# Computerized detection of cancer in multi-parametric prostate MRI

Geert Litjens

This book was typeset by the author using LaTeX2$_\varepsilon$.

Book by: Geert Litjens



ISBN 978-94-6259-481-4

Printed by Ipskamp Drukkers, Nijmegen.

# Computerized detection of cancer in multi-parametric prostate MRI

Proefschrift

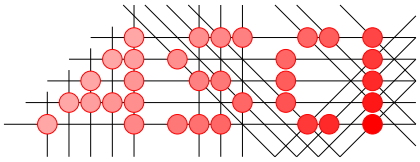
ter verkrijging van de graad van doctor aan de Radboud Universiteit Nijmegen op gezag van de rector magnificus prof. dr. Th.L.M. Engelen, volgens besluit van het college van decanen in het openbaar te verdedigen op vrijdag 23 januari 2015 om 12.30 uur precies

door

Gerardus Johannes Silvester Litjens

geboren op 4 april 1985 te Venlo

| | |
|---|---|
| Promotoren: | **Prof. dr. ir. N. Karssemeijer** |
| | **Prof. dr. J.O. Barentsz** |
| Copromotor: | **Dr. ir. H.J. Huisman** |
| Manuscriptcommissie: | **Prof. dr. J.A. Witjes** |
| | **Prof. dr. W.J. Niessen** (Erasmus MC) |
| | **Dr. J. Veltman** (Ziekenhuisgroep Twente) |




The research described in this thesis was carried out at the Diagnostic Image Analysis Group, Radboud University Medical Center (Nijmegen, The Netherlands).
This work was carried out in the ASCI graduate school (ASCI dissertation series number 319).
This work was funded by grant **KUN 2007-3971** of the Dutch Cancer Society.

Financial support for publication of this thesis was kindly provided by the department of Radiology at the Radboud University Medical Center, and the Advanced School for Computing and Imaging (ASCI).

TABLE OF CONTENTS

# Introduction

1

Prostate cancer is the most commonly diagnosed malignancy and the second leading cause of cancer death among men in the Netherlands. Due to the shortcomings of the current diagnostic pathway for prostate cancer, especially with respect to assessing cancer aggressiveness, alternative strategies are being investigated. Magnetic resonance imaging (MRI) has emerged as an important modality to assist and potentially replace (part of) the current diagnostic pathway. The high complexity of prostate MRI and the lack of sufficient expertise among the radiological community at large has opened the door for (semi-)automated analysis of prostate MRI by computer systems, with or without human intervention. This thesis will cover the development and evaluation of such a system in a clinical setting.

## 1.1 Prostate cancer

### 1.1.1 Prostate anatomy

The prostate is a walnut-sized organ in the male pelvis, located between the pelvic bones. The apex (caudal part) of the prostate is supported by the pelvic floor muscles. The base (cranial part) of the prostate borders the bladder. The urethra, originating in the bladder, and two ejaculatory ducts, originating at the seminal vesicles, pass through the prostate. This is depicted schematically in Figure 1.1. The prostate plays a role in the male reproductive system; it secretes an alkaline fluid, which is added to the spermatozoa and the seminal vesicle fluid, increasing the motility and lifespan of the spermatozoa. Furthermore, smooth muscle cells within the prostate help expel the semen during ejaculation. Structurally, the prostate is often

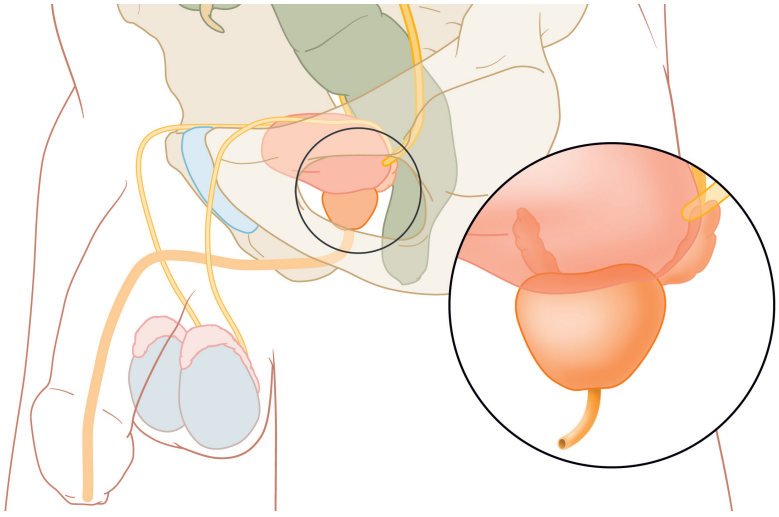


Figure 1.1: Anatomy of the male pelvis with the prostate circled.

divided into three distinct zones (Figure 1.2)[1]. The central zone surrounds the ejaculatory ducts at the base of the prostate and encompasses around 25% of the glandular tissue in a healthy prostate. However, only 2.5% of prostate cancers originate in this zone[2]. The transition zone

surrounds the proximal urethra and contains around 5% of the glandular tissue. Between 10 and 20% of cancers originate here[2]. As individuals age, the transition zone often undergoes hyperplasia, causing it to grow substantially in size. Lastly, the peripheral zone encompasses up to 70% of all glandular tissue in a healthy prostate and it occupies the posterior and lateral parts of the gland. Approximately 70% of all prostate cancers are found in this zone[2]. In addition to these glandular zones, the prostate usually contains an area of fibromuscular stroma, typically located at the ventral side of the prostate.

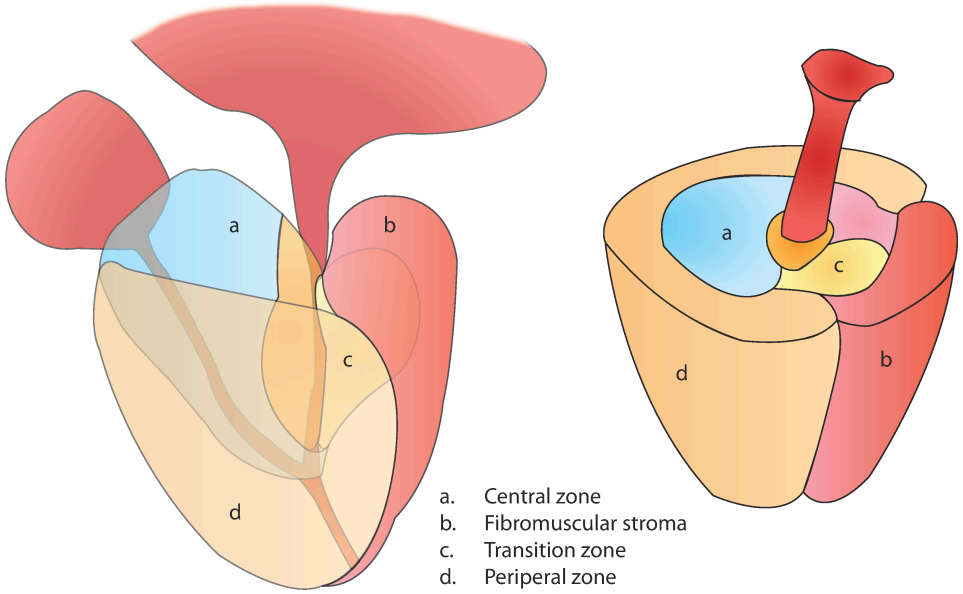



Figure 1.2: Schematic overview of the zonal anatomy of the prostate. Based on a figure by De Marzo et al.[3]

### 1.1.2 Epidemiology

Around 11000 men were diagnosed with prostate cancer in 2011 in the Netherlands and this incidence is rising (IKNL, `http://www.cijfersoverkanker.nl/`). Annually, approximately 2400 - 2500 men will die because of prostate cancer. Globally, the estimates are around 900000 new cases and 250000 deaths per year[4]. From these numbers it can be observed that prostate cancer has a high incidence-to-mortality ratio. The main reason for the high incidence-to-mortality ratio is that most of the diagnosed prostate cancers are indolent, i.e. will not kill the patient in their lifetime. This is evidenced by the high 10-year survival rate (77%) but also by the large number of prostate cancers discovered on autopsy in men who died of other causes[5,6].

Almost 75% of all prostate cancer cases are diagnosed in developed countries. The two main reasons are the higher average age of the general populace in developed countries compared to developing countries (70% of all new prostate cancer patients in the Netherlands are older than 65 years) and secondly, the advent of prostate specific antigen (PSA) based testing in the 1990s[7].

### 1.1.3 Current diagnostic pathway

The current diagnostic pathway for prostate cancer consists of a combination of PSA blood tests, digital rectal examinations (DRE) and trans-rectal ultrasound-guided (TRUS) biopsies (IKNL, Guidelines on Prostate Cancer). PSA tests measure the concentration of PSA in the blood in nanograms per milliliter (ng/mL). In individuals with healthy prostates, the PSA level will be low, as the PSA will be contained in the prostate glands. In individuals with prostate cancer or other prostate disorders (e.g. benign prostatic hyperplasia or prostatitis) the PSA level can increase because the integrity of the prostate glands is compromised. The American Cancer Society suggest a threshold of four ng/mL to refer patients for further examination. In the Netherlands current guidelines suggest using a threshold of three ng/mL, although PSA testing itself is discouraged.

In digital rectal examinations urologists will use a lubricated, gloved finger to inspect the surface of the prostate. Prostate cancer tends to feel as a stony, asymmetrical lump compared to soft, smooth healthy prostate tissue. However, DREs have a very limited area of coverage (ventrally located tumors are missed) and tend to miss smaller tumors. As such DRE has a limited sensitivity and specificity[8].

After initial suspicion has arisen due to either PSA or DRE usually a TRUS biopsy is performed. As most prostate cancers are invisible on ultrasound[9], TRUS biopsies are performed in a systematic way, usually with between six and twelve cores covering part of the prostate. Biopsy specimens are subsequently evaluated by a pathologist using the Gleason Scoring System.

The Gleason Scoring System is named after Donald Gleason, who developed it with other colleagues at the Minneapolis Veterans Affairs Hospital during the 1960s[10]. The system was subsequently updated in 2005 by International Society of Urological Pathology[11]. A Gleason score is assigned by the pathologist by summing two numbers; the first number indicates the grade of the most common tumor pattern in the specimen, the second number indicates the second most common pattern. If there are more than two patterns present in the specimen, the second number should refer to the remaining pattern with the highest grade (which contains at least 5% of the total tumor volume). Gleason grades range from 1 - 5, where 5 is considered the most aggressive pattern. The descriptions of the different patterns are:

1. Very well differentiated, small, closely packed, uniform, glands in essentially circumscribed masses.

2. Similar (to pattern 1) but with moderate variation in size and shape of glands and more atypia in the individual cells; cribriform patterns may be present, still essentially circumscribed, but more loosely arranged.

3. Similar to pattern 2 but marked irregularity in size and shape of glands, with tiny glands

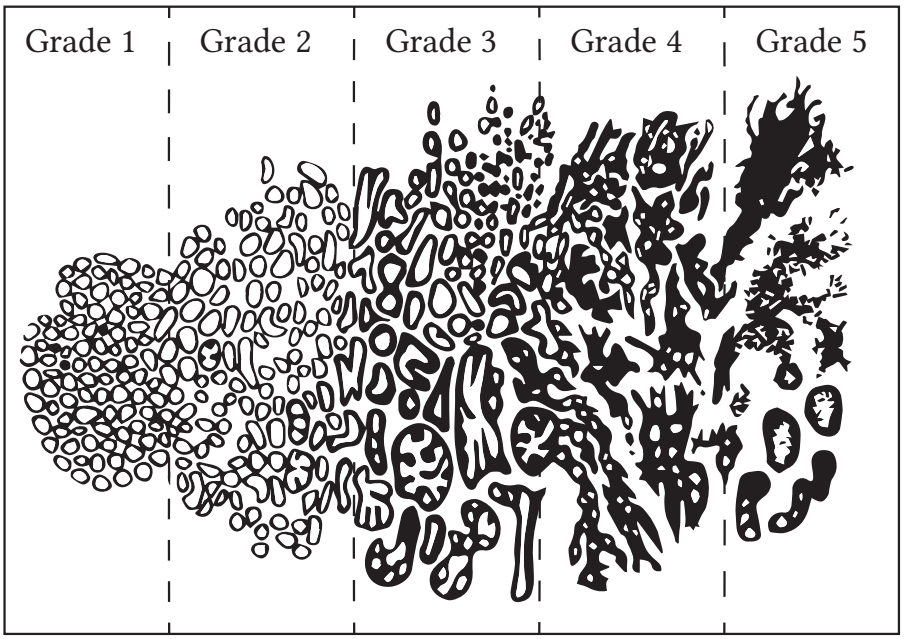



Figure 1.3: Schematic representation of the different Gleason grades. Image adapted from[12]

or individual cells invading stroma away from circumscribed masses, or solid cords and masses with easily identifiable glandular differentiation within most of them. May be papillary or cribriform, which vary in size and may be quite large, but the essential feature is the smooth and usually rounded edge around all the circumscribed masses of tumor.

4. Large clear cells growing in a diffuse pattern resembling hypernephroma; may show gland formation. Raggedly infiltrating, fused-glandular tumor; glands are not single and separate, but coalesce and branch.

5. Very poorly differentiated tumors; usually solid masses or diffuse growth with little or no differentiation into glands. Can resemble comedocarcinoma of the breast; almost absent gland pattern with few tiny glands or signet cells.

These descriptions are illustrated in Figure 1.3. Gleason scores 2 (1+1), 3 (2+1 or 1+2) and 4 (2+2) are generally discouraged after the Gleason Scoring System revision in 2005[11]. As such, in clinical practice, only Gleason score of 3+2 or higher are encountered, of which 3+3 is by far the most common pattern.

The attending physician will use the biopsy Gleason score and other clinical parameters (e.g. PSA level, number of positive biopsy cores) to decide on the best management of the disease.

### 1.1.4 Management and treatment

Several management and treatment options exist for prostate cancer, the most common choices are active surveillance, radical prostatectomy, radiotherapy, or focal therapy.

Active surveillance is an ideal option for men with low grade, localized, well-differentiated prostate cancer. Several studies have shown that in men with untreated, low-risk prostate cancer the 10- to 20-year survival rates are similar to an age-matched group of men without prostate cancer[13–15]. Based on these results several groups have implemented active surveillance protocols using PSA tests and TRUS biopsies for follow-up. Initial results are difficult to compare, as inclusion criteria differ between studies. The study by Klotz et al.[16], including 450 patients, has the longest median follow-up (6.8 years) and has shown promising results with a cancer specific survival of 97%.

Radical prostatectomy, external beam radiotherapy (EBRT) or brachytherapy are all treatment options with intent to cure. In the United States the use of these three treatment modalities is approximately equal. Cancer recurrence and survival rates are relatively similar, and as such other factors like disease stage (lymph node involvement), expected side-effects and patient opinion are important factors in which treatment is chosen[17]. One advantage of radical prostatectomy over the other two modalities is the potential to completely stage the prostate cancer (complete Gleason grading, extra-capsular extension) and the easier biochemical follow-up using PSA (PSA level should reduce to near zero after removal of the prostate).

Research interest into focal therapy has risen over the past decade, with several options available, e.g. laser interstitial thermotherapy (LITT), cryo-ablation or high-intensity focused ultrasound (HIFU). Although initial results with these therapy options are promising, especially for localized, low-grade prostate cancer, they are currently not yet widely available clinically. All treatment options for prostate cancer with intent to cure have the potential to cause side-effects among which incontinence and sexual dysfunction are the most common[18–20]. Two studies reported that between around 40 - 45% of all patients have issues with erectile dysfunction after prostatectomy[18,20]. Urinary problems are reported less consistently, with the percentage of patients affected ranging from 7 - 49%. Reported side-effects for both EBRT and brachytherapy are lower with around 13 - 22% of patients suffering from erectile dysfunction and 11 - 18% of patients having urinary problems. However, patients who underwent brachytherapy or EBRT had more issues with diarrhea or blood in their stool (10%). Due to the impact of these side effects it is important to only diagnose and treat prostate cancer that will cause premature death.

### 1.1.5 Drawbacks of the current diagnostic pathway

The relative simplicity of PSA testing and subsequent TRUS biopsies has lead to trials investigating the potential role of these two techniques in a screening setting for prostate cancer. In

2009 Schröder et al. published the results of a large European screening trial using PSA testing and TRUS biopsies to detect prostate cancer[21]. An update was published in 2012[22]. Their results showed that there is potential for prostate cancer screening, with a reduction in the relative risk of death due to prostate cancer of around 29%. However, the poor sensitivity and specificity of PSA testing and TRUS biopsies would cause large amounts of over-diagnosis and over-treatment. To save one life, 1055 men had to be screened and 37 men had to be treated for prostate cancer.

In this screening setup, PSA testing is essentially used as a triage test for the more invasive procedure of TRUS biopsies. As such the cut-off value selected for this test limits the overall maximum sensitivity of the screening program. A large study by Thompson et al. investigated the sensitivity and specificity of PSA at different cut-off values in over 8500 patients with an initial PSA of 3.0 ng/mL or lower with 7 year follow-up. After 7 years all patients received a end-of-study prostate biopsy. They found that at thresholds of 1.1, 2.1, 3.1 and 4.1 the sensitivities of the PSA test for predicting any cancer were 83.4, 52.6, 32.2 and 20.5% with 38.9, 72.5, 86.7 and 93.8% as the respective specificities. This is visualized in an receiver-operating characteristic curve in Figure 1.4. A review study by the American Cancer Society showed similar results[23]. This indicates that the maximum sensitivity for the screening program is between 32.2 and 20.5 depending on the cut-off value used. Lower cut-off values as they would result in a very large amount of unnecessary TRUS biopsies in healthy men, as the prevalence of prostate cancer in the screening group of men aged between 55 and 74 years is only 8.2%[21]. For cancer with a Gleason score larger than or equal to 7 the sensitivities were 92.8, 75.6, 57.6 and 40.4 with 37.0, 67.3, 82.3 and 90.0 as the respective specificities. These numbers are arbitrarily more relevant, as these are the cancers that need to be detected as they have a relatively poor prognosis.

The subsequent TRUS biopsies and pathological analysis of the biopsy specimens are the end-result of the screening program. Patient handling is based on these results. However, the reported detection rates of TRUS are relatively poor. The most common sextant biopsy strategy typically misses 15 - 34% of prostate cancers at the first biopsy[24–26]. Furthermore, TRUS biopsies often under- or over-estimate cancer grade. Underestimation occurs in around 46% of the cases and overestimation in 18% of the cases when compared to pathological analysis of radical prostatectomy specimens[27,28].

Combining the results for PSA testing and TRUS it becomes apparent that the current tools are inadequate for prostate cancer screening. The high threshold of 4 ng/mL on the initial PSA test causes almost 80% of cancers to be missed at a relatively low screening specificity of 93.8%. For the high-grade cancers, which are clinically most relevant, 59.6% at a specificity of 90.0% would be missed. To put this into perspective: if 100000 men are screened, 2307 men would have a high-grade cancer and require biopsy. Of these cancers 1375 would be missed due to the poor sensitivity of PSA at this cut-off. Furthermore, an additional 232 cancers would be missed

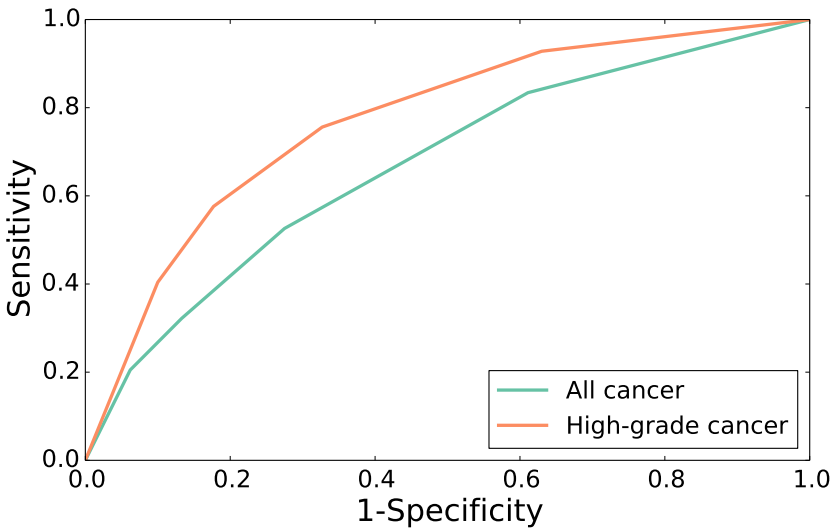



Figure 1.4: Receiver operating characteristic curve for prostate specific antigen blood tests in identifying prostate cancer.

due to the poor detection rate of the TRUS biopsies and 322 of the cancers would be under-graded. Of the 100000 men 10000 would receive an unnecessary biopsy, who would have either no cancer (9170 men) or a low-grade cancer (830 men). The men with low-grade cancer will most likely receive some sort of (unnecessary) treatment and the associated side-effects. The 9170 men with no cancer will have to undergo an unpleasant procedure, suffer from anxiety of potentially having prostate cancer and the associated morbidity of the biopsy. Furthermore, even after a negative result, they will be unsure whether they are actually cancer-free due to the poor sensitivity of TRUS biopsies and thus will most likely undergo repeat biopsies in the future.

Summarizing, although there is potential for prostate cancer screening, the currently accepted clinical tools for diagnosis are inadequate. If an alternative technology can improve the overall sensitivity, and especially the specificity of the diagnostic pathway, and thus reduce over-diagnosis and over-treatment, screening might become feasible. One technology which has shown potential is magnetic resonance imaging (MRI).

## 1.2 MRI for prostate cancer

### 1.2.1 General concepts

The first magnetic resonance image (MRI) was created by Paul Lauterbur in 1973[29]. He expanded on the pioneering work by Herman Carr in 1952. Since then MRI has become a mainstay imaging modality in hospitals worldwide. Magnetic resonance imaging uses the quantum mechanical concept of spin (intrinsic angular moment of particles) to create images. Unpaired protons, neutrons and electrons all possess a spin quantified as 1/2. Combinations of particles

have different amounts of spin, depending on their interactions, the so called net spin. MRI can only visualize particles with a net spin which is non-zero. The rest of this section will specifically be on hydrogen MRI (hydrogen atoms have a net spin of 1/2).

Particles with a net spin possess a magnetic moment. When a external magnetic field is applied the magnetic moment starts precessing about the direction of the magnetic field with a spin-dependent angle. The rate of precession is governed by the gyromagnetic ratio ($\gamma$, an intrinsic property of the particle) and the magnetic field strength ($B_0$), resulting in a precession at the so called Lamor frequency ($\omega$):

$$\omega = B_0\gamma \tag{1.1}$$

For particles with a net spin of 1/2 this precession can occur in a low- (parallel to the magnetic field) or a high-energy (anti-parallel to the magnetic field) state. Transitions between the two states are possible by absorbing a photon with the exact same energy as the energy difference between the states. This energy difference is determined by the strength of the magnetic field; with a higher magnetic field strength a higher energy difference between states exists. When grouping a set of particles with certain spin at room temperature their will be slightly more spins in the lower energy state than in the higher energy state. As the signal in MRI is acquired using the difference in energy absorbed by the spins making the transition from the lower to the higher energy state and the energy released by the spins moving from the higher to the lower energy state, the amount of signal is proportional to the distribution of the spins in the high- and low-energy states.

Because of the difference between the distribution of the spins over the different energy levels a net magnetization vector is present along the direction of the magnetic field. By applying an electromagnetic wave of photons (a radiofrequency pulse) with energy equal to the energy difference between the two states the orientation of the net magnetization vector can be changed and pushed in the transverse plane. When a long enough pulse is given the difference in particles in the energy states can become zero and the net magnetization vector is fully in the transverse plane (a so-called 90-degree pulse). Over time the spin states will return to their original distribution with a time constant T1. This recovery is called spin-lattice relaxation and is governed by the equation:

$$M_z(t) = M_0\left(1 - e^{\frac{-t}{T1}}\right) \tag{1.2}$$

A second relaxation effect occurs in the transverse plane and is called spin-spin relaxation. Immediately after the 90-degree radiofrequency pulse the magnetic moment of the spins will be in phase and the net magnetization vector will start precessing along the transverse plane. As the spin magnetic moments will dephase over time, the net magnetization vector will decay with a time constant T2 governed by the equation:

$$M_{xy}(t) = M_{xy0}\left(e^{\frac{-t}{T2}}\right) \tag{1.3}$$

Dephasing of the spins is caused by magnetic dipole-dipole interactions. Another effect which causes the dephasing is inhomogeneities in the $B_0$ field. The combination of the dipole interactions and the inhomogeneity effects leads to a time constant $T2^*$.

By using different sequences of radiofrequent pulses signals can be measured which are either weighted relative to T1, T2 or spin density at a spatial location. Spatial localization of the signals can be performed by using gradients, which give each spatial location either a slightly different frequency or a slightly different phase. By encoding spatial locations using phase and frequency a Fourier transformation can be used to obtain images.

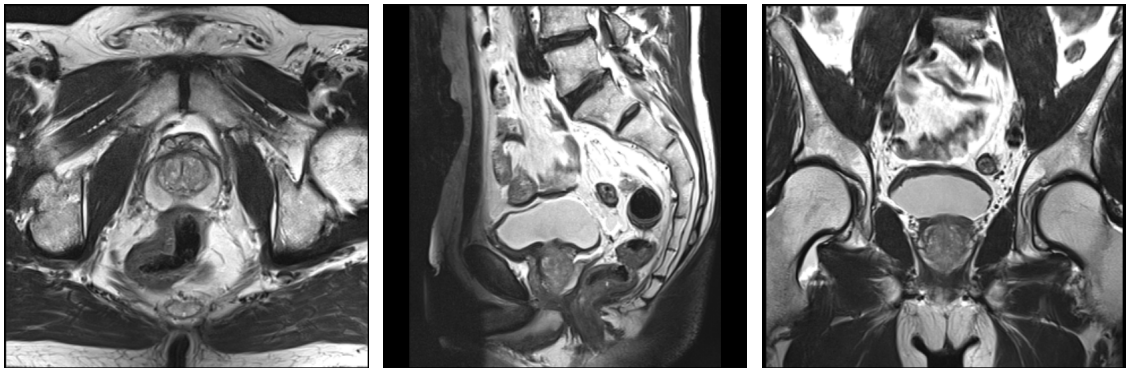
Tissue usually has distinctly different MRI properties (e.g. T1 or T2 times), which allows for high soft tissue contrast in MR images. Depending on the organ or disease of interest one might have a preference for T1-weighted, T2-weighted or more advanced types of imaging.

### 1.2.2 Prostate MRI

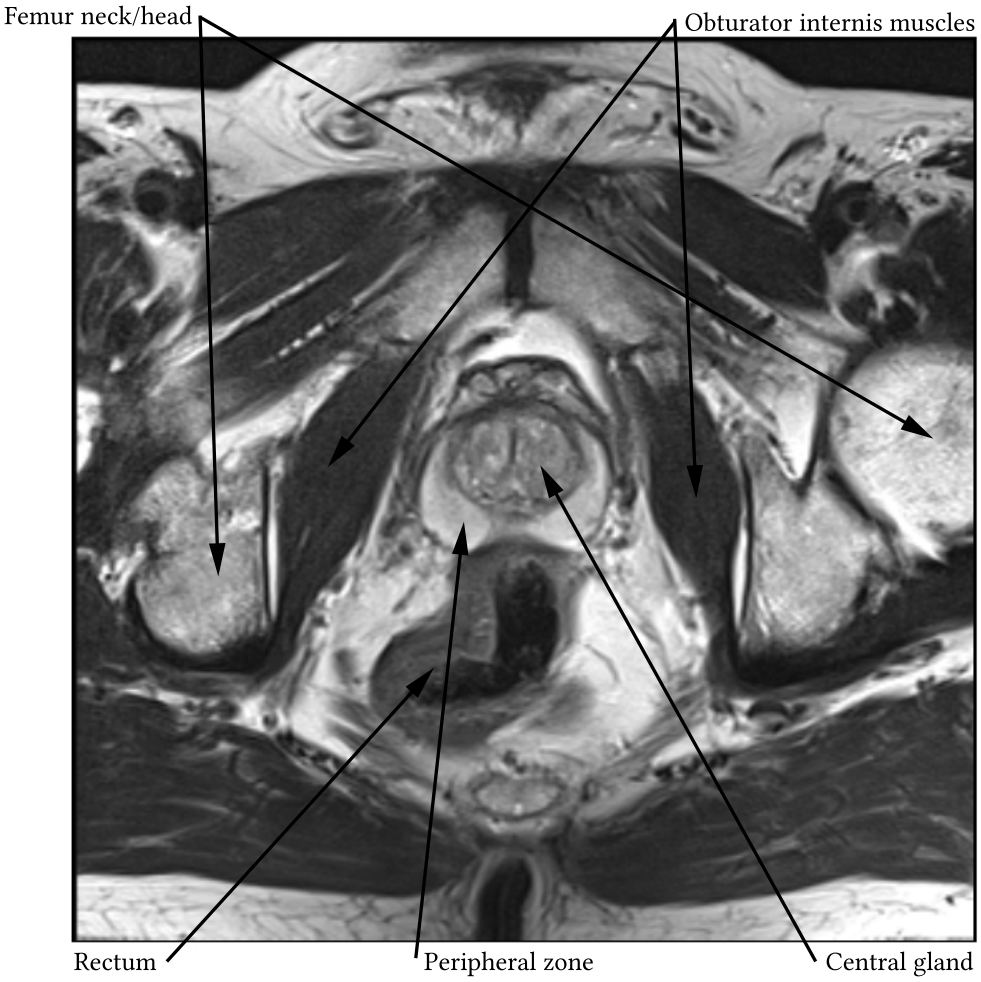
MRI for prostate cancer diagnosis has been around since the 1980s. Initially only focused on T2-weighted imaging[30] due to the high tissue contrast and spatial resolution, it has since expanded to include multiple MR parameters. Modern prostate MRI can consist of T2-weighted imaging, dynamic contrast-enhanced imaging (T1-weighted), diffusion-weighted imaging and spectroscopic imaging[31–33]. Prostate MRI is usually acquired using 1.5 or 3 tesla MRI with either a pelvic phased-array coil or an endo-rectal coil. The decision whether an endo-rectal coil is required depends on the required resolution and signal-to-noise ratio (SNR) of the images.

T2-weighted imaging is considered the standard for anatomical assessment of the prostate, as it has the highest resolution and best tissue contrast of all the modalities. Usually, three orthogonal directions are acquired: sagittal, axial and coronal with a high in-plane resolution (around 0.2 - 0.6mm) and a relatively large slice thickness (2 - 4 mm). On T2-weighted imaging the peripheral zone of the prostate usually appears as a bright, relatively homogenous region. The central zone, transition zone and fibromuscular stroma are usually hardly discernible radiologically and are thus often grouped together as a single zone: the central gland[34]. The central gland is usually of a much darker appearance with a chaotic texture (caused by benign disease like benign prostatic hyperplasia). An example of T2-weighted imaging is shown in Figure 1.5.

Dynamic contrast-enhanced (DCE) imaging is a combination of several T1-weighted image acquisitions over time following contrast agent injection. MRI has several types of contrast agents, one of which are T1-shortening agents, usually containing paramagnetic metals. The most common type of these agents are the gadolinium chelates. Administering a gadolinium contrast agent will result in a T1 shortening proportional to the concentration of agent at the specific location. A shortening of T1 will result in a higher signal intensity on T1-weighted imaging and as such higher contrast between areas with a high concentration of contrast agent

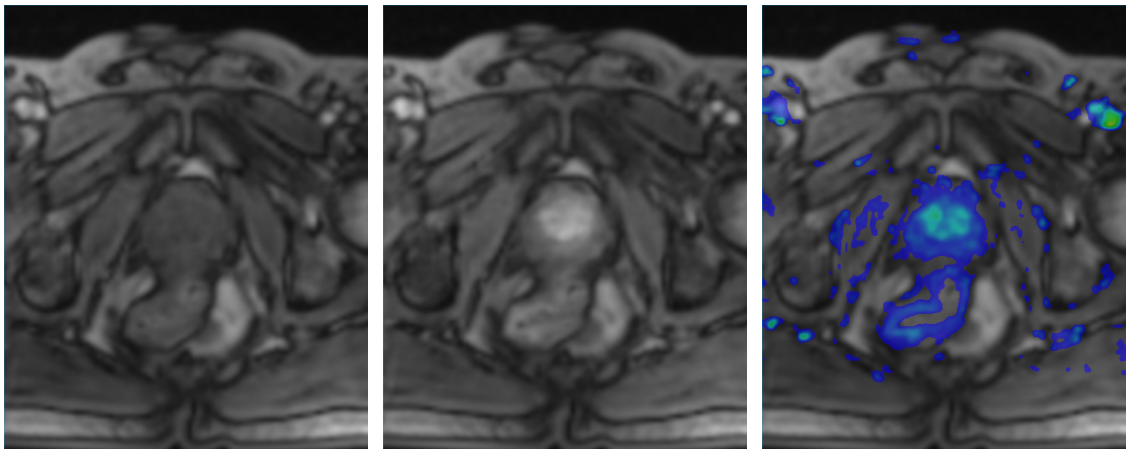


(a) (b) (c)




(d)

Figure 1.5: Example of T2-weighted prostate MRI in three orthogonal directions. Figure (a), (b) and (c) show the transversal, sagittal and coronal view respectively. Figure (d) shows the transversal view with several anatomical structures annotated. These images were acquired on a 3T MRI scanner using a pelvic phased array coil.



(a) (b) (c)

Figure 1.6: Example of dynamic contrast-enhanced MRI. Figure (a) shows a pre-contrast image, Figure (b) a post-contrast image. In Figure (c) a $K^{trans}$ parameter map is overlayed on a pre-contrast image. These images were acquired on a 3T MRI scanner using a pelvic phased array coil.

and areas with low concentrations. An example of a T1-weighted prostate image before and after administration of contrast is visualized in Figures 1.6a and 1.6b. Cancer tends to be rapidly growing tissue needing lots of nutrients and oxygen. This causes cancer cells to stimulate blood vessel growth, however, due to the fast growth vessel integrity is usually sub-optimal, resulting in 'leaky' vessels. The increased vascularity and permeability of the blood vessels causes contrast agent to diffuse out into cancerous tissue more easily than into normal tissue, resulting in higher concentrations and thus higher signal intensity on T1-weighted imaging. By inspecting the signal-intensity-over-time curves diagnostic information can be obtained. For example, the increase of signal intensity over time is related to the rate of uptake of contrast agent. To accurately describe the dynamic behavior of tissue with respect to contrast agent uptake over time a high temporal resolution is typically needed, usually in the order of seconds. High temporal resolution data also allows quantitative extraction of tissue parameters using pharmacokinetic modeling. Pharmacokinetic modeling tries to capture quantitative physiological parameters from DCE MRI signal-intensity-over-time curves by compartmental modeling of tissue. The original models proposed by Tofts[35] and Brix[36] assume that any tissue voxel consists of two compartments that have exchange of contrast agent. As gadolinium chelates cannot enter the interior of the cells, there is only exchange between the blood plasma and the extra-vascular, extra-cellular space or 'leakage' space (EES) through the vessel wall. This can be described using the following equation:

$$C_t(t) = \nu_p C_p(t) + K^{trans} \int C_p(\tau) e^{-k_{ep}(t-\tau)} d\tau \tag{1.4}$$

$$k_{ep} = \frac{K^{trans}}{\nu_e} \tag{1.5}$$

here $C_t(t)$ is the concentration of contrast agent in the tissue at time t, $\nu_p$ is the fraction of

plasma within the tissue, $C_p(t)$ is the concentration of contrast agent in the blood plasma at time t, $K^{trans}$ is the volume transfer constant between blood plasma and EES, $k_{ep}$ is the rate constant between EES and blood plasma and $v_e$ is the fraction of EES within the tissue. In most practical implementations of this model $v_p$ is neglected because it is usually assumed to be close to zero. Furthermore, it requires very high temporal resolution data to correctly estimate[37]. $K^{trans}$ and $v_e$ are of interest because they describe inherent tissue properties where $K^{trans}$ is related to vessel permeability and $v_e$ is related to microscopic tissue structure.

Diffusion-weighted imaging is the newest parameter in prostate cancer MRI and tries to capture information on the Brownian motion of protons[38,39]. Diffusion-weighted MRI is based on the pioneering work of Stejskal and Tanner[40]. It has has been related to cellular density in tissue (higher cell density means restricted diffusion)[41]. Cellular density is usually increased in cancer due to rapid proliferation of cells. MRI is made sensitive to diffusion effects by applying a spatially varying gradient pulse and then after a certain time $\delta$ applying the exact same pulse in the opposite direction. If there was no diffusion, the net effect on the spins would be zero. However, due to diffusion the spins are affected differently by the gradient pulses, resulting in signal loss. As signal loss is also related to T1 and T2 relaxation effects, multiple diffusion-weighted acquisitions are made with varying gradient pulse parameters (for example strength, or duration). The amount of signal loss over the different diffusion-weighted acquisitions can then be calculated using:

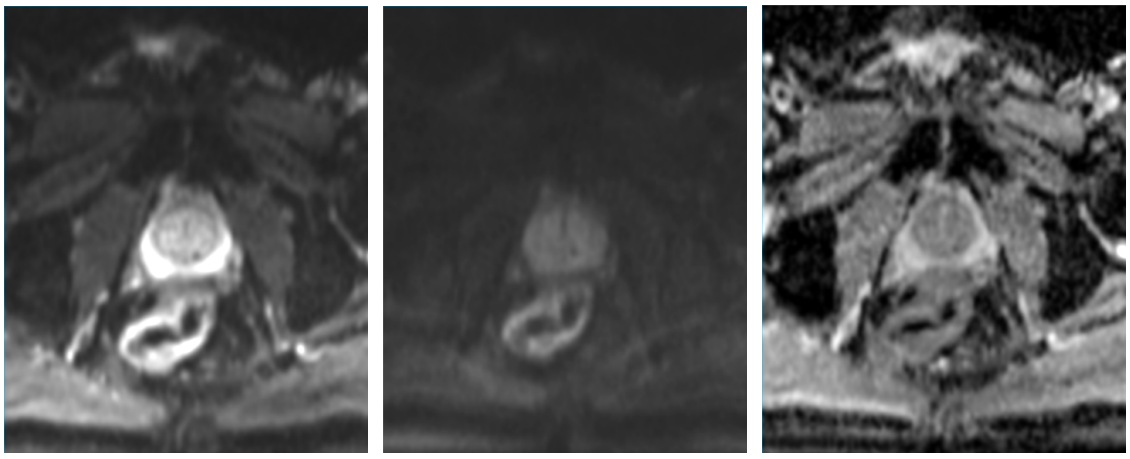
$$S(b) = S(0)\exp^{-b\,ADC} \tag{1.6}$$

where b is a parameter which summarizes the gradient pulse acquisition settings (strength, duration and time between pulses), $S(b)$ is the signal intensity at a certain value of b, and ADC the apparent diffusion coefficient, which is a tissue property. Given at least two different b-value acquisitions, the ADC at a certain location $(x, y, z)$ can be calculated using:

$$ADC(x, y, z) = \frac{\ln \frac{S(b_2,x,y,z)}{S(b_1,x,y,z)}}{b_1 - b_2} \tag{1.7}$$

Subscripts 1 and 2 indicate the acquisition number. Using least squares optimization this equation can be extended to an arbitrary number of acquisitions and b-values. The ADC is in principle sensitive to the gradient pulse direction and as such usually multiple acquisition in orthogonal directions are made and averaged to obtain a direction-independent ADC. An example of two different b-value images and an ADC map are shown in Figure 1.7.

Finally, MR spectroscopic imaging uses the principle of chemical shift (hydrogen protons attached to different molecules exhibit a slightly different resonance frequency due to their environment) to measure the concentration of different metabolites in vivo. Voxel sizes tend to be much larger than in the other MR parameters (5x5x5mm for example), however, at each voxel a complete resonance spectrum is obtained. Visualization usually happens by overlaying a voxel grid over the T2-weighted imaging, which allows clinicians to correlate the spectrum



(a) (b) (c)

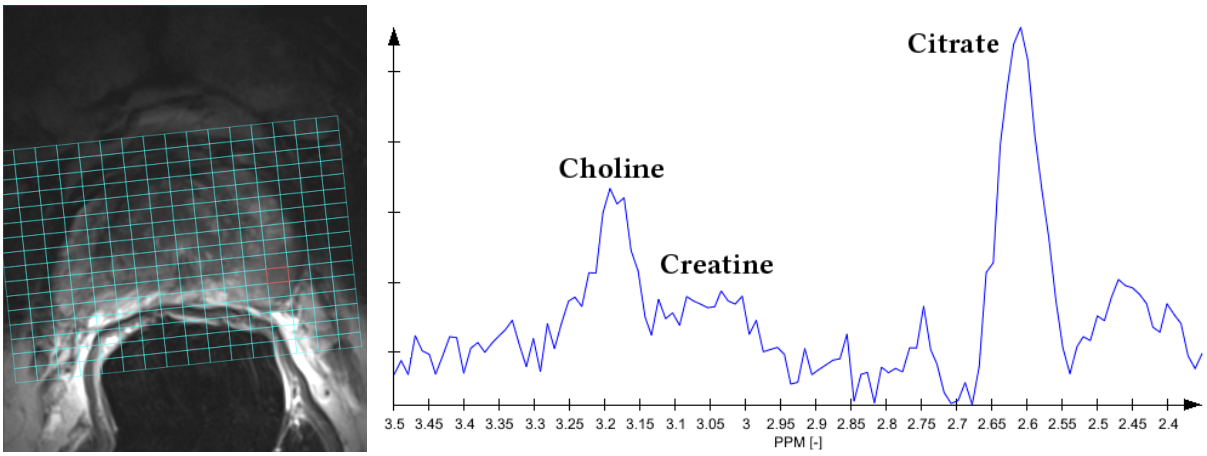
Figure 1.7: Example of diffusion-weighted MRI. Figure a and b show a transversal slice with low (a) and high (b) b-value diffusion-weighting. Image c shows an example of an apparent diffusion coefficient map calculated from multiple b-value images. These images were acquired on a 3T MRI scanner using a pelvic phased array coil.

to a spatial location. An example is shown in Figure 1.8. In prostate MR, spectroscopy allows us to measure concentrations of for example choline, creatine and citrate. MR spectroscopic imaging is currently not as widely used in clinical practice as the other modalities due to the complexity of the acquisition and post-processing, which results in reduced reproducibility.

### 1.2.3 Prostate cancer diagnosis on MRI

A large body of literature exists describing the use of individual and combinations of different MR parameters for prostate cancer detection. For T2-weighted imaging alone only a moderate sensitivity (57 - 84%) and specificity (50 - 94%) are reported[42–44]. Dynamic contrast-enhanced imaging is able to achieve 59 - 73% sensitivity and 74 - 94% specificity[45,46]. Additionally, in a recent study Vos et al. investigated the use of DCE MRI to assess cancer aggressiveness and reported an area under the receiver operating characteristic (ROC) curve of 0.72 in discriminating low-grade cancer from intermediate-to-high-grade cancer[47]. Diffusion-weighted imaging has a reported sensitivity of 84- 90% and a corresponding specificity of 81-84% for the detection of prostate cancer[48–50]. However, stand-alone performance of diffusion-weighted imaging was only evaluated in a limited amount of patients. Additionally, several studies have shown a clear relation between the value of the apparent diffusion coefficient and Gleason grade[51,52]. However, this holds mostly for the peripheral zone as in the central gland the diagnosis is more difficult due to overlapping imaging characteristics of BPH and prostate cancer with respect to diffusion-weighted imaging. Finally, one study also included the stand-alone performance of spectroscopic imaging for the detection of prostate cancer with an area under the ROC of 0.81 versus 0.80 for T2-weighted imaging[42].

Several groups have evaluated the added value of the functional imaging techniques in




(a) (b)

Figure 1.8: Example of visualization of spectroscopic imaging in prostate MRI. Figure (a) shows the spectroscopy grid overlayed on a transversal T2-weighted image. Figure (b) shows the obtained spectrum for the selected voxel (indicated in red in Figure (a)). These images were acquired with an endo-rectal coil.

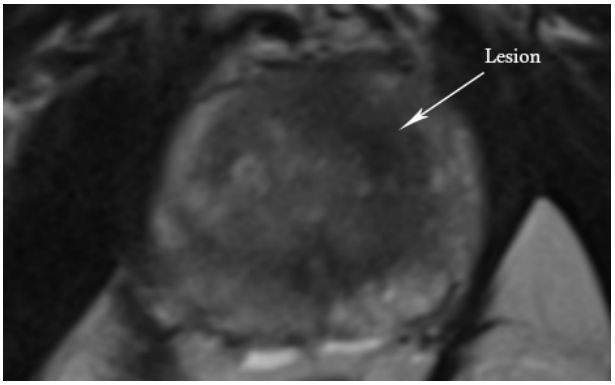



Figure 1.9: A dark, homogenous region with indistinct boundaries (erased charcoal sign) can be identified in the the top-right part of the image. This lesion was scored a 5 on T2-weighted imaging following the ESUR guidelines.

addition to T2-weighted imaging. In[32] an area under the ROC curve (AUC) of 0.84 was reported for T2-weighted imaging alone. The addition of DWI resulted in an AUC of 0.89 and the addition of DCE in an AUC of 0.88. Combining all three parameters resulted in an AUC of 0.90. Similar results were obtained by[31] with an AUC for T2-weighted imaging alone of 0.77. The addition of only DWI resulted in an AUC of 0.90, a combination of all parameters resulted in an AUC of 0.97. The combination of MR spectroscopy with T2-weighted imaging resulted in an AUC of 0.85 versus 0.80 for T2-weighted imaging alone[42].

Although these results show that MRI for prostate cancer has potential, a major issue with broad adaptation was the professional disagreement on what MR parameters to use, how to interpret them and when to use MRI[53,54] (as evidenced by the large variation in reported sensitivities and specificities). Recently, an effort has been made by the European Society for Urogential Radiology (ESUR) to standardize prostate MRI[33]. Currently, this is being formal-

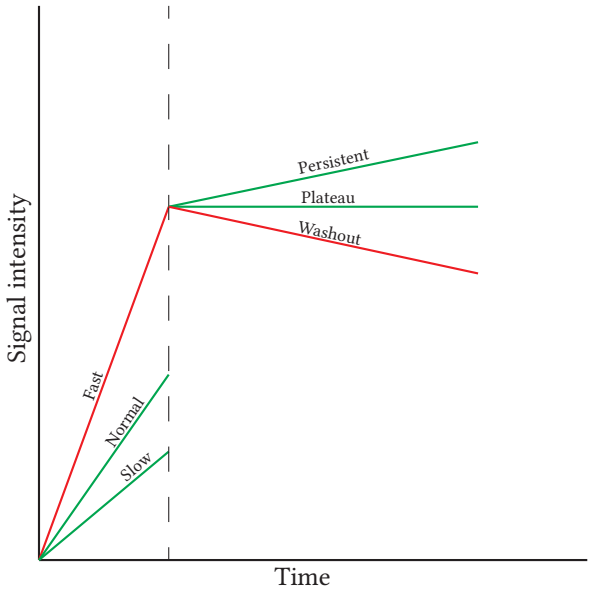
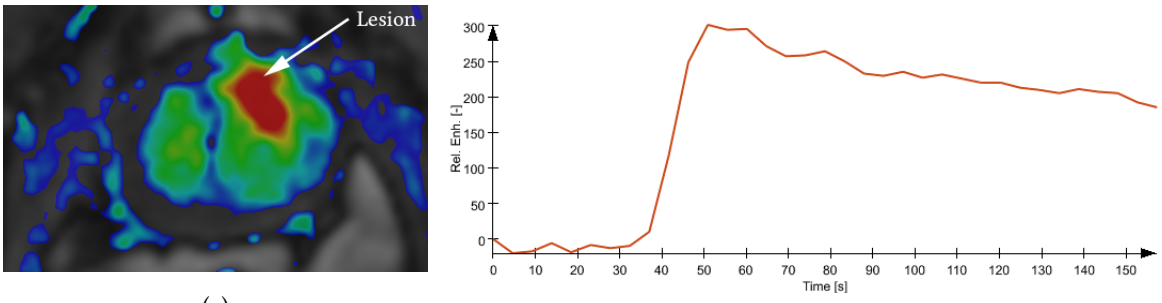



Figure 1.10: Schematic visualization of the three different curve types that can be classified in DCE MRI of the prostate. Curve type 3 (red curve) is a sign of malignancy.




(a) (b)

Figure 1.11: A focal lesion with high $K^{trans}$ is visualized in Figure a. The corresponding curve (type 3) is shown in Figure b. This lesion was assigned a score of 5 according to the DCE reporting guidelines of the ESUR.

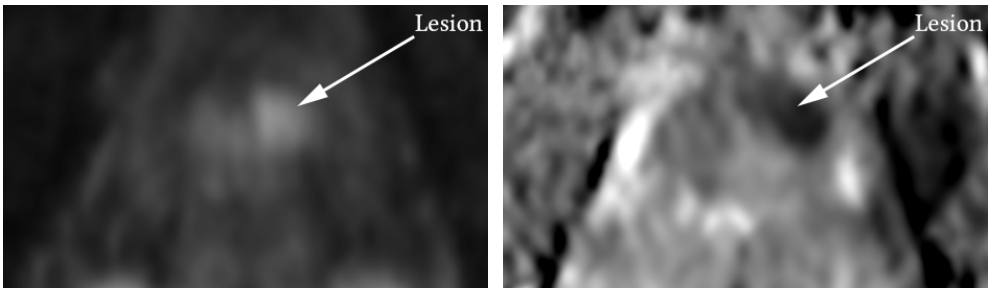
ized in a Prostate Imaging Reporting and Data Standard (PIRADS) by the American College of Radiology. Table 1.1 gives a summary of the scoring criteria presented by ESUR.

For T2-weighted imaging there are separate instructions for the detection of cancer in the peripheral zone and the central gland. As mentioned in section 1.2.2, the appearance of both zones is markedly different and the co-occurence of benign disease like BPH in the central gland make special guidelines for this zone a necessity. In the peripheral zone prostate cancer typically manifests as a round or ill-defined, low-signal-intensity focus. In the CG tumor often shows a homogeneous signal mass with indistinct margins (the so called “erased charcoal sign”). An example of a tumor which was scored a 5 on T2-weighted imaging according to the ESUR guidelines is shown in Figure 1.9.

For DCE imaging the basic analysis described in the ESUR guidelines uses curve types.

| Score | Criteria |
|---|---|
| T2-weighted imaging for the peripheral zone | |
| 1 | Uniform high signal intensity |
| 2 | Linear, wedge shaped, or geographic areas of lower SI, usually not well demarcated |
| 3 | Intermediate appearances not in categories 1/2 or 4/5 |
| 4 | Discrete, homogeneous low signal focus/mass confined to the prostate |
| 5 | Discrete, homogeneous low signal intensity focus with extra-capsular extension/invasive behavior or mass effect on the capsule (bulging), or broad (>1.5 cm) contact with the surface |
| T2-weighted imaging for the transition zone | |
| 1 | Heterogeneous TZ adenoma with well-defined margins: “organised chaos” |
| 2 | Areas of more homogeneous low SI, however well marginated, originating from the TZ/BPH |
| 3 | Intermediate appearances not in categories 1/2 or 4/5 |
| 4 | Areas of more homogeneous low SI, ill defined: “erased charcoal sign” |
| 5 | Same as 4, but involving the anterior fibromuscular stroma or the anterior horn of the PZ, usually lenticular or water-drop shaped. |
| Diffusion-weighted imaging | |
| 1 | No reduction in ADC compared with normal glandular tissue. No increase in SI on any high b-value image ($\geqslant$ b800) |
| 2 | Diffuse, hyper SI on ≥b800 image with low ADC; no focal features, however, linear, triangular or geographical features are allowed |
| 3 | Intermediate appearances not in categories 1/2 or 4/5 |
| 4 | Focal area(s) of reduced ADC but iso-intense SI on high b-value images ($\geqslant$ b800) |
| 5 | Focal area/mass of hyper SI on the high b-value images ($\geqslant$ b800) with reduced ADC |
| Dynamic contrast-enhanced imaging | |
| 1 | Type 1 enhancement curve |
| 2 | Type 2 enhancement curve |
| 3 | Type 3 enhancement curve |
| +1 | For focal enhancing lesion with curve type 2–3 |
| +1 | For asymmetric lesion or lesion at an unusual place with curve type 2–3 |
| Magnetic resonance spectroscopic imaging | |
| 1 | Citrate peak height exceeds choline peak height >2 times |
| 2 | Citrate peak height exceeds choline peak height times $>1$, $<2$ times |
| 3 | Choline peak height equals citrate peak height |
| 4 | Choline peak height exceeds citrate peak height $>1$, $<2$ times |
| 5 | Choline peak height exceeds citrate peak height >2 times |

Table 1.1: ESUR guidelines for scoring multi-parametric prostate MRI per modality.




(a) (b)

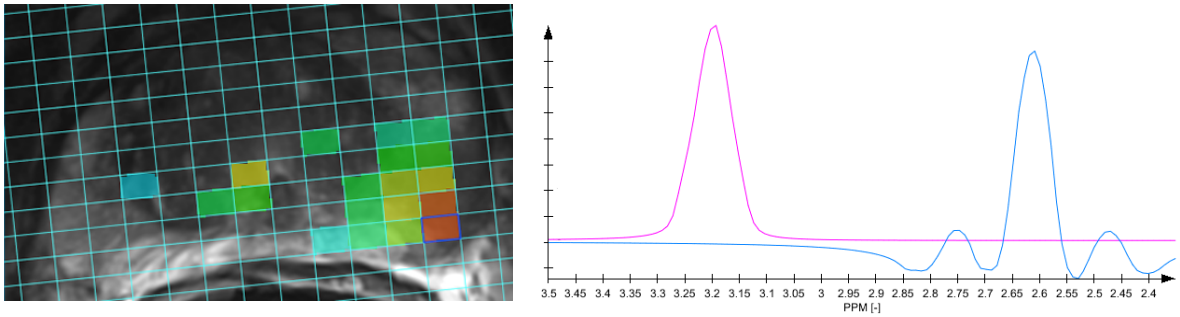
Figure 1.12: A lesion with restricted diffusion, evidenced by the high signal intensity on the high b-value image (a) and low apparent diffusion coefficient (b). This lesion was assigned a score of 5 in concordance with the ESUR guidelines for diffusion-weighted imaging.

The three different curve types are visualized schematically in Figure 1.10. Curve type 1 represents slow-to-moderate initial enhancement and persistent enhancement in the latter part of the curve. Curve type 2 is slow-to-moderate initial enhancement with a subsequent enhancement plateau. Curve type 3 is fast initial enhancement with a subsequent drop in enhancement (wash-out). Curve type 3 is considered a sign of malignancy. At the Radboud University Medical Center pharmacokinetic analysis has been used since 2001[55–57] to provide additional analysis tools. As explained in section 1.2.2, pharmacokinetic modelling allows us to calculate tissue parameters related to the local vasculature. These parameters can be presented as image overlays to the radiologist. In addition to their quantitative value (high values of $K^{trans}$ correspond to higher permeability, thus higher risk of cancer) they also make it easier to assess the focality or symmetry of lesions. An example of a lesion which was scored a 5 on DCE MRI is presented in Figure 1.11.

For DWI imaging the scoring of the lesion is performed using a high b-value image and the apparent diffusion coefficient map. A high b-value image ($\geqslant$800) is needed to minimize the T2 shine-through effect. Prostate cancer usually has a high signal intensity on the high b-value image and a low ADC value combined with a focal appearance on both images. An example of a 5-scored lesion is presented in Figure 1.12.

Last, spectroscopic imaging gives information about the metabolites citrate, choline and creatine. In prostate cancer choline concentrations tend to be higher than in normal tissue due to increased cell proliferation. The ESUR guidelines for spectroscopy specify the relative heights of the peaks for these metabolites with respect to the scores. Typically, the individual peaks are extracted using post-processing software and visualized as ratio parameter maps (similar to visualization of pharmacokinetic modeling in DCE MRI). An example is the CC-over-C ratio (choline and creatine over citrate), which is presented in Figure 1.13, with the corresponding post-processed spectrum.

The ESUR guideline specify that at least two functional modalities (DWI, DCE or spec-




(a) (b)

Figure 1.13: Overlay of the choline + creatine over citrate ratio on the T2-weighted imaging, including the spectroscopy grid. The selected voxel (dark blue outline) shows a high ratio of choline over citrate (a) and the corresponding spectrum shows a slightly higher peak for choline (pink) than for citrate (blue). As such a score of 4 was assigned for this lesion, in concordance with the ESUR guidelines.

troscopy) should be used in conjunction with T2-weighted imaging. They also indicate that a total score between 1 - 5 should be given for each lesion[33]. However, in the initial paper by Barentsz et al. no rules are specified for turning the per-parameter scores into a final lesion score. Some research papers have investigated using the sum of the parameter scores linearly (e.g. sum of 13 - 15 corresponds to 5, 10 - 12 corresponds to 4, etc.)[58,59]. However, in personal communication with Dr. Barentsz, he has expressed the preference assigning either the DWI score or the T2W score as the overall lesion score depending on the zone (DWI for the peripheral zone, T2W for the central gland).

After the guidelines were published several groups evaluated their performance[58–61]. Portalez et al.[58] showed that using the ESUR guidelines a AUC of 0.87 was obtainable in differentiating normal/benign and prostate cancer regions. Using a cut-off of 9 for the sum-of-parameter-scores the obtained sensitivity and specificity were 69.1 and 92.2%. Schimmöller et al. investigated the inter-reader agreement among three readers using Cohens' kappa in addition to reader performance in 67 consecutive patients. They found moderate-to-good inter-reader agreement between radiologists (T2W, $\kappa$=0.55; DWI, $\kappa$=0.64; DCE, $\kappa$=0.65). They found a sensitivity and specificity of 85.7 and 67.6% for a sum score cut-off of 10 and 92.2 and 47.1% for a cut-off of 9. Rosenkrantz et al.[59,60] also investigated both the inter-reader agreement and the overall performance of the guidelines. They found similar performance characteristics, however they also found that agreement between experienced and inexperienced reader was much less than between experienced readers (Concordance Correlation Coefficient of 0.609 versus 0.340 - 0.471).

### 1.2.4 MRI as a screening tool for prostate cancer

Currently, MRI is mostly used as a tool to diagnose patients with persistently high PSA levels and negative TRUS biopsies. However, given the much improved performance of MRI

over TRUS biopsies, it could also supplant them in a screening setting. Some studies have investigated the use of PSA and MRI as a screening tool. Thompson et al.[62] found a sensitivity of 97% for the detection of clinically significant prostate cancer (Gleason score $\geqslant$3+4) with a specificity of 0.50. Furthermore, Pokorny et al.[63] found that using multi-parametric MRI and MR-guided biopsy in biopsy-naïve men results in a large reduction of over-diagnosis of low-grade cancer (82%) while detecting 17% more intermediate/high-grade cancers than TRUS-guided biopsy.

One often mentioned argument is that MRI is too expensive for screening. Recently, de Rooij et al.[64] published a cost-effectiveness study comparing the current diagnostic pathway to one where TRUS biopsies are replaced with MRI and MR-guided biopsies. They showed that the overall cost is almost similar (2423 euro for MRI compared to 2392) and that the cost per quality-adjusted life year was reduced by 323 euros. The main reason for the lower cost per quality-adjusted life year is the substantially reduced amount of unnecessary biopsies and treatment. Moreover, markedly lower patient morbidity is expected.

As such, MRI is not only the best performing option for prostate cancer screening (potentially after initial PSA testing at the appropriate cut-off), but also of similar cost as PSA/TRUS. However, there are still challenges before MRI-based screening can become a reality. One issue is the availability of experienced radiologists to read the subsequent flood of MRIs, especially if one wants to implement double reading like in mammography screening. A potential solution for this issue is the implementation of computer-aided detection algorithms to function as a first or second reader.

## 1.3 Computer-aided detection and diagnosis

Computer-aided detection and diagnosis (CAD) in medical imaging (often named CADe and CADx to differentiate the two) is a field at the cross-roads of image analysis, medicine and machine learning. A very general description of CAD research is: "the use of computer algorithms to aid the image interpretation process"[65]. The very first publication on the use of computer algorithms to aid image interpretation was published in 1963 and focused on the analysis of pulmonary lesions in chest radiographs[66]. Since then, availability of computers, increase in computing power and the digitization of radiological images has led to an increase in publications in the field of CAD. The main subject of this thesis is the research and development of fully automated computer programs to detect and characterize prostate cancer on multi-parametric MRI.

### 1.3.1 General concepts

A complete CAD program is usually referred to as a "CAD system". This naming is apt, as a CAD system is not a single algorithm, but a pipeline involving multiple, distinct algorithms

which can individually be replaced without changing the rest of the system (algorithmically). Typically, CAD systems share many similar steps (or modules); the most common ones are: pre-processing, segmentation, feature calculation and classification. These are set up in a pipeline fashion with the result of one module feeding into the next. However, more complex pipelines with feedback loops are also used. Each of these modules is designed to tackle a specific problem.

Pre-processing steps are usually designed to make the original medical images suitable for the subsequent modules. Typical examples are noise reduction, contrast improvement, or edge enhancement. In the analysis of MRI images intensity differences between different scanner vendors or protocols are an issue as computers usually interpret image values absolutely, instead as relative as humans usually do. A pre-processing step could involve removing scanner and protocol dependencies from the signal intensities. An example of such a strategy is used in chapters 3 and 6. Another type of pre-processing is assessing whether the CAD system can actually perform subsequent steps given the input image. For example, if the image is of too poor quality it might be better to save computation time and flag the image as unusable. This prevents meaningless output, which in the end might reduce the confidence of the end-user in the system.

Segmentation is commonly meant to reduce the complexity of the classification task for the rest of the CAD system. As an example, forcing a CAD system to detect cancer in the entire abdomen is much more difficult than detecting cancer in just the prostate, not only because the search area is increased substantially, but also because some structures might exhibit similar characteristics as prostate cancer. For example, muscle appears dark on T2-weighted imaging, as does prostate cancer. Segmentation of the prostate is covered in chapter 2. However, segmentation is not only performed on an organ-basis. Lesions are also segmented with the goal of extracting features from the lesion area, or from the segmentation itself. Finally, segmentations do not have to be binary, they can also be probabilistic, giving a likelihood per pixel/voxel of belonging to a certain anatomical structure. In this way it can for example be used as a feature. This concept will be used in chapters 3 and 6.

Features are imaging characteristics which separate different classes, in this thesis prostate cancer from normal prostate tissue and benign prostate disease. Feature design for CAD systems is usually based on some inherent knowledge of the task. Extensive use was made of the features described by the ESUR guidelines. However, it is often difficult for humans to describe the features they subconsciously use. For example, it is hard to teach a computer what an "erased charcoal sign" is. As such feature design sometimes also depends on more general descriptors of image structure, an example of which are Gaussian derivatives[67,68]. This type of feature is less intuitive by itself, but a subsequent classifier can use the combination of features to obtain a meaningful class separation. Additionally, the recently renewed interest in "deep learning" has made sparse auto-encoders a popular way to discover features from images.

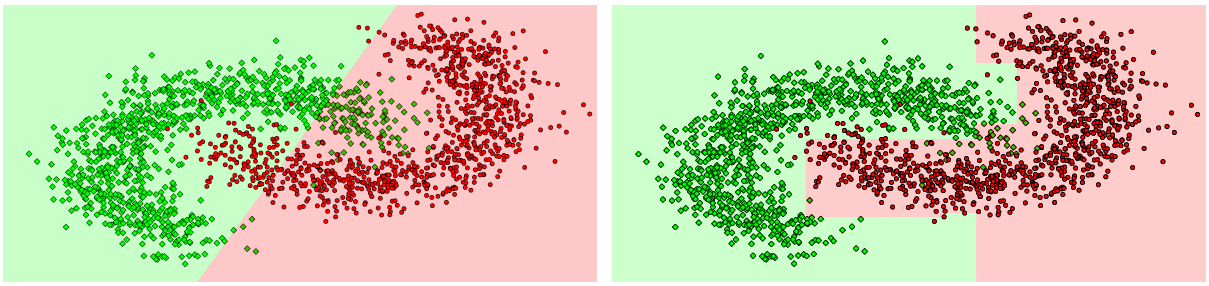
Sparse auto-encoders can use unlabeled images to automatically extract relevant features[69]. Both intuitive features and basic image descriptors for classification are used throughout this thesis. In addition to the features themselves, several dimensionality reduction and feature selection schemes are typically used in CAD system development to circumvent the so called curse of dimensionality. The curse of dimensionality refers to the exponential increase in the size of feature space and as such the required amount of training samples to cover this space. Dimensionality reduction or feature selection help reduce the size of the feature space and thus make a subsequent classification problem more manageable. Examples of dimensionality reduction are principle component analysis and sparse coding. Maximum relevance minimum redundancy (mRMR) feature selection[70], correlation feature selection[71] (CFS) and sequential forward floating feature selection (SFFS) are example of feature selection strategies, of which the latter is used in Chapter 4.

Classification is the task of assigning labels to unlabeled samples, in which each sample has one or more feature values. Classifiers do this by constructing a decision boundary, a hyperplane through feature space which separates different classes. Classifiers can be supervised or unsupervised, in this thesis we focus on supervised classification. Supervised classification requires a training step in which the classifiers learns the decision boundary from labeled data. There are roughly two types of supervised classifiers: parametric and non-parametric. Parametric classifiers impose a distribution on the data, usually a Gaussian distribution. Two examples of such classifiers are the linear discriminant classifier (LDC) and the quadratic discriminant classifier[72]. Non-parametric classifiers do not assume a specific data distribution, examples of such classifiers are the k-Nearest Neighbor (kNN) classifier[72], the GentleBoost classifier (GBC)[73] and the random forest classifier (RFC)[74]. Classifiers also differ in complexity, the linear discriminant classifier is a relatively simple classifier with a linear decision boundary whereas a GentleBoost or random forest classifier can learn much more complex decision boundaries. A general rule of thumb is that the more complex the classifier, the more training data is needed is to obtain an accurate decision boundary. An example of a decision boundary constructed by both an LDC and GBC is shown in Figure 1.14.

After classification usually a label or likelihood is obtained. In the first case the classifier will simply predict the most likely class label for a sample, in the latter case it will give a likelihood per class. This output can then be used as is, or presented to a human interpreter to aid in the decision making process.

### 1.3.2 CAD in clinical practice

The first computer-aided detection system was approved by the Food and Drug Administration (FDA) for commercial use in 1998. This CAD system indicated location of potential breast cancer in mammograms after an initial read by a radiologist. As such it was designed to



(a) (b)

Figure 1.14: Example of classifier decision boundaries for artificially generated data. Figure (a) shows the decision boundary of a linear discriminant classifier (LDC) and Figure (b) of a GentleBoost classifier (GBC). One can observe that the linear nature of the LDC decision boundary is not adequate for correctly classifying this type of data.

reduce the number of missed cancers. Since then several CAD systems have made their way to the clinic for a wide variety of applications ranging from analysis of hand radiographs for bone age reading[75] to detection of lung nodules in chest radiographs[76,77].

The added value of computer-aided detection has been evaluated in several clinical trials, with mixed results[78–82]. Most trials focus on the use of computer-aided detection in mammography, where it is historically most widely used due to the implementation of breast cancer screening programs. In the study by Gilbert et al. it was found that single reading with CAD was as effective as double reading by two radiologist[78]. However, Fenton et al. found that reading with CAD resulted in increased recall rates without any significant effect on early-detected breast cancer[82].

Due to the inconclusive results on the usefulness of the current implementation of CAD algorithms, several research groups have investigated alternative approaches to CAD. One such approach is the use of CAD systems as an independent first reader or second reader[83–85]. Another option is the interactive use of CAD, which advocates the idea that missed lesions are the result of a wrong characterization instead of the lesion being overlooked[86,87]. In chapter 7 the use of CAD as an independent first/second reader was investigated.

### 1.3.3 CAD for prostate cancer on MRI

Computer-aided detection of prostate cancer on multi-parametric MRI is a relatively young field, with the first papers appearing in 2003. Since then several research groups have shown interest, which has lead to a number of publications. Chan et al. were the first to implement a multi-parametric CAD system for the detection of prostate cancer[88] . In their approach they used line-scan diffusion, T2 and T2-weighted images in combination with a support vector machine (SVM) classifier to classify predefined areas of the peripheral zone of the prostate for the presence of prostate cancer. Langer et al.[89] included dynamic-contrast enhanced images and pharmacokinetic parameter maps as extra features to a CAD system for prostate

peripheral zone cancer. Puech et al.[90] implemented a computer-aided diagnosis system for both the peripheral and transition zones based on the dynamic contrast enhanced images. Artan et al.[91] used cost-sensitive conditional random fields to assess the detection performance of a multi-parametric CAD system compared to a CAD system based on the individual MR images. Liu et al.[92] presented a CAD system for peripheral zone prostate cancer which does not need an explicit segmentation of the peripheral zone by including anatomical position features. Tiwari et al.[93,94] investigated the use of magnetic resonance spectroscopy in combination with T2-weighted imaging to identify the spectroscopy voxels that are affected by prostate cancer. They also introduced the use of wavelet embedding to map MRS and T2-W texture features into a common space. This work was further expanded and evaluated in[94]. Viswanath et al.[95] showed that cancer in different zones has quantifiable differences in appearance. Lastly, Vos et al.[96] recently implemented a two-stage computer-aided detection system for prostate cancer using an initial blob detection approach combined with a candidate segmentation and classification using statistical region features. Hambrock et al. assessed the potential of computer-aided diagnosis in improving reader performance in prostate MRI[97]. They showed that both inexperienced readers and experienced readers could increase their area under the receiver operating characteristic curve when assessing whether pre-defined regions were prostate cancer or not. Overall, inexperienced readers were able to achieve near-expert performance (AUC=0.91 versus 0.93) when using CAD.

However, what all these methods lack is fully automated analysis of prostate MRI images for all prostate zones at a performance level close to the radiologist. The only fully automated system that takes a multi-parametric MRI as input and outputs a number of regions with associated cancer likelihoods was the system presented by Vos et al, however, at a performance which was yet too low for clinical implementation. A fully automated system is needed before efficiency and the lack of experienced radiologists can be addressed. Such a system could be used independently of the radiologist to characterize prostate MRI at a high sensitivity level, thus reducing the workload for the experienced radiologists. Furthermore, using a different operating point, it could potentially replace the role of a second observer in a double reading setting. Last, it could also be used as an aid to the radiologist reading prostate MRI by providing a 'second opinion' for radiologist-indicated regions. By operating in these three settings the CAD system could solve the problem of the lack of experienced radiologists.

## 1.4 Performance evaluation and statistical tests

### 1.4.1 General concepts

To investigate whether research findings are meaningful and not caused by chance, statistical significance testing is used. In this thesis one of two strategies was employed: logistic re-

gression analysis or receiver operating characteristic analysis. The former was used to assess the added value of features and clinical parameters in the diagnosis of prostate cancer and the latter to compare the diagnostic performance of humans and computer-aided detection systems.

*Logistic regression*

Logistic regression measures the relationship between a categorical dependent variable (for example prostate cancer/not prostate cancer) and one or more independent variables[98]. Binomial logistic regression uses the logistic function to map any combination of continuous descriptors (with any possible value between $-\infty$ and $+\infty$) to an output between 0 and 1, which can be interpreted as a likelihood. The logistic function is defined as:

$$F(t) = \frac{e^t}{e^t + 1} = \frac{1}{1 + e^{-t}} \tag{1.8}$$

$$t = \beta_0 + \beta_1 x_1 + \dots \tag{1.9}$$

where $t$ can be any linear combination of explanatory descriptors. Given a set of samples, maximum likelihood estimation is used to fit the logistic regression model, usually using optimization based on Newton's method. In general, when enough samples are used relative to the amount of descriptors, the optimization procedure will converge and result in estimates for the regression coefficients $\beta_n$. Using the likelihood ratio test one can assess whether a model is a significant improvement over a model with less descriptors. To perform the likelihood ratio test the deviance needs to be calculated:

$$D = -2 \ln \frac{\text{likelihood of the fitted model}}{\text{likelihood of the saturated model}} \tag{1.10}$$

here the likelihood of the saturated model is the likelihood of a model with perfect fit to the data. The difference in the deviances of the two models one wants to compare can then be tested using a chi-square distribution with degrees of freedom equal to difference in the number of descriptors estimated.

The value of the regression coefficient can be used to interpret the size of the effect a descriptor has on the overall output. Assuming that the descriptors are similarly scaled, the further the corresponding regression coefficient deviates from zero, the bigger the impact of the predictor on the output.

*Sensitivity, specificity and the receiver operating characteristic curve*

Any classification task will lead to errors, whether it is performed by humans or by a computer system. In a two-class classification problem there are four outcomes: a sample is correctly classified as class 1, a sample is incorrectly classified as class 1, a sample is correctly classified

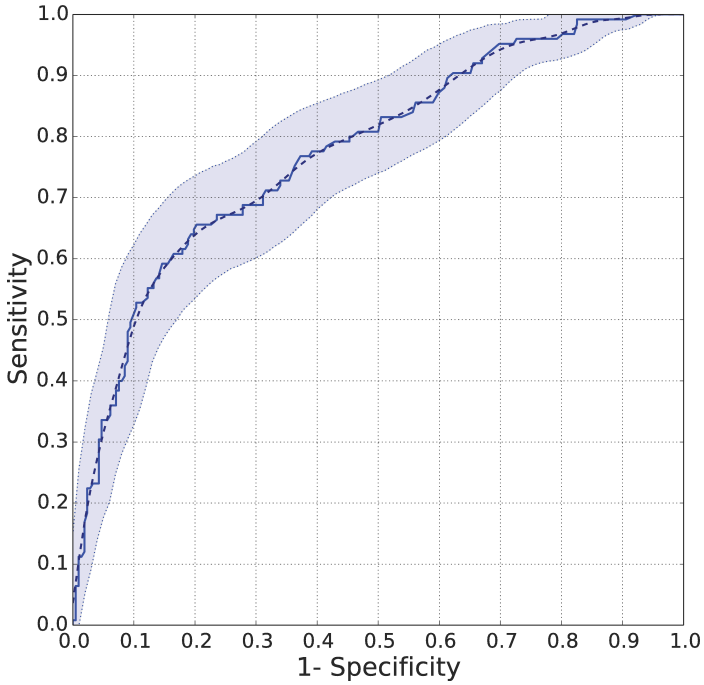



Figure 1.15: An example of a receiver operating characteristic curve. In this figure the mean bootstrap curve (dashed line) and the 95% confidence intervals (transparent area) are also plotted.

as class 0 or sample is incorrectly classified as class 0. Assuming class 1 is the target class (in this thesis prostate cancer) these outcomes are a true positive (TP), false positive (FP), true negative (TN) or false negative (FN) respectively. Given the total number of TP, FP, TN and FNs across a set of data the sensitivity and specificity of the classifier can be calculated. The sensitivity and specificity are defined as:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP+FN}} \tag{1.11}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN+FP}} \tag{1.12}$$

Sensitivity can be interpreted as the percentage of samples in class 1 which were correctly classified as class 1, or in clinical terms, the cases having a disease correctly identified as having the disease. Specificity is the percentage of samples in class 0 which are correctly identified as class 0, or again in clinical terms, the cases which are healthy correctly identified as being healthy. Usually there is a direct trade-off between sensitivity and specificity, increasing one will decrease the other.

When a classification task doesn't give a binary prediction, but an ordinal one (e.g. the ESUR lesion score, or likelihood obtained from a logistic regression model) multiple pairs of sensitivity and specificity can be generated by changing the threshold at which a sample is classified as class 1. Using these pairs a receiver operating characteristic (ROC) curve can be generated (Figure 1.15). The area under this curve (AUC) can be calculated, which is the probability that, given a random positive and negative sample, the positive sample will have

the higher score. The minimum value of the AUC is 0 and the maximum value 1, although typically 0.5 is considered the worst value as it corresponds to straight-out guessing. The maximum value of 1 corresponds to perfect classification. The AUC is not always the most appropriate measure to evaluate a diagnostic test, sometimes single pair of sensitivity/specificity or a partial area under the curve can be more informative, depending on the task. For example, in a screening setting, where the vast majority of patients is healthy, high sensitivity is only relevant within the part of the curve with high specificity.

Sometimes it can be informative to explicitly include the prevalence of the classes in the performance measure to get a better understanding what it means in a practical scenario. In those cases a positive predictive value (PPV) and negative predictive value (NPV) can be calculated, which are defined as:

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP+FP}} \tag{1.13}$$

$$\mathrm{NPV} = \frac{\mathrm{TN}}{\mathrm{TN+FN}} \tag{1.14}$$

In addition to ROC analysis, which assumes one outcome per case, free-response receiver operating characteristic (FROC) is often used. In FROC instead of specificity, the number of false positives per case are used, as multiple false detection per scan can typically occur. From FROC results a ROC can be calculated by converting the multiple detections per case into a single likelihood.

Statistically evaluating the performance of different diagnostic tests or CAD systems in thesis is usually performed using bootstrapping[99,100]. Bootstrapping is a non-parametric way to statistically evaluate differences in performance. In bootstrapping the output data is re-sampled with replacement. Each re-sample is one bootstrap sample and from this sample the performance metric can be calculated, e.g. the AUC. When two methods are statistically compared they are re-sampled in exactly the same way so a paired test can be performed. The p-value can then be calculated by counting the amount of bootstrap samples of one method which give a higher result than the other method.

In addition to bootstrapping in chapter 5 the ROCKIT software package was used to statistically compare ROC curves[101]. This package first fits a bi-normal model to diagnostic test outputs for each class and can then subsequently calculate the ROC curve, AUC and the 95% confidence intervals.

## 1.5 Thesis outline

The main objective of this thesis was to design a computer-aided detection system to detect cancer in prostate MRI which could be used in clinical practice. Although several papers already describe these types of systems numerous challenges remain: the system needs be

fully automated, it needs to work both in the peripheral zone and the central gland and it should approach the performance of a human expert. Meeting these requirements will allow us to solve one of the issues currently prohibiting MRI-based prostate cancer screening: the lack and cost of experienced radiologists. Furthermore, a computer-aided detection system could help improve the performance of individual radiologists. This thesis tries to meet and solve the presented challenges and our results are presented in the subsequent chapters.

In chapter 2 the preparation, organization and results of a prostate MRI segmentation challenge (PROMISE12) are discussed. The goal of the challenge was to identify the performance of different categories and implementations of algorithms that segment the prostate. Multi-center, multi-protocol and multi-vendor data was used to evaluate the methods to ensure that algorithms which showed good performance would generalize well to different centers.

Chapter 3 subsequently focuses on the use of pattern recognition methods to segment the peripheral zone and central gland, both probabilistically and binary. It also includes a method to reduce inter-scanner and inter-protocol variation on T2-weighted imaging. Segmentation results were compared to those of three different observers.

In chapter 4 and 5 the design of discriminative features for assessing cancer aggressiveness and separating cancer from benign confounding classes like BPH is presented. Prostatectomy specimens were mapped to the MRI to accurately delineate specific lesions. The use of logistic regression and SFFS feature selection to identify useful features from a large initial feature set was investigated.

Chapter 6 is concerned with the development of the CAD pipeline, including segmentation, features, classification and initial evaluation on a large retrospective data set. A two-stage classification pipeline is built using initial voxel classification and subsequent false positive reduction.

Chapter 7 details the evaluation of the CAD system on a prospective set of data and the usefulness of the system in a clinical setting where it could function as an independent second reader. The focus is on the improvement in reader performance and the correlation of the CAD system likelihoods with cancer aggression.

# Segmentation

# Segmentation of the prostate capsule

# 2

Geert Litjens, Robert Toth, Henkjan Huisman, Anant Madabhushi et al.

*Original title:* Evaluation of prostate segmentation algorithms for MRI: the PROMISE12 challenge

*Published in:* Medical Image Analysis (2014);18(2);359–373

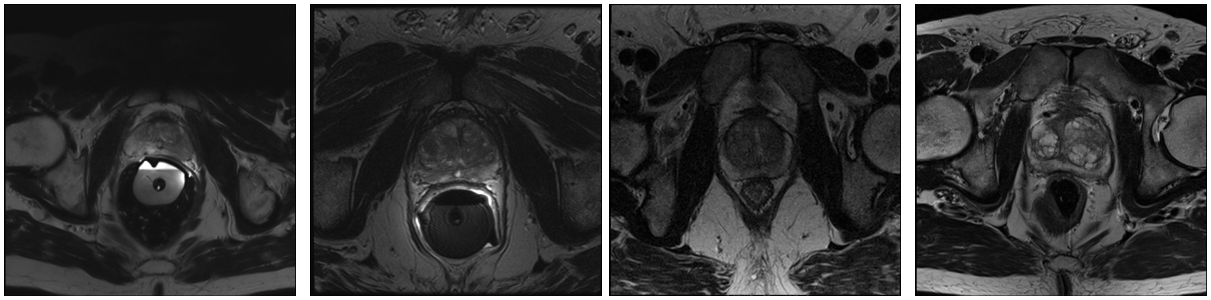
## 2.1 Introduction

Prostate MRI image segmentation has been an area of intense research due to the increased use of MRI as a modality for the clinical workup of prostate cancer, e.g. diagnosis and treatment planning[28,31,32,102,103]. Segmentation is useful for various tasks: to accurately localize prostate boundaries for radiotherapy[104], perform volume estimation to track disease progression[105], to initialize multi-modal registration algorithms[106] or to obtain the region of interest for computer-aided detection of prostate cancer[94,96], among others. As manual delineation of the prostate boundaries is time consuming and subject to inter- and intra-observer variation, several groups have researched (semi-)automatic methods for prostate segmentation[104,107–112]. However, as most algorithms are evaluated on proprietary datasets a meaningful comparison is difficult to make.

This problem is aggravated by the fact that most papers cannot include a comparison against the state-of-the-art due to previous algorithms being either closed source or very difficult to implement without help of the original author. Especially in MRI, where signal intensity is not standardized and image appearance is for a large part determined by acquisition protocol, field strength, coil profile and scanner type, these issues present a major obstacle in further development and improvement of prostate segmentation algorithms.

In recent years several successful 'Grand Challenges in Medical Imaging' have been organized to solve similar issues in the fields of liver segmentation on CT[113], coronary image analysis[114], brain segmentation on MR[115], retinal image analysis[116] and lung registration on CT[117]. The general design of these challenges is that a large set of representative training data is publicly released, including a reference standard for the task at hand (e.g. liver segmentations). A second set is released to the public without a reference standard, the test data. The reference standard for the test data is used by the challenge organizers to evaluate the algorithms. Contestants are then allowed to tune their algorithms to the training data after which their results on the test data are submitted to the organizers who calculate predefined evaluation measures on these test results. The objective of most challenges is to provide independent evaluation criteria and subsequently rank the algorithms based on these criteria. This approach overcomes the usual disadvantages of algorithm comparison, in particular, bias.

The Prostate MR Image Segmentation (PROMISE12) challenge presented in this paper tries to standardize evaluation and objectively compare algorithm performance for the segmentation of prostate MR images. To achieve this goal a large, representative set of 100 MR images was made available through the challenge website: `http://promise12.grand-challenge.org/`. This set was subdivided into training (50), test (30) and live challenge (20) datasets (for further details on the data, see section 2.2). Participants could download the data and apply their own algorithms. The goal of the challenge was to accurately segment the prostate capsule. The calculated segmentations on the test set were then submitted to the challenge



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

Figure 2.1: Slice of a data set from different centers to show appearance differences. Figure (a) is from Haukeland University Hospital, Norway and was acquired at 1.5T with an endorectal coil. Figure (b) is from Beth Israel Deaconess Medical Center, USA and was acquired at 3.0T with an endorectal coil. Figure (c) if from University College London, United Kingdom acquired at 1.5T and 3.0T without an endorectal coil. Figure (d) is from the Radboud University Medical Centre, The Netherlands and was acquired at 3.0T without an endorectal coil.

organizers through the website for independent evaluation. Evaluation of the results included both boundary and volume based metrics to allow a rigorous assessment of segmentation accuracy. To calculate an algorithm score based on these metrics, they were compared against human readers. Further details about generation of the algorithm score can be found in section 2.3.2.

This paper will describe the setup of the challenge and the initial results obtained prior to and at the workshop hosted by the MICCAI2012 conference in Nice, where a live challenge was held between all participants. New results, which can still be submitted through the PROMISE12 website, can be viewed online.

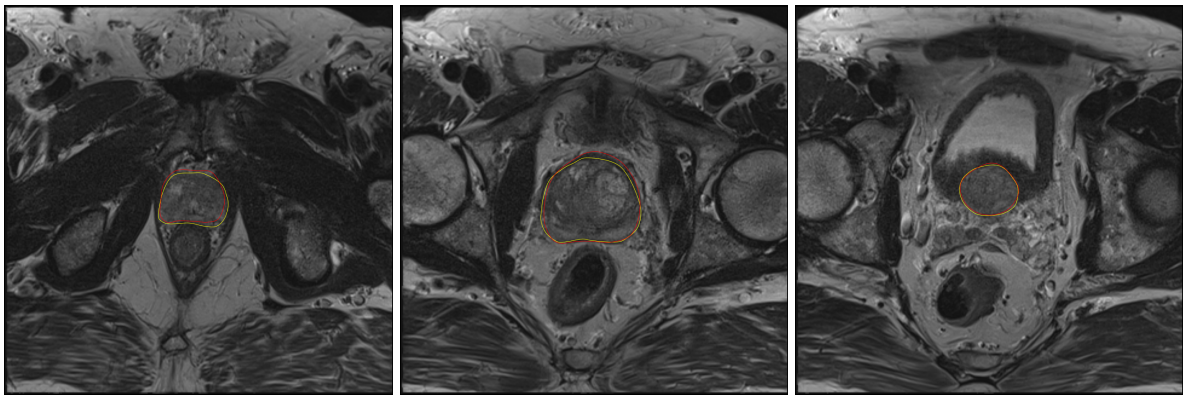
## 2.2 Materials

### 2.2.1 MRI images

In MRI images, the pixel/voxel intensities and therefore appearance characteristics of the prostate can greatly differ between acquisition protocols, field strengths and scanners[33,54]. Example causes of appearance differences include the bias field[118,119], signal-to-noise ratio[120,121] and resolution[31,32], especially through-plane. Additionally, signal intensity values are not standardized[122,123]. Therefore a segmentation algorithm designed for use in clinical practice needs to deal with these issues[124,125]. Consequently, we decided to include data from four different centers: Haukeland University Hospital (HK) in Norway, the Beth Israel Deaconess Medical Center (BIDMC) in the US, University College London (UCL) in the United Kingdom and the Radboud University Nijmegen Medical Centre (RUNMC) in the Netherlands. Each of the centers provided 25 transverse T2-weighted MR images. This resulted in a total of 100 MR images. Details pertaining to the acquisition can be found in Table 2.1. Additionally, a cen-

| Center | Field Strength | Endorectal coil | Resolution (in-/through-plane in mm) | Manufacturer |
|---|---|---|---|---|
| HK | 1.5T | Yes | 0.625 / 3.6 | Siemens |
| BIDMC | 3T | Yes | 0.25 / 2.2 - 3 | GE |
| UCL | 1.5 and 3T | No | 0.325 - 0.625 / 3 - 3.6 | Siemens |
| RUNMC | 3T | No | 0.5 - 0.75 / 3.6 - 4.0 | Siemens |

Table 2.1: Details of the acquisition protocols for the different centers. Each center supplied 25 T2-weighted MR images of the prostate.



(a) (b) (c)

Figure 2.2: Example T2-weighted transverse prostate MRI images displaying an apical, central and basal slice. The reference standard segmentation is shown in yellow and the second observer segmentation in red. Figures (a), (b) and (c) are at the apex, center and base of the prostate respectively.

tral slice of a data set for each of the centers is shown in Figure 2.1 to show the appearance differences. These scans where acquired either for prostate cancer detection or staging purposes. However, the clinical stage of the patients and the presence and location of prostate cancer is unknown to the organizers. Transverse T2-weighted MR was used because these contain most anatomical detail[33], are used clinically for prostate volume measurements[103,105] and because most current research papers focus on segmentation on T2-weighted MRI. The data were then split randomly into 50 training cases, 30 test cases and 20 live challenge cases. Although the selection process was random, it was stratified according to the different centers to make sure no training bias towards a certain center could occur.

### 2.2.2 Segmentation Reference Standard

Each center provided a reference segmentation of the prostate capsule performed by an experienced reader. All annotations were performed on a slice-by-slice basis using a contouring tool. The contouring tool itself was different for the different institutions, but the way cases were contoured was similar. Contouring was performed by annotating spline-connected points in either 3DSlicer (`www.slicer.org`) or MeVisLab (`www.mevislab.de`). The reference segmen-

tations were checked by a second expert, C.H., who has read more than 1000 prostate MRIs, to make sure they were consistent. This expert had no part in the initial segmentation of the cases and was asked to correct the segmentation if inconsistencies were found. The resulting corrected segmentations were used as the reference standard segmentation for the challenge. An example of a reference segmentation at the base, center and apex of the prostate is shown in Figure 2.2.

### 2.2.3 Second Observer

For both the testing and the live challenge data a relatively inexperienced nonclinical observer (W.v.d.V, two years of experience with prostate MR research) was asked to manually segment the prostate capsule using a contouring tool. The second observer was blinded to the reference standard to make sure both segmentations were independent. The second observer segmentations were used to transform the evaluation metrics into a case score, as will be explained in section 2.3.2. An example of a second observer segmentation is shown in Figure 2.2.

## 2.3 Evaluation

### 2.3.1 Metrics

The metrics used in this study are widely used for the evaluation of segmentation algorithms:

1. the Dice coefficient (DSC)[108,113]
2. the absolute relative volume difference, the percentage of the absolute difference between the volumes (aRVD)[113]
3. the average boundary distance, the average over the shortest distances between the boundary points of the volumes (ABD)[113]
4. the 95% Haussdorf distance (95HD)[111]

All evaluation metrics were calculated in 3D. We chose both boundary and volume metrics to give a more complete view of segmentation accuracy, i.e. in radiotherapy boundary based metrics would be more important, whereas in volumetry the volume metrics would be more important. In addition to evaluating these metrics over the entire prostate segmentation, we also calculated them specifically for the apex and base parts of the prostate, because these parts are very important to segment correctly, for example in radiotherapy and TRUS/MR fusion. Moreover, these are the most difficult parts to segment due the large variability and slice thickness. To determine the apex and base the prostate was divided into three approximately equal parts in the slice dimension (the caudal 1/3 of the prostate volume was considered apex,

the cranial 1/3 was considered base). If a prostate had a number of slices not dividable by 3 (e.g. 14), the prostate would be divided as 4-6-4 for the base, midgland and apex respectively.

The DSC was calculated using:

$$D(X,Y) = \frac{2|X \cap Y|}{|X| + |Y|} \tag{2.1}$$

where $|X|$ is the number of voxels in the reference segmentation and $|Y|$ is the number of voxels in the algorithm segmentation.

The relative volume difference was calculated as:

$$\text{RVD(X,Y)} = 100 \times \left( \frac{|X|}{|Y|} - 1 \right) \tag{2.2}$$

and thus the absolute relative volume difference is

$$\text{aRVD(X,Y)} = |\text{RVD}(X,Y)| \tag{2.3}$$

Note that although we use the aRVD to measure algorithm performance (both under- and over-segmentation are equally bad), in the results we will present the RVD, which makes it possible to identify if algorithms on average tend to over- or under-segment the prostate.

For both the 95th percentile Hausdorff distance and the average boundary distance we first extract the surfaces of the reference segmentation and the algorithm segmentation. The regular Hausdorff distance is then defined as:

$$\text{HD}_{\text{asym}}(X_s, Y_s) = \max_{x \in X_s} \left( \min_{y \in Y_s} d(x,y) \right) \tag{2.4}$$

$$\text{HD}(X_s, Y_s) = \max \left( \text{HD}_{\text{asym}}(X_s, Y_s), \text{HD}_{\text{asym}}(Y_s, X_s) \right) \tag{2.5}$$

where $X_s$ and $Y_s$ are the sets of surface points of the reference and algorithm segmentations respectively. The operator d is the Euclidean distance operator. As the normal Hausdorff distance is very sensitive to outliers we use the 95th percentile of the asymmetric Hausdorff distances instead of the maximum.

Finally, the average boundary distance (ABD) is defined as:

$$\text{ABD}(X_s, Y_s) = \frac{1}{N_{X_s} + N_{Y_s}} \left( \sum_{x \in X_s} \min_{y \in Y_s} d(x,y) + \sum_{y \in Y_s} \min_{x \in X_s} d(y,x) \right) \tag{2.6}$$

### 2.3.2 Score

Algorithms were ranked by comparing the resulting evaluation measures to the second observer and the reference segmentation in a way similar to Heimann et al. [113]. First, the metrics of the second observer segmentations are calculated with respect to the reference segmentation. Then we average each metric over all cases and define a mapping function:

$$\text{score}(x) = \max(ax + b, 0) \tag{2.7}$$

This function maps a metric value $x$ to a score between 0 and 100. The equation is solved for $a$ and $b$ by setting a score of 100 to a perfect metric result, e.g. a DSC of 1.0 and setting a score of 85 to a metric result equal to the average metric value of the second observer. This will give us two equations to solve the two unknowns, $a$ and $b$. Additionally, a score of zero was set as the minimum because otherwise cases with a very poor or missing segmentation could bias the final score of an algorithm too much. As an example, if the second observer segmentations have an average DSC of 0.83, $a$ and $b$ are 88.24 and 11.76 respectively. As such, if an algorithm obtains a DSC of 0.87 on a case the score will be 88.53. This approach is applied to all metrics. The scores for all metrics were averaged to obtain a score per case. Then the average over all cases was used to rank the algorithms.

A relatively high reference score of 85 was chosen for the second observer because her segmentations were in excellent correspondence with the reference standard. An even higher score than 85 would not be warranted, as the segmentations still contain errors experienced observers would not make. The average metric scores for the second observer are presented in Tables 2.6 and 2.7. Comparing these metric scores to scores reported in literature for inter-observer variability we can see that they are at approximately at the same level[104,107–112].

The main reason to use this approach is that it allows us to incorporate completely different, but equally important metrics like average boundary distance and the Dice coefficient. Furthermore, in addition to allowing us to rank algorithms, the scores themselves are also meaningful, i.e. higher scores actually correspond to better segmentations. An alternative approach could have been to rank algorithms per metric and average the ranks over all metrics. However, such an average rank is not necessarily related to a segmentation performance: the best ranking algorithm could still show poor segmentation results that are much worse than the second observer.

## 2.4 Methods

This section gives an overview of all the segmentation methods that participated in the challenge. A short description for each algorithm is given. More detailed descriptions of the algorithms can be found in peer-reviewed papers submitted to the PROMISE12 challenge, available at: `http://promise12.grand-challenge.org/Results`. Algorithms were categorized as either automatic (no user interaction at all), semi-automatic (little user interaction, e.g. setting a single seed point) or interactive (much user interaction, e.g. painting large parts of the prostate). The algorithm categories and additional details can be found in Tables 2.3 and 2.8. The names in subsection titles are the team names chosen by the participants and are as such not related to the method themselves. Most names are either abbreviations of group names or company names. Links to the websites of the individual groups can also be found on the PROMISE12-website.

### 2.4.1 Fully automatic segmentation of the prostate using active appearance models - Imorphics

Vincent et al.[126] of Imorphics Ltd. have developed a generic statistical modeling system designed to minimize any bespoke development needed for different anatomical structures and image modalities.

The Imorphics system generates a set of dense anatomical landmarks from manually segmented surfaces using a variant of the Minimum Description Length approach to Groupwise Image Registration[127]. The correspondence points and associated images are used to build an Appearance Model. The Appearance Model is matched to an unseen image using an Active Appearance Model (AAM) which optimizes the model parameters to generate an instance which matches the image as closely as possible[128].

Active Appearance Models require an initial estimate of the model parameters including position, rotation and scale. The system uses a multi-resolution gridded search method. This is started at a low image and model resolution with a small number of measured residuals to make it reasonably fast. The results of these searches are ranked according to the sum of squares of the residual, and a proportion removed from consideration. The remaining search results are used to initialize models at a higher resolution, and so on. Finally, the single best result at the highest resolution gives the segmentation result.

### 2.4.2 Region-specific hierarchical segmentation of MR prostate using discriminative learning - ScrAutoProstate

The segmentation pipeline developed by Birkbeck et al.[129] addresses the challenges of MR prostate segmentation through the use of region-specific hierarchical segmentation with discriminative learning.

First, an intensity normalization is used to adjust for global contrast changes across the images. Images with an endorectal coil are then further enhanced by flattening the intensity profile on the bright regions near the coil using an automatic application of Poisson image editing[130].

In the next phase of the pipeline, a statistical model of mesh surface variation learned from training data is aligned to the normalized image. The pose parameters of the shape model are extracted through the use of marginal space learning[131], which decomposes the estimation of pose into sequential estimates of the position, orientation, scale, and then the first few modes of variation. The estimation of each set of pose parameters relies on a probabilistic boosting tree classifier to discriminatively model the relationship between the image data and the unknown parameters being estimated. During training, each classifier automatically selects the most salient features from a large feature pool of Haar and steerable features. After the statistical mesh model has been aligned to the input image using marginal space learning, the

segmentation is refined through a coarse-to-fine boundary refinement that uses surface varying classifiers to discriminate the boundary of the prostate from adjacent soft tissue. The mesh from this final refinement stage is constrained by the statistical shape model.

### 2.4.3 Smart paint - CBA

Malmberg et al.[132] have developed an interactive segmentation tool called Smart Paint. The user segments the organ of interest by sweeping the mouse cursor in the object or background, similar to how an airbrush is used. Areas are painted with a semi-transparent color which gives immediate feedback in the chosen interaction plane. As the paint is applied in 3D, when the user moves to another plane using the mouse thumbwheel the effect of the painting is seen also there.

The algorithm works by taking both the spatial distance to the cursor and the image content (intensity values) into account. The image I and the segmentation function f are mappings from elements of a three dimensional voxel set to the interval [0,1]. A voxel x belongs to the foreground if $f(x) \geqslant 0.5$, and to the background otherwise. Initially, $f = 0$. The brush tool has a value v that is either 1 (to increase the foreground) or 0 (to increase the background). A single brush stroke centered at voxel x affects the segmentation at all nearby voxels y according to

$$f(y) \leftarrow (1 - \alpha(x, y))f(y) + \alpha(x, y)v \tag{2.8}$$

$$\alpha(x, y) = \beta(1 - |I(y) - I(x)|)^k \max(\frac{(r - d(x, y))}{r}, 0) \tag{2.9}$$

where $d(x, y)$ is the Euclidean distance between the voxel centers of $x$ and $y$, $r$ is the brush radius specified by the user and $\beta$ and $k$ are constants.

Additionally, the user can smooth the current segmentation using a weighted average filter. The algorithm is not very sensitive to the values selected for the $\beta$ and $k$ constants. Values for $\beta$ were in the range 0.01 - 0.1 and for $k$ in the range 1-5 and influence the behavior of the brush. These variables could be changed by the user

### 2.4.4 Multi-atlas segmentation of the prostate: a zooming process with robust registration and atlas selection - SBIA

The multi-atlas based segmentation framework designed by Ou et al.[133] automatically segments the prostate in MR images. Atlases from 50 training subjects are nonrigidly registered to the target image. The calculated deformations are used to warp expert annotated prostate segmentations of the atlases into the target image space. The warped prostate annotations are then fused by the STAPLE strategy[134] to form a single prostate segmentation in the target image.

The main challenge in this multi-atlas segmentation framework is image registration. To account for the registration challenges, three measures are taken in the multi-atlas segmenta-

tion framework. First, the DRAMMS image registration algorithm is used[135]. DRAMMS establishes anatomical correspondences by using high dimensional texture features at each voxel. Voxel texture features are more distinct than just using intensity, which helps to improve registration accuracy. Second, a two-phase strategy is used. In phase 1 the entire prostate images from training subjects are used to compute an initial segmentation of the prostate in target image. Phase 2 focuses only on the initially segmented prostate region and its immediate neighborhood. Third, in each phase, atlas selection is used. Those atlases having high similarity with the target image in the prostate regions after registration are kept. Similarity is measured using the correlation coefficient, mutual information, as well as the DSC between the warped prostate annotation and the tentative prostate segmentation.

### 2.4.5 Automatic prostate segmentation in MR images with a probabilistic active shape model - Grislies

Kirschner et al.[136] segment the prostate with an Active Shape Model (ASM)[128]. For training the ASM, meshes were extracted from the ground truth segmentations using Marching Cubes[137]. Correspondence between the meshes was determined using a nonrigid mesh registration algorithm. The final ASM has 2000 landmarks and was trained using principal component analysis (PCA). The actual segmentation is done with a three step approach, consisting of 1) image preprocessing, 2) prostate localization and 3) adaption of the ASM to the image.

In the preprocessing step, the bias field is removed using coherent local intensity clustering, and the image intensities are normalized[138]. Prostate localization is done using the sliding window approach: a boosted classifier based on 3D Haar-like features is used to decide whether the subimage under the current detector window position contains the prostate or not. This approach is similar to the Viola-Jones algorithm for face detection in 2D images[139], .

The actual segmentation is done with a Probabilistic ASM. In this flexible ASM variant, shape constraints are imposed by minimizing an energy term which determines a compromise between three forces: an image energy that draws the model towards detected image features, a global shape energy that enforces plausibility of the shapes with respect to the learned ASM, and a local shape energy that ensures that the segmentation is smooth. For detection of the prostate's boundary, a boosted detector using 1D Haar-like features is used, which classifies sampled intensity profiles into boundary and nonboundary profiles.

### 2.4.6 An efficient convex optimization approach to 3D prostate MRI segmentation with generic star shape prior - Robarts

The work by Yuan et al.[140] proposes a global optimization-based contour evolution approach for the segmentation of 3D prostate MRI images, which incorporates histogram matching and a variational formulation of a generic star shape prior.

The proposed method overcomes the existing challenges of segmenting 3D prostate MRIs: heterogeneous intensity distributions and a wide variety of prostate shape appearances. The proposed star shape prior does not stick to any particular object shape from learning or specified parameterized models, but potentially reduces ambiguity of prostate segmentation by ruling out inconsistent segments; it provides robustness to the segmentation when the image suffers from poor quality, noise, and artifacts.

In addition, a novel convex relaxation based method is introduced to evolve a contour to its globally optimal position during each discrete time frame, which provides a fully time implicit scheme to contour evolution and allows a large time step size to accelerate the speed of convergence.

Moreover, a new continuous max-flow formulation is proposed, which is dual to the studied convex relaxation formulation and derives a new efficient algorithm to obtain the global optimality of contour evolution. The continuous max-flow based algorithm is implemented on GPUs to significantly speed up computation in practice.

### 2.4.7 An automatic multi-atlas based prostate segmentation using local appearance specific atlases and patch-based voxel weighting - ICProstateSeg

Gao et al.[141] present a fully automated segmentation pipeline for multi-center and multi-vendor MRI prostate segmentation using a multi-atlas approach with local appearance specific voxel weighting.

An initial denoising and intensity inhomogeneity correction is performed on all images. Atlases are classified into two categories: normal MRI scans $A_n$ and scans taken with a transrectal coil $A_m$. This is easily achieved by examining the intensity variation around the rectum since the transrectal coil produces significant physical distortion but also has a characteristic bright appearance in the local region near the coil itself. The subatlas database whose atlas appearance is closest to the new target is chosen as the initial atlas database. After that, the top N similar atlases are further chosen for atlas registration by measuring intensity difference in the region of interest around prostate.

After all the selected atlases are nonrigidly registered to a target image, the resulting transformation is used to propagate the anatomical structure labels of the atlas into the space of the target image. Finally, a patch-based local voxel weighting strategy is introduced, which was recently proposed for use in patch-based brain segmentation[142] and improved by introducing the weight of the mapping agreement from atlas to target. After that, the label that the majority of all warped labels predict for each voxel is used for the final segmentation of the target image.

### 2.4.8 Prostate image segmentation using 3D active appearance models - Utwente

The segmentation method proposed by Maan and van der Heijden[143] is an adaptation of the work presented by Kroon et al.[144] by using a Shape Context based non-rigid surface registration in combination with 3D Active Appearance Models (AAM).

The first step in AAM training is describing the prostate surface in each training case by a set of landmarks. Every landmark in a training case must have a corresponding landmark in all other training cases. To obtain the corresponding points Shape Context based nonrigid registration of the binary segmentation surfaces was used[143,144]. PCA is applied to determine the principal modes of the shape variation. The appearance model can be obtained in a similar way: first each training image is warped so that its points correspond to the mean shape points. Subsequently, the grey-level information of the region covered by the mean shape is sampled. After normalization, a PCA is applied to obtain the appearance model. The combined shape and appearance model can generalize to almost any valid example.

During the test phase, the AAM is optimized by minimizing the difference between the test image and the synthesized images. The mean model is initialized by manually selecting the center of the prostate based on visual inspection. Subsequently, the AAM is applied using two resolutions with both 15 iterations.

### 2.4.9 A multi-atlas approach for prostate segmentation in MR images - DIAG

Litjens et al.[145] investigated the use of a multi-atlas segmentation method to segment the prostate using the Elastix registration package. The method is largely based on the work of Klein et al.[108] and Langerak et al.[146]. The 50 available training data sets are used as atlases and registered to the unseen image using localized mutual information as a metric. Localized mutual information calculates the sum of the mutual information of image patches instead of the mutual information of the entire image. This approach reduces the effect of magnetic field bias and coil profile on the image registration.

The registration process consists of two steps: first a rough initial alignment is found, after which an elastic registration is performed. The 50 registered atlases are then merged to form a signal binary segmentation using the SIMPLE optimization algorithm[146]. SIMPLE tries to automatically discard badly registered atlases in an iterative fashion using the correspondence of the atlas to the segmentation result in the previous iteration. The DSC was used as the evaluation measure in the SIMPLE algorithm.

### 2.4.10 Deformable landmark-free active appearance models: application to segmentation of multi-institutional prostate MRI data - Rutgers

Toth and Madabhushi[147] propose a Multi-Feature, Landmark-Free Active Appearance Model (MFA) based segmentation algorithm, based on[148]. The MFA contains both a training module

and a segmentation module. The MFA is constructed by first aligning all the training images using an affine transformation. Second, the shape is estimated by taking the signed distance to the prostate surface for each voxel, which represents a levelset, such that a value of 0 corresponds to the voxels on the prostate surface. Third, principal component analysis is used to map the shape and intensity characteristics of the set of training images to a lower dimensional space. Then a second PCA is performed on the joint set of lower dimensional shape and appearance vectors to link the shape and appearance characteristics.

To segment an unseen image, the image must be registered to the MFA, resulting in transformation T mapping the input image to the MFA. This is performed by first calculating the PCA projection of the intensities learned from the training data. Then the linked projections are reconstructed and subsequently the intensities and shape. The normalized crosscorrelation between the reconstruction and the original image are calculated and the transform T is optimized to obtain maximum normalized cross-correlation. The shape corresponding to the optimal transformation was thresholded at 0 to yield the final segmentation.

While the original algorithm[148] defined "T" as an affine transformation, to account for the high variability in the prostate shape and appearance (e.g. with or without an endorectal coil), a deformable, b-spline based transform was used to define "T". This resulted in a more accurate registration than affine, although further studies suggest that separate subpopulation based models could potentially yield more accurate segmentations, given enough training data.

### 2.4.11 A random forest based classification approach to prostate segmentation in MRI - UBUdG

The method proposed by Ghose et al.[149] has two major components: a probabilistic classification of the prostate and the propagation of region based levelsets to achieve a binary segmentation. The classification problem is addressed by supervised random decision forest.

During training, the number of slices in a volume containing the prostate is divided into three equal parts as apex, central and base regions. The individual slices are resized to a resolution of 256x256 pixels and a contrast-limited adaptive histogram equalization is performed to minimize the effect of magnetic field bias. Each feature vector is composed of the spatial position of a pixel and the mean and standard deviation of the gray levels of its 3 x 3 neighborhood. Three separate decision forests are built corresponding to the three different regions of the prostate the apex, the central region and the base. Only 50% of the available training data was used for each of the regions.

During testing the first and the last slices of the prostate are selected and the test dataset is divided into the apex, the central and the base regions. Consecutively preprocessing is done on in the same way as for the training images. Decision forests trained for each of the regions are applied to achieve a probabilistic classification of the apex, the central and the base slices.

Finally evolution of the Chan and Vese levelsets on the soft classification ensures segmentation of the image into prostate and the background regions.

### 2.4.12 Combinations of algorithms

It is well known that combining the results of multiple human observers often leads to a better segmentation than using the segmentation of only a single observer[134]. To investigate whether this is also true for segmentation algorithms, different types of combinations were tried. First, combining all the algorithm results using a majority voting approach was explored. The majority voting combination considered a voxel part of the prostate segmentation if the majority of the algorithms segmented the voxel as a prostate voxel. Second, only the top 5 (expert) algorithms were combined based on the overall algorithm score. A 'best combination' reference was also included by selecting the algorithm with the maximum score per case, for both the top 5 and all algorithms.

## 2.5 Results

### 2.5.1 Online challenge

The results of the online challenge are summarized in Tables 2.3, 2.6 and Figure 2.3. In Table 2.3 the average algorithm scores and standard deviations are presented, which are used to rank the algorithms. The ordering of the algorithms represents the ranking after both the online and live challenges. The online and live components were weighted equally to determine the final ranking. Metric values and scores for all algorithms on the online challenge data are presented in Table 2.6. In Figure 2.3 we provide the results per algorithm per case to give a more complete view of algorithm robustness and variability.

### 2.5.2 Live challenge

Tables 2.3, 2.7 and Figure 2.4 show the results of the live challenge at the MICCAI2012 workshop. In Table 2.3 (column 2) the average scores for each algorithm are presented including standard deviations. Metric values and scores for all algorithms on the live challenge data are presented in 2.7. Figure 2.4 shows the scores per case per algorithm for the cases processed at the live challenge. Algorithms that were unable to segment all cases during the period of the challenge (4 hours), or produced segmentations that were considered to be a failure according to algorithm-specific checking criteria or the group, are indicated with an asterisk in Table 2. Unsegmented or failed cases were given a score of 0.

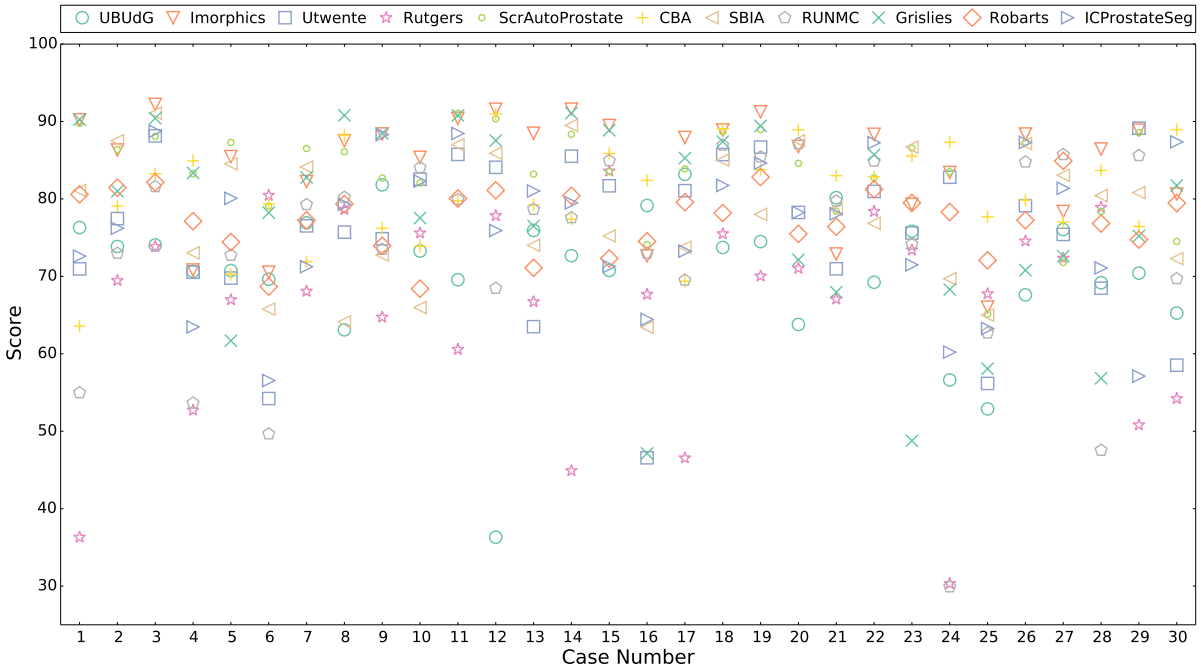



Figure 2.3: Results of the online challenge. The overall score is on the vertical axis and the case number on the horizontal axis. Teams are given a different symbol and color. Case distributions per center were: 1:7 RUNMC, 8:14 BIDMC, 15:22 UCL, 23:30 HK.

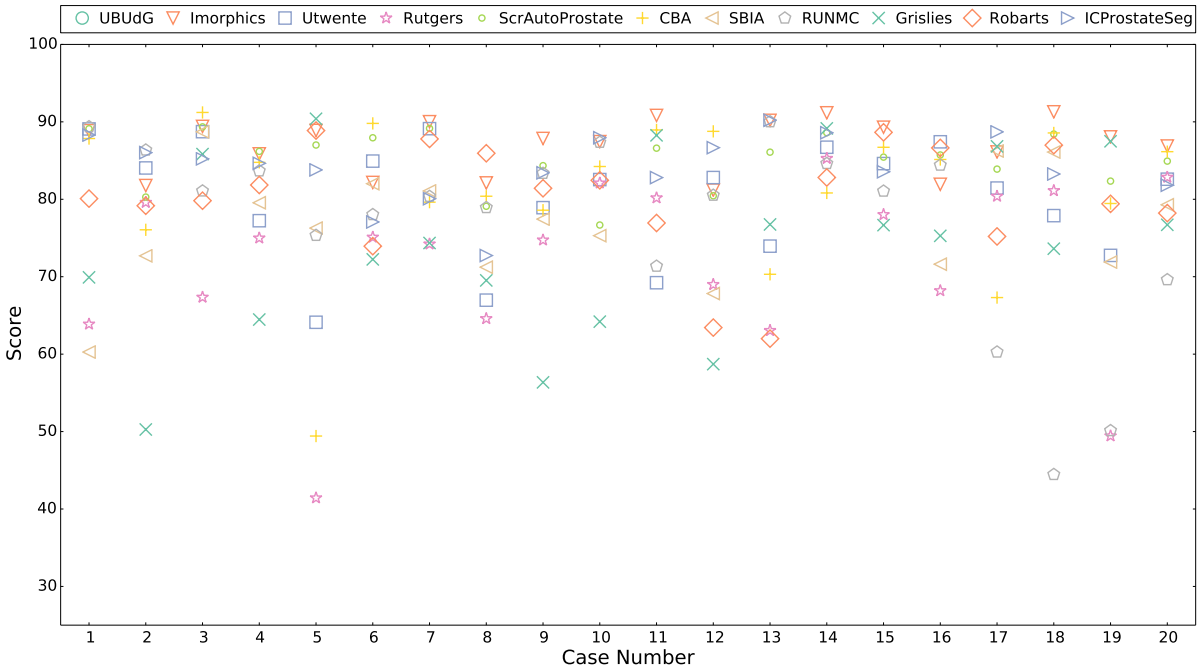



Figure 2.4: Results of the live challenge. The overall score is on the vertical axis and the case number on the horizontal axis. Teams are given a different symbol and color. Case distributions per center were: 1:5 UCL, 6:10 HK, 11:15 BIDMC, 16:20 RUNMC.

### 2.5.3 Overall

The overall ranking of the algorithms is presented in Table 2.3. Additionally, the results of the algorithm combinations are shown in Table 2.2. Furthermore, statistical analysis on the complete set of case scores was also performed to determine which algorithms are significantly better than other algorithms. As a test repeated measures ANOVA was used in combination with Bonferroni correction at a significance level of 0.05. The results indicated that the top 2 algorithms by Imorphics and ScrAutoProstate are significantly better then every algorithm outside of the top 3. This also holds for both combination strategies. However, none of the algorithms or combinations strategies performed significantly better than the second observer. Finally, the robustness of the algorithms against multi-center data was also tested using ANOVA, but the center did not have a significant impact on the overall algorithm score (p=0.118). The average scores and standard deviations for the algorithms on a per-center basis are presented in Tables 2.4 and 2.5

| Name | Online | Live | Average |
|---|---|---|---|
| Imorphics | $84.36 \pm 7.11$ | $87.07 \pm 3.36$ | $85.72 \pm 5.90$ |
| All combined | $82.96 \pm 8.25$ | $87.70 \pm 3.11$ | $85.33 \pm 6.68$ |
| Top 5 combined | $85.38 \pm 6.13$ | $87.09 \pm 3.22$ | $86.24 \pm 5.16$ |
| Maximum | $87.57 \pm 3.37$ | $88.88 \pm 1.73$ | $88.23 \pm 2.83$ |

Table 2.2: Results for the single best algorithm and combinations of algorithms, average over all cases including standard deviation.

## 2.6 Discussion

### 2.6.1 Challenge setup and participation

The images used in the challenge are a good representation of what would be encountered in a clinical setting, with large differences in acquisition protocol, prostate appearance and size. Additionally, the images originated from different centers and scanner manufacturers. The training and test sets were also large enough to draw statistical conclusions on algorithm performance.

The reference standard was constructed by 3 different observers, who each segmented a part of the data. These segmentations were subsequently inspected by the experienced observer for correctness and consistency. Obtaining additional observers for each case would be preferable, however recruiting multiple observers to spend time contouring 100 prostate MR cases is extremely challenging.

| Rank | Team Name | Type | Online | Live | Average |
|---|---|---|---|---|---|
| 1 | Imorphics | Automatic | 84.36 ± 7.11 | 87.07 ± 3.36 | 85.72 ± 5.90 |
| 2 | ScrAutoProstate | Automatic | 83.49 ± 5.92 | 85.08 ± 3.54 | 84.29 ± 5.10 |
| 3 | CBA | Interactive | 80.66 ± 6.46 | 81.21 ± 9.60 | 80.94 ± 7.86 |
| 4 | Robarts | Semi-automatic | 77.32 ± 4.04 | 80.08 ± 7.18 | 78.70 ± 5.51 |
| 5 | Utwente | Semi-automatic | 75.23 ± 10.53 | 80.26 ± 7.30 | 77.75 ± 9.37 |
| 6 | Grislies | Automatic | 77.56 ± 12.60 | 74.35 ± 11.28 | 75.96 ± 12.08 |
| 7 | ICProstateSeg | Automatic | 76.06 ± 9.40 | 75.74 ± 8.81* (84.16 ± 4.43) | 75.90 ± 9.17 |
| 8 | DIAG | Automatic | 73.30 ± 13.69 | 77.01 ± 12.09 | 75.16 ± 13.07 |
| 9 | SBIA | Automatic | 78.34 ± 8.22 | 61.38 ± 28.22* (76.72 ± 7.44) | 69.86 ± 18.95 |
| 10 | Rutgers | Automatic | 65.97 ± 13.13 | 71.77 ± 11.02 | 68.87 ± 12.32 |
| 11 | UBUdG | Semi-automatic | 70.44 ± 9.12 | 00.00 ± 0.0* | 35.22 ± 9.12 |
| - | SecondObserver | - | 85.00 ± 4.50 | 85.00 ± 4.91 | 85.00 ± 4.67 |

Table 2.3: Overall challenge results. The last three columns contain the scores including standard deviations. These scores are an average of all individual metric scores over all cases, as explained in section 2.3.2. For the live challenge scores with an asterisk, teams had either missing or incomplete segmentations for some cases. Incomplete or failed cases were assigned a score of 0. The scores of these groups over all completed cases is shown in brackets. The UBUdG team did not participate in the live challenge and as such received a zero score.

| | RUNMC | BIDMC | UCL | HK |
|---|---|---|---|---|
| Imorphics | 82.55 ± 8.72 | 89.05 ± 2.29 | 84.78 ± 7.52 | 81.44 ± 7.41 |
| ScrAutoProstate | 85.76 ± 3.56 | 86.26 ± 3.73 | 83.12 ± 4.95 | 79.47 ± 8.46 |
| CBA | 76.05 ± 7.71 | 80.82 ± 6.37 | 83.16 ± 6.17 | 82.06 ± 4.94 |
| Robarts | 77.38 ± 4.73 | 76.34 ± 5.13 | 77.57 ± 3.55 | 77.88 ± 3.77 |
| Utwente | 72.52 ± 10.27 | 78.85 ± 8.11 | 76.50 ± 13.02 | 73.16 ± 11.46 |
| Grislies | 81.10 ± 9.69 | 86.10 ± 6.35 | 77.99 ± 14.82 | 66.54 ± 10.99 |
| ICProstateSeg | 72.70 ± 10.58 | 82.12 ± 4.71 | 77.37 ± 7.49 | 72.40 ± 11.92 |
| DIAG | 66.60 ± 13.25 | 77.48 ± 5.09 | 81.45 ± 6.76 | 67.51 ± 20.15 |
| SBIA | 81.02 ± 8.77 | 77.04 ± 10.41 | 77.31 ± 7.32 | 78.15 ± 8.19 |
| Rutgers | 63.98 ± 14.82 | 67.00 ± 11.99 | 69.98 ± 11.02 | 62.79 ± 16.46 |
| UBUdG | 73.17 ± 2.88 | 67.52 ± 14.90 | 74.31 ± 6.39 | 66.73 ± 8.33 |
| Average | 75.69 ± 8.63 | 78.96 ± 7.19 | 78.50 ± 8.09 | 73.47 ± 10.19 |

Table 2.4: Average scores and standard deviations per team over the different centers for the online challenge.

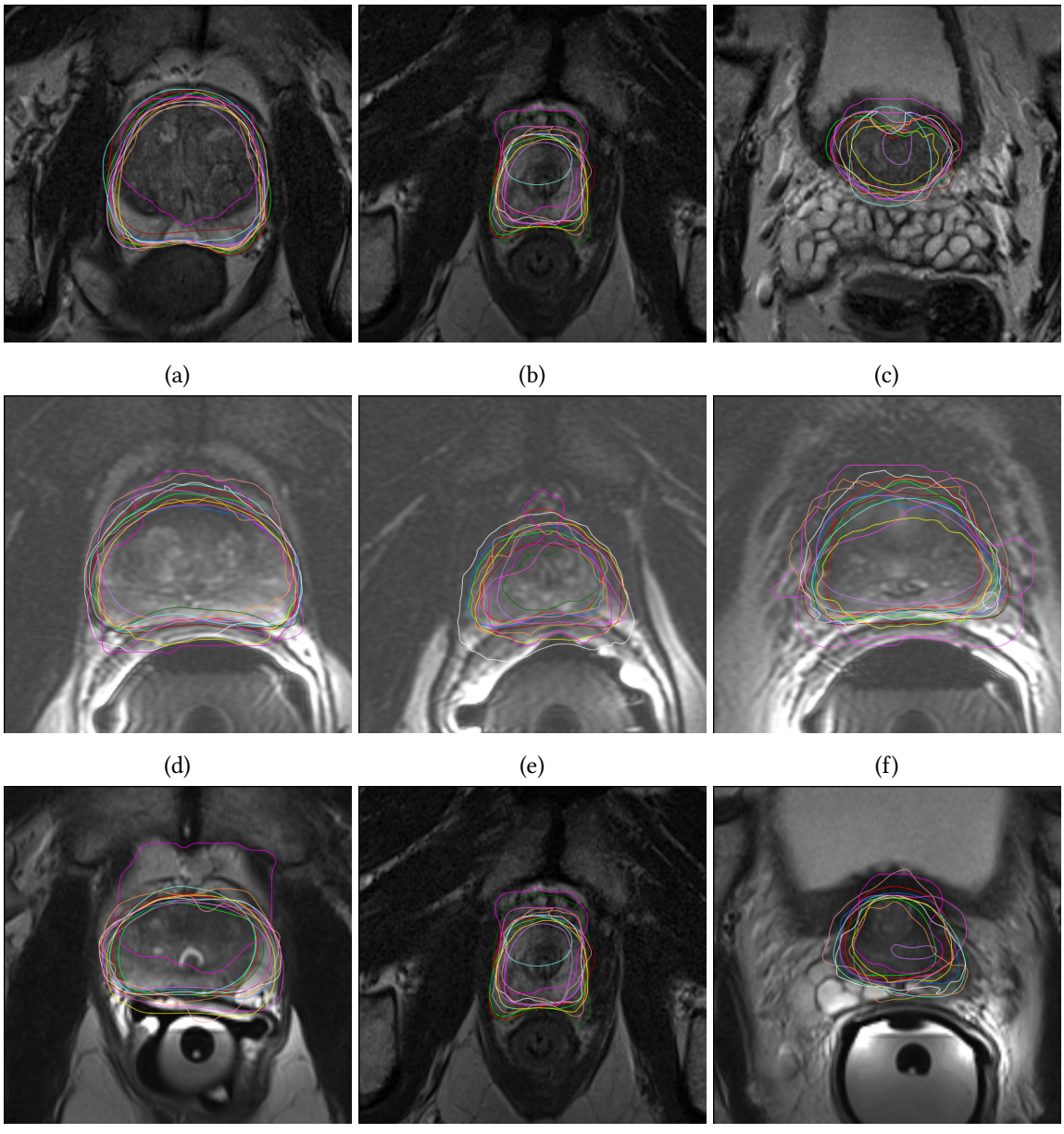


Figure 2.5: Qualitative segmentation results of case 3 (a, b, c), case 10 (d, e, f) and case 25 (g, h, i) at the center (a, d, g), apex (b, e, h) and base (c, f, i) of the prostate. Case 3 had the best, case 10 reasonable and case 25 the worst algorithm scores on average. The different colors indicate the results for the different teams.

| | RUNMC | BIDMC | UCL | HK |
|---|---|---|---|---|
| Imorphics | 86.86 ± 3.39 | 88.54 ± 4.17 | 86.96 ± 3.22 | 85.92 ± 3.59 |
| ScrAutoProstate | 85.06 ± 2.26 | 85.44 ± 3.03 | 86.39 ± 3.67 | 83.44 ± 5.44 |
| CBA | 81.32 ± 8.52 | 83.11 ± 7.88 | 77.86 ± 16.87 | 82.53 ± 4.59 |
| Robarts | 81.29 ± 5.27 | 74.77 ± 11.76 | 81.96 ± 3.99 | 82.31 ± 5.34 |
| Utwente | 80.42 ± 5.48 | 79.46 ± 7.51 | 80.64 ± 10.40 | 80.50 ± 8.42 |
| Grislies | 79.98 ± 6.64 | 77.91 ± 12.30 | 72.18 ± 16.30 | 67.33 ± 7.22 |
| ICProstateSeg | 82.75 ± 4.67 | 86.36 ± 3.18 | 85.60 ± 1.71 | 80.25 ± 5.83 |
| RUNMC | 61.77 ± 15.90 | 81.51 ± 6.82 | 83.16 ± 5.35 | 81.61 ± 3.80 |
| SBIA | 79.03 ± 7.21 | 13.57 ± 30.33 | 75.50 ± 10.38 | 77.41 ± 4.39 |
| Rutgers | 72.39 ± 14.09 | 75.10 ± 8.95 | 65.45 ± 14.78 | 74.14 ± 6.24 |
| Average | 79.09 ± 7.34 | 74.58 ± 9.59 | 79.57 ± 8.67 | 79.54 ± 5.49 |

Table 2.5: Average scores and standard deviations per team over the different centers for the live challenge. Note that team UBUdG did not participate in the live challenge and as such is not included here.

The metrics that were used result in a good separation between algorithms and the conversion into per case scores keeps these differences intact. Other metrics were also considered, for example the Jaccard index, sensitivity/specificity and regular Hausdorff distance. Jaccard index is a volume-based metric with similar characteristics as the Dice coefficient, however, in prostate segmentation literature, the Dice coefficient is more often used. To allow better comparison to existing and future literature we chose the Dice coefficient. Sensitivity and specificity are generally not useful in prostate segmentation because specificity will not be very discriminative: the prostate is always a relative small part of the total image volume. Finally, the modified 95% Hausdorff distance was used because the regular Hausdorff distance can be harsh and sensitive to noise: a single pixel can determine overall image segmentation outcome.

One issue with basing case scores on observer reference standards is that very high scores end up in the realm of inter-observer variability. A score higher than 85 is probably still indicative of improved performance, as the second observer segmentations are less accurate than the reference standard, but it is difficult to say whether a score of e.g. 94 is indeed better or just different and equally accurate than a score of 92. However, in general, the algorithms in this challenge do not obtain these scores on average, so this is not an issue. Visual inspection of the segmentation results also confirms this, the largest segmentation errors made by the algorithms would not be made by an experienced observer.

An alternative scoring approach that is not sensitive to inter-observer variability is to rank algorithms based on their average rank for each of the sub-scores over all algorithms (e.g. if an algorithm has the highest average Dice of all algorithms, it will have rank 1 for Dice. If the

| **Team Name** | **Average Boundary Distance** | | | | | |
|---|---|---|---|---|---|---|
| | Overall | Base | Apex | Score (Overall) | Score (Base) | Score (Apex) |
| Imorphics | 2.10 ± 0.68 | 2.18 ± 1.14 | 1.96 ± 0.80 | 82.66 ± 5.60 | 85.20 ± 7.75 | 88.44 ± 4.71 |
| ScrAutoProstate | 2.13 ± 0.48 | 2.23 ± 0.70 | 2.18 ± 0.68 | 82.42 ± 3.93 | 84.87 ± 4.73 | 87.17 ± 3.98 |
| CBA | 2.33 ± 0.59 | 2.60 ± 1.47 | 2.44 ± 0.81 | 80.77 ± 4.88 | 82.31 ± 9.96 | 85.62 ± 4.75 |
| Robarts | 2.65 ± 0.37 | 2.92 ± 0.88 | 3.49 ± 0.95 | 78.09 ± 3.06 | 80.14 ± 5.97 | 79.45 ± 5.58 |
| Utwente | 3.03 ± 1.06 | 3.45 ± 1.96 | 2.68 ± 0.98 | 74.96 ± 8.73 | 76.54 ± 13.34 | 84.20 ± 5.79 |
| Grislies | 2.96 ± 1.55 | 3.19 ± 2.00 | 2.46 ± 1.26 | 75.55 ± 12.80 | 78.35 ± 13.59 | 85.50 ± 7.42 |
| ICProstateSeg | 2.86 ± 0.82 | 3.18 ± 1.32 | 2.89 ± 1.05 | 76.34 ± 6.78 | 78.38 ± 9.00 | 82.99 ± 6.21 |
| DIAG | 3.40 ± 1.72 | 4.23 ± 3.06 | 2.72 ± 1.75 | 71.90 ± 14.18 | 71.29 ± 20.81 | 84.01 ± 10.33 |
| SBIA | 2.85 ± 0.72 | 2.82 ± 1.02 | 2.13 ± 0.80 | 76.47 ± 5.94 | 80.86 ± 6.93 | 87.44 ± 4.74 |
| Rutgers | 4.06 ± 1.80 | 4.82 ± 2.64* | 3.71 ± 1.26* | 66.47 ± 14.87 | 63.06 ± 23.71 | 74.68 ± 16.56 |
| UBUdG | 4.26 ± 1.58 | 4.21 ± 1.42 | 4.53 ± 1.71 | 64.84 ± 13.09 | 71.40 ± 9.63 | 73.33 ± 10.08 |
| All combined | 2.06 ± 0.78 | 2.60 ± 1.53 | 2.04 ± 0.81 | 82.96 ± 6.46 | 82.30 ± 10.36 | 87.98 ± 4.76 |
| Top 5 combined | 1.94 ± 0.48 | 2.10 ± 0.82 | 1.77 ± 0.62 | 84.00 ± 3.95 | 85.70 ± 5.56 | 89.57 ± 3.63 |
| Maximum | 1.78 ± 0.35 | 1.82 ± 0.52 | 1.58 ± 0.35 | 85.28 ± 2.91 | 87.66 ± 3.51 | 90.70 ± 2.06 |
| SecondObserver | 1.82 ± 0.36 | 2.21 ± 0.80 | 2.55 ± 1.08 | 85.00 ± 2.93 | 85.00 ± 5.42 | 85.00 ± 6.34 |
| **Team Name** | **95% Hausdorff Distance** | | | | | |
| | Overall | Base | Apex | Score (Overall) | Score (Base) | Score (Apex) |
| Imorphics | 5.94 ± 2.14 | 5.45 ± 2.58 | 4.73 ± 1.68 | 84.20 ± 5.70 | 86.98 ± 6.15 | 88.84 ± 3.97 |
| ScrAutoProstate | 5.58 ± 1.49 | 5.60 ± 2.35 | 4.93 ± 1.38 | 85.15 ± 3.98 | 86.63 ± 5.62 | 88.37 ± 3.25 |
| CBA | 6.57 ± 2.11 | 6.64 ± 4.07 | 5.75 ± 1.91 | 82.50 ± 5.61 | 84.15 ± 9.73 | 86.43 ± 4.52 |
| Robarts | 6.48 ± 1.56 | 6.83 ± 2.26 | 7.36 ± 2.11 | 82.76 ± 4.15 | 83.70 ± 5.39 | 82.62 ± 4.98 |
| Utwente | 7.32 ± 2.44 | 7.69 ± 3.75 | 5.89 ± 1.93 | 80.52 ± 6.48 | 81.64 ± 8.94 | 86.11 ± 4.57 |
| Grislies | 7.90 ± 3.83 | 7.61 ± 4.11 | 5.82 ± 2.82 | 78.97 ± 10.19 | 81.85 ± 9.81 | 86.26 ± 6.65 |
| ICProstateSeg | 7.20 ± 1.96 | 7.27 ± 2.92 | 6.51 ± 2.31 | 80.84 ± 5.21 | 82.64 ± 6.97 | 84.62 ± 5.46 |
| DIAG | 8.59 ± 4.00 | 9.00 ± 4.62 | 5.91 ± 3.68 | 77.15 ± 10.66 | 78.52 ± 11.04 | 86.05 ± 8.69 |
| SBIA | 7.73 ± 2.68 | 6.99 ± 2.25 | 4.60 ± 1.31 | 79.43 ± 7.14 | 83.32 ± 5.37 | 89.14 ± 3.10 |
| Rutgers | 9.25 ± 3.76 | 9.88 ± 4.04* | 7.58 ± 2.35* | 75.37 ± 10.00 | 71.18 ± 21.41 | 78.82 ± 16.23 |
| Rutgers | 9.25 ± 3.76 | 9.88 ± 4.04* | 7.58 ± 2.35* | 75.37 ± 10.00 | 71.18 ± 21.41 | 78.82 ± 16.23 |
| UBUdG | 9.17 ± 3.48 | 9.06 ± 2.71 | 9.54 ± 3.52 | 75.59 ± 9.27 | 78.38 ± 6.46 | 77.48 ± 8.30 |
| All combined | 5.43 ± 2.18 | 6.00 ± 3.06 | 4.97 ± 1.94 | 85.55 ± 5.81 | 85.67 ± 7.30 | 88.26 ± 4.57 |
| Top 5 combined | 5.30 ± 1.60 | 5.37 ± 2.38 | 4.22 ± 1.25 | 85.91 ± 4.26 | 87.19 ± 5.67 | 90.04 ± 2.94 |
| Maximum | 4.63 ± 1.06 | 4.32 ± 1.28 | 3.67 ± 0.70 | 87.67 ± 2.81 | 89.68 ± 3.05 | 91.34 ± 1.64 |
| SecondObserver | 5.64 ± 1.73 | 6.28 ± 2.95 | 6.36 ± 2.40 | 85.00 ± 4.61 | 85.00 ± 7.04 | 85.00 ± 5.66 |
| **Team Name** | **Dice Coefficient** | | | | | |
| | Overall | Base | Apex | Score (Overall) | Score (Base) | Score (Apex) |
| Imorphics | 0.88 ± 0.04 | 0.86 ± 0.08 | 0.85 ± 0.08 | 81.96 ± 6.62 | 84.76 ± 8.93 | 88.57 ± 6.13 |
| ScrAutoProstate | 0.87 ± 0.04 | 0.86 ± 0.04 | 0.83 ± 0.07 | 81.14 ± 5.39 | 85.02 ± 4.58 | 87.79 ± 5.23 |
| CBA | 0.87 ± 0.04 | 0.84 ± 0.07 | 0.80 ± 0.11 | 79.80 ± 5.36 | 82.87 ± 8.07 | 85.46 ± 7.98 |
| Robarts | 0.84 ± 0.03 | 0.81 ± 0.05 | 0.71 ± 0.12 | 75.32 ± 4.25 | 79.77 ± 5.82 | 78.70 ± 8.84 |
| Utwente | 0.82 ± 0.07 | 0.78 ± 0.13 | 0.78 ± 0.09 | 72.97 ± 9.77 | 76.12 ± 13.85 | 84.10 ± 6.44 |
| Grislies | 0.83 ± 0.08 | 0.81 ± 0.11 | 0.82 ± 0.10 | 75.10 ± 12.38 | 79.17 ± 11.85 | 86.65 ± 7.09 |
| ICProstateSeg | 0.82 ± 0.06 | 0.76 ± 0.13 | 0.74 ± 0.13 | 72.68 ± 9.40 | 74.12 ± 14.15 | 80.47 ± 9.41 |
| DIAG | 0.80 ± 0.09 | 0.71 ± 0.22 | 0.79 ± 0.12 | 69.62 ± 14.20 | 68.38 ± 23.42 | 84.82 ± 8.77 |
| SBIA | 0.84 ± 0.06 | 0.81 ± 0.08 | 0.84 ± 0.07 | 75.29 ± 8.27 | 79.29 ± 9.07 | 88.11 ± 5.31 |
| Rutgers | 0.74 ± 0.10 | 0.61 ± 0.25 | 0.66 ± 0.17 | 61.05 ± 15.36 | 57.75 ± 25.70 | 74.93 ± 12.60 |
| UBUdG | 0.71 ± 0.11 | 0.71 ± 0.12 | 0.63 ± 0.14 | 56.73 ± 16.09 | 68.17 ± 12.80 | 72.53 ± 10.20 |
| All combined | 0.88 ± 0.05 | 0.81 ± 0.13 | 0.81 ± 0.11 | 81.29 ± 7.55 | 78.90 ± 14.20 | 86.31 ± 8.39 |
| Top 5 combined | 0.89 ± 0.03 | 0.87 ± 0.05 | 0.87 ± 0.06 | 83.65 ± 4.82 | 85.79 ± 5.96 | 90.32 ± 4.63 |
| Maximum | 0.90 ± 0.02 | 0.89 ± 0.03 | 0.88 ± 0.03 | 85.08 ± 3.55 | 88.20 ± 3.80 | 91.46 ± 2.50 |
| SecondObserver | 0.90 ± 0.03 | 0.86 ± 0.06 | 0.80 ± 0.11 | 85.00 ± 3.82 | 85.00 ± 6.14 | 85.00 ± 8.39 |
| **Team Name** | **Relative Volume Difference** | | | | | |
| | Overall | Base | Apex | Score (Overall) | Score (Base) | Score (Apex) |
| Imorphics | 2.92 ± 15.71 | 1.01 ± 19.56 | 0.65 ± 30.68 | 72.53 ± 25.31 | 84.03 ± 16.94 | 84.20 ± 16.97 |
| ScrAutoProstate | 11.53 ± 14.05 | 9.65 ± 16.52 | 14.08 ± 34.25 | 68.18 ± 27.94 | 82.67 ± 14.82 | 82.52 ± 18.44 |
| CBA | 12.75 ± 13.99 | 18.85 ± 24.88 | 0.41 ± 28.63 | 63.48 ± 25.38 | 72.51 ± 24.00 | 82.04 ± 11.91 |
| Robarts | 10.31 ± 17.92 | 12.69 ± 26.26 | -3.27 ± 39.09 | 61.70 ± 28.63 | 70.65 ± 18.41 | 74.96 ± 15.61 |
| Utwente | 22.30 ± 27.88 | 27.52 ± 41.86 | 15.10 ± 41.30 | 50.19 ± 32.42 | 57.94 ± 31.74 | 77.45 ± 23.46 |
| Grislies | 19.81 ± 31.93 | 23.12 ± 44.71 | 15.46 ± 43.71 | 59.25 ± 38.47 | 64.73 ± 31.20 | 79.31 ± 23.00 |
| ICProstateSeg | -2.61 ± 24.86 | -4.47 ± 35.14 | -13.31 ± 43.42 | 57.96 ± 34.16 | 66.62 ± 25.50 | 75.09 ± 20.77 |
| DIAG | 4.66 ± 28.30 | -9.34 ± 43.13 | 11.66 ± 54.14 | 51.04 ± 31.02 | 60.62 ± 31.86 | 76.15 ± 24.37 |
| SBIA | 16.19 ± 25.35 | 13.47 ± 30.78 | 11.26 ± 35.57 | 51.63 ± 35.95 | 67.71 ± 23.49 | 81.33 ± 21.19 |
| Rutgers | -5.83 ± 30.81 | -22.11 ± 57.39 | -16.68 ± 46.37 | 52.18 ± 30.04 | 44.52 ± 31.99 | 71.58 ± 24.00 |
| UBUdG | -5.16 ± 21.40 | -7.33 ± 28.05 | -14.55 ± 33.25 | 59.02 ± 24.71 | 69.96 ± 16.63 | 77.87 ± 16.16 |
| All combined | -10.02 ± 14.62 | -15.45 ± 25.94 | -19.44 ± 22.45 | 67.17 ± 25.33 | 73.19 ± 23.89 | 81.67 ± 13.00 |
| Top 5 combined | 7.63 ± 13.45 | 7.32 ± 18.53 | 6.37 ± 27.31 | 73.70 ± 25.02 | 82.15 ± 15.60 | 86.50 ± 16.37 |
| Maximum | 2.76 ± 3.05 | 4.50 ± 4.80 | 4.23 ± 4.21 | 93.48 ± 7.19 | 94.61 ± 5.76 | 96.78 ± 3.21 |
| SecondObserver | -1.87 ± 7.32 | -6.17 ± 13.49 | -16.24 ± 21.13 | 85.00 ± 9.23 | 85.00 ± 9.23 | 85.00 ± 13.57 |

Table 2.6: Averages and standard deviations for all metrics for all teams in the online challenge. Entries indicated with an asterisk had cases with infinite boundary distance measures removed from the average, which could occur due to empty base or apex segmentation results.

| Team Name | Average Boundary Distance | | | | | |
|---|---|---|---|---|---|---|
| | Overall | Base | Apex | Score (Overall) | Score (Base) | Score (Apex) |
| Imorphics | 1.95 ± 0.36 | 2.45 ± 0.65 | 1.83 ± 0.53 | 85.53 ± 2.70 | 87.12 ± 3.41 | 88.21 ± 3.39 |
| ScrAutoProstate | 2.18 ± 0.36 | 2.34 ± 0.78 | 2.16 ± 0.70 | 83.86 ± 2.65 | 87.73 ± 4.12 | 86.05 ± 4.50 |
| CBA | 2.56 ± 0.96 | 2.48 ± 1.55 | 119.28 ± 522.54 | 81.03 ± 7.10 | 86.95 ± 8.13 | 80.06 ± 21.12 |
| Robarts | 2.67 ± 0.62 | 2.66 ± 0.90 | 3.93 ± 2.42 | 80.23 ± 4.56 | 86.01 ± 4.74 | 74.64 ± 15.57 |
| Utwente | 2.87 ± 0.79 | 3.47 ± 1.33 | 2.43 ± 0.72 | 78.79 ± 5.87 | 81.76 ± 6.96 | 84.32 ± 4.62 |
| Grislies | 4.17 ± 2.35 | 3.75 ± 2.25 | 2.82 ± 1.06 | 69.14 ± 17.43 | 80.31 ± 11.80 | 81.81 ± 6.84 |
| ICProstateSeg | 2.35 ± 0.99 | 2.62 ± 1.37 | 1.95 ± 0.96 | 82.63 ± 7.35 | 86.23 ± 7.21 | 87.46 ± 6.16 |
| DIAG | 3.21 ± 1.39 | 237.53 ± 718.80 | 2.31 ± 0.71 | 76.26 ± 10.29 | 71.09 ± 27.15 | 85.11 ± 4.57 |
| SBIA | 3.13 ± 0.74* | 3.13 ± 0.64* | 2.89 ± 1.03* | 61.49 ± 31.92 | 66.83 ± 34.41 | 65.10 ± 33.91 |
| Rutgers | 3.84 ± 1.37 | 3.70 ± 1.12* | 4.21 ± 1.83 | 71.54 ± 10.18 | 72.52 ± 25.41 | 72.87 ± 11.80 |
| All combined | 1.97 ± 0.34 | 2.18 ± 0.64 | 1.82 ± 0.53 | 85.43 ± 2.51 | 88.55 ± 3.35 | 88.28 ± 3.41 |
| Top 5 combined | 1.90 ± 0.32 | 2.15 ± 0.80 | 1.92 ± 0.64 | 85.93 ± 2.37 | 88.70 ± 4.18 | 87.61 ± 4.14 |
| Maximum | 1.87 ± 0.30 | 1.82 ± 0.45 | 1.53 ± 0.30 | 86.17 ± 2.20 | 90.44 ± 2.36 | 90.17 ± 1.88 |
| SecondObserver | 2.03 ± 0.50 | 2.86 ± 1.26 | 2.33 ± 1.35 | 85.00 ± 3.73 | 85.00 ± 6.63 | 85.00 ± 8.69 |
| **Team Name** | **95% Hausdorff Distance** | | | | | |
| | Overall | Base | Apex | Score (Overall) | Score (Base) | Score (Apex) |
| Imorphics | 5.54 ± 1.74 | 6.09 ± 1.61 | 4.58 ± 1.36 | 86.35 ± 4.28 | 87.96 ± 3.19 | 87.03 ± 3.86 |
| ScrAutoProstate | 6.04 ± 1.67 | 5.64 ± 2.17 | 4.60 ± 1.39 | 85.11 ± 4.12 | 88.84 ± 4.29 | 86.96 ± 3.94 |
| CBA | 7.34 ± 3.08 | 6.29 ± 3.03 | 122.28 ± 523.16 | 81.90 ± 7.59 | 87.55 ± 6.00 | 80.72 ± 20.05 |
| Robarts | 7.15 ± 2.08 | 6.12 ± 2.14 | 7.76 ± 3.20 | 82.38 ± 5.12 | 87.89 ± 4.22 | 78.01 ± 9.06 |
| Utwente | 6.72 ± 1.42 | 7.42 ± 2.38 | 5.68 ± 1.66 | 83.43 ± 3.51 | 85.33 ± 4.71 | 83.91 ± 4.70 |
| Grislies | 11.08 ± 5.85 | 8.68 ± 4.61 | 6.88 ± 2.21 | 72.68 ± 14.42 | 82.83 ± 9.11 | 80.49 ± 6.27 |
| ICProstateSeg | 5.89 ± 2.59 | 5.64 ± 2.73 | 4.58 ± 2.35 | 85.48 ± 6.38 | 88.83 ± 5.41 | 87.00 ± 6.67 |
| DIAG | 7.95 ± 3.21 | 242.13 ± 719.85 | 4.74 ± 1.34 | 80.40 ± 7.91 | 75.30 ± 26.69 | 86.56 ± 3.79 |
| SBIA | 7.07 ± 1.64* | 7.21 ± 1.96* | 5.93 ± 1.69* | 66.05 ± 34.07 | 68.59 ± 35.35 | 66.54 ± 34.40 |
| Rutgers | 8.48 ± 2.53 | 242.00 ± 719.42 | 7.82 ± 2.42 | 79.09 ± 6.23 | 75.29 ± 26.10 | 77.82 ± 6.86 |
| All combined | 5.67 ± 1.82 | 5.14 ± 1.40 | 4.46 ± 1.46 | 86.01 ± 4.49 | 89.84 ± 2.78 | 87.35 ± 4.13 |
| Top 5 combined | 5.49 ± 1.54 | 5.48 ± 2.24 | 4.56 ± 1.51 | 86.45 ± 3.80 | 89.16 ± 4.43 | 87.07 ± 4.27 |
| Maximum | 4.80 ± 1.02 | 4.20 ± 0.94 | 3.53 ± 0.76 | 88.17 ± 2.52 | 91.69 ± 1.86 | 90.13 ± 2.10 |
| SecondObserver | 6.08 ± 2.23 | 7.58 ± 3.90 | 5.29 ± 2.53 | 85.00 ± 5.50 | 85.00 ± 7.71 | 85.00 ± 7.17 |
| **Team Name** | **Dice Coefficient** | | | | | |
| | Overall | Base | Apex | Score (Overall) | Score (Base) | Score (Apex) |
| Imorphics | 0.89 ± 0.03 | 0.84 ± 0.06 | 0.86 ± 0.07 | 85.51 ± 3.92 | 86.98 ± 5.21 | 89.15 ± 5.66 |
| ScrAutoProstate | 0.87 ± 0.03 | 0.85 ± 0.06 | 0.83 ± 0.10 | 83.17 ± 3.53 | 87.35 ± 5.20 | 86.81 ± 7.47 |
| CBA | 0.85 ± 0.08 | 0.85 ± 0.10 | 0.77 ± 0.23 | 79.69 ± 10.77 | 87.82 ± 8.16 | 82.13 ± 17.39 |
| Robarts | 0.84 ± 0.04 | 0.84 ± 0.06 | 0.67 ± 0.22 | 78.82 ± 5.40 | 86.62 ± 4.90 | 74.31 ± 17.27 |
| Utwente | 0.83 ± 0.06 | 0.77 ± 0.10 | 0.79 ± 0.10 | 77.46 ± 7.61 | 81.40 ± 7.81 | 84.12 ± 7.47 |
| Grislies | 0.77 ± 0.12 | 0.78 ± 0.12 | 0.79 ± 0.09 | 70.04 ± 16.09 | 81.93 ± 9.79 | 83.82 ± 7.30 |
| ICProstateSeg | 0.76 ± 0.26 | 0.72 ± 0.26 | 0.74 ± 0.26 | 71.70 ± 25.03 | 77.24 ± 21.30 | 80.26 ± 20.36 |
| DIAG | 0.80 ± 0.07 | 0.63 ± 0.30 | 0.82 ± 0.07 | 73.81 ± 9.43 | 69.73 ± 24.09 | 86.18 ± 5.71 |
| SBIA | 0.65 ± 0.34 | 0.64 ± 0.34 | 0.63 ± 0.33 | 60.41 ± 31.93 | 70.99 ± 27.41 | 71.78 ± 25.83 |
| Rutgers | 0.75 ± 0.10 | 0.68 ± 0.25 | 0.62 ± 0.22 | 67.41 ± 13.75 | 73.93 ± 20.13 | 70.85 ± 17.08 |
| All combined | 0.89 ± 0.03 | 0.87 ± 0.05 | 0.86 ± 0.08 | 86.10 ± 3.30 | 89.01 ± 4.10 | 88.93 ± 5.88 |
| Top 5 combined | 0.89 ± 0.02 | 0.87 ± 0.06 | 0.85 ± 0.09 | 86.12 ± 2.90 | 89.03 ± 4.94 | 88.21 ± 6.58 |
| Maximum | 0.90 ± 0.02 | 0.89 ± 0.03 | 0.89 ± 0.03 | 86.51 ± 2.47 | 90.97 ± 2.82 | 91.90 ± 1.97 |
| SecondObserver | 0.89 ± 0.03 | 0.82 ± 0.10 | 0.81 ± 0.15 | 85.00 ± 4.18 | 85.00 ± 8.32 | 85.00 ± 11.56 |
| **Team Name** | **Relative Volume Difference** | | | | | |
| | Overall | Base | Apex | Score (Overall) | Score (Base) | Score (Apex) |
| Imorphics | -1.50 ± 9.15 | -8.31 ± 18.08 | -1.03 ± 23.97 | 86.31 ± 13.01 | 87.15 ± 7.70 | 87.55 ± 10.37 |
| ScrAutoProstate | 10.05 ± 11.56 | 7.77 ± 22.01 | 9.59 ± 30.51 | 73.96 ± 17.56 | 86.55 ± 11.38 | 84.60 ± 15.29 |
| CBA | 12.26 ± 17.73 | 24.75 ± 41.69 | -7.05 ± 39.63 | 63.49 ± 24.70 | 81.63 ± 24.91 | 81.50 ± 20.08 |
| Robarts | -1.72 ± 17.47 | 5.30 ± 25.52 | -29.19 ± 37.14 | 71.84 ± 21.87 | 86.46 ± 14.29 | 73.77 ± 18.61 |
| Utwente | 12.62 ± 22.25 | 20.75 ± 37.43 | 0.66 ± 28.70 | 62.15 ± 30.81 | 75.02 ± 20.62 | 85.40 ± 12.77 |
| Grislies | 43.13 ± 65.32 | 36.41 ± 58.73 | 7.23 ± 38.19 | 37.72 ± 40.30 | 72.42 ± 29.35 | 79.01 ± 15.76 |
| ICProstateSeg | -8.49 ± 34.17 | -14.15 ± 34.88 | -14.88 ± 36.55 | 69.10 ± 29.32 | 81.82 ± 22.02 | 80.77 ± 18.68 |
| DIAG | -12.34 ± 18.38 | -38.10 ± 32.87 | 1.61 ± 28.65 | 64.59 ± 25.81 | 70.54 ± 24.65 | 84.60 ± 11.74 |
| SBIA | 6.55 ± 59.45 | 2.66 ± 57.32 | 12.12 ± 68.31 | 30.65 ± 34.34 | 64.84 ± 24.89 | 62.50 ± 28.06 |
| Rutgers | -14.59 ± 26.52 | -24.79 ± 31.88 | -24.37 ± 47.01 | 50.76 ± 27.17 | 76.87 ± 20.18 | 72.31 ± 22.95 |
| All combined | 2.69 ± 9.75 | -0.16 ± 13.09 | -2.25 ± 24.49 | 83.77 ± 12.66 | 91.64 ± 5.14 | 87.48 ± 10.93 |
| Top 5 combined | 4.69 ± 9.95 | 6.89 ± 20.16 | -3.07 ± 26.74 | 82.19 ± 13.65 | 88.57 ± 11.35 | 86.10 ± 11.74 |
| Maximum | 1.80 ± 1.43 | 3.65 ± 3.24 | 3.58 ± 3.99 | 96.28 ± 2.94 | 97.21 ± 2.47 | 97.52 ± 2.74 |
| SecondObserver | -5.72 ± 7.44 | -17.49 ± 18.12 | -17.97 ± 22.90 | 85.00 ± 12.07 | 85.00 ± 11.93 | 85.00 ± 13.03 |

Table 2.7: Averages and standard deviations for all metrics for all teams in the live challenge. Entries indicated with an asterisk had cases with infinite boundary distance measures removed from the average, which could occur due to empty segmentation results.

same algorithm has rank 5 for average boundary distance over all algorithms, his average total rank would be 3). This approach has its own disadvantage, e.g. high ranks do not mean good segmentations and algorithm ranking is not only based on the performance of the algorithm itself, but also on the results of other algorithms, i.e. a new algorithm which does very poorly on all metrics except one might influence the ranking of all other algorithms by changing their average rank.

Participation in the initial phase of the challenge was similar to what we have seen in other segmentation challenges, for example[113] and the knee cartilage segmentation challenge (SKI10, `http://www.ski10.org`). The literature on prostate segmentation is well represented by the competing algorithms, which include active shape models, atlas-based methods, pattern recognition algorithms and variants.

We specifically chose to allow only single submissions per algorithm instead of allowing each group to submit results with different parameter settings, to make sure there would be 'no training on the test set'.

### 2.6.2 Challenge results

All algorithms submitted to the challenge produced reasonable to excellent results on average (online and live challenge combined scores ranging from 68.97 - 85.72). One point to note is that although some algorithms may have received the same average score, the variability can differ substantially, as shown in Tables 2.6, 2.7 and 2.3. For example, the algorithm presented by Robarts[140] scored 77.32 and 80.08 in the online and live challenge respectively, but has a very low variability: 5.51 score standard deviation overall. This is much lower than the algorithms that had similar scores, for example 7.86 for CBA[132] and 9.37 for Utwente[143]. Depending on the purpose for which an algorithm is used in the clinic, this can be a very important aspect. As such, it might be good to incorporate performance variability directly in algorithm ranking in future challenges.

It is worth noting that the top 2 algorithms by Imorphics[126] and ScrAutoProstate[129] were completely automatic and even outperformed the completely interactive method presented by CBA. Whereas the algorithm by Imorphics performed best overall, the algorithm by ScrAutoProstate should be noted for its exceptionally fast segmentation speed (2.3 seconds, Table 2.8), the fastest of all algorithms. Further details about interaction, implementation details and computation time can be found in Table 2.8. Algorithm computation times varied, with the active shape model based approaches often having computation times in the order of minutes, whereas the atlas based approaches required substantially more time or computing power (e.g. clusters, GPU). It is important to note that some algorithms were implemented in higher level programming languages like Matlab, whereas some where implemented in low-level languages like C++, computation time is thus not only dependent on algorithm efficiency but also

on the development platform.

Inspecting the illustrative results in Figure 2.5 one can see that algorithms can differ quite substantially per case. In this figure we present the best, worst and a reasonable case with respect to average algorithm performance. Case 25 was especially tricky as it had a large area of fat around the prostate, especially near the base which appears very similar to prostate peripheral zone. Most algorithms oversegmented the prostatic fat, and as the prostate was relatively small, this results in large volumetric errors. However, if one inspects case 25 carefully, it is possible to make the distinction between fat and prostate, especially if you go through the different slices. It is thus no surprise that the interactive segmentation technique of CBA performed the best. Further inspection of the results shows that in the cases with low average algorithm performance the interactive method is usually the best algorithm (e.g. (Figure 2.3): cases 4, 16 and 21 of the online challenge). This indicates that these cases cause problems for automated methods.

In this challenge we explicitly included segmentation results at the base and the apex of the prostate into the algorithm scoring because these areas are usually the most difficult to segment. This can also be observed in the results, especially Tables 2.6 and 2.7. Every algorithm performed worse on the apex and base if we look at the metric values (especially the Dice coefficient and the relative volume difference) themselves; however, as these areas are also the most difficult for the human observer, the scores for apex and base tend to be higher than the overall score. Interesting to note is that the top 2 algorithms outperform the second observer at almost every metric for both apex and base, whereas the overall score is lower than the second observer. For the live challenge the Imorphics algorithm even outperforms the second observer in the overall score. This indicates that for this part of the prostate automatic algorithms might improve over human observers.

Interestingly, similar to the SLIVER07-challenge, active shape based algorithms seemed to give the best results (places 1, 2, 4 and 5), although two of these systems are semi-automatic. Looking at the results in more detail, we can see that the atlas based systems comparatively have more trouble with cases which are not well represented by the training set, for example case 23, which has a prostate volume of 325 mL, while the average is around 50 mL.

One interactive method was included (team CBA) which on average scored 80.94, which is considerably lower than the second observer. This is mostly caused by over-segmentation at the base of the prostate, often the seminal vesicles were included in the prostate segmentation. Thus this algorithm is very dependent on the operator; in principle the algorithm should be able to get close to expert performance given an expert reader.

There were several semi-automatic algorithms (e.g. Robarts, UTwente and UBUdG) which needed manual interaction to initialize the algorithms. The interaction types and the influence this interaction has on segmentation accuracy will differ between the algorithms. Although none of the teams have explicitly tested the robustness to different initializations, some gen-

eral comments can be made. For the Robarts algorithm a number of points on the prostate boundary have to be set (8 to 10) to initialize a shape and the initial foreground and background distributions. As such, the algorithm is robust to misplacing single points. For the Utwente algorithm, the prostate center has to be indicated to initialize the active appearance and shape models. Big deviations in point selection can cause problems for active appearance and shape models, however in general they are pretty robust against small deviations[128]. For the UBUdG method, the user has to select the first and last slice of the prostate. As such, the algorithm will be unable to segment the prostate if it extends beyond those slices, which is an issue if users cannot correctly identify the start and end slice of the prostate.

Another aspect which plays a role in this challenge was the robustness of the algorithms to multi-center data. The image differences between the centers were actually quite large, especially between the endorectal coil and non-endorectal coil cases, as can be seen in Figure 2.1. Differences include coil artifacts near the peripheral zone, coil profiles, image intensities, slice thickness and resolution. However, if we look at for example Tables 2.4, 2.5 and Figure 2.3, it can be seen that all submitted algorithms are at least reasonably robust against these differences. We could not find any significant differences in the performance of the algorithms relative to the different centers using ANOVA (p=0.118).

We also investigated whether segmentation performance could be improved by making several algorithm combinations. First, a majority voting on the segmentation results of all algorithms and the top 5 best performing was calculated. Second, to get a reference for the best possible combination we took the best performing score per case. The summary results of these combinations can be found in Table 2.2. Taking the best results per case results in a substantially better average score than the best performing algorithms. This might be an indication that certain cases might be better suited to some algorithms, and as such, that algorithm selection should be performed on a case-by-case basis. The combinations of algorithms using majority voting also shows that given the correct combination, algorithm results can be improved (84.36 to 85.38 for the online challenge and 87.07 to 87.70 for the live challenge). Although the increase in score is small, it is accompanied by a reduction of the standard deviation (for the top 5 combination strategy, Table 2.2), as the improvements especially occur in poor performing cases. These scores and the reduction in standard deviation thus show that combining algorithms might result in more robust segmentation. These scores also show that there still is room for improvement for the individual algorithms. How to combine and which algorithms to combine is a nontrivial problem and warrants further investigation.

Finally, to assess the statistical significance of differences in algorithm performance we used repeated measures ANOVA with Bonferroni correction. The methods by Imorphics and ScrAutoProstate perform significantly better than all the algorithms outside of the top 3 ($p < 0.05$).

| Team Name | Avg. time | System | MT | GPU |
|---|---|---|---|---|
| Imorphics | 8 minutes | 2.83GHz 4-cores | No | No |
| ScrAutoProstate | 2.3 seconds | 2.7GGz 12-cores | Yes | |
| CBA | 4 minutes | 2.7GHz 2-cores | No | No |
| Robarts | 45 seconds | 3.2GHz 1-core, 512 CUDA-cores | No | Yes |
| Utwente | 4 minutes | 2.94GHz 4-cores | Yes | No |
| Grislies | 7 minutes | 2.5GHz 4-cores | No | No |
| ICProstateSeg | 30 minutes | 3.2GHz 4-cores, 96 CUDA-cores | No | Yes |
| DIAG | 22 minutes | 2.27GHz 8-cores | No | No |
| SBIA | 40 minutes | 2.9GHz, 2 cores | No | No |
| Rutgers | 3 minutes | 2.67GHz, 8-cores | Yes | No |
| UBUdG | 100 seconds | 3.2GHz 4-cores | No | No |

| Team Name | Availability | Remarks |
|---|---|---|
| Imorphics | Commercially available (`http://www.imorphics.com/`). | |
| ScrAutoProstate | Not available | No |
| CBA | Binaries available at: `http://www.cb.uu.se/~filip/SmartPaint/` | Fully interactive painting |
| Robarts | Available at `http://goo.gl/ZAbPpC` | User indicates 8 to 10 points on prostate surface |
| Utwente | Not available | User indicates prostate center |
| Grislies | Not available | |
| ICProstateSeg | Not available | |
| DIAG | Registration algorithm available on `http://elastix.isi.uu.nl/` | Runs algorithm on a cluster of 50 cores, average time without cluster 7 minutes per atlas |
| SBIA | Registration algorithm available on `http://www.rad.upenn.edu/sbia/software/dramms/` | Runs algorithm on a cluster of 140 cores, average time without cluster 25 minutes per atlas |
| Rutgers | Not available | |
| UBUdG | Not available | User selects first and last prostate slice |

Table 2.8: Details on computation time, interaction and computer systems used for the different algorithms. If algorithms where multi-threaded (MT) or used the GPU this is also indicated.

## 2.7 Future work and concluding remarks

Although in general the segmentation algorithms, especially the top 2, gave good segmentation results, some challenges still remain. As we could see in case 25 (Figure 2.5), algorithms sometimes struggle with the interface between the prostate and surrounding tissue. This is not only true for peri-prostatic fat, but also for the interface between the prostate and the rectum, the bladder and the seminal vesicles. Part of these challenges could be addressed by increasing through-plane resolution, but integration of these structures into the segmentation algorithms might also improve performance. Examples included coupled active appearance models[150] or hierarchical segmentation strategies[151]. Furthermore, the enormous volume differences that can occur in the prostate can also be problematic: case 23 had a volume which was approximately 6 times as large as the average. Automatically selecting appropriate atlas sets or appearance models based on an initial segmentation could be a solution. In the difficult cases the interactive segmentation method of team CBA was often the best. This shows that automated performance could still be improved.

Future work on prostate segmentation might also focus on the segmentation of related prostatic structures or substructures. Examples are segmentation of the prostatic zones (transition, central and peripheral), the neurovascular bundles or the seminal vesicles.

Solving these remaining issues might lead to algorithms which, for any case, can replace the tedious task of manually outlining by humans without any intervention. Until we are at that level, the challenge itself will remain online for new submissions and can thus be used as a reference for algorithm performance on multi-center data. As such it could lead to more transparency in medical image analysis.

# Segmentation of the prostate zones

# 3

Geert Litjens, Oscar Debats, Wendy van de Ven, Nico Karssemeijer, and Henkjan Huisman

*Original title:* A pattern recognition approach to zonal segmentation of the prostate on MRI

*Published in:* Lecture notes in computer science (2012);7511(1):413–420

## 3.1 Introduction

Prostate cancer is a major health problem in the Western world, with one in six men affected during their lifetime[152]. Multi-parametric magnetic resonance imaging (MPMR) has been shown to play an important role in the diagnosis of prostate cancer[32]. A typical MR exam contains T2-weighted, dynamic-contrast-enhanced and diffusion-weighted imaging. Interpretation of MPMR prostate studies is challenging, and therefore the use of computer-aided diagnosis techniques has been investigated[88]. For correct interpretation of MPMR knowledge about the zonal anatomy of the prostate is required, because the occurrence and appearance of cancer is dependant on its zonal location[95]. From a radiological point of view the prostate is usually considered to have two visible zones on MRI, the central gland (CG) and the peripheral zone (PZ)[153]. We are exploring options to integrate knowledge about the zonal anatomy into CAD systems. For this automated segmentation of the zones is the first step. The availability of zonal segmentation is also mandatory for those CAD methods in literature that focus on the PZ only, as for example in[88].

Although much research has been done on prostate segmentation[108,110], only recently the first study on segmentation of the individual zones was published by Makni et al.[154]. In their study they investigated the use of an evidential C-means clustering (ECM) approach to cluster voxels into their respective zones. In addition, they extended the ECM approach to incorporate the spatial relation between voxels. Using this method they obtained good results on their data set ($0.87 \pm 0.04$ mean Dice coefficient for the central gland compared to a simultaneous truth and performance level estimation (STAPLE) obtained ground truth[134]). To the best of the authors knowledge their paper remains the only published paper evaluating prostate zone segmentation.

The purpose of this paper was to investigate a pattern recognition algorithm to segment the prostate zones. The pattern recognition approach uses several image features with a voxel classifier to detect the zones. This is a method that has been explored in many other segmentation problems. We compare it to a multi-parametric multi-atlas approach which is used to simultaneously segment the prostate and the prostate zones. Additionally, we will compare our results to inter-observer variability and the results obtained by Makni et al.[154]

## 3.2 Methods

### 3.2.1 Multi-parametric multi-atlas segmentation

Multi-atlas segmentation is an accurate method for prostate segmentation, as has been shown by Klein et al.[108] We have chosen a similar approach, but extended it to use multi-parametric data. We evaluated the atlas method with both majority voting and STAPLE[134] to obtain the final binary segmentation.

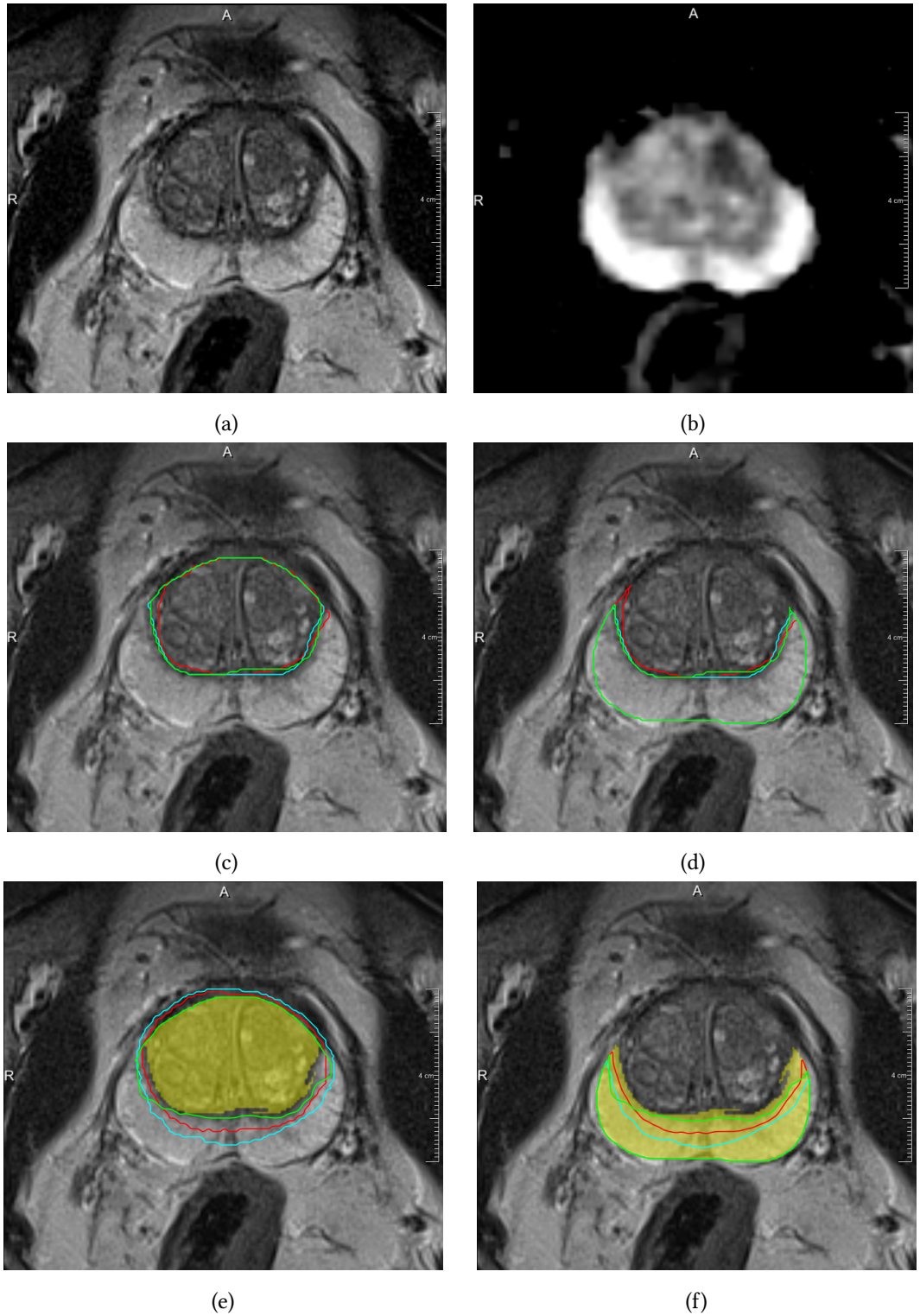



Figure 3.1: Example data set with T2-W image and ADC map in a and b and segmentation results in c, d, e and f. Figure c shows a central gland observer segmentation (red, cyan and green contours for observer 1,2 and 3). Figure d shows the corresponding peripheral zone segmentation. Figures e and f show the automatic segmentation results for the central gland and peripheral zone respectively. Contours are colored red, cyan and green for atlas (voting), atlas (STAPLE) and voxel classification respectively, the STAPLE constructed 'true' segmentation is overlayed in yellow

The registration of the atlases to the new case is performed using the elastix software package[155]. For the registration we use local normalized mutual information as a similarity metric. We register both the T2-weighted image and the quantitative apparent diffusion coefficient (ADC) map simultaneously. We chose to add the ADC map to the registration because it contains additional information on the zonal distribution within the prostate. In a previous experiment we investigated the added value of the ADC in zonal segmentation and we noticed that it improved performance. The cost function we then optimize can be expressed as

$$C(T_\mu; I_F, I_M) = \frac{1}{\sum_{i=1}^{N} \omega_i} \sum_{i=1}^{N} \omega_i C(T_\mu; I_F^i, I_M^i) \tag{3.1}$$

were C is the cost function, $T_\mu$ is the registration transformation, $I_F$ is the fixed (the unknown case) and $I_M$ the moving image (the atlas). Furthermore, $\omega_i$ is the weight for each of the multi-parametric images $i$ were $i = 1$ is the T2-weighted image and and $i = 2$ is the ADC map. We chose $\omega$ to be 0.5 for both $i$.

The registration consists of two distinct steps. In the first step we register using only a translation transform to align the images to the new case. The second step is an elastic registration using a b-spline transformation. After the registration the obtained transformation is used to transform the known binary segmentation to the target image space. These can subsequently be used to construct the unknown binary segmentation. Several approaches exist in literate, of which majority voting is the simplest and best known method[108]. We compare this approach with optimizing the segmentation by using STAPLE[134].

### 3.2.2 Voxel classification segmentation

For the voxel classification segmentation we determined a set of features that represent the difference between the two zones. These features can be separated into three categories: anatomy (positional), intensity and texture.

For the anatomy features we use the information we know from the normal prostate composition. The peripheral zone is usually situated at the dorsal side of the prostate, getting thicker towards the apex of the prostate. We chose to model this by developing a set of three relative position and distance features. Given the whole prostate mask we can calculate a relative position in each direction for each voxel, resulting in a value between 0 and 1. We calculate this feature in the ventrodorsal direction and the craniocaudal direction. In addition, the relative distance (also between 0 and 1) to the prostate boundary is given as a feature.

Two intensity features are included in the voxel classification step. The first intensity feature we use is the apparent diffusion coefficient (ADC) for each voxel, which itself should be a quantitative feature. The second intensity feature we use is a calculated T2 value for each voxel. Using the T2 relaxation time instead of the T2-weighted voxel values will make this feature much more robust to changes in scan parameters. To this end we used the following

signal model equation for turbo-spin-echo sequences:

$$T2_p = -TE \left( \log e^{\frac{-TE}{T2_m}} \frac{S_m^{PD} S_p^{T2W}}{S_m^{T2W} S_p^{PD}} \right)^{-1} \tag{3.2}$$

Here T2 is the estimated T2 relaxation time, TE is the echo time for the MR pulse sequence, S the signal intensity. The superscript PD and T2W represent either the proton density weighted image or the T2-weighted image. The subscript p and m denote prostate and muscle respectively. Using this equation and a region of interest placed in a skeletal muscle we can calculate the true T2 relaxation time for each voxel given the proton density and T2-weighted images.

The muscle ROI is automatically selected using a search method. Starting from the bottom slice of the T2-weighted image an Otsu threshold is performed to separate the dark areas (including the muscles) from the bright areas. We are looking for the two muscles alongside the prostate, so we suppress the center of the image with a rectangular block. Then a connected component analysis is used to find individual dark components in the image. The two largest connected components should correspond to the left and right muscle. We make sure this is the case by investigating the shape and symmetry of the two connected components. The muscle are less wide than long and they should have approximately the same shape on the left and right. We mirror the left connected component and investigate the Jaccard index with the right connected component. The minimum value for width divided by the length is 0.75 and the threshold for the Jaccard index is 0.5. The resulting connected components are eroded to ensure that the ROI is completely in the muscle.

The third set of features consists of five texture features. The first two features are homogeneity and correlation calculated using the co-occurrence matrix[156]. We used 16 gray value bins for the histogram and took the average over all 2D directions. The third and fourth feature are entropy and texture strength, based on the Neighborhood Gray-Tone Difference Matrix[157]. Here also 16 gray level bins were used, in combination with an evaluation distance of 1. For all of these features the kernel size was 10x10x1 voxels. The fifth feature was the local binary pattern at each voxel[158], which was calculated over a 3x3x1 voxel neighborhood. For this feature the images were down-sampled using Gaussian re-sampling such that a 3x3x1 neighborhood corresponded to a 12x12x1 neighborhood.

After calculating the features a balanced training set is constructed. Hard classification using a linear discriminant classifier is performed to obtain a binary segmentation of the central gland. To smoothen the initial boundary some post-processing is performed. Firstly, connected component analysis is used to select the largest connected component. Erosion and dilation are then performed to remove small objects attached to the segmentation. Finally the edge voxels between the central gland and the peripheral zone are selected and a thin plate spline is fitted through these voxels. This results in our final segmentation.

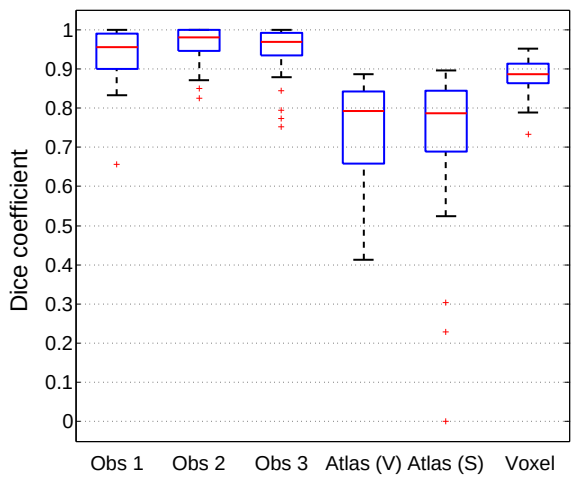
## 3.3 Validation

For validation we used 48 multi-parametric MR studies with manual segmentations of the whole prostate. For each case the transversal T2-weighted scan (resolution 0.6x0.6x4 mm) and the apparent diffusion coefficient map (2x2x4 mm) were used. In addition, for the voxel classification step, the proton density weighted image was used to calculate the T2 values. The ADC and proton density images were inspected to assess the alignment with the T2-weighted image. If needed, they were corrected to obtain good alignment.

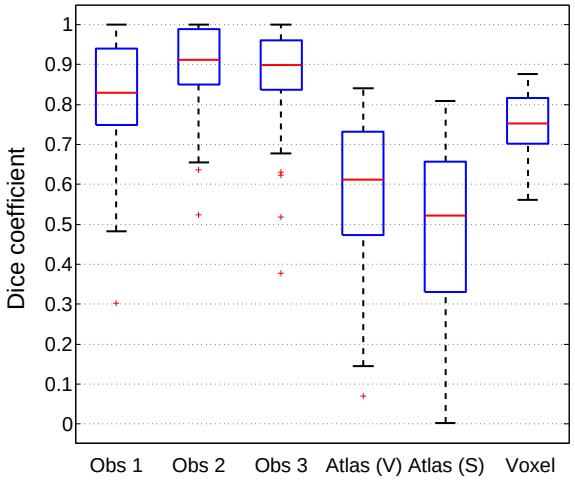
The ground truth was constructed by STAPLE[134] to merge the manual segmentations done by three observers. The observers made manual segmentations by indicating the zonal boundary on each T2-weighted image slice given the the manual whole prostate segmentation. We validated the automatic segmentations by calculating three similarity measures: the Jaccard index (JI), the Dice similarity coefficient (DSC) and the volume difference (VD). The Jaccard index is given as $J = \frac{|V_1 \cap V_2|}{|V_1 \cup V_2|}$, were $V_1$ and $V_2$ are the automated segmentation and the STAPLE ground truth respectively. The Dice coefficient is similar to the Jaccard index and can be expressed as $D = \frac{2|V_1 \cap V_2|}{|V_1|+|V_2|}$. Lastly, the volume difference can be expressed as $VD = |V_1| - |V_2|$. Validation was performed in a leave-one-out-manner, thus the case to be segmented was removed from the set of atlases for the atlas method and from the training data for the voxel classification.

## 3.4 Results

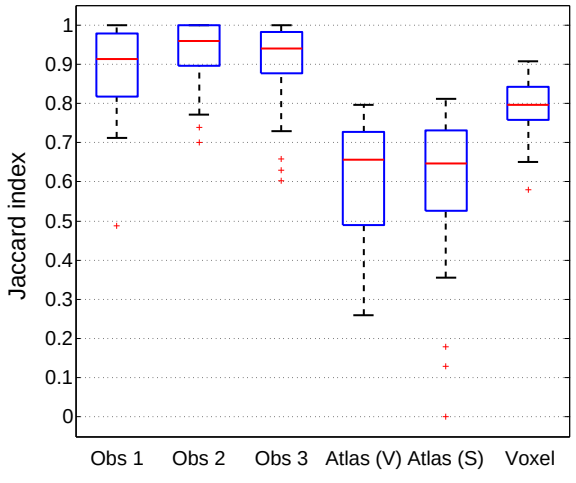
In figures 3.2a, 3.2c and 3.2e the results of the segmentations of the central gland are presented. An example case is also shown in figure 3.1. We can see that the observers all perform well with respect to the STAPLE ground truth. For the segmentation methods the voxel classification approach outperforms the atlas based methods (mean DSC 0.89 $\pm$ 0.03 vs 0.80 $\pm$ 0.013 for majority voting and 0.80 $\pm$ 0.17 for STAPLE), although it is not as good as the human observers (mean DSC's 0.95 $\pm$ 0.06, 0.97 $\pm$ 0.05, 0.96 $\pm$ 0.06). The JI and VD (figure 3.2b and figure 3.2c) show similar results. The VD results show that our methods in general under-segment the central gland. If we compare our results to those in Makni et al.[154] we perform slightly better using our voxel classification approach, as they report a mean DSC of 0.87 $\pm$ 0.04. For the peripheral zone we see similar results (figures 3.2b, 3.2d and 3.2f). Our pattern recognition approach outperforms the atlas based method and is relatively close to the observer scores. Here the pattern recognition approach has a mean DSC of 0.75 $\pm$ 0.07 compared to 0.82 $\pm$ 0.15, 0.89 $\pm$ 0.12 and 0.86 $\pm$ 0.11 for the observers. The atlas methods both perform poorly with respect to the peripheral zone with a mean DSC of 0.57 $\pm$ 0.19 and 0.48 $\pm$ 0.22. Compared to the state of the art we perform slightly worse, with a mean DSC of 0.76 $\pm$ 0.06 compared to our 0.75 $\pm$ 0.07.



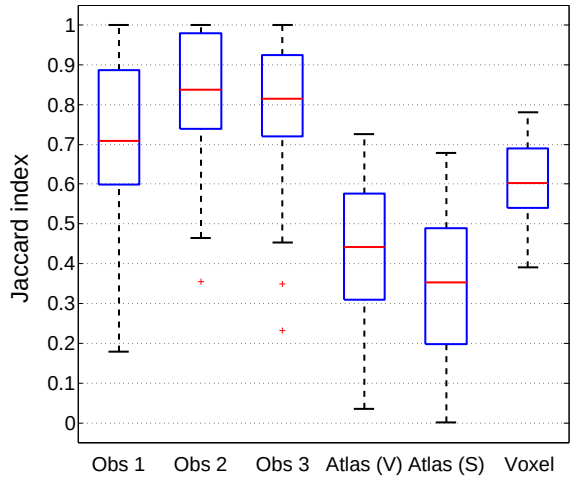

(a) Dice Coefficient (CG)



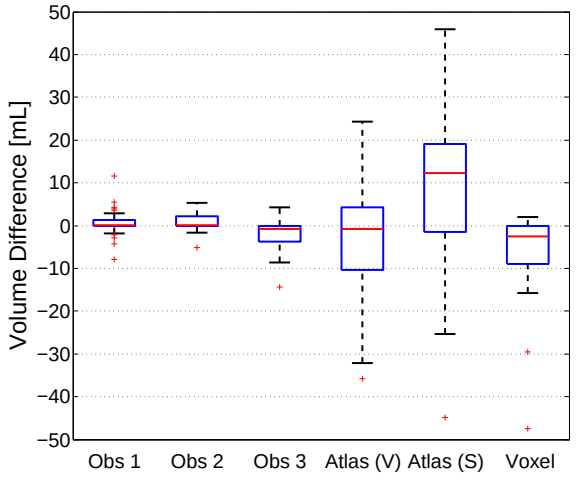

(b) Dice Coefficient (PZ)



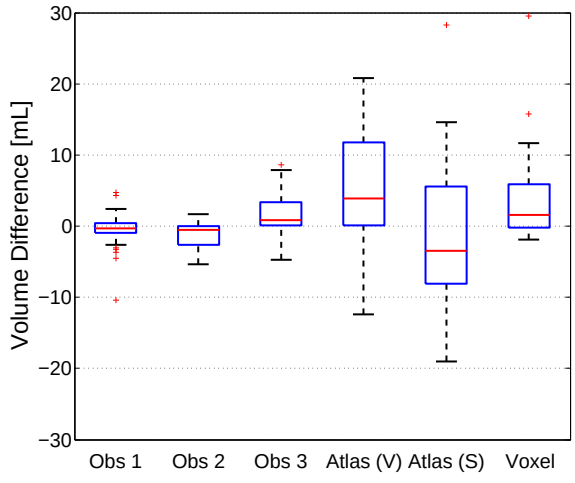

(c) Jaccard index (CG)




(d) Jaccard index (PZ)




(e) Volume difference (CG)




(f) Volume difference (PZ)

Figure 3.2: Results of the segmentation methods. The captions on the x-axes correspond to observers 1, 2 and 3, the atlas method using majority voting, the atlas method using STAPLE and the voxel classification approach.

## 3.5 Discussion

In this paper we investigated a pattern recognition approach to zonal segmentation of the prostate. We compared our method to an atlas based method and to the method published by Makni et al. Our results show that the voxel classification method outperforms the atlas based method. It also shows similar performance compared to the method published by Makni et al. We believe the pattern recognition approach outperforms the atlas-based method because it is less restrictive as an atlas, which is limited to the shapes available within the atlases. Additionally, patter recognition allow for non-linear combination of all features, including texture features.

This study also has limitations. A true comparison with the results from Makni et al. is difficult, mostly due to differences in the data used, for example in resolution. Additionally, for the atlas method we did not use the manual whole prostate segmentations because this method segments the whole prostate and the zones at the same time. This might cause some bias compared to the voxel classification approach were we did use the whole prostate manual segmentation. We did investigate using the manual whole prostate mask for the atlas method by only evaluating the registration metric within the mask. However, this approach gave worse results than not using the whole prostate mask at all. Both methods performed worst when the peripheral zone is very thin, then partial volume effects and unclear boundaries between the zones make it difficult to segment them. Finally, our voxel classification approach might be improved by incorporating additional texture features (e.g. Gaussian or Gabor based texture features) or by incorporating global information like prostate volume[154].

Summarizing, a new pattern recognition approach to segment the prostate zones was presented, incorporating anatomical, intensity and texture features. It outperforms an atlas based method, is relatively close to the inter-observer performance and shows similar performance compared to the state of the art.

# Features

# Features to discriminate benign disease from prostate cancer

# 4

Geert Litjens, Robin Elliott, Natalie Shih, Michael Feldman, Thiele Kobus, Christina Hulsbergen - van de Kaa, Jelle Barentsz, Henkjan Huisman, Anant Madabhushi

*Original title:* Computer-extracted features can distinguish benign confounding disease from prostatic adenocarcinoma on multi-parametric MRI.

*Submitted to:* Radiology

## 4.1 Introduction

MRI is becoming an increasingly popular tool for prostate cancer diagnosis[31,32,44,49], leading to the release of standardized guidelines for acquiring, reading and reporting prostate MRI by the European Society of Urogential Radiologists (ESUR): the Prostate Imaging and Reporting Data Standard (PIRADS)[33]. Initial results on reporting prostate MRI using the ESUR guidelines have been promising, both with respect to overall performance and inter-reader agreement[58–60,62]. However, these initial studies have also shown a large trade-off between sensitivity and specificity, depending on the PIRADS score used as a threshold for biopsy[58,59,62].

In a recent publication by Rosenkrantz et al.[159], 5 out of the 10 named pitfalls in prostate MRI are related to benign confounders, i.e. benign disease mimicking the appearance of cancer. Some typical confounders are prostatic intra-epithelial neoplasia (PIN), atrophy, inflammation and benign prostatic hyperplasia (BPH). For example, it is well known that the apparent diffusion coefficient (ADC) obtained from diffusion-weighted imaging (DWI) is a very capable feature in identifying cancer in the peripheral zone[51,160]. However, it is much more difficult to use in the transition zone[103,161], where the presence of BPH, which tends to have similar appearance to adenocarcinoma on ADC maps, is a major confounder. Additionally, prostatitis and other inflammatory processes within the prostate have been known to cause similar appearance to cancer on dynamic-contrast enhanced (DCE) MRI[162]. Improved understanding of the imaging characteristics of these confounders across MRI parameters might help radiologists improve their diagnostic ability.

Previous research has only peripherally focused on identifying discriminatory features to separate cancer from specific confounders[163–166]. Oto et al. investigated the use of the apparent diffusion coefficient (ADC) to differentiate between central gland tumors and glandular and stromal hyperplasia by visually registering the pathology slides to the MR images. They were able to achieve an area under the curve of 0.78 and 0.99 for differentiation of stromal and glandular hyperplasia and prostate cancer, respectively. Liu et al. designed a bi-exponential diffusion model using 10 b-values to characterize central gland lesions as prostate cancer and BPH. They found that the bi-exponential model (AUC of 0. 92 for ADCs) significantly improved the discriminative performance of DWI compared to a mono-exponential model (ADC of 0.8 for ADC). Chesnais et al. used a multi-parameter approach to differentiate central gland nodules. They found that the T2-weighted features and the ADC values appeared to play a substantial role in characterizing central gland nodules and that DCE imaging did not seem to provide complementary information. However, they did not evaluate the accuracy of different feature subsets.

The goal of this work is to identify the best combination of MRI parameters and computer-extracted features for each pair-wise classification task, i.e. cancer vs. BPH, PIN, inflammation and atrophy, respectively. This approach has recently seen successful application in discrim-

inating different types of breast cancer[167]. Our hypothesis is that each confounding class requires a distinct set of features to be able to discriminate it from prostate cancer successfully and that this pair-wise classification approach will yield improved discriminability compared to a monolithic classifier attempting to distinguish cancer from all benign classes simultaneously.

## 4.2 Materials and methods

### 4.2.1 Patient data

Pre-operative multi-parametric MRIs and radical prostatectomy specimens were included retrospectively for 70 patients at the Radboud University Nijmegen Medical Centre. MRIs were acquired between January 1st 2009 and June 1st 2013. The Institutional Review Board waived the need for informed consent.

### 4.2.2 MRI acquisition

MRI acquisition was performed using a 3 Tesla MRI scanner (either a TrioTim or a Skyra; Siemens, Erlangen, Germany). Cases were acquired both with and without an endorectal coil. A pelvic phased array coil was always used. The multi-parametric protocol consisted of three T2-weighted images in orthogonal directions, diffusion-weighted imaging (three b-values averaged over three orthogonal directions, 50, 400-500 and 800) and dynamic contrast-enhanced imaging (15 mL of Dotarem; Guerbet, France). The transversal T2-weighted images were acquired perpendicular to the rectal wall, the diffusion-weighted imaging and the dynamic contrast-enhanced imaging were acquired in the same orientation. Further acquisition details can be found in Table 4.1.

### 4.2.3 Prostatectomy slide selection and annotation

staining the specimens were evaluated by one expert urological pathologist (C.H.v.d.K, with 20 years of experience). The pathology slides were cut in the same orientation as the acquisition of the transversal MRI to remove angulation errors in subsequent registration steps. Tumors were outlined on the microscopic slides and subsequently transferred to the macroscopic photographs of the specimens.

The H&E stained slide containing the tumor with the highest Gleason score was selected to be digitized using a digital slide scanner (VS120-S5, Olympus, Japan) at 10x or 20x, corresponding to a resolution of 0.6 um and 0.3 um respectively. If multiple slices contained a tumor with the same Gleason grades, the slice with the largest tumor volume was digitized. Approximately half of the specimens were whole-mount slides, the other half consisted of parts (usually two or four). In case the specimen consisted of parts, all parts belonging to

| | PS | SR | ST | NS | ET | RT | FA | Other |
|---|---|---|---|---|---|---|---|---|
| T2W | Turbo spin-echo | 0.28 – 0.6 mm | 3.0 – 4.0 mm | 13 - 19 | 101 – 104 ms | 3540 – 6840 ms | 120 - 160 | Acquired in three orthogonal directions: transversal, sagittal and coronal. |
| DWI | Echo planar | 1.6 - 2 mm | 3 mm - 4 mm | 15 - 20 | 61 – 81 ms | 2300 – 3600 ms | 90 | 3 b-values: 50, 400 – 500, 800 averaged over 3 directions. Apparent diffusion coefficient map calculated by the scanner software. Some scans also include b-value 0. |
| DCE | Turbo Flash | 1.5 – 1.8 mm | 3.2 – 5 mm | 12 - 15 | 1.41 - 1.47 ms | 36 ms | 10 - 14 | Temporal resolution of 3.38 – 4.65 seconds, 36 – 50 timepoints. 15 mL contrast agent used (Dotarem, Guerbet, France) |

Table 4.1: MRI sequence details for the different types of acquisitions. PS = pulse sequence, SR = spatial resolution, ST = slice thickness, NS = number of slices, ET = echo time, RT = repetition time, FA = flip angle.

one slide were digitized. After digitization the digital slides were annotated using the Aperio ImageScope software (Aperio, USA) for the presence of cancer, BPH, PIN, atrophy or inflammation by one of two urological pathologists (N.S. with 8 years of experience or R.E. with 7 years of experience).

### 4.2.4 Co-registration of prostatectomy specimens and MRI

To map the annotations on the histopathology sections to the corresponding MRI sections, the MRI and the pathology slide have to be registered. First, the slice in the MRI corresponding to the prostatectomy slide has to be established[168]. The number of slices in the MRI the prostate was visible on were counted. Subsequently, the number of slides in the prostatectomy was counted. Using the number of the prostatectomy slide the most likely corresponding MRI slice is then given as:

$$S_{MR} = \frac{T_{MR}}{T_P} S_P \tag{4.1}$$

where $S_{MR}$ is the slice number in the MRI, $T_{MR}$ the total number of prostate slices in the MRI, $T_P$ the total number of prostate slices in pathology and $S_P$ the slice number of the selected pathology slice. This is similar to the approach presented by Hambrock et al.[51]. The selected MR and pathology slices where subsequently visually assessed for correspondence by a medical imaging researcher (G.L., four years of experience with prostate MRI) and corrected if deemed necessary. After establishing the corresponding slice it was registered to the MRI using an interactive b-spline elastic registration method, which has successfully been applied in

a number of studies[94,95]. To drive the registration corresponding points on the boundary of the prostate on the MRI and the pathology were selected by a medical imaging researcher (G.L., four years of experience with prostate MRI). After the corresponding points were established, the registration algorithm mapped the prostatectomy slide and the annotations to the corresponding MRI section. During selection of the boundary points, the researcher was blinded to the pathology annotations. An example of the process is illustrated in Figure 4.1.

### 4.2.5 Computer-extracted features

Following co-registration, a number of MRI and computer-extracted features were obtained from within the regions corresponding to the cancer, BPH, PIN, atrophy and inflammation. To obtain a single feature vector per region of interest (ROI) mapped onto the MRI, the median value of each feature across the voxels within the ROI is calculated. All features are calculated in 2D, as we register a single prostatectomy slide to the MRI, resulting in 2D annotations. A listing of these features and their associated descriptions can be found in Table 4.2.

| Category | Feature name | Calculated on | Parameters |
|---|---|---|---|
| Intensity | T2Ws[169] | T2W (Transversal) | |
| | ADC | DWI | |
| | b800 | DWI | |
| Texture | 2D multi-scale Gaussian derivatives | T2Ws | Up to 2nd order, $\sigma$=2.0, 2.7, 4.1 and 6.0mm |
| | 2D multi-angle Gabor | T2Ws | $\theta$=0, $\frac{\pi}{4}$, $\frac{\pi}{2}$, $\frac{3\pi}{4}$. $\lambda$=2, 3 and 4 mm |
| | 2D Li multi-scale blobness[170] | T2Ws, ADC, b800, $K^{trans}$, $k_{ep}$, $\nu_e$, time-to-peak, maximum enhancement, wash-out rate | $\sigma$=2.0, 2.7, 4.1 and 6mm |
| Pharmacokinetic | Time-to-peak[55] | DCE | |
| | Maximum enhancement[55] | DCE | |
| | Wash-out rate[55] | DCE | |
| | $K^{trans}$[57] | DCE | |
| | $\nu_e$[57] | DCE | |
| | $k_{ep}$[57] | DCE | |

Table 4.2: Overview of all the features that are used in this paper including references to the relevant papers.

*MR intensity features*

MR intensity features are extracted from the transversal T2-weighted image volume and the diffusion-weighted imaging. In T2-weighted imaging the non-standardness of the MR intensities, especially between endorectal coil and non-endorectal coil cases, can cause problems for quantitative computerized analysis. As such we developed a method which uses MR pulse sequence equations, a proton-density-weighted image and an automatically segmented muscle ROI to remove most of the T2-weighted intensity non-standardness[169].

In addition to the standardized T2-intensity we included the apparent diffusion coefficient (ADC) as a feature in combination with the image intensity of the b800 image.

*Texture features*

We calculated several popular texture filters, namely Gaussian derivatives and Gabor features, which have shown to be successful in discriminating prostate cancer from other tissue in previous studies[94,95]. To make sure these feature do not also suffer from intensity non-standardness we calculated them on the standardized T2W image.

Furthermore, to assess the focality of lesion appearance on the different MRI parameters, several blobness features were calculated using the techniques presented by Li et al[170]. Parameter settings for these features are listed in Table 4.2.

*Pharmacokinetic features*

DCE MRI has been shown to differentiate between inflammation and prostate cancer relative to normal tissue[162]. In clinical diagnosis often the shape of the enhancement curve is used to assess lesion malignancy. However, several groups have developed methods to more quantitatively evaluate the tissue curves, including pharmacokinetic modelling[55,162,171]. We use the methods presented by Huisman et al.[55] and Vos et al.[57] to calculate pharmacokinetic features.

### 4.2.6 Feature selection and classification

We used sequential forward floating feature selection (SFFS,[172]) in combination with a linear discriminant classifier to assess the most discriminative features. SFFS is a feature selection technique in which at each step one feature is added or removed based on a performance metric; we used the area under the receiver-operating characteristic curve (AUC). In our setup we force the feature selection to find the 5 most relevant features for each pair-wise classification task (cancer vs. BPH, atrophy, inflammation and PIN, respectively).

We repeated the SFFS procedure to investigate whether the selected features are influenced by cancer grade. We specifically looked at intermediate- and high-grade cancer. Intermediate-grade cancer was defined as cancer with a Gleason grade 3+4 and high-grade cancer was defined as any cancer with a major 4 or any 5 component.

In order not to bias the feature selection scheme, a two-fold, patient-stratified, cross-validation scheme was employed to select the features and the procedure was repeated 100 times. We identified the 5 features that most frequently appeared in the top 5 selected features.

Subsequently we wanted to assess whether using these features could result in accurate classification. A ten-fold, patient-stratified, cross-validation scheme was employed to train a random forest classifier (RFC)[74] in conjunction with the top ranked features identified for each of the pair-wise classification tasks (BPH, PIN, inflammation and atrophy vs. cancer). This experiment was also performed specifically for the subsets of the intermediate-grade and high-grade cancers to assess whether any differences in classification performance could be observed. The performance of the classifiers was evaluated using the area under the receiver operating characteristic curve (AUC). Bootstrapping was used to obtain the 95% confidence intervals for the AUCs. Finally, we also performed the feature selection and classification experiments for a monolithic classifier that attempted to distinguish cancer from all the benign confounders, grouped as a single benign category.

## 4.3 Results

### 4.3.1 Patient data

After annotation and co-registration of the prostatectomy slides and the multi-parametric MRI for all patients 92 PIN, 64 atrophy, 120 inflammation and 73 BPH ROIs were identified. In addition, 128 cancer ROIs were identified: 33 Gleason 3+3, 55 Gleason 3+4, 23 Gleason 4+3, 8 Gleason 4+4 and 9 Gleason 4+5. These results are also summarized in Table 4.3. Additionally, two example results for the MRI/pathology fusion are illustrated in Figure 4.1.

### 4.3.2 Feature selection

The results for the feature selection experiments are shown in Table 4.4, Figure 4.2 and Figure 4.3. In Table 4.4 it is interesting to note that the most highly ranked feature is different for all pair-wise classification tasks. Furthermore, distinctly different feature sets are chosen for each of the confounder classes. To gain some insight into why these features are selected, we present some quantitative and qualitative results in Figure 4.2 and Figure 4.3. The graphs in Figure 4.2 represent fitted histograms for the feature value distributions of a specific confounder, all benign confounders grouped together, and cancer. Figures 4.2a, 4.2b, 4.2c and 4.2d show atrophy, BPH, inflammation and PIN, respectively. The plots illustrate that the value distribution for the top selected feature in all pair-wise classification tasks shows a distinctly different distribution than when all the confounders are grouped together. For example, in Figure 4.2d we show that the ADC value of PIN has less overlap with cancer than all the confounders

| | |
|---|---|
| Number of patients | 70 |
| PSA level, mg/ml, median (range) | 9.2 (1 – 76) |
| Age, yr, median (range) | 62 (48 – 70) |

| **Gleason score** | **No. of lesions** | **Category** | |
|---|---|---|---|
| Normal/Benign | 349 | Total | 477 |
| 3 + 3 | 33 | Atrophy | 64 |
| 3 + 4 | 55 | Inflammation | 120 |
| 4 + 3 | 23 | BPH | 73 |
| 4 + 4 | 8 | PIN | 92 |
| 4 + 5 | 9 | Cancer | 128 |

Table 4.3: Characteristics of patients and identified lesions. PSA ranges were determined on 49 patients, for 21 patients PSA levels prior to MRI were unknown.

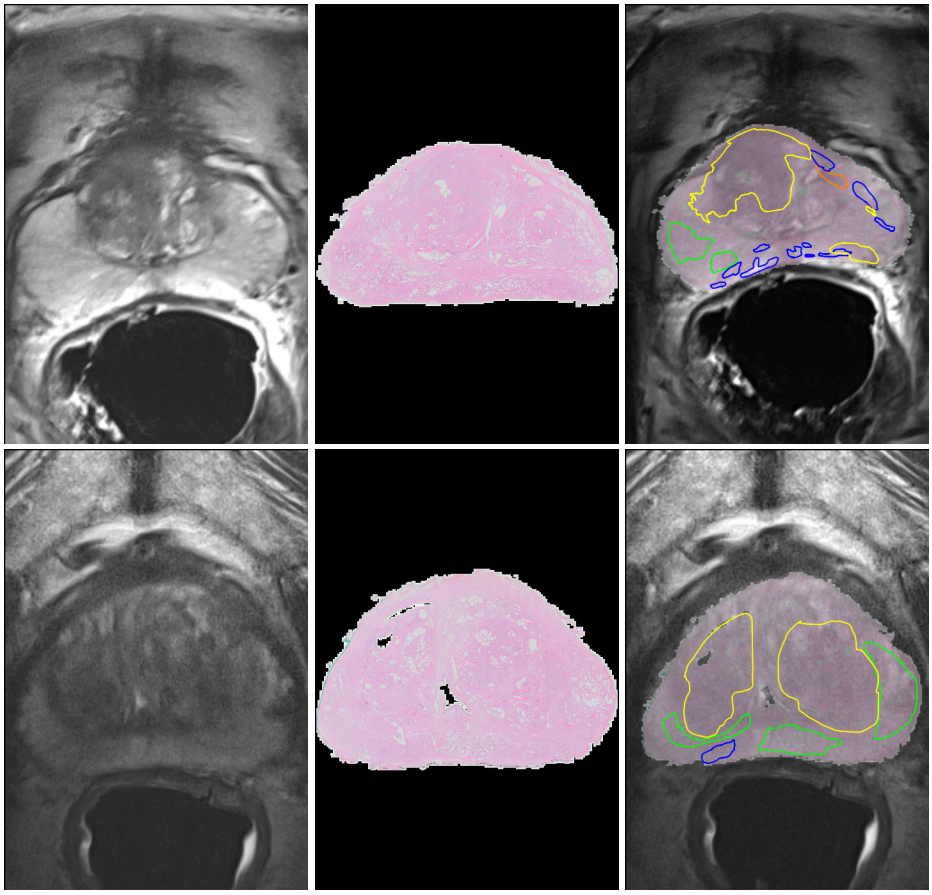


Figure 4.1: Two examples of co-registration results of the MRI and the prostatectomy slide. The last column shows the annotations made on the pathology images overlayed on the MRI/pathology fused images.

grouped together. In Figure 4.3 qualitative results for the top three features for each of the benign classes versus cancer are presented. Here one can appreciate the visual difference for the feature values in cancer and the individual benign confounding classes. The influence of cancer grade on the selected features was also assessed and is presented in Table 4.4. It may be appreciated that for high-grade cancer, ADC is the most important feature across all pair-wise classification tasks. The variety in selected features across the different pair-wise classification tasks is reduced compared to discriminating confounders from all cancer grades grouped together, with a focus mostly on T2-weighted (texture) and diffusion-weighted imaging features. For intermediate grade cancer (Gleason score 3+4) the types of selected features show more variety between the pair-wise classification tasks (similar to all cancer grades grouped together) and also include more dynamic contrast-enhanced features.

### 4.3.3 Classification

Table 4.5 shows the quantitative classification results illustrating the performance of the pair-wise classifiers when using only the selected five features.

The average area under the curve for the pair-wise classification was 0.70 (BPH, PIN, inflammation and atrophy versus cancer had AUCs of 0.69, 0.73, 0.63 and 0.75, respectively) and for the monolithic classifier (cancer versus all benign classes) 0.62 using only the top-selected features.

The results for discriminating confounders from high-grade and intermediate-grade cancer are also presented in Table 4.5 (2nd and 3rd rows). On average the AUCs for discriminating high-grade cancer from the benign confounders are a bit higher, whereas the AUCs for discriminating intermediate-grade cancers are a bit lower.

## 4.4 Discussion

Presence of benign confounding disease (e.g. BPH, inflammation, PIN, atrophy) is the most common reason for false positives in prostate cancer diagnosis on multi-parametric prostate MRI[159]. In this study we used computerized image analysis and feature extraction to identify sets of features to best separate each of the confounding classes (BPH, PIN, inflammation and atrophy) from prostate cancer on MP-MRI.

Our study shows that the appearance of a lesion on a high b-value image might have a higher discriminatory value than the ADC when BPH is present (Figures 4.2b and 4.3, second row, second column), which also is in line with previous reports in literature[103,161]. Additionally, if the patient is at high risk of prostate inflammation at the time of the MRI, for example due to recent biopsy, the results suggest that looking at lesion focality (i.e. roundness, diffuse or well-defined edges) on both T2-weighted imaging and dynamic contrast-enhanced imaging (Table 4.4, Figure 4.2c) could help in discriminating inflammatory processes from prostate can-

| **All** | **PIN** | **N** | **Atrophy** | **N** | **Inflammation** | **N** | **BPH** | **N** | **Monolithic** | **N** |
|---|---|---|---|---|---|---|---|---|---|---|
| *1* | ADC | 117 | Blob. ($v_e$) | 117 | Blob. (T2Map) | 127 | Blob. (b800) | 105 | ADC | 179 |
| *2* | Gauss. Deriv. (XX, σ=4.1) | 99 | Blob. (b800) | 116 | ADC | 83 | Blob. (T2Map) | 94 | Blob. (b800) | 101 |
| *3* | T2Map | 81 | $K^{trans}$ | 109 | Blob. ($v_e$) | 69 | ADC | 86 | Gabor (b=1, λ=2, θ=0.4) | 79 |
| *4* | Blob. ($K^{trans}$) | 66 | Washout | 91 | Blob. ($K^{trans}$) | 63 | Blob. ($k_{ep}$) | 86 | Blob. (T2Map) | 77 |
| *5* | Gauss. Deriv. (XX, σ=6) | 65 | ADC | 87 | Washout | 52 | Time-to-peak | 59 | Blob (Max. Enh.) | 75 |
| **HG** | **PIN** | **N** | **Atrophy** | **N** | **Inflammation** | **N** | **BPH** | **N** | **Monolithic** | **N** |
| *1* | ADC | 117 | ADC | 114 | ADC | 83 | ADC | 117 | ADC | 149 |
| *2* | Blob. (Time-to-peak) | 88 | Blob. (Max. Enh.) | 52 | Blob. (ADC) | 80 | Blob. (T2Map) | 82 | Blob. (ADC) | 106 |
| *3* | Blob. (ADC) | 84 | Gabor (b=1, λ=2, θ=0) | 52 | Blob. (T2Map) | 76 | Blob. (b800) | 50 | Gauss. Deriv. (YY, σ=4.1) | 50 |
| *4* | Gabor (b=1, λ=4, θ=0.4) | 70 | $K^{trans}$ | 52 | Gauss. Deriv. (XY, σ=6) | 72 | Gauss. Deriv. (YY, σ=4.1) | 50 | $v_e$ | 49 |
| *5* | Gauss. Deriv. (Y, σ=6) | 57 | $k_{ep}$ | 51 | Blob. ($k_{ep}$) | 71 | Gauss. Deriv. (XY, σ=6) | 49 | Blob. (Time-to-peak) | 47 |
| **Inter.** | **PIN** | **N** | **Atrophy** | **N** | **Inflammation** | **N** | **BPH** | **N** | **Monolithic** | **N** |
| *1* | Gauss. Deriv. (XX, σ=4.1) | 128 | Gabor (b=1, λ=2, θ=0) | 105 | ADC | 84 | Blob. (b800) | 145 | Blob. (T2Map) | 121 |
| *2* | ADC | 98 | Washout | 84 | $K^{trans}$ | 73 | ADC | 86 | Blob. ($v_e$) | 120 |
| *3* | Gabor (b=1, λ=2, θ=0) | 86 | $K^{trans}$ | 80 | Blob. ($v_e$) | 57 | Gabor (b=1, λ=2, θ=0) | 70 | ADC | 109 |
| *4* | Blob. (Time-to-peak) | 67 | Blob. (b800) | 77 | Gabor (b=1, λ=2, θ=0.8) | 53 | Blob. (T2Map) | 66 | Blob. (b800) | 84 |
| *5* | Gauss. Deriv. (XX, σ=6) | 58 | Time-to-peak | 76 | Gabor (b=1, λ=4, θ=0.8) | 51 | Gabor (b=1, λ=4, θ=0) | 62 | Blob. ($K^{trans}$) | 77 |

Table 4.4: Selected features for each of the different pair-wise classification tasks considered in this work (e.g. PIN versus cancer, BPH versus cancer).The N columns show how often a feature was selected during the feature selection phase. The last two columns show the features selected for the monolithic classification task. The top part of the table shows the results for discriminating the benign confounders from all cancer. The middle part for discriminating confounders from only high-grade cancer and the bottom part for discriminating confounders from intermediate-grade cancer. For features which were calculated with different parameters the parameter values are presented in addition to the feature, e.g. Gaussian derivatives show the order and the scale on which they are calculated.

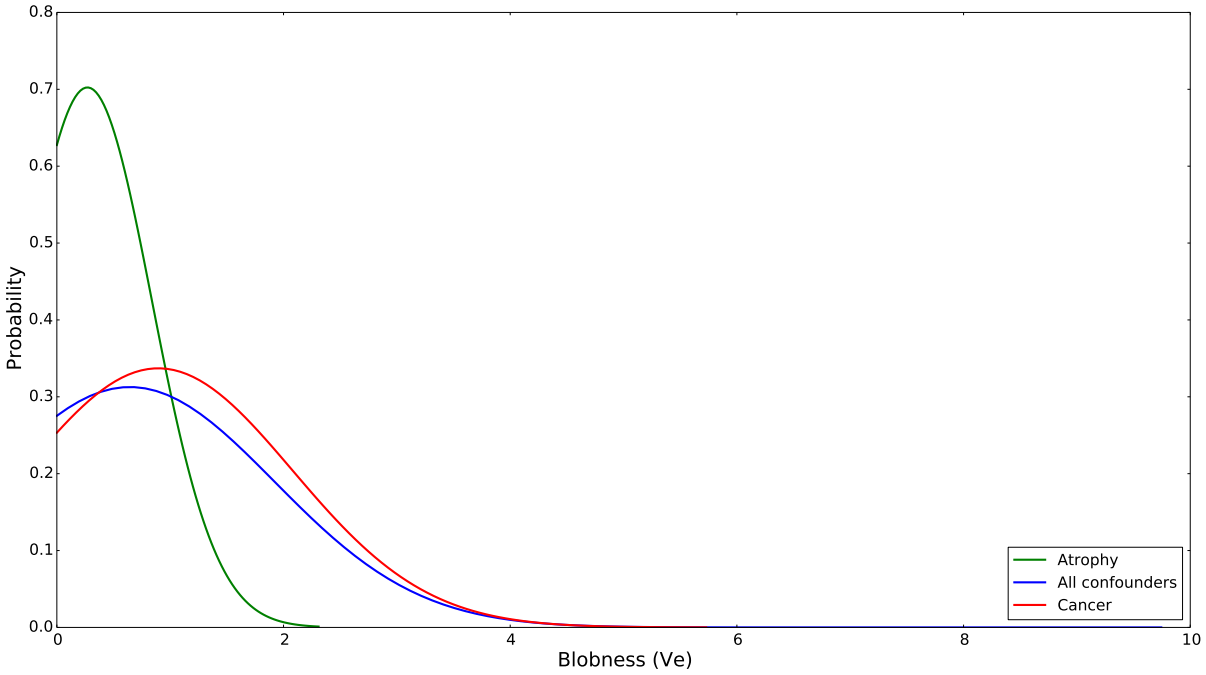
| Grade | PIN | Atrophy | Inflammation | BPH | Monolithic |
|---|---|---|---|---|---|
| All | 0.73 (0.67 – 0.80) | 0.75 (0.69 – 0.82) | 0.63 (0.54 – 0.70) | 0.69 (0.61 – 0.76) | 0.62 (0.53 – 0.70) |
| High | 0.73 (0.63 – 0.81) | 0.77 (0.66 – 0.86) | 0.77 (0.67 – 0.84) | 0.69 (0.55 – 0.82) | 0.64 (0.55 – 0.74) |
| Intermediate | 0.65 (0.57 – 0.72) | 0.70 (0.61 – 0.79) | 0.57 (0.47 - 0.66) | 0.63 (0.54 – 0.73) | 0.62 (0.56 – 0.69) |

Table 4.5: Classification performance (area under the receiver operating characteristic curve, AUC) of a random forest classifier for each of the pair-wise classification tasks (PIN, atrophy, inflammation and BPH vs. cancer). Furthermore, the AUC for a monolithic classification (grouping all benign confounders together as a single benign class) is presented in the last column. The second and third row show the AUCs when looking at a high-grade or intermediate-grade subset of cancer.

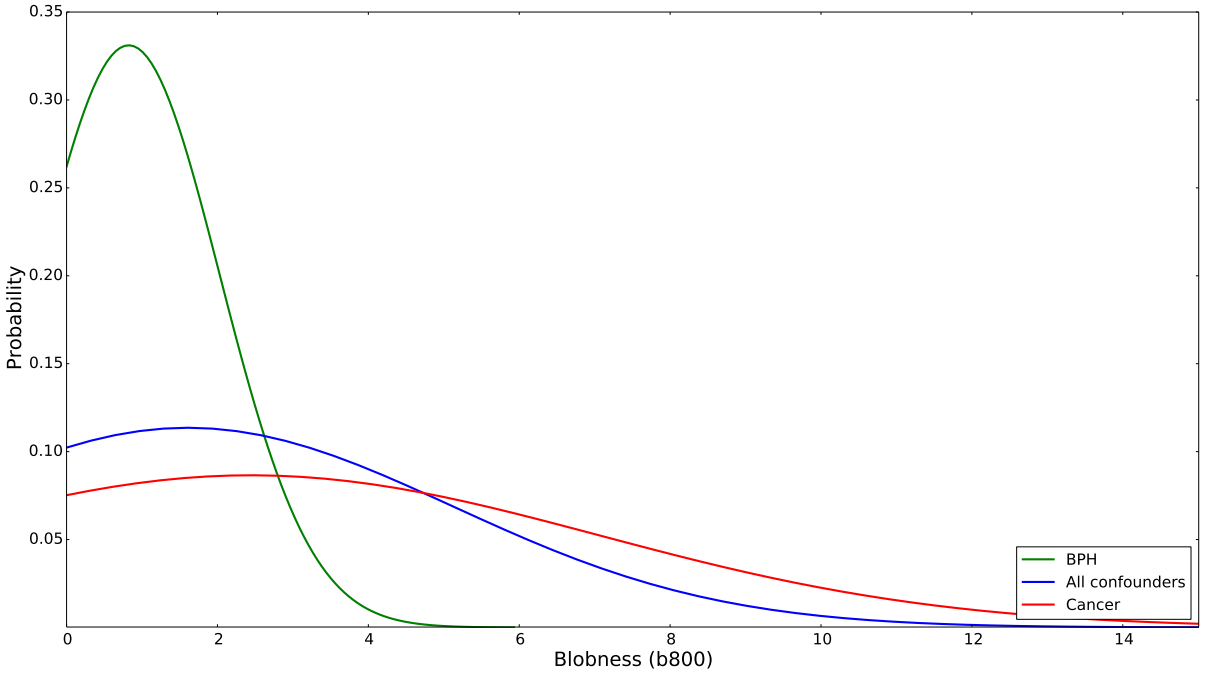
cer. This appears to support our initial hypothesis that each benign confounding class appears to have a distinct set of imaging descriptors that can help characterize them. This is further confirmed by the classification results presented in Table 4.5, where the average and individual AUCs of the pair-wise classification tasks seem higher than the AUC for the monolithic classification.

If we inspect the influence of cancer grade on the selected features it is apparent that for high-grade cancer the variety of selected features across the pair-wise classification tasks (Table 4.4) is reduced. This is accompanied by an increased AUC for the diagnosis of high-grade cancer relative to all cancer grades grouped together (Table 4.5). Both these results indicate that high-grade cancer seems to have its own distinct imaging characteristics (low ADC, distinctly different texture) compared to all other classes (even to low- and intermediate-grade cancer) and is easier to discriminate from benign disease than intermediate-grade cancer (which has a lower AUC on average). Results presented in literature on ADC also show that the difference between benign/normal prostate lesions and high-grade cancer is so large that it is relatively easy to discriminate between the two, even in the presence of BPH, by just using ADC[51,163], whereas the overlap in ADC between intermediate-grade cancer and benign disease is much larger, requiring more and more specific features to allow discrimination.

We acknowledge that our study had its limitations. There is invariably a time lapse between the MRI and the prostatectomy, as such sometimes the appearance or size of a lesion is not completely similar between the two. Additionally, we limited our analysis to a single prostatectomy slide, as 3D reconstruction of an entire prostatectomy and subsequent mapping to the MRI is difficult and often impossible in current diagnostic practice. Choyke et al. presented a method using a 3D printed mold, which could be an avenue for future research[173]. Additionally, to keep the data unbiased we did not edit or remove annotations based on size or MR visibility. Some annotations on the prostatectomy resulted in lesions which are only a couple of voxels large and suffer from partial volume effect, making it difficult to register and characterize them correctly. No statistical significance test was performed on the difference between the monolithic and the pairwise classification. The monolithic classification contains

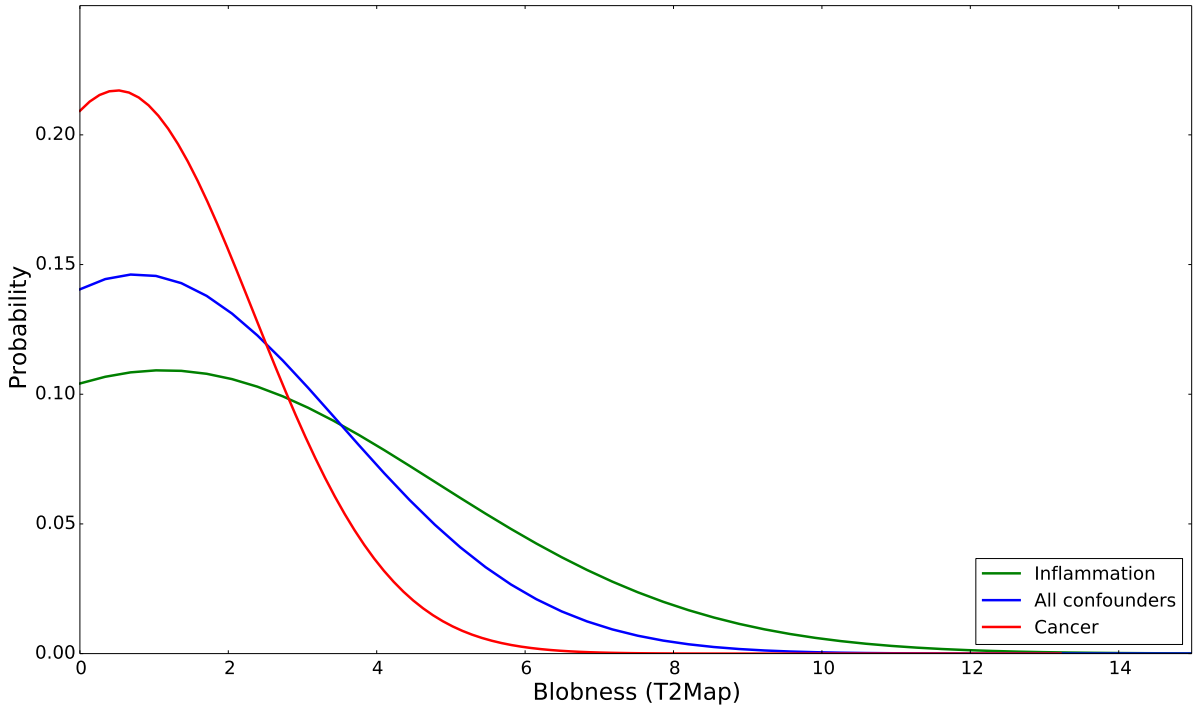



(a)

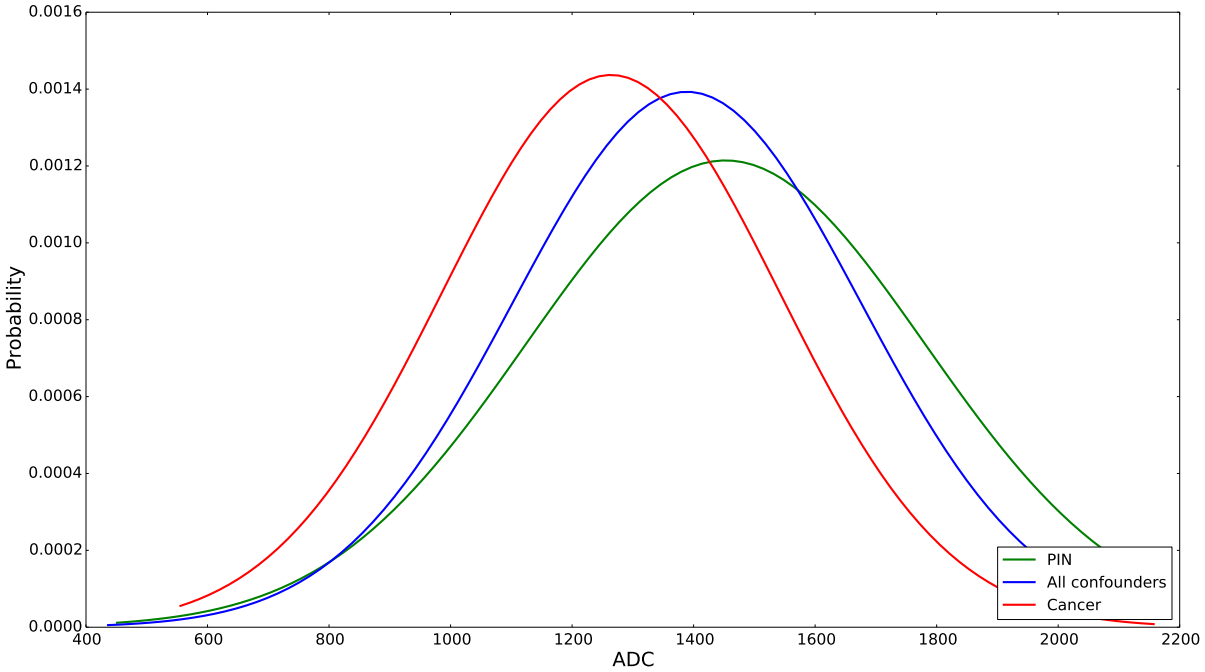



(b)

Figure 4.2: Distributions of the feature values of the top selected feature of each of the classification tasks for cancer, all benign confounders and the confounder to discriminate. Figures (a) and (b) show atrophy and BPH respectively,




(c)




(d)

Figure 4.2: Distributions of the feature values of the top selected feature of each of the classification tasks for cancer, all benign confounders and the confounder to discriminate. Figures (c) and (d) show inflammation and PIN, respectively.

all benign samples, whereas each pair-wise classification only contains the specific benign confounder class, making it impossible to perform a paired t-test. As the samples are not independent a regular t-test is also not applicable. In the future this could be solved by including a completely separate test set for the monolithic classification. Last, multi-center evaluation is an important pre-requisite for further validation of the results.

## 4.5 Concluding remarks

We explored and showed the utility of computerized image and feature analysis in conjunction with multi-parametric MRI to distinguish between prostate cancer and benign confounders. For each pair of HGPIN, atrophy, BPH, inflammation versus cancer we identified a unique set of features which could help improve the differential diagnosis of prostate cancer.

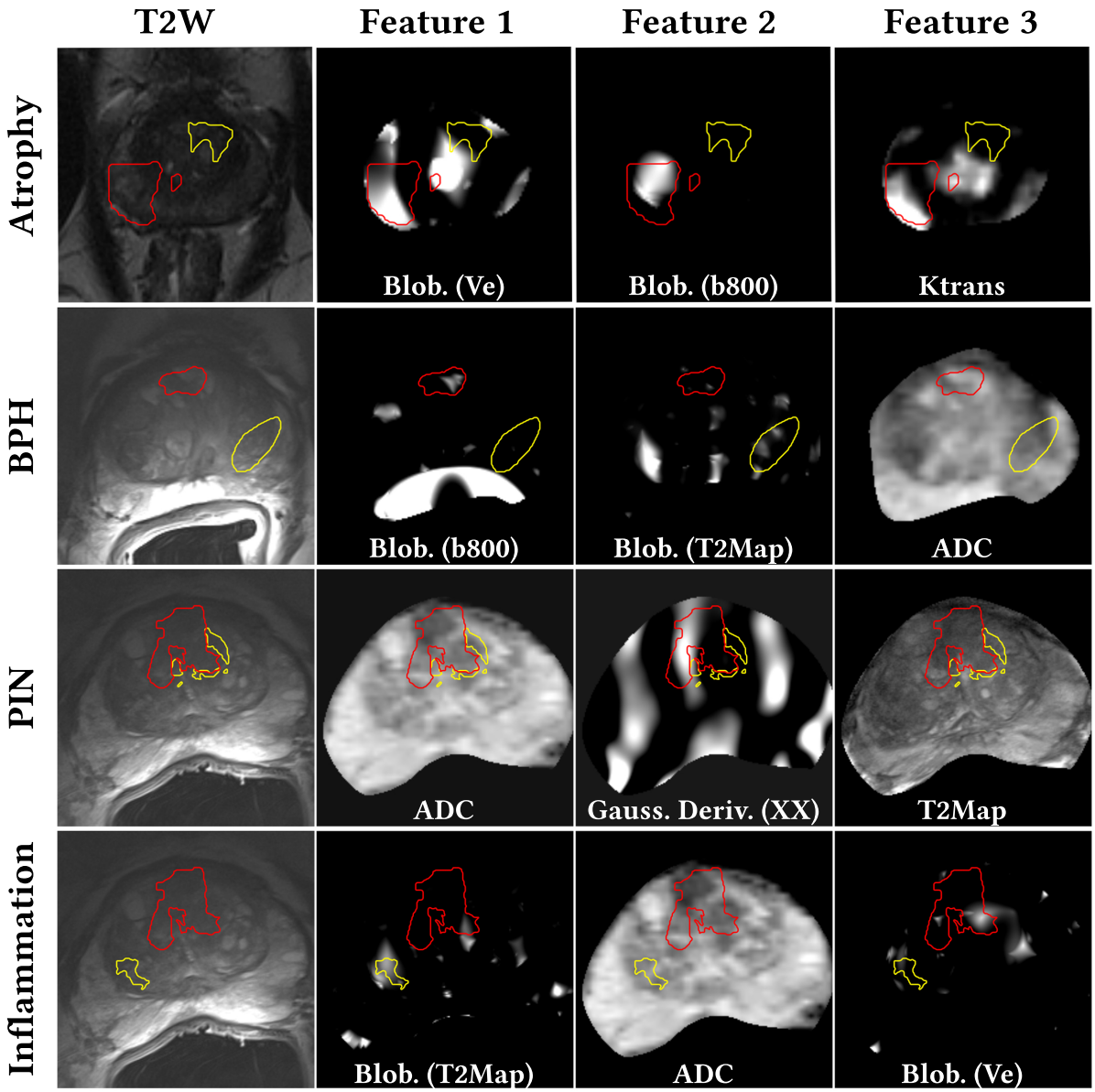



Figure 4.3: Feature maps of the top 3 selected features for each of the classification tasks. Cancer is contoured in red and the specific benign confounder in yellow. The first column contains the axial T2-weighted image as a reference. The first row is atrophy, the second row is BPH, the third row is PIB and last row is inflammation.

# Features to determine cancer grade

5

Geert Litjens, Thomas Hambrock, Christina Hulsbergen - van de Kaa, Jelle Barentsz and Henkjan Huisman

*Original title:* The effect of inter-patient normal peripheral zone Apparent Diffusion Coefficient variation on the Prediction of Prostate Cancer Aggressiveness.

*Published in:* Radiology (2012);265(1);260–266

## 5.1 Introduction

Only 15% of men diagnosed with prostate cancer show a disease specific mortality. The mortality in the US in 2010 was 30000, with 220000 new prostate cancer cases diagnosed[174]. Thus in order to tailor treatment from more radical therapy to active surveillance protocols, accurate cancer aggressiveness risk stratification is very important. The most useful estimator of cancer aggressiveness is the Gleason score (GS), a histopathological scoring system used on biopsy and prostatectomy specimens. It has become such an integral part in prostate cancer evaluation, that patient management is largely influenced by the assessment thereof[175–177].

Recently, the apparent diffusion coefficient (ADC) values determined in diffusion-weighted magnetic resonance imaging (DWI-MRI) showed to be inversely correlated to GS[51,178,179]. As a result, ADC has been proposed as a useful non-invasive biomarker for prostate cancer aggressiveness. However, the discriminative power of ADC depends in part on the variability of the ADC measurement. This variability is machine – i.e. vendor, settings, noise - and patient dependent, the latter caused by natural tissue heterogeneity. Based on the large inter-patient distribution of normal PZ ADC values ($1.2 - 2.2 \times 10^{-3} \frac{mm^2}{s}$) observed on a single MR scanner, we hypothesize that a substantial histo-physiological heterogeneity between patients must exist (inter-patient variation)[51,160].

Inter-patient ADC variation could affect the discriminative power of ADC both for prostate cancer localization as well as for the determination of prostate cancer aggressiveness. Since normal prostate PZ tissue fluctuates significantly in ADC value, the ADC values of an aggressive tumor may show similar fluctuations. Considering normal PZ and tumor ADC simultaneously may lead to better estimates of aggressiveness.

The purpose of this study was to determine the inter-patient variability of prostate peripheral zone (PZ) apparent diffusion coefficient values (ADC) at 3T and the effect this has on the assessment of prostate cancer aggressiveness.

## 5.2 Materials and Methods

### 5.2.1 Patients

Imaging data of two retrospective patient cohorts was used in our experiments. The requirement to obtain institutional review board approval was waived for both cohorts. To determine the significance of the inter-patient variance relative to the measurement variability we included 10 patients (February 2008 to June 2011, interval between scans 6 – 12 months) who had repeated measurements of normal PZ ADC values at three separate MR imaging sessions at 3T. The indication for the studies was continuously high PSA level and at least one negative transrectal ultrasound biopsy. Patients were followed up if PSA level remained high. In these patients no peripheral zone cancer was found in all three imaging sessions by an expert

radiologist (J.O.B., with 18 years of experience). If a suspicious lesion was indicated by the radiologist subsequent MR-guided biopsy found no traces of tumor.

In addition, to determine the effect of the inter-patient variation of ADC on the prediction of prostate cancer aggressiveness a second cohort was included. Between August 2006 and January 2009, 70 consecutive patients with biopsy proven PZ prostate cancer , scheduled for radical prostatectomy, were referred from the departments of urology at the Radboud University Nijmegen Medical Centre and the Canisius Wilhelmina Hospital in Nijmegen, The Netherlands, for clinically routine pre-operative MRI of the prostate.

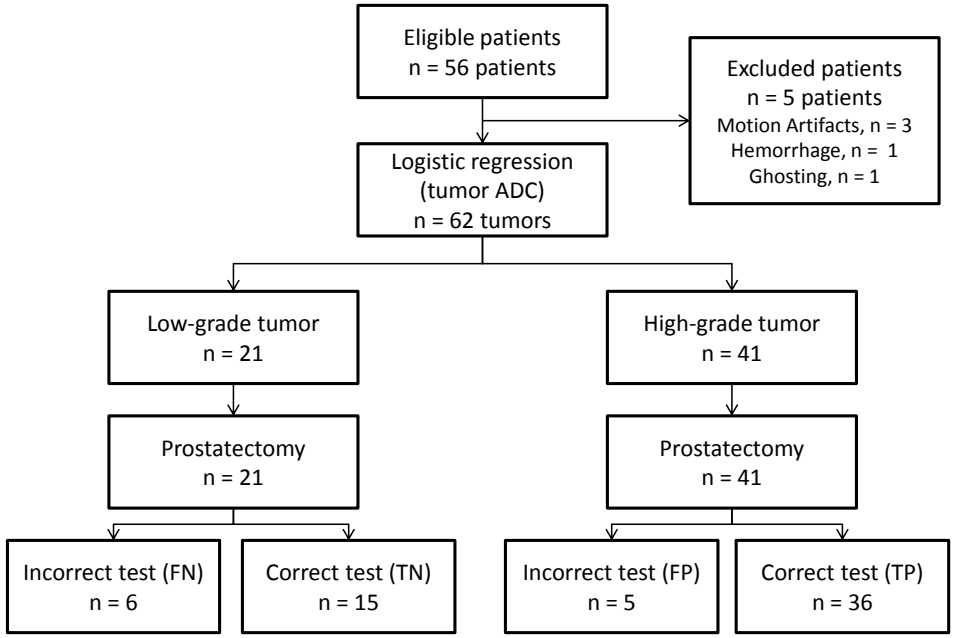
| **Clinical characteristics** | |
|---|---|
| Prostate Specific Antigen (PSA) level (ng/mL) | 6.8 (1.7 – 42) |
| Age (y) | 64 (49 – 69) |
| **Pathologic characteristics (per patient)** | |
| Stage T2a | 5 |
| Stage T2c | 23 |
| Stage T3a | 18 |
| Stage T3b | 4 |
| Stage T4 | 1 |
| **Gleason grade (per tumor)** | |
| 3 + 2 | 3 |
| 3 + 3 | 18 |
| 2 + 4 | 1 |
| 3 + 4 | 13 |
| 3 + 4 + 5 | 4 |
| 4 + 3 | 13 |
| 4 + 3 + 5 | 5 |
| 4 + 4 | 2 |
| 4 + 5 | 3 |

Table 5.1: Summary of clinical and pathologic characteristics for the second cohort of 51 patients.

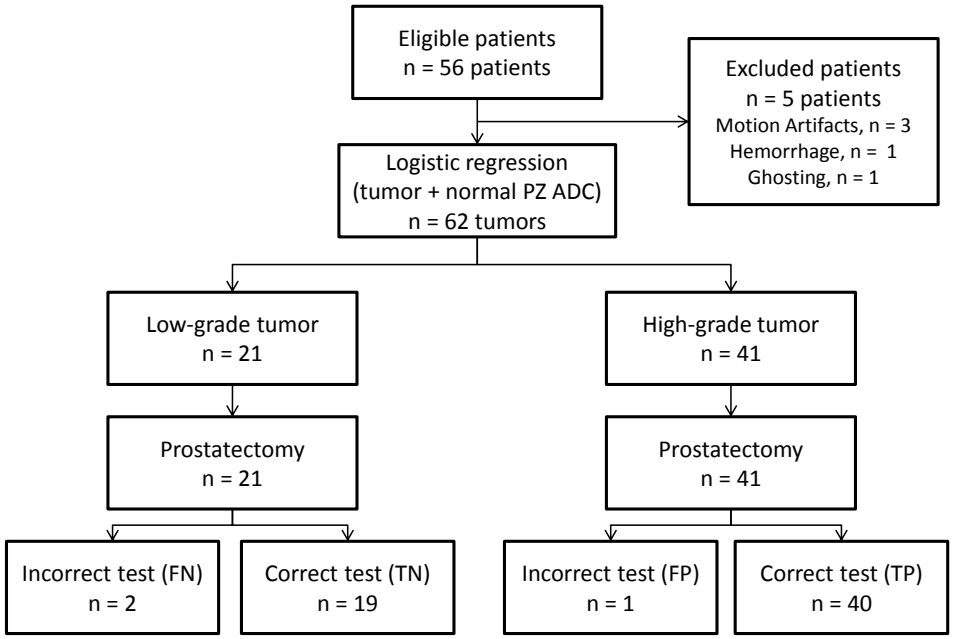
### 5.2.2 MR Imaging Protocol

MR imaging of the prostate was performed using a 3T MR scanner (Siemens Trio Tim, Erlangen, Germany). The first cohort of 10 patients was scanned with only the pelvic phased array coils.

The second cohort was scanned with the use of combined endorectal coil (ERC) (Medrad, Pittsburgh, U.S.A) and pelvic phased array coils. The ERC was filled with a 40-mL Perfluorocarbon preparation (Fomblin, Solvay-Solexis, Milan, Italy)




(a)




(b)

Figure 5.1: Flowcharts expressing the diagnostic accuracy of (a) the method including only tumor ADC and (b) the method incorporating both tumor and normal PZ ADC. FN = false negative, FP = false positive, TN = true negative, TP = true positive.

In both cohorts peristalsis was suppressed with an intramuscular administration of 20-mg Butylscopolaminebromide (Buscopan, Boehringer-Ingelheim, Ingelheim, Germany) and 1 mg of glucagon (Glucagen, Nordisk, Gentofte, Denmark).

The MR imaging protocol included: anatomical T2-weighted turbo spin echo sequences in axial, sagittal and coronal planes covering the entire prostate and seminal vesicles. Axial diffusion weighted imaging was performed using a single-shot-echo-planar imaging sequence with diffusion modules and fat suppression pulses implemented. Water diffusion was measured in 3-scan trace mode using b-values of 0, 50, 500, and 800 $\frac{s}{mm^2}$. ADC-maps were automatically calculated by the scanner software using all b-values. Complete pulse sequence details can be found in Table 5.2 for the first cohort containing 10 patients with repeated measurements and Table 5.3 for the second cohort.

| | PS | ST | NS | SR | RT | ET | AV | Other |
|---|---|---|---|---|---|---|---|---|
| T2W | TSE | 3.5 - 4 mm | 13-19 | 0.6 mm | 3540 ms | 104 ms | 2 -3 | Three orthogonal directions (axial, sagittal, coronal) |
| DWI | SE-EPI | 3.5 - 4 mm | 15-20 | 1.6 – 2.0 mm | 2300 ms | 61 ms | 6 - 10 | b-values used 0, 50, 500, 800 $\frac{mm^2}{s}$ |
| DCE | Turbo FLASH | 3.5 - 4 mm | 14 | 1.8 mm | 37 ms | 1.47 ms | 1 | - |

Table 5.2: Pulse sequence details for the first patient cohort with repeated measurements. In-plane resolution is the same in both directions. PS = pulse sequence, SR = spatial resolution, ST = slice thickness, NS = number of slices, ET = echo time, RT = repetition time, FA = flip angle, AV = number of averages.

| | PS | ST | NS | SR | RT | ET | AV | Other |
|---|---|---|---|---|---|---|---|---|
| T2W | TSE | 4 mm | 15-19 | 0.4 - 0.5 mm | 3540 - 3810 ms | 105 ms | 2 | Three orthogonal directions (axial, sagittal, coronal) |
| DWI | SE-EPI | 4 mm | 15-19 | 2.0 mm | 2800 ms | 81 ms | 10 | b-values used 0, 50, 500, 800 $\frac{mm^2}{s}$ |
| DCE | Turbo FLASH | 4 mm | 14 | 1.8 mm | 37 ms | 1.47 ms | 1 | - |

Table 5.3: Pulse sequence details for the second patient cohort with repeated measurements. In-plane resolution is the same in both directions. PS = pulse sequence, SR = spatial resolution, ST = slice thickness, NS = number of slices, ET = echo time, RT = repetition time, FA = flip angle, AV = number of averages.

### 5.2.3 Whole-Mount Step-Section Preparation

The second cohort of patients underwent radical prostatectomy after imaging. After the radical prostatectomy, prostate specimens were uniformly processed and submitted for histologic

investigation in their entirety. After histologic staining, all specimens were evaluated by one expert urological pathologist (C.A.H.v.d.K. with 17 years of experience). Each individual tumor was graded according to the 2005 International Society of Urological Pathology Modified Gleason Grading System.

Peripheral zone tumors, with a size of >0.5 cc in volume, were divided in two groups, and classified as low- and high-grade tumors. Tumors with a Gleason grade 4 or 5 component were defined as high-grade. Low-grade tumors were defined as tumors harboring only Gleason grades 2 and 3.

### 5.2.4 Annotation of MR images

All annotations were performed using an in-house developed MR viewing and reporting system. In the first cohort the center slice of the prostate in the axial direction was used to annotate the peripheral zone. For this slice the whole peripheral zone was annotated and the median ADC value was extracted from this annotation.

For the second cohort, ADC maps were acquired in the same orientation and of similar thickness as the histopathology step-section. A previously described translation technique was used to match every tumor containing histopathology step-section to a corresponding ADC map[51]. Using histopathology as gold standard, a region of interest (ROI) was placed by one radiologist (T.H. with four years of experience) and one urologist (with one year of experience) in consensus, on the ADC maps. The size and extent of the ROI were chosen such that it matched the tumor size and extent obtained from histological examination as closely as possible. Median ADC values were extracted for each tumor slice separately. In clinical practice, the ADC slice revealing the lowest signal intensity for tumor alerts radiologists. Therefore, for each individual PZ tumor, the tumor slice revealing the lowest ADC values was used for further assessment.

Lastly, to determine the effect of incorporating normal PZ ADC values on the prediction of cancer aggressiveness, an ROI was placed in the normal PZ tissue of every patient. This region was always selected adjacent to the tumor, in order to be the most representative area of normal PZ ADC value at the tumor location. This was done to attempt to minimize intra-patient heterogeneity. Median ADC values were extracted from all ROIs. Median values were used because they are more robust to image artifacts that might occur due to ADC calculation by the scanner.

### 5.2.5 Statistical analysis

Our first hypothesis is that there is a significant degree of inter-patient variation in normal PZ ADC values. This was assessed using a repeated-measures ANOVA. Mauchly's sphericity test will be performed to test the hypothesis of sphericity. If sphericity can not be assumed the

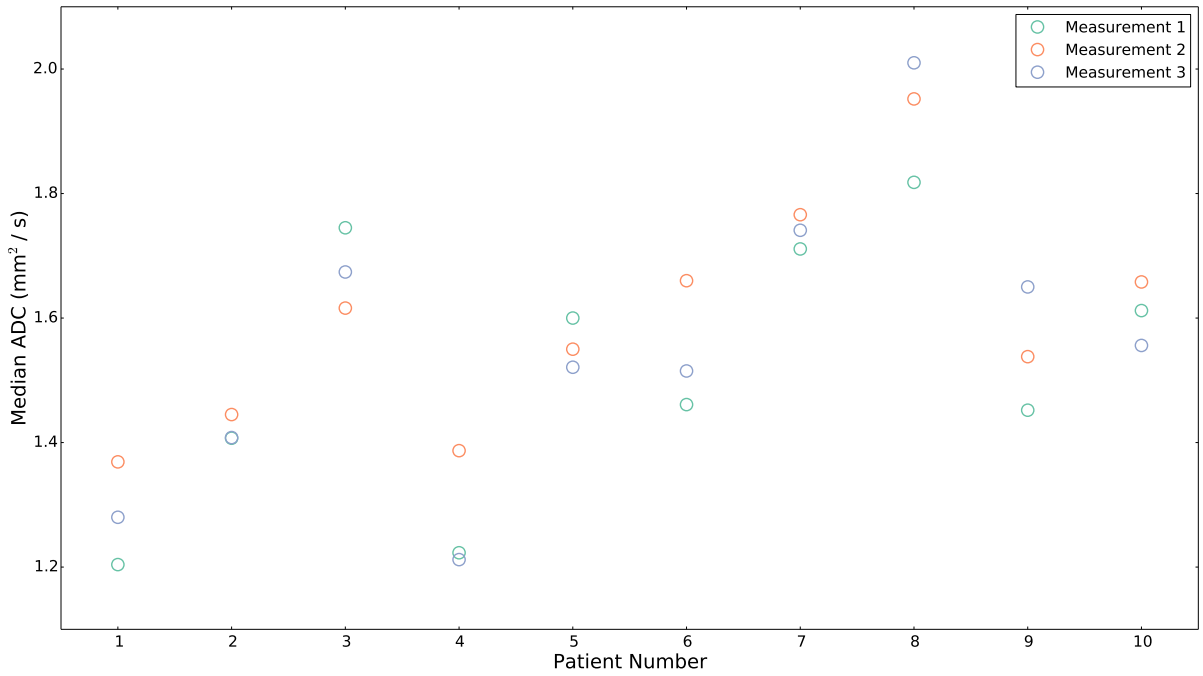



Figure 5.2: Three median ADC measurements of the peripheral zone of 10 patients. The circles represent the individual measurements, the vertical axis shows the median ADC value, the horizontal axis shows to which patient the measurement belongs.

Greenhouse-Geisser corrected p-value will be reported. The repeated measure was the median ADC value of normal PZ tissue, which was obtained three times for each of the 10 patients in the first cohort.

Our second hypothesis is that joint analysis of the normal PZ ADC values and the tumor ADC values will result in an improved prediction of cancer aggressiveness, because this implicitly corrects for the inter-patient variations in normal PZ ADC. Multivariate linear logistic regression was used to test this hypothesis. We can express a regression model of cancer grade as:

$$z = C + \beta_T ADC_T + \beta_N ADC_N \tag{5.1}$$

$$p = \frac{e^z}{1 + e^z} \tag{5.2}$$

The p indicates the probability that a cancer is high-grade and the ADC variables indicate the median ADC of the corresponding ROI. Subscripts T and N are tumor and normal PZ respectively. The beta terms are the regression coefficient corresponding to these variables. Equation 2 represents the conversion from $z$ to the probability p.

The linear logistic regression results in values for $\beta_T$ and $\beta_N$ and the significance of these variables in the regression model. Two regression models were created to compare diagnostic performance: using only tumor ADC values and using tumor and normal PZ ADC values. SPSS (SPSS, version 16.0.01, Chicago, U.S.A.) was used for the statistical analysis. Furthermore, a

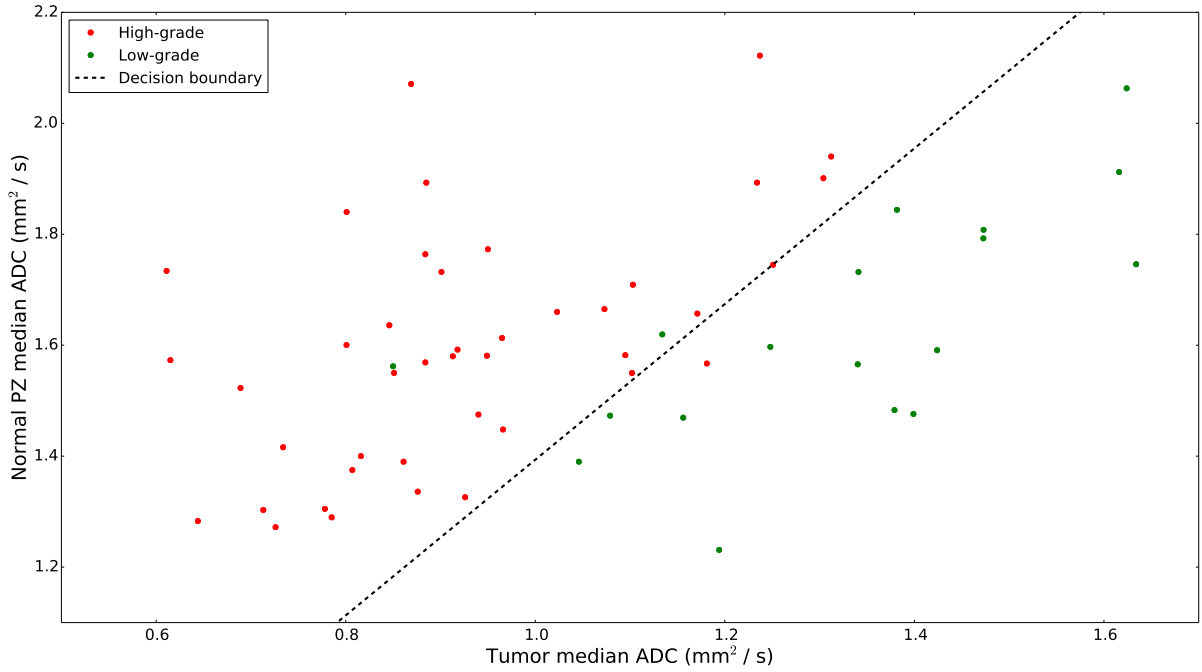



Figure 5.3: Decision Boundary at p=.5 of the logistic regression model. The line represents the decision boundary, the green dots the low-grade cancer and the red dots the high-grade cancers.

visual assessment is given for the correlation between tumor ADC and normal peripheral zone ADC by plotting the low- and high-grade tumors with respect to their ADC values and the corresponding normal PZ ADC values.

Our third hypothesis is that the improved prediction of prostate cancer aggressiveness may result in a significant improvement in diagnostic accuracy in separating low- and high-grade cancer. Receiver-operating characteristic (ROC) curves were constructed for a standalone tumor ADC regression model and the regression model, which incorporates normal PZ ADC values. The areas under the ROC curves were tested for significant differences using the ROCKIT software package (Kurt Rossmann Laboratories, University of Chicago, Chicago).[180]

### 5.2.6 Nomogram construction

Additionally, the regression model incorporating tumor and normal PZ ADC can be used to construct a nomogram by evaluating the obtained equation for a range of ADC values. The ranges used to construct the nomogram are $0.5 - 1.7\times10^{-3}\,\frac{\mathrm{mm}^2}{\mathrm{s}}$ for the tumor ADC value and $0.8 - 2.2\times10^{-3}\,\frac{\mathrm{mm}^2}{\mathrm{s}}$ for the normal PZ ADC value. These ranges are slightly larger than the ranges found in this study to accommodate more extreme values.

## 5.3 Results

For the first cohort of 10 patients, no patients were excluded. The median patient age was 58.5 years (47 – 67). The median PSA at the time of the first MRI was 8.25 (1.8 – 26).

For the second cohort of 70 consecutive patients, 56 patients had clinically significant tumor (>0.5mL). Of the remaining 14 patients 11 had a tumor in the central gland and 3 had a peripheral zone tumor smaller than 0.5 mL. Of the 56 patients 5 patients were excluded due to severe motion artifacts (n=3), hemorrhage (n=1) or ghosting (n=1). Characteristics of these patients are reported in Table 5.1. In these 51 patients a total of 62 different peripheral zone tumors were found. Of these tumors 21 were low-grade tumors and 41 were high grade tumors. The mean ADC for the low-grade tumors was $1.35 \pm 0.26 \times 10^{-3} \frac{mm^2}{s}$ and $0.926 \pm 0.18 \times 10^{-3} \frac{mm^2}{s}$ for the high-grade tumors. The mean value of the normal peripheral zone for patients with a low grade tumor was $1.65 \pm 0.21 \times 10^{-3} \frac{mm^2}{s}$ and $1.60 \pm 0.21 \times 10^{-3} \frac{mm^2}{s}$ for patients with a high grade tumor.

### 5.3.1 Assessment of inter-patient variation of normal PZ ADC values

Normal PZ ADC values differed significantly between patients relative to measurement variability (Mauchly's sphericity test p-value $< 0.0001$, Greenhouse-Geisser corrected p-value = 0.0058) as assessed using the repeated measures ANOVA. The ADC measurements are plotted in Figure 5.2.

### 5.3.2 Effect of including normal PZ ADC values in the prediction of cancer aggressiveness

Normal peripheral zone ADC correlates with ADC of high-grade tumors. Its addition to the regression model results in a significantly improved prediction of aggressiveness (p = 0.013). This was determined using the logistic regression procedure; the results are summarized in Table 5.4.

Both regression models show a significant contribution of the tumor ADC (p = 0.003). Normal PZ ADC values also show a significant contribution to the regression model (p = 0.013). The regression model using standalone tumor ADC values can then be expressed as:

$$z = 10.76 - 9.103 \mathrm{ADC_T} \tag{5.3}$$

and the model combining tumor and normal PZ ADC values can be expressed as:

$$z = 0.126 - 18.82 \mathrm{ADC_T} + 13.43 \mathrm{ADC_N} \tag{5.4}$$

In combination with equation 2 these models result in a probability that a given sample is a high-grade cancer. The model incorporating normal PZ ADC (Eq. 4), together with the data

| | Tumor median ADC | | Tumor and normal PZ median ADC | |
|---|---|---|---|---|
| **Parameter** | **Value** | **p** | **Value** | **p** |
| $ADC_T$ | 9.103 | $< 0.001$ | -18.82 | 0.003 |
| $ADC_N$ | - | - | 13.43 | 0.013 |
| $C$ | 10.76 | $< 0.001$ | 0.126 | 0.978 |

Table 5.4: Result of the linear logistic regression for three regressions based on equation 1 and 2. Regressions performed: using only tumor ADC and using tumor and normal PZ median ADC. The second row shows the values used in each regression. The regression parameters are presented in the bottom three rows, their value and significance respectively for each regression.

used in the regression, is shown in Figure 5.3. This plot indicates that a relatively high tumor ADC value might still constitute a high-grade tumor if the normal PZ ADC is high. In addition one can appreciate that using a static threshold on tumor ADC (a vertical line/contour in Figure 5.3) to determine cancer aggressiveness could result in incorrect diagnosis in some patients.

### 5.3.3 Diagnostic performance of the regression models

Including normal PZ significantly (p = 0.0401) improved diagnostic accuracy. The ROC curves for the regression models in equations 3 and 4 are shown in Figure 5.5. The area under the curve increases from 0.91 to 0.96.

The constructed nomogram is shown in Figure 5.4. This nomogram can be used in a clinical setup to quickly look up the chance that a certain region within the peripheral zone is an aggressive cancer.

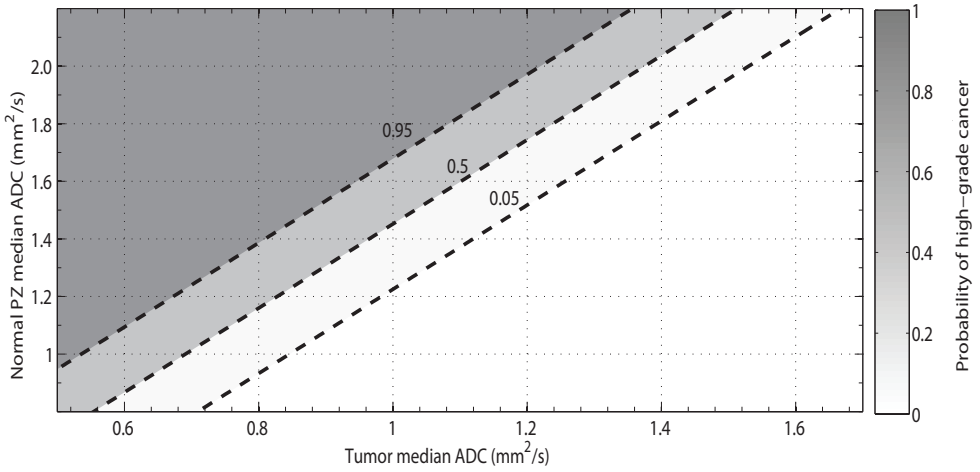



Figure 5.4: Contour of the probabilities of having an aggressive cancer given the adjacent PZ tissue ADC (vertical axis) and the tumor ADC (horizontal axis). The point corresponding to these two values will correspond to the probability of a high-grade cancer. The probability values are specified along the contours and in the color bar on the right of the figure.

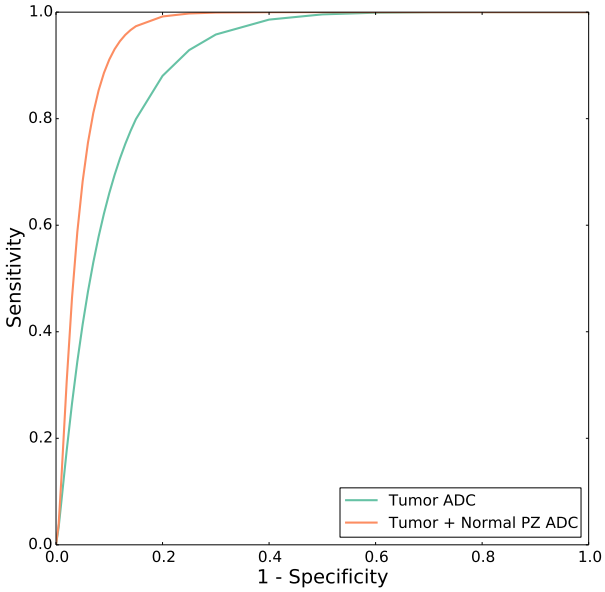



Figure 5.5: ROC curve of the regression models. The red line shows the diagnostic accuracy when including the adjacent PZ tissue median ADC in addition to the tumor ADC, the blue line show the diagnostic accuracy when only using tumor ADC.

## 5.4 Discussion

In this study we have shown that there is a significant inter-patient variation in normal peripheral zone ADC values ($1.2 - 2.0 \times 10^{-3} \frac{mm^2}{s}$), which cannot be solely attributed to measurement variability (average measurement standard deviation $0.068 \pm 0.027 \times 10^{-3} \frac{mm^2}{s}$) . We hypothesize that the inter-patient variations arise from natural variations in prostate physiology.

Secondly, adding normal PZ ADC values to the linear logistic regression, results in a significantly improved prediction of cancer aggressiveness ($p = 0.013$). This suggests that tumor ADC values should not be considered absolute but that these values are influenced by "background" variation of normal PZ tissue composition.

Thirdly, the improvement also results in an increased area under the ROC curve, from 0.91 to 0.96 ($p < 0.05$), thus an improved diagnostic accuracy.

This study has a number of limitations. First, the use of ADC to assess aggressiveness of transition zone tumors has not been investigated in this study. Second, this study was limited to the peripheral zone. This was done because it is known that ADC in peripheral and transition zone tumors can differ substantially. However, the majority of prostate tumors arise in the PZ. Third, the annotation of ROIs was performed by a single observer; the effect of the inter-observer variability on the regression model was not assessed. Our nomogram must be tested and validated in a prospective multi-reader study.

In conclusion, there is a large inter-patient variation in prostate peripheral zone ADC values. This variation propagates into tumor ADC values. Compensating for this variation by

combining tumor and normal PZ ADC when assessing cancer grade significantly increases diagnostic performance.

# CAD system

# Development of a computer-aided detection system for prostate cancer in MRI

6

Geert Litjens, Oscar Debats, Jelle Barentsz, Nico Karssemeijer, Henkjan Huisman

*Original title:* Computer-aided detection of prostate cancer in MRI

*Published in:* Transactions on Medical Imaging (2014);33(5);1083–1092

## 6.1 Introduction

Prostate cancer is one of the major causes of cancer death for men in the Western world[181]. Due to the increased ageing of the general population the incidence of prostate cancer is steadily rising[181]. Current clinical practice for the diagnosis of prostate cancer is to perform a transrectal ultrasound (TRUS) biopsy, which usually is performed due to a positive prostate specific antigen (PSA) blood test. A large screening trial using PSA and TRUS has shown that is is possible to reduce prostate cancer mortality by 20-30%[21]. However, these studies have also shown that PSA testing in combination with TRUS biopsies has a relatively low specificity. Additionally, cancers are often undergraded in TRUS biopsies[28]. These problems lead to overdiagnosis and overtreatment of patients and are prohibiting screening for prostate cancer.

MRI is increasingly used to diagnose prostate cancer as it has improved sensitivity and specificity over PSA and TRUS[32]. Currently, MRI is most often used as a second-line modality after repeat negative TRUS biopsies. One of the reasons MRI has not yet progressed to a first line modality for prostate cancer diagnosis is that it requires substantial expertise from the radiologist to read prostate MRI and such expertise is not widely available. Additionally, due to the large amount of 3D images, reading prostate MR is quite time consuming.

Automated computer-aided detection and diagnosis (CAD) of prostate cancer could help reduce both of these problems and open the door to prostate cancer screening using MRI. In the past several other areas have seen successful CAD applications, such as mammography[182], CT colonography[183] and retinopathy[184]. In the last decade several researchers have investigated the use of these techniques for prostate MRI. Therefore, computer-aided detection and diagnosis of prostate cancer is becoming an active field of research[88,89,93,96].

Chan et al.[88] were the first to implement a multi-parametric CAD system for the diagnosis of prostate cancer. In their approach they used line-scan diffusion, T2 and T2-weighted images in combination with an SVM classifier to identify predefined areas of the peripheral zone of the prostate for the presence of prostate cancer. Langer et al.[89] included dynamic-contrast enhanced images and pharmacokinetic parameter maps as extra features to a CAD system for prostate peripheral zone cancer. They evaluated their system in predefined regions of interest, but on a per-voxel basis. Tiwari et al.[93] investigated the use of magnetic resonance spectroscopy in combination with T2-weighted imaging to identify the voxels that are affected by prostate cancer. They also introduced the use of wavelet embedding to map MRS and T2-W texture features into a common space. This work was further expanded and evaluated in[94]. Niaf et al.[185] presented the use of computer-aided diagnosis in the peripheral zone of the prostate (similar to Vos et al.[57]. They confirmed the results in discriminating prostate cancer from normal regions (area under the ROC curve (AUC)=0.89) and discriminating prostate cancer from suspicious benign regions (AUC of 0.82). This is a limited CAD method as it is constrained to predefined regions of interest in only the peripheral zone. Firjani et al.[186] in-

vestigated the use of computer-aided detection in single-parameter MRI using DWI imaging with transrectal ultrasound guided biopsies as ground truth. They included registration of different b-values to obtain a less motion sensitive apparent diffusion coefficient map. Lastly, Vos et al.[96] recently implemented a two-stage computer-aided detection system for prostate cancer using an initial blob detection approach combined with a candidate segmentation and classification using statistical region features.

In this paper we investigate a fully automated computer-aided detection system including a novel combination of segmentation, voxel classification, candidate extraction and candidate classification, which expands on the work published in[187]. Other novel aspects include a voxel classification stage in combination with a candidate classification stage and inclusion of symmetry, local contrast and anatomical features like peripheral zone likelihood. Feature design was based on the standardized guidelines for reading prostate MR, PI-RADS, and include texture, pharmacokinetic, shape and anatomy, among others[33]. Furthermore, to the best of the authors' knowledge this is the first prostate MRI CAD system that is evaluated on a per-patient basis and compared to the prospective performance of radiologists. The system was validated on a large cohort of 347 patients using per-region FROC and per-patient ROC to show the value of a two stage approach incorporating both voxel and candidate classification.

## 6.2 Materials and Methods

### 6.2.1 MRI data

In our hospital we collected a total of 165 consecutive studies with prostate cancer (187 lesions) and 183 cases without prostate cancer on which to evaluate our CAD-system for a total of 348 studies of 347 patients.

Each MR study was read and reported by or under the supervision of an expert radiologist, J.B. with more than 20 years of experience in prostate MR. The radiologist indicated areas of suspicion with a score per modality using a point marker. If an area was considered likely for cancer a biopsy was performed. All biopsies were performed under MR-guidance and confirmation scans of the biopsy needle in situ were made to confirm accurate localization. Biopsy specimen were subsequently graded by a pathologist and these results were used as ground truth.

All studies included T2-weighted (T2W), proton density-weighted (PD-W), dynamic contrast enhanced (DCE) and diffusion-weighted (DW) imaging. It is currently established clinical consensus that prostate cancer should be diagnosed by T2-weighted imaging with at least two functional modalities (from DWI, DCE and spectroscopic imaging)[32,33]. The images were acquired on two different types of Siemens 3T MR scanners, the MAGNETOM Trio and Skyra. T2-weighted images were acquired using a turbo spin echo sequence and had a resolution of

around 0.5 mm in plane and a slice thickness of 3.6 mm. The DCE time series was acquired using a 3D turbo flash gradient echo sequence with a resolution of around 1.5 mm in-plane, a slice thickness of 4 mm and a temporal resolution of 3.5 seconds. The proton density weighted image was acquired prior to the DCE time series using the same sequence with different echo and repetition times and a different flip angle. Finally, the DWI series were acquired with a single-shot echo planar imaging sequence with a resolution of 2 mm in-plane and 3.6 mm slice thickness and with diffusion-encoding gradients in three directions. Three b-values were acquired (50, 400 and 800), and subsequently, the ADC map was calculated by the scanner software. All images were acquired without an endorectal coil, as per the PI-RADS guidelines for acquisition of prostate MRI[33]. Although an endorectal coil would allow for further improved resolution of the images, the added value is considered negligible compared to added patient comfort when only using a pelvic phased array coil. The transversal T2W image, the PD-W image, the entire DCE time series and from the DWI the apparent diffusion coefficient map and the b800-image were used in this study. These images were chosen because they are explicitly incorporated in the PI-RADS standard, except the PD-W image, which was added purely for feature calculation purposes.

### 6.2.2 Overview of the CAD pipeline

The pipeline of the CAD system is visualized schematically in figure 6.1 and follows a two stage approach.

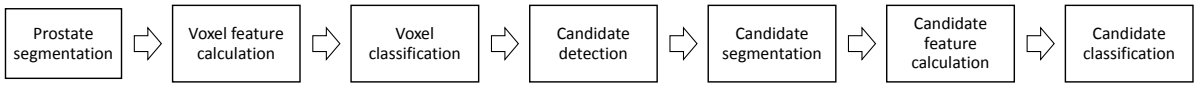



Figure 6.1: Flowchart showing the different steps of the computer-aided detection system

The first (detection) stage consists of segmentation of the prostate on the transversal T2-weighted image, extraction of voxel features from the image volumes, classification of the voxels and candidate selection. The second (diagnosis) stage consists of candidate segmentation, candidate feature extraction and candidate classification. Each of the steps will be described in more detail in the corresponding subsections.

### 6.2.3 Segmentation

The segmentation of the prostate is required to reduce the complexity of the detection task for the classifiers in the later stages. In our system we use an atlas based segmentation approach similar to the one presented in[108], using the atlas selection mechanism presented in[146], named SIMPLE (Selective and Iterative Method for Performance Level Estimation).

Let the image to be segmented be denoted as $I(x)$, where $x$ is a spatial location within the image. A labeled image, $L(x)$ has to be determined. The following steps are similar in most

multi-atlas based systems. A set of N manually labeled images $A$ is non-rigidly registered to the unknown case $I(x)$. The i-th atlas in this set $A$ is denoted as $A_i(x)$, where each $A_i(x)$ consists of the atlas image and the label image: $A_i(x) = \{I_i(x), L_i(x)\}$. After registration of the atlas image $I_i(x)$ the obtained transformation $T_i(x)$ is applied to the label image of the atlas, $L_i(x)$.

An important part of the registration procedure is the similarity metric that is used. In this approach we used the localized mutual information metric[188]. After atlas registration we have a transformed label image $L_i \circ T_i(x)$ for each atlas. We use the SIMPLE method, presented by Langerak et al[146], to combine these label images into one final segmentation. Example results are shown in figure 6.2. This algorithm competed in the prostate MR image segmentation (PROMISE12) challenge (`http://promise12.grand-challenge.org`), where it obtained a 9th place out of 12. Overall segmentation results were still reasonably good, with a median Dice's coefficient of 0.83.

### 6.2.4 Voxel features

After prostate segmentation we calculated voxel features from the image volumes. The types of features can be categorized in intensity, pharmacokinetic, texture, blobness and anatomical features. A complete overview of the voxel features is given in table 6.1, implementation details are given in the corresponding subsections.

*Intensity*

One of the major issues in image analysis for MRI is the absence of a standardized signal intensity, like Hounsfield units in CT. This usually means that an algorithm will give different results as scanners, sequences or even sequence parameters are changed. To mitigate this issue we developed several algorithms to standardize signal intensity in the different MR modalities. First, for the T2-weighted imaging a T2-estimate map is generated by using the MR signal equation, the proton density image and a reference tissue. This process was automated and is explained in more detail in[169]. This map was added as a voxel feature in addition to the original transversal T2-weighted image. Second, the MR scanner software automatically calculates the apparent diffusion coefficient map from the diffusion-weighted images, by fitting a mono-exponential function to the signal decay across the different b-values. Furthermore, studies have shown that the highest b-value image has additional diagnostic value, therefore, the b800 image was also added as a feature.

*Anatomical*

For the anatomical features we used the relative distance to the prostate boundary:

| Name | Type | Description |
|---|---|---|
| T2W | Intensity | T2-weighted voxel grey value, related to voxel T2 |
| ADC | Intensity | Apparent diffusion coefficient, measure for cellular density |
| b800 | Intensity | High b-value image, areas with low diffusivity appear bright |
| T2-map | Intensity | T2-map based on proton density and transversal T2W image [169] |
| x-pos | Anatomical | Relative cumulative position within the prostate in the x-direction |
| y-pos | Anatomical | Relative cumulative position within the prostate in the y-direction |
| z-pos | Anatomical | Relative cumulative position within the prostate in the z-direction |
| Distance | Anatomical | Relative distance to the prostate boundary between 0 and 1 |
| PZ likelihood | Anatomical | Likelihood of being a peripheral zone voxel between 0 and 1 [169] |
| $K^{trans}$ | Pharmacokinetic | Parameter related to vessel permeability |
| $k_{ep}$ | Pharmacokinetic | Parameter related to permeability and extracellular volume |
| tau | Pharmacokinetic | Time-to-peak of contrast agent concentration |
| LateWash | Pharmacokinetic | Curve parameter related to the washout of contrast agent |
| Gaussian texture bank | Texture | Calculate multi-scale Gaussian derivatives on the T2W image |
| ADC Blob | Blobness | Multi-scale blob detection using the Li blobness filter [170] |
| $K^{trans}$Blob | Blobness | Multi-scale blob detection using the Li blobness filter [170] |
| LateWash Blob | Blobness | Multi-scale blob detection using the Li blobness filter [170] |
| tau Blob | Blobness | Multi-scale blob detection using the Li blobness filter [170] |

Table 6.1: Overview of voxel features used in the CAD system.

$$B(x) = \min_{y \in P_b} d(x, y) \tag{6.1}$$

$$RD(x) = \frac{B(x)}{\max_{x \in P} B(x)} \tag{6.2}$$

where $x$ is the position of a voxel, $d$ is the Euclidean distance operator, $P$ is the set of prostate voxels and $P_b$ the set of prostate boundary voxels.

Additionally, we also use relative position features in x, y, and z directions. The relative position features are defined as:

$$RP(x_i) = \frac{x_i - \min_{x \in P} x_i}{\max_{x \in P} x_i - \min_{x \in P} x_i} \tag{6.3}$$

where $x$ is the position of a voxel and $i$ is the image axis, either x, y or z, and $P$ the set of prostate voxels. Both the relative distance and the relative position features are calculated with respect to the prostate mask obtained through the multi-atlas method. Finally, we also implemented

a peripheral zone probability feature, which gives a likelihood per voxel that it belongs to the peripheral zone. This feature uses a pattern recognition framework incorporating intensity, texture and anatomical features. This results in a likelihood for each voxel within a prostate mask of belonging to either the peripheral zone or the central gland. More implementation details of this filter can be found in [169]. This feature is important because we know from clinical practice that prostate cancer appearance can differ substantially between the peripheral zone and the central gland [33].

*Pharmacokinetic*

In the clinic it is common practice to use the DCE time curve to diagnose prostate cancer [33]. The approach used is described as the curve type method, were the radiologist looks at the curve and assesses two characteristics based on the first, last, and peak enhancement time points [33]. These characteristics are whether there is fast initial enhancement and if there is persistent enhancement, an enhancement plateau or wash-out. Slow initial, persistent enhancement (curve type I) or slow initial, constant enhancement (curve type II) are associated with normal and benign findings whereas fast initial enhancement combined with washout are indicative for malignancy (curve type III).

The traditional analysis is incorporated in our CAD system by using a curve-fitting technique to fit, per voxel, a bi-exponential curve to the time data, as presented in [55]. Of these curve parameters we incorporated the parameter tau (which corresponds to time-to-peak of the enhancement curve) and the parameter LateWash (which corresponds to the slope of the last part of the curve). There are two major problems with only using this type of analysis. First, the assessment of the curve is scanner and patient dependent, e.g. different protocols or patient anxiety (which increases blood flow). Second, not all information present in the curve is used. To counter these disadvantages pharmacokinetic modeling of the contrast agent concentrations has been proposed and applied in breast, brain and prostate MRI [57,189]. We implemented the standard Tofts pharmacokinetic model [35] including an automated reference tissue method to estimate the arterial input function, as proposed in [55,57]. This model provides us with three parameter maps for the DCE time series. The parameters represent the permeability of the micro-vasculature, $K^{trans}$, the fraction of extracellular, extravascular space, $v_e$ and the quotient of the two, $k_{ep}$. Due to fast and sloppy vessel construction and tightly packed cells in a cancerous region it is expected that $K^{trans}$ and $v_e$ will differ between cancerous and normal/benign tissue.

*Texture and blobness*

Most cancers show textural distortions in T2-weighted images [33,95]. To capture these characteristics in features we use a Gaussian texture bank. For the Gaussian feature bank we used

5 different scales, from 2 mm to 8 mm exponentially and derivatives up to the second order. This scale range was selected to encompass the typical size ranges of lesions in prostate MRI (between 5mm and 20mm in diameter)[190–192]. Due to the large slice thickness the features were calculated on a slice-by-slice basis. This results in a total of 30 Gaussian texture features.

Prostate cancer tends to appear as a focal, blob-like lesion in diffusion-weighted and dynamic contrast enhanced MRI. This characteristic has been previously used to detect prostate cancer[96]. There are many different blobness measures, we chose to incorporate the blobness-filter presented by Li et al.[170] because this filter incorporates both a shape term and a blobness strength term. The blobness feature was calculated with scales ranging from 2 mm to 8 mm, with 5 different, exponentially increasing scales. Again, this range encompasses the size of most lesions encountered in prostate MRI[190–192]. The maximum (bright blobs) or minimum (dark blobs) value of the blobness output across scales at each voxel was used as the final blobness measure. Blobness was calculated on the ADC, tau and LateWash images (dark blobs) and on the Ktrans and Kep images (bright blobs).

### 6.2.5 Voxel classification

After feature calculation a voxel classification is performed, which results in a likelihood between 0 and 1 per voxel, 0 indicating no suspicion of prostate cancer and 1 indicating very high suspicion of prostate cancer. In this step we experimented with three different classifiers, a linear discriminant classifier, a GentleBoost classifier[73] (with regression stumps as weak learners) and a RandomForest-classifier with regression trees[74]. Both the GentleBoost and the RandomForest classifiers are very robust to over-training[73,74], thus parameter optimization is usually not needed[93]. Therefore, both the RandomForest and the GentleBoost classifier were left at the default settings. For the RandomForest the default settings are that a minimum of 0.1 percent of all samples in the dataset is required to split a tree node, the square root of the number of features is used as the number of active variables at each node and the maximum tree depth was equal to the number of features. For both classifiers the number of weak learners has to be set, where, as explained in[73] and[74], adding more weak learners does not result in over-fitting, but produces a limiting value of the generalization error. We did a small pilot experiment using two fold cross-validation to roughly determine the amount of weak learners needed to achieve the minimal generalization error. This resulted in around 100 regression stumps for the GentleBoost classifier and 300 trees for the RandomForest classifier.

We compared the performance of the different classifiers using ROC-analysis. The output of the classifier with the highest area under the ROC curve was used for further analysis. An example of a obtained likelihood map is shown in figure 6.2.

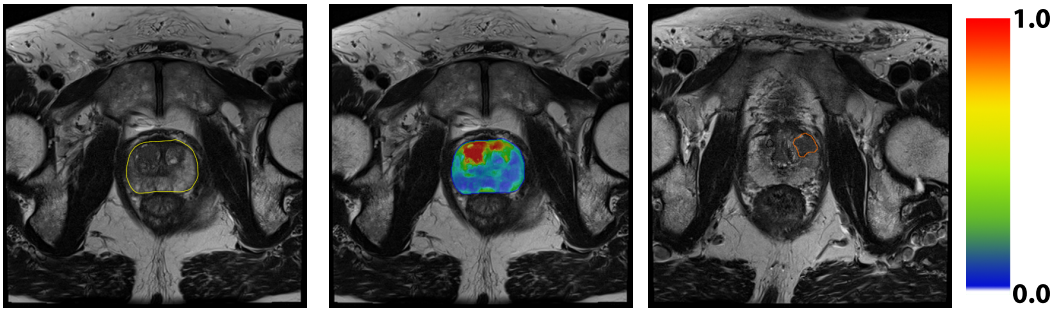



Figure 6.2: Example results for the CAD system for different patients. Figure (a) shows a prostate segmentation result, Figure (b) a likelihoodmap after voxel classification and Figure (c) a region classification result.

### 6.2.6 Candidate selection

After voxel classification a likelihood map is obtained, indicating per voxel the likelihood that it contains cancer. On this likelihood-map we perform local maxima detection using a spherical window with a diameter of 10 mm, which is about the average lesion size in prostate MRI[190–192]. After initial local maxima detection the local maxima which are less than 10 mm apart are merged. This merging step leaves only the local maximum with the highest probability within the 10 mm range. This is iterated until no more merging occurs.

### 6.2.7 Candidate segmentation

For each of the local maxima obtained in the previous step, a region segmentation will be performed. The SmartOpening-algorithm, which has had successful applications in both nodule segmentation in CT and cancer segmentation in breast MRI, was used[193,194]. The segmentation was performed on the likelihood-map itself instead of one of the original clinical images. The main reason for this is that lesions can show slight deviations in size and even position between the different diagnostic images. The likelihood map is essentially a combination of all original images through a classification step and thus should give a good approximation of the lesion extent across all diagnostic images. After initial segmentation, regions which overlapped for more than 50% were merged.

### 6.2.8 Candidate features

After candidate segmentation and merging new candidate features can be calculated given the original feature images and the candidate segmentation. These can be categorized as: statistical (voxel feature statistics), local contrast, symmetry and shape features.

*Statistical*

Statistical candidate features are calculated within the candidate segmentation. Statistics include mean, standard deviation and histogram percentiles. On all the initial voxel features we

calculate the mean and standard deviation of the feature values within the candidate segmentation. Additionally, for the ADC and pharmacokinetic features we calculate either the 25th or 75th percentiles, depending on whether low or high values are indicative of malignancy. The percentiles are calculated because 60% of all tumors are heterogeneous[191], with a more aggressive hot spot within the tumor that for example has lower ADC values. In addition, we also calculated the mean, standard deviation and 75th percentile of the voxel likelihood.

*Local Contrast*

In previous work we have shown that relating tumor feature values to those of surrounding normal tissue can lead to improved characterization of tumor aggressiveness[195]. We incorporate this knowledge into our CAD system by using local contrast candidate features. The local contrast feature is calculated by dilating the original segmentation and then subtracting the original to obtain a rim of tissue outside the candidate. The local contrast is then obtained by taking the quotient of the average candidate and the average rim intensities. We use a 2D kernel with a size of 3 mm for dilation. The local contrast feature is calculated on the b800, ADC, Ktrans, Kep, tau and LateWash voxel feature maps. Additionally, it was also calculated on the voxel likelihood map.

*Symmetry*

A normal prostate has a distinct symmetric appearance in the transversal plane. Radiologist have reported that symmetry in prostate MRI can be important to detect prostate cancer[33]. The CAD system incorporates this knowledge by including a symmetry feature. We take the relative position of a candidate along the x-axis in the transversal plane and mirror it to the other side of the prostate (e.g. if the relative position is 0.25 we map the mirrored candidate segmentation to a relative position of 0.75). Then we calculate the mean intensity value for both the mirrored and the original candidate segmentation and take the quotient. The result is used as the symmetry feature. We calculate this symmetry feature on the b800, ADC, Ktrans, Kep, tau, LateWash voxel feature maps and the voxel likelihood map.

*Shape*

The last candidate feature set are the shape features. Prostate lesions tend to be somewhat spherical and compact. During initial stages of development, most false positives we encountered were due to small segmentation errors , large non-spherical areas of low ADC due to extensive benign prostatic hyperplasia and small artifacts caused by the scanner. By incorporating shape features like volume, sphericity and compactness the classifier can easily remove these false positives from the data. The sphericity is calculated as the ratio of the volume of a sphere having the same diameter as the maximum bounding box length of the candidate

segmentation and the total volume of the candidate segmentation. The compactness is calculated as the candidate segmentation volume divided by the volume of the bounding box of the candidate segmentation.

### 6.2.9 Candidate classification

After candidate feature extraction the final classification is performed. Three different classifiers were tried to obtain the best possible performance, a linear discriminant classifier, a GentleBoost classifier (with regression stumps as weak learners) and a RandomForest classifier with regression trees. The settings we used at this stage were the same as in the voxel classification stage. After classification we obtain a likelihood between 0 and 1 per candidate, 0 indicating no prostate cancer and 1 indicating definite prostate cancer. Examples of a final candidate result can be seen in figure 6.2

### 6.2.10 Relative feature and MR sequence importance in voxel and candidate classification

To establish the importance of individual features and MR sequences to the overall classification results, we performed two experiments at both the voxel and candidate levels. First, using the selected classifiers, we established the classification performance of each feature individually based on area under the ROC curve (using leave-one-patient-out crossvalidation). Subsequently, we repeated this experiment on a per-sequence basis, i.e. only include features calculated using one MR sequence, for example only using T2-texture features or only using DWI features.

### 6.2.11 Validation

*Training data*

For the voxel classification stage voxels in a 10 mm area around the radiologist annotation were extracted as prostate cancer samples. This area was truncated by the prostate mask, to ensure no voxels outside the prostate were included in the training set. Furthermore, we only selected voxels which had specific feature characteristics: the ADC value had to be below the median of the area and the pharmacokinetic features had to be above the median of the area. We know from clinical experience and literature that these are usually good characteristics of prostate cancer and reduces the chance of sampling normal voxels into the malignant class. For the normal class we randomly sample within the prostate mask of normal patients. The resultant voxel dataset is used to train the voxel classifiers.

In the candidate classification stage we extract candidate features from the initially detected true positives and the false positives in normal patients after initial classification. The definition of true and false positives is given in the next section.

*FROC analysis*

The detection performance of the CAD system is evaluated using free-response receiver operating characteristic curve (FROC) analysis. FROC analysis provides the number of false positives per normal patient for a given sensitivity (i.e. the percentage of cancer detected). The occurrence of false positives in normal patients is one of the most relevant problems in prostate cancer diagnosis on MRI as each false positive in a normal patient has the potential to lead to an unnecessary biopsy, and thus patient morbidity and healthcare cost. As such, the number of false positives should be as low as possible at reasonable sensitivity. FROC analysis can be used both after the initial and final stage, which also allows us to assess the performance gained by the second stage of the CAD system. For evaluation of the first stage the criterion for a true positive is that a local maximum should be within 10 mm of the marker annotated by the radiologist. 10 mm corresponds to the average lesions size[190–192]. For the final classification a true positive is defined as a candidate segmentation which has a center of gravity within 10 mm of the marker. Each candidate segmentation is only allowed to correspond to one annotation. This rule is chosen to make sure the system does not have a bias toward large segmentations, i.e. a candidate segmentation covering the entire prostate would cover all lesions, but would generally not result in an accurate localization. We evaluated the system both for the detection of all tumors and the detection of high-grade tumors (first or secondary Gleason component $> 3$). In the second setting a hit on a low-grade tumor is not considered a false positive, the reasoning for this is that in principle low-grade prostate cancer will not require treatment, but it is not detrimental for the patient to detect it.

*ROC analysis and comparison to the radiologists*

In addition to FROC analysis we also performed patient based receiver operating characteristic (ROC) analysis both after the initial voxel classification and after the candidate classification. This is relevant evaluation in a screening setting, where the first thing a clinician wants to know is whether a patient has cancer or not (i.e. the localization aspect captured by the FROC analysis in the previous section is of secondary importance). A CAD system could play a role here by separating out the easy from the difficult to diagnose patients, which could improve the efficiency of the radiologist. In each patient, the voxel (voxel stage) or candidate (candidate stage) with the highest likelihood is used as the patient score, both for patients with prostate cancer and normal patients. In this setup the CAD system can stratify patient as requiring a biopsy or not requiring a biopsy.

Additionally, we compare the system to the overall radiologist performance on this data set. In total 10 radiologists read cases in our patient cohort, each case was read prospectively by one radiologist. Therefore, we can compare the system performance to the actual prospective clinical performance in our hospital.

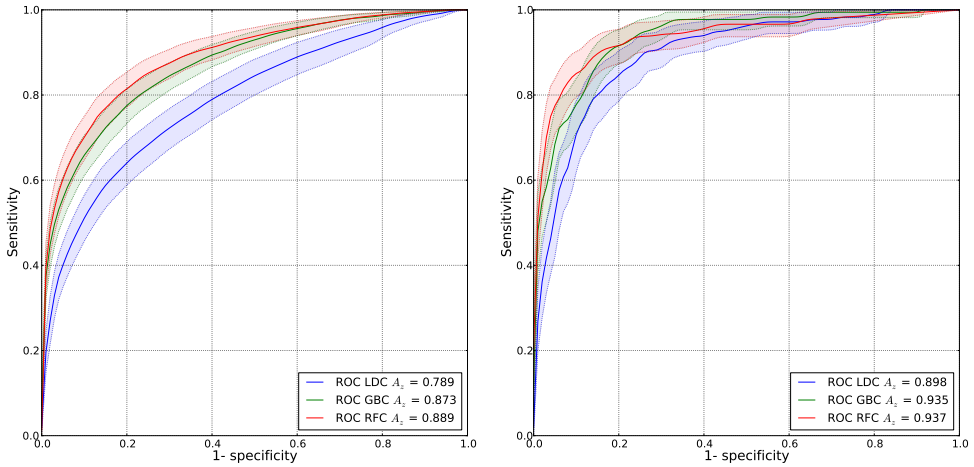



Figure 6.3: Classifier comparison using leave-one-patient-out ROC analysis at the voxel and candidate levels. 95% confidence intervals estimated using bootstrapping are shown as transparent areas around the mean curves. Input for the comparison at the voxel level are voxel in a cancerous areas and voxels in the prostate of normal patients. Input for the comparison at the candidate level is true and false positives after initial voxel classification. Figure a shows the voxel classification results and Figure b the region classification results.

## 6.3 Results

### 6.3.1 Classifier comparison and selection

Bootstrapping and ROC analysis were used to compare classifiers for both CAD stages: the voxel classification stage and the candidate classification stage. For both stages we performed a leave-one-patient-out cross-validation on the training data. Results are shown in figure 6.3. Statistical significance testing was performed using the area under the ROC curve. Both the RandomForest and the Gentleboost classifier performed significantly better than the linear discriminant classifier in both stages ($p < 0.001$). For the voxel classification stage the random forest classifier also performed significantly better than the Gentleboost classifier ($p < 0.01$). Further analysis of the system was performed using the RandomForest classifier for the voxel stage. Although the mean area under curve was higher for the RandomForest classifier than the Gentleboost classifier in the candidate stage, this was not significantly different. Because the mean area under the curve was slightly higher we chose to use the RandomForest classifier for the candidate stage.

### 6.3.2 Relative feature and MR sequence importance

The rankings for each feature and each MR sequence based on there individual classification performance using the RFC classifer are shown in Tables 6.2 and 6.3. Although the ADC intensity is the single most important feature in the voxel stage, overall the features calculated using the T2W MR sequence are the most important in voxel classification. Additionally, we can see from the performance of individual features in the voxel stage that features from each

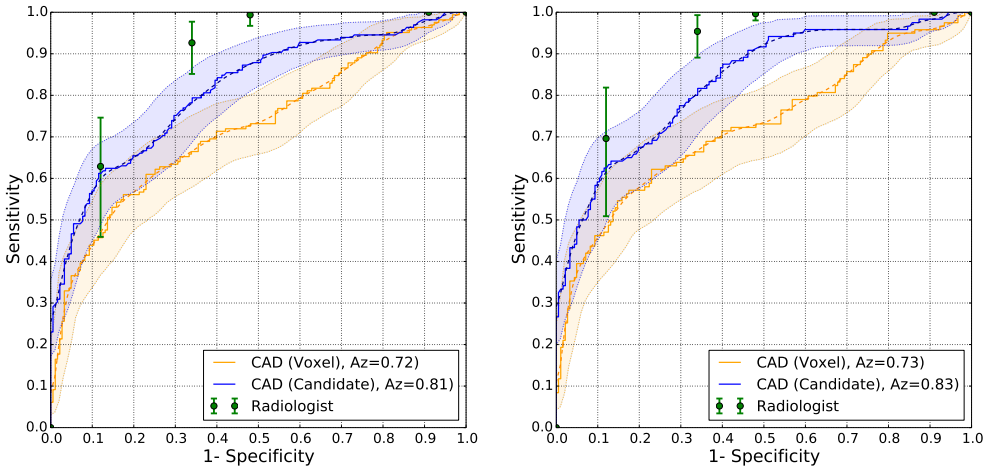



Figure 6.4: ROC analysis on a per-patient level, comparison of the CAD system after the voxel stage and after the candidate stage to the prospective radiologist performance. The raw ROC curve is shown as the solid line and the mean bootstrapped curve as a dashed line. The 95% confidence intervals obtained using bootstrapping are shown as transparent areas around the mean bootstrapped curve. The radiologist ROC curve and confidence intervals are only plotted for the four PIRADS thresholds. Figure (a) shows the results for cancer versus normal/benign and Figure (b) shows the results for high-grade cancer versus normal/benign.

of the MR sequences are selected, showing the importance of using multi-parametric MRI over single-parameter MRI. Finally, the performance per modality is much lower than the overall per-voxel performance (0.76 area under the ROC curve for just T2W and 0.89 when combining all MR sequences).

Inspecting Table 6.3, it is interesting to see that especially heterogeneity of feature values within the candidates have high individual performance. Additionally, in the candidate stage, DWI is by far the best performing individual sequence. Furthermore, the initial voxel likelihood plays an important part in classification in the candidate stage. Finally, in this stage features from the DWI imaging are almost as good as using features from the combination of the three MR sequences.

### 6.3.3 FROC analysis

The FROC curves for detection of prostate cancer and the detection of high-grade prostate cancer are shown in figure 6.5. The results show that adding a candidate classification step reduced the number of false positives at constant sensitivity, e.g. a reduction from approximately 7 to 1.5 false positives per normal case at a sensitivity of 80%. At similar false positive levels, sensitivities were significantly higher after the candidate classification step, e.g. after voxel classification, at 1 false positive per normal case, a sensitivity of 55% is reached, whereas the sensitivity is 75% after the candidate stage ($p < 0.001$). Additionally, the partial area under the curve between 0.1 to 10 false positives per normal case is also significantly higher (7.11 versus 8.74, $p < 0.01$). Furthermore, in figure 6.5b the FROC curves are shown for the de-

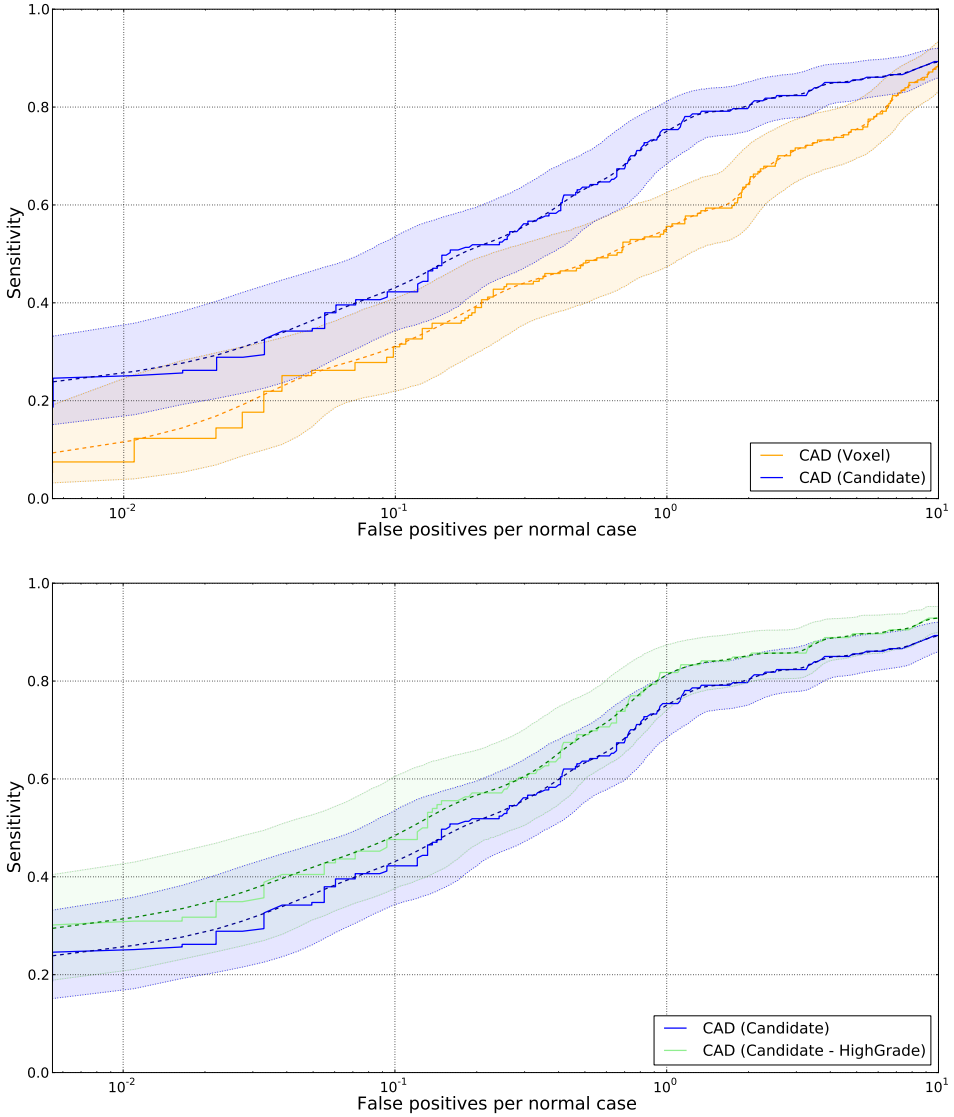



Figure 6.5: FROC analysis of the results of the CAD system. Number of false positives per normal case is shown on a logarithmic scale. The raw FROC curve is shown as the solid line and the mean bootstrapped curve as a dashed line. The 95% confidence intervals obtained using bootstrapping are shown as transparent areas around the mean bootstrapped curve. Figure (a) compares the results after voxel classification and the results after candidate classification. Adding the candidate classification step shows a marked improvement over just voxel classification. Figure (b) shows the results of the candidate classification step for cancer vs. normal/benign and high-grade cancer vs. normal/benign

tection of high-grade cancer vs. normal/benign in addition to the detection of all cancer vs. normal/benign. Here, at one false positive per normal case the sensitivity for detecting high-grade cancer is 0.82 and 0.75 for all cancer. This difference is not significant. Additionally, the partial area under the curve between 0.1 and 10 false positives per normal case is not significantly different (8.74 versus 9.06). The maximum sensitivity reached by the system in cancer versus normal/benign case is 0.94 and 0.97 for the high-grade cancer versus normal/benign case. This is caused by 11 and 5 false negatives in those cases respectively. Examples of a true positive and a false positive are shown in figure 6.7. An example of two false negatives is shown in figure 6.6.

### 6.3.4 ROC analysis

The ROC curves for classifying patients as either having prostate cancer or not having prostate cancer are shown in figure 6.4. Again we also show the improved performance obtained by adding a candidate classification step compared to just using a voxel classification step. In these figures the CAD system is also compared to the clinical diagnosis made by the radiologist for each patient on the basis of the PIRADS system. A radiologist scores each suspicious lesion on a scale from 1 to 5, 1 meaning definitely not cancer and 5 meaning definitely cancer. The patient score is than obtained by taking the highest PIRADS score. For the radiologist only the confidence intervals for the four actual PIRADS thresholds are used for evaluation, as the ROC curve is not well defined at other positions due to the low number of thresholds.

The addition of the candidate classification shows a marked improvement when evaluating on a per-patient basis, with an increase in AUC from 0.722 to 0.81 ($p < 0.01$) and from 0.73 to 0.83 ($p < 0.01$) for high-grade cancer versus normal/benign. At a high specificity (left part of the ROC curve) of 0.88 (PIRADS score 5), there is no significant difference between the radiologist and the CAD system ($p = 0.334$ for detection of cancer, $p = 0.37$ for detection of high-grade cancer). At the other thresholds the radiologist performance is significantly better than the CAD system ($p < 0.01$). The radiologist is significantly better at every PI-RADS threshold compared to CAD system when only using the voxel stage.

## 6.4 Discussion

A CAD system which detects prostate cancer in MRI images was presented in this paper. The performance of the system was evaluated on a large consecutive set of patients, with MR-guided biopsy as a reference standard. Quantitatively, the area under the ROC curve for classifying patients was 0.81. If we investigate the performance from an FROC perspective, at 1 false positive per image we obtain a sensitivity of 0.75 for detecting any cancer lesion and 0.83 for detecting a high-grade cancer lesion. Compared to the radiologist, the system shows no significant differences in performance at high specificity (Figure 6.4, left part of

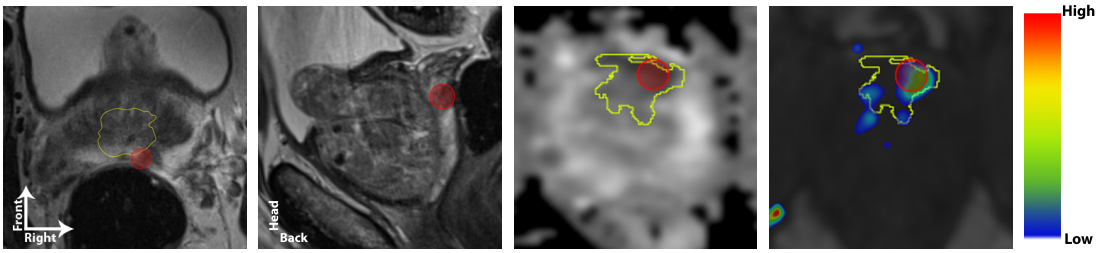



Figure 6.6: Examples of false negatives (FN). The location of the false negatives is indicated with the red circle. False negative 1 (a,b) is caused by segmentation errors. In Figure (a) (axial view) the prostate segmentation is indicated in yellow. The segmentation most likely fails due to the fact that the prostate is growing into the bladder, as can be seen in (b) (sagittal view). False negative 2 is caused by our evaluation criterion, the region segmentation (indicated in yellow in figures (c) (ADC) and (d) (DCE+Ktrans) is quite large and therefore the mark (the red sphere) and the center of gravity of the region segmentation are more than 10mm apart.

the curve). However, at lower specificity the radiologist performs significantly better ($p < 0.01$). Furthermore, in both Figure 6.5 and Figure 6.4 we show that adding a candidate stage in addition to a voxel stage significantly improves performances on both a per-lesion (sensitivity at 1 false positive per normal patient increases from 0.55 to 0.75) and per-patient level (area under the ROC curve increases from 0.72 to 0.81).

If we compare our system to the current state-of-the-art, two types of systems can be distinguished: systems which perform only a voxel-based analysis and systems which perform both a voxel-based analysis and a candidate evaluation step. For the first type of system, Tiwari et al. have shown the best voxel classification performance using a system with manual prostate segmentation and MR spectroscopy. They obtain an average area under the curve of 0.89, which is similar to our results obtained during the classifier comparison at the voxel level figure 6, average area under the curve of 0.889). However, we show in our study set that a voxel classification performance of 0.89 only results in a per-lesion classification performance of 0.55 sensitivity at 1 false positive per normal patient and a per-patient area under the ROC curve of 0.72. Our addition of a subsequent candidate classification step increases the performance of the system by a significant amount (0.75 sensitivity at 1 false positive per normal patient ($p < 0.01$), figure 8 and 0.81 area under the ROC curve for a per-patient analysis ($p < 0.01$). However, as Tiwari et al. did not extend their system to a per-region and per-patient evaluation we cannot directly compare this. We do have to note that this comparison has not been made on the same data set, which is unfortunate, but there is no open availability of a significant amount of multi-parametric prostate MRI data. We are currently considering organizing a prostate cancer detection challenge similar to other grand challenges in medical image analysis to allow our algorithm to fairly compete against others. For the second type of system incorporating a candidate detection and a candidate classification step we can compare our results to Vos et al.[96], which is the only other two-stage system. Instead of a voxel classification step, they use a blob detector to obtain the candidates. At 0.1, 1 and 10 false positives

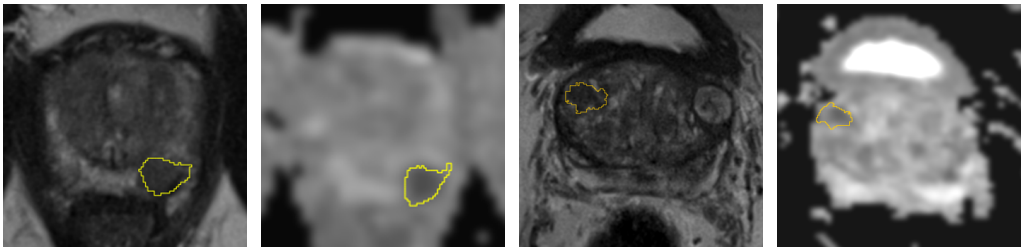


Figure 6.7: Examples of a true positive (a-b) and a false positive (c-d) on T2-weighted imaging (a,c) and the ADC map (b,d).

| **Feature** | **AUC** | **Modality** | **AUC** |
|---|---|---|---|
| ADC | 0.666 | T2W | 0.760 |
| Gauss. Texture (Order=-, Scale=2.0) | 0.633 | DWI | 0.746 |
| tau Blob | 0.632 | DCE | 0.714 |
| $K_{Trans}$ | 0.629 | | |
| T2-map | 0.620 | | |
| $k^{ep}$ | 0.604 | | |
| $K_{Trans}$ Blob | 0.600 | | |
| Gauss. Texture (Order=YY, Scale=5.1) | 0.599 | | |
| Gauss. Texture (Order=-, Scale=3.17) | 0.597 | | |
| b800 | 0.588 | | |

Table 6.2: Results for relative feature and modality importance experiment based on leave-one-patient-out cross-validation for voxel classification.

per normal case they obtained a sensitivity of 0.15, 0.48 and 0.89 where we obtain a sensitivity of 0.42, 0.75 and 0.89. Especially at the lower false positive rates we obtain a substantially better performance.

The use of multi-parametric MRI over single-parameter MRI is already part of clinical guidelines[33]. In this study we investigated the performance of individual features and MR sequences in both the voxel and candidate stages of our CAD system. Especially in the voxel stage, the combined interpretation of all three MR sequences, T2-weighted imaging, diffusion-weighted imaging and dynamic contrast-enhanced imaging shows a large improvement over using any single MR sequence (highest performing single sequences AUC is 0.76, combination is 0.89), and all play about an equal role. In the subsequent candidate stage there is a preference for DWI over DCE over T2W., This experiment also showed that the individual performance of features was relatively low compared to the combination of all features (best performing feature in the voxel stage had a AUC of 0.66, whereas the overall voxel stage AUC was 0.89). These observations confirm clinical practice. A limitation of this study is the fact that the ROC evaluation is positively biased toward the radiologist. Although the reference standard

| Feature | AUC | Modality | AUC |
|---|---|---|---|
| b800 (standard deviation) | 0.805 | DWI | 0.910 |
| ADC (standard deviation) | 0.790 | DCE | 0.814 |
| Voxel likelihood (mean) | 0.765 | T2W | 0.719 |
| LateWash Blob (standard deviation) | 0.750 | | |
| ADC Blob (standard deviation) | 0.741 | | |
| LateWash (standard deviation) | 0.731 | | |
| Voxel likelihood (75th percentile) | 0.727 | | |
| Voxel likelihood (standard deviation) | 0.723 | | |
| $k^{ep}$ (standard deviation) | 0.723 | | |
| Volume | 0.716 | | |

Table 6.3: Results for relative feature and modality importance experiment based on leave-one-patient-out cross-validation for voxel classification.

for cancerous regions is well defined by the MR-guided biopsy specimens, for most of the normal regions we have to depend on the opinion of the radiologist. While we incorporated only data with either negative biopsy results or very low PI-RADS scores (1 and 2) there is still the risk that some areas we deem normal are actually prostate cancer. Furthermore, in the evaluation and the comparison to the radiologist, it may well be that the radiologists did have some false negatives. Recently, prospective preliminary results were published by Thompson et al.[62] They found that the sensitivity for radiologists for detecting high-grade prostate cancer was 96%. If we look at the potential of our CAD system in such a setting, at a operating point with a sensitivity of 96%, the specificity of the CAD system is between 15-40%. This could indicate that between 15-40% of all studies could be read by the CAD system and would potentially require no human intervention, which could reduce the workload of the radiologist substantially. Another limitation is the fact that although the multi-parametric MRI is implicitly registered (all sequences are acquired in one go, without the patient leaving the scanner), registration errors between the different sequences could occur due to patient movement. This was mostly circumvented in our data by a. not using an endorectal coil, which significantly improves patient comfort and as such reduces patient movement and b. by administering Buscopan prior to the MRI to reduce bowel movement and c. using multi-scale features (Gaussian texture, blobness) where exact voxel alignment is less important. Further improvement could be achieved by implementing a registration algorithm for prostate MRI, however, this is currently an unsolved problem. To the best of our knowledge there are currently no publications on the registration of multi-parametric prostate MRI.

The false negatives in our system are mostly caused by prostate segmentation errors. Of the 11 false negatives after candidate classification, 6 are caused by the prostate segmentation

missing the lesion entirely or partly. Most of these issues can be solved by incorporating a segmentation method which is more robust to strange prostate shapes. An example is shown in figure 6.6. Although the prostate segmentation algorithm is a candidate for improvement, it is missing only 6 out of 183 lesions (or 3% of total sensitivity), which is still a reasonable result. The segmentation algorithm only finished 9 out of 12 in the PROMISE12-challenge, however overall segmentations were pretty accurate, with a median Dice's coefficient of 0.83. For the remaining false negatives, in four cases the lesion was detected, but the candidate segmentations were so large that the center of the candidate segmentation and the lesion marker were more than 10 mm apart, thus failing our criterion for a true positive. This can happen because in big lesions the radiologist did not always put the point annotation at the center of the lesion. For one false negative the area was not identified by the voxel classification and the local maxima detection and thus lost to the second part of the system.

The motivation for the development of a CAD system is to aid radiologists by improving efficiency and performance. The amount of cases a radiologist has to read in a screening setting is enormous, and our CAD system will be most useful in such a situation. However, we have not yet evaluated the system on screening data as the data in this study only includes patients with previous negative TRUS biopsies. Therefore we cannot make any claims on the performance of the system in a screening setting. Summarizing, a fully automatic CAD system was developed for the detection of prostate cancer in MRI images. Performance evaluation shows that it outperforms the state-of-the-art, although the comparison has its limitations due to different evaluation data sets. Furthermore, the system is not significantly different from radiologist performance at high specificity. Therefore we believe it to be a potentially valuable tool to aid radiologists in the clinic.

# Evaluation of a computer-aided detection system for prostate cancer in MRI

# 7

Geert Litjens, Jelle Barentsz, Nico Karssemeijer, Henkjan Huisman

*Original title:* Clinical evaluation of a computer-aided diagnosis system for determining cancer aggressiveness in prostate MRI.

*Submitted to:* European Radiology

## 7.1 Introduction

Multi-parametric magnetic resonance imaging (mpMRI) is emerging as an important modality in prostate cancer diagnosis[33,196,197]. Multi-parametric MRI combines T2-weighted, diffusion-weighted, and dynamic contrast-enhanced techniques to provide information, respectively, on abnormal anatomy, cell-density, and neo-vascularity. Several studies have shown that in patients with initial negative trans-rectal ultrasound-guided biopsies (TRUSGB) expert readers using mpMRI find cancer in 38 - 59% of the cases[44,198]. Furthermore, it has been shown that mpMRI upgrades cancer aggression of previously TRUSGB detected cancers in up to 30% of cases[199]. Several other studies found that the negative predictive value of mpMRI is high enough to avoid TRUSGB in 30-50% of men with elevated PSA[62,200]. However, one of the main limitations for broader acceptance of mpMRI is the lack of required expertise, especially in the interpretation of the MR images[31,32,102].

In order to improve the acquisition and interpretation of mpMRI the European Society for Urogenital Radiology (ESUR) established initial guidelines for acquisition and standardized interpretation of mpMRI[33]. These guidelines have been evaluated by several groups, both for detection of cancer prior to biopsy[62,63] and after initial negative TRUSGB[58–61]. Pokorny et al. found that using mpMRI and MR-guided biopsy in biopsy-naïve men results in a large reduction of over-diagnosis of low-grade cancer (82%) while detecting 17% more intermediate/high-grade cancers than TRUSGB. Thompson et al.[62] showed that an mpMRI score of 3 or higher (on a 5 point scale) would result in a sensitivity of 97% and a specificity of 50% for the detection of high Gleason grade cancer using saturation template biopsy as a reference standard. Deferring biopsy to a score of 4 or higher would result in a substantially improved specificity of 92%, but a sensitivity of only 67%.

To further improve prostate mpMRI interpretation, biomarkers and computerized decision aids are actively researched to help detect intermediate/high-grade prostate cancer. Several groups have focused on correlating individual MR parameters (DCE[47], DWI[51,160] or spectroscopy[201]) to cancer grade, estimating their usefulness as quantitative biomarkers of prostate cancer aggressiveness. However, none have yet focused on multi-variate quantitative analysis to determine cancer grade. Various groups are developing computer-aided diagnosis algorithms (CAD) to detect aggressive tumours[94,96,97]. They have shown that CAD can help improve the interpretation, especially for inexperienced radiologists[97,202]. However, the evaluation of such systems has been limited to observer studies and ROIs pre-selected by the researchers[97,202], or to specific sub-parts of the prostate, like the peripheral zone[202].

The purpose of this study is to investigate the clinical effect of a recently developed, state-of-the-art computer-aided diagnosis (CAD) system[203] on the diagnostic accuracy of prostate MRI PIRADS reporting and to study the ability of CAD to help assess prostate cancer aggressiveness.

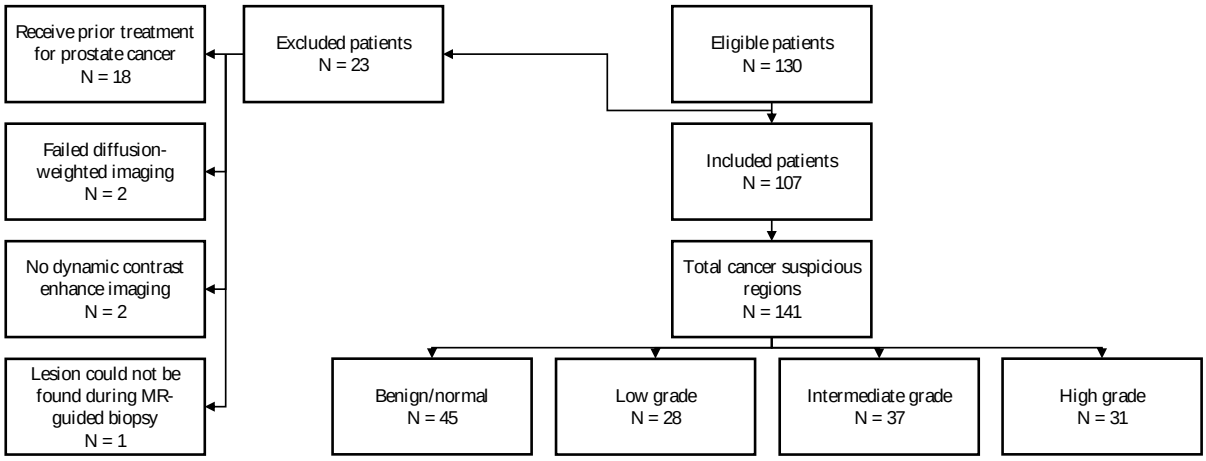



Figure 7.1: STARD diagram of inclusion and exclusion criteria of the prospective patient cohort.

## 7.2 Materials and Methods

### 7.2.1 Patient data

The institutional review board waived the need for informed consent as only imaging data and MR-guided biopsy results obtained from regular care were used. To prospectively evaluate the CAD system we included all 130 patients from January 1st to September 1st 2013 that received both an mpMRI and a subsequent MR-guided biopsy at our institution. The inclusion criteria for the detection MRI were an initial negative TRUSGB and persistently elevated PSA. Multiparametric MRIs were acquired according to the ESUR guidelines and included T2-weighted imaging in three orthogonal directions, diffusion-weighted imaging and dynamic contrast enhanced imaging. All MRIs were performed at a Siemens 3T MRI scanner (TRIOTIM or Skyra) without an endo-rectal coil. Full acquisition details are presented in Table 7.1.

MRIs were read prospectively by one radiologist out of the group of radiologists that report prostate MRI in our clinic. Experience levels of the reporting radiologist ranged from inexperienced (1 year) to very experienced (J.B., 20 years of experience with prostate MRI). MRI studies were read according to the ESUR prostate imaging reporting and data system (PI-RADS) classification including ongoing local and international score refinements[63]. Cancer suspicious regions were given a final PIRADS score from 1 to 5, where 1 means 'definitely not cancer' and 5 means 'definitely cancer'. Per patient all regions with a PIRADS 3, 4 or 5 were indicated. If no lesion with PIRADS 3-5 was present, a lesion or normal tissue with a PIRADS 1 or 2 score was identified. Thus in each study at least one region was indicated. Cancer suspicious regions (PIRADS 4 and 5) were subsequently biopsied under direct MR-guidance. PIRADS 3 lesions were biopsied only if there was a high clinical suspicion for prostate cancer (e.g. extremely high PSA, as assessed by the radiologist). A PIRADS 2 lesion was only biopsied, when a biopsy was already required due to the presence of a PIRADS 3, 4 or 5 lesion. All

| | PS | SR | ST | ET | RT | FA | Other |
|---|---|---|---|---|---|---|---|
| T2W | Turbo spin-echo | 0.28 – 0.6 mm | 3.0 – 3.2 mm | 101 – 104 ms | 4480 – 6840 ms | 120 - 160 | Acquired in three orthogonal directions: transversal, sagittal and coronal. |
| DWI | Echo planar | 2 mm | 3 mm | 63 – 81 ms | 2800 – 3600 ms | 90 | 3 b-values: 50, 400 – 500, 800 averaged over 3 directions. Apparent diffusion coefficient map calculated by the scanner software. |
| DCE | Turbo FLASH | 1.5 – 1.8 mm | 3.2 – 5 mm | 1.41 ms | 36 ms | 10 - 14 | Temporal resolution of 3.38 – 4.65 seconds, 36 – 50 timepoints. 15 mL contrast agent used (Dotarem, Guerbet, France) |

Table 7.1: MRI sequence details for the different types of acquisitions. PS = pulse sequence SR = spatial resolution, ST = slice thickness, ET = echo time, RT = repetition time, FA = flip angle.

other PIRADS 1, 2 or 3 lesions were not biopsied and not further considered in this study.

Lesions were categorized into either benign/low-grade (indolent) or intermediate/high-grade cancer (aggressive) based on the MR-guided biopsy Gleason scores. For brevity, the two categories will be subsequently referred to as indolent and aggressive. The lesion categorization strategy is similar to[51]. Details can be found in Table 7.2.

### 7.2.2 Computer-aided diagnosis system

The computer-aided diagnosis system evaluated in this paper was previously presented in[203]. This system can fully automatically analyze prostate MRIs by first segmenting the prostate. Next, quantitative voxel features are computed based on the PIRADS guidelines (e.g. the apparent diffusion coefficient, the presence and amount of washout in the DCE MRI). Machine learning techniques summarize the voxel features into a likelihood of cancer per voxel. Local maxima detection is then used to identify suspicious regions in the voxel likelihood map. These regions are analyzed in more detail with region-based features (e.g. voxel statistics like the 25th percentile of the ADC within a region, symmetry within the prostate and local contrast). A second machine learning step combines the region features into a region likelihood. The system was trained with retrospective patient data, which had no overlap with the data set used in this study. The retrospective data was acquired in a similar manner as the prospective data and had the same reference standard strategy. Details of the retrospective cohort can be found in Table 7.4.

For the study in this paper one modification was made to the system. The system normally operates autonomously and chooses areas deemed suspicious enough for the second

| Grade | Gleason scores |
|---|---|
| Low | 3+3 or lower, no 4 or 5 component |
| Intermediate | 2+4, 3+4, 2+5 |
| High | 3+5, any cancer with a major 4 or 5 component |

Table 7.2: Mapping of Gleason scores to cancer grade

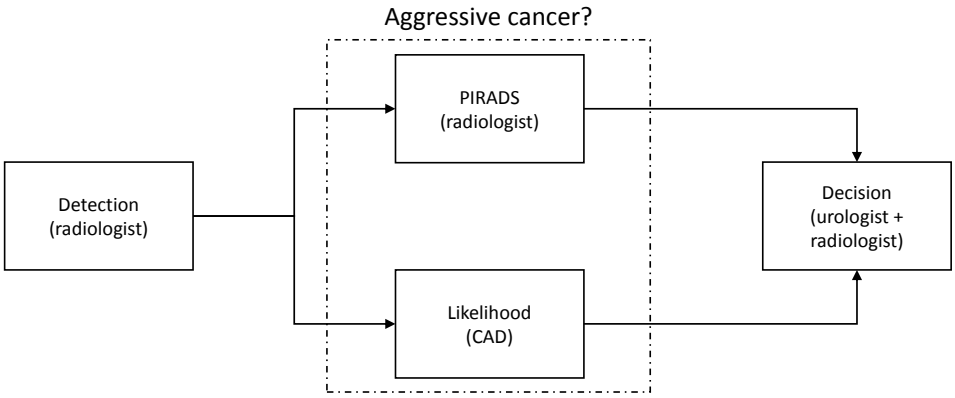



Figure 7.2: Suggested workflow for the proposed CAD system. The biopsy decision can be made by the radiologist, another attending clinician or by using a nomogram (Figure 7.3) to independently combine the PIRADS score and the CAD likelihood.

stage analysis independently. However, for the purpose of this study we used the radiologist-indicated suspicious location(s) as input for the second stage of the CAD system (replacing the local maxima detection). This modification was made to ensure a likelihood from the CAD system is obtained for each region indicated by the radiologist, even if the CAD system itself deemed the region not suspicious enough after the first stage.

### 7.2.3 Combination of the CAD system and the radiologist

The use of the system as suggested in this paper is presented schematically in Figure 1. The initial identification of potential suspicious regions was performed by the radiologist, after which the radiologist and the CAD system gave independent predictions on whether the suspicious lesion is an indolent or aggressive lesion. The radiologist did this by assigning a PIRADS score, while the CAD system assigned a continuous cancer likelihood score between 0 and 1.

The reported scores of the radiologists and CAD were combined afterwards using a nomogram, which is a method to map several distinct measurements and observations to a single outcome variable in an unbiased manner. We developed such a nomogram by using logistic regression to independently combine the PIRADS score of the radiologist and the likelihood from the CAD system. This nomogram was created based on the retrospective data that was also used to train the CAD system and subsequently evaluated on the prospective data pre-

sented in this paper. It is noted that in a clinical workflow alternative methods of incorporating CAD results may be used, such as asking the radiologist to make a final decision based on the two scores.

### 7.2.4 Statistical evaluation

The statistical evaluation consisted of three parts. First, we investigated the potential increase in predictive power of the radiologist/CAD-combination over the radiologist alone using the likelihood ratio test on the logistic regression models.

Second, the diagnostic performance of the CAD system, the radiologist and the combination of radiologist/CAD was evaluated using receiver-operating characteristic (ROC) analysis. The significance of improvement for both sensitivity and specificity was tested using bootstrapping at the points for which the ROC curve of the radiologist was explicitly defined (i.e. the different PIRADS thresholds). A total of 10,000 bootstrap samples was used to obtain the 95% confidence intervals.

Third, we investigated whether the likelihoods of aggressive disease obtained from the CAD system correlate to cancer grade. As cancer grade is an ordinal variable, Spearman's rank correlation coefficient was used to estimate this correlation. Furthermore, the likelihood obtained from the radiologist/CAD-combination nomogram was also correlated to cancer grade.

For the evaluation, a correctly identified aggressive lesion was a true positive result. When an indolent lesion was identified as an aggressive lesion, this was considered a false positive. Note that we explicitly considered diagnosis of low-grade cancer a false positive; identification of low-grade cancer can lead to over-diagnosis and over-treatment.

All analysis was done on a per-lesion basis. For all significance tests a p-value threshold of 0.05 was chosen. SPSS (SPSS, version 20.0.01, Chicago, U.S.A) was used for statistical analysis.

## 7.3 Results

### 7.3.1 Patient data

The prospective cohort composition is detailed using a STARD diagram in Figure 7.1. Information on patient age/PSA ranges and tumor grade distribution are shown in Table 7.3. PSA and age ranges are similar to other studies using patient data with similar inclusion criteria (initial negative TRUS biopsy and persistently elevated PSA)[58–61]. Furthermore, the distribution of cases between the two categories indolent and aggressive was similar (73 versus 68 lesions).

Information on the retrospective cohort that was used to train the CAD system and obtain the nomogram is presented in Table 7.4. A similar distribution of PSA levels, age and cancer grade can be observed between the retrospective training cohort and the prospective evaluation cohort.

| **Clinical characteristics** | | | | |
|---|---|---|---|---|
| PSA level, mg/ml, median (range) | 13 (1 – 56) | | | |
| Age, yr, median (range) | 66 (48 – 83) | | | |
| **Gleason score** | **Grade** | **No. of lesions** | | |
| Normal/Benign | Normal/Benign | 45 | No cancer | 45 |
| 2 + 5 | Intermediate | 1 | Low | 28 |
| 3 + 2 | Low | 2 | Intermediate | 37 |
| 3 + 3 | Low | 26 | High | 31 |
| 3 + 4 | Intermediate | 36 | Total | 141 |
| 4 + 3 | High | 12 | | |
| 4 + 4 | High | 5 | | |
| 4 + 5 | High | 10 | | |
| 5 + 4 | High | 3 | | |
| 5 + 5 | High | 1 | | |

Table 7.3: Characteristics of patients and biopsy specimens for the prospective cohort (107 patients). PSA ranges were determined on 103 patients, for 4 patients PSA levels were unknown.

| **Clinical characteristics** | | | | |
|---|---|---|---|---|
| PSA level, mg/ml, median (range) | 11 (1 – 57) | | | |
| Age, yr, median (range) | 65 (42 – 78) | | | |
| **Gleason score** | **Grade** | **No. of lesions** | | |
| Normal/Benign | - | 151 | No cancer | 151 |
| 2 + 3 | Low | 3 | Low | 61 |
| 2 + 4 | Intermediate | 1 | Intermediate | 67 |
| 3 + 2 | Low | 3 | High | 53 |
| 3 + 3 | Low | 55 | Total | 332 |
| 3 + 4 | Intermediate | 66 | | |
| 4 + 3 | High | 28 | | |
| 4 + 4 | High | 12 | | |
| 4 + 5 | High | 11 | | |
| 5 + 4 | High | 2 | | |

Table 7.4: Characteristics of patients and biopsy specimens for the retrospective cohort (254 patients). PSA ranges were determined on 220 patients, for 34 patients PSA levels were unknown.

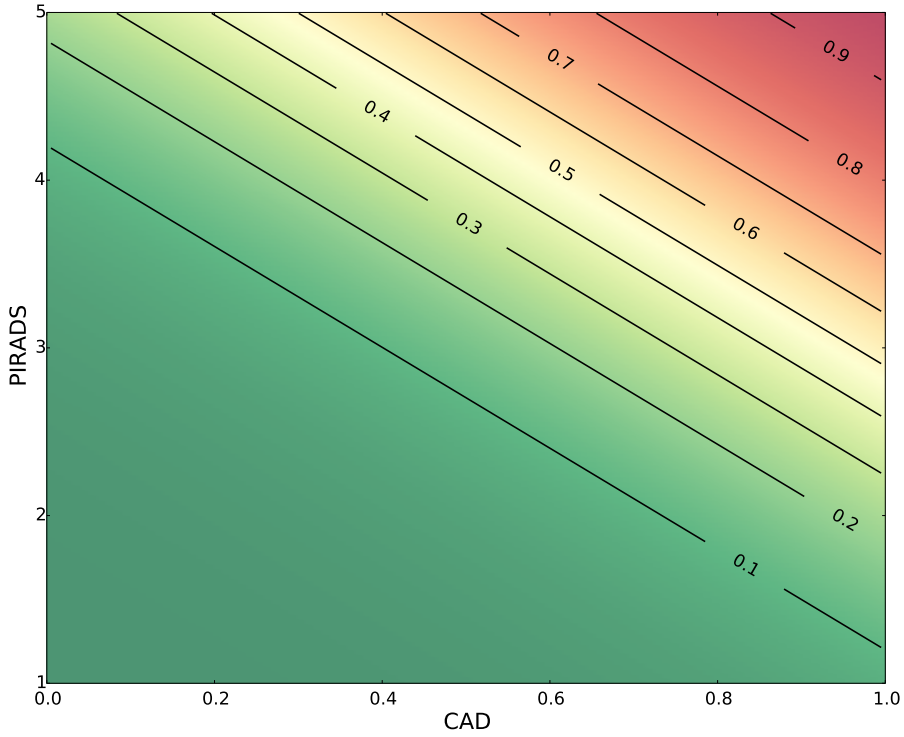



Figure 7.3: Nomogram estimated using logistic regression for the combination of the radiologist and the CAD system. The likelihood for intermediate/high-grade cancer is indicating by the color coding and the contour labels and ranges from 0 to 1. Green indicates low likelihood and red indicates high likelihood.

### 7.3.2 Combination of PIRADS and CAD likelihood

TFirst, the logistic regression procedure showed that including the CAD system likelihood in addition to the radiologist PIRADS score resulted in a model with significantly improved predictive power ($p < 0.001$, likelihood ratio test) for aggressive disease. The nomogram obtained using logistic regression on the retrospective training data of the CAD system is presented in 7.3. The equation describing this nomogram is:

$$\text{Likelihood} = \frac{1}{1 + e^{-t}} \tag{7.1}$$

$$t = -7.629 + 3.886C + 1.295P \tag{7.2}$$

where C is the CAD system likelihood and P is the radiologist PIRADS score. Second, the results of using this nomogram prospectively to combine the radiologist score and the CAD system are presented in Figure 7.4. Here the ROC curves for the CAD system and the combination are shown. The performance of the radiologist is presented at three different PIRADS thresholds.

A significantly improved sensitivity was obtained at the PIRADS 4 threshold (0.98 for the radiologist/CAD-combination versus 0.93 for the radiologist alone, $p = 0.029$). Furthermore a significantly improved specificity was found for the PIRADS 3 threshold (0.25 for the

radiologist/CAD-combination versus 0.09 for the radiologist alone, $p = 0.013$).

### 7.3.3 Correlation of likelihood and cancer grade

Third, the relations between the CAD system likelihood and cancer grade is presented in Figure 7.5a as box-plots. In Figure 7.5b the same result is presented for the CDS/radiologist combination. A clear trend can be observed; higher likelihoods relate to higher cancer grade. The Spearman's rank correlation coefficient was 0.536 when using only the CAD system, 0.582 for the radiologist alone and 0.694 when combining the CAD system and the radiologist. All correlations were significant ($p < 0.01$).

## 7.4 Discussion

The Results of this study indicate that computerized analysis of prostate regions to characterize cancer grade may help improve radiologist performance in selecting biopsy targets in a prospective setting.

The likelihoods of intermediate-to-high-grade cancer of the CAD system (which combines all MRI parameters) significantly correlated with cancer grade. An even higher, significant Spearman's rank correlation coefficient of 0.694 was obtained by using the nomogram (Figure 7.3) combining the radiologist with the CAD system. As far as we are aware, there are currently no prospective studies assessing the correlation of multiple MR parameters with cancer grade.

To translate these results to clinical practice, we tested whether combining the radiologist and CAD system results in improved sensitivity or specificity at the different PIRADS thresholds. Inspecting the ROC curve (Figure 7.4) for the CAD system with respect to the performance of the radiologist at the different PIRADS thresholds, we can appreciate that the performance of both the radiologist and the CAD system seem similar. However, they provided complementary information, as the combination of both predictions (PIRADS score and CAD likelihood) results in an improved ROC curve (blue curve, Figure 7.4). At a PIRADS threshold 3 a significant increase in specificity was found ($p = 0.013$) and at the PIRADS 4 threshold a significant increase in sensitivity was found ($p = 0.029$). Thus, combing radiological expertise with computational methods to characterize prostate cancer results in an improved sensitivity and specificity.

In the study by Pokorny et al. it was already established that MRI before any TRUSGB has the potential to reduce the amount of biopsies by 51% compared to TRUSGB[63]. Improving the performance of the reporting radiologist by adding the CAD system as an independent second reader as presented in this paper might further reduce the amount of biopsies by better characterizing lesions as aggressive or indolent in vivo. Furthermore, the use of computer-aided diagnosis system might make reporting prostate MRI feasible for less experienced radiologists or urologists. Although this was not explicitly investigated in this paper, the high standalone

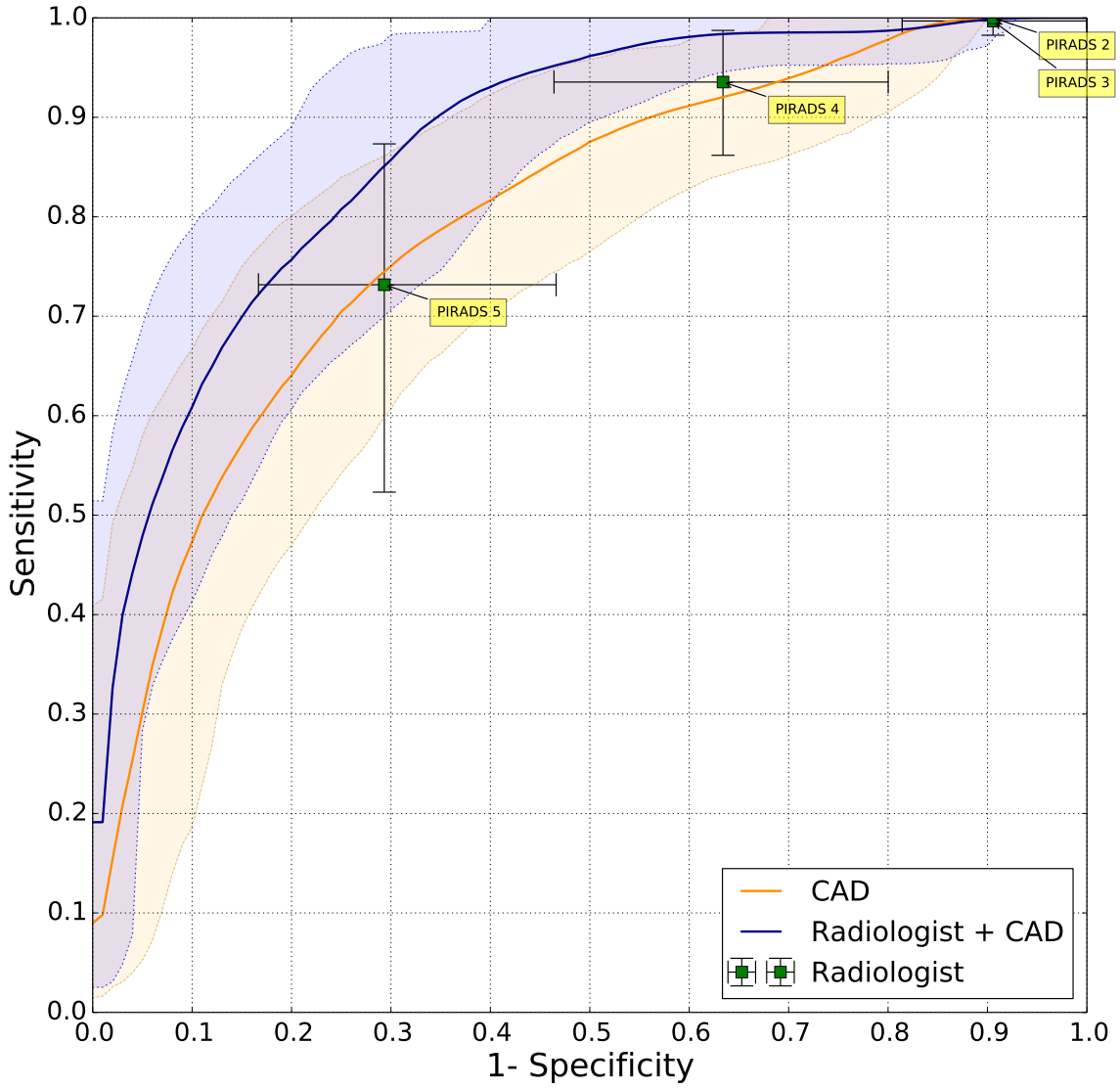
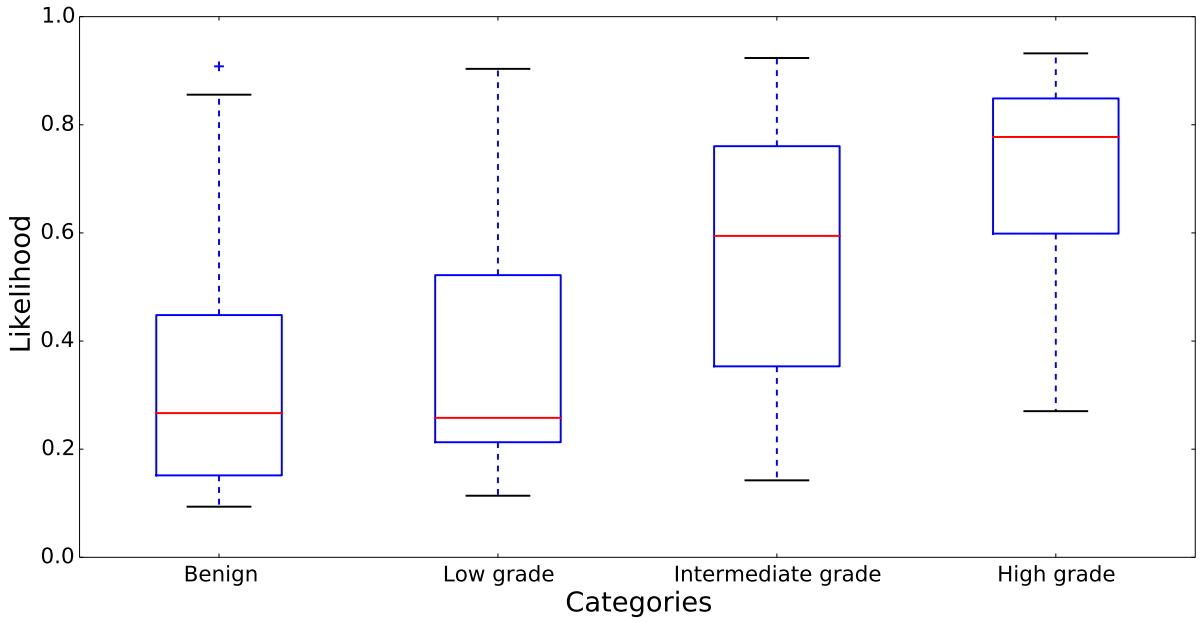
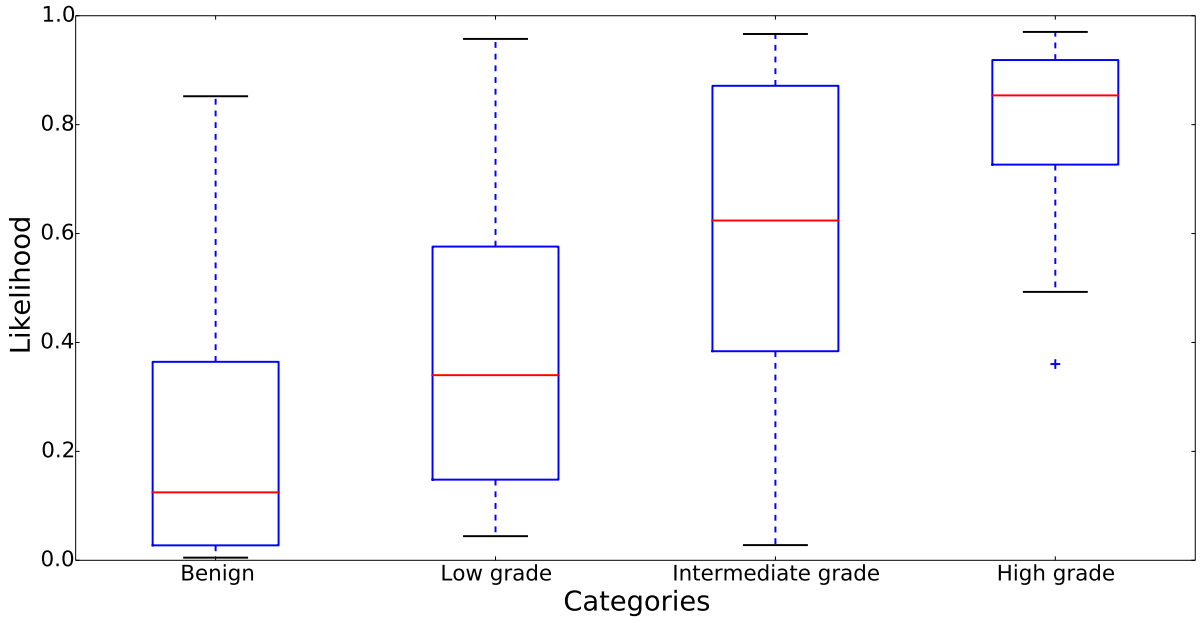



Figure 7.4: Receiver-operating characteristic curve showing the performance of the CAD system (orange) and the radiologist/CAD-system combination (blue). The shaded areas indicated the 95% confidence intervals as calculated using bootstrapping. The radiologist performance is indicated with point for the different PI-RADS thresholds. The vertical error bars indicate the 95% confidence interval on the sensitivity and horizontal error bars indicated the 95% confidence interval on the specificity as estimated by bootstrapping.




(a)




(b)

Figure 7.5: Relation between computer system likelihood and cancer grade (a) and the computer system/radiologist combination and cancer grade (b) when the system is trained to detect only intermediate-to-high-grade cancer.

performance of the CAD system supports this idea.

This study has some limitations. First, although MR-guided biopsy has been described to have a very high concordance with prostatectomy Gleason grade (95% detection rate for Gleason 4 and 5 components), it is not 100% accurate[28]. This potentially implies, that some of the cancers in our study are under- or over-graded by the MR-guided biopsy.

A second limitation is that each case in this study was read by one radiologist. This means that inter-observer variability could not be established in this study. However, a consensus meeting re-evaluated uncertain scores prior to biopsy. In future work it would be beneficial to compare the performance of the CAD system and radiologist to two radiologists performing a first and second read. A third limitation is that our evaluation only pertains to a single center, so we could not test whether our results generalize to different centers.

Last, due to the fact that PIRADS 1 and 2 lesions are generally not biopsied we cannot assess the full performance of the CAD system and the radiologist in negative cases without long-term follow-up or external reference tests like full template biopsy. However, this has little impact on the results of this study. Of all 7 biopsied PIRADS 2 lesions, none was categorized as aggressive. This indicates that radiologists are already reading at a very high sensitivity level and they do not need a computer-aid at the lower PIRADS scores. Literature also confirms this assessment, with the studies by Thompson et al.[62] and Pokorny et al.[63] reporting MRI sensitivities and negative predictive values of 97 and 96.9% respectively when PIRADS 3 and higher lesions are biopsied. The CAD system has most potential in more accurately discriminating which PIRADS 3, 4, or 5 lesions require biopsy and the results at these scores are not affected by the lack of PIRADS 1 or 2 biopsies.

## 7.5 Conclusions

In this paper the use of a computer-aided diagnosis system in conjunction with the radiologist to accurately characterize prostate lesions was investigated. Result showed that a significant increase in diagnostic performance can be achieved when combining the radiologist PIRADS score and the CAD system likelihood. Furthermore, a significant correlation between CAD likelihood and cancer grade exists; this increases further when the PIRADS score and the CAD likelihood are combined using a logistic regression-based nomogram.

# Summary and discussion

# Summary

# 8

The focus of this thesis was the research and development of a CAD system for detection of cancer in prostate MRI. It covers segmentation of relevant structures (chapters 2 and 3), feature discovery (chapters 4 and 5), building the CAD system (chapter 6) and system evaluation (chapter 7).

In chapter 2 the organization of a challenge for prostate segmentation in MRI is discussed. The challenge was setup to allow fair and meaningful comparison of different segmentation algorithms. Challenge design included the acquisition of multi-center, multi-vendor and multi-protocol data and the development of evaluation measures that would allow us to show clear separation between algorithms on the basis of performance. The challenge included both an online component and a live challenge during a workshop at the MICCAI2012-conference in Nice, France. In total 11 teams participated in this initial phase of the challenge, with distinctly different algorithms, ranging from active-shape models to multi-atlas registration approaches. The results indicated that the top-two participating algorithms significantly outperformed all the algorithms outside of the top-three and had an efficient implementation with a run time of 8 minutes and 3 second per case respectively. Overall, active-shape model based approaches seemed to outperform other approaches like multi-atlas registration, both on accuracy and computation time. Average algorithm performance was good to excellent and the Imorphics algorithm even managed to outperform the second human observer on average. However, we showed that algorithm combination might lead to further improvement, indicating that optimal performance for prostate segmentation is not yet obtained.

Segmentation of the prostate zones is a relatively new research topic; initial results of a pattern recognition approach for zonal segmentation are discussed in chapter 3. Zonal segmentation of the prostate into the central gland and peripheral zone is a useful tool in computer-aided detection of prostate cancer because occurrence and characteristics of cancer in both zones differ substantially. We opted for a pattern recognition approach because it can capture the distinct appearance differences through a wide range of quantitative image characteristics and is best suited to deal with the high variability in zonal shapes. The algorithm incorporates three types of features that can differentiate between the two zones: anatomical, intensity and texture. It is evaluated against a multi-parametric multi-atlas based method using 48 multi-parametric MRI studies. Segmentations from three observers were used to assess inter-observer variability and we compared our results against the state of the art. We obtained a mean Dice coefficient of $0.89 \pm 0.03$ for the central gland and $0.75 \pm 0.07$ for the peripheral zone, compared to $0.87 \pm 0.04$ and $0.76 \pm 0.06$ for the state of the art. Summarizing, a pattern recognition approach incorporating anatomy, intensity and texture has been shown to give good results in zonal segmentation of the prostate.

Features to discriminate between specific types of benign disease and prostate cancer are discussed in chapter 4. The presence of benign disease in the prostate acts as a confounder for the diagnosis of prostate cancer. The most common types of benign findings in the prostate

are benign prostatic hyperplasia (BPH), atrophy, inflammation and prostatic intra-epithelial neoplasia (PIN). To establish the imaging characteristics of these classes we used fusion of MRI and histopathology, computer-extracted features and classification. Prostatectomy and pre-operative multi-parametric prostate MRI of 70 patients were included in this study. Intensity, texture and pharmacokinetic features were extracted for each of the confounding classes and prostate cancer. Feature selection was performed for each of the pair-wise classification tasks (cancer vs. BPH, PIN, inflammation and atrophy, respectively) to identify the top five features for each. In total 92 PIN, 64 atrophy, 120 inflammation and 73 BPH lesions, and 128 cancer lesions were annotated. For each of the classification tasks distinct features were identified which provided the best discriminatory performance. Furthermore, for all classification tasks the area under the ROC curve improved significantly compared to monolithic classification (all benign confounders considered a single class).

Chapter 5 assessed whether we could improve the potential of the apparent diffusion coefficient in assessing cancer aggressiveness by explicitly incorporating inter-patient variation in the normal peripheral zone ADC. Intra-, and inter-patient variation of peripheral zone ADC was determined by repeated measurements of normal regions in a retrospective cohort of 10 consecutive patients over three separate MR imaging sessions at 3T. The effect of this intra- and inter-patient variation on assessment of prostate cancer aggressiveness was examined in a second retrospective cohort of 51 patients with prostate cancer who underwent an MRI, prior to prostatectomy. Logistic regression was used to assess whether incorporating normal ADC values improved the prediction of cancer aggressiveness. The effect on the diagnostic performance was assessed using receiver-operating characteristic analysis. The repeated-measures ANOVA revealed that inter-patient variability was significantly larger than measurement variability. Analysis of standalone tumor ADC values showed an AUC of 0.91 for discriminating low- vs. high-grade tumors. Incorporating normal PZ ADC using linear logistic regression, significantly improved the AUC to 0.96.

The development of the two-stage CAD system is detailed in chapter 6. In the first stage we detect initial candidates using multi-atlas-based prostate segmentation, voxel feature extraction, classification and local maxima detection. The second stage segments the candidate regions and using a classifier we obtain cancer likelihoods for each candidate. Features represent pharmacokinetic behavior, symmetry and appearance, among others. In both stages a random forest classifier is used to obtain cancer likelihoods. The system is evaluated on a large consecutive cohort of 347 patients with MR-guided biopsy as the reference standard. This set contained 165 patients with and 182 patients without prostate cancer. Performance evaluation is based on lesion-based FROC and patient-based ROC analysis. The system is also compared to the prospective clinical performance of radiologists. Results show a sensitivity of 0.42, 0.75 and 0.89 at 0.1, 1 and 10 false positives per normal case. Additionally, the patient-based ROC shows no significant difference at high specificity between the CAD system and

the radiologist.

In chapter 7 the independent combination of the CAD system and the radiologist is investigated with respect to the potential to improve radiologist performance and assess cancer aggressiveness. We obtained MRI studies and subsequent histopathologic outcome of MR-guided biopsies of a consecutive set of 130 patients from January 1st to September 1st 2013. Logistic regression combined CAD with the radiologist. Bootstrapping was used to analyze differences in sensitivity and specificity of the CAD/radiologist combination compared to the radiologist alone. Spearmans' rank correlation coefficient was used to assess correlation between CAD likelihood and cancer grade. Of all biopsies performed under MR-guidance, a total of 68% was positive and 32% was negative for prostate cancer. When detecting intermediate-to-high-grade cancer a significantly improved sensitivity was found for the CAD/radiologist combination relative to the radiologist alone (0.98 versus 0.93). Furthermore, a significant correlation was found for the likelihood output of the CAD/radiologist combination and cancer grade, with a rank correlation coefficient of 0.696.

# General discussion

9

We focused on three areas of automated image analysis for prostate cancer MRI: feature discovery, prostate segmentation, and development and evaluation of a computer-aided detection system. In this chapter the major contributions of this thesis are covered and suggestions for future work are given.

### Feature discovery

MRI is becoming an increasingly popular modality for the detection of prostate cancer due to its high sensitivity and specificity. However, results in literature on the performance of prostate MRI vary, caused by differences in acquisition, reader experience, and evaluation criteria. In the past couple of years guidelines on acquisition and reporting of prostate MRI have been published and evaluated; they show potential for standardized reporting of prostate MRI. The first release of these guidelines still lacks granularity and focuses mostly on the detection of any prostate cancer. Two important questions are not addressed: 'how do we discriminate different types of benign disease from prostate cancer?' and 'how can we better discriminate aggressive and indolent prostate cancer in vivo?'

Several other groups have made a start in answering these questions, although most are focused on single parameters or single types of benign disease. In chapter 4 we addressed the first question by combining histopathology and MRI to discover features which help discriminate prostate cancer from four types of benign disease: BPH, inflammation, PIN and atrophy. Multiple unique features for each of these classes were identified which had reasonable performance in discriminating the specific confounding class from prostate cancer. Some of these features may be explicitly incorporated into prostate MR reporting guidelines in the near future.

In chapters 5 and 7 we tried to address the second question of assessing cancer aggressiveness in vivo. Chapter 5 focused on the use of a single feature, the apparent diffusion coefficient, to determine cancer aggressiveness (based on Gleason grading) in the peripheral zone. In previous work it was discovered that the ADC correlates well with cancer grade. However, we discovered that this correlation can be further improved by explicitly taking into account inter-patient variation. In chapter 7 we correlated the output of the CAD system we developed to cancer aggression and found a strong correlation between cancer grade and CAD-generated likelihood.

Prospective evaluation of the features identified in chapter 4 and 5 would further help cement them as useful contributions to the reporting guidelines. Especially how radiologists should interpret the features and how to handle situations where there is uncertainty about more than one class are still unknowns. Expanding the reporting guidelines to also establish instructions for predicting cancer aggressiveness after evaluation of multi-parametric MRI is also an area of further research.

Feature discovery for confounding disease and cancer aggressiveness would benefit immensely from improved registration of histopathology and prostate MRI. The advent of digital pathology and the digitization of whole mount prostatectomy slides has opened the door to more accurate evaluation of MR parameters with respect to the underlying ground truth. The 3D reconstruction of complete prostatectomies and the subsequent mapping to the MRI could result in improved registration. Establishing artificial landmarks using strands and the use of ex vivo prostate MRI might be a path to achieve this. Finally, further automation of the registration process could help remove user variation and dependency.

MR spectroscopy is another method to, in vivo, determine cancer aggressiveness and has in the past obtained good results in this respect. In this thesis we did not investigate the use of spectroscopy, mostly because it is not acquired during the regular prostate cancer detection protocol in our hospital. In general the use of spectroscopy is complex, both on the acquisition side and the post-processing side, and as such its use in clinical practice is still limited. Nevertheless, it definitely has potential as a tool for cancer detection and aggressiveness assessment.

In addition to MR spectroscopy, several groups are investigating novel MRI acquisitions and modeling approaches to discover new features. One such an example is the use of bi-exponential diffusion models to assess slow and fast diffusion within tissue. Other groups have investigated the use of fractional anisotropy for DWI or more complex pharmacokinetic models for DCE MRI. Although there is definite value in more complex models and new acquisition strategies, results are still preliminary. They have not been evaluated in this thesis.

To truly assess cancer aggressiveness, the use of Gleason grading is not enough. Although pathology is often considered the ground truth, it is well known that the inter-observer variability between pathologists assigning Gleason grades is significant. As such there is only one reliable basis for establishing ground truth: final patient outcome, either biochemical recurrence (rise of PSA after treatment) or lack thereof after 5 - 20 years.

### Prostate segmentation

Chapter 2 discusses the standardization of prostate segmentation evaluation in MRI in the form of a 'Grand Challenge'. The concept of challenges has entered medical image analysis in 2007, and has become increasingly popular since. The standardized evaluation of algorithms is especially important in medical imaging, as most evaluation occurs on proprietary data sets, algorithm code is not made publicly available and evaluation measures differ. Furthermore, re-implementing methods from literature is prone to errors without help from the original author as most algorithms tend to be fairly complex, containing many parameters to optimize.

For any challenge to be successful, there are some prerequisites. First of all, the subject the challenge addresses, e.g. prostate segmentation, needs to be relevant, both to ensure enough participation and to generate interest. Second, the data provided in the challenge should have

similar characteristics to the data encountered in an actual clinical setting to make sure the results generalize well. Third, the evaluation measures used should be reasonable, represent a quality which is clinically relevant, and allow separation between competing algorithms. For the prostate MR segmentation challenge (PROMISE12) we tried to adhere to these rules.

The results obtained in the challenge show that several algorithms published in literature already obtain excellent performance relative to humans. Although none of them are as of yet able to automatically segment all the difficult cases correctly, in general the top algorithms will give good results. Depending on the use-case of the algorithm, performance might require no further improvement. For example, for volume estimation or as a pre-processing step for computer-aided diagnosis systems the methods are most likely accurate enough. For some use-cases however, further improvement of the segmentation algorithms for prostate MRI can be very helpful, e.g. automatically detecting extra-prostatic extension of a cancerous lesion. As we expect algorithm performance to be further improved, the challenge is currently still open for new submissions.

Most future work is related to optimizing details and expanding segmentation to the sub-structures of the prostate and across different MR parameters. Segmentation of the prostate sub-structures, for example the prostate zones (Chapter 3), can be very useful. It is for example well known that cancer appearance (and also the appearance and presence of different types of confounders) is prostate zone dependent. As such, for computer-aided detection algorithms the knowledge about the locations of the prostate zones can be very useful. One of the most difficult issues is that the visibility of the zones differs widely from patient to patient. Literature on the segmentation of prostate zones (let alone other sub-structures like the neuro-vascular bundles) is still sparse, and although a challenge on the segmentation of the prostate zones has already taken place, algorithm performance is still far from the performance of the whole prostate segmentation algorithms.

The segmentation of the neuro-vascular bundles and seminal vesicles are areas which are worthwhile to investigate. Patient prognosis for a large part depends on whether the cancer is still organ confined. When we have accurate segmentations of these sub-structures we can start to assess whether the tumor is invading these structures. This may help clinicians to make a more accurate prognosis.

The PROMISE12-challenge focused on the segmentation of the prostate on T2-weighted images as it contains the best anatomical detail (spatial resolution) and contrast (compared to diffusion-weighted imaging and dynamic contrast enhanced imaging). However, the use of the other MRI parameters might allow improvement of segmentation accuracy. In chapter 3 we already use both the ADC map and the axial T2-weighted image to segment the prostate zones, however this has not yet been applied to whole prostate segmentation. Therefore, multi-parametric segmentation of the prostate is still an important topic for future research.

### Development and evaluation of a computer-aided detection system for prostate cancer

Chapters 6 and 7 discuss the development and evaluation of a completely automated computer-aided detection system for prostate cancer on MRI. Although several research groups have investigated the use of CAD systems for prostate MRI, the system presented in this paper is only the second completely automated system incorporating both a detection and a characterization stage. Furthermore, the performance of the system presented in this thesis is better than the performances currently presented in literature. One caveat is that the performance measures have not been obtained on the same data set, which means there is room for a future 'Grand Challenge'.

The building blocks of the system are quite typical for CAD systems and consist of a segmentation of the organ of interest, extraction of voxel features, voxel classification, candidate detection, candidate segmentation and finally candidate classification. The two-stage approach to the system allows us to in the first stage detect all suspicious areas and in the second stage classify these areas, resulting in different and easier tasks for each of the stages. Chapter 6 also shows that the second stage of the system significantly improves the performance.

Although performance of the complete system is good, it is currently not as good as the radiologist, although at high specificity there is no significant difference. There are still some components which could be improved. The algorithm we used for the segmentation of the prostate obtained 7th place out of the 11 participating algorithms in the PROMISE12-challenge. Although the prostate segmentations themselves are quite reasonable, this means we are not currently using the optimal segmentation strategy, causing us to miss some cancers (around 3%). In the challenge active shape based segmentation approaches seemed to outperform atlas approaches. Therefore, replacing our atlas based strategy with a more successful active shape based strategy would be worthwhile. However, implementing such an algorithm is not trivial, and can be the focus of an entire PhD thesis by itself.

Standardization of MR image acquisition is difficult, even within a single institution. Variability between patients, scanners, and protocols make it challenging to create CAD systems which can quantitatively assess MRI. We already incorporate a way to standardize the T2-weighted images in our CAD system, however, also the supposedly quantitative parameters like the ADC differ substantially across different protocols. Furthermore, imaging artifacts, signal-to-noise ratio and lack of resolution can be problematic issues.

Another area of improvement is spatial alignment of the different multi-parametric MR images. Although the different MR images are implicitly registered because they are acquired sequentially (and thus within the same coordinate frame) patient and bowel movement can cause problems. In practice there were only a few cases in our database where there is substantial misregistration within the prostate itself. Co-registration of multiple distinct MR pa-

rameters is no trivial task and algorithm evaluation difficult. One approach could be to first register all the within-parameter images (b-values in DWI, time points in DCE and axial, sagittal and coronal T2-weighted acquisitions) and subsequently perform the between-parameter registration. For evaluation we could potentially use the performance of our CAD system as a surrogate marker for registration success.

Incorporation of new features like bi-exponential diffusion models or spectroscopy might also further improve performance. Additionally, the usefulness of texture features on T2-weighted imaging has not yet been investigated thoroughly. More algorithmic approaches to feature discovery like sparse coders might be useful in discovering improved texture descriptors.

In chapter 7 the CAD system was prospectively evaluated on a consecutive set of patients. We used logistic regression to create a model which combines the radiologist and the CAD system independently. We showed the potential of the CAD system to improve the assessment of the radiologist and found a large correlation between CAD system likelihoods and cancer grade.

Evaluating the CAD system in this way allows us to obtain results that are closer to clinical practice than typical observer studies, as there was no patient selection and the radiologists reported prospectively without knowledge of the outcome. However, this evaluation has its own drawbacks. We only use locations that were actually biopsied and as such the locations which were given a low score by the radiologist are not assessed by the CAD system. In the more controlled setting of an observer study we could have used patients with subsequent prostatectomy to obtain ground truth for the entire prostate. Furthermore, observer studies would have allowed us to assess different usage scenarios of the CAD system (interactive, before/after radiologist scoring) instead of the independent combination that was used now.

CAD systems can theoretically be used as independent readers, either as a first reader or a second reader. As a first reader they could be used as a triage test, e.g. by letting them operate at very high sensitivity (e.g. 99%) and moderate specificity (e.g. 40%). This would reduce the amount of cases that need to be read by a radiologist substantially. The first reader strategy for CAD systems is especially useful in a screening setting, where there are many more healthy men than men with prostate cancer. Currently, several groups are investigating the use of pre-biopsy MRI, i.e. MRI after only an initial PSA test. If such data would become available it would be of great interest to investigate our current CAD system as a first reader.

Last, the data used in this thesis to evaluate the CAD system all originated from a single center, although it does consist of multi-scanner, multi-protocol data. The evaluation on multi-center data would be a logical next step to assess the generalization performance of the CAD system.

## Epilogue

Rising healthcare costs will be the major healthcare problem in the coming decades. As the the average age in the Netherlands (and the world) continuous to rise, more people will put a strain on the healthcare system. Especially for prostate cancer, for which incidence is highly correlated to age, offering suitable diagnosis and treatment to everyone will become more and more difficult. To add insult to injury the current diagnostic pathway for the diagnosis of prostate cancer is invasive, inaccurate, and causes a significant amount of patient morbidity. This currently makes it infeasible to perform screening for prostate cancer, which is unfortunate as early detection of aggressive prostate cancer could lead to earlier (and usually cheaper and less radical) treatment with better outcome for the patient.

With rising healthcare costs MRI might not seem like an ideal solution for the detection and diagnosis of prostate cancer as it is commonly viewed as an expensive modality. However, a recent study by de Rooij et al.[64] has already shown that even in their proposed setting (including a full multi-parametric MRI and an MR-guided biopsy) cost of MRI is almost the same as the cost of PSA and TRUS-guided biopsies. The cost per quality-of-life-year is even lower, mostly because of the reduced number of side effects when using MRI. The main reason for not performing MRI instead of TRUS-guided biopsies is the availability of MRI and the lack of experienced radiologists.

This thesis does not try to answer the issue of MRI availability. However, some general comments can be made. In the Netherlands, the availability of MRI is quite high, with 10.8 MR scanners per 1.000.000 inhabitants, resulting in approximately 180 MR scanners (Brancherapport Nederlandse Vereniging van Ziekenhuizen 2012). In the Dutch population around 1.000.000 men would need to be screened according to the suggestions in Schröder et al.[21]. Performing, for example, a biennial screening of this group using PSA upfront and subsequent MRI (with a PSA threshold of 1, giving us a sensitivity after PSA of 92% for aggressive cancer) would result in 3 - 4 scan per day, per scanner. Using a PSA threshold of 4 (sensitivity for aggressive cancer 40%) would result in 1 scan per day, per scanner. Although this still seems like a lot, it might be feasible when the multi-parametric MRI protocol is modified a bit. For screening only acquiring T2-weighted imaging and diffusion-weighted imaging might be enough to achieve good sensitivity and specificity and would reduce MRI acquisition time, cost and the dependence on contrast agent. With the current protocol acquiring a sagittal and axial T2-weighted image in addition to a diffusion-weighted series including 3 b-values can be performed within 10 minutes. Additionally, new techniques like compressed sensing might revolutionize acquisition speed of MRI in the near future.

The lack of experienced radiologists to read prostate MRI can be partly mediated by training, but still the volume and time required to read all acquired MRIs, perhaps in a double reading setting like in mammography screening, would put a large strain on the radiological

community. This is one of the main problems this thesis tries to address. The current implementation of CAD systems mostly focuses on being an aid to the radiologist, but the potential of computerized tools is not restricted to that. Their use as independent readers may have a larger impact in the long run. This does not mean that the radiologists will be out of a job; it will give them more time so they can focus on the patients that actually need their expertise. To give an example, in the screening setting as suggested in the previous paragraph (PSA $>$ 1), around 250000 MRIs are made each year, of which 85% will be normal. These are not the cases the clinician wants to spend a lot of their time on. If we can design a computerized system which can, at a very high sensitivity, get rid of a substantial percentage of normal cases that would already result in improved efficiency. In the case of prostate MRI, it might make screening using MRI much more feasible. In this thesis we could not yet evaluate the CAD system in this way because we do not yet have MRI screening data, however this would be the ultimate goal.

Of course there are also some ethical considerations to screening for prostate cancer: every screening program has to find a balance between costs and benefits. The current diagnostic pathway already has shown the potential benefits (reduction of prostate cancer mortality), but at too high a cost, mostly in terms of over-diagnosis and over-treatment. The PSA/MRI pathway might offer similar benefits, but the straightforward implementation (all MRIs reported by clinicians) puts too big a strain on the radiological community. Although there are still many unanswered question about how to best use computer aids in screening (both from a practical and a legislative point of view), it does have the potential to make PSA/MRI-based screening for prostate cancer a reality. For me the biggest gain is the fact that we can reduce cancer mortality and simultaneously provide healthy men with a clear yes/no-answer regarding the presence of prostate cancer, reducing anxiety and morbidity relative to the current situation. Maybe the days of patients suffering through more than 45 biopsy needles will then finally be behind us.

# Appendices

# Samenvatting

De focus van deze thesis was het ontwikkelen van een computer-ondersteund detectiesysteem (CAD-systeem) voor het vinden van kanker in prostaat MRI. De benodigde bouwstenen hiervoor zijn het segmenteren van relevante structuren (hoofdstuk 2 en 3), het ontdekken van kenmerken van kanker en benigne ziektepatronen (hoofdstuk 4 en 5), het ontwikkelen van het CAD-systeem (hoofdstuk 6) en de prospectieve evaluatie van het systeem (hoofdstuk 7).

In hoofdstuk 2 werd de organisatie van een internationale wedstrijd voor prostaatsegmentatie op MRI besproken. Deze wedstrijd werd georganiseerd om een eerlijke en betekenisvolle evaluatie van segmentatiealgoritmen mogelijk te maken. Het ontwerp van de wedstrijd hield onder andere in dat er multi-instelling, multi-fabrikant en multi-protocol data verzameld moest worden. Daarnaast moest er gebruik worden gemaakt van evaluatiemethodieken die een duidelijk onderscheid tussen algoritmen lieten zien op basis van accuratesse. De wedstrijd bestond uit een online component en een live component, die laatste werd georganiseerd tijdens een workshop op de MICCAI2012-conferentie in Nice, Frankrijk. In totaal 11 teams deden mee in deze initiële fase, met ieder unieke algoritmen. De algoritmen bestonden onder andere uit active-shapemodellen en multi-atlas-registratiemethodieken. De resultaten laten zien dat de twee beste algoritmes significant beter zijn dan alle andere algoritmes buiten de top drie. Daarbij hadden zij ook een efficiënte implementatie met een segmentatietijd van 8 minuten en 3 seconden per casus. Gemiddeld gezien waren active-shapemodellen beter dan de multi-atlas-registratiemethoden, zowel in prestaties als in rekentijd. Alhoewel de gemiddelde prestaties goed tot uitstekend waren en het Imorphics-algoritme beter presteerde dan de onervaren beoordelaar hebben we ook laten zien dat combinaties van algoritmen tot verdere verbetering zou kunnen leiden. Dit laat zien dat de optimale prestatie in prostaatsegmentatie nog niet gehaald is.

Segmentatie van de prostaatzones is een relatief nieuw onderzoeksveld en de initiële resultaten verkregen via het gebruik van een patroonherkenningsaanpak werden getoond in hoofdstuk 3. Automatische verdeling van de prostaat in de centrale klier en de perifere zone is een zeer bruikbaar gereedschap voor computer-ondersteunde detectie van prostaatkanker omdat de prevalentie en de karakteristieken van kanker in beiden zones substantieel verschillen. De patroonherkenningsaanpak gebruikt drie typen kenmerken om de twee zones uit elkaar te houden: anatomie, intensiteit en textuur. Deze methode werd vergeleken met een multi-atlas-registratietechniek die gebruik maakte van 48 multi-parametrische prostaatstudies. Drie beoordelaars werden ingezet om de inter-beoordelaarvariabiliteit af te schatten en we vergelijken de resultaten met de beste resultaten uit de literatuur. We verkregen een gemiddelde Dice coefficiënt van 0.89 voor de centrale klier en 0.75 voor de perifere zone, vergeleken met 0.87 en 0.76 in de literatuur. Samenvattend, de patroonherkenningsaanpak die anatomie, intensiteit en textuur gebruikt geeft goede resultaten in de zonale segmentatie van de prostaat.

Kenmerken die onderscheidend zijn voor verschillende benigne ziekten en prostaatkanker werden beschreven in hoofdstuk 4. De aanwezigheid van benigne ziekten in de prostaat is

verwarrend voor het diagnosticeren van prostaatkanker. De meest voorkomende typen zijn benigne prostatische hyperplasie (BPH), atrofie, ontsteking en prostatische, intra-epithele neoplasie (PIN). Om vast te stellen wat de beeldkenmerken zijn van deze klassen gebruiken we fusie van MRI en histopathologie en computer-geëxtraheerde kenmerken in combinatie met classificatie. De prostatectomie en de preoperatieve, multi-parametrische MRI van 70 patiënten werden geïncludeerd in deze studie. Intensiteit, textuur en farmacokinetische kenmerken werden geëxtraheerd voor ieder van de benigne klassen en prostaatkanker. Kenmerkselectie werd uitgevoerd voor elke classificatietaak om te bepalen wat de vijf belangrijkste kenmerken waren voor ieder van de benigne klassen. In totaal 92 PIN-, 64 atrofie-, 120 ontstekings- en 73 BPH-laesies werden geannoteerd en daarbij nog 128 prostaatkankerlaesies. Voor elke classificatietaak was het meest belangrijke kenmerk verschillend ten opzichte van de andere taken en elke benigne klasse had verschillende unieke kenmerken. Daarbij verbeterde de oppervlakte onder de 'receiver-operating characteristic' (ROC) curve significant voor elke taak vergeleken met een monolithische classificatie (waarbij alle benigne ziekten als een enkele klasse beschouwd werden).

In hoofdstuk 5 testten we of we het potentieel van de 'apparent diffusion coefficient' (ADC) voor het bepalen van kankeragressiviteit kunnen verbeteren door expliciet de inter-patiënt variabiliteit in de normale perifere zone mee te nemen. Intra- en inter-patiënt variatie van de ADC waarden in de perifere zone werden vastgesteld door middel van herhaalde metingen van de ADC in de normale perifere zone in een retrospectieve cohort van 10 patiënten met drie verschillende MRI sessie op 3 Tesla. Het effect van deze intra- en inter-patiënt variabiliteit op het bepalen van kankeragressiviteit werd bepaald in een tweede cohort van 51 patiënten met prostaatkanker die een MRI ondergingen voor de prostatectomie. Logistische regressie werd gebruikt om te bepalen of het toevoegen van normale ADC waarde de voorspelling van kankeragressiviteit zou kunnen verbeteren. Het effect op de diagnostische prestaties werd bepaald door middel van ROC-analyse. De herhaalde-metingen-ANOVA liet zien dat de inter-patiënt variatie significant hoger is dan de metingvariatie. Analyse van de tumor-ADC-waarden liet een oppervlakte onder de ROC curve zien van 0.91, na het toevoegen van de normale ADC-waarden verbeterde deze significant naar 0.96.

De ontwikkeling van het twee-stadia computer-ondersteund detectiesysteem werd besproken in hoofdstuk 6. In het eerste stadium worden initiële kandidaten gedetecteerd door het gebruik van prostaatsegmentatie, voxelkenmerken, classificatie en lokale-maximadetectie. Het tweede stadium bestaat uit segmentatie van de kandidaten en het verkrijgen van een kankerwaarschijnlijkheid door middel van classificatie. Kenmerken die gebruikt werden zijn onder andere beeldintensiteit, farmacokinetisch gedrag en symmetrie. In beiden stadia wordt een Random-Forestclassificatie gebruikt om kankerwaarschijnlijkheden te berekenen. Het systeem is geëvalueerd op een grote continue cohort van 347 patiënten met MR-geleide biopten als de referentiestandaard. Deze set bevatte 165 patiënten met prostaatkanker en 182 zon-

der prostaatkanker. De evaluatie werd gedaan door laesie-gebasseerde 'free-response receiver operating characteristic' (FROC-)analyse en patiënt-gebasseerde ROC-analyse. Het systeem werd ook vergeleken met de prospectieve klinische prestaties van de radioloog. Resultaten laten zien dat bij een sensitiviteit van 0.45, 0.75 en 0.89 er 0.1, 1 en 10 foutpositieven per normale casus zijn. Daarbij laat de patiënt-gebasseerde ROC analyse zien dat bij hoge specificiteit het systeem niet significant verschilt van de radioloog.

In hoofdstuk 7 onderzochten we de onafhankelijke combinatie van het systeem en de radioloog. Het doel is het potentieel van het systeem te ontdekken met betrekking tot het reduceren van het aantal biopten en bepalen van kankeragressiviteit. We hebben hiervoor MRI-studies en MR-geleide bioptuitkomsten van een continue set van 130 patiënten tussen 1 januari en 1 september 2013 gebruikt. Logistische regressie werd toegepast om de radioloog en het systeem te combineren. Daarna werd bootstrapping ingezet om de mogelijke verbetering in sensitiviteit en specificiteit te bepalen van de systeem/radioloog combinatie ten opzichte van de radioloog alleen. Spearman's rangcorrelatiecoëfficiënt werd gebruikt om de correlatie tussen de CAD-waarschijnlijkheid en kankergraad te bepalen. Van alle biopten waren er 68% positief voor kanker en de rest negatief. Bij het detecteren van gemiddeld-tot-hoog-gradige kankers kunnen we een significant betere sensitiviteit bereiken bij een score van PIRADS 4 (0.93 naar 0.98). Daarnaast werd er een significante correlatie gevonden voor de waarschijnlijkheidsuitkomsten van de systeem/radioloogcombinatie en kankergraad, met een rangcorrelatiecoëfficiënt van 0.696.

# Publications

### Papers in international journals

**G. Litjens**, N. Shih, R. Elliott, T. Kobus, C. Hulsbergen - van de Kaa, H. Huisman, A. Madabhushi. Computer-extracted features can distinguish benign confounding disease from prostatic adenocarcinoma on multi-parametric MRI. *Submitted to Radiology*

**G. Litjens**, J. Barentsz, N. Karssemeijer and H. Huisman. Clinical evaluation of a computer-aided diagnosis system for determining cancer aggressiveness in prostate MRI. *Submitted to European Radiology*

L. Bittencourt, **G. Litjens**, J. Barentsz and E. Gasparetto. Performance of the ESUR/PI-RADS criteria for characterizing extraprostatic extension of prostate cancer using 3.0 T multiparametric magnetic resonance imaging. *Submitted to Radiology*

E. Vos, T. Kobus, **G. Litjens**, T. Hambrock, C. Hulsbergen - van de Kaa, J. Barentsz, M. Maas, T. Scheenen. Multiparametric MRI for discriminating low-aggressive from high-aggressive prostate cancer: an active surveillance tool? *Submitted to Investigative Radiology*

**G. Litjens**, H. Huisman, R. Elliott, N. Shih M. Feldman, S. Viswanath, J. Fütterer, J. Bomers and A. Madabhushi. Quantitative identification of MRI features of prostate cancer response following laser ablation and radical prostatectomy. *Journal of Medical Imaging*, 1(3), 2014.

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

**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. 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*, 18(2):359–373, 2014.

E. Vos, **G. Litjens**, T. Kobus, T. Hambrock, C. Hulsbergen - van de Kaa, J. Barentsz, H. Huisman and T. Scheenen. Assessment of Prostate Cancer Aggressiveness Using Dynamic Contrast-enhanced Magnetic Resonance Imaging at 3T. *European Urology*, 64(1):448–455, 2013.

K. Nagel, M. Schouten, T. Hambrock, **G. Litjens**, C. Hoeks, B. Haken, J. Barentsz and J. Fütterer. Differentiation of Prostatitis and Prostate Cancer by Using Diffusion-weighted MR Imaging and MR-guided Biopsy at 3T. *Radiology*, 267(1):164–172, 2013.

**G. Litjens**, T. Hambrock, C. Hulsbergen - van de Kaa, J. Barentsz and H. Huisman. The effect of inter-patient normal peripheral zone Apparent Diffusion Coefficient variation on the Prediction of Prostate Cancer Aggressiveness. *Radiology*, 265(1):260–266, 2012.

O. Debats, **G. Litjens**, J. Barentsz, N. Karssemeijer and H. Huisman. Automated 3-Dimensional Segmentation of Pelvic Lymph Nodes in Magnetic Resonance Images. *Medical Physics*, 38(11): 6178–6187, 2011.

### Papers in conference proceedings

**G. Litjens**, R. Elliott, N. Shih, M. Feldman, J.O. Barentsz, C. Hulsbergen - van de Kaa, I. Kovacs, H. Huisman and A. Madabhushi. Distinguishing prostate cancer from benign confounders via a cascaded classifier on multi-parametric MRI. In *Medical Imaging, volume 9035 of Proceedings of the SPIE*, pages 903512–903512-14, 2014.

**G. Litjens**, H. Huisman, R. Elliott, N. Shih, M. Feldman, S. Viswanath, J. Fütterer, J. Bomers and A. Madabhushi. Distinguishing benign confounding treatment changes from residual prostate cancer on MRI following laser ablation using feature scoring. In *Medical Imaging, volume 9036 of Proceedings of the SPIE*, pages 90361D–90361D-12, 2014.

**G. Litjens**, O. Debats, W. van de Ven, N. Karssemeijer and H. Huisman. A pattern recognition approach to zonal segmentation of the prostate on MRI. In *Medical Image Computing and Computer-Assisted Intervention, volume 7511 of Lecture Notes in Computer Science*, pages 413–420, 2012.

**G. Litjens**, J. Barentsz, N. Karssemeijer and H. Huisman. Automated computer-aided detection of prostate cancer in MR images: from a whole-organ to a zone-based approach. In *Medical Imaging, volume 8315 of Proceedings of the SPIE*, pages 83150G–83150G-6, 2012.

**G. Litjens**, P. Vos and J. Barentsz, N. Karssemeijer and H. Huisman. Automatic Computer Aided Detection of Abnormalities in Multi-Parametric Prostate MRI. In *Medical Imaging, volume 7963 of Proceedings of the SPIE*, pages 79630T-79630T-7, 2011.

**G. Litjens**, L. Hogeweg, A. Schilham, P. de Jong, M. Viergever and B. van Ginneken. Simulation of nodules and diffuse infiltrates in chest radiographs using CT templates. In *Medical Image Computing and Computer-Assisted Intervention, volume 6362 of Lecture Notes in Computer Science*, pages 3960–403, 2010.

**G. Litjens**, M. Heisen, J. Buurman, B. ter Haar - Romeny. Pharmacokinetic models in clinical practice: what model to use for DCE-MRI of the breast? In *International Symposium on Biomedical Imaging*, pages 185–188, 2010

### Abstracts in conference proceedings

**G. Litjens**, N. Karssemeijer, J. Barentsz and H. Huisman. Computer-aided detection of prostate cancer in multi-parametric magnetic resonance imaging. In *Annual Meeting of the Radiological*

*Society of North America*, 2014.

E. Vos, T. Kobus, **G. Litjens**, T. Hambrock, C. Hulsbergen - van de Kaa, M. Maas and T. Scheenen. Multiparametric MR imaging for the assessment of prostate cancer aggressiveness at 3 Tesla. In *Proceedings of the International Society for Magnetic Resonance in Medicine*, 2014.

**G. Litjens**, N. Karssemeijer, J. Barentsz and H. Huisman. Initial prospective evaluation of the prostate imaging reporting and data standard (PI-RADS): Can it reduce unnecessary MR guided biopsies?. In *Annual Meeting of the Radiological Society of North America*, 2013.

M. Maas, M. Koopman, **G. Litjens**, A. Wright, K. Selnaes, I. Gribbestad, M. Haider, K. Macura, D. Margolis, B. Kiefer, J. Fütterer and T. Scheenen. Prostate Cancer localization with a Multiparametric MR Approach (PCaMAP): initial results of a multi-center study. In *Proceedings of the International Society for Magnetic Resonance in Medicine*, 2013

E. Vos, **G. Litjens**, T. Kobus, T. Hambrock, C. Hulsbergen - van de Kaa, H. Huisman and T. Scheenen. Dynamic contrast enhanced MR imaging for the assessment of prostate cancer aggressiveness at 3T. In *Proceedings of the International Society for Magnetic Resonance in Medicine*, 2012

**G. Litjens**, J. Barentsz, N. Karssemeijer and H. Huisman. Computerized characterization of central gland lesions using texture and relaxation features from T2-weighted prostate MRI. In *Annual Meeting of the Radiological Society of North America*, 2012.

J. Bomers, L. Bittencourt, C. Hoeks, E. Hamoen, M. de Rooij, **G. Litjens**, A. Froemming, J. Fütterer and J. Barentsz. Standardization of Multiparametric Prostate MRI: PI-RADS. In *Annual Meeting of the Radiological Society of North America: Education Exhibits*, 2012.

M. Hicks, **G. Litjens**, N. Karssemeijer, J. Barentsz and H. Huisman. Prostate MR Workstation: An Integrated Work-flow for Case Reading, PI-RADS Structured Reporting, Automatic Pharmacokinetic Modelling and Quantitative Analysis. In *Annual Meeting of the Radiological Society of North America: Quantitative Imaging Reading Room Showcase*, 2012.

**G. Litjens**, N. Karssemeijer and H. Huisman. A multi-atlas approach for prostate segmentation in MRI. In *MICCAI Workshop: Prostate Cancer Imaging: The PROMISE12 Prostate Segmentation Challenge*, 2012.

**G. Litjens**, J. Barentsz, N. Karssemeijer and H. Huisman. Zone-specific Automatic Computer-aided Detection of Prostate Cancer in MRI. In *Annual Meeting of the Radiological Society of North America*, 2011.

O. Debats, T. Hambrock, **G. Litjens**, H. Huisman and J. Barentsz. Detection of Lymph Node

Metastases with Ferumoxtran-10 vs Ferumoxytol. In *Annual Meeting of the Radiological Society of North America*, 2011.

M. Schouten, K. Nagel, T. Hambrock, C. Hoeks, **G. Litjens**, J. Barentsz and J. Fütterer. Differentiation of Normal Prostate Tissue, Prostatitis, and Prostate Cancer: Correlation between Diffusion-weighted Imaging and MR-guided Biopsy. In *Annual Meeting of the Radiological Society of North America*, 2011.

W. van de Ven, **G. Litjens**, J. Barentsz, T. Hambrock and H. Huisman. Required accuracy of MR-US registration for prostate biopsies. In *MICCAI Workshop: Prostate Cancer Imaging. Image Analysis and Image-Guided Interventions*, 2011

H. Huisman, P. Vos, **G. Litjens**, T. Hambrock and J. Barentsz. Computer aided detection of prostate cancer using T2W, DWI and DCE-MRI: methods and clinical applications. In *MICCAI Workshop: Prostate Cancer Imaging: Computer Aided Diagnosis, Prognosis, and Intervention*, 2010

P. Snoeren, **G. Litjens**, B. van Ginneken and N. Karssemeijer. Training a Computer Aided Detection System with Simulated Lung Nodules in Chest Radiographs. In *The Third International Workshop on Pulmonary Image Analysis*, pages 139–149, 2011.

**G. Litjens**, M. Heisen, J. Buurman, A. Wood, M. Medved, G. Karczmar and B. ter Haar - Romeny. T1 Quantification: Variable Flip Angle Method vs Use of Reference Phantom. In *Annual Meeting of the Radiological Society of North America*, 2009.

## Awards

MICCAI 2012 Student Travel Award

SPIE Medical Imaging 2014 Robert Wagner Best Student Paper Award - Runner up

# Bibliography

[1] J. E. McNeal. The zonal anatomy of the prostate. *Prostate*, 2:35–49, 1981.

[2] J. E. McNeal, E. A. Redwine, F. S. Freiha, and T. A. Stamey. Zonal distribution of prostatic adenocarcinoma. correlation with histologic pattern and direction of spread. *Am J Surg Pathol*, 12:897–906, 1988.

[3] A. M. De Marzo, E. A. Platz, S. Sutcliffe, J. Xu, H. Grönberg, C. G. Drake, Y. Nakai, W. B. Isaacs, and W. G. Nelson. Inflammation in prostate carcinogenesis. *Nat Rev Cancer*, 7: 256–269, 2007.

[4] A. Jemal, F. Bray, M. M. Center, J. Ferlay, E. Ward, and D. Forman. Global cancer statistics. *CA Cancer J Clin*, 61:69–90, 2011.

[5] N. B. Delongchamps, A. Singh, and G. P. Haas. The role of prevalence in the diagnosis of prostate cancer. *Cancer Control*, 13:158–168, 2006.

[6] M. Sánchez-Chapado, G. Olmedilla, M. Cabeza, E. Donat, and A. Ruiz. Prevalence of prostate cancer and prostatic intraepithelial neoplasia in Caucasian Mediterranean males: an autopsy study. *Prostate*, 54:238–247, 2003.

[7] G. A. Dijkman and F. M. Debruyne. Epidemiology of prostate cancer. *Eur Urol*, 30: 281–295, 1996.

[8] K. Mistry and G. Cable. Meta-analysis of prostate-specific antigen and digital rectal examination as screening tests for prostate carcinoma. *J Am Board Fam Pract*, 16:95–101, 2003.

[9] P. T. Scardino, R. Weaver, and M. A. Hudson. Early detection of prostate cancer. *Hum Pathol*, 23:211–222, 1992.

[10] D. F. Gleason. Classification of prostatic carcinomas. *Cancer Chemother Rep*, 50:125–128, 1966.

[11] J. I. Epstein, W. C. Allsbrook, M. B. Amin, L. L. Egevad, and ISUP Grading Committee. The 2005 International Society of Urological Pathology (ISUP) consensus conference on Gleason grading of prostatic carcinoma. *Am J Surg Pathol*, 29:1228–1242, 2005.

[12] M. Tannenbaum. *Urologic pathology: The prostate.* Lea & Febiger, 1977. ISBN 081210546X.

[13] M. Borre, B. Nerstrøm, and J. Overgaard. The natural history of prostate carcinoma based on a Danish population treated with no intent to cure. *Cancer*, 80:917–928, 1997.

[14] J. Hugosson and G. Aus. Natural course of localized prostate cancer. a personal view with a review of published papers. *Anticancer Res*, 17:1441–1448, 1997.

[15] P. C. Albertsen, J. A. Hanley, and J. Fine. 20-year outcomes following conservative management of clinically localized prostate cancer. *JAMA*, 293:2095–2101, 2005.

[16] L. Klotz, L. Zhang, A. Lam, R. Nam, A. Mamedov, and A. Loblaw. Clinical results of long-term follow-up of a large, active surveillance cohort with localized prostate cancer. *J Clin Oncol*, 28:126–131, 2010.

[17] J. D. Hall, J. C. Boyd, M. C. Lippert, and D. Theodorescu. Why patients choose prostatectomy or brachytherapy for localized prostate cancer: results of a descriptive survey. *Urology*, 61:402–407, 2003.

[18] J. Hegarty, P. V. Beirne, E. Walsh, H. Comber, T. Fitzgerald, and M. Wallace Kazer. Radical prostatectomy versus watchful waiting for prostate cancer. *Cochrane Database Syst Rev*, page CD006590, 2010.

[19] F. Peinemann, U. Grouven, L. G. Hemkens, C. Bartel, H. Borchers, M. Pinkawa, A. Heidenreich, and S. Sauerland. Low-dose rate brachytherapy for men with localized prostate cancer. *Cochrane Database Syst Rev*, page CD008871, 2011.

[20] L. Budäus, M. Bolla, A. Bossi, C. Cozzarini, J. Crook, A. Widmark, and T. Wiegel. Functional outcomes and complications following radiation therapy for prostate cancer: a critical analysis of the literature. *Eur Urol*, 61:112–127, 2012.

[21] F. H. Schröder, J. Hugosson, M. J. Roobol, T. L. J. Tammela, S. Ciatto, V. Nelen, M. Kwiatkowski, M. Lujan, H. Lilja, M. Zappa, L. J. Denis, F. Recker, A. Berenguer, L. Määttänen, C. H. Bangma, G. Aus, A. Villers, X. Rebillard, T. van der Kwast, B. G. Blijenberg, S. M. Moss, H. J. de Koning, A. Auvinen, and the ERSPC Investigators. Screening and prostate-cancer mortality in a randomized European study. *N Engl J Med*, 360: 1320–1328, 2009.

[22] F. H. Schröder, J. Hugosson, M. J. Roobol, T. L. J. Tammela, S. Ciatto, V. Nelen, M. Kwiatkowski, M. Lujan, H. Lilja, M. Zappa, L. J. Denis, F. Recker, A. Páez, L. Määttänen, C. H. Bangma, G. Aus, S. Carlsson, A. Villers, X. Rebillard, T. van der Kwast, P. M. Kujala, B. G. Blijenberg, U.-H. Stenman, A. Huber, K. Taari, M. Hakama, S. M. Moss, H. J. de Koning, A. Auvinen, and the ERSPC Investigators. Prostate-cancer mortality at 11 years of follow-up. *N Engl J Med*, 366:981–990, 2012.

[23] A. M. D. Wolf, R. C. Wender, R. B. Etzioni, I. M. Thompson, A. V. D'Amico, R. J. Volk, D. D. Brooks, C. Dash, I. Guessous, K. Andrews, C. DeSantis, R. A. Smith, and A. C. S. P.

C. A. Committee. American Cancer Society guideline for the early detection of prostate cancer: update 2010. *CA Cancer J Clin*, 60:70–98, 2010.

[24] M. K. Terris. Sensitivity and specificity of sextant biopsies in the detection of prostate cancer: preliminary report. *Urology*, 54:486–489, 1999.

[25] B. Djavan, V. Ravery, A. Zlotta, P. Dobronski, M. Dobrovits, M. Fakhari, C. Seitz, M. Susani, A. Borkowski, L. Boccon-Gibod, C. C. Schulman, and M. Marberger. Prospective evaluation of prostate cancer detected on biopsies 1, 2, 3 and 4: when should we stop? *J Urol*, 166:1679–1683, 2001.

[26] C. M. A. Hoeks, M. G. Schouten, J. G. R. Bomers, S. P. Hoogendoorn, C. A. Hulsbergen-van de Kaa, T. Hambrock, H. Vergunst, J. P. M. Sedelaar, J. J. Fütterer, and J. O. Barentsz. Three-Tesla magnetic resonance-guided prostate biopsy in men with increased prostate-specific antigen and repeated, negative, random, systematic, transrectal ultrasound biopsies: Detection of clinically significant prostate cancers. *Eur Urol*, 62:902–909, 2012.

[27] M. Noguchi, T. A. Stamey, J. E. McNeal, and C. M. Yemoto. Relationship between systematic biopsies and histological features of 222 radical prostatectomy specimens: lack of prediction of tumor significance for men with nonpalpable prostate cancer. *J Urol*, 166:104–109, 2001.

[28] T. Hambrock, C. Hoeks, C. Hulsbergen-van de Kaa, T. Scheenen, J. Fütterer, S. Bouwense, I. van Oort, F. Schröder, H. Huisman, and J. Barentsz. Prospective assessment of prostate cancer aggressiveness using 3-T diffusion-weighted magnetic resonance imaging-guided biopsies versus a systematic 10-core transrectal ultrasound prostate biopsy cohort. *Eur Urol*, 61:177–184, 2012.

[29] P. Lauterbur. Image formation by induced local interactions: Examples employing nuclear magnetic resonance. *Nature*, 242:190–191, 1973.

[30] H. Hricak, G. C. Dooms, J. E. McNeal, A. S. Mark, M. Marotti, A. Avallone, M. Pelzer, E. C. Proctor, and E. A. Tanagho. MR imaging of the prostate gland: normal anatomy. *AJR Am J Roentgenol*, 148:51–58, 1987.

[31] A. Tanimoto, J. Nakashima, H. Kohno, H. Shinmoto, and S. Kuribayashi. Prostate cancer screening: The clinical value of diffusion-weighted imaging and dynamic MR imaging in combination with T2-weighted imaging. *J Magn Reson Imaging*, 25:146–152, 2007.

[32] K. Kitajima, Y. Kaji, Y. Fukabori, K. Yoshida, N. Suganuma, and K. Sugimura. Prostate cancer detection with 3 T MRI: Comparison of diffusion-weighted imaging and dynamic

contrast-enhanced MRI in combination with T2-weighted imaging. *J Magn Reson Imaging*, 31:625–631, 2010.

[33] J. O. Barentsz, J. Richenberg, R. Clements, P. Choyke, S. Verma, G. Villeirs, O. Rouvière, V. Logager, J. J. Fütterer, and European Society of Urogenital Radiology. ESUR prostate MR guidelines 2012. *Eur Radiol*, 22:746–757, 2012.

[34] F. V. Coakley and H. Hricak. Radiologic anotomy of the prostate: a clinical approach. *Radiol Clin North Am*, 38:15–30, 2000.

[35] P. S. Tofts, G. Brix, D. L. Buckley, J. L. Evelhoch, E. Henderson, M. V. Knopp, H. B. Larsson, T. Y. Lee, N. A. Mayr, G. J. Parker, R. E. Port, J. Taylor, and R. M. Weisskoff. Estimating kinetic parameters from dynamic contrast-enhanced t(1)-weighted MRI of a diffusable tracer: standardized quantities and symbols. *J Magn Reson Imaging*, 10: 223–232, 1999.

[36] G. Brix, W. Semmler, R. Port, L. R. Schad, G. Layer, and W. J. Lorenz. Pharmacokinetic parameters in cns gd-dtpa enhanced MR imaging. *J Comput Assist Tomogr*, 15:621–628, 1991.

[37] G. Litjens, M. Heisen, J. Buurman, and B. ter Haar Romeny. Pharmacokinetic models in clinical practice: what model to use for DCE-MRI of the breast? In *IEEE Int Symp Biomedical Imaging*, pages 185–188, 2010.

[38] P. Gibbs, D. J. Tozer, G. P. Liney, and L. W. Turnbull. Comparison of quantitative t2 mapping and diffusion-weighted imaging in the normal and pathologic prostate. *Magn Reson Med*, 46:1054–1058, 2001.

[39] R. Bammer. Basic principles of diffusion-weighted imaging. *Eur J Radiol*, 45:169–184, 2003.

[40] E. O. Stejskal and J. E. Tanner. Spin diffusion measurements: Spin echoes in the presence of a time-dependent field gradient. *The Journal of Chemical Physics*, 42:288, 1965.

[41] M. Uhl, C. Altehoefer, U. Kontny, K. Il'yasov, M. Büchert, and M. Langer. MRI-diffusion imaging of neuroblastomas: first results and correlation to histology. *Eur Radiol*, 12: 2335–2338, 2002.

[42] G. M. Villeirs, W. Oosterlinck, E. Vanherreweghe, and G. O. De Meerleer. A qualitative approach to combined magnetic resonance imaging and spectroscopy in the diagnosis of prostate cancer. *Eur J Radiol*, 73:352–356, 2010.

[43] N. B. Delongchamps, M. Rouanne, T. Flam, F. Beuvon, M. Liberatore, M. Zerbib, and F. Cornud. Multiparametric magnetic resonance imaging for the detection and localization of prostate cancer: combination of T2-weighted, dynamic contrast-enhanced and diffusion-weighted imaging. *BJU Int*, 107:1411–1418, 2011.

[44] C. M. A. Hoeks, J. O. Barentsz, T. Hambrock, D. Yakar, D. M. Somford, S. W. T. P. J. Heijmink, T. W. J. Scheenen, P. C. Vos, H. Huisman, I. M. van Oort, J. A. Witjes, A. Heerschap, and J. J. Fütterer. Prostate cancer: Multiparametric MR imaging for detection, localization, and staging. *Radiology*, 261:46–66, 2011.

[45] P. Kozlowski, S. D. Chang, E. C. Jones, K. W. Berean, H. Chen, and S. L. Goldenberg. Combined diffusion-weighted and dynamic contrast-enhanced MRI for prostate cancer diagnosis–correlation with biopsy and histopathology. *J Magn Reson Imaging*, 24: 108–113, 2006.

[46] C. K. Kim, B. K. Park, and B. Kim. Localization of prostate cancer using 3t MRI: comparison of t2-weighted and dynamic contrast-enhanced imaging. *J Comput Assist Tomogr*, 30:7–11, 2006.

[47] E. K. Vos, G. Litjens, T. Kobus, T. Hambrock, C. A. Kaa, J. O. Barentsz, H. Huisman, and T. W. Scheenen. Assessment of prostate cancer aggressiveness using dynamic contrast-enhanced magnetic resonance imaging at 3 t. *Eur Urol*, 64:448–455, 2013.

[48] H. Miao, H. Fukatsu, and T. Ishigaki. Prostate cancer detection with 3-t MRI: comparison of diffusion-weighted and t2-weighted imaging. *Eur J Radiol*, 61:297–302, 2007.

[49] M. A. Haider, T. H. van der Kwast, J. Tanguay, A. J. Evans, A.-T. Hashmi, G. Lockwood, and J. Trachtenberg. Combined T2-weighted and diffusion-weighted MRI for localization of prostate cancer. *AJR Am J Roentgenol*, 189:323–328, 2007.

[50] D. M. Somford, J. J. Fütterer, T. Hambrock, and J. O. Barentsz. Diffusion and perfusion MR imaging of the prostate. *Magn Reson Imaging Clin N Am*, 16:685–95, ix, 2008.

[51] T. Hambrock, D. M. Somford, H. J. Huisman, I. M. van Oort, J. A. Witjes, C. A. Hulsbergen-van de Kaa, T. Scheenen, and J. O. Barentsz. Relationship between apparent diffusion coefficients at 3.0-T MR imaging and Gleason grade in peripheral zone prostate cancer. *Radiology*, 259:453–461, 2011.

[52] S. Verma, A. Rajesh, H. Morales, L. Lemen, G. Bills, M. Delworth, K. Gaitonde, J. Ying, R. Samartunga, and M. Lamba. Assessment of aggressiveness of prostate cancer: correlation of apparent diffusion coefficient with histologic grade after radical prostatectomy. *AJR Am J Roentgenol*, 196:374–381, 2011.

[53] G. J. Kelloff, P. Choyke, D. S. Coffey, and P. C. I. W. G. . Challenges in clinical prostate cancer: role of imaging. *AJR Am J Roentgenol*, 192:1455–1470, 2009.

[54] L. Dickinson, H. U. Ahmed, C. Allen, J. O. Barentsz, B. Carey, J. J. Fütterer, S. W. Heijmink, P. J. Hoskin, A. Kirkham, A. R. Padhani, R. Persad, P. Puech, S. Punwani, A. S. Sohaib, B. Tombal, A. Villers, J. van der Meulen, and M. Emberton. Magnetic resonance imaging for the detection, localisation, and characterisation of prostate cancer: recommendations from a European consensus meeting. *Eur Urol*, 59:477–494, 2011.

[55] H. J. Huisman, M. R. Engelbrecht, and J. O. Barentsz. Accurate estimation of pharmacokinetic contrast-enhanced dynamic MRI parameters of the prostate. *J Magn Reson Imaging*, 13:607–614, 2001.

[56] P. C. Vos, T. Hambrock, J. O. Barentsz, and H. J. Huisman. Automated calibration for computerized analysis of prostate lesions using pharmacokinetic magnetic resonance images. In *Med Image Comput Comput Assist Interv*, volume 12 of *Lect Notes Comput Sci*, pages 836–843, 2009.

[57] P. C. Vos, T. Hambrock, J. O. Barentsz, and H. J. Huisman. Computer-assisted analysis of peripheral zone prostate lesions using T2-weighted and dynamic contrast enhanced T1-weighted MRI. *Phys Med Biol*, 55:1719–1734, 2010.

[58] D. Portalez, P. Mozer, F. Cornud, R. Renard-Penna, V. Misrai, M. Thoulouzan, and B. Malavaud. Validation of the European Society of Urogenital Radiology scoring system for prostate cancer diagnosis on multiparametric magnetic resonance imaging in a cohort of repeat biopsy patients. *Eur Urol*, 62:986–996, 2012.

[59] A. B. Rosenkrantz, S. Kim, R. P. Lim, N. Hindman, F.-M. Deng, J. S. Babb, and S. S. Taneja. Prostate cancer localization using multiparametric MR imaging: comparison of Prostate Imaging Reporting and Data System (PI-RADS) and Likert scales. *Radiology*, 269:482–492, 2013.

[60] A. B. Rosenkrantz, R. P. Lim, M. Haghighi, M. B. Somberg, J. S. Babb, and S. S. Taneja. Comparison of interreader reproducibility of the prostate imaging reporting and data system and likert scales for evaluation of multiparametric prostate MRI. *AJR Am J Roentgenol*, 201:W612–W618, 2013.

[61] L. Schimmöller, M. Quentin, C. Arsov, R. S. Lanzman, A. Hiester, R. Rabenalt, G. Antoch, P. Albers, and D. Blondin. Inter-reader agreement of the esur score for prostate MRI using in-bore MRI-guided biopsies as the reference standard. *Eur Radiol*, 23:3185–3190, 2013.

[62] J. Thompson, P. Stricker, P. Brenner, D. Moses, R. Shnier, W. Delprado, A. M. Haynes, and A. Hayen. Magnetic resonance imaging detects significant prostate cancer and could be used to reduce unnecessary biopsies: initial results from a prospective trial. *J Urol*, 189: e910–e911, 2013.

[63] M. R. Pokorny, M. de Rooij, E. Duncan, F. H. Schröder, R. Parkinson, J. O. Barentsz, and L. C. Thompson. Prospective study of diagnostic accuracy comparing prostate cancer detection by transrectal ultrasound-guided biopsy versus magnetic resonance (MR) imaging with subsequent MR-guided biopsy in men without previous prostate biopsies. *Eur Urol*, 66:22–29, 2014.

[64] M. de Rooij, S. Crienen, J. A. Witjes, J. O. Barentsz, M. M. Rovers, and J. P. Grutters. Cost-effectiveness of magnetic resonance (MR) imaging and MR-guided targeted biopsy versus systematic transrectal ultrasound-guided biopsy in diagnosing prostate cancer: A modelling study from a health care perspective. *Eur Urol*, 66:430–436, 2014.

[65] M. L. Giger, N. Karssemeijer, and S. G. Armato. Computer-aided diagnosis in medical imaging. *IEEE Trans Med Imaging*, 20:1205–1208, 2001.

[66] G. S. Lodwick, T. E. Keats, and J. P. Dorst. The coding of Roentgen images for computer analysis as applied to lung cancer. *Radiology*, 81:185–200, 1963.

[67] B. M. ter Haar Romeny and L. M. J. Florack. Front end vision, a multiscale geometry engine. In *Lecture notes in computer science.*, Heidelberg, 2000. Springer.

[68] B. van Ginneken, S. Katsuragawa, B. M. ter Haar Romeny, K. Doi, and M. A. Viergever. Automatic detection of abnormalities in chest radiographs using local texture analysis. *IEEE Trans Med Imaging*, 21:139–149, 2002.

[69] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313:504–507, 2006.

[70] H. Peng, F. Long, and C. Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans Pattern Anal Mach Intell*, 27:1226–1238, 2005.

[71] M. A. Hall. *Correlation-based feature selection for machine learning.* PhD thesis, The University of Waikato, 1999.

[72] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern classification.* John Wiley and Sons, New York, 2nd edition, 2001.

[73] J. Friedman, T. Hastie, and R. Tibshirani. Special invited paper. additive logistic regression: A statistical view of boosting. *Ann Stat*, 28:337–374, 2000.

[74] L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

[75] H. H. Thodberg, S. Kreiborg, A. Juul, and K. D. Pedersen. The bonexpert method for automated determination of skeletal maturity. *IEEE Trans Med Imaging*, 28:52–66, 2009.

[76] S. Kakeda, J. Moriya, H. Sato, T. Aoki, H. Watanabe, H. Nakata, N. Oda, S. Katsuragawa, K. Yamamoto, and K. Doi. Improved detection of lung nodules on chest radiographs using a commercial computer-aided diagnosis system. *AJR Am J Roentgenol*, 182:505–10, 2004.

[77] S. Kligerman, L. Cai, and C. S. White. The effect of computer-aided detection on radiologist performance in the detection of lung cancers previously missed on a chest radiograph. *J Thorac Imaging*, 28:244–252, 2013.

[78] F. J. Gilbert, S. M. Astley, M. G. C. Gillan, O. F. Agbaje, M. G. Wallis, J. James, C. R. M. Boggis, S. W. Duffy, and CADET II Group. Single reading with computer-aided detection for screening mammography. *N Engl J Med*, 359:1675–1684, 2008.

[79] J. J. Fenton, S. H. Taplin, P. A. Carney, L. Abraham, E. A. Sickles, C. D'Orsi, E. A. Berns, G. Cutter, R. E. Hendrick, W. E. Barlow, and J. G. Elmore. Influence of computer-aided detection on performance of screening mammography. *N Engl J Med*, 356:1399–1409, 2007.

[80] M. Gromet. Comparison of computer-aided detection to double reading of screening mammograms: review of 231,221 mammograms. *AJR Am J Roentgenol*, 190:854–859, 2008.

[81] L. A. L. Khoo, P. Taylor, and R. M. Given-Wilson. Computer-aided detection in the United Kingdom National Breast Screening Programme: prospective study. *Radiology*, 237:444–449, 2005.

[82] J. J. Fenton, L. Abraham, S. H. Taplin, B. M. Geller, P. A. Carney, C. D'Orsi, J. G. Elmore, and W. E. Barlow. Effectiveness of computer-aided detection in community mammography practice. *J Natl Cancer Inst*, 103:1152–1161, 2011.

[83] G. Iussich, L. Correale, C. Senore, N. Segnan, A. Laghi, F. Iafrate, D. Campanella, E. Neri, F. Cerri, C. Hassan, and D. Regge. CT colonography: preliminary assessment of a double-read paradigm that uses computer-aided detection as the first reader. *Radiology*, 268:743–751, 2013.

[84] A. Christe, L. Leidolt, A. Huber, P. Steiger, Z. Szucs-Farkas, J. E. Roos, J. T. Heverhagen, and L. Ebner. Lung cancer screening with CT: Evaluation of radiologists and different computer assisted detection software (CAD) as first and second readers for lung nodule detection at different dose levels. *Eur Radiol*, 2013.

[85] L. E. Hogeweg. *Automatic detection of tuberculosis in chest radiographs.* PhD thesis, Radboud Universiteit Nijmegen, 2013.

[86] M. Samulski, R. Hupse, C. Boetes, R. Mus, G. den Heeten, and N. Karssemeijer. Using Computer Aided Detection in Mammography as a Decision Support. *Eur Radiol*, 20: 2323–2330, 2010.

[87] R. Hupse, M. Samulski, M. B. Lobbes, R. M. Mann, R. Mus, G. J. den Heeten, D. Beijerinck, R. M. Pijnappel, C. Boetes, and N. Karssemeijer. Computer-aided detection of masses at mammography: Interactive decision support versus prompts. *Radiology*, 266:123–129, 2013.

[88] I. Chan, W. Wells, 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. *Med Phys*, 30:2390–2398, 2003.

[89] 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. *J Magn Reson Imaging*, 30:327–334, 2009.

[90] P. Puech, N. Betrouni, N. Makni, A.-S. Dewalle, A. Villers, and L. Lemaitre. Computer-assisted diagnosis of prostate cancer using DCE-MRI data: design, implementation and preliminary results. *Int J Comput Assist Radiol Surg*, 4:1–10, 2009.

[91] Y. Artan, M. A. Haider, D. L. Langer, T. H. van der Kwast, A. J. Evans, Y. Yang, M. N. Wernick, J. Trachtenberg, and I. S. Yetik. Prostate cancer localization with multispectral MRI using cost-sensitive support vector machines and conditional random fields. *IEEE Trans Image Process*, 19:2444–2455, 2010.

[92] Y. Liu, H. Liu, X. Bai, Z. Ye, H. Sun, R. Bai, and D. Wang. Differentiation of metastatic from non-metastatic lymph nodes in patients with uterine cervical cancer using diffusion-weighted imaging. *Gynecol Oncol*, 2011.

[93] P. Tiwari, S. Viswanath, J. Kurhanewicz, A. Sridhar, and A. Madabhushi. Multimodal wavelet embedding representation for data combination (maweric): integrating mag-

netic resonance imaging and spectroscopy for prostate cancer detection. *NMR Biomed*, 25:607–619, 2011.

[94] P. Tiwari, J. Kurhanewicz, and A. Madabhushi. Multi-kernel graph embedding for detection, gleason grading of prostate cancer via MRI/mrs. *Med Image Anal*, 17:219–235, 2013.

[95] 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 zone prostate tumors have significantly different quantitative imaging signatures on 3 tesla endorectal, in vivo T2-weighted MR imagery. *J Magn Reson Imaging*, 36:213–224, 2012.

[96] 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. *Phys Med Biol*, 57:1527–1542, 2012.

[97] T. Hambrock, P. C. Vos, C. A. Hulsbergen-van de Kaa, J. O. Barentsz, and H. J. Huisman. Prostate cancer: Computer-aided diagnosis with multiparametric 3-t MR imaging–effect on observer performance. *Radiology*, 266:521–530, 2013.

[98] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, Secaucus, NJ, USA, 2007. ISBN 0387310738.

[99] C. M. Rutter. Bootstrap estimation of diagnostic accuracy with patient-clustered data. *Acad Radiol*, 7:413–419, 2000.

[100] H. Bornefalk and A. B. Hermansson. On the comparison of froc curves in mammography CAD systems. *Med Phys*, 32:412–417, 2005.

[101] C. E. Metz, B. A. Herman, and J. H. Shen. Maximum likelihood estimation of receiver operating characteristic (roc) curves from continuously-distributed data. *Stat Med*, 17: 1033–1053, 1998.

[102] G. M. Villeirs, G. O. D. Meerleer, P. J. D. Visschere, V. H. Fonteyne, A. C. Verbaeys, and W. Oosterlinck. Combined magnetic resonance imaging and spectroscopy in the assessment of high grade prostate carcinoma in patients with elevated PSA: a single-institution experience of 356 patients. *Eur J Radiol*, 77:340–345, 2011.

[103] C. M. A. Hoeks, T. Hambrock, D. Yakar, C. A. Hulsbergen-van de Kaa, T. Feuth, J. A. Witjes, J. J. Fütterer, and J. O. Barentsz. Transition zone prostate cancer: Detection and localization with 3-t multiparametric MR imaging. *Radiology*, 266:207–217, 2013.

[104] D. Pasquier, T. Lacornerie, M. Vermandel, J. Rousseau, E. Lartigau, and N. Betrouni. Automatic segmentation of pelvic structures from magnetic resonance images for prostate cancer radiotherapy. *Int J Radiat Oncol Biol Phys*, 68:592–600, 2007.

[105] R. Toth, B. N. Bloch, E. M. Genega, N. M. Rofsky, R. E. Lenkinski, M. A. Rosen, A. Kalyanpur, S. Pungavkar, and A. Madabhushi. Accurate prostate volume estimation using multifeature active shape models on T2-weighted MRI. *Acad Radiol*, 18:745–754, 2011.

[106] Y. Hu, H. U. Ahmed, Z. Taylor, C. Allen, M. Emberton, D. Hawkes, and D. Barratt. MR to ultrasound registration for image-guided prostate interventions. *Med Image Anal*, 16: 687–703, 2012.

[107] M. J. Costa, H. Delingette, S. Novellas, and N. Ayache. Automatic segmentation of bladder and prostate using coupled 3D deformable models. *Med Image Comput Comput Assist Interv*, 10:252–260, 2007.

[108] S. Klein, U. A. van der Heide, I. M. Lips, M. van Vulpen, M. Staring, and J. P. W. Pluim. Automatic segmentation of the prostate in 3D MR images by atlas matching using localized mutual information. *Med Phys*, 35:1407–1417, 2008.

[109] N. Makni, P. Puech, R. Lopes, A. S. Dewalle, O. Colot, and N. Betrouni. Combining a deformable model and a probabilistic framework for an automatic 3d segmentation of prostate on MRI. *Int J Comput Assist Radiol Surg*, 4:181–188, 2009.

[110] R. Toth, P. Tiwari, M. Rosen, G. Reed, J. Kurhanewicz, A. Kalyanpur, S. Pungavkar, and A. Madabhushi. A magnetic resonance spectroscopy driven initialization scheme for active shape model based prostate segmentation. *Med Image Anal*, 15:214–225, 2011.

[111] S. S. Chandra, J. A. Dowling, K.-K. Shen, P. Raniga, J. P. W. Pluim, P. B. Greer, O. Salvado, and J. Fripp. Patient specific prostate segmentation in 3-D magnetic resonance images. *IEEE Trans Med Imaging*, 31:1955–1964, 2012.

[112] Y. Gao, S. Liao, and D. Shen. Prostate segmentation by sparse representation based classification. *Med Phys*, 39:6372–6387, 2012.

[113] T. Heimann, B. van Ginneken, M. Styner, Y. Arzhaeva, V. Aurich, C. Bauer, A. Beck, C. Becker, R. Beichel, G. Bekes, F. Bello, G. Binnig, H. Bischof, A. Bornik, P. Cashman, Y. Chi, A. Cordova, B. Dawant, M. Fidrich, J. Furst, D. Furukawa, L. Grenacher, J. Hornegger, D. Kainmuller, R. Kitney, H. Kobatake, H. Lamecker, T. Lange, J. Lee, B. Lennon, R. Li, S. Li, H.-P. Meinzer, G. Nemeth, D. Raicu, A.-M. Rau, E. van Rikxoort, M. Rousson, L. Rusko, K. Saddi, G. Schmidt, D. Seghers, A. Shimizu, P. Slagmolen, E. Sorantin, G. Soza, R. Susomboon, J. Waite, A. Wimmer, and I. Wolf. Comparison

and Evaluation of Methods for Liver Segmentation From CT Datasets. *IEEE Trans Med Imaging*, 28:1251–1265, 2009.

[114] M. Schaap, C. T. Metz, T. van Walsum, A. G. van der Giessen, A. C. Weustink, N. R. Mollet, C. Bauer, H. Bogunović, C. Castro, X. Deng, E. Dikici, T. O'Donnell, M. Frenay, O. Friman, M. Hernández-Hoyos, P. H. Kitslaar, K. Krissian, C. Kühnel, M. A. Luengo-Oroz, M. Orkisz, O. Smedby, M. Styner, A. Szymczak, H. Tek, C. Wang, S. K. Warfield, S. Zambal, Y. Zhang, G. P. Krestin, and W. J. Niessen. Standardized evaluation methodology and reference database for evaluating coronary artery centerline extraction algorithms. *Med Image Anal*, 13:701–714, 2009.

[115] D. W. Shattuck, G. Prasad, M. Mirza, K. L. Narr, and A. W. Toga. Online resource for validation of brain segmentation methods. *Neuroimage*, 45:431–439, 2009.

[116] M. Niemeijer, B. van Ginneken, M. J. Cree, A. Mizutani, G. Quellec, C. I. Sánchez, B. Zhang, R. Hornero, M. Lamard, C. Muramatsu, X. Wu, G. Cazuguel, J. You, A. Mayo, Q. Li, Y. Hatanaka, B. Cochener, C. Roux, F. Karray, M. Garcia, H. Fujita, and M. D. Abràmoff. Retinopathy online challenge: automatic detection of microaneurysms in digital color fundus photographs. *IEEE Trans Med Imaging*, 29:185–195, 2010.

[117] K. Murphy, B. van Ginneken, J. M. Reinhardt, S. Kabus, K. Ding, X. Deng, K. Cao, K. Du, G. E. Christensen, V. Garcia, T. Vercauteren, N. Ayache, O. Commowick, G. Malandain, B. Glocker, N. Paragios, N. Navab, V. Gorbunova, J. Sporring, M. de Bruijne, X. Han, M. P. Heinrich, J. A. Schnabel, M. Jenkinson, C. Lorenz, M. Modat, J. R. McClelland, S. Ourselin, S. E. A. Muenzing, M. A. Viergever, D. D. Nigris, D. L. Collins, T. Arbel, M. Peroni, R. Li, G. C. Sharp, A. Schmidt-Richberg, J. Ehrhardt, R. Werner, D. Smeets, D. Loeckx, G. Song, N. Tustison, B. Avants, J. C. Gee, M. Staring, S. Klein, B. C. Stoel, M. Urschler, M. Werlberger, J. Vandemeulebroucke, S. Rit, D. Sarrut, and J. P. W. Pluim. Evaluation of registration methods on thoracic CT: The EMPIRE10 challenge. *IEEE Trans Med Imaging*, 31:1901–1920, 2011.

[118] K. V. Leemput, F. Maes, D. Vandermeulen, and P. Suetens. Automated model-based bias field correction of MR images of the brain. *IEEE Trans Med Imaging*, 18:885–896, 1999.

[119] K. Sung, B. L. Daniel, and B. A. Hargreaves. Transmit B1+ field inhomogeneity and T(1) estimation errors in breast DCE-MRI at 3 tesla. *J Magn Reson Imaging*, 2013.

[120] J. J. Fütterer and J. Barentsz. 3T MRI of prostate cancer. *Appl Radiol*, 38:25–37, 2009.

[121] X. Li, W. Huang, and W. D. Rooney. Signal-to-noise ratio, contrast-to-noise ratio and pharmacokinetic modeling considerations in dynamic contrast-enhanced magnetic resonance imaging. *Magn Reson Imaging*, 30:1313–1322, 2012.

[122] L. G. Nyúl and J. K. Udupa. On standardizing the MR image intensity scale. *Magn Reson Med*, 42:1072–1081, 1999.

[123] L. G. Nyúl, J. K. Udupa, and X. Zhang. New variants of a method of MRI scale standardization. *IEEE Trans Med Imaging*, 19:143–150, 2000.

[124] J. C. Bezdek, L. O. Hall, and L. P. Clarke. Review of MR image segmentation techniques using pattern recognition. *Med Phys*, 20:1033–1048, 1993.

[125] L. P. Clarke, R. P. Velthuizen, S. Phuphanich, J. D. Schellenberg, J. A. Arrington, and M. Silbiger. MRI: stability of three supervised segmentation techniques. *Magn Reson Imaging*, 11:95–106, 1993.

[126] G. Vincent, G. Guillard, and M. Bowes. Fully Automatic Segmentation of the Prostate using Active Appearance Models. *MICCAI Grand Challenge: Prostate MR Image Segmentation 2012*, 2012.

[127] T. F. Cootes, C. J. Twining, V. Petrović, R. Schestowitz, and C. J. Taylor. Groupwise construction of appearance models using piece-wise affine deformations. In *Proceedings of 16th British Machine Vision Conference*, pages 879–888, 2005.

[128] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *IEEE Trans Pattern Anal Mach Intell*, 23:681–685, 2001.

[129] N. Birkbeck, J. Zhang, M. Requardt, B. Kiefer, P. Gall, and S. Kevin Zhou. Region-Specific Hierarchical Segmentation of MR Prostate Using Discriminative Learning. *MICCAI Grand Challenge: Prostate MR Image Segmentation 2012*, 2012.

[130] P. Pérez, M. Gangnet, and A. Blake. Poisson image editing. *ACM Trans Graph*, 22: 313–318, 2003.

[131] B. Zheng, J. Tan, M. A. Ganott, D. M. Chough, and D. Gur. Matching breast masses depicted on different views: a comparison of three methods. *Acad Radiol*, 16:1338–1347, 2009.

[132] F. Malmberg, R. Strand, J. Kullberg, R. Nordenskjöld, and E. Bengtsson. Smart Paint A New Interactive Segmentation Method Applied to MR Prostate Segmentation. *MICCAI Grand Challenge: Prostate MR Image Segmentation 2012*, 2012.

[133] Y. Ou, J. Doshi, G. Erus, and C. Davatzikos. Multi-Atlas Segmentation of the Prostate: A Zooming Process with Robust Registration and Atlas Selection. *MICCAI Grand Challenge: Prostate MR Image Segmentation 2012*, 2012.

[134] S. K. Warfield, K. H. Zou, and W. M. Wells. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. *IEEE Trans Med Imaging*, 23:903 –921, 2004.

[135] Y. Ou, A. Sotiras, N. Paragios, and C. Davatzikos. Dramms: Deformable registration via attribute matching and mutual-saliency weighting. *Med Image Anal*, 15:622–639, 2010.

[136] M. Kirschner, F. Jung, and S. Wesarg. Automatic Prostate Segmentation in MR Images with a Probabilistic Active Shape Model. *MICCAI Grand Challenge: Prostate MR Image Segmentation 2012*, 2012.

[137] W. E. Lorensen and H. E. Cline. Marching cubes: a high resolution 3D surface construction algorithm. In *Computer graphics: SIGGRAPH '87 Conference Proceedings*, volume 21, pages 163–169, 1987.

[138] C. Li, C. Xu, A. W. Anderson, and J. C. Gore. MRI tissue classification and bias field estimation based on coherent local intensity clustering: A unified energy minimization framework. In *Inf Process Med Imaging*, Lect Notes Comput Sci, pages 288–299, 2009.

[139] P. Viola and M. Jones. Robust Real-time Object Detection. In *International Journal of Computer Vision*, 2001.

[140] J. Yuan, W. Qiu, E. Ukwatta, M. Rajchl, Y. Sun, and A. Fenster. An Efficient Convex Optimization Approach to 3D Prostate MRI Segmentation with Generic Star Shape Prior. *MICCAI Grand Challenge: Prostate MR Image Segmentation 2012*, 2012.

[141] Q. Gao, D. Rueckert, and P. Edwards. An automatic multi-atlas based prostate segmentation using local appearance-specific atlases and patch-based voxel weighting. *MICCAI Grand Challenge: Prostate MR Image Segmentation 2012*, 2012.

[142] P. Coupé, J. V. Manjón, V. Fonov, J. Pruessner, M. Robles, and D. L. Collins. Patch-based segmentation using expert priors: Application to hippocampus and ventricle segmentation. *Neuroimage*, 54:940–954, 2011.

[143] B. Maan and F. van der Heijden. Prostate MR image segmentation using 3D Active Appearance Models. *MICCAI Grand Challenge: Prostate MR Image Segmentation 2012*, 2012.

[144] D. Kroon, P. Kowalski, W. Tekieli, E. Reeuwijk, D. Saris, and C. H. Slump. MRI based knee cartilage assessment. In *Medical Imaging*, volume 8315 of *Proceedings of the SPIE*, pages 83151V–83151V–10, 2012.

[145] G. J. S. Litjens, N. Karssemeijer, and H. J. Huisman. A multi-atlas approach for prostate segmentation in MR images. *MICCAI Grand Challenge: Prostate MR Image Segmentation 2012*, 2012.

[146] T. R. Langerak, U. A. van der Heide, A. N. T. J. Kotte, M. A. Viergever, M. van Vulpen, and J. P. W. Pluim. Label fusion in atlas-based segmentation using a selective and iterative method for performance level estimation (SIMPLE). *IEEE Trans Med Imaging*, 29: 2000–2008, 2010.

[147] R. Toth and A. Madabhushi. Deformable Landmark-Free Active Appearance Models: Application to Segmentation of Multi-Institutional Prostate MRI Data. *MICCAI Grand Challenge: Prostate MR Image Segmentation 2012*, 2012.

[148] R. Toth and A. Madabhushi. Multifeature landmark-free active appearance models: application to prostate MRI segmentation. *IEEE Trans Med Imaging*, 31:1638–1650, 2012.

[149] S. Ghose, J. Mitra, A. Oliver, R. Martí, X. Lladó, J. Freixenet, J. C. Vilanova, D. Sidibé, and F. Meriaudeau. A Random Forest Based Classification Approach to Prostate Segmentation in MRI. *MICCAI Grand Challenge: Prostate MR Image Segmentation 2012*, 2012.

[150] T. F. Cootes, K. Walker, and C. J. Taylor. View-based active appearance models. In *Proc. 4th IEEE Int. Conf. on Automatic Face and Gesture Recognition*, pages 227–32. IEEE Comput. Soc, 2000.

[151] R. Wolz, C. Chu, K. Misawa, K. Mori, and D. Rueckert. Multi-organ abdominal CT segmentation using hierarchically weighted subject-specific atlases. *Med Image Comput Comput Assist Interv*, 15:10–17, 2012.

[152] R. Siegel, D. Naishadham, and A. Jemal. Cancer statistics, 2012. *CA Cancer J Clin*, 62: 10–29, 2012.

[153] G. M. Villeirs, K. L. Verstraete, W. J. De Neve, and G. O. De Meerleer. Magnetic resonance imaging anatomy of the prostate and periprostatic area: a guide for radiotherapists. *Radiother Oncol*, 76:99–106, 2005.

[154] N. Makni, A. Iancu, O. Colot, P. Puech, S. Mordon, and N. Betrouni. Zonal segmentation of prostate using multispectral magnetic resonance images. *Med Phys*, 38:6093, 2011.

[155] S. Klein, M. Staring, K. Murphy, M. A. Viergever, and J. P. W. Pluim. elastix: a toolbox for intensity-based medical image registration. *IEEE Trans Med Imaging*, 29:196–205, 2010.

[156] M. Amadasun and R. King. Textural features corresponding to textural properties. *IEEE Trans Syst Man Cybern*, 19:1264–1274, 1989.

[157] H. Li, M. L. Giger, O. I. Olopade, A. Margolis, L. Lan, and M. R. Chinander. Computerized texture analysis of mammographic parenchymal patterns of digitized mammograms. *Acad Radiol*, 12:863–873, 2005.

[158] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans Pattern Anal Mach Intell*, 24:971–987, 2002.

[159] A. B. Rosenkrantz and S. S. Taneja. Radiologist, be aware: ten pitfalls that confound the interpretation of multiparametric prostate MRI. *AJR Am J Roentgenol*, 202:109–120, 2014.

[160] H. A. Vargas, O. Akin, T. Franiel, Y. Mazaheri, J. Zheng, C. Moskowitz, K. Udo, J. Eastham, and H. Hricak. Diffusion-weighted endorectal MR imaging at 3 T for prostate cancer: Tumor detection and assessment of aggressiveness. *Radiology*, 2011.

[161] T. Kobus, P. C. Vos, T. Hambrock, M. De Rooij, C. A. Hulsbergen-Van de Kaa, J. O. Barentsz, A. Heerschap, and T. W. J. Scheenen. Prostate cancer aggressiveness: In vivo assessment of MR spectroscopy and diffusion-weighted imaging at 3 t. *Radiology*, 265: 457–467, 2012.

[162] I. Ocak, M. Bernardo, G. Metzger, T. Barrett, P. Pinto, P. S. Albert, and P. L. Choyke. Dynamic contrast-enhanced MRI of prostate cancer at 3 t: a study of pharmacokinetic parameters. *AJR Am J Roentgenol*, 189:849, 2007.

[163] A. Oto, A. Kayhan, Y. Jiang, M. Tretiakova, C. Yang, T. Antic, F. Dahi, A. L. Shalhav, G. Karczmar, and W. M. Stadler. Prostate cancer: differentiation of central gland cancer from benign prostatic hyperplasia by using diffusion-weighted and dynamic contrast-enhanced MR imaging. *Radiology*, 257:715–723, 2010.

[164] X. Liu, L. Zhou, W. Peng, C. Wang, and H. Wang. Differentiation of central gland prostate cancer from benign prostatic hyperplasia using monoexponential and biexponential diffusion-weighted imaging. *Magn Reson Imaging*, 31:1318–1324, 2013.

[165] A. L. Chesnais, E. Niaf, F. Bratan, F. Mège-Lechevallier, S. Roche, M. Rabilloud, M. Colombel, and O. Rouvière. Differentiation of transitional zone prostate cancer from benign hyperplasia nodules: evaluation of discriminant criteria at multiparametric MRI. *Clin Radiol*, 68:e323–e330, 2013.

[166] K. N. A. Nagel, M. G. Schouten, T. Hambrock, G. Litjens, C. M. A. Hoeks, B. T. Haken, J. O. Barentsz, and J. J. Fütterer. Differentiation of prostatitis and prostate cancer by using diffusion-weighted MR imaging and MR-guided biopsy at 3 t. *Radiology*, 267: 164–172, 2013.

[167] S. C. Agner, M. A. Rosen, S. Englander, J. E. Tomaszewski, M. D. Feldman, P. Zhang, C. Mies, M. D. Schnall, and A. Madabhushi. Computerized image analysis for identifying triple-negative breast cancers and differentiating them from other molecular subtypes of breast cancer on dynamic contrast-enhanced MR images: A feasibility study. *Radiology*, page 121031, 2014.

[168] G. Xiao, B. N. Bloch, J. Chappelow, E. M. Genega, N. M. Rofsky, R. E. Lenkinski, J. Tomaszewski, M. D. Feldman, M. Rosen, and A. Madabhushi. Determining histology-MRI slice correspondences for defining MRI-based disease signatures of prostate cancer. *Comput Med Imaging Graph*, 35:568–578, 2011.

[169] G. Litjens, O. A. Debats, W. J. M. van de Ven, N. Karssemeijer, and H. J. Huisman. A pattern recognition approach to zonal segmentation of the prostate on MRI. In *Med Image Comput Comput Assist Interv*, volume 7511 of *Lect Notes Comput Sci*, pages 413–420, 2012.

[170] Q. Li, S. Sone, and K. Doi. Selective enhancement filters for nodules, vessels, and airway walls in two- and three-dimensional CT scans. *Med Phys*, 30:2040–2051, 2003.

[171] M. Mischi, S. Turco, C. Lavini, K. Kompatsiari, J. J. M. C. H. de la Rosette, M. Breeuwer, and H. Wijkstra. Magnetic resonance dispersion imaging for localization of angiogenesis and cancer growth. *Invest Radiol*, 2014.

[172] P. Somol, P. Pudil, J. Novovicova, and P. Paclik. Adaptive floating search methods in feature selection. *Pattern Recognit Lett*, 20:1157–1163, 1999.

[173] H. Trivedi, B. Turkbey, A. R. Rastinehad, C. J. Benjamin, M. Bernardo, T. Pohida, V. Shah, M. J. Merino, B. J. Wood, W. M. Linehan, A. M. Venkatesan, P. L. Choyke, and P. A. Pinto. Use of patient-specific MRI-based prostate mold for validation of multiparametric MRI in localization of prostate cancer. *Urology*, 79:233–239, 2012.

[174] A. Jemal, R. Siegel, J. Xu, and E. Ward. Cancer statistics, 2010. *CA Cancer J Clin*, 60: 277–300, 2010.

[175] M. L. Blute, E. J. Bergstralh, A. Iocca, B. Scherer, and H. Zincke. Use of gleason score, prostate specific antigen, seminal vesicle and margin status to predict biochemical failure after radical prostatectomy. *J Urol*, 165:119–125, 2001.

[176] L. Egevad, T. Granfors, L. Karlberg, A. Bergh, and P. Stattin. Percent gleason grade 4/5 as prognostic factor in prostate cancer diagnosed at transurethral resection. *J Urol*, 168: 509–513, 2002.

[177] V. Narain, F. J. Bianco, D. J. Grignon, W. A. Sakr, J. E. Pontes, and D. P. Wood. How accurately does prostate biopsy gleason score predict pathologic findings and disease free survival? *Prostate*, 49:185–190, 2001.

[178] Y. Itou, K. Nakanishi, Y. Narumi, Y. Nishizawa, and H. Tsukuma. Clinical utility of apparent diffusion coefficient (ADC) values in patients with prostate cancer: can ADC values contribute to assess the aggressiveness of prostate cancer? *J Magn Reson Imaging*, 33:167–172, 2011.

[179] B. Turkbey, P. A. Pinto, H. Mani, M. Bernardo, Y. Pang, Y. L. McKinney, K. Khurana, G. C. Ravizzini, P. S. Albert, M. J. Merino, and P. L. Choyke. Prostate cancer: value of multiparametric MR imaging at 3 t for detection–histopathologic correlation. *Radiology*, 255:89–99, 2010.

[180] C. E. Metz, B. A. Herman, and C. A. Roe. Statistical comparison of two roc-curve estimates obtained from partially-paired datasets. *Med Decis Making*, 18:110–121, 1998.

[181] R. Siegel, D. Naishadham, and A. Jemal. Cancer statistics, 2013. *CA Cancer J Clin*, 63: 11–30, 2013.

[182] N. Karssemeijer, J. D. M. Otten, H. Rijken, and R. Holland. Computer aided detection of masses in mammograms as decision support. *Br J Radiol*, 79 Spec No 2:S123–S126, 2006.

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

[184] M. D. Abràmoff, M. Niemeijer, M. S. A. Suttorp-Schulten, M. A. Viergever, S. R. Russell, and B. van Ginneken. Evaluation of a system for automatic detection of diabetic retinopathy from color fundus photographs in a large population of patients with diabetes. *Diabetes Care*, 31:193–198, 2008.

[185] E. Niaf, O. Rouvière, F. Mège-Lechevallier, F. Bratan, and C. Lartizien. Computer-aided diagnosis of prostate cancer in the peripheral zone using multiparametric MRI. *Phys Med Biol*, 57:3833–3851, 2012.

[186] A. Firjani, A. Elnakib, F. Khalifa, G. Gimel'farb, M. A. El-Ghar, A. Elmaghraby, and A. El-Baz. A diffusion-weighted imaging based diagnostic system for early detection of prostate cancer. *Journal of Biomedical Science and Engineering*, 6, 2013.

[187] G. Litjens, J. Barentsz, N. Karssemeijer, and H. Huisman. Automated computer-aided detection of prostate cancer in MR images: from a whole-organ to a zone-based approach. In *Medical Imaging*, volume 8315 of *Proceedings of the SPIE*, pages 83150G–83150G–6, 2012.

[188] C. Studholme, C. Drapaca, B. Iordanova, and V. Cardenas. Deformation-based mapping of volume change from serial brain MRI in the presence of local tissue contrast change. *Medical Imaging, IEEE Transactions on*, 25:626–639, 2006.

[189] H. Laue, S. Behrens, F. Giesel, N. Hoggard, C. Zechmann, I. Wilkinson, S. Krass, and H.-O. Peitgen. A statistical comparison of tofts and brix model parameters for glioma and prostate MRI data. In *ESMRMB*, Magnetic Resonance Materials in Physics, Biology and Medicine 19, Supplement 7, page 318, 2006.

[190] T. Wolters, M. J. Roobol, P. J. van Leeuwen, R. C. N. van den Bergh, R. F. Hoedemaeker, G. J. L. H. van Leenders, F. H. Schröder, 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. *J Urol*, 185:121–125, 2011.

[191] W. J. M. van de Ven, C. A. Hulsbergen-van de Kaa, T. Hambrock, J. O. Barentsz, and H. J. Huisman. Simulated required accuracy of image registration tools for targeting high-grade cancer components with prostate biopsies. *Eur Radiol*, 23:1401–1407, 2013.

[192] F. Bratan, E. Niaf, C. Melodelima, A. L. Chesnais, R. Souchon, F. Mège-Lechevallier, M. Colombel, and O. Rouvière. Influence of imaging and histological factors on prostate cancer detection and localisation on multiparametric MRI: a prospective study. *Eur Radiol*, 23:1–11, 2013.

[193] J. M. Kuhnigk, V. Dicken, L. Bornemann, A. Bakai, D. Wormanns, S. Krass, and H. O. Peitgen. Morphological segmentation and partial volume analysis for volumetry of solid pulmonary lesions in thoracic CT scans. *IEEE Trans Med Imaging*, 25:417–434, 2006.

[194] B. Platel, H. Huisman, H. Laue, R. Mus, R. Mann, H. Hahn, and N. Karssemeijer. Computerized characterization of breast lesions using dual-temporal resolution dynamic contrast-enhanced MR images. In *MICCAI Workshop: Breast Image Analysis*, 2011.

[195] G. Litjens, T. Hambrock, C. Hulsbergen-van de Kaa, J. Barentsz, and H. Huisman. Interpatient variation in normal peripheral zone apparent diffusion coefficient: Effect on the prediction of prostate cancer aggressiveness. *Radiology*, 265:260–266, 2012.

[196] A. Heidenreich, J. Bellmunt, M. Bolla, S. Joniau, M. Mason, V. Matveev, N. Mottet, H.-P. Schmid, T. van der Kwast, T. Wiegel, F. Zattoni, and E. A. o. U. . EAU guidelines on prostate cancer. part 1: screening, diagnosis, and treatment of clinically localised disease. *Eur Urol*, 59:61–71, 2011.

[197] J. Thompson, N. Lawrentschuk, M. Frydenberg, L. Thompson, and P. Stricker. The role of magnetic resonance imaging in the diagnosis and management of prostate cancer. *BJU International*, 112:6–20, 2013.

[198] N. Lawrentschuk and N. Fleshner. The role of magnetic resonance imaging in targeting prostate cancer in patients with previous negative biopsies and elevated prostate-specific antigen levels. *BJU International*, 103:730–733, 2009.

[199] M. M. Siddiqui, S. Rais-Bahrami, H. Truong, L. Stamatakis, S. Vourganti, J. Nix, A. N. Hoang, A. Walton-Diaz, B. Shuch, M. Weintraub, J. Kruecker, H. Amalou, B. Turkbey, M. J. Merino, P. L. Choyke, B. J. Wood, and P. A. Pinto. Magnetic resonance imaging/ultrasound-fusion biopsy significantly upgrades prostate cancer versus systematic 12-core transrectal ultrasound biopsy. *Eur Urol*, 64:713–719, 2013.

[200] N. Numao, S. Yoshida, C. Ishii, Y. Komai, T. Kijima, M. Yokoyama, J. Ishioka, Y. Matsuoka, F. Koga, K. Saito, H. Masuda, Y. Fujii, S. Kawakami, and K. Kihara. Potential of prebiopsy multiparametric magnetic resonance imaging to reduce initial biopsies in men with suspected clinically localized prostate cancer. *J Urol*, 189:e602–, 2013.

[201] T. Kobus, A. K. Bitz, M. J. van Uden, M. W. Lagemaat, E. Rothgang, S. Orzada, A. Heerschap, and T. W. J. Scheenen. In vivo (31) p MR spectroscopic imaging of the human prostate at 7 t: Safety and feasibility. *Magn Reson Med*, 68:1683 – 1695, 2012.

[202] E. Niaf, C. Lartizien, F. Bratan, L. Roche, M. Rabilloud, F. Mège-Lechevallier, and O. Rouvière. Prostate focal peripheral zone lesions: Characterization at multiparametric MR imaging-influence of a computer-aided diagnosis system. *Radiology*, page 130448, 2014.

[203] G. Litjens, O. Debats, J. Barentsz, N. Karssemeijer, and H. Huisman. Computer-aided detection of prostate cancer in MRI. *IEEE Trans Med Imaging*, 33:1083–1092, 2014.

# Acknowledgments

En dan nu het leukste en moeilijkste stuk van het promoveren, het schrijven van het dankwoord. Leuk omdat het je de kans geeft om nog eens terug te blikken op de afgelopen vier jaar en aan de mooie dingen die je hebt beleeft, moeilijk omdat je in een paar zinnen niet altijd goed kan uitdrukken wat mensen voor je betekend hebben. Toch ga ik een poging wagen.

Dr. Huisman, beste Henkjan, als dagelijks begeleider heb jij het dichtst bij mijn promotie gestaan. Van jou heb ik geleerd dat technici en medici niet altijd dezelfde taal spreken en hoe belangrijk het is daar rekening mee te houden. Ik heb onze gesprekken en discussies over het onderzoek altijd als heel constructief en prettig ervaren. Daarnaast wil ik je graag bedanken voor je altijd positieve kijk als ik ergens vastzat (er zijn geen problemen, alleen maar uitdagingen), het feit dat je me veel vrijheid gaf om mijn eigen onderzoek in te vullen en je tomeloze inzet als het aankwam op dingen geregeld krijgen.

Prof. Karssemeijer, beste Nico, jij bent in de wereld van de medische beeldanalyse misschien wel de meeste deskundige en ervaren onderzoeker als het aankomt op het ontwikkelen van CAD-systemen die daadwerkelijk een bijdrage kunnen leveren aan kliniek. Ik ben dan ook erg blij dat jij één van mijn promotoren bent. Jouw kritische kijk op de evaluatie en implementatie van CAD-systemen heeft bijgedragen aan een aantal succesvolle publicaties, waarvan ik ons TMI-artikel als grootste succes beschouw.

Prof. Barentsz, beste Jelle, het was voor mij een enorm geluk dat ik in de beste klinische prostaat-MRI-groep ter wereld heb mogen werken. Waar andere onderzoekers vaak te maken hebben met slechte beelden en kleine patiëntengroepen had ik duizenden datasets van hoge kwaliteit tot mijn beschikking. Dit voornamelijk door jouw drive en idee dat alleen het beste goed genoeg is voor de patiënt. Inmiddels maken we door jouw inzet als afdeling ons eigen contrastmiddel en is PIRADS de standaard voor het lezen van prostaatMRIs. Dankzij jou werden mijn papers altijd voorzien van de juiste klinische context en hielp je me mijn onderzoek te focussen op de echte problemen, waarvoor ik je heel erg wil bedanken.

Prof. Witjes, Prof. Niessen en Dr. Veltman, bedankt dat jullie de tijd en moeite hebben genomen om zitting te nemen in mijn manuscriptcommissie en mijn werk te beoordelen, ik kijk er naar uit om met jullie van gedachten te wisselen tijdens mijn verdediging.

Beste Oscar en Wendy, als co-promovendi aan de technische kant van het prostaatonderzoek hebben we een heleboel dingen samen aangepakt en opgelost. Daarnaast hebben we ook lang een kamer gedeeld, eerst in de barak, daarna op onze mooie nieuwe afdeling. Oscar, wij hebben samen veel aan het CAD-systeem gesleuteld, jij voor het ontdekken van lymfeklieruitzaaiingen, ik voor het vinden van de primaire prostaatkanker. Dit heeft tot een tweetal gezamenlijke publicaties geleid, en ik hoop dat je snel het lymfeklier-CAD-systeem kan afronden. Heel erg bedankt voor alle tijd en moeite die je in de evaluatiemodules van het CAD-systeem hebt gestoken. Wendy, alhoewel onze projecten wat minder aan elkaar gerelateerd waren, hebben we toch overlap kunnen vinden. Ik wil je nog bedanken voor jouw bijdrage aan de PROMISE12-challenge, zonder jouw inbreng als ‘second observer’ hadden we nooit zulke

leuke resultaten bereikt.

Beste Joyce, Eline en Thomas, als klinische prostaatonderzoekers hebben we op een aantal verschillende projecten samengewerkt. Joyce, het feit dat jij begon met bijhouden van een database waarin alle resultaten van de prostaatbiopten werden vastgelegd heeft mij erg geholpen bij het kunnen trainen van het CAD-systeem. Eline, wij hebben samen veel gekeken naar de farmacokinetische analyse van de contrast-MRI, waar we de eerste waren die hebben laten zien dat als je het goed doet, je ook die sequenties kan gebruiken om de agressiviteit van prostaatkanker te bepalen. Jouw inzet heeft er toe geleid dat ik een mooie tweede-auteurspublicatie in European Urology op mijn lijst heb staan. Thomas, jouw initiële werk naar het bepalen van de Gleasongraad van prostaatkanker door middel van ADC analyse heeft aan de basis gestaan van mijn eerste publicatie. Dankzij jouw creatieve ideeën, voorwerk en mooie manier van dingen verwoorden en opschrijven is dat artikel een groot succes geworden. Ik wil jullie alle drie bedanken voor jullie bijdragen aan mijn onderzoek.

Marnix, als mijn eerste echte testpersoon voor het gebruiken van het prostaatwerkstation in MeVisLab heb je een behoorlijke hoeveelheid bugs moeten trotseren voordat we een goed werkend prototype hadden. Inmiddels loopt het volgens mij redelijk (toch?) en heeft het in ieder geval geleid tot een aantal leuke ISMRM abstracts over de PCa-MAP studie. Bedankt voor je altijd positieve en constructieve feedback.

Ik wil ook graag alle andere leden van de prostaatwerkgroep bedanken voor hun commentaar en bijdragen aan mijn onderzoek. Ik vond onze tweewekelijkse meeting altijd erg verhelderend en het heeft mij meer inzicht gegeven in waar de problematiek in de prostaat-MRI analyse nog ligt.

Buiten de afdeling Radiologie wil ik ook graag dr. Christina Hulsbergen – van de Kaa en dr. Jeroen van der Laak van de afdeling Pathologie, mijn nieuwe werkgever, bedanken. Beste Christina, ik denk dat we blij kunnen zijn dat we zo'n goede en ervaren prostaatpathologe in ons ziekenhuis hebben. Jouw bereidheid om desnoods preparaten opnieuw te beoordelen en te annoteren heeft niet alleen bij mijn proefschrift geholpen, maar ook bij vele anderen voor mij. Je stond ook onmiddellijk voor me klaar toen ik halsoverkop 200 biopten verwerkt moest hebben omdat ik een maand later naar Cleveland vertrok. Beste Jeroen, bij dat laatste project heb jij me aan de technische kant geholpen met de scanmicroscoop en de beelden. Daarnaast werken we op dit moment prettig samen bij de pathologie, en ik hoop dat daar een sterke onderzoeksgroep uit kan ontstaan.

Prof. Madabhushi, dear Anant, thank you for allowing me to do an internship in your lab over the summer in 2013. During those three months I think we were extremely productive, resulting in two conference publications, one journal publication and another one on the way. This was mostly thanks to your drive and strong belief that things could be done, even if I thought that they couldn't. This not only resulted in those publications, but also in the second place in the student paper awards at SPIE2014. My stay also allowed me to pick up a lot on

pathology research and how to combine pathology and radiology.

Mirabela, we had many a beer together (if you ever visit the Netherlands, I still owe you one) and you introduced me to roller derby, which I still have trouble explaining to other people. Rob, Rachel, Ajay, George, Satish, Pallavi, Sahir, Angel, Eileen, Haibo, Asha, Mahdi, Andrew, Ahmad, Shoshana, Prateek and Ann, thank you for making me feel welcome at Case and in Cleveland.

Prof. van Ginneken, beste Bram, als een van de twee hoogleraren binnen DIAG heb je een grote invloed op het reilen en zeilen binnen de groep. Alhoewel je me nooit direct begeleid hebt binnen mijn PhD wil ik je toch graag bedanken voor de tijd die je hebt gestoken in de gesprekken over mijn toekomst en het vertrouwen in mij als onderzoeker. Ik heb veel opgestoken van hoe jij wetenschappelijke projecten leidt en de rigoureuze doorzaagsessies bij het oefenen van conferentiepraatjes hebben er voor gezorgd dat ik betere presentaties ben gaan geven.

Daarnaast wil ik ook alle andere onderzoekers binnen DIAG bedanken voor het creëren van een heel fijne omgeving om onderzoek te doen. Toen ik begon waren we nog een redelijk klein clubje van 11, maar inmiddels zijn we uitgegroeid tot een groep van bijna 50 man/vrouw. Ik heb het erg leuk gevonden om een klein deel uit te maken van dat proces. Ook hebben we natuurlijk de fantastische DIAG-weekenden gehad, maar ook de Sinterklaasavondjes, vrijdagborrels en andere feesten en partijen.

Colin en Tom, bedankt dat jullie mijn paranimfen willen zijn (al hoop ik dat jullie het niet al te druk zullen krijgen). Colin, jij was één van de eerste studenten uit Eindhoven die na mij bij DIAG terecht kwam. Daarvoor kende ik je natuurlijk al van onze redelijk legendarische ECR uitstapjes met de BMIA-groep uit Eindhoven (gelukkig hebben we de foto's nog). Ik denk dat wij een beetje hetzelfde in het leven staan en over veel dingen hetzelfde denken. Dat zal er mede toe geleid hebben dat we altijd erg veel lol hadden op de vele conferenties waar we zijn geweest. Je gaat inmiddels zelf richting de laatste loodjes van je PhD, heel veel succes daarmee, alhoewel dat jou kennende niet zo'n groot probleem zal zijn.

Tom, wij kennen mekaar al sinds we konden kruipen (zo ongeveer) en zijn al bijna 30 jaar vrienden. Je hebt in je leven een behoorlijk wat ellende meegemaakt en ik heb er altijd enorm veel respect voor gehad hoe je daarmee omging en dat je altijd vrolijk en positief bent gebleven. Vroeger, toen ik wat meer verlegen was, heb je me regelmatig op sleeptouw genomen en ik heb daar veel aan gehad. Ook bedankt daarvoor en binnenkort maar weer eens een avondje naar Van Mol!

Naast Tom kan ik ook de rest van de vriendenclub uit Helden natuurlijk niet vergeten. Bart, Chris, Dennis, Guyon, Jaap, Joep, Marcel, Paul, Rogier, Roy en Sjors, ook wij kennen mekaar al een enorm lange tijd. Door de jaren heen zijn er vaak stapavonden en vakanties geweest, die inmiddels een beetje vervangen worden door house-warmings en baby-borrels. Als ik niet beter wist zou ik bijna zeggen dat we volwassen aan het worden zijn. Er zijn dan nog af en toe

van die vriendenweekenden waar het tegendeel dan weer uit blijkt. Ik prijs mezelf gelukkig met zo'n hechte vriendengroep.

Lieve Lisanne, we zijn alweer een tijdje uit elkaar en je bent begonnen aan je eigen PhD-avontuur. Toch wil ik je bedanken voor het feit dat je er altijd voor me was als ik je nodig had. We zijn bijna mijn hele promotietraject bij elkaar geweest, en we hebben een heel fijne tijd samen gehad, waar ik veel mooie herinneringen aan bewaar, onder andere de vakanties naar China, Amerika en Egypte. Ik wens je alle goeds! Ook wil ik Paul, Marlie, Ceryl, Ramon en Aimée bedanken, ik voelde me altijd erg thuis als ik bij jullie op bezoek was en de BBQ/'De Tent Sjteit'-weekenden zal ik niet snel vergeten.

Ik kan natuurlijk niet mijn vriendjes en vriendinnetjes van Uisge Beatha, oftewel UB, vergeten. Ik denk dat mijn studententijd (en ook nog een flink stuk van mijn promotie) een stuk minder gezellig zou zijn geweest zonder jullie. In het bijzonder wil ik nog even René, Danny en Wai Kam noemen, mijn mede-pre-distillaten waarmee ik samen toetrad tot dit illustere dispuut. Moge we mekaar nog vele reünistenactiviteiten en UBBQs zien.

Lieve Eva, toen wij mekaar voor het eerst ontmoetten woonde jij nog in Los Angeles. Om een of andere vage reden besloot je toch terug te verhuizen naar dit koude kikkerlandje. Sinds een paar maanden zijn wij bij elkaar en kan ik alleen maar heel blij zijn met die vlaag van verstandverbijstering. Ik ben heel gelukkig met jou en ik hoop dat dat nog jaren zo mag blijven.

René, lieve broer, ik wil je laten weten dat ik heel erg trots op je ben. Je hebt het gedurende je opleiding niet altijd makkelijk gehad, maar je hebt je nooit laten ontmoedigen om je droom na te jagen. Dat is voor mij ook een voorbeeld om nooit op te geven als je iets echt wilt bereiken. Inmiddels heb je een vaste baan als leraar en ga je een huis kopen samen Iris, dus je kan wel zeggen dat je het goed voor elkaar hebt! Ik wens jullie heel veel geluk samen.

Opa en oma, ik wil jullie graag bedanken voor het feit dat jullie er altijd voor me zijn. Fijnere grootouders kan niemand zich wensen, van de reepjes chocola en de tien guldens die jullie ons toestopten toen we klein waren tot de telefoontjes om te informeren of ik toch echt niet te lang naar het buitenland wil gaan.

Als laatste, lieve pap en mam, zonder jullie zou ik nooit geworden zijn wie ik nu ben. Jullie hebben me altijd gesteund in mijn keuzes en gestimuleerd om het beste uit mezelf te halen. En als ik eens een keer iets heel doms deed, stonden jullie toch voor me klaar. Het is altijd enorm fijn om thuis te komen.

# Curriculum Vitae

## Curriculum Vitae

Geert Litjens was born in Venlo, the Netherlands, on April 4th 1985. After high school at the Bouwens van der Boije-college (Panningen) he studied Biomedical Engineering at the Eindhoven University of Technology. His MSc.-thesis on "Pharmacokinetic modeling in breast cancer MRI" was a collaboration with Philips Healthcare in Best. As of January 2010, he worked as a PhD-student in the Diagnostic Image Analysis Group. His work on computer-aided detection of prostate cancer is described in this thesis.