 September 2016)

Purpose: In this paper we propose a texton based prostate Computer Aided Diagnosis approach which bypasses the typical feature extraction process such as filtering and convolution which can be computationally expensive. The study focuses the peripheral zone because 75% of prostate cancers start within this region and the majority of prostate cancers arising within this region are more aggressive than those arising in the transitional zone.

Methods: For the model development, square patches were extracted at random locations from malignant and benign regions. Subsequently, extracted patches were aggregated and clustered using k-means clustering to generate textons that represent both regions. All textons together form a texton dictionary, which was used to construct a texton map for every peripheral zone in the training images. Based on the texton map, histogram models for each malignant and benign tissue sample were constructed and used as a feature vector to train our classifiers. In the testing phase, four machine learning algorithms were employed to classify each unknown sample tissue based on its corresponding feature vector.

Results: The proposed method was tested on 418 T2-W MR images taken from 45 patients.Evaluation results show that the best three classifiers were Bayesian Network ($A_z = 92.8\% \pm 5.9\%$), Random Forest ($89.5\% \pm 7.1\%$) and k-NN ($86.9\% \pm 7.5\%$). These results are comparable to the state-of-the-art in the literature.

Conclusions: We have developed a prostate Computer Aided Diagnosis method based on textons using a single modality of T2-W MRI without the need for the typical feature extraction methods, such as filtering and convolution. The proposed method could form a solid basis for a multimodality MRI based systems.

Keywords: Computer Aided Detection, Prostate MRI and Texton

1. INTRODUCTION

Prostate cancer is one of the most common cancers affecting men, with an estimated 1.1 million diagnoses and 307,000 deaths in the world in 2012 [1]. In the same year, there were 41,736 cases reported in the UK with 10,837 deaths. Prostate cancer rates in the UK have at least tripled over the last 35 years, causing prostate cancer to be the most common cancer among British men. In 2015, there were an estimated 220,800 incidences and around 27,500 deaths, making it the second most deadly cancer in the United States [2].

The most common methods used for preliminary screening are the prostate-specific antigen (PSA) test, digital rectal examination (DRE) and transrectal ultrasound (TRUS) guided biopsy. However, these methods have limited sensitivity and specificity. For example, an elevated PSA level does not always indicate the occurrence of prostate cancer because several factors can increase PSA levels such as a urine infection, vigorous exercise and ejaculation in the 48 hours before a PSA test [4]. The DRE test is highly dependent on the experience of the examiner. For example, a more experienced examiner can detect subtle abnormalities in comparison to less experienced clinicians [4,8]. For TRUS guided biopsy, due to its random procedure, the sample needle can miss cancerous and significant tissues, meaning the test result can indicate incorrect results [57].

Since there is still space for improvement in the reliability of clinical methods for screening and detecting prostate cancer, integrating Magnetic Resonance Imaging (MRI) into clinical practices (e.g. MRI/Ultrasound guided biopsy and multimodality image fusion) is becoming popular as it has shown a significant improvement over PSA and TRUS alone [3,4]. Unfortunately such methods require substantial expertise from the radiologist. Previous studies have shown a high degree of inter-observer variability [3,4], indicating a high risk of

human error. They are also time consuming. Computer Aided Diagnosis (CAD) can assist radiologists in the interpretation of medical images by providing a 'second' opinion, which eliminates variability among radiologists, speeds up the analysis of the images and improves diagnosis decision results [5,6].

Developing CAD systems is a difficult task due to variations in the appearance of anatomy in images produced by MRI scanners. Recent studies [4,7–9,42] have reported the deficiencies of T2-weighted (T2-W) MRI including weak texture descriptors that could be affected by noise. In fact, Tiwari *et al.*[10,41] suggested that T2-W MRI texture features alone are insufficient to identify prostate malignancies. Recently, a popular approach to improve the performance of CAD systems is using multiparametric MRI. The use of multimodality MRI in developing CAD systems is a popular way to improve the performances of existing methods. It is noted that using T2-W alone is deemed insufficient, but that T2-W classification can form a solid basis for a multimodality MRI based CAD system.

In 2015, Lemaître *et al.*[11] conducted a review of CAD systems for prostate cancer detection and reported that there are 42 studies in the literature from 2007 until 2014. Most of the methods described used typical feature extraction algorithms based on filtering and convolution, which can be computationally expensive. The large number of extracted features also led to the need for the additional step of feature selection or dimensionality reduction. None of the CADs reviewed in [11] have used textons to discriminate benign and malignant tissue in their studies. Although the term texton was first introduced in the 80's, it did not get much attention until a study of texture classification by Leung and Malik [12] in 2001. Later, similar studies showing promising results in texture classification were conducted by Varma and Zisserman [15,16] in 2005 and 2009, respectively. In medical image analysis textons have been used in retinal vessel segmentation [13] and lung cancer detection [14].

Textons can be seen as a representation of micro-structures in natural images and are considered as the atoms of pre-attentive human visual perception [17]. In the original approach, textons were represented by means of a collection of filter bank responses obtained from large filter banks such as the MR8 [18], LM [12], S [19] filter banks and Gabor filter [20]. All the response vectors were collected and clustered using k-means and the resulting cluster centers were called textons (hence, in a simplest definition textons are the k-means' cluster centers). Nevertheless, the study in [21] showed textons can be generated by directly clustering the image's pixel values from patches without the need for filter banks (hence, speeding up the process of constructing the texton dictionary).

In [15,16], the authors made quantitative comparisons between a typical texton-based approach (using a filter bank) and a texton-based approach without filter bank, which showed that the latter appraoch produced better classification results. The study in [21] suggested there are three reasons for the relatively strong performance of textons generated from the image's pixels in comparison to textons generated from the convolved image's pixels. First, the use of filter banks reduces the number of textons that can be extracted from a texture image [22]. For example, the number of textons are significantly reduced when an image of 250×250 pixel is convolved with a 50×50 filter. This affects the ability of the histogram models to characterise a particular texture (i.e. insufficient information to model the actual representation of the image). Second, the large number of filters leads to small errors in the edge localisations which may significantly change the geometry of the textons, leading to errors in the estimation of the texton frequency histogram [22]. Finally, most filter banks lead to some blurring of the texture which might remove local details in the texture hence resulting in different textons [22]. Based on these reasons the proposed method in this paper did not use any filter banks but took the image's pixels directly to generate textons.

The aim of this paper is to investigate the use of textons without the need for filtering or convolution for feature extraction in prostate cancer CAD system in a single modality T2-W MRI. The novel contributions of our work are:

1. This is the first CAD method which has investigated the use of textons in classifying benign and malignant tissues within the prostate gland in MRI.

2. The proposed method learns directly from image pixels without the need to use a filter bank. In comparison, most prostate cancer CAD systems in the literature compute large numbers of texture descriptors, which are computationally expensive. In fact, computing a large number of texture descriptors also leads to an additional (essential) step of feature selection or dimensionality reduction.

The clinical motivations of our work are three-fold:

1. Finding cancer regions in each MRI image manually by a radiologist is time consuming. A CAD system can potentially speed up this process by delineating cancerous regions automatically.

2. Computer algorithms are deterministic while radiologists' results can be variable. For example, fatigue affects radiologists' performances, potentially resulting in missing cancerous regions.

3. The accuracy of detecting prostate cancer among radiologists varies depending on the level of experience. In comparison, a CAD system can eliminate this issue as it can provide consistent results.

4. Prostate cancer CAD acts as a second opinion which can significantly improve the performance of less experienced radiologists.

2. METHODOLOGY

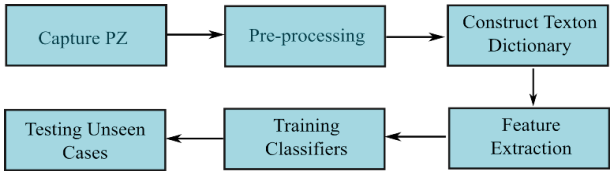

FIG. 1. A general overview of the proposed method.

Figure 1 shows a flowchart of our proposed method. For every input image, we roughly estimate the area of the prostate's peripheral zone (PZ) followed by normalisation and noise reduction. Subsequently, for every training image we randomly extract patches from benign and malignant regions within the PZ and employed k-means clustering to generate textons (the output of this stage in the texton dictionary). Each pixel in every training image is labelled with the texton to which it lies closest, producing texton maps. Using the texton maps, a histogram of textons (the frequency with which each texton occurs) is constructed for every pixel within the PZ. The texton histograms from all pixels are treated as feature vectors and used to train our classifiers. Finally, at the testing phase, every unseen PZ is processed in the same way and the trained classifiers are used to decide, for each pixel, whether it belongs to the benign or malignant class. The next subsections will explain this process in more details.

2.A. Capturing the Peripheral Zone

We employed the 2D model developed by Rampun *et al.*[25] to estimate the area of the PZ. The method uses a quadratic equation based on the central coordinates of the prostate gland, the left-most and right-most coordinates of the prostate gland boundary (each prostate boundary was provided by a radiologist). This allows us to model a *priori* general knowl-

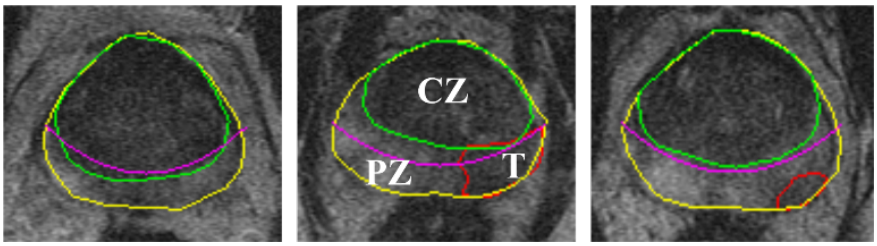

FIG. 2. Example images of prostate MRI with ground truth delineated by an expert radiologist and the estimated PZ region under the magenta line. PZ (peripheral zone), CZ (central zone) and T (Tumour).

edge of radiologists which is similar to the methods of Makni *et al.*[43] and Liu *et al.*[25]. Figure 2 shows example MRI images with the ground truth location of the prostate gland, transitional/central zone (CZ) and tumor (T) represented in red, yellow and green respectively, while the magenta line is the estimated boundary of the PZ based on the method given in [25]. Note that our study is only developed within the segmented PZ which is under the magenta line in Figure 2. Our clinical justifications for currently focusing on the peripheral zone (PZ) is [8,47,48]:

1. More than 75% of prostate cancers are located within this region.

2. The majority of prostate cancers arising within this region are more aggressive than those arising in the transitional zone.

3. Most prostate cancers start to develop in this zone before spreading to the transitional zone.

2.B. Pre-processing

The major problem with MR images is that specific tissues do not have fixed specific intensity values. This is mainly caused by [8,26–28]: a) corruption by thermal noise due to receiver coils, b) different scanning protocols causing large intensity variations and c) poor radio frequency coil uniformity. These can significantly affect the discriminative performance hence need to be corrected [28]. Following the pre-processing procedure method described in [8,29,30], each image is median filtered to preserve edge boundaries. Subsequently, image intensities were normalised to zero mean unit variance and anisotropic diffusion filtering [30,31] was applied to remove noise.

We used the anisotropic diffusion and median filter to eliminate low-level noise and sharp noise (e.g. bright spikes), respectively. The anisotropic diffusion is a robust filter, however studies [8,30] have shown that it is ineffective for eliminating sharp noise. The idea behind anisotropic diffusion is to use a diffusion function to prevent smoothing happening across edges, and therefore it preserves edges in the images. Unfortunately, the gradient to the noise element (e.g. sharp noise) may compete with edge responses and the diffusion function cannot distinguish between image structure and noise contribution. By combining both filters, they work in a complementary way to gradually eliminate the overall noise element without blurring the edges and textures of the image. The filters we use are both of size 3×3 for the entire study.

2.C. Texton Dictionary

Figure 3 shows the summary of steps to construct the texton dictionary. Textons (i.e. $m \times n$ window square, where m and n are rows and columns, respectively) were retrieved

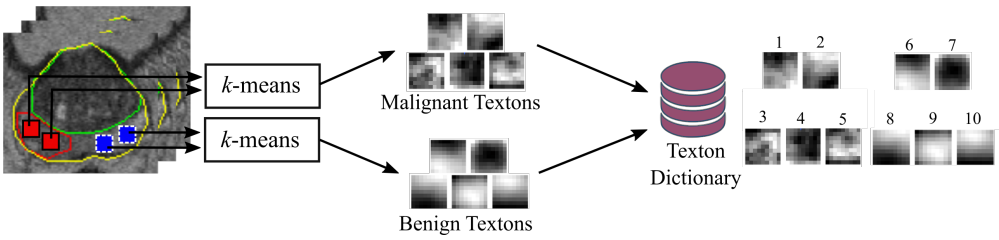

FIG. 3. Generating the texton dictionary. Patches from the same class are aggregated and clustered using the k-means algorithm. Red (black solid line) and blue (white dotted line) patches are malignant and benign samples, respectively.

from benign and malignant regions. To construct the texton dictionary, we followed the work of Varma and Zisserman [15,16,18,21]. For every PZ area in the training images we randomly extracted $m \times n$ patches of raw pixels from benign and malignant regions. Subsequently, all patches extracted from benign regions were clustered using the k-means algorithm. The same process was performed for all patches extracted from malignant regions. A summary of the k-means algorithm can be found in[32]. The cluster centroids produced by the k-means algorithm are the textons. Once all textons from both classes (benign and malignant) were generated, they were combined to form the texton dictionary. As shown in Figure 3, each texton is unique and has its own id ($TX = tx_1, tx_2, tx_3.....tx_n$) saved in a matrix which will be used to construct the texton map for each image.

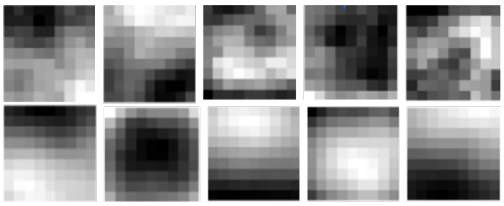

FIG. 4. Example of textons generated from malignant (top row) and benign (bottom row) regions.

Figure 4 shows textons extracted from both classes. Visually it can been seen that textons generated from benign regions (bottom row) look smoother than the ones retrieved

from malignant regions (top row).

2.D. Feature Extraction

In this study, texton histograms are used to model benign and malignant tissue. These histograms are calculated in two main stages. The first stage is to generate the texton map, where every pixel in the image (within the PZ) is assigned to the closest texton (using the Euclidean distance) in the texton dictionary. We used a sliding window, W^T of the same size as the textons, and found the texton, TX, with the shortest Euclidean distance to the image patch W^T. Subsequently, the central pixel in W^{T} is assigned the texton id which is the closest to W^{T}. This process was repeated for every pixel in the image until all pixels were assigned with the corresponding textons 'ids'. By the end of this stage, a texton map was constructed for every PZ which was used in the subsequent stages.

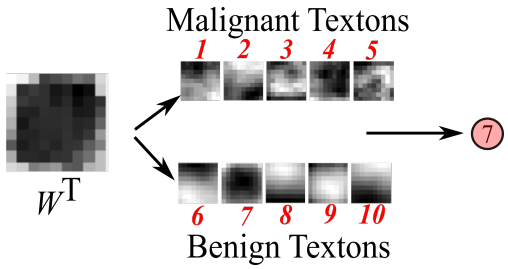

FIG. 5. A graphical illustration on how to construct a texton value of an image patch.

Figure 6 shows examples of texton maps of three PZs generated in this phase. Each pixel within the PZ was replaced with the corresponding texton 'id'. At the second stage, using the texton map we were able to generate a histogram model for each pixel by using a sliding window of the same size. A histogram for each pixel was constructed based on the occurrence of each texton's frequency within the neighborhood of the central pixel (including the central pixel).

FIG. 6. Examples of texton maps of three PZs taken from three different prostates.

Figure 7 shows an example of constructing a histogram for each pixel. In this example, there are 10 textons (5 textons for each class) in the dictionary and each histogram of a tissue was constructed based on 9×9 window size; this means the histogram was constructed based on the texton frequency within 81 pixels (and 25 pixels for 5×5 window size). Note that every histogram was normalised to unity. This yielded histograms for each tissue in the training images, which are used as feature vectors representing every pixel. To this end, each pixel is represented in a *txt* dimension feature space where *txt* is the number of textons in the dictionary (10 in this example). Similarly, if there are 30 textons (15 textons per class), each pixel is represented in a 30 dimensional feature space. It should be noted that the data dimensional is independent of W^T. Finally, the constructed histogram for each pixel was treated as feature vector in the training and testing phases.

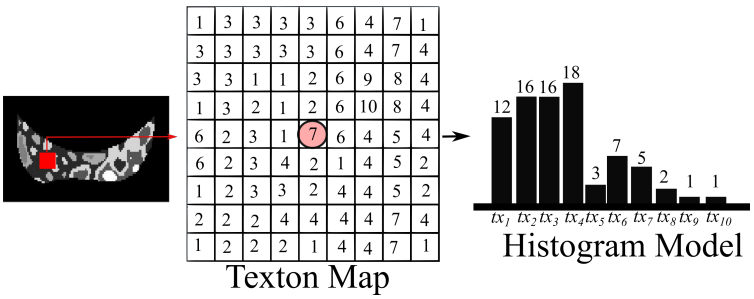

FIG. 7. Constructing a histogram for each pixel in from the texton map.

3. EXPERIMENTAL SETTINGS

3.A. Materials and Dataset

Our dataset consists of 418 T2-W MR images (227 malignant and 191 normal slices resulting in 74,208 malignant pixels and 97,310 normal pixels) taken from 45 patients aged 54 to 74 (all patients had biopsy-proven prostate cancer). Each patient has between 6 to 13 slices covering the top to the bottom of the prostate gland. The prostate gland, malignant and transitional zone were delineated by an expert radiologist with more than 10 years experience in prostate MRI. All sequences with prostate cancer cases were confirmed malignancies based on TRUS biopsy reports. All annotated cases of malignant cancer were confirmed as clinically significant (Gleason score grade 7 and above). Vollmer [55] who conducted a study (526 patients) about the relationship between tumour length and Gleason score found that tumour length was positively and significantly related to Gleason score. The study found that the median tumour length for Gleason scores 4 through 6 was 2.9mm, the median length for Gleason scores 7 through 10 was 13.1mm. These results are similar to the study conducted by Lee *et al.* [56] who found most cancer with Gleason score ≥ 7 have longer cancer core length (based on more than 5000 patients). In another study conducted by Billis *et al.* [54] covering 401 patients with Gleason score ≤ 6 found that the median core cancer length was 1.5mm with the range of 0.5mm-3.0mm. All patients underwent T2-W MR imaging at the Department of Radiology Norfolk and Norwich University Hospital, Norwich, UK. MR acquisitions were performed prior to radical prostatectomy. All images were obtained on a 1.5 Tesla magnet (Sigma, GE Medical Systems, Milwaukee, USA) using a phased array pelvic coil, with a 24×24 cm field of view, 512×512 matrix, 3mm slice thickness, 3.5mm inter-slice gap and 0.47mm pixel spacing.

Approximately 60% (pre-processing, PZ modelling, construction of texton dictionary, feature extraction and segmentation) of the source code was written in Matlab 2012a and 40% (training and testing in machine learning) was written in Java. All experiments were run under the Window 7 operating system with an Intel core i5 processor.

3.B. Training and Testing

All pixels within the radiologist's tumor annotation were extracted as prostate cancer samples. This area was delimited by the tumor mask, to ensure no pixels outside the tumor region were included in the malignant samples. All pixels outside the tumor region and within the PZ were considered benign samples. Similarly, this region was delimited by the tumor or prostate gland masks to ensure no pixels within the tumor region and outside the prostate gland were included as benign samples.

A stratified nine runs 9-fold cross-validation (9-FCV) scheme was employed. The folds were populated on a patient basis to ensure no samples from the same patient were used in the training and testing phases (45 patients in our case, hence each fold contains 5 patients). Each classifier was trained using the histograms of textons from the training partition for that fold. In the testing phase each unseen instance/pixel from the testing data (taken from 5 randomly selected patients) was classified as malignant or benign. During the cross-validation we set aside a small set of patients (in our case 5 patients) as the validation set and trained the classifiers on the remaining 40 patients. This process was randomised and the results presented in our paper are the average of the 9-FCV.

For classification, we employed four machine learning algorithms in WEKA [44] namely the Bayesian Network (BNet) [33], Alternating Decision Tree (ADTree) [34], Random Forest (RF) [35], Linear Discriminant Analysis (LDA) [50] and k-Nearest Neighbours (k-NN) with the

following default settings; Hill climbing search algorithm, number of boosting iterations is 10, number of initial random forests are 100, the ridge parameter is $1e^{-6}$ and k=1, respectively.

4. EXPERIMENTAL RESULTS

The performances were measured using the most popular metrics in the literature: Area Under the Curve (A_z, also known as AUC) and Classification Accuracy (CA). The A_z indicates the trade-off between the true positive rate against the false positive rate, where CA represents the number of pixel classified correctly. The CA can be calculated as $\frac{TP+TN}{TN+TP+FP+FN}$. TP and FP denote the number of true positives and false positives, respectively. Similarly, TN and FN indicate the numbers of true negatives and false negatives. Note that the A_z values in this paper are presented as a percentage $(0-100)$ instead of the automatic normalised range (0-1). The standard deviation measures the dispersion of both metrics across 9-fold cross validation. The t-test statistics was used to compare the best results (both metrics) between the best classifier with the other classifiers. This determines whether the best classifier produced significantly better results in comparison to the other classifiers.

One of the main challenges in developing a texton based approach in texture classification is finding the best window size (ws) for W^T and the number of textons (txt) as these parameters can influence the classification results. For this purpose, we used the following $ws = 5\times5, 7\times7, 9\times9, 11\times11$ and 13×13. For txt, we tested the following values $6, 10, 12, 16, 20, 24$ and 30. Note that these numbers represent the number of textons for both classes (e.g. $txt = 6$, 3 benign textons and 3 malignant textons).

Tables I and II show A_z and CA results, respectively using the ADTree classifier which produced best results of A_z=83.2% and CA=75.3%. At the smaller ws, the classifiers

produced A_z ranging from 74% to 77%. The results are slightly better at $ws = 7 \times 7$ with A_z ranging from 75% to 79%. It produced the best A_z at 9×9 with average A_z>80% regardless of the number of textons. However, in terms of accuracy the ADTree classifiers produced better results on average at 11×11 (CA>70%).

TABLE I. A_z (%) values for ADTree classifier.

	5 × 5	**7 × 7**	**9 × 9**	**11 × 11**	**13 × 13**
6	77.3 ± 12.1	79.4 ± 12.4	**83.8 ± 10.8**	83.2 ± 10.6	82.3 ± 10.7
10	76.5 ± 11.9	77.6 ± 13.0	82.9 ± 9.8	80.7 ± 10.7	80.2 ± 9.9
12	75.9 ± 11.9	77.3 ± 12.2	81.9 ± 9.7	80.2 ± 11.5	80.0 ± 10.7
16	76.2 ± 11.7	77.0 ± 12.0	82.1 ± 10.7	82.5 ± 9.6	81.3 ± 10.2
20	75.1 ± 11.3	76.1 ± 10.1	81.6 ± 10.2	76.5 ± 12.9	76.3 ± 11.9
24	75.6 ± 11.9	75.6 ± 11.8	80.9 ± 9.9	76.6 ± 12.1	75.9 ± 11.5
30	74.3 ± 12.0	75.1 ± 12.1	80.7 ± 10.9	76.3 ± 12.6	75.5 ± 11.7

TABLE II. CA (%) values for ADTree classifier.

	5 × 5	**7 × 7**	**9 × 9**	**11 × 11**	**13 × 13**
6	71.0 ± 13.3	70.3 ± 14.6	71.9 ± 15.1	72.7 ± 14.6	70.9 ± 15.1
10	69.5 ± 15.6	68.2 ± 14.4	69.7 ± 16.7	70.7 ± 16.3	70.1 ± 15.5
12	70.8 ± 15.4	67.4 ± 14.9	69.3 ± 12.8	70.5 ± 16.3	70.3 ± 16.1
16	68.6 ± 16.1	66.9 ± 15.4	69.3 ± 10.7	**75.3 ± 13.6**	74.3 ± 12.9
20	68.3 ± 16.2	68.5 ± 14.5	69.2 ± 14.9	69.8 ± 16.2	68.3 ± 15.3
24	68.1 ± 17.5	65.8 ± 16.3	70.0 ± 14.9	68.4 ± 16.8	67.1 ± 16.1
30	63.9 ± 21.1	65.3 ± 15.1	69.1 ± 15.1	69.7 ± 16.8	68.1 ± 15.3

The results for the BNet classifier can be found in Tables III and IV. The BNet classifier performed best when features were extracted using a larger window (e.g. 11×11). At $ws = 11 \times 11$ and $txt = 16$ the BNet outperformed the other classifiers with A_z=92.8% and CA=84%. At $ws = 13 \times 13$ with the same value of txt the BNet classifier decreased 1.5% and on average. In fact, it can also be seen that all A_z and CA went down regardless of the

number of textons. One noticeable pattern can be seen from the results in Tables III and IV is at $ws = 5 \times 5$ until 11×11 the range of A_z value increased around 4% to 5% at each window size and gradually decreased at 13×13. In terms of the number of textons, it did not affect the performance much on both metrics (variation between 1%-3%).

TABLE III. A_z (%) values for BNet classifier.

	5 × 5	**7 × 7**	**9 × 9**	**11 × 11**	**13 × 13**
6	79.2 ± 11.5	83.7 ± 10.7	89.5 ± 7.9	90.8 ± 7.8	89.4 ± 7.7
10	80.1 ± 11.4	84.3 ± 10.2	90.3 ± 6.7	92.0 ± 6.9	91.3 ± 7.5
12	80.2 ± 11.0	84.8 ± 10.1	90.0 ± 7.1	91.6 ± 6.1	90.7 ± 6.9
16	81.7 ± 10.6	85.0 ± 9.9	90.9 ± 7.1	**92.8 ± 5.9**	91.3 ± 5.7
20	80.5 ± 11.2	85.3 ± 9.7	90.8 ± 7.0	91.4 ± 6.0	90.7 ± 5.9
24	80.9 ± 11.5	85.4 ± 9.6	90.8 ± 6.8	91.5 ± 6.3	90.1 ± 6.1
30	81.0 ± 10.9	84.2 ± 8.9	90.9 ± 7.3	91.8 ± 5.9	91.1 ± 6.0

TABLE IV. CA (%) values for BNet classifier.

	5 × 5	**7 × 7**	**9 × 9**	**11 × 11**	**13 × 13**
6	70.0 ± 15.2	73.6 ± 13.2	78.6 ± 10.2	80.5 ± 10.8	78.1 ± 10.2
10	70.4 ± 14.1	74.0 ± 12.4	80.1 ± 9.9	83.0 ± 8.5	82.3 ± 9.5
12	70.4 ± 14.5	74.5 ± 11.7	80.3 ± 10.1	82.1 ± 10.1	81.7 ± 10.2
16	72.3 ± 14.2	75.1 ± 11.8	81.4 ± 8.9	**84.0 ± 7.0**	82.8 ± 7.5
20	70.2 ± 14.8	75.7 ± 12.6	81.9 ± 8.0	82.0 ± 8.9	81.5 ± 8.3
24	71.6 ± 14.6	75.6 ± 12.1	81.4 ± 8.9	83.0 ± 8.3	82.3 ± 8.9
30	70.8 ± 14.3	75.3 ± 11.9	81.8 ± 8.2	82.8 ± 8.7	81.9 ± 7.9

The k-NN classifier performed well with A_z=86.9% and CA=80.2% as shown in Table V and VI, respectively. In terms of the area under the curve, the classifier performed better with a smaller number of textons regardless of the ws. This can be seen in Table V where most A_z values are above 80% at $txt = 6$, but as the txt value increases the A_z value decreases to around 70%. The lowest accuracy was produced at the largest window of 13×13 with

30 textons in the texton dictionary.

TABLE V. A_z (%) values for k-NN classifier.

	5 × 5	**7 × 7**	**9 × 9**	**11 × 11**	**13 × 13**
6	80.2 ± 11.3	82.7 ± 9.9	**86.9 ± 7.5**	85.2 ± 7.1	85.6 ± 7.4
10	80.0 ± 10.1	80.8 ± 9.1	83.5 ± 6.8	80.1 ± 7.9	79.1 ± 7.9
12	79.4 ± 9.6	78.3 ± 8.7	80.4 ± 7.5	79.0 ± 8.5	78.8 ± 8.3
16	77.0 ± 9.4	74.8 ± 8.4	77.3 ± 7.4	77.7 ± 8.2	76.6 ± 8.1
20	73.6 ± 8.6	72.2 ± 8.4	75.5 ± 7.4	72.8 ± 8.4	74.0 ± 8.9
24	71.7 ± 8.3	70.1 ± 7.8	74.5 ± 7.9	71.6 ± 9.2	69.9 ± 9.5
30	69.8 ± 8.1	69.8 ± 7.3	73.1 ± 7.4	70.7 ± 9.1	69.8 ± 9.3

TABLE VI. CA (%) values for k-NN classifier.

	5 × 5	**7 × 7**	**9 × 9**	**11 × 11**	**13 × 13**
6	74.8 ± 11.3	74.3 ± 11.2	78.1 ± 9.6	**80.2 ± 9.1**	78.9 ± 9.3
10	74.7 ± 11.7	74.2 ± 9.1	75.9 ± 8.5	74.3 ± 9.9	76.3 ± 9.4
12	75.5 ± 10.8	73.2 ± 8.7	74.3 ± 8.5	73.6 ± 10.1	72.1 ± 10.3
16	73.4 ± 10.3	71.1 ± 8.5	73.1 ± 8.3	74.3 ± 9.6	73.2 ± 9.8
20	72.2 ± 9.5	69.6 ± 8.2	71.8 ± 8.0	70.0 ± 9.6	69.3 ± 9.7
24	71.1 ± 9.6	68.7 ± 8.4	71.4 ± 8.4	68.9 ± 10.1	68.5 ± 10.4
30	69.9 ± 9.3	67.8 ± 8.5	70.6 ± 8.2	68.5 ± 10.5	67.7 ± 10.6

Tables VII and VIII show the results of the second best classifier in our experiments which is the RF. In terms of A_z, the RF performed best at larger ws (e.g. 9×9 and 11×11) with $txt = 6$ or 10. In our experiment, using the maximum number of textons ($txt = 30$) decreased the A_z from 89.5% to 81.6%, which is statistically significant ($p < 0.001$). On the other hand, both metrics are highest at $ws = 11 \times 11$ and lowest at $ws = 5 \times 5$.

Tables IX and X show the results of the proposed method using the LDA classifier. The best A_z value was achieved at $ws = 9 \times 9$ which is 80.7%. At the same ws, the proposed method achieved on average 79.9% and higher A_z was achieved $txt > 12$. In terms of

TABLE VII. A_z (%) values for RF classifier.

	5 × 5	**7 × 7**	**9 × 9**	**11 × 11**	**13 × 13**
6	80.3 ± 11.3	83.6 ± 9.8	88.2 ± 7.4	**89.5 ± 7.1**	86.2 ± 8.1
10	80.6 ± 9.9	83.0 ± 8.8	87.2 ± 6.6	86.7 ± 7.8	85.5 ± 8.2
12	80.4 ± 9.6	81.9 ± 8.7	85.7 ± 6.7	86.0 ± 8.0	82.7 ± 7.9
16	79.5 ± 9.3	80.4 ± 8.7	85.1 ± 7.5	86.3 ± 8.1	83.6 ± 8.2
20	77.3 ± 9.1	79.3 ± 8.9	84.4 ± 7.4	82.6 ± 8.9	81.1 ± 8.2
24	76.9 ± 8.9	78.8 ± 9.1	83.9 ± 7.6	82.0 ± 9.4	81.3 ± 9.5
30	75.9 ± 9.1	78.3 ± 9.9	83.4 ± 8.4	81.6 ± 9.7	80.9 ± 8.9

TABLE VIII. CA (%) values for RF classifier.

	5 × 5	**7 × 7**	**9 × 9**	**11 × 11**	**13 × 13**
6	74.9 ± 11.3	74.6 ± 10.9	78.4 ± 9.6	**81.1 ± 9.3**	77.8 ± 9.8
10	74.8 ±11.7	75.2 ± 8.9	77.7 ± 8.7	77.5 ± 9.9	75.3 ± 9.5
12	75.9 ± 10.6	74.9 ± 8.9	76.9 ± 8.7	77.8 ± 9.0	74.3 ± 8.9
16	74.5 ± 10.1	73.9 ± 8.8	76.3 ± 8.7	77.0 ± 10.0	75.1 ± 9.5
20	74.0 ± 9.4	73.4 ± 8.6	75.3 ± 8.2	74.0 ± 9.7	73.0 ± 9.3
24	73.6 ± 9.7	72.7 ± 9.1	75.2 ± 8.4	73.0 ± 10.2	74.0 ± 10.3
30	73.1 ± 9.7	72.5 ± 9.3	74.3 ± 8.6	72.1 ± 11.0	72.2 ± 10.8

accuracy, the highest $CA = 75.8\%$ was achieved at $ws = 11 \times 11$. Our results are within the range of the existing studies [4,42,51,52] in the literature.

Overall, the BNet classifier with A_z=92.8% and CA=84% outperformed the other classifiers in both metrics, which is statistically significant (p<0.001) against all classifiers except the CA of RF classifier (p=0.095) and k-NN classifier (p=0.025). The p value between the best A_z and CA of the BNet classifier against the best results of ADTree is p<0.001. The p value against the best A_z of k-NN is p<0.001. The RF classifier produced the second best results in both metrics with A_z=89.5% and CA=81.1% at $ws = 11 \times 11$ and $txt = 6$. The k-NN and ADTree classifiers achieved A_z>83% but in terms of accuracy the ADTree

TABLE IX. A_z (%) values for LDA classifier.

	5 × 5	**7 × 7**	**9 × 9**	**11 × 11**	**13 × 13**
6	74.6 ± 13.9	75.0 ± 15.4	79.1 ± 12.1	78.2 ± 13.8	77.3 ± 12.5
10	75.2 ± 12.9	74.9 ± 13.7	79.5 ± 11.3	79.2 ± 11.9	78.2 ± 12.3
12	74.9 ± 12.3	75.8 ± 13.3	79.7 ± 11.2	78.6 ± 12.4	77.5 ± 11.9
16	75.9 ± 12.6	75.9 ± 13.1	80.2 ± 11.7	79.1 ± 12.8	78.5 ± 12.5
20	75.2 ± 12.2	75.7 ± 12.7	80.3 ± 11.1	74.6 ± 14.6	73.6 ± 13.2
24	75.5 ± 12.7	75.5 ± 12.7	80.1 ± 11.5	75.8 ± 14.2	73.2 ± 13.7
30	75.3 ± 11.9	75.1 ± 12.6	80.7 ± 11.7	75 .4± 13.1	73.2 ± 13.5

TABLE X. CA (%) values for LDA classifier.

	5 × 5	**7 × 7**	**9 × 9**	**11 × 11**	**13 × 13**
6	65.2 ± 17.6	64.5 ± 18.3	67.7 ± 16.8	67.8 ± 16.9	67.2 ± 16.5
10	65.7 ± 16.3	65.9 ± 18.1	67.9 ± 16.5	69.7 ± 18.1	68.7 ± 15.7
12	64.7 ± 16.1	67.1 ± 15.8	67.6 ± 14.8	70.2 ± 16.6	69.5 ± 16.1
16	67.4 ± 15.9	66.0 ± 16.6	68.7 ± 15.7	75.1 ± 13.2	69.1 ± 13.5
20	64.2 ± 16.6	66.3 ± 16.5	68.8 ± 15.3	68.5 ± 17.1	67.3 ± 17.3
24	66.7 ± 16.2	66.6 ± 16.2	69.4 ± 14.3	75.8 ± 14.2	67.5 ± 15.8
30	66.0 ± 15.7	66.1 ± 16.3	70.1 ± 14.6	68.4 ± 15.3	67.1 ± 15.3

achieved CA<80%. The overall results show that the best results for all classifiers employed were achieved using either $ws = 9 \times 9$ or 11×11 and $txt = 6$ or 16. Considering the best A_z values of all classifiers, results suggest that the proposed method can achieve similar performances to other prostate cancer CAD in the literature. Furthermore, based on the best A_z, our method qualitatively outperformed most of the existing methods. Nevertheless, in terms of accuracy there is space for improvement.

The RF classifier (Tables VII and VIII) perform better than the ADTree and k-NN because of its ability to perform like an ensemble classifier (consider various decisions and use averaging to improve predictive accuracy). The k-NN classifier produced better results

than the ADTree at $txt = 6$ because of the simplicity of its decision rule as well as the data itself (low data dimension). Nevertheless, the ADTree classifier performed better than the k-NN at $txt > 10$ because it employs a 'boosting' approach (building a decision tree iteratively based on the error produced in the previous decision tree). A larger number of textons would be beneficial for the ADTree classifier because a better representation of the problem domain can be created iteratively. Results in Table III and IV show that the BNet classifier produced consistent results regardless of txt and outperformed the other classifiers. The BNet is expected to perform better due to its ability to map the relationships among variables (or features) to build a predictive model without being restricted by the independence condition. Finally, the LDA produced the worst results in our experiments. Our explanations for this behaviour are two-fold. Firstly, the LDA is a linear classifier which means that the data would be linearly separable. Our test/training data are much more complex and the decision boundary between classes is expected to be non-linear. Secondly, the decision rules in the LDA classifier are incapable of dealing with complex data such as prostate MRI. Existing studies [4,42,51,52] whose methods employed the LDA classifier achieved similar results ranging between $A_z = 75\% - 84\%$ whereas our proposed method achieved $A_z = 80.7\%$ which is similar to the current methods.

The top and bottom rows in Figure 8 shows the segmentation results of two different cases produced by the classifiers employed in this study. The BNet, RF and k-NN classifiers produced good segmentation accuracy covering most area of the malignant region. In the second case (bottom row) the RF classifier again produced the highest accuracy followed by the k-NN classifier whereas the BNet and ADTree classifiers generated reasonable segmentation results.

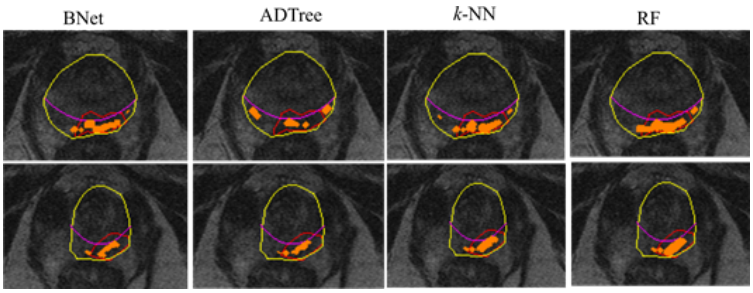

FIG. 8. Segmentation results using different machine learning algorithms.

5. PARAMETER OPTIMISATION

Since the performance of most machine learning algorithms are dependent on the parameters chosen by the users, we further investigated the performance of three classifiers employed in this study which are the k-NN, RF and the ADTree classifier (note that the BNet classifier does not have adjustable parameter in WEKA). Performing parameter optimisation for each of the classifiers are time consuming due to the size of dataset (number of instances more than 170,000) and the complexity of the classifier itself. Note that in this section we used the data with features extracted using $ws = 11 \times 11$ based on the results in the previous section and we have not tested on features extracted using different ws. For the k-NN classifier we tested $k = 1$ up to $k = 41$ (at 2 neighbours interval to ensure odd k values). For the RF and ADTree classifiers, we tested the initial number of random trees (rF) from 5 to 165 (with an interval $rF = 5$) and the number of boosting (nB) from 1 to 41 (with an interval $nB = 2$, $nB = 10$ the WEKA default value), respectively. The purpose of these experiments is to demonstrate the stability of the proposed method when different parameters are used.

Figure 9 shows results for A_z and CA when k is varied from 1 to 41. The data was extracted using $ws = 11 \times 11$ and $txt = 12$ (6 texton per class). In terms of classifica-

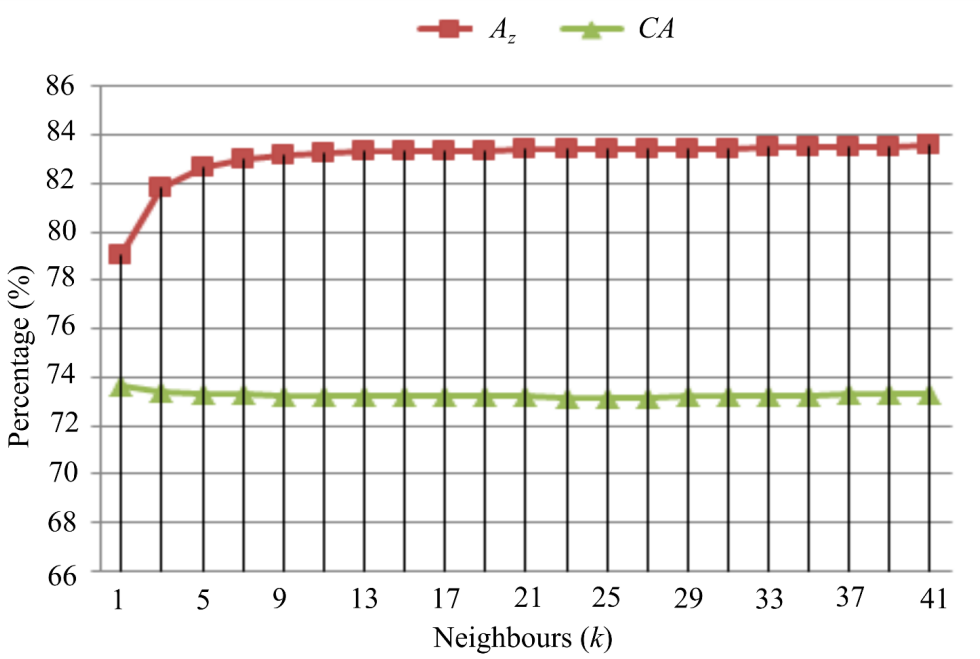

FIG. 9. The A_z and CA values using different k values for the k-NN classifier. Default $k = 1$ in WEKA[44].

tion accuracy no significant difference was noticed as all CA were between 72% to 74%. Nevertheless, there was a significant difference in terms of A_z at $k = 1$ and $k = 5$. The A_z increased to just below 84% as the k increases in comparison to $k = 1$ where the A_z value is around 79%. This indicates that the classifier tends to produce better result when comparing k neighbours instead of just taking the nearest neighbour in classification.

Figure 10 shows the results for the ADTree classifier using 22 different nB values. The classifier produced $A_z \leq 80\%$ and $CA \leq 70\%$ at $nB \leq 9$. At the default nB given in WEKA[44], it produced around A_z=83% and CA=73%. At $nB \geq 11$ both metrics change around 1% before they gradually increase at $nB \geq 17$ and reaches $A_z \leq 83\%$ and $CA \leq 73\%$. As shown in Figure 10, the classifier produced the best A_z and CA at $nB = 10$. Our explaination for this behaviour are two-fold: first it may be caused that an optimal model for the data was achieved at $nB = 10$, which means adding more iterations results in an

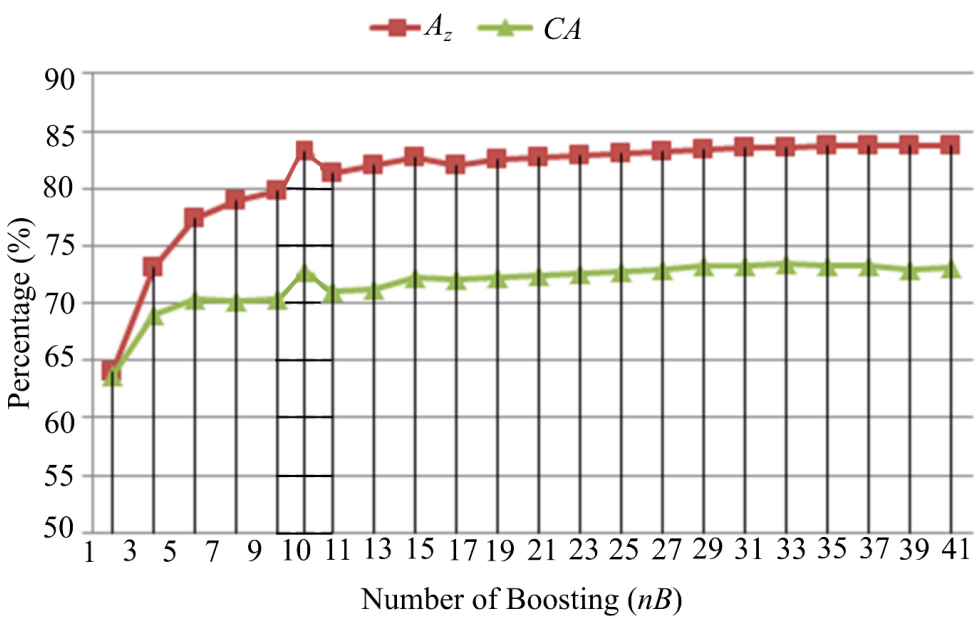

FIG. 10. The A_z and CA values using different nB values for the ADTree classifier. Default $nB = 10$ in WEKA[44].

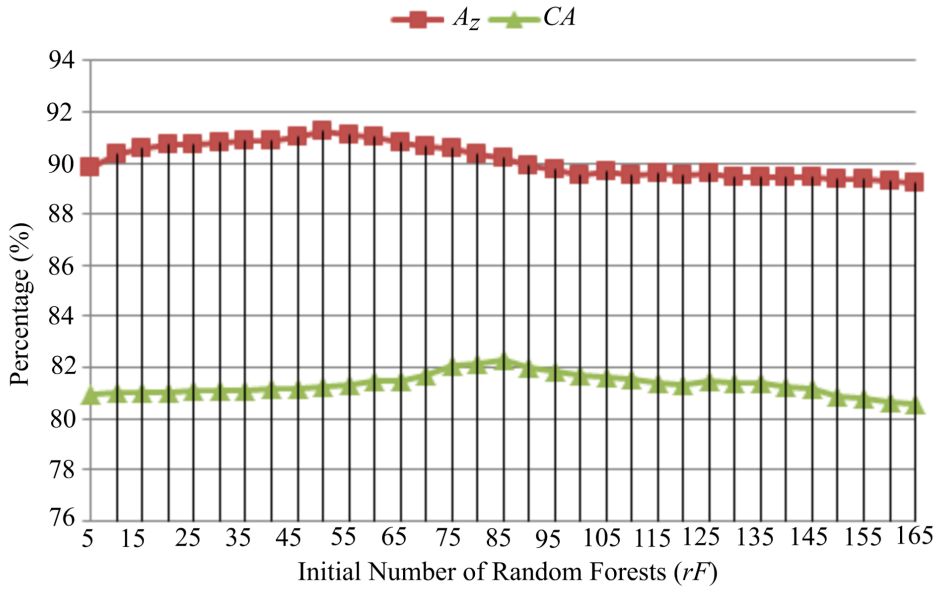

FIG. 11. The A_z and CA values using different rF values for the RF classifier. Default $rF = 100$ in WEKA[44].

overfitted model; second, adding more iterations decreases the training error but increases

the test error, which affects the overall accuracy. In an early study conducted by Freund and Manson [34] they showed that a significant test error was encountered at $nB > 10$.

In Figure 11, the RF classifier achieved the highest A_z=91% at $rF = 50$ with A_z just above 82% with $rF = 85$. Overall, there is no significance difference for both performance metrics using $rF = 5$ to 165, the CA=80%-82% and A_z=89%-91%. In this section, one visible pattern for all the classification results is after an optimal model (or optimal parameter) is achieved, both metrics showed very little change, which may be caused by the size of our data (more than 170,000 instances). For example, 100 misclassified instances has very little effect on the percentage. Oshiro *et al.*[53] conducted a study investigating the correlation between the numbers of random forests and A_z values, and found that in most cases the classifier achieved high AUC value between random forest 8 to 64 (similar to our results in Figure 11). Adding more random forest only increases the computational time.

6. QUALITATIVE COMPARISON

Despite promising results of CAD in assisting radiologists in diagnostic decision making, the major problem is the lack of publicly available datasets, resulting in each group of researchers having their own dataset. Several factors contribute to the difficulty of making quantitative comparisons in prostate cancer CAD:

1. Differences in the type of modalities (different modalities such as T2-weighted (T2-W) MRI, diffusion-weighted (DWI) MRI, dynamic contrast enhanced (DCE) MRI, Magnetic resonance spectroscopy (MRS), etc.) and protocols used in the other studies as the tissues' numerical representations are inconsistent for different modalities.

2. Absence of public datasets also makes quantitative comparisons among CADs in the

literature difficult. Each team of researchers has their own datasets which cause a huge range of variability in terms of noise, acquisition protocol and image quality.

3. Studies were conducted within different regions of the prostate. Visually it is harder to detect and differentiate malignant regions within the CZ in comparison to the ones in the PZ.

4. Finally, another difficulty is that the basis of evaluation have been different (e.g. volume, slice, regions or voxel/pixels). Pixel level evaluation is more challenging as the number of instances increases as the number of pixels increases, resulting in more complex data, whereas the number of instances is limited to the number of regions annotated by the radiologists in region level evaluation.

Nevertheless qualitative comparisons can roughly indicate the relative performance of the proposed method in this paper.

TABLE XI. Qualitative comparison with eight of the existing CAD methods in the literature. Note that # and WP indicate the number of patients and whole prostate, respectively.

Authors	**#**	**Zone**	**Modality**	A_z
Vos *et al.* [36]	29	PZ	T2-W, DCE	97
Lv *et al.* [37]	55	PZ	T2-W	97
Peng *et al.* [38]	48	WP	T2-W, DCE, DWI	95
Our method	**45**	**PZ**	**T2-W**	**93**
Vos *et al.* [39]	29	PZ	T2-W, DCE	91
Tiwari *et al.* [40]	19	WP	T2-W, MRS	91
Tiwari *et al.* [41]	36	WP	T2-W, MRS	90
Litjens *et al.* [4]	347	WP	T2-W, DCE, DWI	89
Kwak *et al.* [45]	244	WP	T2-W, DWI	89
Niaf *et al.* [42]	30	PZ	T2-W, DWI, DCE	89

All methods in Table XI achieved at least A_z=89%. The methods proposed by Vos *et al.*[36] and Lv *et al.*[37] achieved the highest A_z=97%. Vos *et al.*[36] proposed a method using features extracted from quantitative pharmacokinetic (PK) maps and T2-W MRI before training a SVM to calculate the malignancy likelihood of each lesion. However, the method was tested on a small dataset of 87 regions of interest (ROI) taken from 29 patients. On the other hand, Lv *et al.*[37] used analysis of histogram fractal dimension (HFD) and texture fractal dimension (TFD) information on a single modality of T2-W MRI. Although the study covered 55 patients, the actual evaluation was based on 130 selected ROI of 12×12 pixels (which means only a small part of the PZ region was covered). In fact, Lv *et al.*[37] did not perform cross validation to further evaluate their method. In our study, we performed 9-FCV as well as tested the proposed method on 418 PZ regions.

Peng *et al.* [38] reported A_z=95% based on T2-W, DCE, and DWI using the following features: 10th percentile apparent diffusion coefficient (ADC), average ADC, and T2-W skewness. Subsequently, individual image features were combined using linear discriminant analysis (LDA) to perform leave-one-patient-out cross validation. From an evaluation point of view, their study is similar to the studies in [36,37]. Although Peng *et al.*[38] reported that their study covered 48 patients, the actual evaluation was based on 104 ROI (61 malignant ROI, 43 normal ROI). In comparison, our proposed methods achieved similar results qualitatively with some of the methods in the literature regardless of the size of dataset, modality and studied zones.

Another study by Vos *et al.*[36] reported $Az = 91\%$ for malignant and benign discrimination, and 83% for suspicious malignant and benign discrimination, which is similar to the earlier study conducted by them [39]. Niaf *et al.*[42] extracted 140 texture features from 180 ROI (30 patients) and achieved A_z=89% which is similar to the method in [4]. Niaf *et al.*[42]

compared the performance of four different classifiers (SVM, LDA, k-NN and NB) based on four different feature selection methods. Further, their results showed that employing feature selection significantly improved the performance of their method and gradient features showed a high discriminant capability in their study.

Litjens *et al.* [4] conducted a study which covered 347 patients and reported A_z=89%. Their method consisted of two stages: in the first stage the prostate gland was segmented using a multi-atlas-based segmentation method and features based on intensity, anatomical, pharmacokinetic, texture and blobness were calculated. Subsequently, each voxel was classified using GentleBoost and RF classifiers to generate a likelihood map. On each likelihood map local maxima detection was performed to capture ROIs with the highest probability of being malignant. A method by Tiwari *et al.*[40] which is based on multi-kernel graph embedding in T2-W and MRS produced A_z=89% covering 29 patients. The method [40] was also based on a two-stage classification approach: in the first stage, a voxel based classification was performed by employing a random forest classifier in conjunction with the SeSMiK-GE based data representation and a probabilistic pairwise Markov Random Field (MRF) algorithm to identify malignant ROIs. Subsequently, each of the segmented malignant ROIs was classified as either high or low Gleason grade. Using the same method, in a smaller study [40] of 19 patients Tiwari *et al.* reported an A_z=91%. Later, Tiwari *et al.*[41] proposed a data integration framework for T2-W and MRS for prostate cancer detection. Texture descriptors such as Gabor, gradient, first and second order statistical features were extracted from T2-W and wavelet features were extracted from MRS images. Both sets of features were fused (via dimensionality reduction) using their proposed framework before employing a probabilistic boosting tree (PBT), SVM and RF classifiers. They reported an improvement of at least A_z=5% in comparison to the results without using the proposed data integration

framework.

Finally, a recent study by Kwak *et al.* [45] combining texture descriptors (different variations of Local Binary Pattern) in T2-W and b-value in DWI showed similar results to the studies of Litjens *et al.* [4] and Niaf *et al.*[42]. However, the proposed method involved a three-step feature selection process, which can be time consuming. In fact, texture information extracted using local binary patterns and its variants yields a large number of features and extracting unecessary features, both of which can be computationally expensive. In comparison to our proposed method, we achieve A_z approximately 90% using the RF classifier with only 6 textons (only 6 features and feature selection is not necessary) and at smaller rF, our proposed method achieved A_z>91% using the same classifier.

7. DISCUSSION

In our experiments, the results suggest that the number of textons in the dictionary has a significant effect on both A_z and CA. For example ADTree, k-NN and RF classifiers produced better results at a smaller txt value. In contrast, the BNet classifier performed slightly better at txt = 16 or 20. Both metrics are highly influenced by the ws used to construct the histograms (treated as feature vectors) from the texton maps. Furthermore, using the maximum value of ws (in our case 13×13), reduced the performance on both metrics. In terms of selecting the best ws and txt, most classifiers performed well at 9×9 and 11×11 with 6 or 16 textons (3 and 8 textons per class, respectively).

All classifiers produced the best results when the size of the textons generated are either 9×9 or 11×11. Our explanations for this are four-fold. Firstly using a small ws such as 5×5 does not provide sufficient information about the regions (such as limited intensities and grey level variations). Secondly, small textons which contain or affected by noise are

unable to characterise the actual representation of the region. Thirdly, using a medium ws (e.g. 9×9) features tend to be more reliable because 'noisy pixels' are shrunk by the domination of 'reliable pixels' (e.g. malignant pixels). Finally, when using a large ws (e.g. 13×13), the performance tends to decrease because the chance of mixing up pixels from benign and malignant classes is higher, hence altering the actual feature's representation of a particular class.

The number of textons affects the complexity of the predictive model built by the classifier. Most classifiers produced better results using smaller number of textons (e.g. $txt = 6$) because it reduces the data complexity. This makes it much easier for the classifier to create decision boundaries between classes which decreases error rates and increases the accuracy of the model in making a prediction in unseen cases. Nevertheless, for the BNet classifier using a small number of textons (e.g. $txt = 6$) is insufficient to build a network model that can represent the problem. A larger number of textons was needed (e.g. $txt = 16$) to build an optimum network model to represent the problem.

From a computer vision point of view, the appearance of textons are determinded by pixel intensities which means they could be influenced by the image intensities. In contrast, feature-based textons are less affected by image intensities. Therefore, to overcome this issue most of the texton-based approaches perform image normalisation. In our study, we used a similar approach by normalising image intensities to zero mean and unit variance. The standardization gives similar representation for normal and malignant regions for the whole prostate. From a data classification point of view, data imbalance between normal and malignant samples (resulting in a model biased towards the class with a larger number of samples) is another challenge in the development of prostate cancer CAD. Most CAD systems employed stratified cross validation as a standard procedure to build a predictive

model. Recently, Fehr *et al.*[49] proposed an alternative procedure using data augmentation for classifying prostate cancer aggressiveness.

There are two main advantages of the proposed method: first it bypasses the typical feature extraction algorithm such as filtering and convolution which can be computationally expensive; secondly it does not need the additional step of feature selection as the number of textons are already small (e.g. $txt = 6$ or 16). With a large number of features, selecting the best features can be time consuming. Although feature selection can significantly improve classification results, it needs to be applied with a robust feature selection algorithm. The results suggested that even at $txt = 6$, our method can produce A_z>90% with the BNet classifier and using the simplest classifier (k-NN) can still achieve around 87%. In fact, all classifiers employed in this study produced A_z>80%. There are three reasons why most texton-based methods in the literature used larger numbers of textons in comparison to our proposed method. Firstly, most filter banks lead to blurring of the appearance of the texture. This results in a larger number of textons being needed to characterise the actual representation of the texture. Secondly, the number of textons used is depending on the variation (or complexity) of the textures. For example to construct a texton model for a grass image we need more textons because the variations are huge due to different orientations, shapes, colours, sizes, etc. In contrast to a specific cancer (e.g. prostate or lung cancer), which can show limited variations. In retinal vessel segmentation [13], the authors reported that 12 textons are sufficient to get good segmentation results. Similarly, the study in [14] achieved more than 85% accuracy in lung cancer detection using only 10 textons. Finally, most studies investigated a larger number of classes which increase the number of textons in the codebook (texton dictionary) considerably. For example, studies in [15,16,18,21] classified more than 60 different textures resulting in more than a thousand textons generated in the

codebook. In comparison, the studies in [13,14] classified only two classes vessel and non-vessel and healthy and non-healthy resulting in the best accuracy using 12 and 10 textons.

The main limitation of this study is that we are unable to compare the results quantitatively with existing methods due to the absence of public datasets. This is the major problem among the research communities in prostate cancer CAD. Secondly, we are unable to test the classifiers at different parameters with features extracted using different ws. Nevertheless, the experimental results presented in this paper showed the prospect of CAD based on T2-W MRI to achieve similar results with CADs based on multiparametric MRI (and as such would form an excellent basis for multiparametric MRI based CAD) without the need for the typical feature extraction methods such as filtering and convolution. In addition, preliminary results indicated that the proposed method could achieve better results when the classifier's parameters were optimised. Although the proposed method achieved $A_z = 92.8\%$, we would like to emphasise that this does not indicate that our method is better in comparison to the other prostate cancer CAD based on multiparametric MRI. We are aware that several studies [4,9,46] have shown that combining different features from different modalities can increase the CADs performance. It should be noted that the classification accuracies reported in this study were based on the prostate PZ. Therefore the proposed method may not produce the same accuracy when tested within the CZ due to different MR phenotypes (e.g. the tissue contrast between PZ and CZ is different due to higher water content within the PZ). Another limitation of our study is that our ground truth was based on TRUS biopsies report and annotated by a radiologist. Due to the random procedure and limited access on the horns of the PZ, TRUS biopsies can miss cancerous tissues, particularly small lesions. This means some of our training samples might be inaccurate (e.g. some benign lesions might be malignant). In contrast, template biopsy can help detect small

lesions as its procedure tends to obtain more tissue samples from across the prostate. On the other hand, radical prostatectomy allows the whole prostate to be examined thoroughly in the lab which will give definite results about how aggressive the cancer is and how far it may have spread. Future work will investigate combining texton with features from the other modalities such as DCE, MRS and DWI and use patches with deep learning to see if there is a significant effect on both performance metrics.

8. SUMMARY AND CONCLUSIONS

The proposed method consists of the following steps: (a) pre-processing, (b) construction of the texton dictionary, (c) feature extraction, (d) training and testing. We used median and anisotropic diffusion filtering techniques in the pre-processing phase. To construct the texton dictionary we did not use a filter bank as originally proposed by Varma and Zisserman [15]. Instead we followed their later study in [21] which clusters benign and malignant patches directly from the original image pixels. In the feature extraction phase, each pixel is represented as a histogram treated as a feature vector. The constructed histogram for each pixel consists of the frequency of the neighbouring texton occurrence within the ws (or patch size) including the texton at the central pixel. Subsequently we employed 4 classifiers to build predictive models and test the models on unseen cases.

Evaluation results show the proposed method achieved similar results with the state-of-the-art in all performance metrics. Indeed, the texton based approach relies on two parameters ws and txt. In our experiments we found $ws = 9 \times 9$ and 11×11 with 6 and 16 textons produced the best results for most classifiers. The BNet, RF and k-NN classifiers are the best three machine learning algorithms produced A_z >87%.

In conclusion, we have developed a CAD texton based approach based on a single modal-

ity of T2-W MRI and have similar performance with those based on multiparameteric MRI and as such would form an excellent basis for multiparametric MRI based CAD. Experimental results suggest that textons can be used as robust texture descriptors to characterise benign and malignant tissues. The main advantages of the proposed method are: firstly it bypasses the conventional feature extraction methods based on filtering and secondly it avoids dimensionality reduction methods or feature selection which can both be time consuming. To our knowledge this is the first texton based CAD method in the literature applied to prostate cancer detection.

ACKNOWLEDGMENTS

Andrik Rampun is grateful for the awards given by Aberystwyth University under the Departmental Overseas Scholarship (DOS) and Doctoral Career Development Scholarships (DCDS). This work was funded in part by the NISCHR Biomedical Research Unit for Advanced Medical Imaging and Visualization.

DISCLOSURE OF CONFLICTS OF INTEREST

The authors have no relevant conflicts of interest to disclose.

a) Authors to whom correspondence should be addressed. Electronic mails:yar@aber.ac.uk or rrz@aber.ac.uk.

REFERENCES

[1]International agency for research on cancer, "GLOBOCAN 2012: Estimated cancer incidence, mortality and prevalence worldwide in 2012" (2012), http://globocan.iarc.fr/Pages/fact_sheets_cancer.aspx, Accessed 2-August-2015.

[2]American Cancer Society, "Cancer facts & figures 2015", 2015.

[3]F. H. Schroder, J. Hugosson, M. J . Roobol, T. L Tammela, S. Ciatto, V. Nelen, S. Ciatto, V. Nelen, M. Kwiatkowski, M. Lujan, H. Lilja, M. Zappa, L. J. Denis, F. Recker, A. Berenguer, L. Maattanen, C. H. Bangma, G. Aus, A. Villers, X. Rebillard, T. van der Kwast, B. G. Blijenberg, S. M. Moss, H. J. de Koning and A. Auvinen. For the ERSPC Investigators. "Screening and prostate-cancer mortality in a randomized European study", *New England Journal of Medicine*, 360, 1320–1328 (2009).

[4]G. Litjens, O. Debats, J. Barentsz, N. Karssemeijer, H. Huisman. "Computer-aided detection of prostate cancer in MRI", *IEEE Transactions on Medical Imaging*, vol. 33(5), 1083–1092(2014).

[5]K. Doi. "Computer-Aided Diagnosis in Medical Imaging: Historical Review, Current Status and Future Potential", *Computerized medical imaging and graphics: the official journal of the Computerized Medical Imaging Society*, vol. 31(4-5), 198–211, 2007.

[6]J. Shiraishi, Q. Li, D. Appelbaum and K. Doi. "Computer-aided diagnosis and artificial intelligence in clinical imaging", *Seminars in Nuclear Medicine*, vol. 41(6), 449–462(2011).

[7]R. M. Summers, J. Liu , B. Rehani, P. Stafford, L. Brown, A. Louie, D. S. Barlow, D. W. Jensen, B. Cash , J. R. Choi, P. J. Pickhardt and N. Petrick. "CT colonography computer-aided polyp detection: effect on radiologist observers of polyp identification by CAD on both the supine and prone scans", *Academic Radiology*, 17, 948–959 (2010).

[8]Y. Artan and I. S. Yetik. Prostate cancer localization using multiparametric MRI based on semi-supervised techniques with automated seed initialization, *IEEE Transactions on Information Technology*

in Biomedicine, vol. 16(6), 2986–2994 (2012).

[9]S. Viswanatha, B. N. Blochb, J. Chappelowa, P. Patela, N. Rofskyc, R. Lenkinskid, E. Genegad and A. Madabhushi. Enhanced multi-protocol analysis via intelligent supervised embedding (empravise): detecting prostate cancer on multi-parametric MRI, Proc. SPIE Medical Imaging, 7963 (2011).

[10]J. O. Barentsz, J. Richenberg, R. Clements, P. Choyke, S. Verma, G. Villeirs, O. Rouviere, V. Logager and J. J. Futterer. European Society of Urogenital Radiology. ESUR prostate MR guidelines 2012. *European Radiology*, vol. 22(4), 746–757 (2012).

[11]G. Lemaitre, R. Marti, J. Freixenet, J. C. Vilanova, P. M. Walker and F. Meriaudeau. Computer-aided detection and diagnosis for prostate cancer based on mono and multi-parametric MRI: a review. *Computers in Biology and Medicine*, 60, 8–31 (2015).

[12]T. Leung and J. Malik. "Representing and recognizing the visual appearance of materials using three-dimensional textons", *International Journal of Computer Vision*, vol. 43(1), 29–44 (2001).

[13]L. Zhang, M. Fisher and W. Wang. "Retinal vessel segmentation using Gabor filter and textons". Proc. 18th Conference on Medical Image Understanding and Analysis, MIUA'14, 155–160 (2014).

[14]M. Gangeh, L. Sorensen, S. Shaker, M. Kamel, M. de Bruijne and M. Loog. "A Texton-Based Approach for the Classification of Lung Parenchyma in CT Images", Proc. Medical Image Computing and Computer-Assisted Intervention MICCAI 2010, 6363, 595–602 (2010).

[15]M. Varma and A. Zisserman. "A statistical approach to texture classification from single images", *International Journal of Computer Vision*, vol. 62(1), 61–81 (2005).

[16]M. Varma and A. Zisserman. "A statistical approach to material classification using image patch exemplars", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31(11), 2032–2046 (2009).

[17]B. Julesz. "A theory of preattentive texture discrimination based on first-order statistics of textons", *Biological Cybernetics*, vol. 41(2), 131-138 (1981).

[18]M. Varma and A. Zisserman. "Classifying images of materials: achieving viewpoint and illumination independence", Proc. 7th European Conference on Computer Vision, Copenhagen, Denmark, 3, 255–271(2002).

[19]C. Schmid, "Constructing models for content-based image retrieval", Proc. of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2,II–39–II–45 (2001).

[20]D. Gabor. "Theory of communication", *Journal of the Institute of Electrical Engineers*, 93, 429–457(1946).

[21]M. Varma and A. Zisserman. "Texture classification: Are filter banks necessary?", Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2, 691-698(2003).

[22]L. van der Maaten and E. O. Postma. "Texton-based Texture Classification", Proc. of the Belgian-Dutch Artificial Intelligence Conference 2007, 213–220 (2007).

[23]A. Rampun, Z. Chen, P. Malcolm, B. Tiddeman, and R. Zwiggelaar. "Computer-aided diagnosis: detection and localization of prostate cancer within the peripheral zone", *International Journal for Numerical Methods in Biomedical Engineering* (in press), ISSN 2040-7947. doi: 10.1002/cnm.2745 (2015). http://dx.doi.org/10.1002/cnm.2745.

[24]N. Makni, A. Iancu, O. Colot, P. Puech, S. Mordon and N. Betrouni. Zonal segmentation of prostate using multispectral magnetic resonance images, *Medical Physics*, 38, 6093–6105(2011).

[25]X. Liu, M. A. Haider and S. Yetik. Automated Prostate Cancer Localization with MRI without the need of manually extracted peripheral zone, *Medical Physics*, vol. 38(6), 2986–2994(2011).

[26]A. Madabhushi, J. Udupa and A. Souza. "Generalized scale: theory, algorithms, and application to image inhomogeneity correction", *Computer Vision Image Understanding*, vol. 101(2), 100–121(2006).

[27]A. Madabhushi and J. K. Udupa. "New methods of MR image intensity standardization via generalized scale", *Medical Physics*, vol. 33(9), 3426–3434(2006).

[28]A. Madabhushi, J. K. Udupa and G. Moonis. "Comparing MR image intensity standardization against tis-

sue characterizability of magnetization transfer ratio imaging", *Medical Physics*, vol. 24(3), 667–675(2006).

[29]Y. Artan, M. A. Haider, D. L. Langer and I. S. Yetik. "Semi-supervised prostate cancer segmentation with multiparametric MRI", *Proceedings International Symposium Biomedical Imaging*, 648–651 (2010).

[30]J. Liang and A. Bovik. "Smoothing low-SNR molecular images via anisotropic median-diffusion", *IEEE Transactions on Medical Imaging*, vol. 21(4), 377–384(2002).

[31]P. Perona and J. Malik. "Scale-space and edge detection using anisotropic diffusion", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12(7), 629–639(1990).

[32]J. A. Hartigan and M. A. Wong. "Algorithm AS 136: A K-Means Clustering Algorithm", *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 28(1), 100–108(1979).

[33]T. Bayes. "An Essay Toward Solving a Problem in the Doctrine of Chances", *Philosophical Transactions of the Royal Society of London* 53, 370–418(1763).

[34]Y. Freund and L. Mason. "The alternating decision tree learning algorithm", Proc. Sixteenth International Conference on Machine Learning, Bled, Slovenia, 124–133(1999).

[35]L. Breiman. "Random forests", *Machine Learning*, vol. 45(1), 5–32(2001).

[36]P. C. Vos, T. Hambrock, J. Barentsz and H. Huisman. "Computer-assisted analysis of peripheral zone prostate lesions using T2-weighted and dynamic contrast enhanced T1-weighted MR", *Physics in Medicine and Biology*, vol. 55(6), 1719–1734 (2010).

[37]D. Lv, X. Guo, X. Wang, J. Zhang and J. Fang. "Computerized characterization of prostate cancer by fractal analysis in MR images", *Journal of Magnetic Resonance Imaging*, vol. 30(1), 161–168(2009).

[38]Y. Peng, Y. Jiang, C. Yang, J. B. Brown, T. Antic, I. Sethi, C. Schmid-Tannwald, M. L. Giger, S. E. Eggener and A. Oto. "Quantitative analysis of Multiparametric Prostate MR Images: Differentiation between prostate malignant and normal tissue and correlation with gleason score-A Computer-aided diagnosis development study", *Radiology*, vol. 267(3), 787–796(2013).

[39]P. C. Vos, T. Hambrock, J. O. Barenstz and H. J. Huisman, "Combining T2-weighted with dynamic MR images for computerized classification of prostate lesions", Proc. of SPIE 6915, Medical Imaging 2008:Computer-Aided Diagnosis, 69150W-69150W-8(2008).

[40]P. Tiwari, J. Kurhanewicz, M. Rosen and A. Madabhushi,"Semi supervised multi kernel (SeSMiK) graph embedding: identifying aggressive prostate cancer via magnetic resonance imaging and spectroscopy", Proc. Medical Image Computing Computer Assisted Interventions (MICCAI), 666–673 (2010).

[41]P. Tiwari, S. Viswanath, J. Kurhanewicz, A. Shridhar and A. Madabhushi, "Multimodal wavelet embedding representation for data combination (MaWERiC): integrating magnetic resonance imaging and spectroscopy for prostate cancer detection", *NMR in Biomedicine*, 25, 607–619 (2012).

[42]E. Niaf and O. Rouvière and F. Mège-Lechevallier and F. Bratan and C. Lartizien, "Computer aided diagnosis of prostate cancer in the peripheral zone using multiparametric MRI", *Physics in Medicine and Biology*, vol. 57(12), 3833–3851(2012).

[43]N. Makni, A. Iancu, O. Colot, P. Puech, S. Mordon and N. Betrouni. Zonal segmentation of prostate using multispectral magnetic resonance images, *Medical Physics*, 38, 6093–6105 (2011).

[44]M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten, "The WEKA data mining software: an update", *SIGKDD Explorations*, vol. 11(1), 2009.

[45]J. T. Kwak, S. Xu, B. J. Wood, B. Turkbey, P. L. Choyke, P. A. Pinto, S. Wang, R. M. Summers, "Automated prostate cancer detection using T2-weighted and high-b-value diffusion-weighted magnetic resonance imaging", *Medical Physics*, vol. 42(5), 2368-7238 (2015).

[46]P. Tiwari, J. Kurhanewicz and A. Madabhushi. Multi-kernel graph embedding for detection, Gleason grading of prostate cancer via MRI/MRS. *Medical Image Analysis*, vol.17(2), 219–235 (2013).

[47]I. Ocak,M. Bernardo, G. Metzger, T. Barrett, P. Pinto, P. S. Albert, P. L. Choyke. Dynamic contrast-enhanced MRI of prostate cancer at 3 T: a study of pharmacokinetic parameters. American Journal of

721 Roentgenology, vol. 189(4), W192–W201 (2007).

722 [48]J. Carlsson, G. Helenius, M. G. Karlsson, O. Andren, K. Klinga-Levan, B. Olsson. Differences in microRNA
723 expression during tumor development in the transition and peripheral zones of the prostate. BMC Cancer,
724 vol.13(1): pp.6093–6105 (2013).

725 [49]D. Fehr, H. Veeraraghavan, A. Wibmerb, Tatsuo Gondo, K. Matsumoto, H. A. Vargas, E. Sala, H. Hricak
726 and J. O. Deasy. Automatic classification of prostate cancer Gleason scores from multiparametric magnetic
727 resonance images. *Proc. Natl Acad. Sci. USA*, 112, E6265–E6273 (2015).

728 [50]J.H. Friedman. Regularized Discriminant Analysis. *Journal of the American Statistical Association*, vol.
729 84(405), 165–175 (1989).

730 [51]I. Chan, W. Wells III, R. V. Mulkern, S. Haker, J. Zhang, K. H. Zou, S. E. Maier and C. M. C. Tempany.
731 Detection of prostate cancer by integration of line-scan diffusion, T2-mapping and T2-weighted magnetic
732 resonance imaging; a multichannel statistical classifier. *Medical Physics*, vol. 30, pp. 2390–2398 (2003).

733 [52]P. C. Vos, J. O. Barentsz, N. Karssemeijer and H. J. Huisman. Automatic computeraided detection of
734 prostate cancer based on multiparametric magnetic resonance image analysis. *Physics in Medicine and*
735 *Biology*, vol. 57, pp.1527–1542 (2012).

736 [53]T. M. Oshiro, P. S. Perez and J. A Baranauskas. How many trees in a random forest? In: Perner P,
737 editor, Machine Learning and Data Mining in Pattern Recognition, Springer Berlin Heidelberg, vol. 7376
738 of Lecture Notes in Computer Science. 154–168, (2012).

739 [54]A. Billis, M. M. Q. Quintal, L. L. L. Freitas, L. B. E. Costa and U. Ferreira. Predictive criteria of
740 insignificant prostate cancer: what is the correspondence of linear extent to percentage of cancer in a
741 single core?. *International Brazilian Journal Of Urology*, vol. 41 (2), pp. 367–372, 2015.

742 [55]R. T. Vollmer. Tumor Length in Prostate Cancer. *American Journal of Clinical Pathology*, vol. 130,
743 pp.77–82, 2008.

[56]S. Lee, J. K. Lee, J. Keun, C. W. Jeong, S. J. Jeong, S. K. Hong, S. S. Byun, S. E. Lee and H. Lee. Core length as a predictor of Gleason score upgrading in men diagnosed with low risk prostate cancer by contemporary multicore prostate biopsy. *The Journal of Urology*, vol. 189 (4), 2013.

[57]J. H. Yacoub, S. Verma, J. S. Moulton, S. Eggener, and A. Oto. Imaging-guided Prostate Biopsy: Conventional and Emerging Techniques. *Radiographics*, vol.32(3), pp. 819–837, 2012.