Computer-aided detection of prostate cancer in T2-weighted MRI within the peripheral zone

Rampun, Andrik; Zheng, Ling; Malcolm, Paul; Tiddeman, Bernard; Zwiggelaar, Reyer

Published in:
Physics in Medicine and Biology

DOI:
10.1088/0031-9155/61/13/4796

Publication date:
2016

Citation for published version (APA):
Rampun, A., Zheng, L., Malcolm, P., Tiddeman, B., & Zwiggelaar, R. (2016). Computer-aided detection of prostate cancer in T2-weighted MRI within the peripheral zone. *Physics in Medicine and Biology*, *61*(3), 4796-4825. https://doi.org/10.1088/0031-9155/61/13/4796

Computer-Aided Detection of Prostate Cancer in T2-Weighted MRI within the Peripheral Zone

Andrik Rampun[1], Ling Zheng[1], Paul Malcolm[2], Bernie Tiddeman[1], Reyer Zwiggelaar[1]

[1]Department of Computer Science, Aberystwyth University, Aberystwyth SY23 3DB, UK

[2]Department of Radiology, Norfolk Norwich University Hospital, Norwich NR4 7UY, UK

Abstract. In this paper we propose a prostate cancer computer-aided diagnosis (CAD) system and suggest a set of discriminant texture descriptors extracted from T2-weighted MRI data which can be used as a good basis for a multimodality system. For this purpose, 215 texture descriptors were extracted and eleven different classifiers were employed to achieve the best possible results. The proposed method was tested based on 418 T2-weighted MR images taken from 45 patients and evaluated using 9-fold cross validation with five patients in each fold. The results demonstrated comparable results to existing CAD systems using multimodality MRI. We achieved an area under the receiver operating curve (A_z) values equal to 90.0% ± 7.6%, 89.5% ± 8.9%, 87.9% ± 9.3% and 87.4% ± 9.2% for Bayesian Networks, ADTree, Random Forest and Multilayer perceptron classifiers, respectively, while a Meta-voting classifier using average probability as a combination rule achieved 92.7% ± 7.4%.

1. Introduction

According to the latest figures compiled by the American Cancer Society [1], prostate cancer is the fourth most common cancer occurring globally and is only surpased by lung, female breast and bowel cancers. In 2013, more than 237,000 and 40,000 incidences were reported in the United States (US) and United Kingdom (UK), respectively and is estimated to reach 1.7 million cases globally by 2030 in comparison to 1.1 million cases reported in 2012 [2–4]. Current clinical practices such as transrectal ultrasound biopsy (TRUS), prostate specific antigen blood test (PSA) and digital rectal examination (DRE) are among the most popular methods used in hospitals, and such methods have shown the potential to reduce prostate cancer mortality by up

to 30% [5]. However, these methods are associated with several problems, for example PSA testing and TRUS biopsies have relatively low specificity which leads to overdiagnosis and overtreatement of patients [6]. In fact, TRUS biopsy does not provide robust results due to limited length of the needle which could potentially miss tumor regions within the prostate capsule [92].

In an effort to minimise these issues, integrating Magnetic Resonance Imaging (MRI) with other clinical practices (e.g. MRI/Ultrasound guided biopsy and multimodality image fusion) is becoming popular to diagnose prostate cancer and has shown a significant improvement over PSA and TRUS [5, 6]. Unfortunately such methods require substantial expertise from the radiologist, a significant amount of human interaction (which increases the potential of human errors), is time consuming and often suffers from observer variability [5, 6]. Automated Computer Aided Diagnosis (CAD) systems can be used to reduce these problems [84].

Over the last decade, CAD systems have been successfully used in mammography [8, 9] and CT colonoscopy [10]. The development of CAD systems for prostate cancer (CAD-PC) MRI is becoming an active field of research [6, 7]. In 2015, Lemaitre *et al.* [7] conducted a review of CAD-PC and reported that there are 42 studies (using mono and multi-parametric MRI) in the literature from 2007 until 2014. Studies [6, 10–13] have reported the limitation of CAD systems using single T2-weighted (T2-W) MRI including weak texture descriptors and noise. In fact, Tiwari *et al.* [39,55] suggests that T2-W MRI texture features alone might not be sufficient to identify prostate malignancies. Therefore, the use of multimodality MRI in developing CAD systems is a popular way to improve the performances of existing methods. We are aware the fact that using T2-W alone is not sufficient, but that T2-W classification will form a solid basis for a multimodality MRI based system. None of the previous studies have toroughly investigated texture descriptors in T2-W MRI in distinguishing malignant and normal (included benign) voxels.

This study does not attempt to improve the performance of CAD-PC based on multimodality MRI but aims to investigate the performance of CAD-PC using a large number of texture descriptors in T2-W MRI alone within the PZ towards inclusion of a more general multimodality MRI classification platform. As a result, this study will reveal some of the most discriminant texture descriptors in T2-W MRI. Our main motivation of using T2-W MRI is due to its availability in most general hospitals in comparison to the other modalities, such as Dynamic Contrast Enhanced (DCE) and Diffusion Weighted (DW) MRI which are not always available. On the other hand, our study is currently focused within the PZ because 80% of prostate cancers arise within this region and prostate cancer that arises within the PZ is

more aggressive than that which arises in the transitional zone. The novel contributions of our work are the following:

(i) The proposed method incorporates a large number (215) of different texture descriptors from T2-W MRI. This means our study investigates a number of novel feature options that have not previously been applied in CAD-PC. To the authors' knowledge the previously used largest number of 2D texture descriptors in the literature using only T2-W MRI was in a study conducted by Viswanath *et al.* [14] (110 texture descriptors) and 83 texture descriptors by Tiwari *et al.* [39]. Using multimodality MRI Niaf *et al.* [12] extracted 140 texture descriptors from T2-W MRI, diffusion-weighted imaging (DWI), and dynamic contrast enhanced (DCE).

(ii) We extensively compare 9 classifiers' performances with two additional combined classifiers (11 classifiers in total). Again, to our knowledge the largest number of classifiers used in the literature in CAD-PC is in a study conducted by Niaf *et al.* [12] (4 classifiers), and Litjens *et al.* [6] and Ozer *et al.* [17] employed 3 classifiers.

(iii) Since this study involved a large number of texture descriptors, we evaluated all features individually and combined them to improve performances of the proposed method. By evaluating them, we are able to determine which features individually give the best performances on each classifier.

(iv) Finally we investigate the effect of different window sizes on the performance of the proposed method as well as the performance of the features individually. To our knowledge, none of the current CAD-PC in the literature have reported quantitatively the effect of window size on performance. Many studies [12,14,39,41,72] selected window size without a qualitative justification.

These contributions are expected to be beneficial to the research community as they provide state-of-the-art CAD-PC using single modality T2-W MRI. Moreover, this study provides general guidelines on selecting window size, classifier, as well as the set of features used in CAD-PC development.This is the first study in CAD-PC has investigated the effects of window size on the CAD's and features' performances, investigated the largest number of texture features in T2-W MRI and largest number of classifiers with qualitative comparisons. These can be used as a good foundation for the development of CAD-PC based on multimodality data.

2. Methodology

Figure 1 provides a general overview of our method. For every input image, we estimate the area of the prostate's PZ and extract 215 feature descriptors. We normalise each of the selected features. Feature selection was performed to eliminate irrelevant or redundant features and use them to train 11 classifiers. Finally in the testing phase, for every unseen pixel within the PZ the trained classifiers determined whether it belongs to the malignant or benign class.

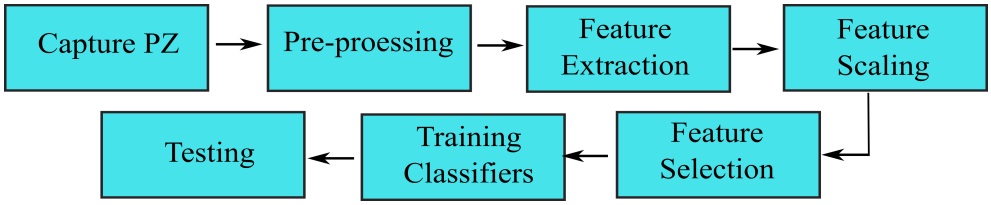

Figure 1: A general overview of the proposed method.

2.1. Capturing the Peripheral Zone

Since segmenting the PZ manually is time consuming, we employed the 2D model developed by Rampun *et al.* [18], which 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. This allows us to model a *priori* general knowledge of radiologists which is similar to the methods of Makni *et al.* [19] and Liu *et al.* [20]. Figure 2 shows example MRI images with the ground truth location of the prostate gland, central/transitional 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 [18]. Our study is only within the segmented PZ which is under the magenta line in Figure 2. Note that in this study we did not perform prostate segmentation because all prostates were already delineated by an expert radiologist. Nevertheless, we are aware that several prostate segmentation methods have been developed in the last decade [84–86].

2.2. Pre-processing

A major challenge in MR image analysis is that intensities do not have a fixed tissue specific numeric meaning even within the same MRI protocol, the same body region, and the same scanner [21–23]. These problems are mainly caused by [11, 21–23]: a) corruption by thermal noise due to receiver coils, b) intensity variations due to different scanning protocols and c) poor radio

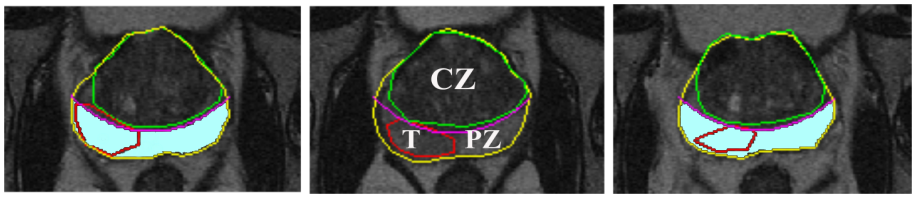

Figure 2: Example images of prostate MRI with the ground truth delineated by an expert radiologist and the estimated PZ region under the magenta line.

frequency coil uniformity. Intensity variations in MR data can significantly affect performances of many image processing techniques, hence, they need to be corrected [23]. Following the pre-processing procedure method described in [11, 24, 28], each image is median filtered to preserve edge boundaries. Subsequently, image intensities were normalised to zero mean unit variance and anisotropic diffusion filtering [28] is applied to remove noise. This method is chosen because it does not cause inter regional blurring [11, 24, 28]. However, [11, 24] suggested that anisotropic diffusion filtering needs to be applied on the median-filtered and normalised images for better results. This three-step pre-processing approach has the following advantages [11, 24] a) while suppressing the noise, it simultaneously preserves the edge boundaries b) it standardise image intensities for all patients avoiding dissimilar intensity values for the same tissue types c) it is a robust denoising method without blurring the tumor edges. Other denoising techniques in the literature could also be investigated [27].

2.3. Feature Extraction

In this study, we extracted a set of 215 image features and their selection was mainly motivated by statistical, psychological and image and signal processing points of view. This means each pixel is represented in a 215 dimensional feature space. These features were selected based on the visual characteristics of malignant regions as indicated by expert pathologists as well as their efficiency at discriminating between malignant and benign regions [14–16]. In this study, for first and second order statistical features, Tamura texture features and grey-level percentile based features estimated for each pixel for a local $n \times m$ window [12] where n and m are rows and columns, respectively. The features extracted in this study can be divided into the following categories:

First order statistical features (F_1). These features mainly rely on image intensity as used in many pattern recognition algorithms. Studies in CAD-PC [12, 14, 17, 30, 41] have used intensity-based features extensively in

the last decade. From a clinical point of view, most malignant regions in the PZ tend to have a dark appearance (low intensity) [31, 36] (we are aware that in some cases low intensity does not represent malignancy (resulting in false positive detection) due to inflammation and post-biopsy scarring [42], therefore other texture descriptors such as a filter bank were used in this study) and radiologists tend to use darker regions as the basis of their a *priori* knowledge to identify abnormality within the PZ [35]. Moreover, several studies have suggested that prostate cancer tissue tends to appear darker on a T2-W MRI image [32–34]. Therefore the selection of intensity-based features is appropriate in this study. Niaf *et al.* [12] used mean, median and standard deviation in their study. On top of that, we extracted mean and median absolute deviation, skewness, kurtosis, the mean of correlation coefficients, local contrast [18], variance and local probability [18] (11 features in total). A study in [37] indicated that many malignant regions are more likely to have low contrast value. On the other hand, a study of Rampun *et al.* [54] used probability images to quantify the likelihood of every pixel/voxel belonging to specific tissues (e.g. the tumor region).

Second order statistical features (F_2). The main motivation for using Haralick's features (Grey Level Co-occurrence Matrix (GLCM)) [38] is that these texture features characterized homogeneity, grey-level transitions, and anatomical structures in the image [7] as well as its simplicity, large range of potential features and its popularity (second after intensity-based features in MRI [7]). Julesz [40] states that first order statistics alone are not sufficient for humans to discriminate between two textures. Hence, in order for a CAD system to be able to discriminate textures Madabhushi *et al.* [41] consider both first-order and second-order statistical features. The GLCM is defined as the joint probability of occurrence of two grey level values at a given offset both in terms of distance and orientations [43]. We extracted all features originally suggested by Haralick *et al.* [38] and all features which were further suggested by Soh and Tsatsoulis [44] and Clausi [45]. To maximise the texture information captured from the co-occurrence matrix we considered four orientations ($\theta = 0°, 45°, 90°$ and $135°$) with distance d limited to 1 (a larger distance increases the potential of losing more information as reported in [47]). In addition, we calculated the mean, variance and standard deviation of four orientations for each of the features (154 features in total). From a clinical point of view, according to [48] malignant regions have a higher degree of uniformity in T2-W MRI. In fact, studies in [49, 50] suggested that the distribution of malignant foci detected by biopsies in the peripheral zone of the prostate is homogeneous. To capture these characteristics in features we use GLCM as we can calculate uniformity (also known as energy) and

homogeneity.

Percentile based features ***(F_3)*** [51] namely lower and higher quartile to quantify the symmetry of the image (or a region of interest) intensity distribution [69] (2 features in total). Vos *et al.* [52] and Niaf *et al.* [12] extracted similar features and found that many malignant regions have smaller values in the upper quartile (2 features in total). In our case we compute these features by replacing the central pixel with lower and upper quartile value of the grey levels within a $n \times m$ sliding window. These features represent the distribution of signal intensity within a specified window (e.g. 7×7 or 9×9). A smaller value of higher quartile indicates the central pixel within a specified window is surrounded by pixels with low intensities which increases the probability of being malignant (most tumors display low signal intensity within the PZ in T2-W MRI [42]). To calculate these features: (a) sort grey level values (X) within a $n \times m$ sliding window in an acending order, (b) the value for lower quartile is the middle value between the smallest grey level and the median of X and (c) the upper quartile is the middle value between the median and the highest grey level value in X. For example, let q be the quartile position in X and lower and upper quartiles are the q^{th} and $(3 \times q)^{th}$ grey levels in X (sorted in an ascending order), respectively.

Tamura texture features ***(F_4)***. In [16] six texture features corresponding to human visual perception were proposed: coarseness, contrast, directionality, line-likeness, regularity, and roughness. However, from experiments testing the significance of these features with respect to human perception, it was concluded that only the first three features are very important [16]. Therefore we only use the first three features in this study which are coarseness, contrast and directionality (3 features in total). Note that in this study we extracted the original (or standard) Tamura texture features. Since many malignant regions are more likely to have low contrast values [37], our initial hypothesis is that Tamura's contrast feature is more effective than the one extracted from the GLCM because Tamura contrast captures the variation of grey-level range and the polarisation of the distribution of black and white whereas GLCM contrast only captures the intensity variations within a $n \times m$ window [78].

Gradient features ***(F_5)***. There are many operators (e.g. Sobel filter, Kirsch filter, etc.) that could be used to extract these features. In this study we only selected the most discriminating ones according to the results by Niaf *et al.* [12], namely image numerical gradient at 0° and 90° orientations and image magnitude. Secondly, using Sobel operators we extracted image gradient at 0°, 90° and diagonal orientations and image magnitude (7 features in total). According to [53], gradient operators perform well in characterising

micro-textures as well as providing more consistent behavior as a descriptor of pathologies than co-occurrence matrices. In previous work [54], we used image magnitude as one of our texture descriptors to segment malignant regions.

Filter bank features ***(F_6)***. From a clinical point of view, most malignant regions show textural distortions in T2-W MRI [14, 55]. Litjens *et al.* [6] captured these characteristics in features using a Gaussian texture bank. However the conventional Gaussian texture bank is a) more sensitive to rotation (hence, rotated versions of malignant textures would be classified as non-malignant unless those rotated versions were included in the training set) and b) it does not incorporate spots/bars and edges. Therefore, we employed a filter bank as proposed by Varma and Zisserman [15] which is rotationally invariant and takes edges and spots/bars into account (38 features in total). The filter bank consist of an edge and a bar filter, at 6 orientations ($\theta = 0°, 30°, 60°, 90°, 120°, 150°$) and 3 scales ($(\sigma_x, \sigma_y) = (1, 3), (2, 6), (4, 12)$), Gaussian and Laplacian of Gaussian filters both with σ =10 pixels (in total 38 responses). Viswanath *et al.* [14] extracted texture features using a bank of Gabor filters but the results from their study showed that Haralick's features were more discriminant in capturing malignant regions within the PZ and features extracted from Gabor filters work better in detecting malignant regions within the CZ.

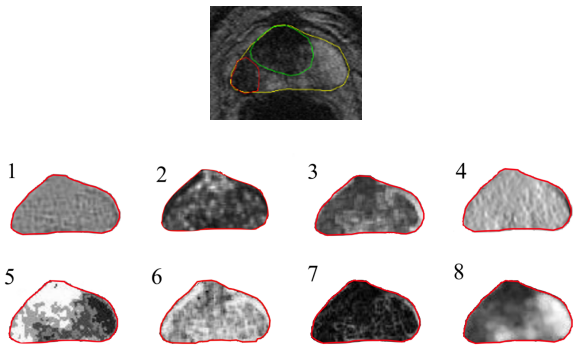

Figure 3: Laplacian of Gaussian (1), GLCM- energy (2), local contrast (3), gradient of Sobel operator (4), probability image (5), GLCM- correlation (6), image magnitude of Sobel operator (7) and upper quartile (8).

Figure 3 shows image responses of some of the features used in this study and Table 1 summarises the list of features used in this study which are divided into six categories.

Table 1: Summary of features used in this study.

Category	Features	Total
F_1:	Mean, median, standard deviation, mean and median of absolute deviation, skewness, kurtosis, mean of correlation coefficients, local contrast, variance and local probability	11
F_2:	GLCM features of Haralick *et al.* [38], Soh and Tsatsoulis [44] and Clausi [45] by taking 4 orientations and mean, variance and standard deviation of 4 orientations	154
F_3:	Grey-level lower and upper quartile	2
F_4:	Tamura's textures features namely coarseness, contrast and directionality	3
F_5:	Image numerical gradient (0° and 90° orientations), image magnitude, using Sobel operator image gradient (0°, 90° and diagonal orientations) and image magnitude.	7
F_6:	Filter bank of Varma and Zisserman [15] which contains an edge and a bar filter, at 6 orientations and 3 scales, a Gaussian and Laplacian of Gaussian filters	38

2.4. Feature Scaling and Feature Selection

Since we have 215 texture descriptors, feature selection is necessary to a) reduce over-fitting when building a classifier model as less data means less chance of making decisions based on noise, b) possibly improve accuracy because only the most relevant attributes are selected to build a classifier model (a study conducted by Niaf *et al.* [12] demonstrated that feature selection significantly improves the discrimination performance between malignant and benign regions) and c) reduce training time because fewer features are used in making decisions. Many feature selection methods have been developed in the literature [87]. However, the main focus of this paper is not to select the best feature selection method but to show the prospect of CAD using a single modality of T2-W MRI for prostate cancer diagnosis. Before feature selection is performed, we normalised each selected feature to avoid that absolute values play a role [56]. Following the suggestion in [56], each of the features was linearly scaled to the range [0,1] and the same was applied for the test data.

We employed the CfsSubsetEval [26] attribute evaluator and the GreedyStepwise search method in WEKA [25]. The CfsSubsetEval [26] method measures the value of a subset of features by considering each feature's predictive ability with the degree of redundancy between the other features within the subset, while the GreedyStepwise search method performs a greedy forward or backward search through the feature space [26]. Firstly, the dataset was seperated into training and testing sets (using 9-fold cross validation) and we performed feature selection based on the training set only. Subsequently, we

use the same selected features in the testing set. This process was repeated 81 times (9 runs in each fold) until the whole 9-fold cross validation is completed. Recently, Chen *et al.* [9] used the same feature selection method in classifying microcalcification clusters in mammograms.

2.5. Data Clasification

Table 2: List of classifiers used in our study. For more details please refer the default parameter settings in WEKA [25].

Classifiers	Summary of default parameters in WEKA
Support Vector Machines (SVM) [57]	SMO procedure with polynomial kernel
Simple Logistic (SL) [58]	Number of boosting iterations=0
Random Forest (RF) [59]	Number of trees=100
Multilayer Perceptron (MLP) [60]	Validation threshold=20, momentum=0.2
Naïve Bayes (NB) [62]	Without kernel estimator
Bayesian Networks (BNet) [62]	Hill climbing search algorithm
k-Nearest Neighbor (k-NN) [63]	k=1, Euclidean distance
C4.5 (also known as J48) [64]	Confidence factor=0.25
Alternating decision tree (ADTree) [65]	Number of boosting iterations=10
Meta-Vote (best 2) [66]	Combination rule=average probability
Meta-Vote (best 3) [66]	Combination rule=average probability

In this study we employed 11 different classifiers to achieve the best possible results using the WEKA data mining suite [25]. These classifiers were selected due to their robustness and popularity in CAD-PC [7]. Since our study evaluates a large number of classifiers, tuning parameters for each of the classifiers are time consuming and computationally expensive. Therefore, all parameters were left on default settings. The classifiers used in this study are presented in Table 2.5 (which includes their abbreviation). For the meta-voting classifiers we combined two or three prediction models which have the best 2 and 3 area under the curve (AUC) values based on the results produced after the training phase from any of the first nine classifiers in Table and use average probability as a combination rule.

3. Experimental Settings

3.1. Materials and Dataset

Our dataset consists of 418 T2-W MR images 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 malignant regions annotated cases were clinically significant cancer (Gleason score grade 7 and above). All patients underwent T2-W MR imaging at the Department of Radiology at the 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, and $0.5mm$ inter-slice gap.

3.2. Training and Testing

All pixels within the radiologist's tumor annotation were extracted as prostate cancer samples (e.g. within the red outlined region in Figure 2). This area was truncated by the tumor mask, to ensure no pixels outside the tumor region were included into the malignant samples. On the other hand, every pixel outside the tumor region and within the PZ (under the magenta line in Figure 2) was considered as benign samples. Similarly, this region is truncated by the tumor and 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 [91]. The evaluation was based at a patient level to ensure no samples from the same patient were used in the training and testing phases. We chose 9 folds instead of 10 folds to ensure each fold has the same number of patients (45 patients in our case, hence each fold contains 5 patients). Each classifier was trained and in the testing phase, each unseen instance/pixel from the testing data (taken from 5 randomly selected patients) was classified as malignant or non-malignant.

4. Experimental Results

In this study the experimental results are divided into four categories: performance based on classifiers, performance using different window sizes, performance evaluation before and after feature selection and feature evaluation (top 20 features).

4.1. Overall performance

This section presents the overall performance of the proposed method using different classifiers. The performance were measured using the most popular metrics in the literature: Area Under the Curve (A_z, also known as AUC), Classification Accuracy (CA), Sensitivity (Sen) and time taken for training and testing (t). 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. On the other hand, Sen measures the proportion of actual positives which are correctly identified (in this case the percentage of malignant pixels which are correctly identified). Sensitivity and accuracy can be calculated as $Sen = \frac{TP}{TP+FN}$ and $CA = \frac{TP+TN}{TN+TP+FP+FN}$, respectively. 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. The time taken by each classifier is measured in minutes to complete both training and testing in 9-FCV. On the other hand, p values indicate the significant difference for all metrics (A_z, CA and Sen) between the Meta-Vote classifier (best 2) in comparison to the other 10 classifiers.

Table 3: Overall performances using different classifiers using a 11×11 sliding window.

Classifiers	A_z (%)	CA (%)	Sen (%)	t
Meta-Vote (best 2)	**92.7 ± 7.4**	**85.5 ± 7.2**	**93.3 ± 9.1**	21.7
Meta-Vote (best 3)	92.3 ± 7.6	85.3 ± 7.6	92.7 ± 9.5	24.3
BNet	90.0 ± 7.6	83.5 ± 7.6	90.8 ± 9.5	3.83
ADTree	89.5 ± 8.9	83.7 ± 8.5	88.9 ± 11.6	53.48
RF	87.6 ± 9.3	84.7 ± 7.2	91.8 ± 7.1	19.22
MLP	87.4 ± 9.2	81.0 ± 9.9	86.5 ± 12.2	294.17
NB	86.9 ± 10.5	80.5 ± 8.7	90.9 ± 10	**1.23**
SL	85.6 ± 9.8	78.6 ± 9.5	80.1 ± 16.8	444.37
C4.5	79.3 ± 9.8	82.4 ± 6.8	86.7 ± 7.6	21.60
SVM	77.3 ± 9.9	78.5 ± 10.4	78.6 ± 17.2	154.98
k-NN	68.8 ± 10.1	72.0 ± 11.9	81.9 ± 9.2	116.78

Table 3 shows that overall performances for each of the classifiers employed in this study and Table 4 contains the p values for the three main metrics between the Meta-vote (best 2) classifier against the other 10 classifiers. The Meta-vote (best 2) classifier outperformed all classifiers in all metrics and statistically significant in terms of A_z with all individual classifiers (at least $p < 0.05$). The performance of meta-vote (best 2) classifier also significantly improved for CA and Sen in comparison to most of the results produced

Table 4: The p values for A_z, CA and Sen between the Meta-Vote classifier (best 2) against the other 10 classifiers from results in Table 3.

Classifiers	A_z	CA	Sen
Meta-Vote (best 2)	-	-	-
Meta-Vote (best 3)	0.4013	0.5517	0.3783
BNet	0.0436	0.1003	0.1020
ADTree	0.0322	0.1401	0.0228
RF	0.0040	0.2981	0.1922
MLP	0.0026	0.0135	0.0026
NB	0.0024	0.0023	0.1170
SL	<0.0001	<0.0001	<0.0001
C4.5	<0.0001	<0.0001	<0.0001
SVM	<0.0001	<0.0001	<0.0001
k-NN	<0.0001	<0.0001	<0.0001

by single classifiers. However, there is no significant difference to Meta-vote (best 3) classifier for A_z, CA and Sen. These results are similar in terms of CA for BNet, ADTree and RF. The results indicate that using several classifiers in making decision often produces better results due to their ability to handle complex data representation (or high dimensional data). For example, when the feature space dimension is large many possible hypotheses could be created by a single classifier to build a prediction model (in our case the number of training instances are round 150,000 instances). This increases the probability that the classifier cannot guarantee finding the best hypothesis and approximation boundary of the target classes [77]. Hence, there is a risk of selecting a hypothesis or class boundary with a low accuracy on unseen data [77]. However, using several classifiers in making a decision increases the chance of selecting the best hypothesis by combining and averaging decisions or class boundaries for a final decision [77].

Individually, the BNet classifier performed best with an A_z=90% followed by the ADTree, RF and MLP classifiers with an A_z=89.5%, 87.6% and 87.4%, respectively. BNet is expected to perform better than NB due to its ability of mapping the relationships among variables (or features) to build a predictive model without being restricted by the independence condition, whereas NB builds a predicitive model based on a certain condition between two variables. Unfortunately, the independence condition is not always met and this leads to a less accurate model. MLP produced good results as it shares similar property with BNet (models the relationships among the input layer (which contains features), the hidden layer (the neuron) and the output layer (the actual class prediction)). On the other hand, RF and ADTree produced promising results

due to their ability to perform like an ensemble classifier (consider various decisions and use averaging to improve predictive accuracy). In fact, ADTree employs boosting to improve the conventional decision tree algorithm (e.g. C4.5). Therefore ADTree produced better results in all three metrics compared to C4.5. In addition, RF and BNet contain efficient approaches for avoiding data over fitting. On the other hand, k-NN performed poorly in our study (A_z =67.5%). This may be caused by the number of neighbourhoods (k=1) as this restricts the algorithm to make decision based on a nearest single neighbourhood instead of based on several neighbours. Similarly, the low performance of SVM (A_z=76%) is expected to have been caused by default parameter settings. It is known that the performance of SVM is heavily affected by its parameters such as the choice of kernel, the kernel's parameters, and soft margin parameter C.

In terms of accuracy, there is still space for improvement (despite all predicitive models produced $CA > 70\%$) as it can be seen that none of the predictive models managed to achieve CA $> 90\%$. The highest accuracy presented in Table tab:overallResults is achieved by the predictive models built by meta-classifiers via a voting approach resulted in just above 85%. From a sensitivity point of view all predictive models achieved more than 80% except the model built by SVM. Both meta classifiers achieved more than 90% with acceptable time taken for training and testing below 25 minutes for 9-FCV. Individually, the predictive models built by BNet, RF and NB classifiers achieved more than 90% sensitivity in comparison to MLP (86.5%) and ADTree (88.9%), where other predictive models achieved reasonable sensitivities. On the other hand, the NB classifier is the fastest predictive model taking less than 2 minutes to complete 9-FCV followed by the BNet classifier, which took less than 4 minutes. The slowest predictive models are the ones built by SL, MLP and SVM classifiers which took 444.37, 294.17 and 154.98 minutes, respectively.

The performances among the classifiers vary greatly due to:

(i) Classifier variation. Each classifier has its own 'decision rules' to build a predictive model of the training set. This makes each classifier only suitable for specific tasks as discussed in [88]. Amancio *et al.* [89] compared 9 classifiers using real and artificial data and found that the accuracies among classifiers vary between 7%-20%. In a study of CAD-PC, Chan *et al.* [72] showed that the variations of the A_z values on three different classifiers are up to 0.24 (e.g. Linear Discriminant 0.84, SVM 0.76 and a single channel maximum-likelihood classifier 0.60). In another study, Litjens *et al.* [6] also showed a variation of up to 10% difference between the Random Forest (0.89) and Linear Discriminant

(0.79) classifiers.

(ii) Parameter settings. Some classifiers are highly dependent on their parameters. For example the SVM classifier is very sensitive to its parameters such as the C (complexity) value and the type of kernel (e.g. RBF/polynomial/linear). Similarly, the performance of the k-NN classifier is dependent on the number of neighbours. On the other hand, Random Forest and Bayesian Network classifiers are more robust and less dependent on their parameters.

(iii) The number of selected features. Amancio *et al.* [89] showed in their study that some classifiers produced better results with a smaller number of features and some classifiers worked better with a larger number of features. This is similar in our study as sometimes the feature selection method selected 25 features and sometimes it selected only 10 features.

4.2. Performance using different window sizes

The selection of window size (ws) is one of the major issues in image processing. Many studies (such as [12, 14, 39, 41, 72]) in CAD-PC did not report how window sizes affect the overall performance of their methods. In this study, we investigated the effects of ws on performance quantitatively. For this purpose we conducted nine experiments using the following ws: 3×3, 5×5, 7×7, 9×9, 11×11, 13×13, 15×15, 17×17 and 19×19. The results presented in Tables 5 and 6 show that ws affects the performance of the proposed method both in terms of A_z and CA values. In general using smaller ws such as 3×3 and 5×5 produced lower results compared to 9×9 and 11×11 due to insufficient information such as limited spatial information, limited intensities and grey level variations, the statistical values calculated from the neighbourhood are affected by noise, etc.

A study [80] suggests that large windows contain more information then small ones (hence, provides better texture characterisation). Moreover, several studies suggest that an appropriate guideline for window sizes are 7×7 and 9×9 [81, 82], where medium window sizes do not increase classification accuracy significantly [83] (in our case most CA and A_z values at 9×9 and 11×11 are very similar in Tables 5 and 6). Nevertheless, an absolute optimal size window is difficult to determine due to the complexity of the problem domain. As shown in Tables 5 and 6 the A_z values for all classifiers decreased after $ws = 11\times11$. A large ws causes the computed feature values within neighboring pixels over-represented the actual characteristic of the region. This can be be clearly seen in Tables 5 and 6 in both A_z and CA. The results change between 5%-10% at $ws = 3\times3$ and 19×19. The proposed method achieved

lower results at $ws = 3 \times 3$ due to insufficient information to characterise tumour regions. Similarly, at $ws = 19 \times 19$ the proposed method produced lowest results. According to Wolters *et al.* [79], the typical size ranges of malignant regions in prostate MRI are $5 - 20mm$ (on average $12.5mm$, using medium sizes ws (e.g. 9×9) are appropriate in our case since it is the average tumor size and $ws = 3 \times 3$ would be too small). For most classifiers there are small improvement (less than 2%) for both metrics using 9×9 and 11×11.

Table 5: The A_z values using different window sizes.

Classifiers	3×3	5×5	7×7	9×9	11×11	13×13	15×15	17×17	19×19
Meta-Vote (best 2)	89.3 ± 7.1	89.6 ± 7.3	90.6 ± 6.8	91.7 ± 6.8	**92.7 ± 7.4**	87.6 ± 6.9	86.3 ± 7.3	84.3 ± 6.9	83.3 ± 8.5
Meta-Vote (best 3)	88.9 ± 7.5	89.2 ± 7.6	90.3 ± 7.3	91.5 ± 6.9	**92.3 ± 7.6**	87.2 ± 6.7	84.5 ± 7.1	83.9 ± 7.1	81.3 ± 9.1
BNet	85.7 ± 9.2	87.1 ± 9.8	88.1 ± 8.5	89.9 ± 9.9	**90.0 ± 7.6**	82.7 ± 9.9	83.1 ±11.6	83.1 ± 13.3	82.3 ± 14.9
ADTree	82.4 ± 9.5	84.4 ± 9.8	85.8 ± 10.3	86.5 ± 10	**89.5 ± 8.9**	83.5 ± 11	82.4 ± 11.6	82.5 ± 12.5	78.6 ± 17
RF	83.2 ± 8.1	85.1 ± 8.6	85.7 ± 9.0	87.8 ± 9.2	**87.9 ± 9.3**	82.2 ± 9.8	82.2 ± 11.9	81.3 ± 12.7	76.8 ± 15.4
MLP	81.2 ± 9.3	83.4 ± 9.7	85.8 ± 8.6	**88.4 ± 7.7**	87.4 ± 9.2	76.2 ±11.4	78.2 ± 13.1	83.2 ± 12.7	78.4 ± 14.3
NB	74.7 ± 10.3	77.1 ± 10.9	83.3 ± 10.5	**87.3 ± 9.6**	86.9 ± 10.5	77.9 ± 13.7	78.9 ± 13.4	75.3 ± 10.7	74.8 ± 11.4
SL	79.9 ± 8.3	80.7 ± 8.8	83.2 ± 8.9	**86.2 ± 7.1**	85.6 ± 9.8	80.2 ± 9.7	80.3 ± 11.9	78.3 ± 11.4	75.9 ± 11.9
C4.5	72.6 ± 9.2	74.1 ± 8.9	74.7 ± 9.5	77.5 ± 10	**79.3 ± 9.8**	78.3 ± 10.4	76.7 ± 12.5	72.3 ± 12.4	72.3 ± 14.4
SVM	69.3 ± 7.8	71.5 ± 7.6	74.6 ± 8.7	76.2 ± 7.9	77.3 ± 9.9	**79.8 ± 10.3**	79.8 ± 11.9	75.3 ± 11.7	74.3 ± 12.9
k-NN	67.7 ± 7.7	68.5 ± 7.3	66.0 ± 8.2	67.5 ± 8.6	68.8 ± 10	**70.7 ± 15.2**	67.8 ± 17.7	71.9 ± 12.9	68.8 ± 11

Table 6: The CA values using different window sizes.

Classifiers	3×3	5×5	7×7	9×9	11×11	13×13	15×15	17×17	19×19
Meta-Vote (best 2)	79.5 ± 7.3	80.1 ± 6.8	82.8 ± 7.1	83.6 ± 7.5	**85.5 ± 7.2**	85.1 ± 8.7	84.7 ± 8.1	80.7 ± 8.4	78.5 ± 10.5
Meta-Vote (best 3)	78.9 ± 7.5	79.3 ± 7.3	82.5 ± 6.7	83.0 ± 8.1	**85.3 ± 7.6**	84.3 ± 8.3	84.1 ± 8.5	80.1 ± 8.1	77.3 ± 11.3
BNet	80.3 ± 6.1	81.4 ± 6.4	82.2 ± 6.6	82.6 ± 8.3	83.5 ± 7.6	**88.1 ± 9.8**	87.3 ± 11.5	82.3 ± 11.7	81.3 ± 15.2
ADTree	78.3 ± 6.1	79.4 ± 6.2	82.6 ± 6.3	82.7 ± 6.0	83.7 ± 8.5	**86.7 ± 11.6**	85.7 ± 13.7	83.1 ± 14.6	77.1 ± 18
RF	81.3 ± 5.9	82.3 ± 5.6	82.7 ± 6.0	**85.2 ± 5.3**	84.7 ± 7.2	83.1 ± 12.9	81.2 ± 15.4	84.3 ± 15.4	79.6 ± 18.7
MLP	77.5 ± 6.5	78.8 ± 6.3	79.4 ± 7.1	83.2 ± 6.3	81.0 ± 9.9	83.2 ± 11.5	**84.1 ± 13.9**	83.6 ± 14.7	80.7 ± 16.8
NB	65.7 ± 19.3	67.2 ± 20.4	67.7 ± 20.8	71.3 ± 19	80.5 ± 8.7	84.1 ± 11.9	**85.9 ± 11.6**	80.1 ± 12.7	73.8 ± 19.1
SL	70.1 ± 8.3	72.8 ± 8.2	76.4 ± 8.2	77.5 ± 9.1	78.6 ± 9.5	**86.9 ± 10.4**	86.7 ± 13.5	77.8 ± 15.8	77.9 ± 19
C4.5	77.2 ± 5.6	78.9 ± 5.4	79.4 ± 6.2	81.8 ± 6.1	**82.4 ± 6.8**	72.4 ± 13.7	72.2 ± 15.5	75.1 ± 17	73.5 ± 23.6
SVM	72.0 ± 8.5	72.2 ± 8.8	75.7 ± 9.2	77.6 ± 9.1	**78.5 ± 10.4**	77.9 ± 11.6	78.2 ± 12.5	75.1 ± 14	74.1 ± 16.5
k-NN	69.9 ± 5.3	72.5 ± 5.7	70.5 ± 8.5	70.3 ± 12.4	72.0 ± 11.9	67.1 ± 13.7	66.4 ± 13.9	**73.8 ± 14.6**	67.3 ± 15.5

Nevertheless, using different ws showed a significant difference for most classifiers in terms of A_z and CA. For the Meta-vote (best 2), increasing $ws = 3 \times 3$ to 11×11 increases CA value from 79.5 ± 7.3 to 85.5 ± 7.2 ($p < 0.0001$). However, some classifiers produced lower results when using a larger ws. For example the SVM classifier produced A_z=77.5% using 5×5 but

produced 3% and 1% lower at 7×7 and 11×11, respectively. Similar for k-NN in terms of accuracy we can see that its best performance is using 5×5 with CA =72.5%. Interestingly, some classifiers produced the best CA at $ws = 13 \times 13$ such as BNet and ADTree. Nevertheless, most classifiers produced poorer CA at $ws \geq 15 \times 15$. Therefore, based on these results using $ws = 9 \times 9$ or 11×11 is an appropriate guideline when selecting window size in our study.

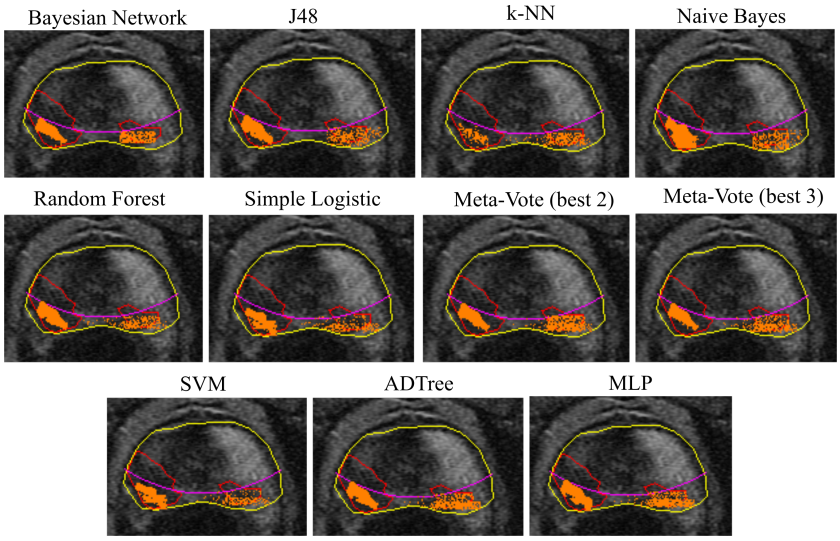

Figure 4: Segmentation results using different machine learning algorithms.

Figure 4 shows segmentation results using different machine learning algorithms employed in this study. There are two malignant regions within the PZ which both are detected and segmented. In this case, the Bayesian Network classifier outperformed the other individual classifiers. However, the Meta-Vote (best 2) combining the Bayesian Network and MLP classifiers produced the best A_z and CA. All classifiers managed to segment the first malignant region (left annotation in red) but a poor segmentation produced by the k-NN classifier. On the other hand, all classifiers produced false positives in detecting the second malignant region (right annotation in red) and the Meta-vote (best 2) classifier produced the highest sensitivity followed by the Bayesian Network. Overall, in this example the top three classifiers are the Bayesian Network, Naïve Bayes and MLP, accordingly.

4.3. Performances before and after feature selection

In a study by Niaf *et al.* [12], they showed that feature selection can deliver a significant improvement in a classification model's performance. Therefore, in this study we investigated the effects (both for A_z and CA) of feature selection

using an 11×11 window size. In the first experiment no feature was excluded and in the next experiment only features selected using CfsSubsetEval [26] were included in training and testing.

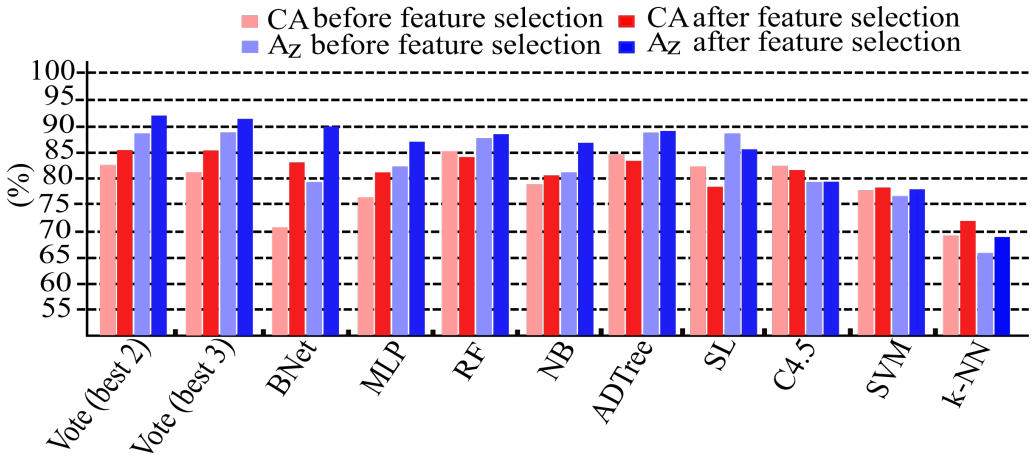

Figure 5: Performance comparisons among classifiers before and after feature selection using 11×11 window size

Figure 5 shows that there is a significant improvement in both A_z and CA ($p < 0.05$) for most classifiers after feature selection is performed due to different levels of data complexity. For example it is easier to build a predictive model in a lower dimensional dataset (e.g. 20 features per instance) than in a higher dimensional dataset (e.g. 215 features per instance). The difficulty to build a predictive model in a higher dimensional dataset increases the error rate (hence reduces predictive accuracy). However, the improvement for most classfiers is less than 3%, except BNet, MLP, SL and NB classifiers improvement of approximately 5% to 10% in both A_z and CA. For example, the significant difference for A_z values before and after feature selection are $p < 0.0001$, $p = 0.0292$ and $p = 0.0072$ for BNet, MLP and NB classifiers, respectively at 11×11. This may have been caused by two reasons:

(i) BNet, MLP and NB classifiers share similar concepts of building a predictive model by mapping the relationships among features. Increasing the number of features means building a more complex predictive model as a larger number of relationships can be created. The complexity increases error rates and decreases the accuracy of the model in making a prediction in unseen cases.

(ii) BNet, MLP and NB classifiers do not have an approach to avoid data over fitting which means even weak and uncorrelated features will be considered in building a classification model.

Therefore performing feature selection is expected to be benefical for BNet, MLP and NB classifiers as it decreases the level of complexity of the model. In comparison, for tree-based classifiers only strong features will be considered in building the predictive model even without performing feature selection (the classifier selects the most correlated features in building a classification model). This suggests that the performance of the RF and ADTree classifiers are less affected by feature selection. From the results shown in Figure 5, differences for RF and ADTree classifiers before and after feature selection are less than 2% for both A_z and C_A. In addition, ADTree employs boosting procedures (uses many weak hyphotheses to build a strong hypothesis) which often produce better results. Similarly, C4.5 is also a tree-based classifier which was not affected significantly by the feature selection (no significant difference $p = 0.4207$ for A_z at 11×11). On the other hand, the results produced by the Meta classifiers are mainly effected by the performance of the top three individual classifiers. In this study, either BNet, MLP, RF, NB and ADTree are among the top 3 classifiers combined in the Meta-Vote (based on the 3 best A_z values in the training phase).

4.4. Features evaluation

We also investigated feature performance individually. For this purpose we conducted experiments by ranking the top 20 features based on the number of selection (ns) over the number of runs in 9-FCV (81 runs in total in this study) at nine different ws. The maximum value of ns is 81. The higher the ns the more frequent the feature has been selected by CfsSubsetEval [26].

Table 7 (top left) shows the list of top 20 most discriminant features based on the ns value for $ws = 3 \times 3$. There were 65 features selected with minimum ns=1 using $ws = 3 \times 3$. Features such as Gaussian filter, Laplacian of Gaussian filter and image magnitude were always selected by CfsSubsetEval [26] in 9-FCV of 81 runs. F_2 and F_6 features (see Table 1) dominate the list at the smallest window where none of Tamura's features were selected. F_1 features such as mean, median, local probability and local contrast are also among the most popular features at $ws = 3 \times 3$. On the other hand, 64 features were selected at least once at $ws = 5 \times 5$. Results in top right show that all features selected at least 80 times in 9-FCV come from F_1, F_2, F_5 and F_6, followed by local probability and local contrast with ns =79. Other features in the top 20 are a bank of bar/spot filters (also from F_6) covering different scales and orientations. Most of the features selected at $ws = 5 \times 5$ are filter bank of Varma and Zisserman [15]. Increasing ws to 7×7 decreased the number of features selected to 59.

The results in the bottom left of Table 7 show that features from category

F_5 and F_6 remain in favour. Variance along with image magnitude of Sobel operator remain among the most discriminant features. Interestingly, the ns value for GLCM: sum of squares variance ($\theta = 135°$) dropped to 68, placing it 16^{th} in the ranking. However, it can been seen that new features appear to be listed in the top 20 such as variance of cluster prominences ($\theta = 0°, 45°, 90°, 135°$), Tamura contrast and kurtosis. Moving up to $ws = 9 \times 9$, Tamura's contrast reached its best performance (ns=81) as it can be seen in bottom right of the table. Similarly for image gradient of the Sobel operator, Gaussian filter and Laplacian of Gaussian filter remain among the top features with maximum ns value. The number of bar/spot filters has decreased and only one GLCM feature was selected in the top 20. New features such as upper quartile and edge filters appeared into the list while variance dropped its ns value from 81 to 75.

Table 8 (top left) shows the results using $ws = 11 \times 11$. The results reveal a similar pattern for Gaussian filter, Laplacian of Gaussian filter, image magnitude of Sobel operator, Tamura contrast, image magnitude and variance. Although some ns values for some of the features (such as local probability, image gradient of Sobel operator and local contrast) dropped by at least 10, they remain in favour among the top 20 most discriminant features out of 215 features extracted in this study. Edge filters which were selected at $ws = 9 \times 9$ are not listed at $ws = 11 \times 11$, instead 3 GLCM features based on the feature's variance of four orientations ($\theta = 0°, 45°, 90°, 135°$) were selected. Results in the top right and bottom left of the Table 8 show that ns values for Gaussian and Laplacian of Gaussian have dropped to 76 and 70, respectively. These features are becoming even less discriminant in larger ws as shown in Table 9.

Other features such as Tamura's contrast, image gradient, image magnitude and some edge filters remain consistent within the top 10. From these experimental results, it can be seen that the ws parameter affects the performance of the features. For example GLCM: sum of squares variance ($\theta = 135°$) appear to be in the top 20 list using smaller ws but did not perform well using larger ws (e.g 9×9 and 11×11). Similarly, Tamura's contrast was listed among the most discriminant feature using medium ws but performed poorly using smaller ws (5×5). In addition, some features performed well only using certain values of ws. For example Kurtosis appeared to be in the list only at $ws = 7 \times 7$, 13×13 and 15×15. This is similar to edge filters on specific scales and orientations. However, there are some features consistent without being affected by ws parameter. For example the Gaussian filter, Laplacian of Gaussian filter, image magnitude of the Sobel operator, image magnitude, local contrast, local probability and variance are always in the top 20.

Table 7: Top 20 most selected features using $ws = 3 \times 3$ up to 9×9.

$ws = 3 \times 3$		$ws = 5 \times 5$	
Features	ns	Features	ns
Gaussian filter, Laplacian of Gaussian filter, image magnitude	81	GLCM: sum of squares variance ($\theta = 135°$), Gaussian filter, Laplacian of Gaussian filter, bar/spot filter ($(\sigma_x, \sigma_y) = (1, 3)$, $\theta = 30°$), bar/spot filter ($(\sigma_x, \sigma_y) = (1, 3)$, $\theta = 0°$), edge filter ($(\sigma_x, \sigma_y) = (1, 3)$, $\theta = 90°$), image magnitude, image magnitude of Sobel operator, variance	81
Standard deviation, GLCM: sum of squares variance ($\theta = 135°$), GLCM: variance of homogeneities ($\theta = 0°, 45°, 90°, 135°$)	78	Image gradient ($\theta = 90°$), image gradient ($\theta = 0°$)	80
Local probability, local contrast, median	76	Local probability, local contrast	79
Variance, GLCM: Dissimilarity ($\theta = 45°$)	74	Bar/spot filter ($(\sigma_x, \sigma_y) = (1, 3)$, $\theta = 150°$), bar/spot filter ($(\sigma_x, \sigma_y) = (4, 12)$, $\theta = 90°$)	77
GLCM: variance ($\theta = 45°$), bar/spot filter ($(\sigma_x, \sigma_y) = (4, 12)$, $\theta = 150°$)	73	Bar/spot filter ($(\sigma_x, \sigma_y) = (2, 6)$, $\theta = 90°$), bar/spot filter ($(\sigma_x, \sigma_y) = (2, 6)$, $\theta = 60°$), bar/spot filter ($(\sigma_x, \sigma_y) = (4, 12)$, $\theta = 150°$)	76
Bar/spot filter ($(\sigma_x, \sigma_y) = (2, 6)$, $\theta = 0°$)	72	Bar/spot filter ($(\sigma_x, \sigma_y) = (2, 6)$, $\theta = 0°$)	75
Bar/spot filter ($(\sigma_x, \sigma_y) = (4, 12)$, $\theta = 90°$),bar/spot filter ($(\sigma_x, \sigma_y) = (1, 3)$, $\theta = 0°$), edge filter ($(\sigma_x, \sigma_y) = (1, 3)$, $\theta = 90°$)	70	Bar/spot filter ($(\sigma_x, \sigma_y) = (2, 6)$, $\theta = 30°$)	74
Image magnitude of Sobel operator, mean	69		
GLCM: Entropy ($\theta = 45°$), GLCM: Energy ($\theta = 90°$)	67		

$ws = 7 \times 7$		$ws = 9 \times 9$	
Features	ns	Features	ns
Gaussian filter, Laplacian of Gaussian filter, bar/spot filter ($(\sigma_x, \sigma_y) = (1, 3)$, $\theta = 0°$), bar/spot filter ($(\sigma_x, \sigma_y) = (1, 3)$, $\theta = 90°$), image magnitude of Sobel operator, image magnitude , variance	81	Gaussian filter, Laplacian of Gaussian filter, bar/spot filter ($(\sigma_x, \sigma_y) = (4, 12)$, $\theta = 0°$), image magnitude of Sobel operator, image gradient of Sobel operator ($\theta = 45°$), image magnitude , image gradient ($\theta = 90°$), Tamura contrast	81
Local contrast	80	Image gradient of Sobel operator ($\theta = 90°$), local probability	80
Bar/spot filter ($(\sigma_x, \sigma_y) = (1, 3)$, $\theta = 30°$)	77	Local contrast	79
Bar/spot filter ($(\sigma_x, \sigma_y) = (1, 3)$, $\theta = 150°$)	74	Bar/spot filter ($(\sigma_x, \sigma_y) = (1, 3)$, $\theta = 0°$)	78
Bar/spot filter ($(\sigma_x, \sigma_y) = (2, 6)$, $\theta = 0°$), local probability	73	Image gradient of Sobel operator ($\theta = 0°$)	76
GLCM: sum of squares variance ($\theta = 135°$)	68	Variance	75
GLCM: variance of cluster prominences ($\theta = 0°, 45°, 90°, 135°$)	67	Upper quartile	74
Bar/spot filter ($(\sigma_x, \sigma_y) = (1, 3)$, $\theta = 90°$)	60	Bar/spot filter ($(\sigma_x, \sigma_y) = (4, 12)$, $\theta = 45°$)	59
Tamura contrast	59	Image gradient ($\theta = 0°$)	42
Image gradient ($\theta = 0°$)	55	An edge filter ($(\sigma_x, \sigma_y) = (2, 6)$, $\theta = 30°$)	36
Bar/spot filter ($(\sigma_x, \sigma_y) = (2, 6)$, $\theta = 60°$)	54	An edge filter ($(\sigma_x, \sigma_y) = (2, 6)$, $\theta = 0°$)	29
Kurtosis	52	GLCM: standard deviation of sum of variances ($\theta = 0°, 45°, 90°, 135°$)	24

Table 8: Top 20 most selected features using different $ws = 11 \times 11$ up to 17×17.

$ws = 11 \times 11$

Features	ns
Gaussian filter, Laplacian of Gaussian filter, image magnitude of Sobel operator, Tamura contrast, image magnitude , variance	81
Bar/spot filter ($(\sigma_x, \sigma_y) = (2, 6)$, $\theta = 0°$), local probability	70
Bar/spot filter ($(\sigma_x, \sigma_y) = (1, 3)$, $\theta = 30°$)	63
Image gradient of Sobel operator ($\theta = 45°$)	54
Bar/spot filter ($(\sigma_x, \sigma_y) = (1, 3)$, $\theta = 150°$)	52
Bar/spot filter ($(\sigma_x, \sigma_y) = (1, 3)$, $\theta = 90°$)	51
Local contrast	50
GLCM: variance of autocorrelations ($\theta = 0°, 45°, 90°, 135°$)	48
GLCM: variance of sum of variances ($\theta = 0°, 45°, 90°, 135°$)	44
Upper quartile	43
GLCM: sum of variance ($\theta = 45°$)	41
An edge filter ($(\sigma_x, \sigma_y) = (1, 3)$, $\theta = 90°$)	36
GLCM: variance of cluster shades ($\theta = 0°, 45°, 90°, 135°$)	33
Image gradient ($\theta = 0°$)	30

$ws = 13 \times 13$

Features	ns
GLCM: sum of squares variance ($\theta = 135°$), bar/spot filter ($(\sigma_x, \sigma_y) = (1, 3)$, $\theta = 60°$), image gradient of Sobel operator ($\theta = 45°$), image magnitude, image gradient ($\theta = 90°$), Tamura contrast, variance and local contrast	81
Image magnitude of Sobel operator and local probability	78
Gaussian filter, Laplacian of Gaussian filter and image gradient ($\theta = 0°$)	76
GLCM: variance of cluster prominences ($\theta = 0°, 45°, 90°, 135°$) and upper quartile	59
An edge filter ($(\sigma_x, \sigma_y) = (1, 3)$, $\theta = 0°$) and an edge filter ($(\sigma_x, \sigma_y) = (1, 3)$, $\theta = 90°$)	46
An edge filter ($(\sigma_x, \sigma_y) = (1, 3)$, $\theta = 150°$) and kurtosis	41
An edge filter ($(\sigma_x, \sigma_y) = (1, 3)$, $\theta = 30°$) and an edge filter ($(\sigma_x, \sigma_y) = (2, 6)$, $\theta = 90°$)	37

$ws = 15 \times 15$

Features	ns
Image gradient of Sobel operator ($\theta = 45°$), image magnitude, image gradient ($\theta = 90°$), Tamura contrast, variance	80
Image magnitude of Sobel operator, GLCM: variance of sum of variance ($\theta = 0°, 45°, 90°, 135°$) and local probability	75
Gaussian filter, Laplacian of Gaussian filter, GLCM: variance of cluster prominences ($\theta = 0°, 45°, 90°, 135°$) and image gradient ($\theta = 0°$)	70
An edge filter ($(\sigma_x, \sigma_y) = (1, 3)$, $\theta = 0°$)	65
GLCM: sum of variance ($\theta = 0°$) and kurtosis	33
GLCM: sum of square variance ($\theta = 135°$)	29
An edge filter ($(\sigma_x, \sigma_y) = (2, 6)$, $\theta = 120°$)	26
GLCM: variance of sum of square variances ($\theta = 0°, 45°, 90°, 135°$), image gradient ($\theta = 90°$) and local contrast	24
Upper quartile	14

$ws = 17 \times 17$

Features	ns
GLCM: variance of sum of variances ($\theta = 0°, 45°, 90°, 135°$), variance, Tamura contrast, image gradient ($\theta = 0°$)	81
Image magnitude of Sobel operator and image magnitude	77
GLCM: variance of sum of square variances ($\theta = 0°, 45°, 90°, 135°$), local contrast	74
An edge filter ($(\sigma_x, \sigma_y) = (1, 3)$, $\theta = 60°$), image gradient of Sobel operator ($\theta = 90°$) and local probability	73
Gaussian filter, Laplacian of Gaussian and An edge filter ($(\sigma_x, \sigma_y) = (2, 6)$, $\theta = 120°$)	43
GLCM: autocorrelation (135°)	33
GLCM: variance of cluster prominences ($\theta = 0°, 45°, 90°, 135°$)	23
GLCM: variance of dissimilarities ($\theta = 0°, 45°, 90°, 135°$), Variance of autocorrelations ($\theta = 0°, 45°, 90°, 135°$), image gradient ($\theta = 90°$)	10

Table 9: Top 20 most selected features using different $ws = 19 \times 19$ and top 20 common features across different ws.

$ws = 19 \times 19$		Common features across different ws	
Features	ns	Features	total ns
Image gradient (0°), image magnitude and Tamura contrast	80	Image magnitude	723
GLCM: variance of sum of square variances ($\theta = 0°, 45°, 90°, 135°$), An edge filter ($(\sigma_x, \sigma_y) = (1, 3)$, $\theta = 60°$), variance and GLCM: variance of sum of variances ($\theta = 0°, 45°, 90°, 135°$)	70	Image magnitude of Sobel operator	687
Image gradient of Sobel operator ($\theta = 90°$) and image magnitude of Sobel operator	64	Local probability	668
Local probability	58	Variance	623
Local contrast	49	Gaussian filter and Laplacian of Gaussian	616
Image gradient of Sobel operator ($\theta = 0°$)	39	Local contrast	601
GLCM: variance of contrasts ($\theta = 0°, 45°, 90°, 135°$), Gaussian and Laplacian of Gaussian filters	22	Tamura contrast	543
GLCM: variance of dissimilarities ($\theta = 0°, 45°, 90°, 135°$)	17	Image gradient ($\theta = 0°$)	469
GLCM: variance of cluster prominences ($\theta = 0°, 45°, 90°, 135°$)	10	Image gradient ($\theta = 90°$)	403
		Bar/spot filter ($(\sigma_x, \sigma_y) = (1, 3)$, $\theta = 0°$)	310
GLCM: variance of autocorrelations ($\theta = 0°, 45°, 90°, 135°$) and An edge filter ($(\sigma_x, \sigma_y) = (2, 6)$, $\theta = 120°$)	9	GLCM: sum of square variance $\theta = 135°$)	308
		Bar/spot filter ($(\sigma_x, \sigma_y) = (2, 6)$, $\theta = 0°$)	290
		Image gradient of Sobel operator ($\theta = 45°$)	242
		Edge filter ($(\sigma_x, \sigma_y) = (1, 3)$, $\theta = 90°$)	233
		GLCM: variance of sum of variances ($\theta = 0°, 45°, 90°, 135°$)	226
		Bar/spot filter ($(\sigma_x, \sigma_y) = (1, 3)$, $\theta = 60°$)	224
		Bar/spot filter ($(\sigma_x, \sigma_y) = (1, 3)$, $\theta = 30°$)	221
		Image gradient of Sobel operator ($\theta = 90°$)	217
		Bar/spot filter ($(\sigma_x, \sigma_y) = (1, 3)$, $\theta = 150°$)	203

Our experimental results support an earlier study conducted by Kovalev *et al.* [53] who claimed that image magnitude and gradient have a more consistent behavior as a descriptor compared to GLCM features. Tamura contrast, local probability, variance and upper quartile are among the most promising features but they need to be used using particular window sizes (e.g. 9×9 and 11×11). This is simialar to the other features such as Gaussian and Laplacian of Gaussian (best at $ws \leq 13 \times 13$). Table 9 (right side) presents all the common features across different ws based on the total ns. The image magnitude has the highest $ns = 723$ (maximum $ns = 729$) followed by the image magnitude of Sobel operator ($ns = 687$). This indicate that these features are fairly consistent regardless to the ws (similar to the study claimed by Kovalev *et al.* [53]).

Other features such as local probability, local contrast, Tamura contrast, Gaussian and Laplacian of Gaussian filters also consistent and less dependent to the ws. On the other hand, the GLCM features are among the texture descriptors which are heavily dependent on the ws together with the bar/spot filters.

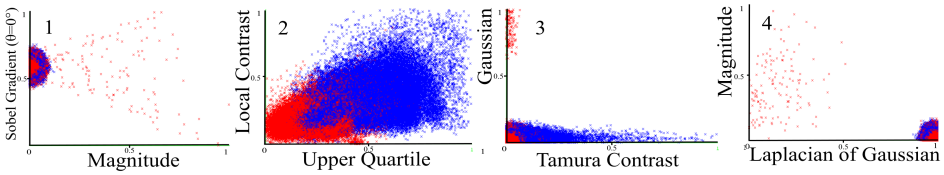

Figure 6: Pairwise scatterplots for four pairs of features. (1) Gradient of the Sobel operator $\theta = 0°$) versus magnitude, (2) local contrast versus upper quartile, (3) Gaussian versus Tamura contrast feature and (4) magnitude versus Laplacian of Gaussian (each graph represents 152, 445 instances taken using 9×9 sliding window).

Figure 6 shows the distribution of features in 2D feature space. Most malignant and benign data are quite separated in Gaussian versus Tamura contrast feature spaces. For image 2, although there are fairly large amount of data overlapping in local contrast versus upper quartile the number of data separated are still quite noticeable.

5. DISCUSSION

A successful CAD system can improve clinical dignostic decision making by acting as a 'second opinion' reducing the possibility of expert/non-expert missing cancerous regions. There are many factors which influence the radiologist's performance such as fatigue due to the large amount of MRI images needing to be analysed, the appearances of some malignant regions are obscure due to noise or unwanted artifacts and different levels of experience [7,11,12,31]. CAD can overcome these problems by speeding up the process of decision making regardless of the number of images and delineating cancerous regions which could not be detected via human visual perception or less experienced radiologists. Although the proposed method achieved $A_z = 93\%$, we would like to emphasise that this does not indicate that our method is better in comparison to CAD-PC based on multiparametric MRI. We are aware that studies [6, 13, 71] based on multiparametric MRI have shown that combining different features from different modalities can increase the CADs performance. In fact, the number of patients covered in this study does not allow us to draw

a conclusion whether the results are consistent or not when tested with a larger dataset.

However, our study suggests a set of discriminant texture descriptors extracted from T2-W modality which can be used in the development of CAD-PC. These descriptors can be combined with other features from the other modalities such as DWI, MRS and DCE. On the other hand, the results of our study indicate the possibility of developing a multi-scale classifier which might be better performing overall. In our study using $ws = 9 \times 9$ and 11×11 produced the best results.

5.1. Qualitative Comparisons

We would like to emphasis that a full quantitative comparison is impossible due to:

(i) Differences in datasets (different modalities such as T2-weighted (T2-W) MRI, diffusion-weighted (DWI) MRI, dynamic contrast enhanced (DCE) MRI, Magnetic resonance spectroscopy (MRS), etc) and frameworks used in the other studies.

(ii) Absence of public datasets also makes a quantitative comparison of methodologies in the literature difficult. Each team of researchers has their own datasets which cause a huge range of variability in terms of noise and image quality.

(iii) Studies were conducted within different regions of the prostate. For example, some studies were conducted within the prostate PZ only and some took the whole prostate gland into account.

(iv) Evaluation has been at volume, slice, regions or voxel level.

Table 10 shows a qualitative comparisons with the state-of-the-art in the literature. The methods proposed by Vos *et al.* [75] and Lv *et al.* [68] achieved the highest accuracy of $A_z = 97\%$. Vos *et al.* [75] 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 region of interests (ROI) taken from 29 patients. In an earlier study, using similar features Vos *et al.* [75] reported an A_z =92% based on 90 ROI taken from 34 patients. Later Vos *et al.* [90], reported a lower A_z =83% based on a larger number of ROI (6227) extracted from 200 patients. On the other hand, Lv *et al.* [68] 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 selected 130

Table 10: Qualitative comparisons with existing methods. Dynamic Contrast Enhanced (DCE), Diffusion Weighted (DW) and Magnetic Resonance Spectroscopy (MRS).

Authors	Patients	Studied Zones	Modality (MRI)	A_z (%)
Vos *et al.* (2010) [75]	29	PZ only	T2-W + DCE	97
Lv *et al.* (2009) [68]	55	PZ only	T2-W	97
Peng *et al.* (2013) [69]	48	PZ and CZ	T2-W + DCE + DW	95
Our method	**45**	**PZ only**	**T2-W**	**93**
Lopes *et al.* (2011) [70]	17	PZ only	T2-W	93
Vos *et al.* (2008) [52]	34	PZ only	T2-W + DCE	92
Tiwari *et al.* (2012) [39]	36	PZ and CZ	T2-W + MRS	90
Niaf *et al.* (2012) [12]	30	PZ only	T2-W+ DW + DCE	89
Tiwari *et al.* (2013) [71]	29	PZ and CZ	T2-W + MRS	89
Litjens *et al.* (2014) [6]	347	PZ and CZ	T2-W + DCE + DW	89
Viswanath *et al.* (2012) [14]	22	PZ and CZ	T2-W	86
Chan *et al.* (2003) [72]	15	PZ only	T2-W + DW	84
Vos *et al.* (2012) [90]	200	PZ and CZ	T2-W + DW	83
Puech *et al.* (2009) [74]	100	PZ and CZ	DCE	77
Langer *et al.* (2009) [73]	25	PZ only	T2-W + DCE + DW	71

ROI of 12×12 pixels (which means only a small part of the PZ region was covered). In fact, Lv *et al.* [68] 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. In fact, since the proposed is based on pixel-based classification it means the number of instances are around 152,000 (in comparison the number of instances used in [75] and [68] are 90 and 130, respectively).

Peng *et al.* [69] reported A_z =95% using three modalities of T2-W, DCE, and DW-MRI and extracted 10th percentile 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 [68, 75]. Although Peng *et al.* [69] reported that their study covered 48 patients, the actual evaluation was based on 104 ROI (61 malignant ROIs, 43 normal ROIs). In a smaller study of 17 patients, Lopes *et al.* [70] concluded that classical texture features (such as Haralick, wavelet, and Gabor filters) are less discriminant in classifying malignant and benign regions in comparison to fractal and multifractal features. In their study, they combined fractal and multifractal features and employed SVM and AdaBoost classifiers to get A_z =93% in comparison to the combination of classical texture features (A_z =88%).

Niaf *et al.* [12] extracted 140 texture features from 180 ROIs (30 patients) and achieved A_z =89% which is similar to the methods in [6, 12, 73]. Niaf *et al.* [12] 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. In contrast, the methods in [6, 12, 73] attempted to cover the whole prostate gland. Recently, Litjens *et al.* [6] 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 is 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.* [71] which is based on Multi-kernel graph embedding in T2-W and MRS produced A_z=89% covering 29 patients. The method [71] 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. Subsequently, Tiwari *et al.* [39] 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.

A study by Viswanath *et al.* [14] attempted to differentiate the textural characteristics of malignant regions within the CZ and PZ. They extracted 110 texture descriptors and reported that Haralick's features (e.g sum entropy and difference average) achieved A_z=73% in differentiating cancer regions within the PZ, whereas Gabor features performed better within the CZ (A_z=73%). In contrast to the study conducted by Chan *et al.* [72], who introduced a multichannel statistical classifier and applied it in prostate malignancy detection. The proposed method achieved a reasonable A_z=84% considering their texture features are only based on image intensity obtained from T2-W

and DW MRI. The proposed methods of Puech *et al.* [74] and Langer *et al.* [73] achieved A_z <80% covering 100 and 25 patients, respectively. Puech *et al.* [74] used a scoring algorithm approach as part of their CAD system. The scoring algorithm was firstly developed based on median and maximum wash-in and wash-out slope values and they used it to assign a malignancy likelihood score for each of the ROIs. Langer *et al.* [73] combined three MRI modalities via several image fusion techniques and achieved a reasonable A_z=71% based on 25 patients.

5.2. Qualitative Comparison With Human Performance

Niaf *et al.* [12] reported preliminary results for radiologists performances for two and 15 years experience with A_z=80% and 86%, respectively. Litjens *et al.* [6] reported the performances of 10 radiologists was between A_z=81% to 83%. From these studies, we have a general idea of what typical human performance is in comparison with CAD-PC. In our case, the proposed method achieved A_z=93% which is significantly better (p <0.001) than human performance and some CAD systems based on multimodality MRI [12, 39, 52, 73].

5.3. Study Limitations

The limitations of our study are: firstly, we are unable to compare our results quantitatively with existing methods in the literature mainly due to the absence of public datasets. Every group of researchers has their own datasets which are not publicly available. This is currently one of the major issues in CAD-PC causing most studies in the literature to make only qualitative comparisons. This issue also limits the interpretation and meaning of the results with respect to clinical utility. However, a recent study based on 347 patients conducted by Litjens *et al.* [6] indicated the feasibility of developing CAD systems which are able to discriminate between malignant and normal regions. Secondly, we are unable to compare our results quantitatively with the actual prospective clinical performance due to absence of radiologist performance for our dataset. Nevertheless, the studies in [6,12] suggest that a radiologist performance typically ranges between A_z=80% to 86%. Although these values are based on their datasets, this roughly indicates that CAD-PC have the potential to assist radiologists as a second or first reader setting [6]. Thirdly, since our study employed 11 different classifiers to get the best possible results, performing parameter optimisation for each of the classifiers is time consuming and computationally expensive (therefore all parameters in this study were left on the default settings in WEKA). Feature performance results are based on only one feature selection method (CfsSubsetEval [26]).

Therefore the results may be different using different feature selection methods which could be investigated in future work. However, from our experimental results we achieved several similarities with studies in [12,53] in terms of feature consistency and performance particularly for F_5 features. Finally, only voxels within the malignant regions which are visible in T2-W MRI were considered as cancerous voxels. This means voxels within the cancerous regions which did not appear in T2-W were considered normal. This issue could be addressed by such obscure cancer regions in DWI Apparent Diffusion Coefficient (ADC) imaging which was not performed in our study. However this part of our future research work (see below).

5.4. Future work

This study will be extended to cover the whole prostate gland using additional modalities such as DWI and DCE. Differentiating malignant regions within the CZ is more challenging due to the intensities being very similar between the CZ and the malignant regions [76]. Although the study of Viswanath *et al.* [14] reported significant differences between malignant regions within the CZ and PZ, there are limited studies which have been conducted to differentiate textural characteristics between malignant and benign regions within the CZ. Secondly, although many existing methods [6, 12, 39, 52, 68, 71, 73] did not perform parameter optimisation for the classifiers they have used, parameter selection is one of the most important steps in CAD-PC development. Therefore, performing parameter optimisation for each of the classifiers used in this study could be investigated as future work.

6. CONCLUSIONS

The challenges of developing CAD-PC remain open due to its complexity and limitations both in single and multimodality imaging. Whether using a single modality, image fusion or using clinical features, none of these methods provide superior results. Therefore developing CAD-PC detection and localisation remains a challenge and there is still space for improvement both in A_z and CA. In conclusion, we have presented a novel method for prostate cancer detection or localisation within the PZ using the single modality of T2-W MRI. Performance evaluation shows that despite the limitations of T2-W MRI, the proposed method achieved similar results with existing methods in the literature, although the comparison was made qualitatively mainly due to different evaluation datasets. We are aware that T2-W MRI alone is insufficient in developing a more robust CAD-PC system as multimodality MRI can provide more informative data (e.g. physiological tissue characteristics

and metabolites composition) which are not available in a conventional MRI. Nevertheless, this study identifies a set of discriminant texture descriptors which can be combined with features from the other modalities, hence provide a solid basis for a CAD-PC based on multimodality. In this study we have shown that:

(i) CAD-PC systems based on single modality T2-W MRI could achieve similar results to those based on multimodality MRI. However, we would like to emphasise to the reader that studies [6, 13, 71] based on multiparametric MRI have shown that combining different features from different modalities can increase the CADs performance. As such, it is expected that our T2-W approach would form a good starting point for a modality MRI based system.

(ii) Combining different classifiers produces better results in A_z, especially when dealing with high dimensional data.

(iii) In this study feature selection improved the performance of the developed CAD-PC system. This further supports an earlier study conducted by Niaf *et al.* [12].

(iv) This study suggests a set of discriminant texture descriptors in the development of CAD-PC.

Therefore, although there is still space for improvement in the development of CAD-PC, this study further supports the potential of CAD-PC systems to be an invaluable tool to assist radiologists as mentioned in existing studies [6, 12, 71, 84].

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.

References

[1] R. Siegel, J. Ma, Z. Zou, and A. Jemal. Cancer statistics, 2014. *CA: A Cancer Journal for Clinicians*, vol.64(1), pp. 9–29, 2014.

[2] http://prostatecanceruk.org/information/prostate-cancer-facts-and-figures/, accessed 26-April-2015.

[3] N. Howlader, A. M. Noone, M. Krapcho, J. Garshell, N. Neyman, S. Altekruse, C. Kosary, M. Yu, J. Ruhl,Z. Tatalovich, A. Cho, H. Mariotto, D. Lewis, H. Chen, E. Feuer, and K. Cronin. (2013) Seer cancer Statistics Review,1975-2010, National

cancer Institute. http://seer.cancer.gov/archive/csr/1975_2010/, accessed 16-May-2015.

[4] R. Chou, J. M. Croswell, T. Dana, C. Bougatsos, I. Blazina, R. Fu, K. Gleitsmann, H. C. Koenig, C. Lam, A. Maltz, J. B. Rugge and K. Lin. A Review of the Evidence for the U.S. Preventive Services Task Force, [October 2010] http://www.uspreventiveservicestaskforce.org/uspstf12/prostate/prostateart.htm accessed 15-April-2015.

[5] 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. Screening and prostate-cancer mortality in a randomized European study. New England Journal of Medicine, vol. 360(13), pp. 1320-1328, 2009. doi: 10.1056/NEJMoa0810084. http://dx.doi.org/10.1056/NEJMoa0810084. PMID: 19297566.

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

[7] 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*, vol. 60, pp. 8-31, 2015.

[8] N. Karssemeijer, J. D. M. Otten , H. Rijken and R. Holland. Computer aided detection of masses in mammograms as decision support, *The British Journal of Radiology*, vol. 79, pp.S123-S126, 2006.

[9] Z. Chen, H. Strange, A. Oliver, E. R. Denton, C. Boggis and R. Zwiggelaar. Topological modeling and classification of mammographic microcalcification clusters. IEEE Transactions Biomedical Engineering, vol. 62(4) pp.1203-1214, 2015.

[10] 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. CTcolonography computer-aided polyp detection: Effect on radiologist observers of polyp identification by CAD on both the supine and prone scans, *Academic Radiology*, vol. 17, pp.948-959, 2010.

[11] 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), pp. 2986–2994, 2012.

[12] E. Niaf, O. Rouviere, F. Mege-Lechevallier, F. Bratan and C.Lartizien. Computer-aided diagnosis of prostate cancer in the peripheral zone using multi-parametric MRI. *Physics in Medicine and Biology*, vol. 57, pp. 3833-3851, 2012.

[13] 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. In *Proceedings SPIE Medical Imaging*, vol. 7963, 2011.

[14] S.E. Viswanath, N. B. Bloch, J. C. Chappelow, R. Toth, N. M. Rofsky, E. M. Genega, R.E. Lenkinski and A. Madabhushi. Central gland and peripheral pone prostate tumors have significantly different quantitative imaging signatures on 3 Tesla endorectal, invivo T2-weighted MR imagery, *Journal of Magnetic Resonance*

Imaging, vol. 36(1), pp.213–224, 2012.

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

[16] H. Tamura S. Mori, and Y. Takashi. Textural features corresponding to visual perception. *IEEE Transaction on Systems, Man, and Cybernetics*, vol. 8(6), pp. 460–472, 1978.

[17] S. Ozer, D. L. Langer, X. Liu, M. A. Haider, T. H. van der Kwast, A. J. Evans, Y. Yang, M. N. Wernick and I. S. Yetik. Supervised and unsupervised methods for prostate cancer segmentation with multispectral MRI. *Medical Physics*. vol. 37(4), pp.1873–1883, 2010.

[18] A. Rampun, P. Malcolm and R. Zwiggelaar. Detection and localisation of prostate cancer within the peripheral zone using scoring algorithm. In *Proceedings 16th Irish Machine Vision and Image Processing*, pp. 75–80, 2014.

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

[20] 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), pp. 2986–2994, 2011.

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

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

[23] A. Madabhushi, J. K. Udupa and G. Moonis. Comparing MR image intensity standardization against tissue characterizability of magnetization transfer ratio imaging. *Journal of Magnetic Resonance Imaging*, vol. 24(3), pp. 667–75, 2006.

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

[25] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, vol. 11(1), pp.10–18, 2009.

[26] M. A. Hall. Correlation-based feature selection for discrete and numeric class machine learning. In *Proceedings 17th International Conference of Machine Learning*, pp. 359–366, 2000.

[27] Y. Wang, G. W. Wei and S. Yang. Partial differential equation transform - Variational formulation and Fourier analysis. *International Journal for Numerical Methods in Biomedical Engineering*, vol. 27(12), pp. 1996–2020, 2011.

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

[29] 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), pp. 2032–2046, 2009.

[30] D. Ampeliotis, A. Anonakoudi, K. Berberidis and E. Z. Psarakis. Computer aided detection of prostate cancer using fused information from dynamic contrast enhanced and morphological magnetic resonance images. *IEEE International*

Conference on Signal Processing and Communications, pp.888–891, 2007.

[31] E. J. Halpern, D. L. Cochlin, and B. Goldberg, *Imaging of the Prostate*. London, UK: Martin Dunitz Ltd., 2002.

[32] M. B. Garnick, A. MacDonald, R. Glass, and S. Leighton, Harvard Medical School 2012 *Annual Report on Prostate Diseases*. Harvard , US: Harvard Medical School, 2012.

[33] D. T. Ginat, S. V. Destounis, R. G. Barr, B. Castaneda, J. G. Strang, and D. J. Rubens. Us elastography of breast and prostate lesions, *Radiographics*, vol. 29 (7), pp. 2007–2016, 2009.

[34] H. Choi, E. Loyer, H. Kaur and P. M. Silverman. Imaging neoplasms of the abdomen and pelvis. In: Kufe DW, Pollock RE, Weichselbaum RR, et al., editors. Holland-Frei cancer Medicine. 6th edition. Hamilton (ON): BC Decker; 2003. Chapter 36d. Available from: http://www.ncbi.nlm.nih.gov/books/NBK13197/

[35] S. S. Taneja. Imaging in the diagnosis and management of prostate cancer, *Reviews in Urology*, vol. 6 (3), p. 101,2004.

[36] S. B. Edge, D. R. Byrd, C. Compton, A. G. Fritz, F. L. Greene, and A. Trotti (2010). AJCC cancer Staging Manual (7th Edition). Springer., Chicago, US.

[37] S. S. Mohamed, E. F. El-Saadany, A. Fenster, D. B. Downey and K. Rizkalla. Region of interest identification in prostate TRUS images based on Gabor filter. *IEEE 46th Midwest Symposium on Circuits and Systems*, vol. 1, pp. 415–419, 2003.

[38] R. M. Haralick, K. Shanmugam and I. Dinstein. Textural features of image classification. *IEEE Transactions on Systems, Man and Cybernetics*, vol. 6, pp. 610–621, 1973.

[39] 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*, vol. 25, pp. 607–619, 2012.

[40] B. Julesz. Experiments in the visual perception of texture. *Scientific American*, vol. 232, pp. 34–43, 1975.

[41] A. Madabhushi, M. D. Feldman, D. N. Metaxas, J. Tomaszeweski, D. Chute. Automated detection of prostatic adenocarcinoma from high-resolution ex vivo MRI. *IEEE Transaction Medical Imaging*, vol. 24(12), pp. 1611–1625, 2005.

[42] C. M. Moore, A. Ridout and M. Emberton. The role of MRI in active surveillance of prostate cancer. *Current Opinion in Urology*, vol. 23(3), pp. 261–267, 2003.

[43] L. de O.Bastos, P. Liatsis and A. Conci. Automatic texture segmentation based on k-means clustering and efficient calculation of co-occurrence features. *International Conference on Systems, Signals and Image Processing (IWSSIP) 2008*, pp.141–144, 2008.

[44] L. Soh and C. Tsatsoulis. Texture analysis of SAR sea ice imagery using gray level co-occurrence matrices. *IEEE Transactions on Geoscience and Remote Sensing*, vol. 37(2), pp. 780–795, 1999.

[45] D. A. Clausi. An analysis of co-occurrence texture statistics as a function of grey level quantization. *Canadian Journal of Remote Sensing*, vol. 28(1), pp. 45–62, 2002.

[46] J. R. Ferguson. Using the grey-level co-occurrence matrix to segment and classify radar imagery. ProQuest, Information and Learning Company, Ann Arbor, MI, US, 1 edition, 2007

[47] A. Rampun, H. Strange and R. Zwiggelaar. Texture segmentation using different

orientations of GLCM features. In *Proceedings 6th International Conference on Computer Vision/Computer Graphics Collaboration Techniques and Applications (MIRAGE)*. pp. 519–528, 2013.

[48] D. J. A. Margolis. Multiparametric MRI for localized prostate cancer: lesion detection and staging. *BioMed Research International*, vol. 2014, Article ID 684127, 11 pages, 2014. doi:10.1155/2014/684127.

[49] O. Aki, E. Sala, C. S. Moskowitz, K. Kuroiwa, N. M. Ishill, D. Pucar, P. T. Scardino and H. Hricak. Transition zone prostate cancers: features, detection, localization, and staging at endorectal MR imaging. *Radiology*, vol. 239(3), pp. 784–792, 2006.

[50] C. Brossner, A. Winterholer, M. Roehlich, E. Dlouhy-Schutz, V. Serra, M. Sonnleithner, K. H. Grubmuller, K. Pummer and E. Schuster. Distribution of prostate carcinoma foci within the peripheral zone: analysis of 8,062 prostate biopsy cores. *World Journal of Urology*, vol. 21(3), pp. 163-166, 2003.

[51] E. Langford. Quartiles in Elementary Statistics. *Journal of Statistics Education*, vol. 14(3), 2006.

[52] P. C. Vos, T. Hambrock, C. A. Hulsbergen-van de Kaa, J. J. Futterer, J. O. Barentsz and H. J. Huisman. Computerized analysis of prostate lesions in the peripheral zone using dynamic contrast enhanced MRI. *Medical Physics*, vol.35(3), pp. 888–899, 2008.

[53] V. Kovalev, M. Petrou and Y. Bondar. Texture anisotropy in 3-D images. *IEEE Transaction Image Processing*, vol. 8 (3), pp. 34–43, 1999.

[54] A. Rampun, Z. Chen and R. Zwiggelaar. Detection and localisation of prostate abnormalities. In *Proceedings 3rd International Conference on Computational Mathematical Biomedical Engineering (CMBE13)*, pp. 205–208, 2013.

[55] 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.

[56] C. Hsu, C. Chang, and C. Lin. A practical guide to support vector classification, 2010.

[57] J. C. Platt. Fast training of support vector machines using sequential minimal optimization, Advances in kernel methods: support vector learning, *MIT Press*, Cambridge, MA, 1999

[58] N. Landwehr, M. Hall and E. Frank. Logistic model trees. *Machine Learning*, vol. 59 (1-2), pp. 161–205, 2005.

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

[60] E. B Baum. On the capabilities of multilayer perceptrons. *Journal of Complexity*, vol. 4 (3), pp. 193–215, 1988.

[61] G. H. John and P. Langley. Estimating continuous distributions in Bayesian classifiers. In *11th Conference on Uncertainty in Artificial Intelligence*, San Mateo, pp. 338–345, 1995.

[62] N. Friedman, D. Geiger and M. Goldszmidt. Bayesian network classifiers. *Machine Learning*, vol. 29(2-3), pp. 131–163, 1997.

[63] D. Aha and D. Kibler. Instance-based learning algorithms. *Machine Learning*, vol. 6, pp. 37–66, 1991.

[64] J. R. Quinlan. C4.5: Programs for Machine Learning. Morgan Kaufmann, San Francisco, CA, 1993.

[65] Y. Freund and L. Mason. The alternating decision tree learning algorithm. In *Proceeding of the Sixteenth International Conference on Machine Learning*, Bled,

Slovenia, pp. 124–133, 1999.

[66] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20(3), pp. 226–239, 1998.

[67] V. Vapnik and A. Lerner. Pattern recognition using generalized portrait method. *Automation Remote Control* vol.24, pp. 774–780, 1964.

[68] 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), pp. 161–168, 2009.

[69] 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 cancer and normal tissue and correlation with gleason score-A Computer-aided diagnosis development Study. *Radiology*, vol. 267(3), pp. 787–796, 2013.

[70] R. Lopes, A. Ayache, N. Makni, P. Puech, A. Villers, S. Mordon, N. Betrouni. Prostate cancer characterization on MR images using fractal features. *Medical Physics*, vol. 38(1), pp. 83–95, 2011.

[71] 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), pp. 219–235, 2013.

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

[73] D. L. Langer, T. H. van der Kwast, A. J. Evans, J. Trachtenberg, B. C. Wilson and M. A. Haider. Prostate cancer detection with multi-parametric MRI: Logistic regression analysis of quantitative T2, diffusion-weighted imaging, and dynamic contrast-enhanced MRI. *Journal of Magnetic Resonance Imaging*, vol. 30(2), pp. 327–334, 2009.

[74] P. Puech, N. Betrouni, N. Makni, A. Dewalle, A. Villers and L. Lemaitre. Computer-assisted diagnosis of prostate cancer using DCE-MRI data: design, implementation and preliminary results. *International Journal of Computer Assisted Radiology and Surgery*, vol. 4(1), pp. 1–10, 2009.

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

[76] Y. J. Choi, J. K. Kim, N. Kim, K. W. Kim, E. K. Choi and K. S. Cho. Functional MR imaging of prostate cancer. *Radiographics*, vol. 27(1), pp.63–75, 2007.

[77] T. G. Dietterich. Ensemble learning. The handbook of brain theory and neural networks 2, M.A. Arbib (Ed.) (Cambridge, MA: The MIT Press), pp. 110–125, 2002.

[78] P. Howarth and S. Ruger. Evaluation of texture features for content-based image retrieval. In *Proceedings of the International Conference on Image and Video Retrieval*, pp. 326–334, 2004.

[79] T. Wolters, M. J. Roobol, P. J. van Leeuwen, R. C. van den Bergh, R. F. Hoedemaeker, G. J. van Leenders, F. H. Schroder, and T. H. van der Kwast. A critical analysis of the tumor volume threshold for clinically insignificant prostate cancer using a data set of a randomized screening trial. Journal of Urology, vol.

185(1), pp. 121–125, 2010.

[80] D. Puig and M. A. Garcia. Determining optimal window size for texture feature extraction methods. In *proceedings of the IX Spanish Symposium on Pattern Recognition and Image Analysis*, pp. 237–242, 2001.

[81] P. Gong and P. J. Howarth. Frequency-based contextual classification and gray-level vector reduction for land-use identification. *Photogrammetric Engineering and Remote Sensing*, vol.58(4), pp. 423–437, 1992.

[82] J. R. Jensen. Introductory digital image processing a remote sensing perspective. *Photogrammetric Engineering and Remote Sensing*. Prentice Hall, New Jersey (1996).

[83] M. E. Hodgson. What size window for image classification? A cognitive perspective. *Photogrammetric Engineering and Remote Sensing*, vol. 64(8), pp. 797–807, 1998.

[84] Y. Zhu, S. Williams and R. Zwiggelaar. Computer technology in detection and staging of prostate carcinoma: a review. *Medical Image Analysis*, vol. 10(2), pp. 178–179, 2006.

[85] R. Zwiggelaar, Y. Zhu and S. Williams. Semi-automatic segmentation of the prostate. In *Pattern Recognition and Image Analysis*, pp. 1108–1116, Springer Berlin Heidelberg, 2003.

[86] G. Litjens, R. Toth, W. van de Ven, C. Hoeks, S. Kerkstra, B. van Ginneken, G. Vincent, G. Guillard, N. Birbeck, J. Zhang, R. Strand, F. Malmberg, Y. Ou, C. Davatzikos, M. Kirschner, F. Jung, J. Yuan, W. Qiu, Q. Gao, P. E. Edwards, B. Maan, F. van der Heijden, S. Ghose, J. Mitra, J. Dowling, D. Barratt, H. Huisman, and A. Madabhushi. Evaluation of prostate segmentation algorithms for MRI: The PROMISE12 challenge. *Medical Image Analysis*, vol. 18(2), pp. 359–373, 2014.

[87] G.Brown, A.Pocock, M.Lujan and M.-J.Zhao. Conditional Likelihood Maximisation: A Unifying Framework for Information Theoretic Feature Selection. *Journal of Machine Learning Research*, vol. 13, pp. 27–66, 2012.

[88] L-C. Huang, S-E. Hsu and E. Lin. A comparison of classification methods for predicting Chronic Fatigue Syndrome based on genetic. *Journal of Translational Medicine*, vol. 7(1), pp.1–8, 2009. doi:10.1186/1479-5876-7-81

[89] D. R. Amancio, C. H. Comin, D. Casanova and G. Travieso and O. M. Bruno, F. A. Rodrigues, L. d. F. Costa. A Systematic Comparison of Supervised Classifiers. PLoS ONE 9, e94137, 2014. doi:10.1371/journal.pone.0094137

[90] P. C. Vos, J. O. Barentsz, N. Karssemeijer and H. J. Huisman. Automatic computer-aided detection of prostate cancer based on multiparametric magnetic resonance image analysis. *Physics in Medicine and Biology*, vol. 57, pp.1527–1542, 2012.

[91] K. H. Esbensen and P. Geladi. Principles of Proper Validation: use and abuse of re-sampling for validation. *ournal of Chemometrics*, vol. 24(3–4), pp.168–187, 2010.

[92] M. McCormack, A. Duclos, M. Latour, M. H. McCormack, D. Liberman, O. Djahangirian, J. Bergeron, L. Valiquette, and K. Zorn. Effect of needle size on cancer detection, pain, bleeding and infection in TRUS-guided prostate biopsies: a prospective trial. *Canadian Urological Association Journal*, vol. 6(2), pp. 97–101, 2012