Doctoral theses at NTNU, 2021:264

Doctoral thesis

Mohammed Rasem Sadeq Sunoqrot

Computer-Aided Diagnosis of Prostate Cancer Using Multiparametric MRI: Pre-processing, Segmentation and Quality Control

NTNU
Norwegian University of Science and Technology
Thesis for the Degree of
Philosophiae Doctor
Faculty of Medicine and Health Sciences
Department of Clinical and Molecular Medicine

NTNU
Norwegian University of
Science and Technology

Mohammed Rasem Sadeq Sunoqrot

Computer-Aided Diagnosis of Prostate Cancer Using Multiparametric MRI: Pre-processing, Segmentation and Quality Control

Thesis for the Degree of Philosophiae Doctor

Trondheim, August 2021

Norwegian University of Science and Technology
Faculty of Medicine and Health Sciences
Department of Clinical and Molecular Medicine

NTNU
Norwegian University of
Science and Technology

NTNU
Norwegian University of Science and Technology

Thesis for the Degree of Philosophiae Doctor

Faculty of Medicine and Health Sciences
Department of Clinical and Molecular Medicine

© Mohammed Rasem Sadeq Sunoqrot

ISBN 978-82-326-6531-0 (printed ver.)
ISBN 978-82-326-6417-7 (electronic ver.)
ISSN 1503-8181 (printed ver.)
ISSN 2703-8084 (online ver.)

Doctoral theses at NTNU, 2021:264

Printed by NTNU Grafisk senter

Sammendrag

Dataassistert diagnostikk av prostatakreft ved bruk av Multiparametrisk MRI: Forbehandling, segmentering og kvalitetskontroll

Prostatakreft er den vanligste kreftformen hos menn og den nest hyppigste årsaken til kreftrelaterte dødsfall hos menn på verdensbasis. På grunn av fremskritt innen teknologi og diagnostiske metoder har overlevelsesraten for prostatakreft de siste årene økt og dødeligheten har sunket. Tidlig diagnostikk av prostatakreft er viktig for bedre behandling av sykdommen. Den tradisjonelle diagnostiske prosessen inkluderer måling av forhøyet prostata spesifikt antigen (PSA) i blodet etterfulgt av prøvetaking av prostata biopsi og histopatologisk analyse. Multi-parametrisk magnetisk resonans avbildning (mpMRI) og etablering av internasjonale retningslinjer for bildeopptak og tolkning har bidratt til bedre nøyaktighet i diagnostikken, men tolkningen av MR-bildene er fortsatt i stor grad kvalitativ. Dette har noen begrensninger, for eksempel at tolkningen krever erfarne radiologer, variasjon mellom observatører og at det er tidkrevende arbeid. Med innføring av pakkeforløp for prostatakreft i Norge har antallet MR undersøkelser som gjennomføres for deteksjon av prostatakreft økt kraftig, og det er krevende å skalere opp de nødvendige radiolog-ressursene for å holde tidsrammene som er angitt i pakkeforløpet. Automatiske dataassisterte deteksjons- og diagnosesystemer (CAD) har potensial til å overvinne disse begrensningene ved å bruke MR-bildene i kvantitative modeller som automatiserer, standardiserer og støtter reproduserbar tolkning av radiologiske bilder.

Den automatiserte CAD-arbeidsflyten består av flere trinn, for eksempel normalisering og segmentering, før bildene så kan benyttes til å etablere diagnostiske modeller basert på maskinlæring (ML) eller dyp læring (DL). For å sikre effektiv og pålitelig beslutningsstøtte, må alle trinn i arbeidsflyten være generaliserbare, transparente og robuste.

CAD for diagnostikk av prostatakreft har ennå ikke blitt innlemmet i klinisk praksis. Målet med denne avhandlingen var derfor å legge til rette for dette ved å utvikle og evaluere nye metoder for bildebehandling, segmentering og kvalitetskontroll for å forbedre generaliserbarheten, gjennomsiktigheten og robustheten til arbeidsflyten i CAD.

Denne avhandlingen er basert på tre artikler. I Artikkel I ble en ny automatisert metode for normalisering av T2-vektede (T2W) MR-bilder av prostata utviklet og evaluert ved bruk av to referansevev (fett og muskler). Metoden reduserer intensitetsforskjeller mellom ulike MR-bilder og forbedrer med dette den kvantitative vurderingen av prostatakreft. Artikkel II og III fokuserer på segmenteringsmetoder basert på DL. I Artikkel II ble et helautomatisk kvalitetskontrollsystem for DL-basert prostata-segmentering fra T2-vektete MR-bilder etablert og evaluert. Kvalitetskontrollen identifiserer når segmenteringen blir unøyaktig, og hindrer dermed at senere trinn i CAD-systemet baseres på feilaktig informasjon. I Artikkel III blir reproduserbarheten av DL-basert segmentering av hele prostatakjertelen og prostatasoner vurdert. Dette er spesielt viktig for applikasjoner hvor pasienten følges opp med flere MR-undersøkelser over tid (aktiv overvåkning). Forskningsresultatene viser at reproduserbarheten til den beste DL-baserte prostata-segmenteringsmetoden er sammenlignbar med manuell segmentering.

Kort oppsummert viser avhandlingen hvordan avanserte, generaliserte og kontrollerte metoder for bildeforbehandling og kvalitetskontroll kan bidra til å forbedre ytelsen og tilliten til CAD-basert beslutningstøtte for diagnostikk av prostatakreft, noe som er et viktig skritt mot klinisk implementering.

Kandidat: Mohammed Rasem Sadeq Sunoqrot

Institutt: Institutt for Sirkulasjon og bildediagnostikk

Veiledere: Mattijs Elschot, Tone F. Bathen, Kirsten M. Selnæs, Harald Martens

Finansiering: NTNU Biotechnology

Ovennevnte avhandling er funnet verdig til å forsvares offentlig for graden
Philosophiae Doctor (PhD) i medisinsk teknologi.
Disputas finner sted digitalt via Zoom,
Tirsdag 24. August 2021 kl. 12:15.

Acknowledgement

The work presented in this thesis was carried out at the MR Cancer group, Department of Circulation and Medical Imaging, Faculty of Medicine and Health Sciences, Norwegian University of Science and Technology (NTNU), between March 2017 and April 2021. Financial support was provided by the NTNU Biotechnology (grant number 81770928).

I would like to thank everyone who supported me and contributed in any way to the completion of this thesis. I would especially like to thank all the patients who allowed us to use their data for scientific research, without you this work would not have been possible.

First and foremost, I would like to express my sincere gratitude to my supervisors, Dr. Mattijs Elschot, Prof. Tone F. Bathen, Dr. Kirsten M. Selnæs and Prof. Harald Martens. Thank you for always being there for discussions, encouraging me and guiding me through this journey. Mattijs, I cannot thank you enough for your guiding, you were always there for my questions no matter how silly they were, you helped me with every step of the work, you always listened to my ideas, thoughts and problems and you made sure that I have the continuous support and supervision to finish my thesis. Tone, thank you for welcoming me to the group, guiding me and helping me in every way possible. Thank you for creating this great research environment that I can't imagine would be possible without you. Kirsten, thank you for always being there when I needed you, helping me with data collection and being open to my questions even when you left the group. Harald, I know we didn't have the opportunity to work together as much as we had planned, but your philosophy on transparent artificial intelligence algorithms has left its mark on this work.

I would also like to thank all the internal and external collaborators with whom I enjoyed working. In particular, I would like to thank Dr. Elise Sandsmark who put a lot of effort into segmenting all these images. I would also like to thank all those who made their data and methods publicly available.

A sincere thank you to all my current and former colleagues at the MR center, you have created a warm, positive and dynamic work environment. Debbie thank you for being the

social engine of the group. Torill, you are, as a wise man once said, the silent engine behind the group, thank you for all your help. Gabriel, the wise man, thank you for always helping me when needed. Daniel, thanks for being a good friend and for dragging me to bouldering. Thanks to the sushi guys (Alex, Torfinn and Chris) for being there for the sushi exploration trip. Alex, I'd also like to thank you for all the coffee breaks and chats. I would also like to thank the rest of my office mates: Hanna Maja, Maren, Kaia and Bendik.

Finally, I would like to thank my family and friends for their support and love. The sincerest gratitude, thank and love to my father, Rasem, and my mother, Rola, without whom I would never have gotten to this point. Their prayers, support, care and love have been the reason behind everything good I have had in life. My siblings Ayat, Momen, Janat, Gadeer and Asem, thank you for always being there for me. To my uncles, aunts, cousins and friends, thank you.

والحمد لله الذي علم الإنسان ما لم يعلم

Mohammed Rasem Sadeq Sunoqrot

Trondheim, April 2021

Summary

Computer-Aided Diagnosis of Prostate Cancer Using Multiparametric MRI: Pre-processing, Segmentation and Quality Control

Prostate cancer is the most commonly diagnosed cancer in men and the second leading cause of cancer-related deaths in men worldwide. In recent years, and due to advances in technology and diagnostic procedures, prostate cancer survival rates have increased and mortality rates have decreased. Early diagnosis of prostate cancer is critical for better treatment of the disease. The traditional diagnostic process includes measuring elevated prostate-specific antigen (PSA) in the blood followed by prostate biopsy sampling and histopathology analysis. The addition of multiparametric magnetic resonance imaging (mpMRI) and the establishment of international guidelines for image acquisition and interpretation have improved prostate cancer diagnosis. Typically, interpretation of mpMR images is performed qualitatively by a radiologist. This approach has a number of limitations, such as high inter-observer variability, time-consuming nature, dependence on reader opinion and lack of scalability of the manual data processing approach as demand increases. Automated computer-aided detection and diagnosis (CAD) systems have the potential to overcome these limitations and utilize mpMRI by implementing quantitative models to automate, standardize and support reproducible interpretation of radiological images.

The automated CAD workflow typically consists of a machine learning algorithm, preceded by several stages of image processing, including pre-processing, segmentation, registration, feature extraction and classification. Each stage depends on the previous stages to finally produce an accurate diagnosis. Errors in any of the stages of the workflow, but especially in the early pre-processing stages, will propagate through the pipeline and can lead to a misdiagnosis of the patient. Consequently, to provide an efficient and trustworthy diagnosis, each stage of a CAD system should be generalizable, transparent and robust.

Despite a growing body of evidence showing potential, CAD of prostate cancer has not yet been integrated into clinical practice. This is mainly due to the lack of generalizability, transparency and robustness, which causes a lack of confidence of the radiologists in the capabilities of CAD. To increase the confidence in CAD, its performance should be improved, controlled and generalized. Therefore, the aim of this thesis was to facilitate the integration of automated CAD systems for prostate cancer using mpMRI into clinical practice by developing and evaluating new image normalization, segmentation and quality control methods to improve the generalizability, transparency and robustness of the CAD workflow.

This thesis is based on three papers. In Paper I, a novel automated method for prostate T2-weighted (T2W) MR image normalization using dual-reference tissue (fat and muscle) was developed and evaluated. The method was shown to reduce T2W intensity variation between scans and to improve quantitative assessment of prostate cancer on MRI. Papers II and III focused on deep learning (DL)-based prostate segmentation. In Paper II, a fully automated quality control system for DL-based prostate segmentation on T2W MRI was established and evaluated. The system was able to assign an appropriate score based on extracted image features, reflecting the quality of the generated segmentations. This score can be used to distinguish between acceptable and poor DL-based segmentations. In Paper III, the reproducibility of the DL-based segmentations of the whole prostate, peripheral zone, and remaining prostate zones was investigated. This is important for implementing DL-based segmentation methods in CAD system for clinical applications that depend on multiple scans. The study showed that the reproducibility of the best performing DL-based prostate segmentation methods is comparable to that of manual segmentations.

In summary, in this thesis advanced image pre-processing and quality control methods were developed and evaluated for CAD of prostate cancer using mpMRI. Ultimately, these automated methods can help improve the performance of and increase the confidence in CAD systems, which is an important step towards their implementation in clinical practice.

Symbols and Abbreviations

^{1}H	Hydrogen
ACF	Aggregate channel features
ADC	Apparent diffusion coefficient
AI	Artificial intelligence
AUC	Area under the receiver operating characteristic curves
AutoRef	Dual-reference tissue normalization
B_0	Main magnetic field strength
B_1	Strength of temporarily applied radiofrequency pulse
BPH	Benign prostatic hyperplasia
CAD	Computer-aided detection and diagnosis
CNN	Convolutional neural network
DCE	Dynamic contrast enhanced
DL	Deep learning
DRE	Digital rectal examination
DSC	Dice similarity coefficient
DWI	Diffusion-weighted imaging
FAIR	Findability, accessibility, interoperability, and reusability
FID	Free induction decay
GDPR	General Data Protection Regulation
GPU	Graphics processing unit
ICC	Intra-class correlation coefficient
LASSO	Least absolute shrinkage and selection operator
M	Non-zero net magnetization vector
M_0	Non-zero net magnetization vector under thermal equilibrium
ML	Machine learning
mpMRI	Multiparametric magnetic resonance imaging
MRI	Magnetic resonance imaging
NMR	Nuclear magnetic resonance
non-PZ	Central, transition, and anterior fibromuscular stroma zones of the prostate, combined
PACS	Image Archiving and Communication System
PI-RADS	Prostate imaging-reporting and data system
PROMISE12	Prostate MR image segmentation challenge
PSA	Prostate-specific antigen
PZ	Peripheral zone
QC	Quality control
RF	Radio frequency
ROI	Region-of-interest
SVM	Support vector machine
T2W	T2-weighted imaging

TE	Echo time
TNM	Tumor Node Metastasis
TR	Repetition time
TRUS	Transrectal ultrasound
VOI	Volume-of-interest
WP	Whole prostate
γ	Gyromagnetic ratio
ω0	Larmor frequency

List of papers

Paper I

Automated reference tissue normalization of T2-weighted MR images of the prostate using object recognition

Mohammed R. S. Sunoqrot, Gabriel A. Nketiah, Kirsten M. Selnæs, Tone F. Bathen, Mattijs Elschot.

Magnetic Resonance Materials in Physics, Biology and Medicine 2021; 34(2):309-321.

Paper II

A quality control system for automated prostate segmentation on T2-weighted MRI

Mohammed R. S. Sunoqrot, Kirsten M. Selnæs, Elise Sandsmark, Gabriel A. Nketiah, Olmo Zavala-Romero, Radka Stoyanova, Tone F. Bathen, Mattijs Elschot.

Diagnostics 2020; 10(9):714.

Paper III

The reproducibility of deep learning-based segmentation of the prostate gland and zones on T2-weighted MR images

Mohammed R. S. Sunoqrot, Kirsten M. Selnæs, Elise Sandsmark, Sverre Langørgen, Helena Bertilsson, Tone F. Bathen, Mattijs Elschot.

Submitted

Contents

1 Introduction

1.1 Cancer

Cancer is a general term for a large group of heterogeneous, convoluted diseases characterized by unregulated cell division and growth in the body [1]. There are more than 100 different types of cancer that can affect humans [1], which are thought to share a number of molecular, biochemical and cellular characteristics that ensure the survival, proliferation and spread of cancer cells [2]. Hanahan and Weinberg referred to these characteristics as *hallmarks of cancer* and listed them as “self-sufficiency in growth signals”, “insensitivity to growth-inhibitory signals”, “evasion of apoptosis”, “limitless replicative potential”, “sustained angiogenesis”, and “tissue invasion and metastasis” [2]. In 2011, two *emerging hallmarks* – “deregulating cellular energetics” and “avoiding immune destruction” – and two *enabling characteristics* – “genome instability and mutation” and “tumour promoting inflammation” – were added to this list [3].

Cancer is one of the leading causes of premature death worldwide, with 9.6 million cancer deaths and 18.1 million estimated new cancer cases in 2018 [4]. The most commonly diagnosed cancers are breast, colorectal and lung cancer in women and lung, prostate and colorectal cancer in men [4]. Despite the complexity of cancer and the high incidence rates, mortality rates have decreased in recent years [5], which can be attributed to improvements in cancer diagnosis and treatment procedures.

1.2 Prostate anatomy and function

The human prostate is a walnut-sized accessory genital gland composed of 70% glandular tissue and 30% fibromuscular or stromal tissue, surrounded by a thin fibrous capsule. It is part of the male reproductive tract and is located anterior to the rectal ampulla between the bladder neck at the base and the pelvic floor at the apex and surrounds the uppermost part of the urethra (Figure 1.1 A) [6-8].

The prostate is divided into four histological zones: peripheral zone, central zone, transition zone and anterior fibromuscular stroma (Figure 1.1 B). The peripheral zone is a horseshoe-shaped region composed of branched glands; it occupies approximately 70% of the prostate volume in young men and covers the distal prostatic urethra at the apex and extends posterolaterally to the base. The central zone is an inverted cone-shaped region composed of periurethral mucosal glands; it occupies about 25% of the prostatic volume and is located posterior to the urethra, surrounds the ejaculatory ducts, and makes up most of the gland base.

The transition zone is an annular region consisting of periurethral submucosal glands; it occupies about 5% of the prostate volume and is located in the glandular centre surrounding the urethra and makes up a large portion of the midgland. As men age, the transition zone tends to enlarge and develop a non-cancerous condition called benign prostatic hyperplasia (BPH). The anterior fibromuscular stroma is a thickened area composed of muscle fibres and fibrous connective tissue that surrounds the anterior and anterolateral surfaces of the prostate [6-9].

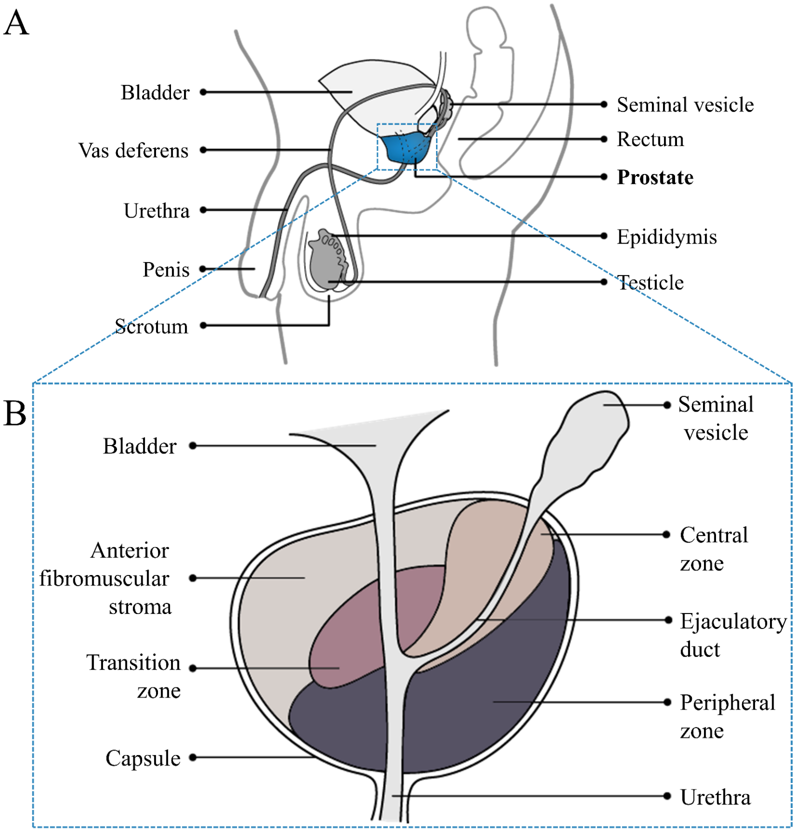


Figure 1.1: Sagittal view of the location and anatomy of the prostate.

A) The anatomical location of the prostate, between the bladder neck and the pelvic floor. B) The four histological zones of the prostate: peripheral zone, central zone, transition zone and anterior fibromuscular stroma. *Adapted and edited from [10].*

The main function of the prostate is to secrete a slightly alkaline prostatic fluid containing calcium, citrate ions, phosphate ions, a coagulating enzyme and a profibrinolysin. This fluid is added to the semen during ejaculation. The properties of the prostatic fluid help enhancing the sperm fertility and the ability of spermatozoa to move independently [11].

1.3 Prostate cancer

Prostate cancer is a heterogeneous type of cancer that begins in the mucus-producing glandular cells and ranges from slow-growing and indolent to very aggressive [12]. About 70-80%, 20-25% and 5-10% of prostate cancers originate in the peripheral, transition and central zone, respectively [6,7,12].

Prostate cancer is the second most commonly diagnosed cancer and the second leading cause of cancer-related deaths in men worldwide, with an estimated 358,989 deaths and 1,276,106 new cases in men in 2018 [4]. In Norway, prostate cancer has the highest cancer incidence rate in total with 4,877 new cases in 2019, which is slightly lower (about 7.4%) than in previous years, indicating a stabilization in the incidence rate. The stabilization may be due to the decrease in elevated prostate specific antigen (PSA) testing. Despite the increase in the prostate cancer incidence rate over the last two decades, it has been shown that the mortality rate has decreased while the 5-year survival rate has increased (Figure 1.2). This shift can be attributed to early and improved detection and treatment of prostate cancer [13].

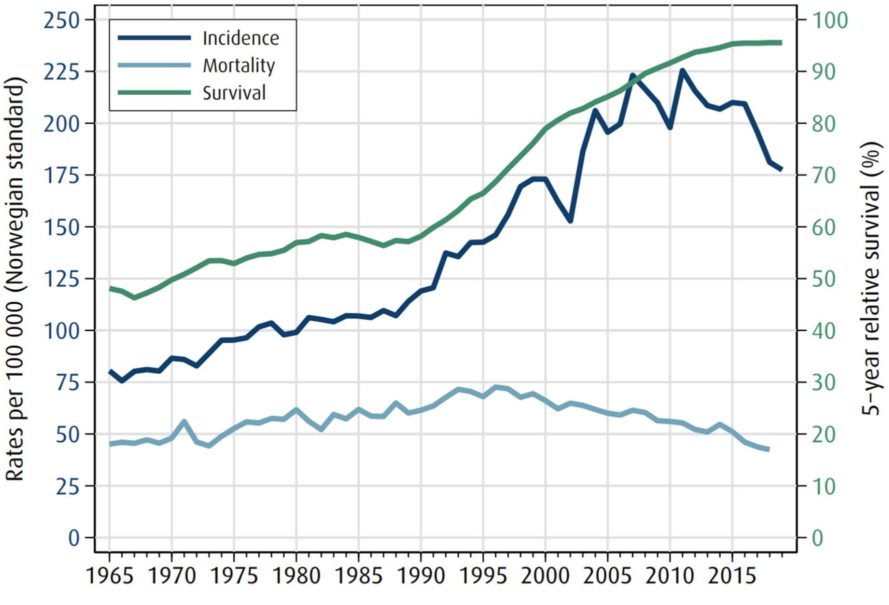


Figure 1.2: Trends in incidence and mortality rates and 5-year relative survival rate of prostate cancer in Norway.

Incidence, mortality and 5-year relative survival rate of cancer in Norway for the last 54 years. Since the mid-1990s, incidence and survival rates have increased, while mortality rate have decreased. *Adapted from [13] with permission.*

1.3.1 Risk factors

Age, ethnicity and family history are established risk factors for prostate cancer [12,14-16]. Prostate cancer is rare in men younger than 50 years, while it is more likely in men aged 65-75 years [12]. The likelihood of developing prostate cancer has been shown to be higher in men of African descent, while it is lower in men of Asian descent [12,14-16]. The reason for this is unclear, but it has been speculated that it may be related to the gene pool [12]. Family history is an important factor, as the presence of a first-degree relative with prostate cancer history increases the risk by twofold [17] and the presence of multiple relatives with prostate cancer history increases the risk by up to fivefold due to the presence of multiple risk alleles [12]. High fat, high processed carbohydrate diet, low physical activity level, unhealthy lifestyle, harmful work environment, obesity and smoking have also been associated with the risk of developing prostate cancer [14-16,18,19]. Therefore, a combination of several factors increases the risk of developing prostate cancer.

1.3.2 Clinical presentation

The symptoms of prostate cancer are similar to those of BPH. They may include one or more of the following symptoms: Urinary tract obstruction, dysuria, urinary incontinence, nocturia or hematuria. These symptoms occur as the cancer progresses, whereas most prostate cancers are asymptomatic at the time of diagnosis [12,20,21]. The impact of prostate cancer on urinary function is due to the location of the prostate gland, as it surrounds the urethra and abuts the bladder neck. Due to the prostate's function of secreting prostatic fluid and mixing it with seminal fluid from the vas deferens, the changes in the prostate resulting from the developed cancer can lead to complications related to sexual function and performance, such as difficulty achieving an erection or painful ejaculation [12,20]. Bone pain is the presenting symptom in men with metastatic prostate cancer, but the initial diagnosis of such condition is rare, with only 6% of men with prostate cancer having metastatic disease at diagnosis [22].

1.3.3 Detection and diagnosis

The main diagnostic procedure of prostate cancer includes measuring the elevated PSA level in the blood, digital rectal examination (DRE), transrectal ultrasound (TRUS)-guided prostate biopsy sampling and histopathological analysis [23]. More recently, the use of multiparametric magnetic resonance imaging (mpMRI) was also added to the recommended diagnostic procedure [24].

Testing for elevated PSA is the most commonly used test in prostate cancer diagnosis and management [12]. PSA is a protein produced by the glandular cells of the prostate. When prostate cancer attacks the tissue barriers, PSA leaks into the bloodstream causing an elevated PSA level in the blood serum. Normal PSA level is usually below 4 ng/ml in old men and below 3 ng/ml in young men [21,23]. The elevation of PSA level is not limited to the development of cancer; it can also be caused by BPH, prostatitis or urinary tract infection [12,25]. In addition, prostate cancer may still exist despite low PSA level [26,27]. Although PSA testing improves the initial diagnosis of prostate cancer patients, the sensitivity and specificity are still low. Because of its low specificity, PSA may lead to overdiagnosis and overtreatment in some men [28-30]. Therefore, the Norwegian Directorate of Health, the Unites States Preventive Services Task Force and the European Society of Urogenital Radiology do not recommend PSA screening in healthy men [23,31,32]. DRE examination, which is performed in addition to PSA testing, is part of the usual primary care routine in men [33,34]. The DRE examination is a test in which the physician inserts a gloved finger into the rectum to palpate the prostate and examine for lumps or abnormalities. This exam can help detect some of the aggressive tumours that do not have an elevated PSA. However, DRE might fail to detect localized and less aggressive tumours, raising questions about its role in early detection of prostate cancer [35]. Similar to the elevated PSA test, DRE testing has shown a high false-positive rate, leading to overdiagnosis and overtreatment [36,37].

Due to the limitations of the PSA and DRE tests, suspicious findings must be confirmed by TRUS-guided biopsy sampling followed by histopathologic analysis. In TRUS-guided biopsy sampling, an ultrasound-guided needle is used to schematically sample 10-12 cores from the prostate [38]. Because prostate tumours are heterogeneous and multifocal, the underlying structures are often difficult to capture with a needle biopsy. This sometimes leads to differences between the aggressiveness assessment from TRUS-guided biopsies and subsequent radical prostatectomy specimens [39,40]. TRUS is also unable to visualize most prostate cancer tumours and may fail to detect up to 35% of carcinomas at initial biopsy, which pushed towards utilising a high resolution MRI scan prior to biopsy [41]. The MRI scan can then be used to guide biopsy sampling by model-based MRI-ultrasound fusion, MRI-directed cognitive fusion or directly in the MRI scanner [41-43]. The MRI scan before biopsy is usually evaluated according to the standardized guidelines "Prostate Imaging-Reporting and Data System (PI-RADS)" and the clinical suspicion of the presence of cancer to decide whether biopsy sampling is required [44,45].

1.3.4 Histopathological evaluation

The aggressiveness of prostate cancer is assessed by the Gleason score. Gleason score is assigned by a pathologist after viewing a biopsy or prostatectomy material. The Gleason score is a histologic scoring system that describes the appearance, patterns and organisational structure of the cancerous epithelial cells of prostate cancer [46]. The Gleason scoring system ranges from 1 to 5, with 1 representing a nearly normal cells pattern and appearance and 5 representing the presence of exclusively abnormal cancerous epithelial cells. The Gleason score contains two grades, the most common and the second most common pattern Gleason grade in the biopsy, that together make up the total score, with the lowest total score being 2 (1 + 1) and the highest being 10 (5+5) [47]. For more accurate assessment, the International Society of Urologic Pathology proposed a new classification system "Grade Groups", which was adapted by the World Health Organisation in 2016. The new system includes 5 grade groups (1-5) with prognostic differences corresponding to Gleason scores 3+3, 3+4, 4+3, 8 (4+4; 3+5; 5+3) and 9-10 (4+5; 5+4; 5+5), respectively [48,49].

1.3.5 Staging and prognostics

Determining the stage of prostate cancer is important to define the prognosis of the disease and to choose the appropriate therapy. The TNM classification system – primary tumour (T), regional lymph nodes (N), and distant metastases (M) – (Table 1.1) is the most common system for prostate cancer staging. T stage is determined based on findings from DRE, number and location of positive TRUS biopsies and MRI [50].

To aid in treatment decisions, prostate cancer prognostics are grouped into four stages based on PSA level, Gleason grade group and TNM categories [50]. Table 1.2 shows and describes each of these stages. The same clinical variables are used by the European Association of Urology to group patients with a similar risk of biochemical recurrence (see *Section 1.3.7*) after initial treatment. Accordingly, the patients are stratified into low-, intermediate- and high-risk groups (Table 1.3) [45].

Table 1.1: Tumour Node Metastasis (TNM) Classification system for prostate cancer.

Category	Definition/Criteria
T – Primary tumour	
TX	Primary tumour cannot be assessed
T0	No evidence of primary tumour
T1	Clinically inapparent tumour that is not palpable
T2	Tumour is palpable and confirmed within the prostate
T2a	Tumour involves one-half of one side or less
T2b	Tumour involves more than one-half of one side but not both sides
T2c	Tumour involves both sides
T3	Extraprostatic extension
T4	Tumour is fixed or invades adjacent structures other than seminal vesicles
R – Regional lymph nodes	
NX	Regional lymph nodes were not assessed
N0	No positive regional lymph nodes
N1	Metastases in regional lymph node(s)
M – Distant metastasis	
M0	No distant metastasis
M1	Distant metastasis

Adapted from [50,51] with permission.

Table 1.2: Prognostic stage grouping for prostate cancer.

Stage	T status	N status	M status	PSA level (ng/mL)	Grade Group	Spreading
I	T1, T2a	N0	M0	<10	1	**Localized**
IIA	T1, T2a-c	N0	M0	≥10, <20	1	
IIB	T1, T2	N0	M0	<20	2-4	
IIIA	T1, T2	N0	M0	≥20	1-4	**Locally advanced**
IIIB	T3, T4	N0	M0	Any	1-4	
IIIC	Any T	N0	M0	Any	5	
IVA	Any T	N1	M0	Any	Any	
IVB	Any T	Any N	M1	Any	Any	**Metastatic**

Adapted from [50,51] with permission.

Table 1.3: Risk groups for biochemical recurrence of prostate cancer.

Risk group	Definition				
	PSA (ng/mL)		Gleason score		Clinical stage
Low-risk	<10	AND	<7	AND	T1, T2a
Intermediate-risk	10–20	OR	7	OR	T2b
High-risk	≥20 Any	OR	>7 Any	OR OR	T2c T3

Adapted from [45,51] with permission.

1.3.6 Management and treatment

The next step for men diagnosed with prostate cancer is to proceed with either disease management or treatment. Early-stage patients with indolent or slow-growing cancer, or patients with short life expectancy will proceed with disease management, which is divided into active surveillance and watchful waiting [52]. In active surveillance, the patient is followed-up to monitor the disease progression so that intervention can be made as early as possible if the cancer begins to behave more aggressively. Monitoring in active surveillance may include PSA testing, DRE, biopsy sampling and MRI scans. In watchful waiting, the patient is treated for symptoms only and no palliative treatment is given unless advanced symptomatic disease develops [52,53].

Treatment of prostate cancer is determined based on disease progression and location, in addition to other factors such as age, life expectancy and side effects. Treatment may include one or a combination of external beam radiotherapy, brachytherapy, cryosurgery, high-intensity focused ultrasound, and prostatectomy if the cancer has not spread from the prostate. In case of development of metastatic cancer, chemotherapy and hormonal therapy are the usual treatment choices [54-57]. Each of these treatments has side effects, such as erectile dysfunction, rectal bleeding and urinary incontinence, in addition to the constant stress, anxiety and lifestyle changes [58]. Therefore, accurate diagnosis and assessment of prostate cancer is necessary to select the most appropriate disease management and treatment and to avoid over- or under-treatment.

1.3.7 Biochemical recurrence

The PSA level is expected to drop a few weeks after treatment until it becomes undetectable or returns to baseline levels, so an increase in PSA may be an indicator of prostate cancer recurrence [59]. Biochemical recurrence is the continuous rise in PSA after treatment. It is defined as serum PSA ≥ 0.02 ng/ml in two independent measurements after radical prostatectomy or an increase in serum PSA ≥ 2 ng/ml above baseline after radiotherapy [60,61]. Biochemical recurrence occurs in 20-40% of patients after radical prostatectomy and in 30-50% of patients after radiotherapy within 10 years of treatment [62,63]. Patients with biochemical recurrence are considered to have prostate cancer recurrence, even in the absence of symptoms and signs of local or metastatic disease [64]. The management of biochemical recurrence is challenging, as the spread of the cancer should be stopped without over-treating the patient or negatively affecting his quality of life [64].

1.4 Magnetic resonance imaging

Magnetic resonance imaging (MRI) is a noninvasive medical imaging modality based on the principles of nuclear magnetic resonance (NMR) that uses nonionizing radiation to produce images of the anatomy and functional and physiological processes of the body. MRI is typically used to scan soft tissues because of its exceptional soft tissue contrast and high sensitivity to a variety of tissue properties [65,66]. These characteristics have made MRI a useful tool for diagnosis and repeated assessment of the progression of various diseases, including cancer. Therefore, MRI has become a popular tool for prostate cancer diagnosis, active surveillance monitoring and treatment evaluation [67]. In Norway, MRI examination is currently the first step in the standardized care path for patients suspected of having prostate cancer based on PSA test and/or DRE [68].

In 1938, Isidor Rabi first described NMR. He realized that atomic nuclei, when exposed to a strong magnetic field, can absorb or emit radio waves [69]. In 1946, Felix Bloch and Edward Purcell observed the NMR phenomenon in liquids and solids; they discovered that atomic nuclei with angular momentum (spin) can interact with a magnetic field [70,71]. In 1973, Peter Mansfield and Paul Lauterbur described how NMR can be used to generate images [72,73]; this can be considered the basis of what is now known as MRI.

1.4.1 Basics of nuclear magnetic resonance

NMR principles are based on the property of spinning motion of atomic nuclei. Inside the nucleus, the protons and neutrons spin in opposite directions with a value of ½. The nucleus with an even number of protons and neutrons ends up with a net spin of zero, while the nucleus with an odd mass number ends up with a non-zero net spin and thus a magnetic moment. MR uses spin -½ nuclei, e.g. hydrogen-1 (^{1}H), carbon-13 (^{13}C) and oxygen-17 (^{17}O). ^{1}H, which contains only one proton, is abundant in biological tissues, so it is used in medical MRI [65,66,74].

The magnetic moments of the nuclei are randomly aligned unless an external magnetic field (B_0) is applied (Figure 1.3 A). B_0 then forces the magnetic moments to align parallel or antiparallel to it (Figure 1.3 B). The principle of thermal equilibrium will result in a slightly higher number of parallel aligned magnetic moments, producing a non-zero net magnetization vector (M) along the z-axis (longitudinal plane), referred to as M_0. The spinning nucleus precesses around the B_0 axis at the Larmor frequency (ω_0), is proportional to B_0 strength (Figure 1.3 C) and governed by equation **(1.1)** [65,66,74].

$$\omega_0 = \gamma B_0 \qquad \textbf{(1.1)}$$

where γ is the gyromagnetic ratio, which is specific for each nucleus (42.57 MHz/T for ^{1}H).

To generate MR signal (Figure 1.3 D), the thermal equilibrium state must be disturbed by exposing the nuclei to a high radiofrequency (RF) pulse, also called an 'excitation pulse', with a frequency equal to ω_0 resulting in a resonance in which the spins absorb energy and precess in phase. Due to the resonance, M will not equal M_0 anymore and it will have an angle (flip angle) that depends on the duration and magnitude of the RF pulse. When a 90° excitation pulse disturbs the thermal equilibrium state, M flips from the longitudinal plane to x-y space (transverse plane). When the excitation pulse is turned off, the relaxation process begins. M will try to realign with B_0, the longitudinal plane will gradually become more magnetized (T1 relaxation), with the nuclei releasing the absorbed RF energy to the surrounding lattice. At the same time, the magnetization of the transverse plane decreases (T2 relaxation), while the spin goes out of phase due to the interaction between the magnetic fields of the neighbouring nuclei. The T2 decay causes a decrease in the current voltage of the receiving coil, leading to the generation of the free induction decay signal (FID), which represents the recorded MR signal.[65,66,74].

The time required for T1 and T2 relaxation varies depending on the surrounding environment. This property, in addition to the proton density (i.e., number of protons per unit volume), allows contrasting and distinguishing different tissues and thus generating anatomical images. Furthermore, the properties of blood perfusion and water diffusion can be detected and help in the generation of functional images [65,66,74].

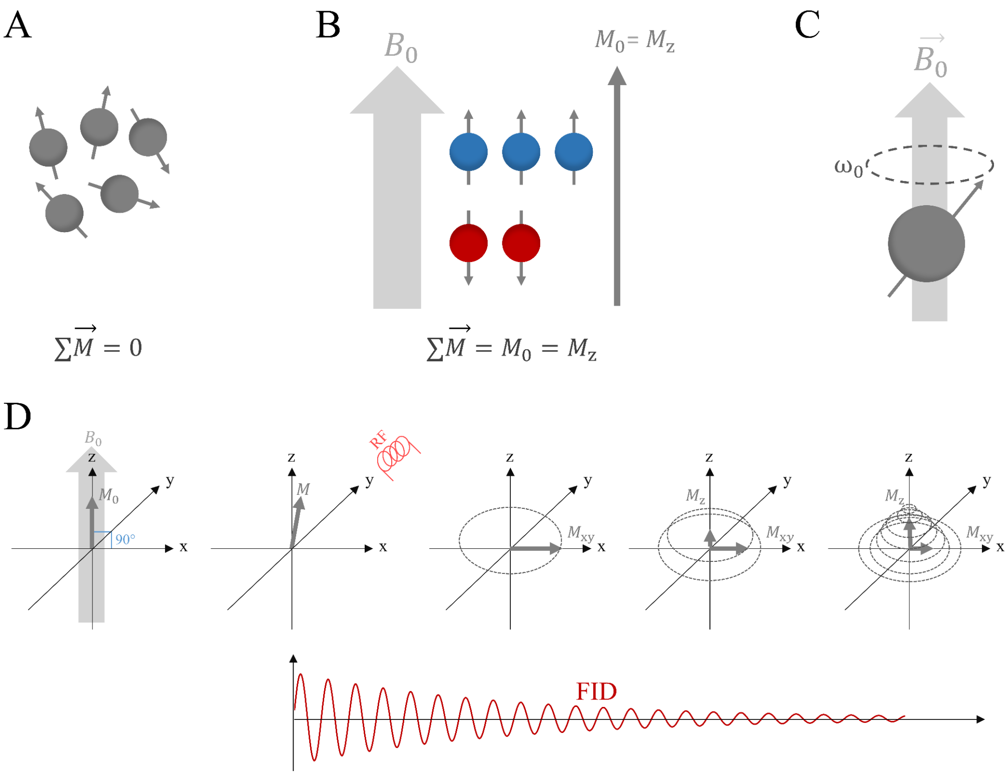


Figure 1.3: Basics of NMR signal generation.

A) The magnetic moments of the nuclei are randomly aligned in the absence of the external magnetic field (B_0). B) When B_0 is present, the magnetic moments will align parallel or antiparallel to it. Due to thermal equilibrium theory, a slightly higher number of magnetic moments will align parallel to B_0, and a non-zero net magnetization vector (M) will be produced along the z-axis. C) A spinning proton precesses around B_0 axis at the Larmor frequency (ω_0). D) Signal generation begins by perturbing the thermal equilibrium state by exposing the spins to a radiofrequency pulse (RF).90° RF pulse will flip M from the longitudinal plane to precess in the transverse plane, resulting in a net transverse magnetization (Mxy) and inducing a current (FID) in the receiver coil. When the RF turns off, relaxation begins and the longitudinal magnetization re-establishes, resulting in a decrease in the FID signal. *Adapted and modified from [51] with permission.*

1.4.2 Image formation and spatial encoding

To create an image, the origin of the FID signal needs to be located in space, this is called spatial encoding. For spatial encoding, special magnetic coils (gradients) are used to create a magnetic field of different strength at different locations. The gradients are superimposed on the external homogeneous magnetic field of the MRI scanner. Three gradients are needed to acquire an image: the slice gradient, the frequency-encoding gradient and the phase-encoding gradient. The slice gradient ensures that the ^{1}H protons experience different magnetic fields and thus have different ω_0. By manipulating the slice gradient, images of different slices in different planes can be acquired. The RF pulse frequency range and the bandwidth of the gradient field strength determine the thickness of these slices. The frequency-encoding gradient will cause the protons to have different precession frequencies, while the phase-encoding gradient will result in phase-shifted proton precession. Using the three gradients together helps to identify the exact point in space of each FID signal. The signals are then sampled and mapped into an array (k-space). The signal frequency components along the horizontal and vertical axis of the image are mapped into the x- and y-axis of the k-space, respectively. The inverse Fourier transform of the entire array yields the MR image (Figure 1.4 C) [65,66,75].

1.4.3 MRI pulse sequences

To obtain an MR image, RF pulses and gradients are used to control the contrast through pulse sequences. The pulse sequence is a combination of RF pulses, second FID signals (echo) generated by refocusing the spin through the process of dephasing followed by rephasing, and an intervening recovery phase. The echo is generated either by an additional RF pulse (spin-echo) or by additional gradient application (gradient-echo). These two means serve as the basis for all MRI pulse sequences [65,66,74,76]. There are many different types of sequences, but spin-echo and gradient-echo sequences are described here as they are considered the fundamental pulse sequences.

Spin-echo sequence

The spin-echo sequence (Figure 1.4 A) consists of an excitation pulse and a refocusing pulse. The excitation pulse (90°) rotates M from the longitudinal plane to the transverse plane. When the excitation pulse is turned off, M will try to realign with B_0, the spins will dephase, and thus the FID will decay exponentially. Then a refocusing pulse (180°) is applied, which rotates the dephasing magnetization vectors around the y-axis. In the case of static magnetic field inhomogeneities, the protons regain their precession frequency and the magnetization vectors will rephase an echo at echo time (TE). To enable phase-encoding, this sequence is repeated with different gradients for each repetition. The time between each excitation pulse is called the repetition time (TR). The scanning time in the conventional spin-echo sequence is relatively long; therefore, the fast or turbo spin-echo sequence is often used in practice. In the fast spin-echo sequence, multiple rephasing pulses (180°) are applied per TR to generate a train of echoes and perform multiple phase encoding steps, resulting in more k-space lines being filled per TR [65,66,77].

Gradient-echo sequence

The gradient-echo sequence (Figure 1.4 B) consists of an excitation pulse and a dephasing/rephasing gradient. After the excitation pulse, the frequency-encoding gradient is used to force a dephasing of the magnetization in the transverse plane. The same gradient, but in the opposite direction, is then turned on to rephase the spin and produce a gradient echo. To allow faster image acquisition, the waiting time for longitudinal relaxation before the next acquisition must be reduced, which can be achieved by using flip angles smaller than 90° [65,66].

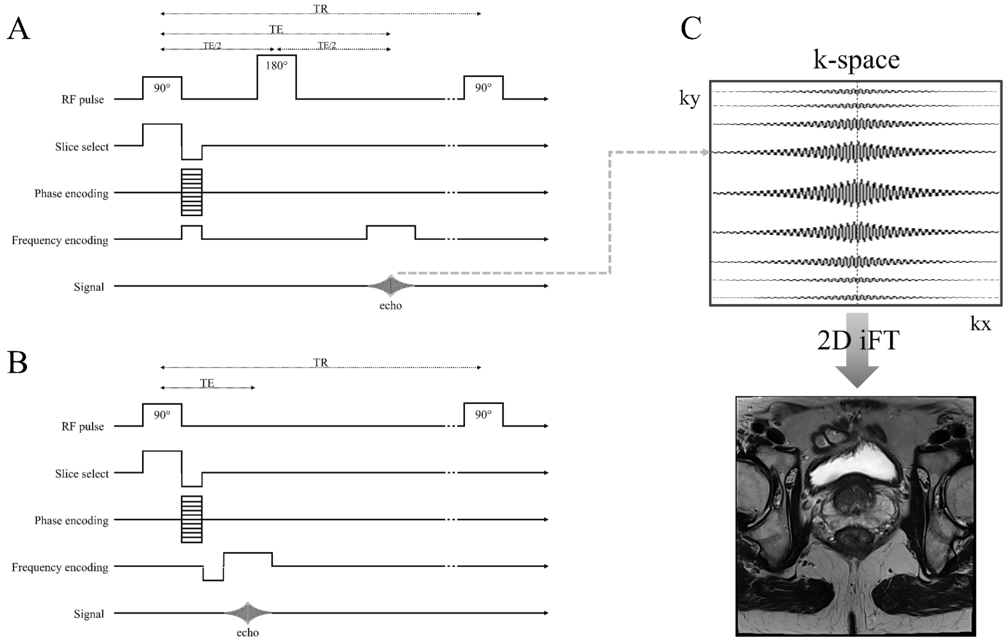


Figure 1.4: Illustration of MRI pulse sequences and image formation.

A) Diagram of spin-echo sequence containing an excitation pulse (90°) and a refocusing pulse (180°) to produce the echo. B) Gradient-echo sequence diagram, where the frequency-encoding gradient is used for dephasing and rephrasing to generate an echo. C) K-space representation. The k-space is an array filled with the signals that are assigned an exact position within the array using the gradients. The frequency components of the signal along the horizontal and vertical axis of the image are mapped into kx and ky, respectively. The array is then used to generate the final image by implementing the 2D inverse Fourier transform (2D iFT) help. TR: repetition time; TE: echo time.

1.4.4 Multiparametric MRI in prostate cancer diagnosis

In recent years, MRI has become an indispensable tool for the diagnosis of prostate cancer because it provides excellent soft tissue contrast, is a non-invasive technique and offers the ability to assess multiple physiologic parameters [65,66]. Advances in technology led to the development of multiparametric MRI (mpMRI), which involves the acquisition and integration of multiple MRI sequences and provides images with different types of functional and anatomical contrast [23]. To improve prostate cancer diagnosis, the use of mpMRI has been established by international guidelines [23,44,78]. mpMRI is being used to detect, localize and stage prostate cancer in order to select a more appropriate treatment strategy for patients [79-84]. In addition, mpMRI has been employed in active surveillance programs to follow up patients with indolent lesions [85], prostate cancer risk calculators [86] and treatment response monitoring [87]. Moreover, mpMRI has demonstrated the ability to reduce overdiagnosis of inconspicuous prostate cancer [83,88].

The mpMRI protocols include T2-weighted imaging (T2W), diffusion-weighted imaging (DWI) and dynamic contrast-enhanced (DCE) MRI [23]. The T2W sequence provides anatomical and structural information; the DWI sequence produces high-contrast images based on water molecule motion variation, while DCE can be used to study vascularity characteristic of the tissue [23]. Figure 1.5 shows an example case where the mpMRI sequences have been used to scan a prostate cancer patient.

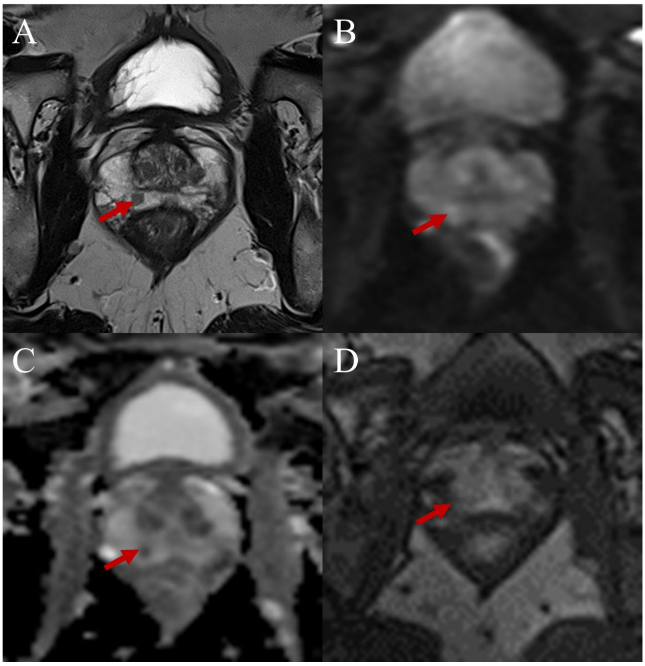


Figure 1.5: An example of multiparametric MRI scans.

An example case of a patient with biopsy-confirmed prostate cancer (pointed with the red arrow; PI-RADS 4, Gleason = 4+4). The example shows the middle slice of the prostate gland on T2W (A), DWI b800 (B), ADC (C) and DCE (D) MRI.

T2-weighted imaging

T2W imaging is the main sequence to visualize the anatomy of the prostate gland and zones [89]. In T2W images, contrast depends mainly on the differences in T2 relaxation times between fat and water. T2W images require a long TE, to give the fat and water enough time to decay [23,66]. In T2W images, fatty tissue will appear darker than the water due to the shorter T2 relaxation time of the fat. Therefore, the peripheral zone, which is fluid rich, will have moderately high and uniform signal intensity, while the transition and central zones will have lower signal intensity. Due to the increased cell density and loss of glandular ducts, prostate cancer appears hypo-intense on T2W images compared to normal prostate tissue, which tends to appear hyper-intense in the peripheral zone [23]. However, low signal intensities may also be caused by BPH, prostatitis, scarring, or post-biopsy haemorrhage [90]. In prostate cancer, T2W does not serve as an independent sequence due to the non-specificity of the intensity signal.

Diffusion-weighted imaging

The DWI uses diffusion weighting gradients to probe the movement of water molecules in the extracellular space due to thermal motion known as diffusion [66,91]. Diffusion is dependent on tissue structure. In normal prostate tissue, water molecules move more freely than in cell-dense malignant tissues [91]. The apparent diffusion coefficient (ADC) represents the total displacement of molecular diffusion in the tissue and is higher in areas where there is no restricted diffusion [66]. The ADC map can be calculated from DWIs with different gradient strengths (b values) [83] . In contrast to DWI, a suspicious cancer area has a low signal intensity in the ADC map [44,92]. In prostate cancer, the extracellular space is reduced, therefore ADC and DWI can help provide quantitative and qualitative information to aid in the detection and staging of the cancer. Combining DWI with T2W imaging has increased the sensitivity and specificity of prostate cancer detection [93] and improved transition zone characterization [94].

Dynamic contrast-enhanced imaging

DCE imaging is used to assess tissue vascularity by calculating perfusion parameters. DCE imaging is performed by following the time-course of the contrast agent (usually gadolinium-based) by sequentially acquiring T1-weighted images. The contrast agent shortens the T1 relaxation time, giving rise to increased signal. Cancer is characterized by angiogenesis and the new vessels are more permeable. Therefore, on DCE images, tumour areas typically exhibit rapid wash-in and wash-out of contrast agent, which can be seen as a rapid signal increase followed by a signal decrease when the signal is tracked as a function of time [95,96]. DCE information can help in the diagnosis of prostate cancer and the assessment of response to treatment [97]. However, recently the added value of DCE has been debated [98-100] and therefore DCE acquisition for prostate cancer assessment is no longer embedded in all mpMRI procedures.

1.4.5 Interpretation

The mpMR images of prostate cancer are usually interpreted qualitatively by a radiologist to find signs and patterns of the disease (detection) and/or to identify the nature of the disease (diagnosis). Until 2012, the variability and lack of reliability of the radiologists' reporting and assessment systems was high [83]. Therefore, in 2012, the first version of PI-RADS was introduced by European Society of Urogenital Radiology to standardize the prostate mpMRI reporting process [23]. To overcome some of the problems of the first version, PI-RADS v2 was introduced in 2014 [44]. In 2019, an updated version, PI-RADS v2.1, was released to simplify PI-RADS assessment and improve inter-reader variability [78]. PI-RADS categorizes suspected prostate cancer according to the likelihood of clinically significant cancer (PI-RADS 1 = very unlikely to PI-RADS 5 = very likely). The studies demonstrated the utility and improvements in prostate cancer assessment with PI-RADS [101-103], which increased the confidence in PI-RADS, which is widely used in the clinic nowadays [83].

Despite the improvements in mpMRI reporting systems, traditional qualitative radiological interpretation of images still has a number of limitations, such as high inter-observer variability [104], time-consuming nature [105], dependence on reader opinion [104,106] and lack of scalability of the manual data processing approach as demand increases [107]. Automated computer-aided detection and diagnosis (CAD) systems, discussed in more detail in *Section 1.6*, have the potential to overcome the limitations of traditional radiological reading by implementing quantitative models to automate, standardize and support reproducible interpretation of radiological images [105,107-109].

1.4.6 Radiomics and quantitative analysis

The mpMRI images contain information that goes beyond the qualitative observations of a radiologist. Quantitative analysis of mpMRI images provides numerical data from which various useful parameters can be extracted [110,111]. These parameters, called features, contain valuable information about the characteristics of the tissue and thus can be used to improve prostate cancer diagnosis [112,113]. The process of extracting and analysing a large number of advanced quantitative radiological features through high-throughput computations is referred to as *radiomics* (Figure 1.6). In radiomics, imaging data is converted into a high-dimensional space that enables feature mining using automated statistical models to develop decision support tools [110-115]. Radiomics features include, but are not limited to, first order (histogram-based), shape and higher order (textural) features [116,117]. The most common textural features are those from the gray level co-occurrence matrix [118], gray level run length matrix [119], gray level size zone matrix [120], gray level dependence matrix [121] and neighbouring gray tone difference matrix [122]. Radiomics feature extraction requires the determination of some variables and settings, which are detailed in the Image Biomarker Standardization Initiative [117].

The implementation of radiomics may lead to a better assessment of prostate tumours by providing quantitative features for intra- and inter-tumoral heterogeneity [113]. Although the field of radiomics is relatively young, several studies have shown that it has potential for prostate cancer detection, staging and monitoring of treatment response [110,112,123-125]. Although radiomics can be performed as a stand-alone process, it is usually implemented as a part of a CAD system [111]. The implementation of radiomics in CAD systems has improved the performance of CAD systems [107,108,110].

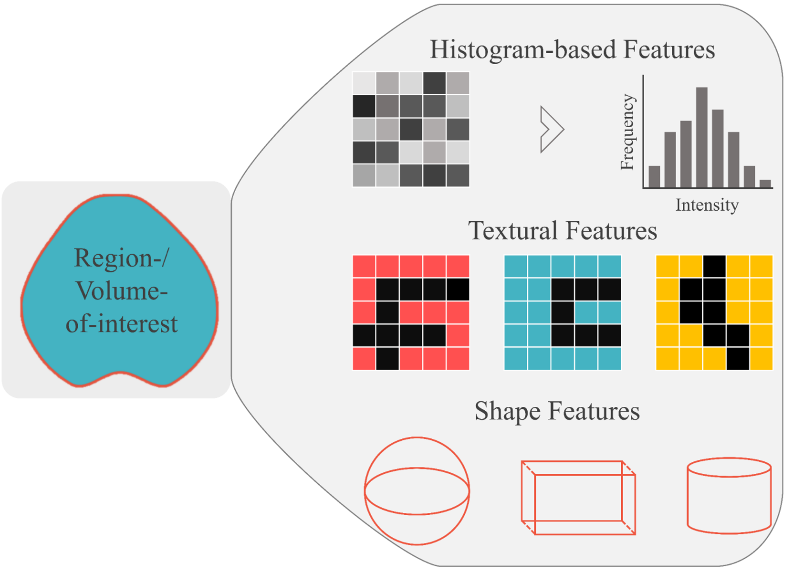


Figure 1.6: Illustration of radiomics features.

Radiomics features extracted from the region/volume of interest typically include statistical (histogram-based), textural and morphological (shape-based) features. The combination of these features enables the development of more efficient statistical models. *Adapted and modified from [126] with permission.*

1.5 Artificial intelligence

CAD systems can be regarded as a form of artificial intelligence (AI). AI is the implementation of computerized systems to mimic human intelligence to perform tasks that require the ability to learn, reason, and respond to situations that are not programmed into the machine's system [127,128]. Machine learning (ML), which is a subfield of AI, and deep learning (DL), which is a subfield of ML, have been used in various areas of medical imaging, including mpMRI of prostate cancer, and have shown great potential for a variety of applications [129-134].

1.5.1 Machine learning

ML is a branch of AI (Figure 1.7 A) that uses statistical and mathematical models to improve the performance of computer systems. ML models learn from training data to make predictions for unseen data [129,132]. CAD systems are highly dependent on the ML methods [129], which are classified into three types: supervised learning, unsupervised learning and reinforcement learning. Supervised learning, most commonly used in medical imaging applications, depends on labelled data, where the model is trained on pairs of inputs and the corresponding output [129,132]. Examples of supervised learning approaches include linear regression [135], logistic regression [136], least absolute shrinkage and selection operator [137], decision tree [138], random forest [139], naive Bayes [136], support vector machine (SVM) [140], k-nearest neighbour [141] and neural network [142]. Unsupervised learning, on the other hand, does not require the corresponding outputs of the training data. It categorizes the input data based on the recognized patterns [129,132]. Examples of unsupervised learning approaches include hierarchical clustering [143], fuzzy C-means clustering [144], Gaussian mixture modelling [145] and K-means clustering [146]. Reinforcement learning is based on the reward principle, where a classifier is created with labelled data and used with unlabelled data to further improve the performance of the classifier using the returned feedback [129,132]. Examples of reinforcement learning approaches include Markov decision process and Q-learning [147].

1.5.2 Deep learning

DL is a subfield of ML (Figure 1.7 A), which is based on the use of multilayer artificial neural networks to learn a large number of features using Big Data to improve the performance of computer systems [129,148]. Due to the advances in computing technology, the development of graphics processing units (GPUs) and the increase in the amount of available data, DL has become very popular in recent years [149]. DL has shown promising results in various fields of medical imaging [133,149,150]. In mpMRI of the prostate, DL has been implemented in image acquisition and reconstruction [151-155], pre-processing [156-162], prostate cancer diagnosis

[163], detection [164-167] and staging [168-170]. Since it is a subfield of ML, DL can also be part of the CAD workflow [171]. DL can be based on supervised or unsupervised models using different architectures such as recurrent neural networks, long short-term memory networks and deep belief networks [148]; however, the most common architecture in medical imaging is the convolutional neural network (CNN) [172]. As Figure 1.7 B illustrates, a CNN takes images as input; each image then goes through a sequence of convolutional layers along with filters, activation functions and pooling layers, extracting features from the images; then the output (features) of the last convolutional layer is fed into fully connected layer and activation function to classify the object with probabilistic values [172,173].

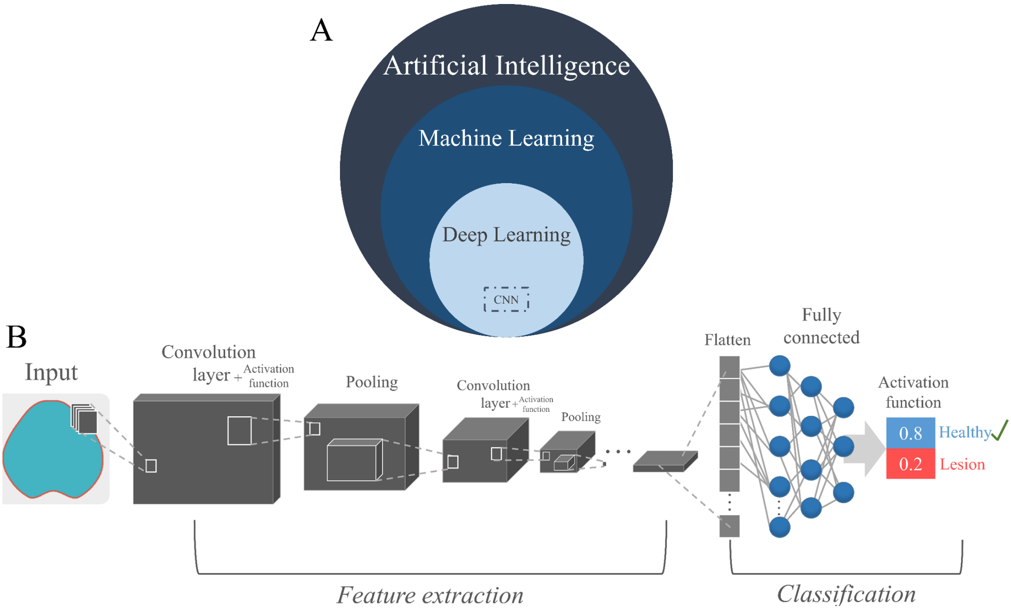


Figure 1.7: Illustration of artificial intelligence subfields and convolutional neural network (CNN) architecture.

A) Deep learning is a subfield of machine learning, which is a subfield of artificial intelligence. B) A common CNN architecture shows how a series of convolutional layers with activation functions and pooling layers are used to extract features and pass them to a fully connected layer to classify the input (e.g., healthy/lesion tissue) using an activation function.

1.6 Computer-aided detection and diagnosis

CAD systems have emerged from the field of computer vision with the aim of assisting radiologists in making clinical decisions by facilitating the detection or/and diagnosis of disease from medical images [107,108,174]. Automated CAD systems offer promising solutions to overcome the limitations of qualitative image interpretation. They can shorten reading time, reduce required radiological reading expertise, standardize and support reproducible interpretation of radiological images [105,107-109,175]. CAD systems have been developed to assist radiologists in the detection and diagnosis of various diseases, such as breast cancer [176], colorectal cancer [177], lung cancer [178] and prostate cancer [107-109]. For prostate cancer, CAD systems using prostate mpMRI have shown promising results in detecting and diagnosing the disease [175,179-184]. For prostate mpMRI, a CAD system can use some or all of the mpMRI sequences (i.e., T2W, DWI and DCE images) as input to the CAD workflow. A typical CAD workflow (Figure 1.8) consists of pre-processing, segmentation, registration, feature extraction and selection, classification and diagnosis [107-109]. These stages are usually performed with the assistant of ML, either traditional methods such as linear regression models or more recently DL methods such as CNNs. DL methods are also capable of combining two or more of the CAD stages, e.g. feature extraction and classification (Figure 1.7 B).

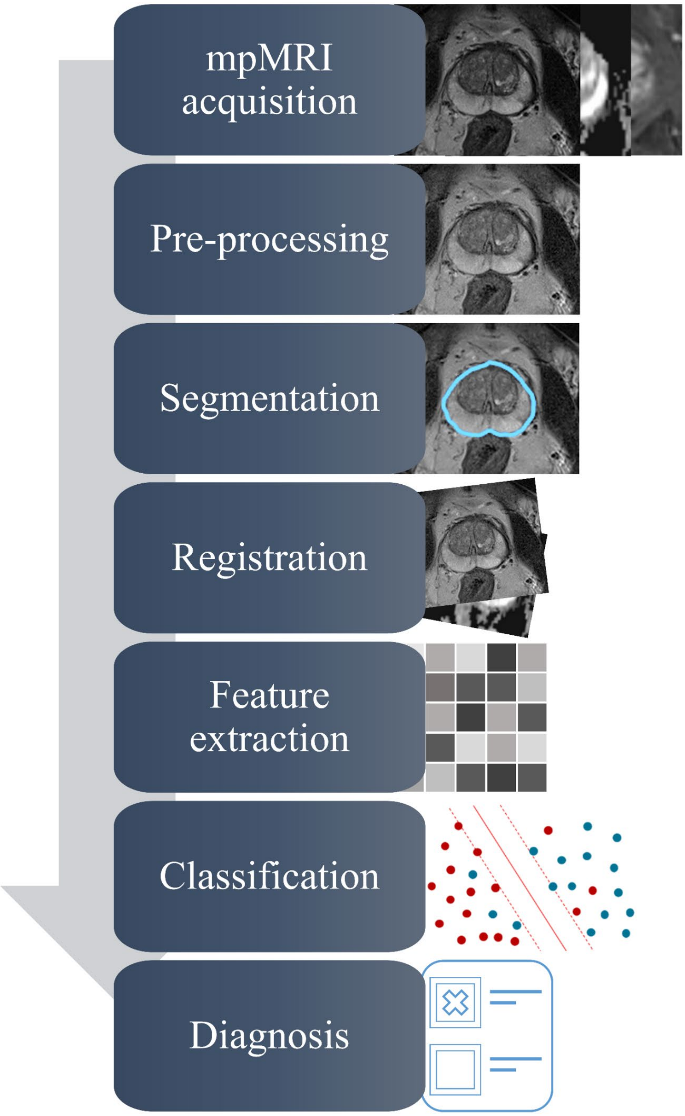


Figure 1.8: Computer-aided detection and diagnosis (CAD) system workflow.

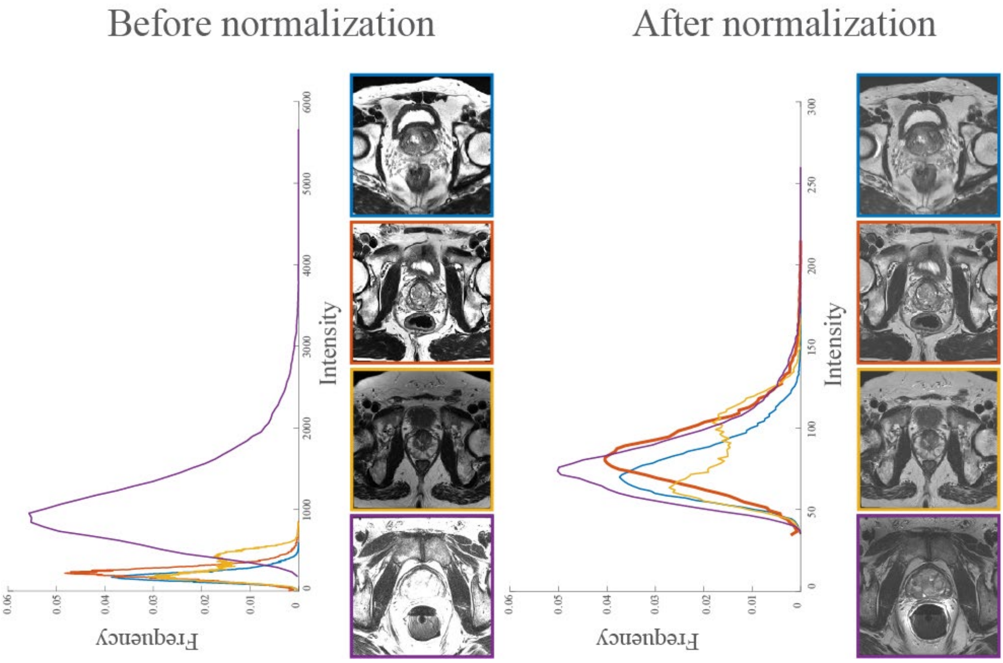
A typical CAD workflow takes input images (e.g., prostate gland T2W, DWI and DCE MR images) and proceed with pre-processing, segmentation, registration, feature extraction and selection and classification to end up with a diagnosis or detection.

1.6.1 Pre-processing and normalization

Image pre-processing is an essential stage in CAD systems as it prepares and transforms the images into a domain where the data can be processed quantitatively [107]. The most important pre-processing steps for mpMRI images are bias field correction and signal intensity normalization. Bias field correction involves the correction of low spatial frequencies variations in signal intensities arising from inhomogeneity of the MRI field and the sensitivity profile of the receiver coils [185]. Excluding this step from CAD pre-processing will increase the difficulty of performing the next processing steps.

Another important step in CAD pre-processing is intensity normalization, which eliminates signal intensity variations between images [107,108]. Intensity normalization is often required to use T2W images for quantitative analysis because of a lack of standardization between scanners. The lack of intensity standardization is due to scanning parameters such as coil type, field strength and acquisition protocols, among others [186-189]. Intensity normalization allows comparison of T2W image values from different patients (inter-patient comparison), tracking patients on multiple scans over time (intra-patient comparison), and performing tissue classification tasks [190-192].

The intensity normalization approaches can be divided into histogram-based and reference tissue-based methods. Although simple to implement, histogram-based methods, which depend on pre-set histogram landmarks to deform or rescale intensity, have their limitations [188,193]. A promising alternative is reference tissue-based normalization, which is based on scaling the intensity of the original T2W image by the intensity in the corresponding region-of-interest (ROI) of the reference tissue [194,195]. A drawback is that this approach traditionally requires manual segmentation of the reference tissues. Figure 1.9 shows an example of a number of cases normalized using the fully automated reference tissue-based approach proposed in Paper I.


Before normalization

After normalization

Figure 1.9: An example of a number of cases of a T2W MR image of the prostate acquired from three different datasets before (left) and after (right) normalization.

The example cases were normalized using the normalization approach proposed in Paper I (see *Section 4.1*). The figure shows the changes of the images after normalization qualitatively and quantitatively (stacked prostate intensity histogram). In both panels, the images were window-levelled from 0 to 2 times the mean prostate intensity of all images in the example.

1.6.2 Segmentation

Volume-of-interest (VOI) segmentation (e.g., prostate gland or zones) is an essential pillar for any CAD system. It helps remove redundant image information and enables the subsequent extraction of quantitative image features from sub-volumes such as tumours for further analysis or diagnosis [107,108]. Figure 1.10 gives an example of T2W MRI segmentation.

Accurate and precise segmentation is crucial as the following stages of a CAD system depend on it. It is also necessary for clinical applications that are sensitive to segmentation errors, such as MRI-ultrasound fusion for targeted prostate biopsies, which is currently becoming a standard clinical procedure [41], and prostate-targeted MR-guided radiotherapy, which has been used in the treatment of prostate cancer patients in recent years [196]. However, manual segmentation of the prostate, which is traditionally performed on T2W MR images by radiologists, is a time-consuming task. Recently, DL-based segmentation methods have shown great promise to fully automate this stage [158-161,197], which would save valuable time and could facilitate the integration of CAD systems in clinical practice.

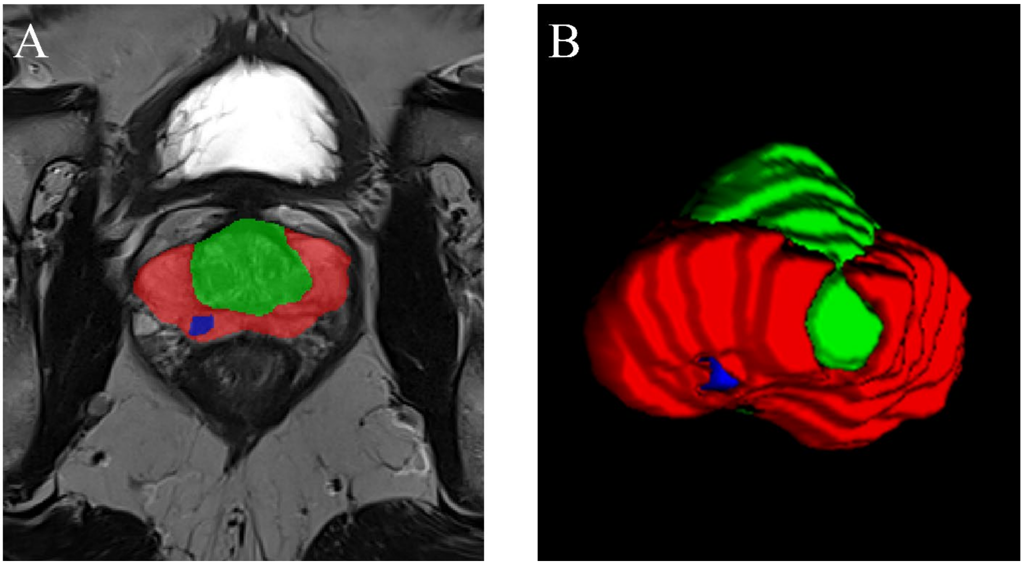


Figure 1.10: An example on manually segmented T2W MR images of the prostate.

A) Shows an axial 2D view of the scan middle slice, in which the peripheral zone (red), the remaining zones (central zone, transition zone and anterior fibromuscular stroma, combined; green) and the lesion (blue) are segmented. B) A 3D view of the whole prostate gland segmentation.

1.6.3 Registration

Registration is the process of bringing different imaging modalities (e.g., MRI, ultrasound, computed tomography) or sequences (e.g., T2W, DWI, DCE) into the same spatial position and aligning them [107,108,198]. Registration is performed in 2D or 3D by aligning a moving image with a fixed image by geometric transformation to maximize the similarity of the two images [198]. The geometric transformation can be categorized into linear and non-linear [198]. Examples of linear transformation include rigid transformation, affine transformation, and projective transformation [198]. Examples of non-linear transformation include B-splines, fluid flow, and optical flow [198]. The purpose of the registration is to allow feature extraction from the same VOI using different modalities or sequences, which will improve the performance of the classification process and thus the diagnosis by providing more representative quantitative information. In mpMRI of the prostate, it is common to register DWI or DCE images to the T2W images and use the VOI segmentation mask generated on T2W to extract features from the VOI in the moving image.

1.6.4 Feature extraction and selection

Feature extraction is the step where the quantitative image features (radiomics) that characterize the VOI, see *Section 1.4.6*, are computed to serve as input to the classification stage [107,108]. The feature extraction process in traditional ML methods is hand-crafted, i.e., the required features are first identified and then computed. In DL-based systems, a larger number of features than the hand-crafted ones in ML are automatically extracted, without prior identification, and fed into an integrated fully connected layer to perform the classification [199]. To simplify the classification model, a feature selection strategy can be used to select a subset of the extracted features to be used in training and testing the classification model [108]. For feature selection, the traditional ML-based systems could for example rank the features in order of importance and then select the most important ones [200], while the DL-based systems could use a dropout layer after the fully connected layer to randomly exclude a certain percentage of the extracted features from further analysis [201].

1.6.5 Classification

Classification is the final stage in the workflow of a CAD system that leads to disease detection or/and diagnosis [107,108]. In this stage, the selected features from the previous stage and the generated VOI segmentations are used to train and test models that perform a specific task, such as distinguishing healthy prostate tissue from malignant lesions. Training and testing the classifier depends on the training approach (supervised/unsupervised) and whether it is traditional ML-based or DL-based. Examples of traditional ML-based and DL-based classifiers can be found in *Section 1.5.1* and *Section 1.5.2* respectively.

2 Objectives

The overall aim of this thesis was to facilitate the integration of automated computer-aided detection and diagnosis (CAD) systems of prostate cancer using multiparametric MRI into clinical practice by developing and evaluating new image pre-processing, segmentation and quality control methods to improve the performance of the CAD workflow.

The specific focus of the thesis was to:

- Develop and evaluate a novel automated method for prostate T2-weighted MR image normalization using dual-reference (fat and muscle) tissue (Paper I).

- Establish a fully automated quality control system for deep learning-based prostate segmentation on T2-weighted MRI (Paper II).

- Investigate the reproducibility of deep learning-based segmentations of the whole prostate gland, peripheral zone and the remaining prostate zones (Paper III).

3 Materials and methods

This chapter briefly summarizes the materials and methods used in the three papers that make up this thesis. Further details are provided in the corresponding papers. All methods were carried out in accordance with the relevant guidelines and regulations. Table 3.1 provides an overview of the datasets, methods, and statistical analyses conducted for each of the three papers.

Table 3.1: Overview of the datasets, methods and statistical analyses used in the papers that make up this thesis.

		Paper I	Paper II	Paper III
Datasets	*In-house*	N = 60	N = 246	N = 244
	PROMISE12	N = 80	N = 50	–
	PROSTATEx	N = 202	N = 339	–
Methods	*Pre-processing*	N4 Bias field correction Intensity rescaling Image resizing	N4 Bias field correction AutoRef normalization DL network requirements	DL network requirements
	Segmentation	Manual Object detection-based	Manual DL-based (U-Net, V-Net, nnU-Net-2D, nnU-Net-3D)	Manual DL-based (V-Net, nnU-Net-2D, nnU-Net-3D)
	Extracted features	First order (N = 1)	First order (N = 18) Texture (N = 75) Shape (N = 14)	Shape (N = 14)
	Models	AFC object detectors Linear scaling model Logistic regression model	Linear mapping function LASSO model	–
Statistical analysis	*Statistical difference*	Wilcoxon signed-rank test Wilcoxon rank-sum test Two-sample *t* test DeLong's method	–	Wilcoxon signed-rank test Permutation test
	Correction for multiple testing	Benjamini–Hochberg false discovery rate	–	Benjamini–Hochberg false discovery rate
	Correlation	–	Spearman's rank test	Spearman's rank test
	Evaluation	Qualitative analysis Histogram intersection Area under receiver operating characteristic curve	Mean absolute error Bland–Altman analysis Dice similarity coefficient Absolute relative volume difference 95% Hausdorff distance Average Symmetric surface distance	Bland–Altman analysis Dice similarity coefficient Intra-class correlation coefficient

AFC: aggregate channel features; LASSO: least absolute shrinkage and selection operator; AutoRef: the normalization method proposed in Paper I; DL: deep learning.

3.1 Datasets

The research conducted for the three papers that make up this thesis relied on one or more of three datasets: PROMISE12, PROSTATEx and In-house. The Regional Committee for Medical and Health Research Ethics (REC Mid Norway) approved the use of the in-house collected dataset (identifiers 2013/1869 and 2017/576). All the in-house collected dataset patients signed informed consent prior to the initiation of the study, whereas the two other datasets were publicly available. An overview of how and where each of the datasets was used can be found in Figure 3.1.

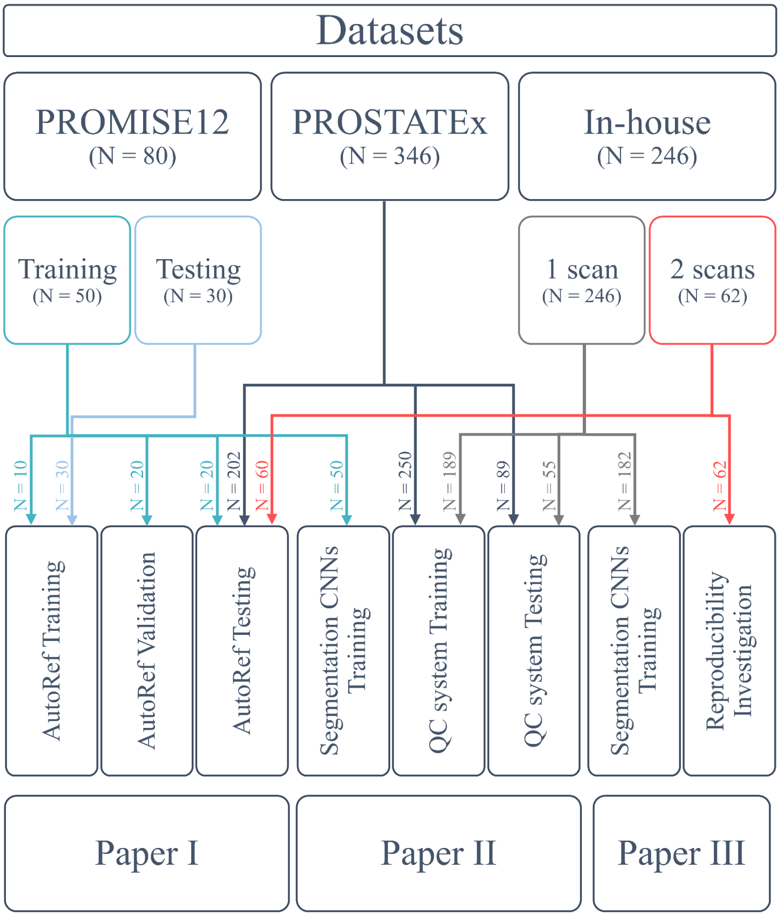


Figure 3.1: Overview of the datasets that used in the papers that make up this thesis and where they were used.

Seven patients were excluded from the PROSTATEx dataset due to technical issues. Note that patients with 2 scans in the in-house collected dataset were also counted among those with 1 scan, but they were used separately in Paper III. *AutoRef: the normalization method proposed in Paper I; CNN: convolutional neural network; QC: quality control.*

PROMISE12

The prostate MR image segmentation (PROMISE12) challenge dataset [202] is a multi-centre and multi-vendor dataset that consists of transverse T2W images of both patients with prostate cancer and benign disease acquired with different field strengths, acquisition protocols and coils for the purpose of prostate cancer detection or staging. Table 3.2 provides details of the PROMISE12 dataset collection.

Table 3.2: Details of PROMISE12 acquisition protocols.

Centre	HUH	BIDMC	UCL	RUNMC
Patients number	20	20	20	20
Field strength (T)	1.5	3	1.5 & 3	3
Manufacturer	Siemens	GE	Siemens	Siemens
Endorectal coil used	Yes	Yes	No	No
In-plane resolution (mm^2)	0.625	0.25	0.325 – 0.625	0.5 – 0.75
Slice thickness (mm)	3.6	2.2 – 3	3 – 3.6	3.6 – 4

HUH: Haukeland University Hospital, Bergen, Norway; BIDMC: Beth Israel Deaconess Medical Center, Boston, US; UCL: University College London, London, UK; RUNMC: Radboud University Nijmegen Medical Centre Nijmegen, Netherlands. Siemens: Siemens Healthineers, Erlangen, Germany. GE: General Electric, Boston, US. Adapted and modified from [202] with permission.

PROSTATEx

The PROSTATEx challenge dataset [203] consists of pre-biopsy mpMRI sequences (T2W, DWI and DCE) from 346 patients (median age = 66; range: 48 – 83 years) acquired at Radboud University Medical Centre, Nijmegen, Netherlands. Targeted biopsy cores results were available for 202 patients, which were used in Paper I to distinguish between healthy and malignant tissue. The use of this dataset was limited to the transverse T2W images, which were acquired using a turbo spin-echo sequence and had an in-plane resolution of 0.5 mm and a slice thickness of 3.6 mm. 7 patients were excluded from this dataset due to technical issues related to the field of view of the images.

In-house

The in-house collected dataset consists of pre-biopsy mpMRI sequences (T2W, DWI and DCE) from 246 patients (median age = 65; range: 44 – 76 years) examined at St. Olavs Hospital, Trondheim University Hospital, Trondheim, Norway between March 2015 and December 2017. The use of this dataset was limited to the transverse T2W images, which were performed on a Magnetom Skyra 3 T MRI system (Siemens Healthineers, Erlangen, Germany) with a turbo spin-echo sequence. 62 of the patients had two scans acquired at two different time points: first, at the initial visit for detection of prostate cancer, and second, during an MR-guided biopsy procedure. The interval between the two scans ranged from 1 – 71 (median = 7) days. The details of the scan parameters of the dataset are shown in Table 3.3.

Table 3.3: Details of the in-house collected dataset scanning parameters.

	Patients with multiple scans		**Rest of patients**
	Scan 1	***Scan 2***	
Repetition time (ms)	4800 – 9520	5660 – 7740	4450 – 9520
Echo time (ms)	101 – 104	101 – 104	101 – 108
Flip angle (degree)	152 – 160	152 – 160	145 – 160
Number of averages	3	3 – 6	1 – 3
Matrix size	320×320 – 384×384	320×320 – 384×384	320×320 – 384×384
Slices	24 – 32	17 – 24	24 – 36
Slice thickness (mm)	3	3	3 – 3.5
In plane resolution (mm^2)	0.5×0.5 – 0.6×0.6	0.5×0.5 – 0.6×0.6	0.5×0.5 – 0.6×0.6

3.2 Methods

For each of the papers, the study workflow was similar to that of CAD (Figure 1.8). After the images were collected, they were pre-processed, segmented and features were extracted to develop statistical models (Paper I and Paper II) or to investigate reproducibility (Paper III). All processing and subsequent statistical analysis was performed using MATLAB R2019b (Mathworks, Natick, MA, USA) unless otherwise stated. In the spirit of transparent science, the code for the proposed algorithms has been made publicly available. For Paper I, it can be found at www.github.com/ntnu-mr-cancer/AutoRef. For Paper II, it can be found at www.github.com/ntnu-mr-cancer/SegmentationQualityControl.

3.2.1 Pre-processing

In Paper I, 3D T2W images were pre-processed using N4 bias field correction [204] to correct for MR image distortion caused by MRI field inhomogeneity (see *Section 1.6.1*); rescaling to the 99th percentile intensity value to exclude the extreme intensity values that could have a negative impact on the performance of the proposed normalization method (AutoRef); and resizing the transverse slices to 384x384 pixels with 0.5x0.5 mm in-plane resolution to feed into object detectors that require a fixed input size. The bias field correction and rescaling were based on an optimization process aimed at finding the optimal pre- and post-processing settings that lead to the best performance of AutoRef. In Paper II, the 3D T2W images were pre-processed with the N4 bias field correction [204] and normalized with the AutoRef method (Paper I) to prepare the images for quantitative analysis. In Paper II and Paper III, before training the prostate segmentation CNNs, each network was pre-processed according to its requirements as implemented in the code provided by the respective authors.

3.2.2 Segmentation

In this thesis, the segmentation of VOIs was an essential stage. Segmentation was performed both manually, as a gold standard, and automatically to develop or evaluate the performance of the method-of-interest.

Manual segmentation

For the PROMISE12 dataset, manual expert segmentations of the whole prostate (WP) were publicly available for 50 patients (training subset). Segmentation was performed using either 3DSlicer (www.slicer.org) [205] or MeVisLab (www.mevislab.de). For the PROSTATEx dataset, the manual segmentation was performed using MIM (MIM Software Inc., Cleveland, OH, USA) by imaging experts with a combined experience of more than 25 years in prostate imaging and reviewed by radiation oncologists at Miller School of Medicine, Miami, FL, USA. The segmentations included the WP, peripheral zone (PZ), non-PZ (central, transition and anterior fibromuscular stroma zones, combined), and cancer-suspicious VOIs (based on the targeted biopsy locations provided by the PROSTATEx challenge organizers). The results of the targeted biopsy cores were used to label each cancer-suspicious VOI as a true positive (Gleason score >3+3) or false positive (Gleason score ≤3+3) radiological finding, while the prostate remnant was considered healthy tissue. For the in-house collected dataset, the WP, PZ and non-PZ were segmented using ITK-SNAP (www.itksnap.org) [206] by a radiology resident at St. Olavs Hospital, Trondheim University Hospital, Trondheim, Norway, under the supervision of a radiologist with more than 10 years' experience in prostate imaging.

For Paper I, manual segmentation of areas within fat and muscle tissue was required for a training set of T2W images. Segmentation was performed using ITK-SNAP [206] by a researcher with three years of experience in prostate imaging. The same researcher performed manual segmentations of the prostate for 50 cases randomly selected from a combination of the PROSTATEx and in-house collected datasets to be used for developing a mapping function in Paper II.

Automated segmentation

In Paper I, the automated segmentation of the fat and the levator ani muscle was performed using two trained separate aggregate channel features (ACF) object detectors [207] that generate rectangular ROIs. Each ROI was then post-processed by Otsu thresholding [208] and morphological opening (disk shape with one pixel radius, based on an optimization process) to extract the largest contiguous bright (for fat) or dark (for muscle) structures in the detected rectangle.

In Paper II and Paper III, DL-based segmentation of the prostate was performed with CNNs (Figure 1.7 B). All CNNs are variants of the famous U-Net with skip connections [209]. In Paper II and Paper III, V-Net [159], nnU-Net-2D [158] and nnU-Net-3D [158] were used, while U-Net [210] was used only in Paper II. Table 3.4 gives an overview of these CNNs and their usage.

Table 3.4: Overview of the CNNs used for automated segmentation.

	Paper II				Paper III		
	U-Net	***V-Net***	***nnU-Net-2D***	***nnU-Net-3D***	***V-Net***	***nnU-Net-2D***	***nnU-Net-3D***
Base	2D slice-by-slice	3D volume	2D slice-by-slice	3D volume	3D volume	2D slice-by-slice	3D volume
Pre-processing	According to [210]	According to [159]	According to [158]	According to [158]	According to [159]	According to [158]	According to [158]
Platform	Keras (v. 2.3.0) + TensorFlow (v. 1.9.0)	PyTorch (v. 1.4.0)	PyTorch (v. 1.4.0)	PyTorch (v. 1.4.0)	PyTorch (v. 1.4.0)	PyTorch (v. 1.7.0)	PyTorch (v. 1.7.0)
Software	Python (v. 2.7.12)	Python (v. 3.6.9)	Python (v. 3.6.9)	Python (v. 3.6.9)	Python (v. 3.6.9)	Python (v. 3.6.10)	Python (v. 3.6.10)
System	Ubuntu 16.04.6 LTS	Ubuntu 16.04.6 LTS	Ubuntu 16.04.6 LTS	Ubuntu 16.04.6 LTS	Ubuntu 18.04.4 LTS	Ubuntu 18.04.4 LTS	Ubuntu 18.04.4 LTS
GPU	Single NVIDIA Tesla P100 PCIe 16 GB	Single NVIDIA Tesla P100 PCIe 16 GB	Single NVIDIA Tesla P100 PCIe 16 GB	Single NVIDIA Tesla P100 PCIe 16 GB	Single NVIDIA Tesla P100 PCIe 16 GB	Single NVIDIA Tesla P100 PCIe 16 GB	Single NVIDIA Tesla P100 PCIe 16 GB
Model to segment	WP	WP	WP	WP	1. WP 2. PZ	1. PZ 2. non-PZ	1. PZ 2. non-PZ
Note	–	–	–	–	The models were used to generate the non-PZ masks by subtraction	The models were used to generate the WP masks by merging	The models were used to generate the WP masks by merging

Keras: Keras API (www.keras.io); TensorFlow: TensorFlow (www.tensorflow.org); PyTorch: PyTorch (www.pytorch.org) [211]; Python : Python (Python Software Foundation, Wilmington, DE, USA); Ubuntu: Ubuntu (www.ubuntu.com); NVIDIA: NVIDIA (Santa Clara, CL, USA). WP: Whole prostate; PZ: peripheral zone (PZ); non-PZ: central, transition and anterior fibro-muscular zones, combined.

3.2.3 Feature extraction

In Paper I, the mean intensity feature was calculated using MATLAB for the WP, in addition to healthy and malignant PZ and non-PZ to evaluate the performance of the proposed normalization method. In Paper II, 107 radiomics features (first order (N = 18), texture (N = 75), shape (N = 14); see *Section 1.4.6*) were extracted from the 3D segmentation masks (manual or DL-based) of the WP using Pyradiomics (v. 2.2.0; an open-source Python package) [116] to train, optimize and test the proposed segmentation quality control (QC) system. In Paper III, 14 shape features were extracted from the 3D segmentation masks (manual or DL-based) of WP, PZ and non-PZ and WP using Pyradiomics (v. 3.0) [116] to investigate the reproducibility of the DL-based segmentations over multiple scans in time.

3.2.4 Models

New methods based on statistical models were developed and evaluated in Paper I and Paper II. Note that the segmentation models were developed by others, as described in *Section 3.2.2.*

In Paper I, the ACF object detector was trained in two training stages using manually selected rectangular ROIs. The ACF object detector works as follows: It computes multiple channels from an input image, each channel being a registered feature map of the input image, and then sums and smooths each group of pixels in the channels to produce lower resolution channels. Features are then extracted from each pixel in the aggregated channels and used to train boosted decision trees to distinguish the object (fat/muscle) from the background [207]. The core of Paper I was a linear scaling function, which is based on multiplying each value (i.e., pixel intensity) by a constant plus an additive term. The scaling equation (3.1) scales the fat (I^{fat}) and muscle (I^{muscle}) reference intensity values, calculated as the 90th (for fat) and 10th (for muscle) percentiles of the intensity values in the extracted ROIs, to their respective T2 values at 3T from the literature ($T2^{fat}$ = 121 ms and $T2^{muscle}$ = 40 ms) [212]. In this process all 3D image intensities ($I(x,y,z)$) are normalized to pseudo T2 values ($pT2(x,y,z)$).

$$pT2(x,y,z) = \frac{I(x,y,z) - I^{muscle}}{I^{fat} - I^{muscle}} \times (T2^{fat} - T2^{muscle}) + T2^{muscle} \quad \textbf{(3.1)}$$

To evaluate the performance of the normalization method, a simple logistic regression model was trained and tested to discriminate healthy from malignant tissue based on mean intensity values in ROIs in the PZ and non-PZ. Logistic regression is a type of binary classification that uses predictors to determine a probability value for belonging to two possible values (e.g., healthy vs malignant tissue), using log-odds and sigmoid functions [213].

In Paper II, the manual segmentations were used to develop a mapping function to calculate representative reference segmentation quality scores. The function was in agreement with the mapping function proposed by Litjens et al [202] and uses a combination of metrics that reflect segmentation performance: the dice similarity coefficient (DSC) [214], absolute relative volume difference [215], average symmetric surface distance [216] and 95% Hausdorff distance [217]. They were separately obtained from the whole prostate, apex and base by comparing DL-based segmentations with manual segmentations. The core of Paper II was a least absolute shrinkage and selection operator (LASSO) [137] with the aim of assigning an automatically estimated segmentation quality score. LASSO is an appropriate choice when dealing with a large number of radiomics features, as it performs feature selection using L1 regularization, which adds a penalty equal to the absolute value of the magnitude of the coefficients that leads to the elimination of the useless input variables, to improve model accuracy and interpretability [218]. LASSO is a type of linear regression model, which makes it a simple and fully transparent ML-based model.

3.3 Statistical analysis

Wilcoxon signed rank tests (non-parametric) [219] were used to assess statistical differences between two related samples, whereas Wilcoxon rank sum tests (non-parametric) [219], also known as Mann-Whitney *U* tests, were used for independent samples. Two-sample *t*-tests (parametric) [220] were used for the continuous independent samples with the assumption of equal means. In Paper I, the performance of the logistic regression model was evaluated using the area under the receiver operating characteristic curves (AUC) [221]. To assess statistical differences between AUCs, the DeLong's method [222] was used. In Paper III, to assess the difference in feature reproducibility before and after the implementation of the segmentation QC system, a permutation test [219] with 1000 runs was performed. In all papers, the Benjamini-Hochberg correction for multiple comparisons [223] was performed at a false discovery rate of 0.05. *p*-values of less than 0.05 after correction for multiple comparisons were considered statistically significant in all papers.

In Paper II and Paper III, Spearman's rank tests [219] were performed to assess correlations, and Bland-Altman analyses [224] were performed to visually assess bias in the data distribution.

Evaluation metrics were used to assess the performance of the methods or features. In Paper I, histogram intersections [225] were calculated to evaluate inter- and intra-patient normalization performance. In Paper II, mean absolute error [226] was used to evaluate the QC system (LASSO model). In Paper III, DSC was used to evaluate the segmentation performance, and the two-way random, single score intra-class correlation coefficient (ICC) [227,228] was used to measure the inter-scan reproducibility of the radiomics shape features.

4 Summary of papers

4.1 Paper I

Automated reference tissue normalization of T2-weighted MR images of the prostate using object recognition

Mohammed R. S. Sunoqrot, Gabriel A. Nketiah, Kirsten M. Selnæs, Tone F. Bathen, Mattijs Elschot.

Magnetic Resonance Materials in Physics, Biology and Medicine 2021; 34(2):309-321.

T2W MRI is considered an essential pillar of mpMRI for prostate cancer diagnosis due to its high spatial resolution and the anatomical details it provides. However, T2W images are hindered by non-standard signal intensity, which limits their use to qualitative analysis. To enable quantitative analysis and facilitate comparison between and within patients, intensity normalization, an essential step of CAD, is required. Several normalization approaches have been proposed for prostate imaging, but the most promising has been multi-reference tissue normalization, where the intensity from two or more reference tissues is used to scale the intensity of the image. A disadvantage is that the method requires manual segmentation of the reference tissues. Therefore, the aim of this work was to develop and evaluate a novel method (Figure 4.1) for automated dual-reference tissue normalization of T2W images of the prostate, referred to as *AutoRef*, based on object recognition to automatically extract the reference tissue ROIs.

In this study, transverse T2W images from the publicly available PROMISE12 (N = 80) and PROSTATEx (N = 202) challenge datasets and an in-house collected dataset (N = 60) were used. ACF object detectors were trained to detect reference regions for fat and muscle tissue, which were processed and utilized to normalize the 3D images to pseudo T2 values by linear scaling. To evaluate the performance of Autoref, mean pseudo T2 values of the prostate after normalization were compared with literature values. Inter-patient histogram intersections of voxel intensities in the prostate were compared between the proposed method, the original images, and other commonly used normalization methods. The classification performance of healthy and malignant tissue was compared before and after normalization.

The results showed that the prostate pseudo T2 values of the three tested datasets (mean±standard deviation = 78.49±9.42, 79.69±6.34 and 79.29±6.30 ms) were in good agreement with T2 values from the literature (80±34 ms). AutoRef was also found to result in significantly higher ($p < 0.001$) inter-patient histogram intersections (median = 0.746) than the original images (median = 0.417) and most other normalization methods. In addition, there was a significant improvement ($p < 0.001$) in classification of healthy vs. malignant tissue in PZ (AUC = 0.826 vs. 0.769) and non-PZ (AUC = 0.743 vs. 0.678).

In conclusion, in this study, an automated dual-reference tissue normalization method of T2W images of the prostate was proposed, which has been shown to reduce T2W intensity variation between scans and could improve quantitative assessment of prostate cancer on MRI.

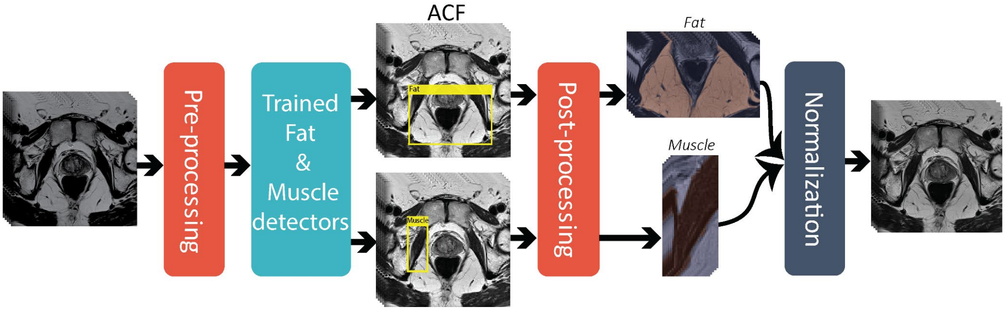


Figure 4.1: Overview of the proposed normalization method.

T2W images were first pre-processed including bias field correction, rescaling and resizing. Rectangles containing fat/muscle were then detected slice by slice using trained aggregate channel features (ACF) detectors. The three slices that contained rectangular regions with the highest probability of fat/muscle were identified and post-processed by Otsu thresholding and morphological opening to extract the largest connected fat/muscle region-of-interest (ROI). Fat/muscle reference intensities were obtained from these ROIs for normalization of 3D image intensities.

4.2 Paper II

A quality control system for automated prostate segmentation on T2-weighted MRI

Mohammed R. S. Sunoqrot, Kirsten M. Selnæs, Elise Sandsmark, Gabriel A. Nketiah, Olmo Zavala-Romero, Radka Stoyanova, Tone F. Bathen, Mattijs Elschot.

Diagnostics 2020; 10(9):714.

Fully automated segmentation of the prostate is a crucial step of CAD for prostate cancer. This step helps focusing on the relevant image information and facilitates the subsequent extraction of radiomics features from sub-volumes for further analysis or diagnosis. DL-based methods seem to be most promising for this purpose, but are not perfect yet. Consequently, visual inspection of the segmentation results is still required to detect poorly segmented cases. Therefore, the aim of this work was to establish a fully automated QC system for prostate segmentation based on T2W MRI (Figure 4.2).

Four different DL-based segmentation methods (U-Net, V-Net, nnU-Net-2D and nnU-Net-3D) were trained using 50 cases from the PROMISE12 challenge dataset. These methods were then used to segment the prostate for a dataset (N = 585) resulting from the combination of the PROSTATEx dataset (N = 339) and the in-house collected dataset (N = 246). T2W images were bias field corrected and normalized using AutoRef (the method proposed in Paper I) to facilitate feature extraction. First order (N = 18), shape (N = 14) and textural (N = 75) radiomics features were extracted from the segmented prostate masks. A reference quality score was calculated for each automated segmentation in comparison to its corresponding manual segmentation. A LASSO was trained and optimized on a randomly assigned training dataset (N = 1756, 439 cases from each segmentation method) to create a generalizable linear regression model based on the radiomics features that best estimated the reference quality score. To evaluate the performance of the QC system, the mean absolute error and Spearman's rank correlation tests were used.

The mean±standard deviation absolute error between the estimated and reference quality scores was 5.47±6.33 on a scale of 0 to 100. Furthermore, a strong correlation was found between the estimated and reference quality scores (rho = 0.70).

In conclusion, in this study, a fully automated and transparent QC system was developed to estimate the quality of automated segmentation of the prostate in T2W MR images, which could be an important step towards the clinical implementation of CAD for prostate cancer.

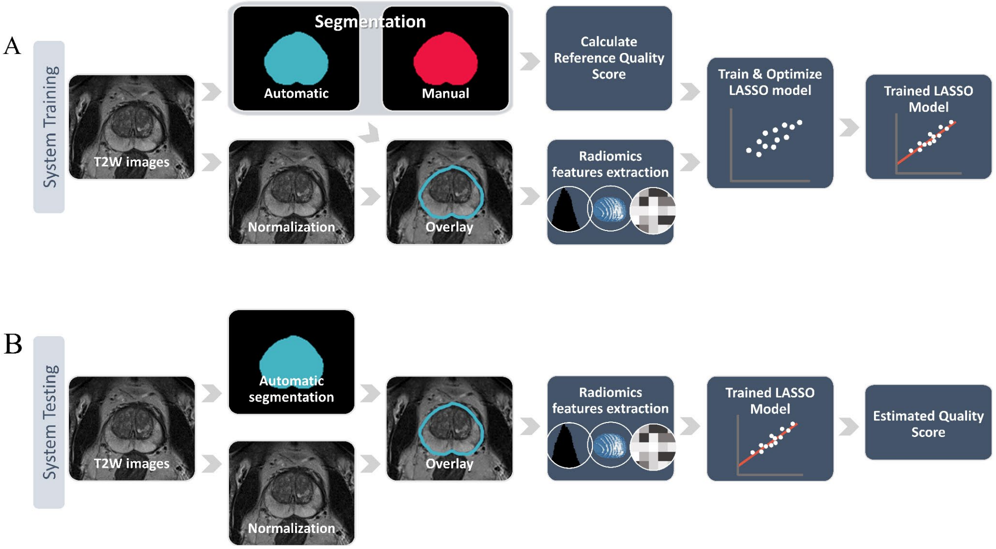


Figure 4.2: The pipeline of training (A) and testing (B) the proposed quality control system.

The system training (A) starts from the normalized T2W image stack with the corresponding manual prostate segmentation and the automated segmentation provided by a deep learning-based segmentation method. These two segmentations are used to compute the reference quality score, and the automated segmentation is also overlaid on the normalized image stack to extract various radiomics features. The reference quality score and the features are then fed into a least absolute shrinkage and selection operator (LASSO) to train and optimize a linear regression model that predicts the quality scores based on the imaging features. During system test (B), the trained model uses the radiomics features extracted from the overlaid automated segmentation on the normalized image stack to estimate a quality score for a previously unseen case.

4.3 Paper III

The reproducibility of deep learning-based segmentation of the prostate gland and zones on T2-weighted MR images

Mohammed R. S. Sunoqrot, Kirsten M. Selnæs, Elise Sandsmark, Sverre Langørgen, Helena Bertilsson, Tone F. Bathen, Mattijs Elschot.

Submitted

Although the performance of DL-based prostate segmentation on single scans is well described, little is known about the reproducibility of these methods for clinical MRI scans. Yet good reproducibility is important for the clinical implementation of automated CAD systems to automate, standardize and support interpretation of radiological images, and paramount for clinical applications based on multiple scans in time, such as active surveillance. Therefore, in this work, the reproducibility of DL-based segmentations of WP, PZ and non-PZ was investigated by comparing radiomics shape features from T2W MR images acquired with short time intervals.

In this work, the in-house collected dataset (N = 244) was used. The dataset (T2W images) was split into a training set (N = 182) to train the DL-based segmentation networks, and an investigation set (N = 62) acquired at two different time points (interval time median = 7 days) to investigate the intra-patient reproducibility of 14 radiomics shape features extracted from the segmented prostate masks of WP, PZ and non-PZ (Figure 4.3). The DL-based segmentation was performed and compared using three different CNNs: V-Net, nnU-Net-2D and nnU-Net-3D. To measure the inter-scan reproducibility of each feature for each CNN and manual segmentation, the two-way random, single score ICC was used.

The reproducibility of all investigated DL-based methods was found to be comparable to that of manual segmentations (14/14 features), except for the V-Net in the PZ (7/14 features). The ICC score for segmentation volume was 0.888, 0.607, 0.819 and 0.903 in PZ; 0.988, 0.967, 0.986 and 0.983 in non-PZ; and 0.982, 0.975, 0.973 and 0.984 in WP for manual, V-Net, nnU-Net-2D and nnU-Net-3D, respectively.

In conclusion, in this study, the reproducibility of shape features extracted from DL-based segmentations of the prostate gland and zones on T2W MR images acquired at short time intervals was investigated. The results demonstrate the feasibility of embedding DL-based segmentation into CAD systems based on multiple T2W MR scans of the prostate, which is an important step towards clinical implementation.

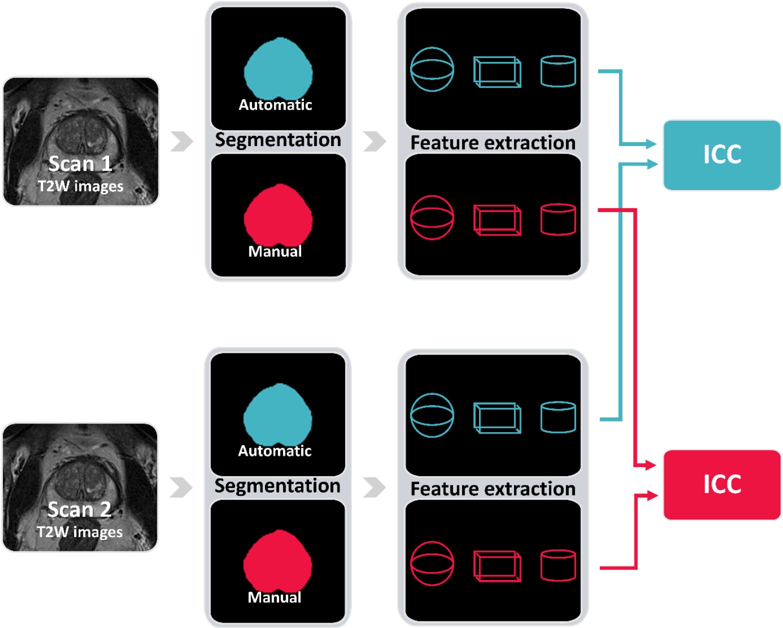


Figure 4.3: The pipeline to investigate the reproducibility of deep learning-based segmentation. The T2W MR images from each of the scans were segmented manually and with deep learning-based segmentation methods. The segmentations are then used to extract shape features. The two-way random, single score intra-class correlation coefficient (ICC) was then used to measure the inter-scan reproducibility of each feature for each of the three deep learning-based segmentation methods and the manual segmentation. Segmentations of the whole prostate, peripheral zone and rest of the zones were investigated.

5 Discussion

Automated CAD of prostate cancer using mpMRI can overcome many of the limitations of the traditional diagnostic approach. Its promise is the fully automated, standardized, reproducible and rapid diagnosis of patients with suspected prostate cancer [105,107-109,175]. However, to facilitate the implementation of CAD systems in clinical practice, further improvement of CAD stages is required, and a relationship of trust must be established. To increase trust in CAD systems, the methods embedded in them should be generalizable, transparent, controlled, reproducible and robust. The aim of this thesis was therefore to develop, evaluate and investigate new methods to achieve this goal.

5.1 Multiparametric MRI interpretation

The addition of mpMRI scanning of prostate cancer patients has significantly improved the diagnostic process of the disease [79-84]. Initial high inter-reader variability of image interpretation has led to the establishment of international guidelines and the proposal of PI-RADS to standardize image interpretation [23]. Standardized interpretation leads to standardized decision-making for the right treatment strategy for the patient [23]. PI-RADS has led to more standardized interpretation [101-103], but it has not eliminated inter-reader variability, which is still a concern [78,229,230]. Despite the establishment of the guidelines, the human factor still has an impact on diagnosis and treatment. The radiologist is still the one who ultimately decides, typically based on qualitative information, whether the perceived patterns/areas in the images meet one of the definitions of PI-RADS. In addition, radiologists manually segment the VOIs on the images to be used in clinical applications such as MRI-ultrasound fusion for targeted prostate biopsies [41], targeted MR-guided radiotherapy of the prostate [196] and PSA density measurement for prostate cancer risk calculators [86].

Automated interpretation, i.e. CAD, of mpMR images in accordance with PI-RADS could help minimize the influence of the human factor. In this way, the diagnostic process becomes standardized and less prone to human error [175]. However, this does not mean that the radiologist will be replaced by a CAD system, because the aim of the system is to support the radiologist in the diagnostic process, not to replace. Of course, the role of the radiologist will change, the focus will probably shift towards the most difficult and complex cases, i.e. the cases that the system has problems with [231].

5.2 The potential and challenges of computer-aided diagnosis of prostate cancer

Automated CAD systems for prostate cancer have the potential to overcome traditional reading problems [105,107-109,175]. CAD extracts and utilizes quantitative (radiomics) information in mpMR images. This information, which is impossible to obtain with the traditional manual approach, is paramount to provide a better interpretation of the patient images [112,113]. The entire CAD processing workflow is automated and thus the diagnostic process is less prone to human error with the aid of CAD [105,175]. The implementation of CAD in the diagnostic process can help overcome the variability between and within readers that results from the reader-dependent nature of the traditional diagnostic approach [107]. Furthermore, the addition of CAD can help the less experienced observers to significantly improve their ability to discriminate between benign and malignant lesions and achieve similar performance to experienced observers [232]. Overcoming the variability problems leads to a more standardized diagnosis and thus a more standardized decision-making [44]. The traditional diagnostic approach requires a high degree of focus, is not scalable to handle the increasing demand for prostate cancer mpMRI and is time-consuming [105,107]. With CAD, a large number of cases can be processed quickly, leaving time for radiologists to focus on the difficult cases that require further investigation or care [231].

In recent years, several CAD systems for prostate cancer have been developed. In 2003, Chan et al. [233] implemented a CAD system with mpMRI for the first time. They integrated the information from T2W, ADC, T2 map and proton density sequences with anatomical and texture features extracted from manually delineated VOIs. A linear discriminant analysis classifier was used to generate a cancer probability map for the PZ, and an average AUC of 0.839 was obtained. To generate a similar map, Shah et al. [234] used T2W, ADC and DCE images to create a combination of features from the manually delineated VOIs. The features were fed into a SVM classifier and an F-score of 0.89 was obtained. To distinguish between benign and malignant tissues for WP, Liu et al. [180] used the T2W, ADC, and DCE images to train a SVM classifier. Intensity, shape and texture features were extracted from the images and fed into the classifier, resulting in an AUC of 0.82. Peng et al. [235] chose to extract the 10th percentile and average ADC values, DCE transfer constant, and histogram-based features and fed them into a linear discriminant analysis classifier, resulting in an AUC of 0.95. Vos et al. [184] developed a fully automated two-stage CAD system to detect cancer in WP. Instead of manually delineating VOIs, they first performed voxel classification using a Hessian blob detection algorithm on the ADC map along with an automatic prostate segmentation method to detect possible lesion candidates. Histogram-based features are then computed from the lesion candidates on the T2W, ADC and DCE images and fed into a classifier with linear discriminant analysis. The results showed sensitivities of 0.41, 0.65 and 0.74 with false positives of 1, 3 and 5 per patient, respectively. The two-stage strategy was also used by Litjens et al. [105] to detect cancer in WP. In the first stage, they used an atlas-based method to segment the prostate on T2W images, extracted voxel features from the segmented VOI, and classified the voxels with a random forest classifier to select candidate areas. In the second stage, T2W, DW, DCE and proton density weighted images were used to extract statistical, local contrast, symmetry and shape features from candidate areas and fed them to a random forest classifier to obtain a cancer probability score of the candidate area. The results showed sensitivities of 0.42, 0.75 and 0.89 with 0.1, 1 and 10 false positives per normal case, respectively. To determine whether or not the patient has prostate cancer, Ishioka et al. [236] developed a CAD system based on DL. They fed a CNN architecture combining U-Net (17 layers) with ResNet50 with labelled T2W images and obtained an AUC of 0.645. To increase the prediction accuracy, Song et al. [237] incorporated an extended prediction method into their optimized patch-based CNN model (based on VGGNet) and obtained an AUC of 0.944. CAD systems for grading prostate cancer have also been developed. In their work, Abraham and Niar [238] developed a CAD system for predicting the Gleason Grade Group for prostate cancer. The lesion centres were defined; therefore, they cropped the area around the lesion centre and used the T2W, ADC and high B-value DW images to extract histogram-based and textural radiomics features. The radiomics features were then fed into a stacked sparse autoencoder with three hidden layers for latent feature extraction. The laten features were then fed into a softmax classifier and a square weighted kappa score of 0.2326 was obtained. de Vente et al. [239] developed a CAD system that used T2W and ADC images as input to a 2D U-Net (5 layers) and generated lesion segmentation maps that encoded Gleason Grade Group as output. The system included placement of a rectangular ROI around the prostate gland and automatic segmentation of PZ and non-PZ with 3D U-Net. The system achieved a quadratic-weighted kappa score of 0.13. 2D U-Net was also used by Schelb et al. [170] to discriminate between clinically significant and non-significant lesions using T2W, ADC and high B-value DW images. The network composed of 34 layers and achieved a sensitivity of 0.92 and a specificity of 0.47 when the cut threshold was set to 0.33. It should be noted that the performance of CAD systems depends on how the system is trained and tested. The earlier studies used leave-one-out cross-validation [105,234,235] and *k*-fold cross-validation [184,233], whereas the later studies [170,180,236,237,239] set up completely separate training, validation, and testing sets in which the same patient data are not used in more than one of the sets. In summary, the studies show that the performance of CAD is improved when a combination of features was used and when DL was included in the classification stage.

Despite its potential, CAD faces several challenges that hinders its implementation in clinical practice, including system compatibility, processing power, machine error, generalizability, transparency, familiarity and building a relationship of trust [240,241]. Translating CAD into the clinic requires the development of compatible systems that can easily communicate with a variety of systems and data structures [241,242]. A suitable hardware infrastructure that enables high computational performance will also be required, especially when DL-based methods are embedded [133,240]. The workflow of CAD consists of several stages that build on each other, each of which typically embeds one or more different ML-based methods [107], which means that any error through the pipeline can be propagated and lead to a misdiagnosis. The errors are to be expected, there is no perfect CAD system. The different stages are trained with data processed and labelled by humans [107,108,241]; thus, human errors can eventually lead to machine errors. Therefore, there should be QC systems for various CAD steps to ensure that mistakes are detected and corrected, or forwarded to radiologists for correction. This was addressed in Paper II, where a QC system was developed for the segmentation stage. Furthermore, the generalizability of the systems is very important for clinical implementation [240,243]. The automated systems should be able to adapt to different types of unseen data. They should be able to perform well in patients with different backgrounds, lifestyles and health conditions [242,243]. This means that CAD systems need big and diverse data for training to be able to accurately diagnose a wide range of patients [241,244]. Paper I addressed this problem by providing a generalizable normalization method for the T2W images. Another important aspect is the transparency of CAD systems [245]. Ideally, it should be clear how the algorithms work and what features they rely on [246,247]. There are fewer transparency concerns with traditional ML approaches than with DL-based methods, since in many of the DL-based methods the decision-making mechanism is a black box [246]. It is difficult to gain a complete understanding of what is going on in the black box [248]. Therefore, traditional ML can be used to control the output of the DL-based methods and determine when it goes wrong. In other words: If we cannot understand how it works, we can at least control it to prevent it from making mistakes. Paper II has adopted this strategy and shown its potential. Another important issue is the reproducibility of CAD systems [249]. For clinical applications based on multiple scans in time, such as active surveillance, it is crucial that the implemented CAD systems are reproducible [250]. If they are not, this could have a negative impact on patient diagnosis and thus treatment. This motivated the investigation of the automated segmentation reproducibility in Paper III. All these reasons, in addition to lack of familiarity and the limited number of studies that have prospectively evaluated the performance of CAD in the clinic, have raised concerns among radiologists about whether CAD can be trusted [241]. In order to build a trustworthy relationship between CAD and radiologists, the aforementioned challenges should be addressed and considered in system development, and more prospective studies should be conducted with an aim at evaluating the performance of the CAD system in clinic [241].

In this thesis, the overall goal was to make CAD of prostate cancer more trustworthy for implementation in the clinic by ensuring the implementation and control of the best performing ML-/DL-based methods in the early stages of the workflow. Focusing on the early stages should reduce the risk of propagated errors. In this thesis, a new normalization approach was proposed (Paper I), a QC system for DL-based segmentations was developed (Paper II) and a reproducibility study for the DL-based segmentations was performed (Paper III). The proposed methods aimed to be generalizable, transparent and robust. Although clinical data were used in this work, the methods still need to be prospectively evaluated in a clinical setting to test their compatibility, efficiency, accuracy and ease of use, among others. Such a step will require many efforts, including obtaining ethical, organizational, legal and patient approvals, rewriting code to be compatible, developing easy-to-use graphical interfaces and recruiting radiologists willing to invest time to use and evaluate the methods.

5.3 Improving T2-weighted MRI normalization

The normalization method proposed in Paper I helps facilitate more accurate quantitative analysis by increasing the homogeneity of signal intensity between and within cohorts. The method could help standardize the T2W images used at different stages of CAD workflow. In addition, it may help to increase confidence in the representativeness of the extracted radiomics features from the normalized images. In Paper I, AutoRef, the proposed method, was compared to some of the commonly used histogram-based normalization methods and was found to outperform them. This might be due to the fact that these methods are dependent on the overall 2D/3D image values, which is a weakness as these are subject to variation due to differences in scan settings (e.g., field of view) and patient-related factors (e.g., bladder filling) [251]. Normalization using single or multiple reference tissues, on the other hand, is potentially less sensitive to variations in scan settings and patient-related factors. The single reference tissue normalization approach is based on scaling the intensity of the original T2W image by the intensity in the corresponding ROI of the reference tissue [181,194,195,252,253]. A common example of this in the prostate is normalization to the intensity of the obturator internus muscle or the levator ani muscle [194,238,247,254,255]. In contrast, multi-reference tissue normalization uses the intensities of multiple reference tissues to build a linear/non-linear regression model to estimate normalized T2W image values [195,252]. In Paper I, AutoRef, an approach to normalization using two reference tissues, was found to outperform normalization using only one reference tissue (levator ani muscle).

Reference tissue-based normalization requires labelling of the reference tissues to enable intensity extraction from them. This is usually done manually [195,252], which is a time-consuming and tedious process. Automating the labelling task, e.g. using object detectors as in Paper I, makes reference tissue normalization more efficient and could potentially facilitate its integration into clinical practice. Automation of the labelling task can also be achieved using other methods, for example by semantic segmentation methods. Compared to semantic segmentation, object recognition requires less computational power, time and data [150,207,238]. To provide a fully automated normalization method, AutoRef relies on ACF object detectors to detect the ROIs of the reference tissues, which are then post-processed to obtain a segmented region within the ROIs.

AutoRef has already been used in studies requiring quantitative image analysis of mpMRI. In Paper II of this thesis, AutoRef was used in pre-processing the T2W images for standardized feature extraction for the QC system. Earlier versions of Paper II skipped image normalization or performed it with variations of scaling to the histogram median [256]. It was observed that replacing these normalization approaches with AutoRef improved the performance of the developed system. In their work, Patsanis et al. [257] evaluated generative adversarial networks for prostate cancer detection. They found that an automated end-to-end pipeline, which is highly dependent on pre-processing parameters, gave the best results (AUC = 0.878) when AutoRef was implemented. These results are consistent with the comparison of AutoRef's ability to improve discrimination between healthy and malignant tissue performed in Paper I. Dewi et al. [258] included AutoRef in their study on the influence of pre-processing configurations on the reproducibility of radiomic features in T2W MRI of the prostate. Similarly, they showed that the inclusion of AutoRef in the pre-processing of T2W images increased the reproducibility of first order and textural radiomics features extracted from T2W MRI. In their preliminary study aimed at evaluating the diagnostic relevance of T2W MRI-derived textural features in prostate cancer with Gleason score 3+4 and 4+3, Nketiah et al. [125]

used the histogram-based normalization approach proposed by Nyúl et al. [188] to pre-process the T2W images. However, in their subsequent multicentre study which aimed at evaluating the potential of T2W MRI-derived textural features for quantitative assessment of peripheral zone aggressiveness [259], histogram equalization was replaced by AutoRef due to its higher performance and ability to achieve more homogeneous intensities within and between cohorts.

As suggested in the discussion of Paper I, the performance of AutoRef was further investigated using a large, multicentre, multivendor cohort. Sørland et al. [260] used AutoRef to normalize T2W images from a cohort of ten scanners (three manufacturers) located at three different institutions in three different countries. The study confirmed that AutoRef performed well across scanners and centres. In Paper I, quantitative T2 maps were not available for the study patients, which hindered direct comparison of pseudo T2 values with a gold standard. Therefore, Sørland et al. [261] acquired the quantitative T2 maps for 7 asymptomatic volunteers with the aim of comparing the gold standard T2 values with the pseudo T2 values generated by AutoRef. The work concluded that Autoref can reproduce both the prostate T2 values and the contrast between the prostate zones. However, since the cohort size is small and consist of relatively young, asymptomatic volunteers (median = 28.5 years), further confirmation in a clinical cohort is required.

Whereas AutoRef performed well for the inter-patient normalization, little additional value was shown for normalization of two scans of the same patients. One explanation for this could be the limited variability of the dataset used in the test, as both scans were acquired at the same centre, on the same scanner, with similar protocols and with a short time interval in between. Moreover, the general performance of AutoRef was close to that of Gaussian normalization, which is much easier to implement and faster than AutoRef (1 second vs. 35 seconds). This might raise the question of whether it is worth implementing AutoRef when Gaussian normalization can perform sufficiently well. However, unlike Gaussian normalization, AutoRef is able to produce pseudo T2 values that are comparable to the T2 values reported in the literature [212]. In addition, normalization with a single reference tissue was also close to AutoRef. Single reference tissue normalization is also easier to implement than AutoRef, but unlike AutoRef, single reference tissue normalization was not able to map prostate T2 values from the literature.

One potential disadvantage of AutoRef in comparison to histogram-based normalization is that the detection of reference tissue ROIs can fail. In paper I, it was shown that this was mostly a problem in patients scanned with an endorectal coil. However, body surface coils are currently recommended for prostate imaging with 3T MRI scanners [262], on which the vast majority of patients is scanned. The high sensitivity of fat detection in T2W images acquired without an endorectal coil was confirmed by Sørland et al. [260] whose results indicated that the object detectors for fat and muscle are stable, but the fat detector has a higher probability (≈ 2.2%) of failure than muscle (≈ 1.8%). Interestingly, the study suggested that an object detector for femoral head can be used instead of fat if fat detection fails. Extending AutoRef to include more tissues could lead to a more robust method. The femoral head, pelvic bone and urinary bladder might be good candidates [195,260,261]. However, further investigation should be conducted to explore their potentials and effects on the performance of AutoRef.

5.4 Towards deep learning-based segmentation

Automated segmentation of the prostate is of great importance for automated CAD systems, as it can reduce human error, standardize output and save time [107,108]. DL-based segmentation of the prostate has shown excellent performance in this regard [158-161,197]. Inter-observer variability has been shown to be approximately the same between DL-based segmentation methods and experienced radiologists [170]. Nevertheless, each of the proposed segmentation methods will occasionally lead to unpredictable suboptimal contours in some cases. Thus, manual verification of contours by radiologists remains a necessary step. This verification limits the automated DL-based prostate segmentation methods implementation in clinical practice. A QC system that automatically provides an assessment of segmentation quality could help overcome this limitation and standardize segmentation quality decisions. Such a QC system has been proposed in Paper II. However, little is known about the reproducibility of DL-based segmentation methods for clinical MRI scans [263], which was addressed in Paper III.

Segmentation quality control

The proposed segmentation QC system in Paper II is a transparent and flexible (i.e., easily trainable on different datasets) safety net. The results shown in Paper II indicate that the system performance is acceptable and could prevent poorly segmented cases from continuing through the CAD system. These cases are red flagged and forwarded to the radiologist for correction. This indicates that the intervention of the radiologist will still be necessary from time to time even if automated systems are implemented.

The proposed QC system can also be very helpful in the development of new CAD systems for prostate imaging, as it simplifies the labelling process by integrating DL-based segmentation methods. This saves time, as it allows automated generation of prostate segmentations with acceptable quality. Sørland et al. [260] used the QC system to discard data with low-quality DL-based segmentations from their test set. Patsanis et al. [257] used the QC system to choose between segmentations generated by two different DL-based methods. The selected mask was not only of acceptable quality, but also the one with the highest quality score from either of the networks. This shows that the QC system can also be used to automatically select the best segmentation from a set of segmentations generated by different networks. Incorporating multiple DL-based segmentation methods into one CAD system and followed by an educated selection process can potentially reduce the number of cases requiring radiologist intervention.

Radiomics features were used to train the QC system. Some features such as the wavelet features were not included even though they could improve the performance of the model. These features were excluded because they are expected to increase the complexity of the model and hence the processing time. The combination of radiomics and LASSO has been shown to work well, as LASSO performs feature selection and assigns appropriate weights to the features to increase the model accuracy and interpretability [218].

The proposed QC system was only developed for WP segmentation. The proposed system could be specifically useful for clinical applications that are sensitive to errors in WP segmentation, such as MRI-ultrasound fusion for targeted biopsies [41], and prostate-targeted MR-guided radiotherapy [196]. The performance of DL-based segmentation methods was shown to be comparable to that of radiologists for WP segmentation [170]. Recently, DL networks such as nnU-Net have also shown good performance for prostate zones segmentation [158]. Therefore, and for future work, the proposed QC system could be extended to cover the DL-based segmentation models for prostate zones, which would make it useful for more clinical applications.

One of the concerns about the proposed QC system is processing time. The total time required to generate a mask using a DL-based method and check its quality is about one minute. Of course, this time may vary depending on the computational power of the device, but in the end, this time will still be less than the time required for a radiologist to perform the same tasks. Most importantly, it may help implementing DL-based segmentation methods in the clinic, as it helps detect the segmentation failures.

A potential drawback of the proposed QC system is that it performs differently on different datasets. This could be due to the unbalanced and biased reference quality scores, which are dependent on the manual segmentation quality, used to train the model. The general model used by the QC system showed that the combination of the different datasets in training increased the overall robustness and generalizability of the model. Despite the good performance of the general model, there were some outliers, indicating that the system is not perfect and may over- or underestimate the quality score. To obtain the best possible performance for a new dataset, it might be necessary to retrain the model with a balanced subset of that dataset. For future work, the proposed system could be improved by training it with a large and diverse dataset containing segmentations generated by several radiologists and a variety of DL-based segmentation methods.

A clinical evaluation of the CAD systems, integrating DL-based prostate segmentation and the proposed QC system, is still required. Such an evaluation will identify any compatibility or integration difficulties. It will also allow radiologists to explore automated segmentation, with its capabilities and pitfalls, and the potential added value of the QC system in this context.

Segmentation reproducibility

For clinical applications based on multiple scans in time, such as active surveillance, it is critical that the CAD systems used are reproducible [249,250]. The lack of reproducibility could be a reason not to use CAD in the clinic. Currently, very little is known about the reproducibility of DL-based segmentation methods [263], which are an important component of the fully automated CAD system.

In Paper III the reproducibility of DL-based segmentation was investigated by comparing 14 radiomics shape features from two T2W MR scans acquired with short time intervals (median = 7 days). The investigation led to the conclusion that the overall reproducibility of the DL-based segmentations was comparable to manual segmentations. The exception was the V-Net segmentation of PZ, which was found to be significantly less reproducible than manual for 7/14 features. The study also highlighted the influence of the biopsy guiding probe on prostate deformation, reducing the reproducibility of Elongation, Flatness and Sphericity features in WP and non-PZ for the manual and automated segmentations.

The study also showed that the inclusion of a post-processing step for DL-based segmentation, where only the largest connected component is retained, can remarkably increase reproducibility. Implementing this post-processing step costs no more than a few seconds in processing time, and thus its inclusion in CAD is recommended. Similarly, implementing the QC system proposed in Paper II and excluding cases with low quality segmentations leads to a more reproducible DL-based segmentation.

In addition to WP segmentation, the reproducibility of PZ and non-PZ segmentations was investigated, as the DL-based segmentation methods for PZ and non-PZ recently started reporting good results [158]. The reported DSCs in Paper III show that overall, but specifically for PZ, the networks that require 3D input images perform better than those that perform the segmentation slice-by-slice (2D).

Manual segmentation in Paper III was performed by a single radiologist. This raises concerns about the possibility of bias since the same radiologist provided the masks for training the CNNs. Multiple readers may be needed to ensure that there is no bias or that the CNNs are not simply imitating the style of just one radiologist. In addition, the study used a dataset that came from a single centre. For a better overall understanding of the reproducibility of DL-based segmentation, a multicentre dataset with manual segmentations from multiple readers would be needed.

5.5 Registration

Registration is one of CAD workflow stages and it can be performed before or after segmentation, depending on the application [107]. Registration can be very useful in clinical applications, such as MRI-ultrasound fusion for targeted biopsies, where the suspicious lesions are segmented on mpMR images and overlaid on the ultrasound images, allowing the operator to locate the areas to be biopsied [41,83]. It may also be useful to facilitate the extraction of radiomics features from the different mpMRI sequences to improve the performance of the classifiers. In that case, VOIs are segmented manually or automatically on one sequence, usually the T2W sequence, and then the generated mask is overlaid on the images of the other sequences to extract features [108].

Since some CAD systems use registration to allow one image segmentation to be used by another, the quality of the segmentation is critical. If the overlaid mask was faulty, this would result in unrepresentative features being extracted from the registered images. Furthermore, this could result in suspicious areas not being properly detected when MRI-ultrasound fusion is used for targeted biopsies. Such a problem might be avoided by implementing the segmentation QC system proposed in Paper II.

Another important aspect is the accuracy of the registration process and the quality of the registered images. There are several traditional ML-based [107,108,264,265] and DL-based [266-268] methods have been developed for medical image registration. However, as in the segmentation case, the quality of these methods needs to be controlled to avoid the patient misdiagnosis and to increase CAD robustness. Consequently, developing a QC system for mpMRI registration is an interesting topic for future work.

5.6 Feature extraction and radiomics

A key characteristic of a successful CAD system is extracting distinctive features for the task at hand [107]. Radiomics features have demonstrated the ability to exploit the big data generated by mpMRI [110,111]. Combining features can improve classifier performance and thus the performance of CAD [112,113]. Therefore, the interest in radiomics for prostate cancer diagnosis and treatment has increased in the last few years [110,112,123].

Radiomics features can be divided into hand-crafted (traditional ML-based) and non-hand-crafted (DL-based) [199]. The employment of hand-crafted features in a CAD system, make it easier to understand which features were selected to be fed into the model and how the classifiers reach their decision, which is not the case with DL-based radiomics [199,269]. The DL-based radiomics features are the features that a deep artificial neural network extracts through multiple layers with different filters before passing them to the classification part of the network architecture [199]. By their nature, artificial neural networks extract a larger number of features than traditional hand-crafted radiomics [270].

Due to the lack of understanding of what and why the artificial neural network extracts, the use of DL-based radiomics has raised trustworthy concerns despite their excellent performance [269]. However, although it is difficult to understand what is going on in a neural network, it is possible to visualize the extracted features after each layer [271]. The visualized features from the images shows that shape-related features seems to be extracted in the shallow layers of the network, and a wide range of divergent features, including textural features, seem to be extracted in the deep layers [272]. This might explain why most of the shape features were selected by LASSO during the development of the segmentation QC system proposed in Paper II.

Normalization is an essential step in radiomics feature extraction because it increases the robustness and reproducibility of radiomics features [258,273-275]. This was one of the motivations for developing the normalization method proposed in Paper I. It can be hypothesized that an improved normalization method will lead to more robust and reproducible the radiomics features will be, and hence the better the performance of the classifier. AutoRef was studied in Dewi et al. [258] and found to increase the reproducibility of the extracted radiomics features from T2W-MR.

5.7 Classification

Classification is the final stage of the CAD workflow. After extracting the features, a traditional ML-based or DL-based classifier is trained with the features as predictors and the reference classes as responses [107,108]. The output of the classifier in CAD systems is usually disease detection or diagnosis, depending on the goal of the system [107,108].

In recent years, several traditional ML-based and DL based classifiers have been developed [107,108,171], see *Section 1.5.1* and *Section 1.5.2*. The performance of DL-based classifiers for prostate cancer detection and diagnosis is promising [171,276]. However, DL-based feature extraction and classification techniques still require pre-processing and segmentation. The robustness and accuracy of these two stages have a great impact on the final output of the CAD system, regardless of the classification technique. Recently, Patsanis et al. [257] developed an automated end-to-end pipeline for evaluating generative adversarial networks for prostate cancer detection. In their work, they implemented the normalization method proposed in Paper I and the QC system proposed in Paper II and found that they improved the classification performance.

Classification is the last stage of the CAD workflow and its performance is influenced by the accuracy and quality of the preceding stages. Lemaître et al. [191] showed in their work that selecting a good normalization approach can improve classification accuracy for traditional ML-based classification. Swiderska-Chadaj et al. [277] came to the same conclusion with respect to DL-based classification. Gao et al. [278] showed that segmentation optimization can also improve classification performance. In this thesis, methods were proposed to reduce the possibility of errors in the normalization (Paper I) and segmentation (Paper II and Paper III) stages. The better the normalization and segmentation, the more representative, accurate and reproducible are the extracted features, i.e. the predictors of the classifier.

Another important aspect is the generalizability of the classification. A generalizable classifier should be balanced and applicable to the new unseen dataset, regardless of which institution or population it came from [279]. Generalizable classifiers allow CAD systems to be used across institutions without fear of system bias, which can lead to over- or under-diagnosis and consequently incorrect patient treatment [241]. Training the classifier with large and diverse data and avoiding overfitting and underfitting the model can help develop a generalizable classifier [280]. The developed classifier should also have balanced results and avoid high false-positive and false-negative rates to avoid overdiagnosis or underdiagnosis of patients [107,281]. Open access repositories are a good source of large and diverse data that can be used to increase the accuracy and generalizability of the classifier, and benchmark it [282]. These repositories could save the hassle of acquiring medical images and all the legal and technical issues involved. In this thesis, the open access datasets of the PROMISE12 [202] and PROSTATEx [203] challenges were used in Paper I and Paper II. It is a good practice to separate the training, validation, and testing sets of the classifier [283]. This separation can help avoid the risk of overfitting, increase the estimation accuracy of the model performance and improve the generalizability of the classifier [283]. One possible approach to train more generalizable classifiers with variant data from different countries and institutions is federated learning. Federated learning performs at the client level, so there is no need to transfer the data to have it in one place [284,285]. In federated learning, each client (e.g., institution/hospital) can train the classifier on its own data. The weights of the classifier are then uploaded to a server that is shared by all clients, so that the other clients can download the weights and continue the training process of the classifier with their data [284,285].

5.8 Research ethics, data management and privacy aspects

The use of medical images to develop AI-based systems, e.g., CAD, is subject to a specific procedure including collecting ethical approvals, data access, querying data, data de-identification, data transfer and storage, QC, structuring data, and labelling data [286].

Medical images and all supporting clinical data are considered sensitive data. Data that should not be collected without ethical approvals [286]. Approvals are usually granted by institutional and/or local ethics committees before the study begins. Ethics committee members evaluate the benefits, harms, and risks of the systems to be developed [286]. For most medical image analysis studies, informed or passive consent must be obtained from patients [286]. Researchers and developers of AI systems have an obligation to respect the dignity and rights of patients [287]. Patient data should be secured, not sold, and not used in a way that contradicts the ethical consent given [287]. In this thesis, the ethical perspective was explored before starting the work. The ethical approvals to collect and use the in-house collected dataset were obtained from the Regional Committee for Medical and Health Research Ethics (REC Mid Norway; identifiers 2013/1869 and 2017/576), and signed informed consent was obtained from patients. Patients retained the right to withdraw consent to the use of their images at any time during the study period. To ensure the patients involvement, the thesis work was discussed during the users' panel meetings with the researchers at the MR Cancer group at the Norwegian University of Science and Technology. The panel (https://www.ntnu.edu/isb/mr-cancer-user-involvement), which was established in close dialogue with the Norwegian Cancer Society, includes four former patients, two breast cancer and two prostate cancer patients, who provide important insights and participate in ethical and scientific discussions.

Medical images are typically accessed, queried and retrieved through Image Archiving and Communication System (PACS) [287]. Each medical institution has its own PACS, which requires access permission, granted after reviewing the access request and ensuring that ethical approval has been obtained. Locating the desired data in PACS is a tedious and time-consuming process that must be done carefully. Researchers should not view data for which they have not granted access permission. In this thesis, these guidelines were carefully followed and the permission to access and retrieve data from PACS was granted by the department of Radiology and Nuclear Medicine, St. Olavs Hospital, Trondheim University Hospital, Trondheim, Norway.

Before using the data to develop the CAD system, the data should be de-identified [286]. This includes anonymizing/pseudonymizing the metadata of the images and renaming the cases [288,289]. If a list is needed to link patients to their new pseudonymized identifiers, this list should be carefully stored in a different location than the data. Once de-identified, the data can be transferred to a secure storage point for later use in developing the system [286]. In this thesis, the European General Data Protection Regulation (GDPR) act [290] was followed, the data was pseudonymized and uploaded to a secure server on HUNT Cloud [291] and the link list was stored securely in a different location. HUNT Cloud is am ISO-certified digital infrastructure that allows data controllers and researchers to store, access and analyse sensitive data in controlled environments. HUNT Cloud is in compliance with the European GDPR and Norwegian acts and regulations for research and data security.

After transferring the data to a secure server, it should be systematically structured and checked for quality before being used for AI system development to avoid inherited errors [286]. All or part of the data should be labelled by experts to provide references for AI system development and evaluation [286]. All these aspects have been considered and followed in this work.

To ensure a good data management protocol in this thesis, the data and the AutoRef and segmentation QC code was treated according to the FAIR (Findability, Accessibility, Interoperability, and Reusability) principles [292]. AutoRef and the segmentation QC code have been made publicly available on GitHub (https://github.com/ntnu-mr-cancer) with clear instructions for their use

6 Conclusions and future perspectives

This thesis aimed to facilitate the integration of automated CAD systems of prostate cancer using mpMRI into clinical practice by developing and evaluating new image pre-processing, segmentation and quality control methods to improve the performance of the CAD workflow.

CAD systems have the potential to overcome many of the pitfalls of traditional prostate cancer diagnostics. Especially when integrated with mpMRI, which provides multiple anatomical and functional parameters and quantitative information that can improve the diagnostic process. CAD usually consists of a chain of steps, which implement ML-based methods to achieve a specific task. Each step depends on the performance of the previous step, i.e., if one of the steps fails or commits an error, the following steps are prone to propagate that error, potentially leading to misdiagnosis. Therefore, the implemented methods should be automated, accurate and transparent. The work in this thesis focused on the early steps of the CAD workflow, in particular the T2W MRI normalization and prostate segmentation, as ensuring high performance and error control of these steps reduces the risk of propagated errors. This could not only improve the performance of CAD, but also increase the confidence of the radiologists in these systems.

T2W MR images require normalization of signal intensity to allow quantitative analysis, which is the direction CAD and related statistical models follow. Several normalization methods have been proposed for this purpose, but with limitations. In this thesis a new dual-reference tissue normalization approach that automatically extracts the signal intensity of the fat and muscle around the prostate to normalize the image was proposed. To the best of our knowledge, this is the first multi-reference tissue normalization approach where the delineation of ROIs is fully automated. The proposed method was found to increase the intensity homogeneity between patients and within patients scanned multiple times. Moreover, it showed better performance than other commonly used normalization methods. The method was also shown to improve classification between healthy and malignant prostate tissue in PZ and non-PZ. The proposed method is generalizable, transparent, easy to implement and made publicly available.

Another important step in the CAD workflow is VOI segmentation, in this case of the prostate. This step, which defines the VOIs to be used later for feature extraction, could be efficiently performed automatically using DL-based methods. Despite the overall good performance, these methods can still produce poor segmentation masks in some cases, which calls for a QC step. In this thesis, a generalizable, transparent, publicly available segmentation QC system was developed. The system assigns a score to each segmentation related to its quality, which can be used to distinguish between acceptable and poor segmentations. This system is an important step towards implementing DL-based segmentation methods in clinical practice and reducing human intervention.

DL-based segmentations could also be used in clinical applications that rely on multiple scans in time, such as for patients on active surveillance. Therefore, it is extremely important that the segmentations generated by DL-based methods are not only accurate but also reproducible. In this thesis, the reproducibility of DL-based segmentations of the prostate and its zones was investigated. The reproducibility of the best-performing DL-based methods were found to be comparable to that of manual segmentations.

In conclusion, this thesis shows that the performance of the early steps of automated CAD for prostate cancer can be improved and controlled, leading to more generalizable, transparent and trustworthy systems. This is seen as an important step towards the integration of CAD systems into clinical practice.

This thesis could be fundamental for further research to improve robust, generalizable and transparent CAD systems for prostate cancer. Normalization can be further improved by developing new methods that build on AutoRef and perhaps incorporate additional reference tissues. The segmentation QC system can be extended to include prostate zones segmentation. Developing a similar QC system for mpMRI registration would be helpful and may improve the performance of CAD. Conducting studies to investigate the reproducibility of the various radiomics features and pre-processing steps would be very informative and would provide useful suggestions to extract features correctly. CAD systems have the potential to decide whether or not biopsy sampling is necessary, help detect the suspicious areas and help targeting them when biopsy sampling is necessary. Despite the existence of several CAD systems aimed at detecting or grading prostate cancer, there is still room for the development of more robust and trustworthy systems. In the era of open science, these systems should benefit from previous research and methods developed for the various CAD steps to ensure better performance than the previous systems. However, the most important step for the future is to test the various CAD systems in the clinic and ensure that they meet the radiologists' expectations. This is crucial for building a trust relationship between the radiologists and the CAD systems, which will hopefully lead to the actual implementation of CAD in the clinic.

7 Bibliography

1. Stratton, M.R.; Campbell, P.J.; Futreal, P.A. The cancer genome. *Nature* **2009**, *458*, 719-724, doi:10.1038/nature07943.
2. Hanahan, D.; Weinberg, R.A. The hallmarks of cancer. *Cell* **2000**, *100*, 57-70, doi:10.1016/s0092-8674(00)81683-9.
3. Hanahan, D.; Weinberg, R.A. Hallmarks of cancer: the next generation. *Cell* **2011**, *144*, 646-674, doi:10.1016/j.cell.2011.02.013.
4. Bray, F.; Ferlay, J.; Soerjomataram, I.; Siegel, R.L.; Torre, L.A.; Jemal, A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* **2018**, *68*, 394-424, doi:10.3322/caac.21492.
5. Siegel, R.L.; Miller, K.D.; Jemal, A. Cancer statistics, 2020. *CA Cancer J Clin* **2020**, *70*, 7-30, doi:10.3322/caac.21590.
6. Shah, R.B.; Zhou, M. *Prostate Biopsy Interpretation: An Illustrated Guide*; Springer Berlin Heidelberg: 2012; pp. 1 online resource.
7. Lee, C.H.; Akin-Olugbade, O.; Kirschenbaum, A. Overview of prostate anatomy, histology, and pathology. *Endocrinol Metab Clin North Am* **2011**, *40*, 565-575, viii-ix, doi:10.1016/j.ecl.2011.05.012.
8. Kierszenbaum, A.L.; Tres, L.L. *Histology and cell biology : an introduction to pathology*, 3rd ed.; Elsevier Saunders: Philadelphia, 2012; pp. xiv, 701 p.
9. Verma, S.; Rajesh, A. A clinically relevant approach to imaging prostate cancer: review. *AJR Am J Roentgenol* **2011**, *196*, S1-10 Quiz S11-14, doi:10.2214/AJR.09.7196.
10. Canadian Cancer Society. The prostate. Availabe online: https://www.cancer.ca/en/cancer-information/cancer-type/prostate/prostate-cancer/the-prostate (accessed on 05.Jan.2021).
11. Hall, J.E.; Guyton, A.C. *Guyton and Hall textbook of medical physiology*, Thirteenth edition. ed.; Elsevier: Philadelphia, PA, 2016; pp. xix, 1145 pages.
12. Kumar, V.; Abbas, A.K.; Aster, J.C.; Robbins, S.L. *Robbins basic pathology*, Tenth edition. ed.; Elsevier: Philadelphia, Pennsylvania, 2018; pp. xiv, 935 pages.
13. Cancer Registry of Norway. Cancer in Norway 2019 - Cancer incidence, mortality, survival and prevalence in Norway. *Oslo: Cancer Registry of Norway* **2020**.
14. Pernar, C.H.; Ebot, E.M.; Wilson, K.M.; Mucci, L.A. The Epidemiology of Prostate Cancer. *Cold Spring Harb Perspect Med* **2018**, *8*, doi:10.1101/cshperspect.a030361.
15. Perdana, N.R.; Mochtar, C.A.; Umbas, R.; Hamid, A.R. The Risk Factors of Prostate Cancer and Its Prevention: A Literature Review. *Acta Med Indones* **2016**, *48*, 228-238.
16. Gann, P.H. Risk factors for prostate cancer. *Rev Urol* **2002**, *4 Suppl 5*, S3-S10.
17. Steinberg, G.D.; Carter, B.S.; Beaty, T.H.; Childs, B.; Walsh, P.C. Family history and the risk of prostate cancer. *Prostate* **1990**, *17*, 337-347, doi:10.1002/pros.2990170409.

18. Peisch, S.F.; Van Blarigan, E.L.; Chan, J.M.; Stampfer, M.J.; Kenfield, S.A. Prostate cancer progression and mortality: a review of diet and lifestyle factors. *World J Urol* **2017**, *35*, 867-874, doi:10.1007/s00345-016-1914-3.

19. Calle, E.E.; Rodriguez, C.; Walker-Thurmond, K.; Thun, M.J. Overweight, obesity, and mortality from cancer in a prospectively studied cohort of U.S. adults. *N Engl J Med* **2003**, *348*, 1625-1638, doi:10.1056/NEJMoa021423.

20. Miller, D.C.; Hafez, K.S.; Stewart, A.; Montie, J.E.; Wei, J.T. Prostate carcinoma presentation, diagnosis, and staging: an update form the National Cancer Data Base. *Cancer* **2003**, *98*, 1169-1178, doi:10.1002/cncr.11635.

21. Leslie, S.W.; Soon-Sutton, T.L.; Sajjad, H.; Siref, L.E. Prostate Cancer. In *StatPearls*, Treasure Island (FL), 2020.

22. Saad, F.; Clarke, N.; Colombel, M. Natural history and treatment of bone complications in prostate cancer. *Eur Urol* **2006**, *49*, 429-440, doi:10.1016/j.eururo.2005.12.045.

23. Barentsz, J.O.; Richenberg, J.; Clements, R.; Choyke, P.; Verma, S.; Villeirs, G.; Rouviere, O.; Logager, V.; Futterer, J.J.; European Society of Urogenital, R. ESUR prostate MR guidelines 2012. *Eur Radiol* **2012**, *22*, 746-757, doi:10.1007/s00330-011-2377-y.

24. Mottet, N.; Bellmunt, J.; Bolla, M.; Briers, E.; Cumberbatch, M.G.; De Santis, M.; Fossati, N.; Gross, T.; Henry, A.M.; Joniau, S., et al. EAU-ESTRO-SIOG Guidelines on Prostate Cancer. Part 1: Screening, Diagnosis, and Local Treatment with Curative Intent. *Eur Urol* **2017**, *71*, 618-629, doi:10.1016/j.eururo.2016.08.003.

25. Nadler, R.B.; Humphrey, P.A.; Smith, D.S.; Catalona, W.J.; Ratliff, T.L. Effect of inflammation and benign prostatic hyperplasia on elevated serum prostate specific antigen levels. *J Urol* **1995**, *154*, 407-413, doi:10.1097/00005392-199508000-00023.

26. Thompson, I.M.; Pauler, D.K.; Goodman, P.J.; Tangen, C.M.; Lucia, M.S.; Parnes, H.L.; Minasian, L.M.; Ford, L.G.; Lippman, S.M.; Crawford, E.D., et al. Prevalence of prostate cancer among men with a prostate-specific antigen level < or =4.0 ng per milliliter. *N Engl J Med* **2004**, *350*, 2239-2246, doi:10.1056/NEJMoa031918.

27. McGuire, B.B.; Helfand, B.T.; Loeb, S.; Hu, Q.; O'Brien, D.; Cooper, P.; Yang, X.; Catalona, W.J. Outcomes in patients with Gleason score 8-10 prostate cancer: relation to preoperative PSA level. *BJU Int* **2012**, *109*, 1764-1769, doi:10.1111/j.1464-410X.2011.10628.x.

28. Loeb, S.; Catalona, W.J. What is the role of digital rectal examination in men undergoing serial screening of serum PSA levels? *Nat Clin Pract Urol* **2009**, *6*, 68-69, doi:10.1038/ncpuro1294.

29. Kirby, R. The role of PSA in detection and management of prostate cancer. *Practitioner* **2016**, *260*, 17-21, 13.

30. Bangma, C.H.; Roemeling, S.; Schroder, F.H. Overdiagnosis and overtreatment of early detected prostate cancer. *World J Urol* **2007**, *25*, 3-9, doi:10.1007/s00345-007-0145-z.

31. Helsedirektør. Nasjonalt handlingsprogram med retningslinjer for diagnostikk, behandling og oppfølging av prostatakreft: NASJONALE FAGLIGE

RETNINGSLINJER. *Helsedirektoratet, Avdeling spesialisthelsetjenester, Oslo* **2020**.

32. U. S. Preventive Services Task Force; Grossman, D.C.; Curry, S.J.; Owens, D.K.; Bibbins-Domingo, K.; Caughey, A.B.; Davidson, K.W.; Doubeni, C.A.; Ebell, M.; Epling, J.W., Jr., et al. Screening for Prostate Cancer: US Preventive Services Task Force Recommendation Statement. *JAMA* **2018**, *319*, 1901-1913, doi:10.1001/jama.2018.3710.
33. Naji, L.; Randhawa, H.; Sohani, Z.; Dennis, B.; Lautenbach, D.; Kavanagh, O.; Bawor, M.; Banfield, L.; Profetto, J. Digital Rectal Examination for Prostate Cancer Screening in Primary Care: A Systematic Review and Meta-Analysis. *Ann Fam Med* **2018**, *16*, 149-154, doi:10.1370/afm.2205.
34. Wallner, L.; Frencher, S.; Hsu, J.W.; Loo, R.; Huang, J.; Nichol, M.; Jacobsen, S. Prostate cancer screening trends in a large, integrated health care system. *Perm J* **2012**, *16*, 4-9.
35. Gosselaar, C.; Kranse, R.; Roobol, M.J.; Roemeling, S.; Schroder, F.H. The interobserver variability of digital rectal examination in a large randomized trial for the screening of prostate cancer. *Prostate* **2008**, *68*, 985-993, doi:10.1002/pros.20759.
36. Cui, T.; Kovell, R.C.; Terlecki, R.P. Is it time to abandon the digital rectal examination? Lessons from the PLCO Cancer Screening Trial and peer-reviewed literature. *Curr Med Res Opin* **2016**, *32*, 1663-1669, doi:10.1080/03007995.2016.1198312.
37. Ilic, D.; Neuberger, M.M.; Djulbegovic, M.; Dahm, P. Screening for prostate cancer. *Cochrane Database Syst Rev* **2013**, 10.1002/14651858.CD004720.pub3, CD004720, doi:10.1002/14651858.CD004720.pub3.
38. Donovan, J.; Hamdy, F.; Neal, D.; Peters, T.; Oliver, S.; Brindle, L.; Jewell, D.; Powell, P.; Gillatt, D.; Dedman, D., et al. Prostate Testing for Cancer and Treatment (ProtecT) feasibility study. *Health Technol Assess* **2003**, *7*, 1-88, doi:10.3310/hta7140.
39. Hambrock, T.; Hoeks, C.; Hulsbergen-van de Kaa, C.; Scheenen, T.; Futterer, J.; Bouwense, S.; van Oort, I.; Schroder, F.; Huisman, H.; Barentsz, J. 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* **2012**, *61*, 177-184, doi:10.1016/j.eururo.2011.08.042.
40. Kvale, R.; Moller, B.; Wahlqvist, R.; Fossa, S.D.; Berner, A.; Busch, C.; Kyrdalen, A.E.; Svindland, A.; Viset, T.; Halvorsen, O.J. Concordance between Gleason scores of needle biopsies and radical prostatectomy specimens: a population-based study. *BJU Int* **2009**, *103*, 1647-1654, doi:10.1111/j.1464-410X.2008.08255.x.
41. Jayadevan, R.; Zhou, S.; Priester, A.M.; Delfin, M.; Marks, L.S. Use of MRI-ultrasound Fusion to Achieve Targeted Prostate Biopsy. *J Vis Exp* **2019**, 10.3791/59231, doi:10.3791/59231.
42. Moore, C.M.; Robertson, N.L.; Arsanious, N.; Middleton, T.; Villers, A.; Klotz, L.; Taneja, S.S.; Emberton, M. Image-guided prostate biopsy using magnetic

resonance imaging-derived targets: a systematic review. *Eur Urol* **2013**, *63*, 125-140, doi:10.1016/j.eururo.2012.06.004.

43. Murphy, I.G.; NiMhurchu, E.; Gibney, R.G.; McMahon, C.J. MRI-directed cognitive fusion-guided biopsy of the anterior prostate tumors. *Diagn Interv Radiol* **2017**, *23*, 87-93, doi:10.5152/dir.2016.15445.
44. Weinreb, J.C.; Barentsz, J.O.; Choyke, P.L.; Cornud, F.; Haider, M.A.; Macura, K.J.; Margolis, D.; Schnall, M.D.; Shtern, F.; Tempany, C.M., et al. PI-RADS Prostate Imaging - Reporting and Data System: 2015, Version 2. *Eur Urol* **2016**, *69*, 16-40, doi:10.1016/j.eururo.2015.08.052.
45. Mottet, N.; van den Bergh, R.C.N.; Briers, E.; Van den Broeck, T.; Cumberbatch, M.G.; De Santis, M.; Fanti, S.; Fossati, N.; Gandaglia, G.; Gillessen, S., et al. EAU-EANM-ESTRO-ESUR-SIOG Guidelines on Prostate Cancer-2020 Update. Part 1: Screening, Diagnosis, and Local Treatment with Curative Intent. *Eur Urol* **2021**, *79*, 243-262, doi:10.1016/j.eururo.2020.09.042.
46. Gleason, D.F.; Mellinger, G.T. Prediction of prognosis for prostatic adenocarcinoma by combined histological grading and clinical staging. *J Urol* **1974**, *111*, 58-64, doi:10.1016/s0022-5347(17)59889-4.
47. Pierorazio, P.M.; Walsh, P.C.; Partin, A.W.; Epstein, J.I. Prognostic Gleason grade grouping: data based on the modified Gleason scoring system. *BJU Int* **2013**, *111*, 753-760, doi:10.1111/j.1464-410X.2012.11611.x.
48. Chen, N.; Zhou, Q. The evolving Gleason grading system. *Chin J Cancer Res* **2016**, *28*, 58-64, doi:10.3978/j.issn.1000-9604.2016.02.04.
49. Epstein, J.I.; Egevad, L.; Amin, M.B.; Delahunt, B.; Srigley, J.R.; Humphrey, P.A.; Grading, C. The 2014 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason Grading of Prostatic Carcinoma: Definition of Grading Patterns and Proposal for a New Grading System. *Am J Surg Pathol* **2016**, *40*, 244-252, doi:10.1097/PAS.0000000000000530.
50. Buyyounouski, M.K.; Choyke, P.L.; McKenney, J.K.; Sartor, O.; Sandler, H.M.; Amin, M.B.; Kattan, M.W.; Lin, D.W. Prostate cancer - major changes in the American Joint Committee on Cancer eighth edition cancer staging manual. *CA Cancer J Clin* **2017**, *67*, 245-253, doi:10.3322/caac.21391.
51. Nketiah, G.A. Magnetic Resonance Imaging for Improved Prostate Cancer Diagnosis. Doctoral thesis. Norwegian University of Science and Technology (NTNU), Trondheim, 2018.
52. Filson, C.P.; Marks, L.S.; Litwin, M.S. Expectant management for men with early stage prostate cancer. *CA Cancer J Clin* **2015**, *65*, 265-282, doi:10.3322/caac.21278.
53. Romero-Otero, J.; Garcia-Gomez, B.; Duarte-Ojeda, J.M.; Rodriguez-Antolin, A.; Vilaseca, A.; Carlsson, S.V.; Touijer, K.A. Active surveillance for prostate cancer. *Int J Urol* **2016**, *23*, 211-218, doi:10.1111/iju.13016.
54. Jayadevappa, R.; Chhatre, S.; Wong, Y.N.; Wittink, M.N.; Cook, R.; Morales, K.H.; Vapiwala, N.; Newman, D.K.; Guzzo, T.; Wein, A.J., et al. Comparative effectiveness of prostate cancer treatments for patient-centered outcomes: A systematic review and meta-analysis (PRISMA Compliant). *Medicine (Baltimore)* **2017**, *96*, e6790, doi:10.1097/MD.0000000000006790.

55. Mohan, R.; Schellhammer, P.F. Treatment options for localized prostate cancer. *Am Fam Physician* **2011**, *84*, 413-420.

56. Mongiat-Artus, P.; Peyromaure, M.; Richaud, P.; Droz, J.P.; Rainfray, M.; Jeandel, C.; Rebillard, X.; Moreau, J.L.; Davin, J.L.; Salomon, L., et al. [Recommendations for the treatment of prostate cancer in the elderly man: A study by the oncology committee of the French association of urology]. *Prog Urol* **2009**, *19*, 810-817, doi:10.1016/j.purol.2009.02.008.

57. Resnick, M.J.; Lacchetti, C.; Bergman, J.; Hauke, R.J.; Hoffman, K.E.; Kungel, T.M.; Morgans, A.K.; Penson, D.F. Prostate cancer survivorship care guideline: American Society of Clinical Oncology Clinical Practice Guideline endorsement. *J Clin Oncol* **2015**, *33*, 1078-1085, doi:10.1200/JCO.2014.60.2557.

58. Michaelson, M.D.; Cotter, S.E.; Gargollo, P.C.; Zietman, A.L.; Dahl, D.M.; Smith, M.R. Management of complications of prostate cancer treatment. *CA Cancer J Clin* **2008**, *58*, 196-213, doi:10.3322/CA.2008.0002.

59. Ravery, V. The significance of recurrent PSA after radical prostatectomy: benign versus malignant sources. *Semin Urol Oncol* **1999**, *17*, 127-129.

60. Cornford, P.; Bellmunt, J.; Bolla, M.; Briers, E.; De Santis, M.; Gross, T.; Henry, A.M.; Joniau, S.; Lam, T.B.; Mason, M.D., et al. EAU-ESTRO-SIOG Guidelines on Prostate Cancer. Part II: Treatment of Relapsing, Metastatic, and Castration-Resistant Prostate Cancer. *Eur Urol* **2017**, *71*, 630-642, doi:10.1016/j.eururo.2016.08.002.

61. Roach, M., 3rd; Hanks, G.; Thames, H., Jr.; Schellhammer, P.; Shipley, W.U.; Sokol, G.H.; Sandler, H. Defining biochemical failure following radiotherapy with or without hormonal therapy in men with clinically localized prostate cancer: recommendations of the RTOG-ASTRO Phoenix Consensus Conference. *Int J Radiat Oncol Biol Phys* **2006**, *65*, 965-974, doi:10.1016/j.ijrobp.2006.04.029.

62. Kupelian, P.A.; Mahadevan, A.; Reddy, C.A.; Reuther, A.M.; Klein, E.A. Use of different definitions of biochemical failure after external beam radiotherapy changes conclusions about relative treatment efficacy for localized prostate cancer. *Urology* **2006**, *68*, 593-598, doi:10.1016/j.urology.2006.03.075.

63. Roehl, K.A.; Han, M.; Ramos, C.G.; Antenor, J.A.; Catalona, W.J. Cancer progression and survival rates following anatomical radical retropubic prostatectomy in 3,478 consecutive patients: long-term results. *J Urol* **2004**, *172*, 910-914, doi:10.1097/01.ju.0000134888.22332.bb.

64. Artibani, W.; Porcaro, A.B.; De Marco, V.; Cerruto, M.A.; Siracusano, S. Management of Biochemical Recurrence after Primary Curative Treatment for Prostate Cancer: A Review. *Urol Int* **2018**, *100*, 251-262, doi:10.1159/000481438.

65. Brown, R.W.; Haacke, E.M.; Thompson, M.R.; Venkatesan, R.; Cheng, N. *Magnetic resonance imaging : physical principles and sequence design*, Second edition. ed.; Wiley-Blackwell: Hoboken, New Jersey, 2014; pp. xxxii, 944 pages.

66. Westbrook, C.; Roth, C.K.; Talbot, J. *MRI in practice*, 4th ed.; Wiley-Blackwell: Chichester, 2011; pp. vii, 442 p.

67. Rosenkrantz, A.B.; Kim, S.; Lim, R.P.; Hindman, N.; Deng, F.M.; Babb, J.S.; Taneja, S.S. Prostate cancer localization using multiparametric MR imaging: comparison of Prostate Imaging Reporting and Data System (PI-RADS) and Likert scales. *Radiology* **2013**, *269*, 482-492, doi:10.1148/radiol.13122233.
68. The Norwegian Directorate of Health. Prostatakreft. Availabe online: https://www.helsedirektoratet.no/pakkeforlop/prostatakreft (accessed on 12.Jan.2020).
69. Rabi, II; Zacharias, J.R.; Millman, S.; Kusch, P. Milestones in magnetic resonance: 'a new method of measuring nuclear magnetic moment' . 1938. *J Magn Reson Imaging* **1992**, *2*, 131-133, doi:10.1002/jmri.1880020203.
70. Pound, R.V.; Purcell, E.M. Measurement of Magnetic Resonance Absorption by Nuclear Moments in a Solid. *Phys Rev* **1946**, *69*, 681-681.
71. Bloch, F.; Hansen, W.W.; Packard, M. Nuclear Induction. *Phys Rev* **1946**, *69*, 127-127, doi:DOI 10.1103/PhysRev.69.127.
72. Lauterbur, P.C. Image Formation by Induced Local Interactions - Examples Employing Nuclear Magnetic-Resonance. *Nature* **1973**, *242*, 190-191, doi:DOI 10.1038/242190a0.
73. Mansfield, P.; Grannell, P.K. Nmr Diffraction in Solids. *J Phys C Solid State* **1973**, *6*, L422-L426, doi:Doi 10.1088/0022-3719/6/22/007.
74. Pooley, R.A. AAPM/RSNA physics tutorial for residents: fundamental physics of MR imaging. *Radiographics* **2005**, *25*, 1087-1099, doi:10.1148/rg.254055027.
75. Bushberg, J.T. *The essential physics of medical imaging*, 3rd ed.; Wolters Kluwer Health/Lippincott Williams & Wilkins: Philadelphia, 2012; pp. xii, 1030 p.
76. Bitar, R.; Leung, G.; Perng, R.; Tadros, S.; Moody, A.R.; Sarrazin, J.; McGregor, C.; Christakis, M.; Symons, S.; Nelson, A., et al. MR pulse sequences: What every radiologist wants to know but is afraid to ask. *Radiographics* **2006**, *26*, 513-U515, doi:10.1148/rg.262055063.
77. Jung, B.A.; Weigel, M. Spin echo magnetic resonance imaging. *J Magn Reson Imaging* **2013**, *37*, 805-817, doi:10.1002/jmri.24068.
78. Turkbey, B.; Rosenkrantz, A.B.; Haider, M.A.; Padhani, A.R.; Villeirs, G.; Macura, K.J.; Tempany, C.M.; Choyke, P.L.; Cornud, F.; Margolis, D.J., et al. Prostate Imaging Reporting and Data System Version 2.1: 2019 Update of Prostate Imaging Reporting and Data System Version 2. *Eur Urol* **2019**, *76*, 340-351, doi:10.1016/j.eururo.2019.02.033.
79. Delongchamps, N.B.; Rouanne, M.; Flam, T.; Beuvon, F.; Liberatore, M.; Zerbib, M.; Cornud, F. 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* **2011**, *107*, 1411-1418, doi:10.1111/j.1464-410X.2010.09808.x.
80. Hegde, J.V.; Mulkern, R.V.; Panych, L.P.; Fennessy, F.M.; Fedorov, A.; Maier, S.E.; Tempany, C.M. Multiparametric MRI of prostate cancer: an update on state-of-the-art techniques and their performance in detecting and localizing prostate cancer. *J Magn Reson Imaging* **2013**, *37*, 1035-1054, doi:10.1002/jmri.23860.

81. Scheenen, T.W.; Rosenkrantz, A.B.; Haider, M.A.; Futterer, J.J. Multiparametric Magnetic Resonance Imaging in Prostate Cancer Management: Current Status and Future Perspectives. *Invest Radiol* **2015**, *50*, 594-600, doi:10.1097/RLI.0000000000000163.

82. Selnaes, K.M.; Heerschap, A.; Jensen, L.R.; Tessem, M.B.; Schweder, G.J.; Goa, P.E.; Viset, T.; Angelsen, A.; Gribbestad, I.S. Peripheral zone prostate cancer localization by multiparametric magnetic resonance at 3 T: unbiased cancer identification by matching to histopathology. *Invest Radiol* **2012**, *47*, 624-633, doi:10.1097/RLI.0b013e318263f0fd.

83. Stabile, A.; Giganti, F.; Rosenkrantz, A.B.; Taneja, S.S.; Villeirs, G.; Gill, I.S.; Allen, C.; Emberton, M.; Moore, C.M.; Kasivisvanathan, V. Multiparametric MRI for prostate cancer diagnosis: current status and future directions. *Nat Rev Urol* **2020**, *17*, 41-61, doi:10.1038/s41585-019-0212-4.

84. Vos, E.K.; Kobus, T.; Litjens, G.J.; Hambrock, T.; Hulsbergen-van de Kaa, C.A.; Barentsz, J.O.; Maas, M.C.; Scheenen, T.W. Multiparametric Magnetic Resonance Imaging for Discriminating Low-Grade From High-Grade Prostate Cancer. *Invest Radiol* **2015**, *50*, 490-497, doi:10.1097/RLI.0000000000000157.

85. Fascelli, M.; George, A.K.; Frye, T.; Turkbey, B.; Choyke, P.L.; Pinto, P.A. The role of MRI in active surveillance for prostate cancer. *Curr Urol Rep* **2015**, *16*, 42, doi:10.1007/s11934-015-0507-9.

86. Alberts, A.R.; Roobol, M.J.; Verbeek, J.F.M.; Schoots, I.G.; Chiu, P.K.; Osses, D.F.; Tijsterman, J.D.; Beerlage, H.P.; Mannaerts, C.K.; Schimmoller, L., et al. Prediction of High-grade Prostate Cancer Following Multiparametric Magnetic Resonance Imaging: Improving the Rotterdam European Randomized Study of Screening for Prostate Cancer Risk Calculators. *Eur Urol* **2019**, *75*, 310-318, doi:10.1016/j.eururo.2018.07.031.

87. Patel, P.; Mathew, M.S.; Trilisky, I.; Oto, A. Multiparametric MR Imaging of the Prostate after Treatment of Prostate Cancer. *Radiographics* **2018**, *38*, 437-449, doi:10.1148/rg.2018170147.

88. Schoots, I.G.; Roobol, M.J.; Nieboer, D.; Bangma, C.H.; Steyerberg, E.W.; Hunink, M.G. Magnetic resonance imaging-targeted biopsy may enhance the diagnostic accuracy of significant prostate cancer detection compared to standard transrectal ultrasound-guided biopsy: a systematic review and meta-analysis. *Eur Urol* **2015**, *68*, 438-450, doi:10.1016/j.eururo.2014.11.037.

89. Hricak, H.; Dooms, G.C.; McNeal, J.E.; Mark, A.S.; Marotti, M.; Avallone, A.; Pelzer, M.; Proctor, E.C.; Tanagho, E.A. MR imaging of the prostate gland: normal anatomy. *AJR Am J Roentgenol* **1987**, *148*, 51-58, doi:10.2214/ajr.148.1.51.

90. Rosenkrantz, A.B.; Taneja, S.S. Radiologist, be aware: ten pitfalls that confound the interpretation of multiparametric prostate MRI. *AJR Am J Roentgenol* **2014**, *202*, 109-120, doi:10.2214/AJR.13.10699.

91. Somford, D.M.; Futterer, J.J.; Hambrock, T.; Barentsz, J.O. Diffusion and perfusion MR imaging of the prostate. *Magn Reson Imaging Clin N Am* **2008**, *16*, 685-695, ix, doi:10.1016/j.mric.2008.07.002.

92. Hambrock, T.; Somford, D.M.; Huisman, H.J.; van Oort, I.M.; Witjes, J.A.; Hulsbergen-van de Kaa, C.A.; Scheenen, T.; Barentsz, J.O. Relationship

between apparent diffusion coefficients at 3.0-T MR imaging and Gleason grade in peripheral zone prostate cancer. *Radiology* **2011**, *259*, 453-461, doi:10.1148/radiol.11091409.

93. Wu, L.M.; Xu, J.R.; Ye, Y.Q.; Lu, Q.; Hu, J.N. The clinical value of diffusion-weighted imaging in combination with T2-weighted imaging in diagnosing prostate carcinoma: a systematic review and meta-analysis. *AJR Am J Roentgenol* **2012**, *199*, 103-110, doi:10.2214/AJR.11.7634.

94. Jung, S.I.; Donati, O.F.; Vargas, H.A.; Goldman, D.; Hricak, H.; Akin, O. Transition zone prostate cancer: incremental value of diffusion-weighted endorectal MR imaging in tumor detection and assessment of aggressiveness. *Radiology* **2013**, *269*, 493-503, doi:10.1148/radiol.13130029.

95. Hara, N.; Okuizumi, M.; Koike, H.; Kawaguchi, M.; Bilim, V. Dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) is a useful modality for the precise detection and staging of early prostate cancer. *Prostate* **2005**, *62*, 140-147, doi:10.1002/pros.20124.

96. Hylton, N. Dynamic contrast-enhanced magnetic resonance imaging as an imaging biomarker. *Journal of Clinical Oncology* **2006**, *24*, 3293-3298, doi:10.1200/Jco.2006.06.8080.

97. Verma, S.; Turkbey, B.; Muradyan, N.; Rajesh, A.; Cornud, F.; Haider, M.A.; Choyke, P.L.; Harisinghani, M. Overview of dynamic contrast-enhanced MRI in prostate cancer diagnosis and management. *AJR Am J Roentgenol* **2012**, *198*, 1277-1288, doi:10.2214/AJR.12.8510.

98. Boesen, L.; Norgaard, N.; Logager, V.; Balslev, I.; Bisbjerg, R.; Thestrup, K.C.; Winther, M.D.; Jakobsen, H.; Thomsen, H.S. Assessment of the Diagnostic Accuracy of Biparametric Magnetic Resonance Imaging for Prostate Cancer in Biopsy-Naive Men: The Biparametric MRI for Detection of Prostate Cancer (BIDOC) Study. *JAMA Netw Open* **2018**, *1*, e180219, doi:10.1001/jamanetworkopen.2018.0219.

99. Cosma, I.; Tennstedt-Schenk, C.; Winzler, S.; Psychogios, M.N.; Pfeil, A.; Teichgraeber, U.; Malich, A.; Papageorgiou, I. The role of gadolinium in magnetic resonance imaging for early prostate cancer diagnosis: A diagnostic accuracy study. *PLoS One* **2019**, *14*, e0227031, doi:10.1371/journal.pone.0227031.

100. Jambor, I.; Bostrom, P.J.; Taimen, P.; Syvanen, K.; Kahkonen, E.; Kallajoki, M.; Perez, I.M.; Kauko, T.; Matomaki, J.; Ettala, O., et al. Novel biparametric MRI and targeted biopsy improves risk stratification in men with a clinical suspicion of prostate cancer (IMPROD Trial). *J Magn Reson Imaging* **2017**, *46*, 1089-1095, doi:10.1002/jmri.25641.

101. Smith, C.P.; Turkbey, B. PI-RADS v2: Current standing and future outlook. *Turk J Urol* **2018**, *44*, 189-194, doi:10.5152/tud.2018.12144.

102. Woo, S.; Suh, C.H.; Kim, S.Y.; Cho, J.Y.; Kim, S.H. Diagnostic Performance of Prostate Imaging Reporting and Data System Version 2 for Detection of Prostate Cancer: A Systematic Review and Diagnostic Meta-analysis. *Eur Urol* **2017**, *72*, 177-188, doi:10.1016/j.eururo.2017.01.042.

103. Zhang, L.; Tang, M.; Chen, S.P.; Lei, X.Y.; Zhang, X.L.; Huan, Y. A meta-analysis of use of Prostate Imaging Reporting and Data System Version 2 (PI-

RADS V2) with multiparametric MR imaging for the detection of prostate cancer. *European Radiology* **2017**, *27*, 5204-5214, doi:10.1007/s00330-017-4843-7.

104. Ruprecht, O.; Weisser, P.; Bodelle, B.; Ackermann, H.; Vogl, T.J. MRI of the prostate: interobserver agreement compared with histopathologic outcome after radical prostatectomy. *Eur J Radiol* **2012**, *81*, 456-460, doi:10.1016/j.ejrad.2010.12.076.

105. Litjens, G.; Debats, O.; Barentsz, J.; Karssemeijer, N.; Huisman, H. Computer-aided detection of prostate cancer in MRI. *IEEE Trans Med Imaging* **2014**, *33*, 1083-1092, doi:10.1109/TMI.2014.2303821.

106. Girometti, R.; Giannarini, G.; Greco, F.; Isola, M.; Cereser, L.; Como, G.; Sioletic, S.; Pizzolitto, S.; Crestani, A.; Ficarra, V., et al. Interreader agreement of PI-RADS v. 2 in assessing prostate cancer with multiparametric MRI: A study using whole-mount histology as the standard of reference. *J Magn Reson Imaging* **2019**, *49*, 546-555, doi:10.1002/jmri.26220.

107. Wang, S.; Burtt, K.; Turkbey, B.; Choyke, P.; Summers, R.M. Computer aided-diagnosis of prostate cancer on multiparametric MRI: a technical review of current research. *Biomed Res Int* **2014**, *2014*, 789561, doi:10.1155/2014/789561.

108. Lemaitre, G.; Marti, R.; Freixenet, J.; Vilanova, J.C.; Walker, P.M.; Meriaudeau, F. Computer-Aided Detection and diagnosis for prostate cancer based on mono and multi-parametric MRI: a review. *Comput Biol Med* **2015**, *60*, 8-31, doi:10.1016/j.compbiomed.2015.02.009.

109. Liu, L.; Tian, Z.; Zhang, Z.; Fei, B. Computer-aided Detection of Prostate Cancer with MRI: Technology and Applications. *Acad Radiol* **2016**, *23*, 1024-1046, doi:10.1016/j.acra.2016.03.010.

110. Sun, Y.; Reynolds, H.M.; Parameswaran, B.; Wraith, D.; Finnegan, M.E.; Williams, S.; Haworth, A. Multiparametric MRI and radiomics in prostate cancer: a review. *Australas Phys Eng Sci Med* **2019**, *42*, 3-25, doi:10.1007/s13246-019-00730-z.

111. Gillies, R.J.; Kinahan, P.E.; Hricak, H. Radiomics: Images Are More than Pictures, They Are Data. *Radiology* **2016**, *278*, 563-577, doi:10.1148/radiol.2015151169.

112. Smith, C.P.; Czarniecki, M.; Mehralivand, S.; Stoyanova, R.; Choyke, P.L.; Harmon, S.; Turkbey, B. Radiomics and radiogenomics of prostate cancer. *Abdom Radiol (NY)* **2019**, *44*, 2021-2029, doi:10.1007/s00261-018-1660-7.

113. Stoyanova, R.; Takhar, M.; Tschudi, Y.; Ford, J.C.; Solorzano, G.; Erho, N.; Balagurunathan, Y.; Punnen, S.; Davicioni, E.; Gillies, R.J., et al. Prostate cancer radiomics and the promise of radiogenomics. *Transl Cancer Res* **2016**, *5*, 432-447, doi:10.21037/tcr.2016.06.20.

114. Lambin, P.; Leijenaar, R.T.H.; Deist, T.M.; Peerlings, J.; de Jong, E.E.C.; van Timmeren, J.; Sanduleanu, S.; Larue, R.; Even, A.J.G.; Jochems, A., et al. Radiomics: the bridge between medical imaging and personalized medicine. *Nat Rev Clin Oncol* **2017**, *14*, 749-762, doi:10.1038/nrclinonc.2017.141.

115. Thawani, R.; McLane, M.; Beig, N.; Ghose, S.; Prasanna, P.; Velcheti, V.; Madabhushi, A. Radiomics and radiogenomics in lung cancer: A review for the clinician. *Lung Cancer* **2018**, *115*, 34-41, doi:10.1016/j.lungcan.2017.10.015.

116. van Griethuysen, J.J.M.; Fedorov, A.; Parmar, C.; Hosny, A.; Aucoin, N.; Narayan, V.; Beets-Tan, R.G.H.; Fillion-Robin, J.C.; Pieper, S.; Aerts, H. Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer Res* **2017**, *77*, e104-e107, doi:10.1158/0008-5472.CAN-17-0339.

117. Zwanenburg, A.; Vallieres, M.; Abdalah, M.A.; Aerts, H.; Andrearczyk, V.; Apte, A.; Ashrafinia, S.; Bakas, S.; Beukinga, R.J.; Boellaard, R., et al. The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping. *Radiology* **2020**, *295*, 328-338, doi:10.1148/radiol.2020191145.

118. Haralick, R.M.; Shanmugam, K.; Dinstein, I. Textural Features for Image Classification. *Ieee T Syst Man Cyb* **1973**, *Smc3*, 610-621, doi:Doi 10.1109/Tsmc.1973.4309314.

119. Chu, A.; Sehgal, C.M.; Greenleaf, J.F. Use of Gray Value Distribution of Run Lengths for Texture Analysis. *Pattern Recogn Lett* **1990**, *11*, 415-419, doi:Doi 10.1016/0167-8655(90)90112-F.

120. Thibault, G.F., B.; Navarro, C.; Pereira, S.; Cau, P.; Levy, N.; Sequeira, J.; Mari, J. Texture Indexes and Gray Level Size Zone Matrix Application to Cell Nuclei Classification. Minsk, Belarus, 19–21 May 2009; pp. 140-145.

121. Sun, C.J.; Wee, W.G. Neighboring Gray Level Dependence Matrix for Texture Classification. *Comput Vision Graph* **1983**, *23*, 341-352, doi:Doi 10.1016/0734-189x(83)90032-4.

122. Amadasun, M.; King, R. Textural Features Corresponding to Textural Properties. *Ieee T Syst Man Cyb* **1989**, *19*, 1264-1274, doi:Doi 10.1109/21.44046.

123. Delgadillo, R.; Ford, J.C.; Abramowitz, M.C.; Dal Pra, A.; Pollack, A.; Stoyanova, R. The role of radiomics in prostate cancer radiotherapy. *Strahlenther Onkol* **2020**, *196*, 900-912, doi:10.1007/s00066-020-01679-9.

124. Fehr, D.; Veeraraghavan, H.; Wibmer, A.; Gondo, T.; Matsumoto, K.; Vargas, H.A.; Sala, E.; Hricak, H.; Deasy, J.O. Automatic classification of prostate cancer Gleason scores from multiparametric magnetic resonance images. *Proc Natl Acad Sci U S A* **2015**, *112*, E6265-6273, doi:10.1073/pnas.1505935112.

125. Nketiah, G.; Elschot, M.; Kim, E.; Teruel, J.R.; Scheenen, T.W.; Bathen, T.F.; Selnaes, K.M. T2-weighted MRI-derived textural features reflect prostate cancer aggressiveness: preliminary results. *Eur Radiol* **2017**, *27*, 3050-3059, doi:10.1007/s00330-016-4663-1.

126. Coroller, T.P. Combining data science and medical imaging: Advancing cancer precision medicine with radiomics. Maastricht University, Datawyse / Universitaire Pers Maastricht., 2017.

127. Langlotz, C.P.; Allen, B.; Erickson, B.J.; Kalpathy-Cramer, J.; Bigelow, K.; Cook, T.S.; Flanders, A.E.; Lungren, M.P.; Mendelson, D.S.; Rudie, J.D., et al. A Roadmap for Foundational Research on Artificial Intelligence in Medical Imaging: From the 2018 NIH/RSNA/ACR/The Academy Workshop. *Radiology* **2019**, *291*, 781-791, doi:10.1148/radiol.2019190613.

128. Rajkomar, A.; Dean, J.; Kohane, I. Machine Learning in Medicine. *N Engl J Med* **2019**, *380*, 1347-1358, doi:10.1056/NEJMra1814259.

129. Erickson, B.J.; Korfiatis, P.; Akkus, Z.; Kline, T.L. Machine Learning for Medical Imaging. *Radiographics* **2017**, *37*, 505-515, doi:10.1148/rg.2017160130.

130. Huang, S.; Yang, J.; Fong, S.; Zhao, Q. Artificial intelligence in cancer diagnosis and prognosis: Opportunities and challenges. *Cancer Lett* **2020**, *471*, 61-71, doi:10.1016/j.canlet.2019.12.007.

131. Shimizu, H.; Nakayama, K.I. Artificial intelligence in oncology. *Cancer Sci* **2020**, *111*, 1452-1460, doi:10.1111/cas.14377.

132. Cuocolo, R.; Cipullo, M.B.; Stanzione, A.; Ugga, L.; Romeo, V.; Radice, L.; Brunetti, A.; Imbriaco, M. Machine learning applications in prostate cancer magnetic resonance imaging. *Eur Radiol Exp* **2019**, *3*, 35, doi:10.1186/s41747-019-0109-2.

133. Lundervold, A.S.; Lundervold, A. An overview of deep learning in medical imaging focusing on MRI. *Z Med Phys* **2019**, *29*, 102-127, doi:10.1016/j.zemedi.2018.11.002.

134. Mazurowski, M.A.; Buda, M.; Saha, A.; Bashir, M.R. Deep learning in radiology: An overview of the concepts and a survey of the state of the art with focus on MRI. *J Magn Reson Imaging* **2019**, *49*, 939-954, doi:10.1002/jmri.26534.

135. Montgomery, D.C.; Peck, E.A.; Vining, G.G. *Introduction to linear regression analysis*, 5th ed.; Wiley: Hoboken, N.J., 2012; pp. xvi, 645 p.

136. Hosmer, D.W.; Lemeshow, S.; Cook, E.D. *Applied logistic regression*, 2nd ed.; Wiley: New York ; Chichester, 2000; pp. xii, 373 p.

137. Tibshirani, R. Regression shrinkage and selection via the Lasso. *J Roy Stat Soc B Met* **1996**, *58*, 267-288, doi:DOI 10.1111/j.2517-6161.1996.tb02080.x.

138. Quinlan, J.R. Simplifying Decision Trees. *Int J Man Mach Stud* **1987**, *27*, 221-234, doi:Doi 10.1016/S0020-7373(87)80053-6.

139. Breiman, L. Random forests. *Mach Learn* **2001**, *45*, 5-32, doi:Doi 10.1023/A:1010933404324.

140. Cristianini, N.; Shawe-Taylor, J. *An introduction to Support Vector Machines : and other kernel-based learning methods*; Cambridge University Press: Cambridge, 2000; pp. xi, 189 p.

141. Zhou, C.Y.; Chen, Y.Q. Improving nearest neighbor classification with cam weighted distance. *Pattern Recogn* **2006**, *39*, 635-645, doi:10.1016/j.patcog.2005.09.004.

142. Hornik, K.; Stinchcombe, M.; White, H. Multilayer Feedforward Networks Are Universal Approximators. *Neural Networks* **1989**, *2*, 359-366, doi:Doi 10.1016/0893-6080(89)90020-8.

143. Dueck, D.; Frey, B.J. Non-metric affinity propagation for unsupervised image categorization. *Ieee I Conf Comp Vis* **2007**, 198-205.

144. Bezdek, J.C.; Ehrlich, R.; Full, W. Fcm - the Fuzzy C-Means Clustering-Algorithm. *Comput Geosci* **1984**, *10*, 191-203, doi:Doi 10.1016/0098-3004(84)90020-7.

145. Roberts, S.J.; Husmeier, D.; Rezek, I.; Penny, W. Bayesian approaches to Gaussian mixture modeling. *Ieee T Pattern Anal* **1998**, *20*, 1133-1142, doi:Doi 10.1109/34.730550.

146. Krishna, K.; Murty, M.N. Genetic K-means algorithm. *Ieee T Syst Man Cy B* **1999**, *29*, 433-439, doi:Doi 10.1109/3477.764879.

147. Matiisen, T. DEMYSTIFYING DEEP REINFORCEMENT LEARNING. Availabe online: https://neuro.cs.ut.ee/demystifying-deep-reinforcement-learning/ (accessed on 25.Jan.2021).

148. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436-444, doi:10.1038/nature14539.

149. Shen, D.; Wu, G.; Suk, H.I. Deep Learning in Medical Image Analysis. *Annu Rev Biomed Eng* **2017**, *19*, 221-248, doi:10.1146/annurev-bioeng-071516-044442.

150. Litjens, G.; Kooi, T.; Bejnordi, B.E.; Setio, A.A.A.; Ciompi, F.; Ghafoorian, M.; van der Laak, J.; van Ginneken, B.; Sanchez, C.I. A survey on deep learning in medical image analysis. *Med Image Anal* **2017**, *42*, 60-88, doi:10.1016/j.media.2017.07.005.

151. Chen, F.; Taviani, V.; Malkiel, I.; Cheng, J.Y.; Tamir, J.I.; Shaikh, J.; Chang, S.T.; Hardy, C.J.; Pauly, J.M.; Vasanawala, S.S. Variable-Density Single-Shot Fast Spin-Echo MRI with Deep Learning Reconstruction by Using Variational Networks. *Radiology* **2018**, *289*, 366-373, doi:10.1148/radiol.2018180445.

152. Eo, T.; Jun, Y.; Kim, T.; Jang, J.; Lee, H.J.; Hwang, D. KIKI-net: cross-domain convolutional neural networks for reconstructing undersampled magnetic resonance images. *Magn Reson Med* **2018**, *80*, 2188-2201, doi:10.1002/mrm.27201.

153. Knoll, F.; Hammernik, K.; Kobler, E.; Pock, T.; Recht, M.P.; Sodickson, D.K. Assessment of the generalization of learned image reconstruction and the potential for transfer learning. *Magn Reson Med* **2019**, *81*, 116-128, doi:10.1002/mrm.27355.

154. Schlemper, J.; Caballero, J.; Hajnal, J.V.; Price, A.N.; Rueckert, D. A Deep Cascade of Convolutional Neural Networks for Dynamic MR Image Reconstruction. *IEEE Trans Med Imaging* **2018**, *37*, 491-503, doi:10.1109/TMI.2017.2760978.

155. Zhu, B.; Liu, J.Z.; Cauley, S.F.; Rosen, B.R.; Rosen, M.S. Image reconstruction by domain-transform manifold learning. *Nature* **2018**, *555*, 487-+, doi:10.1038/nature25988.

156. Benou, A.; Veksler, R.; Friedman, A.; Riklin Raviv, T. Ensemble of expert deep neural networks for spatio-temporal denoising of contrast-enhanced MRI sequences. *Med Image Anal* **2017**, *42*, 145-159, doi:10.1016/j.media.2017.07.006.

157. Kustner, T.; Liebgott, A.; Mauch, L.; Martirosian, P.; Bamberg, F.; Nikolaou, K.; Yang, B.; Schick, F.; Gatidis, S. Automated reference-free detection of motion artifacts in magnetic resonance images. *MAGMA* **2018**, *31*, 243-256, doi:10.1007/s10334-017-0650-z.

158. Isensee, F.; Jaeger, P.F.; Kohl, S.A.A.; Petersen, J.; Maier-Hein, K.H. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods* **2020**, 10.1038/s41592-020-01008-z, doi:10.1038/s41592-020-01008-z.

159. Milletari, F.; Navab, N.; Ahmadi, S. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. *2016 Fourth International Conference on 3D Vision (3DV), Stanford, CA* **2016**, 10.1109/3DV.2016.79, 565-571, doi:10.1109/3DV.2016.79.

160. Wang, B.; Lei, Y.; Tian, S.; Wang, T.; Liu, Y.; Patel, P.; Jani, A.B.; Mao, H.; Curran, W.J.; Liu, T., et al. Deeply supervised 3D fully convolutional networks with group dilated convolution for automatic MRI prostate segmentation. *Med Phys* **2019**, *46*, 1707-1718, doi:10.1002/mp.13416.

161. Zavala-Romero, O.; Breto, A.L.; Xu, I.R.; Chang, Y.C.; Gautney, N.; Dal Pra, A.; Abramowitz, M.C.; Pollack, A.; Stoyanova, R. Segmentation of prostate and prostate zones using deep learning : A multi-MRI vendor analysis. *Strahlenther Onkol* **2020**, *196*, 932-942, doi:10.1007/s00066-020-01607-x.

162. Shao, W.; Banh, L.; Kunder, C.A.; Fan, R.E.; Soerensen, S.J.C.; Wang, J.B.; Teslovich, N.C.; Madhuripan, N.; Jawahar, A.; Ghanouni, P., et al. ProsRegNet: A deep learning framework for registration of MRI and histopathology images of the prostate. *Med Image Anal* **2021**, *68*, 101919, doi:10.1016/j.media.2020.101919.

163. Reda, I.; Khalil, A.; Elmogy, M.; Abou El-Fetouh, A.; Shalaby, A.; Abou El-Ghar, M.; Elmaghraby, A.; Ghazal, M.; El-Baz, A. Deep Learning Role in Early Diagnosis of Prostate Cancer. *Technol Cancer Res Treat* **2018**, *17*, 1533034618775530, doi:10.1177/1533034618775530.

164. Alkadi, R.; Taher, F.; El-Baz, A.; Werghi, N. A Deep Learning-Based Approach for the Detection and Localization of Prostate Cancer in T2 Magnetic Resonance Images. *J Digit Imaging* **2019**, *32*, 793-807, doi:10.1007/s10278-018-0160-1.

165. Padhani, A.R.; Turkbey, B. Detecting Prostate Cancer with Deep Learning for MRI: A Small Step Forward. *Radiology* **2019**, *293*, 618-619, doi:10.1148/radiol.2019192012.

166. Yoo, S.; Gujrathi, I.; Haider, M.A.; Khalvati, F. Prostate Cancer Detection using Deep Convolutional Neural Networks. *Sci Rep-Uk* **2019**, *9*, doi:10.1038/s41598-019-55972-4.

167. Wang, X.G.; Yang, W.; Weinreb, J.; Han, J.; Li, Q.B.; Kong, X.C.; Yan, Y.L.; Ke, Z.; Luo, B.; Liu, T., et al. Searching for prostate cancer by fully automated magnetic resonance imaging classification: deep learning versus non-deep learning. *Sci Rep-Uk* **2017**, *7*, doi:10.1038/s41598-017-15720-y.

168. Bulten, W.; Pinckaers, H.; van Boven, H.; Vink, R.; de Bel, T.; van Ginneken, B.; van der Laak, J.; Hulsbergen-van de Kaa, C.; Litjens, G. Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study. *Lancet Oncol* **2020**, *21*, 233-241, doi:10.1016/S1470-2045(19)30739-9.

169. Lucas, M.; Jansen, I.; Savci-Heijink, C.D.; Meijer, S.L.; de Boer, O.J.; van Leeuwen, T.G.; de Bruin, D.M.; Marquering, H.A. Deep learning for automatic Gleason pattern classification for grade group determination of prostate biopsies. *Virchows Arch* **2019**, *475*, 77-83, doi:10.1007/s00428-019-02577-x.

170. Schelb, P.; Kohl, S.; Radtke, J.P.; Wiesenfarth, M.; Kickingereder, P.; Bickelhaupt, S.; Kuder, T.A.; Stenzinger, A.; Hohenfellner, M.; Schlemmer, H.P., et al. Classification of Cancer at Prostate MRI: Deep Learning versus

Clinical PI-RADS Assessment. *Radiology* **2019**, *293*, 607-617, doi:10.1148/radiol.2019190938.

171. Wildeboer, R.R.; van Sloun, R.J.G.; Wijkstra, H.; Mischi, M. Artificial intelligence in multiparametric prostate cancer imaging with focus on deep-learning methods. *Comput Methods Programs Biomed* **2020**, *189*, 105316, doi:10.1016/j.cmpb.2020.105316.
172. Soffer, S.; Ben-Cohen, A.; Shimon, O.; Amitai, M.M.; Greenspan, H.; Klang, E. Convolutional Neural Networks for Radiologic Images: A Radiologist's Guide. *Radiology* **2019**, *290*, 590-606, doi:10.1148/radiol.2018180547.
173. Yasaka, K.; Akai, H.; Kunimatsu, A.; Kiryu, S.; Abe, O. Deep learning with convolutional neural network in radiology. *Jpn J Radiol* **2018**, *36*, 257-272, doi:10.1007/s11604-018-0726-3.
174. Castellino, R.A. Computer aided detection (CAD): an overview. *Cancer Imaging* **2005**, *5*, 17-19, doi:10.1102/1470-7330.2005.0018.
175. Greer, M.D.; Lay, N.; Shih, J.H.; Barrett, T.; Bittencourt, L.K.; Borofsky, S.; Kabakus, I.; Law, Y.M.; Marko, J.; Shebel, H., et al. Computer-aided diagnosis prior to conventional interpretation of prostate mpMRI: an international multi-reader study. *Eur Radiol* **2018**, *28*, 4407-4417, doi:10.1007/s00330-018-5374-6.
176. Karssemeijer, N.; Otten, J.D.; Rijken, H.; Holland, R. Computer aided detection of masses in mammograms as decision support. *Br J Radiol* **2006**, *79 Spec No 2*, S123-126, doi:10.1259/bjr/37622515.
177. Summers, R.M.; Liu, J.; Rehani, B.; Stafford, P.; Brown, L.; Louie, A.; Barlow, D.S.; Jensen, D.W.; Cash, B.; Choi, J.R., et al. CT colonography computer-aided polyp detection: Effect on radiologist observers of polyp identification by CAD on both the supine and prone scans. *Acad Radiol* **2010**, *17*, 948-959, doi:10.1016/j.acra.2010.03.024.
178. Firmino, M.; Angelo, G.; Morais, H.; Dantas, M.R.; Valentim, R. Computer-aided detection (CADe) and diagnosis (CADx) system for lung cancer with likelihood of malignancy. *Biomed Eng Online* **2016**, *15*, 2, doi:10.1186/s12938-015-0120-7.
179. Campa, R.; Del Monte, M.; Barchetti, G.; Pecoraro, M.; Salvo, V.; Ceravolo, I.; Indino, E.L.; Ciardi, A.; Catalano, C.; Panebianco, V. Improvement of prostate cancer detection combining a computer-aided diagnostic system with TRUS-MRI targeted biopsy. *Abdom Radiol (NY)* **2019**, *44*, 264-271, doi:10.1007/s00261-018-1712-z.
180. Liu, P.; Wang, S.J.; Turkbey, B.; Grant, K.; Pinto, P.; Choyke, P.; Wood, B.J.; Summers, R.M. A prostate cancer computer-aided diagnosis system using multimodal magnetic resonance imaging and targeted biopsy labels. *Proc Spie* **2013**, *8670*, doi:10.1117/12.2007927.
181. Niaf, E.; Rouviere, O.; Mege-Lechevallier, F.; Bratan, F.; Lartizien, C. Computer-aided diagnosis of prostate cancer in the peripheral zone using multiparametric MRI. *Phys Med Biol* **2012**, *57*, 3833-3851, doi:10.1088/0031-9155/57/12/3833.
182. Rampun, A.; Zheng, L.; Malcolm, P.; Tiddeman, B.; Zwiggelaar, R. Computer-aided detection of prostate cancer in T2-weighted MRI within the peripheral zone. *Phys Med Biol* **2016**, *61*, 4796-4825, doi:10.1088/0031-9155/61/13/4796.

183. Vos, E.K.; Litjens, G.J.; Kobus, T.; Hambrock, T.; Hulsbergen-van de Kaa, C.A.; Barentsz, J.O.; Huisman, H.J.; Scheenen, T.W. Assessment of prostate cancer aggressiveness using dynamic contrast-enhanced magnetic resonance imaging at 3 T. *Eur Urol* **2013**, *64*, 448-455, doi:10.1016/j.eururo.2013.05.045.

184. Vos, P.C.; Barentsz, J.O.; Karssemeijer, N.; Huisman, H.J. Automatic computer-aided detection of prostate cancer based on multiparametric magnetic resonance image analysis. *Phys Med Biol* **2012**, *57*, 1527-1542, doi:10.1088/0031-9155/57/6/1527.

185. Styner, M.; Brechbuhler, C.; Szekely, G.; Gerig, G. Parametric estimate of intensity inhomogeneities applied to MRI. *IEEE Trans Med Imaging* **2000**, *19*, 153-165, doi:10.1109/42.845174.

186. Loizou, C.P.; Pantziaris, M.; Seimenis, I.; Pattichis, C.S. Brain MR Image Normalization in Texture Analysis of Multiple Sclerosis. *2009 9th International Conference on Information Technology and Applications in Biomedicine* **2009**, 131-+.

187. Madabhushi, A.; Udupa, J.K. New methods of MR image intensity standardization via generalized scale. *Med Phys* **2006**, *33*, 3426-3434, doi:10.1118/1.2335487.

188. Nyul, L.G.; Udupa, J.K.; Zhang, X. New variants of a method of MRI scale standardization. *Ieee T Med Imaging* **2000**, *19*, 143-150, doi:Doi 10.1109/42.836373.

189. Simmons, A.; Tofts, P.S.; Barker, G.J.; Arridge, S.R. Sources of Intensity Nonuniformity in Spin-Echo Images at 1.5-T. *Magn Reson Med* **1994**, *32*, 121-128, doi:DOI 10.1002/mrm.1910320117.

190. Ahdoot, M.; Wilbur, A.R.; Reese, S.E.; Lebastchi, A.H.; Mehralivand, S.; Gomella, P.T.; Bloom, J.; Gurram, S.; Siddiqui, M.; Pinsky, P., et al. MRI-Targeted, Systematic, and Combined Biopsy for Prostate Cancer Diagnosis. *N Engl J Med* **2020**, *382*, 917-928, doi:10.1056/NEJMoa1910038.

191. Lemaitre, G.; Rastgoo, M.; Massich, J.; Vilanova, J.C.; Walker, P.M.; Freixenet, J.; Meyer-Baese, A.; Meriaudeau, F.; Marti, R. Normalization of T2W-MRI Prostate Images using Rician a priori. *Medical Imaging 2016: Computer-Aided Diagnosis* **2015**, *9785*, doi:10.1117/12.2216072.

192. Schwier, M.; van Griethuysen, J.; Vangel, M.G.; Pieper, S.; Peled, S.; Tempany, C.; Aerts, H.; Kikinis, R.; Fennessy, F.M.; Fedorov, A. Repeatability of Multiparametric Prostate MRI Radiomics Features. *Sci Rep* **2019**, *9*, 9441, doi:10.1038/s41598-019-45766-z.

193. Ge, Y.L.; Udupa, J.K.; Nyul, L.G.; Wei, L.G.; Grossman, R.I. Numerical tissue characterization in MS via standardization of the MR image intensity scale. *Journal of Magnetic Resonance Imaging* **2000**, *12*, 715-721, doi:Doi 10.1002/1522-2586(200011)12:5<715::Aid-Jmri8>3.0.Co;2-D.

194. Peng, Y.H.; Jiang, Y.L.; Oto, A. Reference-Tissue Correction of T-2-weighted Signal Intensity for Prostate Cancer Detection. *Medical Imaging 2014: Computer-Aided Diagnosis* **2014**, *9035*, doi:10.1117/12.2043585.

195. Stoilescu, L.; Maas, M.C.; Huisman, H.J. Feasibility of multireference tissue normalization of T2-weighted prostate MRI. *ESMRMB Annual Scientific Meeting, Barcelona* **2017**.

196. Salembier, C.; Villeirs, G.; De Bari, B.; Hoskin, P.; Pieters, B.R.; Van Vulpen, M.; Khoo, V.; Henry, A.; Bossi, A.; De Meerleer, G., et al. ESTRO ACROP consensus guideline on CT- and MRI-based target volume delineation for primary radiation therapy of localized prostate cancer. *Radiother Oncol* **2018**, *127*, 49-61, doi:10.1016/j.radonc.2018.01.014.

197. Khan, Z.; Yahya, N.; Alsaih, K.; Ali, S.S.A.; Meriaudeau, F. Evaluation of Deep Neural Networks for Semantic Segmentation of Prostate in T2W MRI. *Sensors (Basel)* **2020**, *20*, doi:10.3390/s20113183.

198. Song, G.L.; Han, J.D.; Zhao, Y.W.; Wang, Z.; Du, H.B. A Review on Medical Image Registration as an Optimization Problem. *Curr Med Imaging* **2017**, *13*, 274-283, doi:10.2174/1573405612666160920123955.

199. Avanzo, M.; Wei, L.S.; Stancanello, J.; Vallieres, M.; Rao, A.; Morin, O.; Mattonen, S.A.; El Naqa, I. Machine and deep learning methods for radiomics. *Medical Physics* **2020**, *47*, E185-E202, doi:10.1002/mp.13678.

200. Saeys, Y.; Inza, I.; Larranaga, P. A review of feature selection techniques in bioinformatics. *Bioinformatics* **2007**, *23*, 2507-2517, doi:10.1093/bioinformatics/btm344.

201. Poernomo, A.; Kang, D.K. Biased Dropout and Crossmap Dropout: Learning towards effective Dropout regularization in convolutional neural network. *Neural Networks* **2018**, *104*, 60-67, doi:10.1016/j.neunet.2018.03.016.

202. Litjens, G.; Toth, R.; van de Ven, W.; Hoeks, C.; Kerkstra, S.; van Ginneken, B.; Vincent, G.; Guillard, G.; Birbeck, N.; Zhang, J., et al. Evaluation of prostate segmentation algorithms for MRI: the PROMISE12 challenge. *Med Image Anal* **2014**, *18*, 359-373, doi:10.1016/j.media.2013.12.002.

203. Armato, S.G., 3rd; Huisman, H.; Drukker, K.; Hadjiiski, L.; Kirby, J.S.; Petrick, N.; Redmond, G.; Giger, M.L.; Cha, K.; Mamonov, A., et al. PROSTATEx Challenges for computerized classification of prostate lesions from multiparametric magnetic resonance images. *J Med Imaging (Bellingham)* **2018**, *5*, 044501, doi:10.1117/1.JMI.5.4.044501.

204. Tustison, N.J.; Avants, B.B.; Cook, P.A.; Zheng, Y.; Egan, A.; Yushkevich, P.A.; Gee, J.C. N4ITK: improved N3 bias correction. *IEEE Trans Med Imaging* **2010**, *29*, 1310-1320, doi:10.1109/TMI.2010.2046908.

205. Bruns, N. 3D Slicer Universal 3D-Visualization-Software. *Unfallchirurg* **2019**, *122*, 662-663, doi:10.1007/s00113-019-0654-4.

206. Yushkevich, P.A.; Piven, J.; Hazlett, H.C.; Smith, R.G.; Ho, S.; Gee, J.C.; Gerig, G. User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. *Neuroimage* **2006**, *31*, 1116-1128, doi:10.1016/j.neuroimage.2006.01.015.

207. Dollar, P.; Appel, R.; Belongie, S.; Perona, P. Fast Feature Pyramids for Object Detection. *Ieee T Pattern Anal* **2014**, *36*, 1532-1545, doi:10.1109/Tpami.2014.2300479.

208. Otsu, N. Threshold Selection Method from Gray-Level Histograms. *Ieee T Syst Man Cyb* **1979**, *9*, 62-66, doi:Doi 10.1109/Tsmc.1979.4310076.

209. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. *Lect Notes Comput Sc* **2015**, *9351*, 234-241, doi:10.1007/978-3-319-24574-4_28.

210. Mirzaev, I. Fully Convolutional Neural Network with Residual Connections for Automatic Segmentation of Prostate Structures from MR Images. Availabe online: https://grand-challenge-public.s3.amazonaws.com/evaluation-supplementary/40/d70ba7d1-bc95-439e-a81e-7f1a4ed5fda0/18_MBIOS.pdf (accessed on 11.Feb.2021).
211. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.M.; Gimelshein, N.; Antiga, L., et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Adv Neur In* **2019**, *32*.
212. Bojorquez, J.Z.; Bricq, S.; Brunotte, F.; Walker, P.M.; Lalande, A. A novel alternative to classify tissues from T 1 and T 2 relaxation times for prostate MRI. *MAGMA* **2016**, *29*, 777-788, doi:10.1007/s10334-016-0562-3.
213. Tolles, J.; Meurer, W.J. Logistic Regression Relating Patient Characteristics to Outcomes. *Jama-J Am Med Assoc* **2016**, *316*, 533-534, doi:10.1001/jama.2016.7653.
214. Klein, S.; van der Heide, U.A.; Lips, I.M.; van Vulpen, M.; Staring, M.; Pluim, J.P. Automatic segmentation of the prostate in 3D MR images by atlas matching using localized mutual information. *Med Phys* **2008**, *35*, 1407-1417, doi:10.1118/1.2842076.
215. Heimann, T.; van Ginneken, B.; Styner, M.A.; Arzhaeva, Y.; Aurich, V.; Bauer, C.; Beck, A.; Becker, C.; Beichel, R.; Bekes, G., et al. Comparison and evaluation of methods for liver segmentation from CT datasets. *IEEE Trans Med Imaging* **2009**, *28*, 1251-1265, doi:10.1109/TMI.2009.2013851.
216. Heimann, T.; van Ginneken, B.; Styner, M.A.; Arzhaeva, Y.; Aurich, V.; Bauer, C.; Beck, A.; Becker, C.; Beichel, R.; Bekes, G., et al. Comparison and Evaluation of Methods for Liver Segmentation From CT Datasets. *Ieee T Med Imaging* **2009**, *28*, 1251-1265, doi:10.1109/Tmi.2009.2013851.
217. Chandra, S.S.; Dowling, J.A.; Shen, K.K.; Raniga, P.; Pluim, J.P.W.; Greer, P.B.; Salvado, O.; Fripp, J. Patient Specific Prostate Segmentation in 3-D Magnetic Resonance Images. *Ieee T Med Imaging* **2012**, *31*, 1955-1964, doi:10.1109/Tmi.2012.2211377.
218. Friedman, J.; Hastie, T.; Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw* **2010**, *33*, 1-22, doi:DOI 10.18637/jss.v033.i01.
219. Gibbons, J.D.; Chakraborti, S. *Nonparametric statistical inference*, 5th ed.; Taylor & Francis: Boca Raton, 2011; pp. xx, 630 p.
220. Kalpić, D.; Hlupić, N.; Lovrić, M. Student's t-Tests. In *International Encyclopedia of Statistical Science*, Lovric, M., Ed. Springer Berlin Heidelberg: Berlin, Heidelberg, 2011; 10.1007/978-3-642-04898-2_641pp. 1559-1563.
221. Hanley, J.A.; McNeil, B.J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **1982**, *143*, 29-36, doi:10.1148/radiology.143.1.7063747.
222. Delong, E.R.; Delong, D.M.; Clarkepearson, D.I. Comparing the Areas under 2 or More Correlated Receiver Operating Characteristic Curves - a Nonparametric Approach. *Biometrics* **1988**, *44*, 837-845, doi:Doi 10.2307/2531595.
223. Benjamini, Y.; Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical*

Society: Series B (Methodological) **1995**, *57*, 289-300, doi:https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

224. Bland, J.M.; Altman, D.G. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* **1986**, *1*, 307-310.
225. Swain, M.J.; Ballard, D.H. Color Indexing. *Int J Comput Vision* **1991**, *7*, 11-32, doi:Doi 10.1007/Bf00130487.
226. Willmott, C.J.; Matsuura, K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Res* **2005**, *30*, 79-82, doi:DOI 10.3354/cr030079.
227. McGraw, K.O.; Wong, S.P. Forming inferences about some intraclass correlation coefficients. *Psychological Methods* **1996**, *1*, 30-46, doi:10.1037/1082-989X.1.1.30.
228. Shrout, P.E.; Fleiss, J.L. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull* **1979**, *86*, 420-428, doi:10.1037//0033-2909.86.2.420.
229. Kohestani, K.; Wallstrom, J.; Dehlfors, N.; Sponga, O.M.; Mansson, M.; Josefsson, A.; Carlsson, S.; Hellstrom, M.; Hugosson, J. Performance and inter-observer variability of prostate MRI (PI-RADS version 2) outside high-volume centres. *Scand J Urol* **2019**, 10.1080/21681805.2019.1675757, doi:10.1080/21681805.2019.1675757.
230. Sonn, G.A.; Fan, R.E.; Ghanouni, P.; Wang, N.N.; Brooks, J.D.; Loening, A.M.; Daniel, B.L.; To'o, K.J.; Thong, A.E.; Leppert, J.T. Prostate Magnetic Resonance Imaging Interpretation Varies Substantially Across Radiologists. *Eur Urol Focus* **2019**, *5*, 592-599, doi:10.1016/j.euf.2017.11.010.
231. Neri, E.; deSouza, N.; Brady, A.; Esr. What the radiologist should know about artificial intelligence - an ESR white paper. *Insights into Imaging* **2019**, *10*, doi:10.1186/s13244-019-0738-2.
232. Hambrock, T.; Vos, P.C.; Hulsbergen-van de Kaa, C.A.; Barentsz, J.O.; Huisman, H.J. Prostate Cancer: Computer-aided Diagnosis with Multiparametric 3-T MR Imaging-Effect on Observer Performance. *Radiology* **2013**, *266*, 521-530, doi:10.1148/radiol.12111634.
233. Chan, I.; Wells, W.; Mulkern, R.V.; Haker, S.; Zhang, J.Q.; Zou, K.H.; Maier, S.E.; Tempany, C.M.C. Detection of prostate cancer by integration of line-scan diffusion, T2-mapping and T2-weighted magnetic resonance imaging; a multichannel statistical classifier. *Medical Physics* **2003**, *30*, 2390-2398, doi:10.1118/1.1593633.
234. Shah, V.; Turkbey, B.; Mani, H.; Pang, Y.X.; Pohida, T.; Merino, M.J.; Pinto, P.A.; Choyke, P.L.; Bernardo, M. Decision support system for localizing prostate cancer based on multiparametric magnetic resonance imaging. *Medical Physics* **2012**, *39*, 4093-4103, doi:10.1118/1.4722753.
235. Peng, Y.H.; Jiang, Y.L.; Yang, C.; Brown, J.B.; Antic, T.; Sethi, I.; Schmid-Tannwald, C.; Giger, M.L.; Eggener, S.E.; Oto, A. Quantitative Analysis of Multiparametric Prostate MR Images: Differentiation between Prostate Cancer and Normal Tissue and Correlation with Gleason Score-A Computer-aided Diagnosis Development Study. *Radiology* **2013**, *267*, 787-796, doi:10.1148/radiol.13121454.

236. Ishioka, J.; Matsuoka, Y.; Uehara, S.; Yasuda, Y.; Kijima, T.; Yoshida, S.; Yokoyama, M.; Saito, K.; Kihara, K.; Numao, N., et al. Computer-aided diagnosis of prostate cancer on magnetic resonance imaging using a convolutional neural network algorithm. *Bju International* **2018**, *122*, 411-417, doi:10.1111/bju.14397.

237. Song, Y.; Zhang, Y.D.; Yan, X.; Liu, H.; Zhou, M.X.; Hu, B.W.; Yang, G. Computer-aided diagnosis of prostate cancer using a deep convolutional neural network from multiparametric MRI. *Journal of Magnetic Resonance Imaging* **2018**, *48*, 1570-1577, doi:10.1002/jmri.26047.

238. Abraham, B.; Nair, M.S. Computer-aided classification of prostate cancer grade groups from MRI images using texture features and stacked sparse autoencoder. *Comput Med Imag Grap* **2018**, *69*, 60-68, doi:10.1016/j.compmedimag.2018.08.006.

239. de Vente, C.; Vos, P.; Hosseinzadeh, M.; Pluim, J.; Veta, M. Deep Learning Regression for Prostate Cancer Detection and Grading in Bi-Parametric MRI. *Ieee T Bio-Med Eng* **2021**, *68*, 374-383, doi:10.1109/Tbme.2020.2993528.

240. van Ginneken, B.; Schaefer-Prokop, C.M.; Prokop, M. Computer-aided Diagnosis: How to Move from the Laboratory to the Clinic. *Radiology* **2011**, *261*, 719-732, doi:10.1148/radiol.11091710.

241. Yanase, J.; Triantaphyllou, E. The seven key challenges for the future of computer-aided diagnosis in medicine. *Int J Med Inform* **2019**, *129*, 413-422, doi:10.1016/j.ijmedinf.2019.06.017.

242. Lynch, C.J.; Liston, C. New machine-learning technologies for computer-aided diagnosis. *Nat Med* **2018**, *24*, 1304-1305, doi:10.1038/s41591-018-0178-4.

243. Li, Q. Improvement of bias and generalizability for computer-aided diagnostic schemes. *Comput Med Imag Grap* **2007**, *31*, 338-345, doi:10.1016/j.compmedimag.2007.02.004.

244. Supriya, M.D., A.J. Machine learning approach on healthcare big data: a review. *Big Data and Information Analytics* **2020**, *5*, 58-75, doi:10.3934/bdia.2020005.

245. Lucieri, A.B., M.N.; Dengel, A.; Ahmed, S. Achievements and Challenges in Explaining Deep Learning based Computer-Aided Diagnosis Systems. *arXiv preprint* **2020**, *arXiv:2011.13169*.

246. Singh, A.; Sengupta, S.; Lakshminarayanan, V. Explainable Deep Learning Models in Medical Image Analysis. *J Imaging* **2020**, *6*, doi:10.3390/jimaging6060052.

247. Dikaios, N.; Alkalbani, J.; Abd-Alazeez, M.; Sidhu, H.S.; Kirkham, A.; Ahmed, H.U.; Emberton, M.; Freeman, A.; Halligan, S.; Taylor, S., et al. Zone-specific logistic regression models improve classification of prostate cancer on multi-parametric MRI. *Eur Radiol* **2015**, *25*, 2727-2737, doi:10.1007/s00330-015-3636-0.

248. Price, W.N. Big data and black-box medical algorithms. *Sci Transl Med* **2018**, *10*, doi:10.1126/scitranslmed.aao5333.

249. Malich, A.; Azhari, T.; Bohm, T.; Fleck, M.; Kaiser, W.A. Reproducibility - an important factor determining the quality of computer aided detection (CAD) systems. *European Journal of Radiology* **2000**, *36*, 170-174, doi:Doi 10.1016/S0720-048x(00)00189-3.

250. Lu, H.; Parra, N.A.; Qi, J.; Gage, K.; Li, Q.; Fan, S.X.; Feuerlein, S.; Pow-Sang, J.; Gillies, R.; Choi, J.W., et al. Repeatability of Quantitative Imaging Features in Prostate Magnetic Resonance Imaging. *Front Oncol* **2020**, *10*, doi:10.3389/fonc.2020.00551.

251. Sun, X.F.; Shi, L.; Luo, Y.S.; Yang, W.; Li, H.P.; Liang, P.P.; Li, K.C.; Mok, V.C.T.; Chu, W.C.W.; Wang, D.F. Histogram-based normalization technique on human brain magnetic resonance images from different acquisitions. *Biomedical Engineering Online* **2015**, *14*, doi:10.1186/s12938-015-0064-y.

252. Leung, K.K.; Clarkson, M.J.; Bartlett, J.W.; Clegg, S.; Jack, C.R.; Weiner, M.W.; Fox, N.C.; Ourselin, S.; Initi, A.s.D.N. Robust atrophy rate measurement in Alzheimer's disease using multi-site serial MRI: Tissue-specific intensity normalization and parameter selection. *Neuroimage* **2010**, *50*, 516-523, doi:10.1016/j.neuroimage.2009.12.059.

253. Niaf, E.; Rouviere, O.; Lartizien, C. Computer-Aided Diagnosis for prostate cancer detection in the peripheral zone via multisequence MRI. *Medical Imaging 2011: Computer-Aided Diagnosis* **2011**, *7963*, doi:10.1117/12.877231.

254. Dikaios, N.; Alkalbani, J.; Sidhu, H.S.; Fujiwara, T.; Abd-Alazeez, M.; Kirkham, A.; Allen, C.; Ahmed, H.; Emberton, M.; Freeman, A., et al. Logistic regression model for diagnosis of transition zone prostate cancer on multi-parametric MRI. *Eur Radiol* **2015**, *25*, 523-532, doi:10.1007/s00330-014-3386-4.

255. Engelhard, K.; Hollenbach, H.P.; Deimling, M.; Kreckel, M.; Riedl, C. Combination of signal intensity measurements of lesions in the peripheral zone of prostate with MRI and serum PSA level for differentiating benign disease from prostate cancer. *Eur Radiol* **2000**, *10*, 1947-1953, doi:10.1007/s003300000524.

256. Sunoqrot, M.R.S.S., K. M.; Zavala-Romero, O.; Stoyanova, R. ; Bathen, T. F.; Elschot, M. A quality control system for automated prostate segmentation on T2-weighted MRI. In Proceedings of International Society for Magnetic Resonance in Medicine 27th Annual Meeting, Montréal.

257. Patsanis, A.S., M. R. S.; Sandsmark, E.; Langørgen, S.; Bertilsson, H.; Wang, H.; Bathen , T. F.; Elschot, M. Prostate Cancer Detection on T2-weighted MR images with Generative Adversarial Networks. In Proceedings of International Society for Magnetic Resonance in Medicine 29th Annual Meeting.

258. Dewi, D.E.O.S., M. R. S.; Nketiah, G. A.; Sandsmark, E.; Langørgen, S.; Bertilsson, H.; Elschot, M.; Bathen , T. F. Repeatability of Radiomic Features in T2-Weighted Prostate MRI: Impact of Pre-processing Configurations In Proceedings of International Society for Magnetic Resonance in Medicine 29th Annual Meeting.

259. Nketiah, G.A.; Elschot, M.; Scheenen, T.W.; Maas, M.C.; Bathen, T.F.; Selnaes, K.M.; Consortium, P.C.-M. Utility of T2-weighted MRI texture analysis in assessment of peripheral zone prostate cancer aggressiveness: a single-arm, multicenter study. *Sci Rep* **2021**, *11*, 2085, doi:10.1038/s41598-021-81272-x.

260. Sørland, K.S., M. R. S.; Goa, P. A.; Sandsmark, E.; Langørgen, S.; Bertilsson, H.; Lin, G.; Bathen, T. F.; Elschot, M. Automated reference tissue normalization of prostate T2-weighted MRI on a large, multicenter dataset. In Proceedings of International Society for Magnetic Resonance in Medicine 29th Annual Meeting

261. Sørland, K.G., P. A.; Selnæs, K. M.; Sandsmark, E.; Trimble, C. G.; Bertilsson, H.; Lin, G.; Sunoqrot, M. R. S.; Elschot, M.; Bathen, T. F. Pseudo-T2 mapping of T2-weighted MRI- of the prostate: Comparison to gold standard. In Proceedings of International Society for Magnetic Resonance in Medicine 29th Annual Meeting.

262. Engels, R.R.M.; Israel, B.; Padhani, A.R.; Barentsz, J.O. Multiparametric Magnetic Resonance Imaging for the Detection of Clinically Significant Prostate Cancer: What Urologists Need to Know. Part 1: Acquisition. *European Urology* **2020**, *77*, 457-468, doi:10.1016/j.eururo.2019.09.021.

263. Renard, F.; Guedria, S.; De Palma, N.; Vuillerme, N. Variability and reproducibility in deep learning for medical image segmentation. *Sci Rep-Uk* **2020**, *10*, doi:10.1038/s41598-020-69920-0.

264. Klein, S.; Staring, M.; Murphy, K.; Viergever, M.A.; Pluim, J.P.W. elastix: A Toolbox for Intensity-Based Medical Image Registration. *Ieee T Med Imaging* **2010**, *29*, 196-205, doi:10.1109/Tmi.2009.2035616.

265. Mazaheri, Y.; Bokacheva, L.; Kroon, D.J.; Akin, O.; Hricak, H.; Chamudot, D.; Fine, S.; Koutcher, J.A. Semi-automatic Deformable Registration of Prostate Mr Images to Pathological Slices. *Journal of Magnetic Resonance Imaging* **2010**, *32*, 1149-1157, doi:10.1002/jmri.22347.

266. Fu, Y.B.; Lei, Y.; Wang, T.H.; Curran, W.J.; Liu, T.; Yang, X.F. Deep learning in medical image registration: a review. *Phys Med Biol* **2020**, *65*, doi:10.1088/1361-6560/ab843e.

267. Hu, Y.P.; Modat, M.; Gibson, E.; Li, W.Q.; Ghavamia, N.; Bonmati, E.; Wang, G.T.; Bandula, S.; Moore, C.M.; Emberton, M., et al. Weakly-supervised convolutional neural networks for multimodal image registration. *Medical Image Analysis* **2018**, *49*, 1-13, doi:10.1016/j.media.2018.07.002.

268. Yan, P.; Xu, S.; Rastinehad, A.R.; Wood, B.J. Adversarial Image Registration with Application for MR and TRUS Image Fusion. 2018; p arXiv:1804.11024.

269. Vial, A.; Stirling, D.; Field, M.; Ros, M.; Ritz, C.; Carolan, M.; Holloway, L.; Miller, A.A. The role of deep learning and radiomic feature extraction in cancer-specific predictive modelling: a review. *Translational Cancer Research* **2018**, *7*, 803-816, doi:10.21037/tcr.2018.05.02.

270. Zhang, Z.Q.; Zhao, Y.; Liao, X.K.; Shi, W.Q.; Li, K.L.; Zou, Q.; Peng, S.L. Deep learning in omics: a survey and guideline. *Brief Funct Genomics* **2019**, *18*, 41-57, doi:10.1093/bfgp/ely030.

271. Zurowietz, M.; Nattkemper, T.W. An Interactive Visualization for Feature Localization in Deep Neural Networks. *Frontiers in Artificial Intelligence* **2020**, *3*, doi:10.3389/frai.2020.00049.

272. Zeiler, M.D.; Fergus, R. Visualizing and Understanding Convolutional Networks. *Lect Notes Comput Sc* **2014**, *8689*, 818-833, doi:Doi 10.1007/978-3-319-10590-1_53.

273. Carre, A.; Klausner, G.; Edjlali, M.; Lerousseau, M.; Briend-Diop, J.; Sun, R.; Ammari, S.; Reuze, S.; Andres, E.A.; Estienne, T., et al. Standardization of brain MR images across machines and protocols: bridging the gap for MRI-based radiomics. *Sci Rep-Uk* **2020**, *10*, doi:10.1038/s41598-020-69298-z.

274. Cattell, R.; Chen, S.; Huang, C. Robustness of radiomic features in magnetic resonance imaging: review and a phantom study. *Vis Comput Ind Biomed Art* **2019**, *2*, 19, doi:10.1186/s42492-019-0025-6.

275. Scalco, E.; Belfatto, A.; Mastropietro, A.; Rancati, T.; Avuzzi, B.; Messina, A.; Valdagni, R.; Rizzo, G. T2w-MRI signal normalization affects radiomics features reproducibility. *Med Phys* **2020**, *47*, 1680-1691, doi:10.1002/mp.14038.

276. Bardis, M.D.; Houshyar, R.; Chang, P.D.; Ushinsky, A.; Glavis-Bloom, J.; Chahine, C.; Bui, T.L.; Rupasinghe, M.; Filippi, C.G.; Chow, D.S. Applications of Artificial Intelligence to Prostate Multiparametric MRI (mpMRI): Current and Emerging Trends. *Cancers* **2020**, *12*, doi:10.3390/cancers12051204.

277. Swiderska-Chadaj, Z.; de Bel, T.; Blanchet, L.; Baidoshvili, A.; Vossen, D.; van der Laak, J.; Litjens, G. Impact of rescanning and normalization on convolutional neural network performance in multi-center, whole-slide classification of prostate cancer. *Sci Rep-Uk* **2020**, *10*, doi:10.1038/s41598-020-71420-0.

278. Gao, Y.K., N.;Mas, J.F.; Navarrete, A.; Niemeyer, I. Optimized image segmentation and its effect on classification accuracy. In Proceedings of The 5th International symposium on Spatial Data Quality, SDQ 2007: Modelling qualities in space and time, IT, ITC, Enschede, The Netherlands, 13-15 June, 2007.

279. Lanka, P.; Rangaprakash, D.; Dretsch, M.N.; Katz, J.S.; Denney, T.S.; Deshpande, G. Supervised machine learning for diagnostic classification from large-scale neuroimaging datasets. *Brain Imaging Behav* **2020**, *14*, 2378-2416, doi:10.1007/s11682-019-00191-8.

280. Pham, H.N.A.; Triantaphyllou, E. The Impact of Overfitting and Overgeneralization on the Classification Accuracy in Data Mining. In *Soft Computing for Knowledge Discovery and Data Mining*, Maimon, O., Rokach, L., Eds. Springer US: Boston, MA, 2008; 10.1007/978-0-387-69935-6_16pp. 391-431.

281. Collmann, S.; Lin, J.S.; Freedman, M.T.; Wu, C.; Hayes, W.; Mun, S.K. A design-based approach to ethics in computer-aided diagnosis. *P Soc Photo-Opt Ins* **1996**, *2707*, 610-617.

282. Prior, F.; Almeida, J.; Kathiravelu, P.; Kurc, T.; Smith, K.; Fitzgerald, T.J.; Saltz, J. Open access image repositories: high-quality data to enable machine learning research. *Clin Radiol* **2020**, *75*, 7-12, doi:10.1016/j.crad.2019.04.002.

283. Xu, Y.; Goodacre, R. On Splitting Training and Validation Set: A Comparative Study of Cross-Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning. *J Anal Test* **2018**, *2*, 249-262, doi:10.1007/s41664-018-0068-2.

284. Kairouz, P.; McMahan, H.B.; Avent, B.; Bellet, A.; Bennis, M.; Nitin Bhagoji, A.; Bonawitz, K.; Charles, Z.; Cormode, G.; Cummings, R., et al. Advances and Open Problems in Federated Learning. 2019; p arXiv:1912.04977.

285. Lee, G.H.; Shin, S.Y. Federated Learning on Clinical Benchmark Data: Performance Assessment. *J Med Internet Res* **2020**, *22*, doi:10.2196/20891.

286. Willemink, M.J.; Koszek, W.A.; Hardell, C.; Wu, J.; Fleischmann, D.; Harvey, H.; Folio, L.R.; Summers, R.M.; Rubin, D.L.; Lungren, M.P. Preparing Medical

Imaging Data for Machine Learning. *Radiology* **2020**, *295*, 4-15, doi:10.1148/radiol.2020192224.

287. Larson, D.B.; Magnus, D.C.; Lungren, M.P.; Shah, N.H.; Langlotz, C.P. Ethics of Using and Sharing Clinical Imaging Data for Artificial Intelligence: A Proposed Framework. *Radiology* **2020**, *295*, 675-682, doi:10.1148/radiol.2020192536.

288. Aryanto, K.Y.E.; Oudkerk, M.; van Ooijen, P.M.A. Free DICOM de-identification tools in clinical research: functioning and safety of patient privacy. *European Radiology* **2015**, *25*, 3685-3695, doi:10.1007/s00330-015-3794-0.

289. Moore, S.M.; Maffitt, D.R.; Smith, K.E.; Kirby, J.S.; Clark, K.W.; Freymann, J.B.; Vendt, B.A.; Tarbox, L.R.; Prior, F.W. De-identification of Medical Images with Retention of Scientific Research Value. *Radiographics* **2015**, *35*, 727-735, doi:10.1148/rg.2015140244.

290. Regulation (EU) 2016/679 of the European Parliament and of the Council. The protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (Availabe online: https://eur-lex.europa.eu/eli/reg/2016/679/oj (accessed on 08.April.2021).

291. Norwegian University of Science and Technology. HUNT Cloud. Availabe online: https://www.ntnu.edu/mh/huntcloud (accessed on 08.April.2021).

292. Wilkinson, M.D.; Dumontier, M.; Aalbersberg, I.J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.W.; da Silva Santos, L.B.; Bourne, P.E., et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **2016**, *3*, 160018, doi:10.1038/sdata.2016.18.

Paper I

Magnetic Resonance Materials in Physics, Biology and Medicine (2021) 34:309–321
https://doi.org/10.1007/s10334-020-00871-3

RESEARCH ARTICLE

Automated reference tissue normalization of T2-weighted MR images of the prostate using object recognition

Mohammed R. S. Sunoqrot[1] · Gabriel A. Nketiah[1,2] · Kirsten M. Selnæs[1,2] · Tone F. Bathen[1,2] · Mattijs Elschot[1,2]

Received: 8 April 2020 / Revised: 2 July 2020 / Accepted: 21 July 2020 / Published online: 31 July 2020
© The Author(s) 2020

Abstract

Objectives To develop and evaluate an automated method for prostate T2-weighted (T2W) image normalization using dual-reference (fat and muscle) tissue.

Materials and methods Transverse T2W images from the publicly available PROMISE12 ($N=80$) and PROSTATEx ($N=202$) challenge datasets, and an in-house collected dataset ($N=60$) were used. Aggregate channel features object detectors were trained to detect reference fat and muscle tissue regions, which were processed and utilized to normalize the 3D images by linear scaling. Mean prostate pseudo T2 values after normalization were compared to literature values. Inter-patient histogram intersections of voxel intensities in the prostate were compared between our approach, the original images, and other commonly used normalization methods. Healthy vs. malignant tissue classification performance was compared before and after normalization.

Results The prostate pseudo T2 values of the three tested datasets (mean ± standard deviation = 78.49 ± 9.42, 79.69 ± 6.34 and 79.29 ± 6.30 ms) corresponded well to T2 values from literature (80 ± 34 ms). Our normalization approach resulted in significantly higher ($p<0.001$) inter-patient histogram intersections (median = 0.746) than the original images (median = 0.417) and most other normalization methods. Healthy vs. malignant classification also improved significantly ($p<0.001$) in peripheral (AUC 0.826 vs. 0.769) and transition (AUC 0.743 vs. 0.678) zones.

Conclusion An automated dual-reference tissue normalization of T2W images could help improve the quantitative assessment of prostate cancer.

Keywords Prostate · Reference tissue · Normalization · MRI · Object recognition

Introduction

Prostate cancer is the second most commonly diagnosed cancer and the leading cause of cancer-related deaths among men worldwide [1]. Multiparametric magnetic resonance imaging (mpMRI) has been established as a valuable diagnostic tool for prostate cancer [2, 3]. T2-weighted (T2W) MR imaging is considered an essential pillar of mpMRI for prostate cancer diagnosis due to the high spatial resolution and the superior anatomical details it provides [3–5]. However, unlike other mpMRI sequences such as diffusion-weighted and dynamic contrast-enhanced imaging, the use of T2W imaging has mainly been limited to a qualitative evaluation of prostate anomalies. Its utility for quantitative analysis is hindered by, among other things, non-standard signal intensity (SI) attributed to scanner parameters such as the field strength, coil type, signal amplification, and acquisition protocols [6–9]. To make use of T2W images for quantitative analysis, an image processing step called SI normalization is often required, which theoretically removes the variation in SI between images from different scan sessions. Consequently, SI normalization enables comparing T2W image values from different patients (inter-patient comparison), patient follow-up at multiple scans over time

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10334-020-00871-3) contains supplementary material, which is available to authorized users.

✉ Mohammed R. S. Sunoqrot
mohammed.sunoqrot@ntnu.no

[1] Department of Circulation and Medical Imaging, NTNU, Norwegian University of Science and Technology, 7030 Trondheim, Norway

[2] Department of Radiology and Nuclear Medicine, St. Olavs Hospital, Trondheim University Hospital, 7030 Trondheim, Norway

(intra-patient comparison), and tissue classification tasks in the setting of a radiomics or computer-assisted diagnosis approach [10, 11].

SI normalization is not new, and over the years, different approaches have been proposed for prostate imaging. Due to their simplicity, histogram-based approaches, which typically depend on pre-set histogram landmarks to deform or rescale the SI [7, 12], have become the most commonly used [10, 13–16]. A drawback of these methods is that they usually rely on the content in the complete 2D or 3D image, which is subject to variation due to differences in scan settings (e.g. the field-of-view) and patient-related factors (e.g. bladder filling). Recently, SI normalization utilizing single or multiple reference tissues has shown promise as an alternative to histogram-based methods [17–21]. In single reference tissue normalization, the original T2W image SI is scaled by the SI in the corresponding reference tissue region-of-interest (ROI). One common example of this in the prostate is normalization to the SI of the obturator internus or levator ani muscles [17, 22–24]. Multi-reference tissue normalization, on the other hand, utilizes the SIs of multiple reference tissues to create a linear or non-linear regression model to estimate the normalized T2W image values [18, 19]. The assumption is that reference tissue-based normalization is less sensitive to variations in scan settings and patient-related factors. However, a key aspect of this approach is labelling the reference tissues, to enable SI extraction. Currently, this is done manually, which is a time-consuming and tedious process. Automated delineation of reference tissue ROIs would make the approach more efficient and could possibly facilitate its integration into clinical practice. This can for example be achieved using automated semantic segmentation or object detection methods. In comparison with semantic segmentation, object detection requires less processing power, time and data [25, 26].

The contribution of this work is a novel method for automated dual-reference tissue normalization of T2W images of the prostate, based on object recognition to extract the reference tissue ROIs. We compared the automatically extracted reference tissue intensities with those of manually delineated ROIs, and evaluated the merit of the proposed method for inter- and intra-patient comparison of T2W image intensities and for the classification of malignant lesions versus healthy prostate tissue.

Materials and methods

Datasets

In this study, transverse T2W images from three separate datasets were used: the PROMISE12 grand challenge dataset ($N=80$) [27], the PROSTATEx challenge dataset ($N=202$) [28] and a dataset of in-house collected T2W images from patients who underwent two sequential MRI scans for detection and biopsy-guiding, respectively ($N=60$). The Regional Committee for Medical and Health Research Ethics (REC Mid Norway) approved the use of the in-house collected dataset (identifier 2017/576) and granted permission for passive consent to be used, whereas the two other datasets were publicly available.

The PROMISE12 dataset [27] consists of multi-centre and multi-vendor transverse T2W images obtained with different field strengths, acquisition protocols and coils. It also includes manual expert segmentations of the whole prostate for 50 cases. The PROSTATEx challenge dataset [28] consists of pre-biopsy mpMRI sequences acquired at Radboud University Medical Centre, Nijmegen, Netherlands. The whole prostate, peripheral zone, and cancer-suspicious volumes of interest (VOIs) were manually delineated by radiologists (at Miller School of Medicine, Miami, FL, USA) based on targeted biopsy locations provided by the challenge organizers. The presence of clinically significant prostate cancer (Gleason score $>3+3$) in the targeted biopsy cores was then used to label each cancer-suspicious VOI as a true positive (malignant) or false positive radiological finding. The rest of the prostate was considered healthy tissue.

The in-house collected dataset was obtained from St. Olavs Hospital, Trondheim University Hospital, Trondheim, Norway between March 2015 and December 2017. It consists of pairs of pre-biopsy 3 T images from 60 patients (median age = 65.5 years; range 47–75 years) acquired at two different time points: first, at the initial visit for detection of prostate cancer (scan 1), and second, during an MR-guided biopsy procedure (scan 2). The interval between scans ranged 1–71 days with a median interval of 7 days. T2W imaging was performed on a Magnetom Skyra 3 T MRI system (Siemens, Erlangen, Germany) with a turbo spin-echo sequence (Scan 1: repetition time/echo time = 4800–9520/104 ms, 320×320 – 384×384 matrix size, 26–32 slices, 3 mm slice thickness and 0.5×0.5–0.6×0.6 mm^2 in plane resolution. Scan 2: repetition time/echo time = 5660–7740/101–104 ms, 320×320–384×384 matrix size, 19–26 slices, 3 mm slice thickness and 0.5×0.5–0.6×0.6 mm^2 in plane resolution). The whole prostate volumes were manually delineated by a radiologist in training.

Proposed intensity normalization method

Figure 1 gives an overview of the proposed method, termed AutoRef. The method contains several tuneable parameters, which were optimized as described in the next section. In the final, optimized version, the 3D T2W images were first pre-processed, which included N4 bias field correction [29], rescaling to the 99th percentile intensity value and resizing the transverse slices to

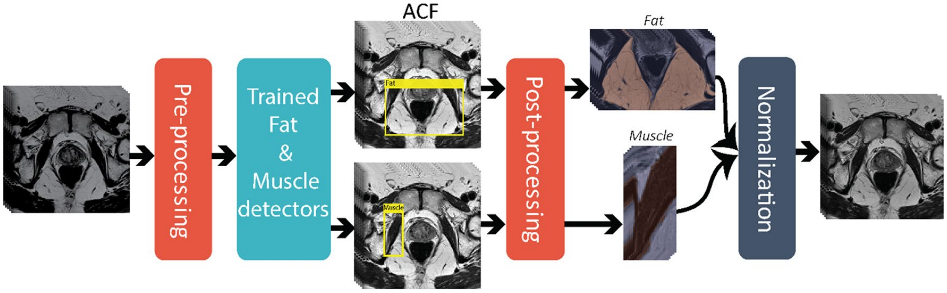


Fig. 1 Overview of AutoRef, the proposed normalization method. The T2W images were first pre-processed including bias field correction, rescaling and resizing. Rectangles containing fat/muscle were then detected slice by slice using trained aggregate channel features (ACF) detectors. The three slices containing rectangular regions with the highest probability of containing fat/muscle were identified and post-processed by Otsu thresholding and morphological opening to extract the largest connected fat/muscle region-of-interest (ROI). From these ROIs, fat/muscle reference intensities were obtained for normalization of the 3D image intensities

384 × 384 pixels with 0.5 × 0.5 mm in-plane resolution. Two separate aggregate channel features (ACF) object detectors [25] were then trained, using two training stages for the iterative training process, to detect rectangular ROIs containing fat and muscle (levator ani muscle) tissue on the 2D transverse slices. Both object detectors were forced to focus on regions where the ROIs were expected to minimize the detection of unwanted structures. For fat, the focus region comprised the lower (posterior) 50% of the image in the lower (inferior) 75% of the slices. For muscle, the focus region comprised the middle (posterior-anterior) 50% of the image in the middle (inferior-superior) 50% of the slices. The three slices containing the rectangular ROIs with the highest probability of fat/muscle were identified, and post-processed by Otsu thresholding [30] and morphological opening, with disk shape of one-pixel radius, to extract the largest connected bright and dark structures in the detected rectangle, representing fat and muscle ROIs, respectively. The fat (I^{fat}) and muscle (I^{muscle}) reference intensity values were then calculated as the 90th and 10th percentiles, respectively, of the intensity values in these ROIs. Subsequently, the 3D image intensities ($I(x, y, z)$) were normalized to pseudo T2 values ($pT2(x, y, z)$) by linearly scaling I^{fat} and I^{muscle} to their respective T2 values at 3 T from literature ($T2^{\text{fat}} = 121$ ms and $T2^{\text{muscle}} = 40$ ms) [31], using Eq. (1):

$$pT2(x,y,z) = \frac{I(x,y,z) - I^{\text{muscle}}}{I^{\text{fat}} - I^{\text{muscle}}} \times (T2^{\text{fat}} - T2^{\text{muscle}}) + T2^{\text{muscle}}. \tag{1}$$

Training, validation and testing

The PROMISE12 dataset was shuffled and split for training ($N = 40$), validation ($N = 20$), and testing ($N = 20$) of AutoRef. Since prostate segmentations were only available for 50 cases, the splitting was semi-random and controlled in a way that ensured that only cases with the required segmentations were included in the validation and test subsets. The PROSTATEx and the in-house collected datasets were used for testing only.

The training and validation subsets were used to train the object detectors and to find the optimal pre- and post-processing settings resulting in the best performance of AutoRef. An overview of the optimization results in the validation subset is provided in Online Resource 1. The trained detectors and optimal parameter settings, as described in the previous section, were subsequently applied to normalize the images in the PROMISE12 test subset, the PROSTATEx dataset and the in-house collected dataset.

Verification of reference tissue intensities

The reference tissue intensities extracted from muscle and fat tissue by AutoRef, I^{fat} and I^{muscle}, respectively, were compared with those of manually drawn ROIs in the PROMISE12 test subset. In the manual approach, a researcher with 3 years of experience with prostate imaging (MRSS) delineated three ROIs in both fat and muscle tissue on what were judged to be representative T2W slices by visual inspection. The 90th and 10th percentiles of the intensity values within the manual fat and muscle ROIs, respectively, were compared to I^{fat} and I^{muscle} and the relative differences and absolute relative differences were calculated. Visual inspection of all automatically extracted fat and muscle ROIs from the PROMISE12 test subset, the PROSTATEx dataset, and the in-house collected dataset was performed by the same researcher to reveal any suboptimal ROIs. A ROI was considered suboptimal when it failed to detect the tissue of interest or covered additional regions not belonging to fat or muscle on any of the three slices.

Inter- and intra-patient performance of normalization

The performance of AutoRef was compared to the original images and three other automated normalization methods, commonly used in literature, i.e. histogram stretching (Eq. (2)) [8], histogram equalization (histeq function from MATLAB®), and Gaussian kernel normalization (Eq. (3)) [8]:

$$I_{\text{normalized}}(x, y, z) = \frac{I(x, y, z) - I_{\min}}{I_{\max} - I_{\min}} \tag{2}$$

where $I_{\max}$ and $I_{\min}$ represent the maximum and minimum intensity values, respectively, in the original image I.

$$I_{\text{normalized}}(x, y, z) = \frac{I(x, y, z) - \mu}{\sigma} \tag{3}$$

where μ and σ represent the mean and standard deviation of the voxel intensities in the original image I, respectively.

Furthermore, the performance of AutoRef using two reference tissues (as proposed) was compared to that of AutoRef$^{\text{muscle}}$ (Eq. 4), which uses only muscle reference intensity values, as by several other studies [17, 22–24]:

$$pT2(x, y, z) = \frac{I(x, y, z)}{I^{\text{muscle}}} \times T2^{\text{muscle}}, \tag{4}$$

where I^{muscle} represents the mean value of the automatically extracted muscle ROIs and $T2^{\text{muscle}}$ the muscle T2 value from literature.

The histogram intersections (Eq. 5) of whole prostate voxel intensities of each pair of patients within the PROMISE12 test subset were used as a metric of inter-patient performance. In addition, the PROSTATEx dataset was used to separately evaluate the inter-patient histogram intersections in the peripheral (PZ) and transition zone (TZ):

$$\text{Intersection}\left(H_x, H_y\right) = \sum_{i=1}^{n} \min\left(H_x(i), H_y(i)\right) \tag{5}$$

where H_x and H_y represent the intensity histograms of patient x and patient y, respectively, and n represents the number of histogram bins (set to 100). H_x and H_y were normalized to the total number of voxels in the prostate or zone.

The in-house collected dataset was used to assess the intra-patient performance, by measuring the whole prostate histogram intersection between the pair of consecutive scans of the same patient (Eq. 6):

$$\text{Intersection}\left(H_1, H_2\right) = \sum_{i=1}^{n} \min\left(H_1(i), H_2(i)\right) \tag{6}$$

where H_1 and H_2 represent the histograms for the first and second scans of the same patient, respectively, and n represents the number of histogram bins (set to 100). H_1 and H_1 were normalized to the total number of voxels in the prostate.

For all datasets, the $pT2(x, y, z)$ values of prostate tissue obtained with AutoRef and AutoRef$^{\text{muscle}}$ were compared to T2 values from the literature [31]. Furthermore, the $pT2(x, y, z)$ values of prostate tissue obtained with AutoRef were compared between patients scanned with and without an endorectal coil.

Classification of malignant lesions versus healthy prostate tissue

Mean intensity values were extracted from the histologically verified malignant lesions and from healthy tissue in the PZ and TZ of the PROSTATEx dataset. The values were used as predictors in logistic regression models to distinguish healthy prostate tissue from malignant lesions in the PZ and TZ, separately. To ensure representative results least influenced by how the data was split, the models were trained and tested using 10 iterations with fivefold cross-validation. In each iteration, the dataset was randomly split, in a controlled way, into training (4 folds) and testing (1 fold) datasets, allowing each fold to be used once for testing. Receiver operating characteristic (ROC) curves were created to evaluate the performance of the classifier at each iteration and the mean and 95% confidence interval (CI) of the area under the curves (AUC) was reported.

Statistical analysis

Wilcoxon signed-rank tests were used to assess statistical differences between the manually and automatically obtained reference tissue intensities, and between the histogram intersections of the various normalization methods. Two-sample t tests were used to assess statistical differences between the pseudo T2 and literature T2 values of the prostate [31], and between the prostate pseudo T2 values of patients scanned with and without an endorectal coil. Wilcoxon rank-sum tests were used to assess statistical differences between the mean intensity values of healthy and malignant regions after normalization. DeLong's method [32] was used to assess statistical differences between AUCs. The tests were followed by Benjamini–Hochberg correction for multiple comparisons [33] with false discovery rate of 0.05. Corrected p values less than 0.05 were considered statistically significant.

All algorithms and analyses were implemented and performed in MATLAB R2019b (The Mathworks, Nattick, MA, USA). The proposed algorithm will be made available on GitHub at https://github.com/ntnu-mr-cancer/AutoRef.

Results

Verification of reference tissue intensities

Figure 2a shows the manually and automatically extracted fat and muscle intensities, respectively, for all cases in the PROMISE12 test subset before normalization. There were significant differences between the reference intensity values from manually and automatically detected fat ($p=0.048$) and muscle ($p=0.018$) ROIs, with relative differences (median (range)) of 2.52% (– 16.21 to 39.86%) for fat and 7.03% (– 20.24 to 23.20%) for muscle. The absolute relative differences [median (range)] between the manual and automated approach were 5.25% (0.17–39.86%) for fat and 9.10% (1.74–23.20%) for muscle intensities. Visual inspection revealed that automated ROIs were suboptimal in 4/20 (20%), 4/202 (2%) and 0/120 (0%) cases for fat and in 0/20 (0%), 3/202 (1.5%) and 0/120 (0%) for muscle ROIs in the PROMISE12 test subset, the PROSTATEx dataset and the in-house collected dataset, respectively, whereas the method performed well in all other cases. In the PROMISE12 test subset, 3/4 (75%) suboptimal ROIs were found in patients with an endorectal coil. Figure 2b shows representative examples of optimal ROIs automatically extracted with our method. All automatically extracted suboptimal ROIs are shown in Online Resource 2. It can be appreciated that the 'suboptimal parts' of the ROIs are often relatively small and of similar image intensity compared to the 'correct parts' of the ROIs, so their impact on the normalization is limited as shown in Online Resource 2.

Inter- and intra-patient evaluation of normalization performance

Figure 3 shows examples from the PROMISE12 test subset, the PROSTATEx dataset, and the in-house collected dataset before and after normalization using AutoRef. The image intensities are more homogeneous within and

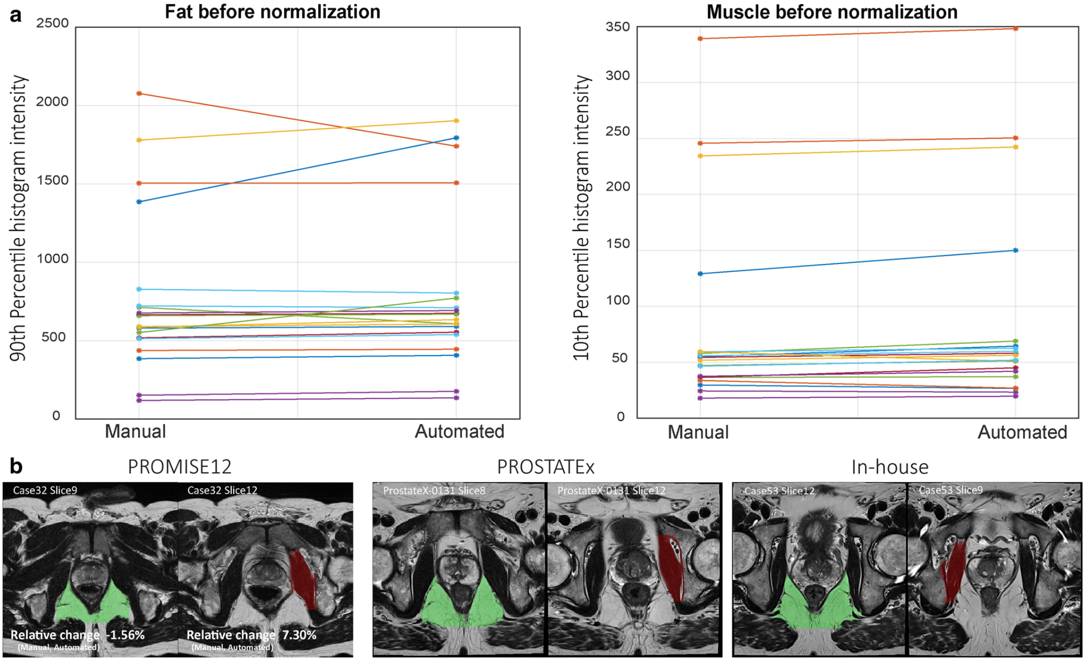


Fig. 2 **a** The 90th and 10th percentiles of the fat and muscle intensities before normalization, respectively, in manually placed and automatically detected ROIs. **b** Representative examples of optimal fat (green) and muscle (red) ROIs automatically extracted with our method

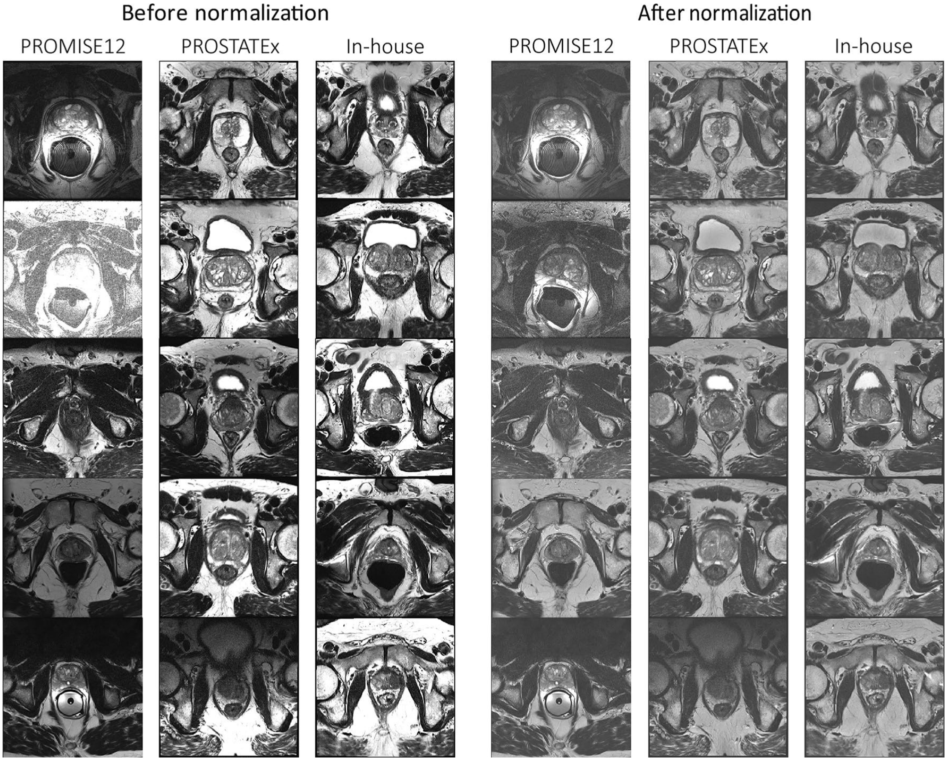


Fig. 3 Central slice through the prostates of five patients from the PROMISE12 test subset, the PROSTATEx dataset and the in-house collected dataset before (left panel) and after normalization (right panel). In both panels, the images were window-levelled from 0 to 2 times the mean prostate intensity of all images in the respective dataset

between the datasets after normalization. This improvement is most obvious in the PROMISE12 dataset, which was acquired with varying protocols, field strengths, and at multiple centres.

The intensity histograms from the original and normalized images of PROMISE12 test subset are displayed in Online Resource 3. Figure 4a and Table 1 show that AutoRef resulted in significantly higher inter-patient intersections than the original data and the other normalization methods, except for AutoRef$^{\text{muscle}}$.

Figure 4b, c and Table 1 also present the inter-patient histogram intersections for PZ and TZ of the PROSTATEx dataset. In both zones, the histogram intersections after normalization with AutoRef were significantly higher than those of the original data and the other normalization methods, except for histogram stretching in TZ.

The intra-patient histogram intersections between scan 1 and scan 2 of the in-house collected dataset are shown in Fig. 4c and Table 2. AutoRef resulted in significantly higher intra-patient intersections than histogram equalization but performed similar to the original data and the other normalization methods.

Figure 5 compares the pseudo T2 values of the whole prostate obtained with AutoRef and AutoRef$^{\text{muscle}}$ with those reported in the literature (80 ± 34 ms) [31]. Using AutoRef, the mean ± standard deviation prostate pseudo T2 values were 78.49 ± 9.42 ms ($p = 0.063$), 79.69 ± 6.34 ms ($p = 0.486$) and 79.29 ± 6.30 ms ($p = 0.161$) for PROMISE12 test subset, the PROSTATEx dataset and the in-house collected dataset, respectively. Pseudo T2 values were not significantly different between patients scanned with (83.15 ± 8.85 ms) or without (79.36 ± 6.41 ms) an endorectal coil ($p = 0.690$). Using AutoRef$^{\text{muscle}}$, the prostate pseudo T2 values were significantly higher ($p < 0.001$) than literature values for all the datasets.

Classification of malignant lesions versus healthy prostate tissue

Figure 6a, b and Table 3 compare the performances (ROC curves and mean AUCs of the 10 iterations, respectively) of

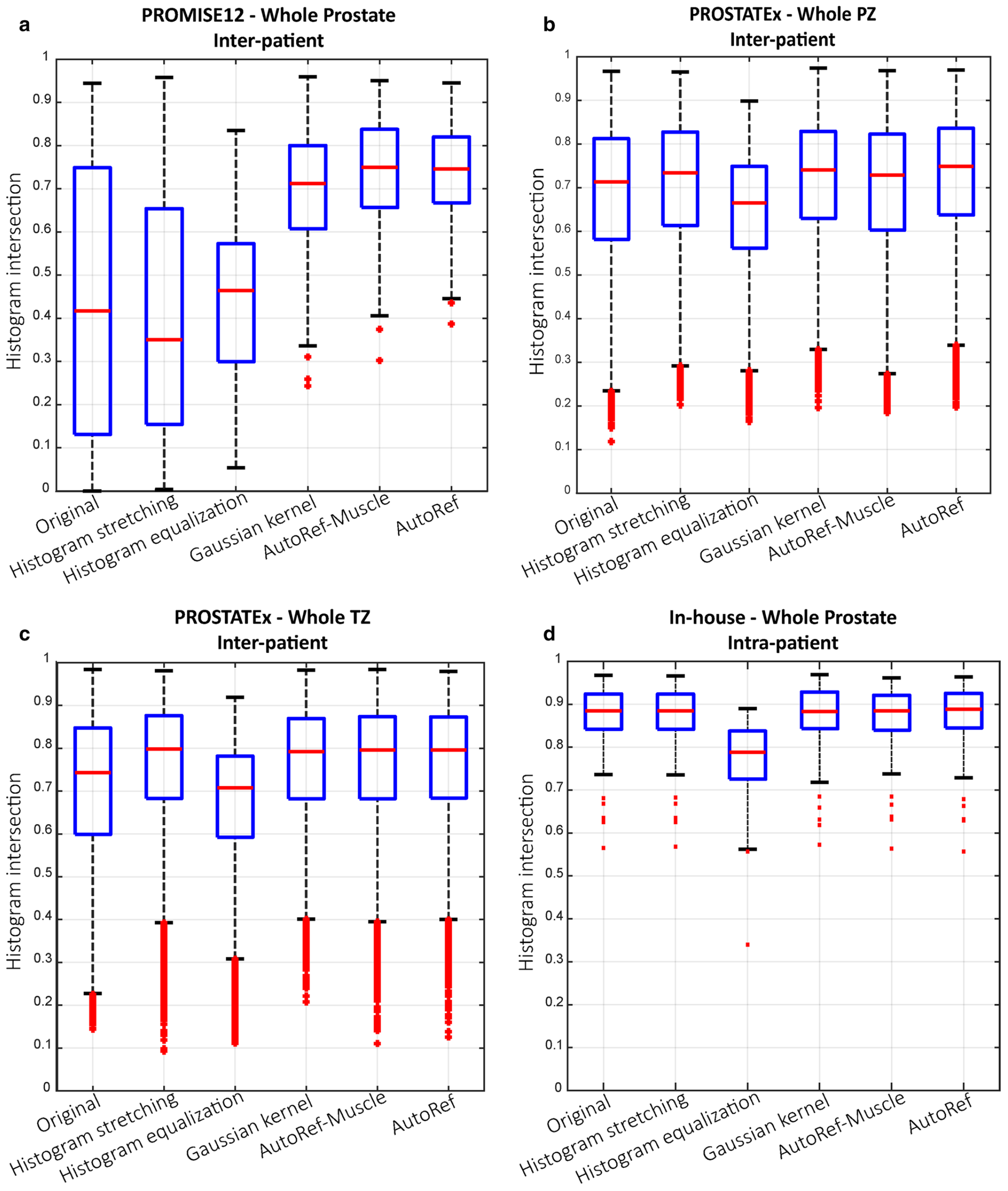


Fig. 4 The inter-patient histogram intersections of the proposed method (AutoRef) compared to original and normalized images for the whole prostate (**a**), the peripheral (PZ; **b**) and transitional zone (TZ; **c**), respectively. The PROMISE12 test subset and PROSTATEx dataset were used in **a**, and **b** and **c**, respectively. AutoRef intersections were significantly higher ($p < 0.001$) than others, except for AutoRef$^{\text{muscle}}$ in **a** ($p = 0.424$) and histogram stretching in **c** ($p = 0.154$). The histogram intersections between scan 1 and scan 2 of the in-house collected dataset (**d**) of AutoRef were significantly higher than for histogram equalization ($p < 0.001$), but similar to those of the other methods

Table 1 The inter-patient histogram intersections before (Original data) and after normalization with our proposed method (AutoRef) and the other investigated methods in the whole prostate, peripheral (PZ) and transition zone (TZ)

	Original data	Histogram stretching	Histogram equalization	Gaussian kernel	AutoRef^{muscle}	AutoRef
Whole prostate						
Median	0.417	0.351	0.465	0.712	0.750	0.746
Range	0.000–0.945	0.003–0.958	0.054–0.835	0.244–0.960	0.302–0.951	0.387–0.945
p value	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	0.424	
PZ						
Median	0.714	0.734	0.665	0.741	0.729	0.749
Range	0.118–0.967	0.202–0.965	0.165–0.898	0.196–0.974	0.185–0.968	0.197–0.970
p value	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	
TZ						
Median	0.743	0.799	0.708	0.792	0.796	0.796
Range	0.144–0.984	0.093–0.981	0.111–0.919	0.208–0.983	0.111–0.984	0.126–0.980
p value	**< 0.001**	0.154	**< 0.001**	**< 0.001**	**0.003**	

The PROMISE12 test subset and PROSTATEx dataset were used in Whole prostate, and PZ and TZ, respectively. The bold values indicate a significant difference from AutoRef after correction for multiple testing

Table 2 The intra-patient histogram intersections between scan 1 and scan 2 of the in-house collected dataset before (Original data) and after normalization with our proposed method (AutoRef) and the other investigated methods

	Original data	Histogram stretching	Histogram equalization	Gaussian kernel	AutoRef^{muscle}	AutoRef
Median	0.884	0.885	0.788	0.883	0.884	0.889
Range	0.565–0.968	0.568–0.966	0.340–0.890	0.573–0.969	0.563–0.961	0.557–0.964
p value	0.640	0.640	**< 0.001**	0.774	0.640	

The bold values indicate a significant difference from AutoRef after correction for multiple testing

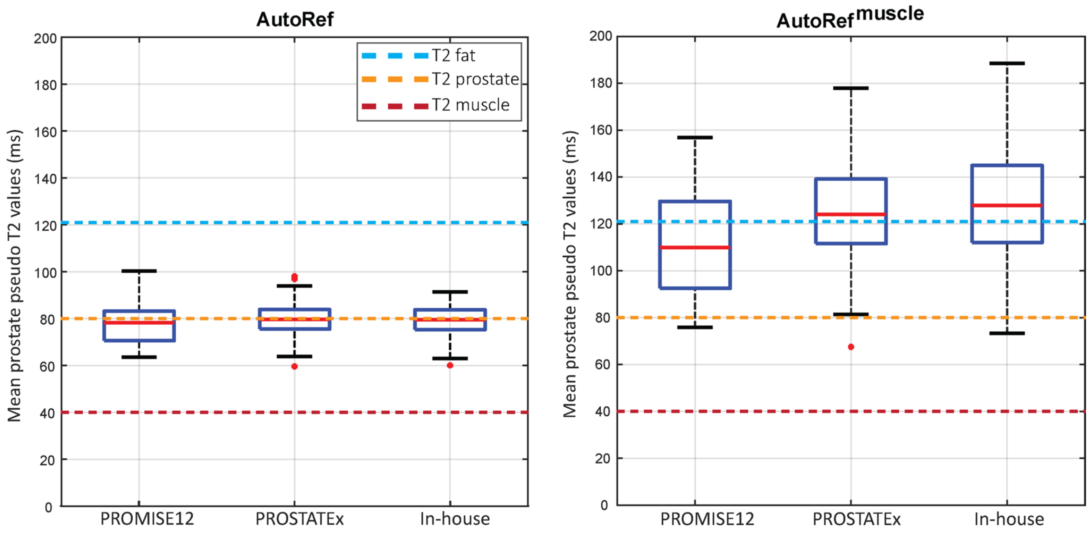


Fig. 5 Box and whisker plots of the mean prostate pseudo T2 values of the patients in the PROMISE12 test subset, the PROSTATEx dataset and the in-house collected dataset after normalization with the proposed dual-reference normalization method (AutoRef) and single reference tissue normalization (AutoRef^{muscle}). The dashed lines correspond to the T2 values reported in literature. All the mean prostate T2 values for AutoRef^{muscle}, but not AutoRef, were significantly higher than those reported in literature ($p < 0.001$)

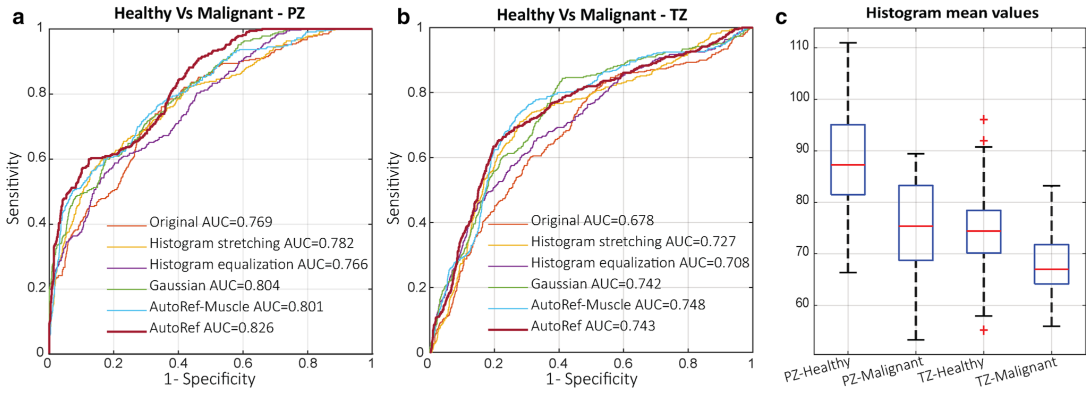


Fig. 6 The receiver operating characteristic curves and areas under the curves (AUC; mean of 10 iterations) for the proposed method (AutoRef), the original images and the other investigated normalization methods in the peripheral (PZ; **a**) and transitional zone (TZ; **b**). In PZ, the AUC for AutoRef was significantly higher than that of the other methods ($p < 0.001$), whereas in TZ it was significantly higher than the original data ($p < 0.001$), histogram stretching ($p = 0.010$) and histogram equalization ($p = 0.007$). The mean pseudo T2 values (**c**) were significantly different between healthy and malignant regions in both the PZ and TZ. ($p < 0.001$)

Table 3 Areas under the curves (AUC; mean of 10 iterations) for the proposed method (AutoRef), the original images and the other investigated normalization methods when classifying healthy versus malignant tissues in the peripheral (PZ) and transition zone (TZ)

	Original data	Histogram stretching	Histogram equalization	Gaussian kernel	AutoRef$^{\text{muscle}}$	AutoRef
PZ						
AUC	0.769	0.782	0.766	0.804	0.801	0.826
95% CI	0.765–0.772	0.778–0.787	0.761–0.771	0.800–0.808	0.797–0.805	0.822–0.830
p value	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	
TZ						
AUC	0.678	0.727	0.708	0.742	0.748	0.743
95% CI	0.672–0.684	0.723–0.730	0.703–0.712	0.739–0.746	0.744–0.751	0.738–0.748
p value	**< 0.001**	**0.010**	**0.007**	0.881	0.559	

The bold values indicate a significant difference from AutoRef after correction for multiple testing

CI confidence interval

AutoRef and other methods in the classification of healthy tissue versus biopsy-confirmed cancer regions. In the PZ, AutoRef performed significantly better than the original data and the other normalization methods. In the TZ, the performance was similar to Gaussian kernel normalization and AutoRef$^{\text{muscle}}$, but significantly better than the original data, histogram stretching and histogram equalization. Figure 6c shows box and whisker plots of the mean pseudo T2 values of healthy and malignant regions after AutoRef normalization, which were significantly different in both the PZ and TZ ($p < 0.001$).

Discussion

In this paper, we propose a new method for automated dual-reference tissue normalization of T2W images of the prostate, which shows promise for quantitative assessment of prostate cancer and could ease the comparison of T2-weighted images between and within patients. The proposed method successfully uses a simple object detector to extract reference tissue intensities from fat and muscle surrounding the prostate, which are subsequently used for intensity normalization of the 3D T2-weighted image. The proposed method generally resulted in higher inter-patient histogram intersections compared to the other investigated automated normalization methods, which indicates that the normalized intensity values in the prostate are more similar between images. Furthermore, the proposed method resulted in images with pseudo T2 values comparable to T2 values reported in the literature [31]. Lastly, as demonstrated by the improved classification of healthy versus malignant tissue, the proposed method successfully reduced the inter-patient variation in T2W image intensities, which could facilitate the extraction and application of meaningful intensity-based image features for quantitative assessment of prostate cancer, e.g. in a radiomics or computer-assisted diagnosis framework [34].

T2W normalization is paramount for the quantitative assessment of prostate cancer, and several methods have been previously proposed in the literature. Liu et al. [13] defined a non-parametric normalization standard as the median image intensity plus two times the inter-quartile range. Artan et al. [14] and Ozer et al. [15] normalized T2W images in a way similar to the Gaussian kernel method investigated here, but with the mean and standard deviation extracted from the PZ instead of the entire image. However, these methods require manual delineation of the PZ and might not be valid if the image intensities do not follow a Gaussian distribution [10]. Lemaitre et al. [10] chose to normalize the images using a parametric model assuming a Rician distribution of the voxel intensities in the whole prostate. Yet differently, Nyúl et al. [7] proposed a two-stage method, wherein the first stage a template histogram with landmarks of interest is created and in the second stage new histograms are mapped via linear transformation to the template. This method assumes that the MR images of the same sequence should have the same intensity distribution, which might not be the case for varying protocols. Vos et al. proposed a sequence-based approach, which depends on the original T2W signal, proton density value, a reference tissue, and a known sequence model to estimate new normalized T2W images [35]. Although this approach performs well, the intricate nature and additional scan time make its practical implementation difficult. Niaf et al. [20, 21] investigated a single reference tissue method that normalizes the image intensities by dividing by the mean intensity value of the bladder. Likewise, Peng et al. [17] normalized the images separately using each of the levator ani muscle, urinary bladder, and pubic bone, and concluded that using levator ani muscle as a single reference tissue gave the best results. In this work, the performance of AutoRef using only muscle reference intensities was shown to be generally inferior to that based on a dual-reference tissue normalization approach, and unable to correctly map the image intensities to literature T2 values. Our method uses fat as a second reference tissue because it typically has high T2W intensity values, thus together with muscle covering the full range of expected prostate intensity values, it is present in all images and less vulnerable to external factors than for example the urinary bladder.

Recently, Stoilescu et al. [19] showed that multi-reference tissue normalization of T2W prostate images significantly improved prostate cancer classification accuracy in comparison to non-normalized images. Four reference tissues were used based on manually annotated ROIs, which currently hinders the implementation of the method in clinical practice. Therefore, in our work, we developed an automated approach for detecting ROIs to enable multi-reference tissue normalization using two reference tissues (fat and levator ani muscle). The ACF detector used in this work is a relatively simple, classical machine learning approach that was able to accurately detect the fat and muscle ROIs in nearly all cases, despite the small training dataset ($N = 40$). Exceptions were found in 8/342 (2%) cases for fat and 3/342 (1%) cases for muscle ROIs when considering all patients, and in 1/331 (0.3%) and 3/331 (1%) cases, respectively, when considering patients scanned without an endorectal coil. The detection of fat thus performs worse in patients scanned with an endorectal coil but this may not pose a problem in clinical practice, as 3 T MRI with body surface coils is currently the recommended and preferred method. The detection of both fat and muscle may be further improved by using a larger dataset for training, while the method can also be extended to include more reference tissues if deemed necessary, which is subject of further investigation.

Although AutoRef generally performed better than or similar to the other normalization methods in all datasets, the largest differences were observed in the multi-centre, multi-vendor PROMISE12 dataset. In this dataset, images were acquired with 1.5 T or 3 T scanners, with or without an endorectal coil, and with different acquisition protocols, all of which are likely to influence the T2W image intensity. An important advantage of our method to the other investigated normalization methods is that the image intensities could be correctly mapped to literature T2 values [31], irrespective of these factors. The pseudo T2 values could be an interesting alternative to quantitative T2 mapping, given the limited scan time available in clinical practice, but this needs further investigation in studies where T2 maps are also acquired. However, it should be noted that AutoRef does not correct for local differences in signal intensities caused by the non-uniform sensitivity of the receiver coils. This effect is especially apparent for images acquired with an endorectal coil, which typically shows an intensity profile inversely related to proximity to the coil. Although we showed that the mean pseudo T2 values of images acquired with an endorectal coil were comparable to those acquired with body surface coils, there may be differences in intensity distribution within the prostate gland that are not accounted for by AutoRef.

In the intra-patient evaluation, AutoRef had similar intra-patient histogram intersections compared to the original data and most of the other investigated methods. This probably reflects the limited variability in the in-house collected dataset, which has been acquired at the same centre, the same scanner, with the same protocols at a relatively short interval between scans. It would be insightful to assess the performance of the method in a dataset where the same patients are systematically scanned at different hospitals, but such data are probably scarce.

Normalization with the proposed method resulted in a significantly higher AUC for the classification of histologically verified PZ lesions compared to the other methods. For TZ lesions, the AUC was significantly higher than the original data, histogram stretching and histogram equalization, and on par with the other normalization methods. However, the differences in the classification performance were relatively small, which again may be the result of the limited variability in a dataset acquired at a single centre and with a single protocol [28]. Furthermore, considerable overlap in pseudo T2 values was still present between healthy tissue and malignant lesions, especially in the TZ, indicating that pseudo T2 values alone may not be sufficient to detect prostate cancer in clinical practice.

Our study has some limitations. Quantitative T2 maps were not available for the patients included in this study, which hindered a direct comparison of the pseudo T2 values with a gold standard. Although we included several commonly applied automated normalization methods in this study, there are still many more described in the literature, as discussed above, that may perform better than those included here. In addition, it would be interesting to compare the performance of the proposed object detector to that of semantic segmentation for detecting ROIs, which will be subject to further research. Despite these limitations, we have shown that our proposed method for automated dual-reference tissue normalization performed equal to or better than other automated normalization methods. The method requires no manual input and the resulting images can be used for both quantitative and qualitative assessment of prostate cancer.

Conclusion

We successfully developed a method for automated dual-reference tissue normalization of T2W MR images of the prostate using object recognition. The method was shown to reduce T2W intensity variation between scans and could improve the quantitative assessment of prostate cancer on MRI.

Acknowledgements Open Access funding provided by NTNU Norwegian University of Science and Technology (incl St. Olavs Hospital - Trondheim University Hospital). We would like to thank Radboud University Nijmegen in addition to the organizers of the PROSTATEx and PROMISE12 challenges for making their datasets available. We would specifically like to thank Prof. Radka Stoyanova and her team at Miller School of Medicine (Miami, FL, USA) for providing us with the prostate delineations for the PROSTATEx dataset. We would like to thank Dr. Elise Sandsmark from St. Olavs Hospital, Trondheim University Hospital (Trondheim, Norway), for providing us with prostate delineations for the in-house collected dataset.

Author contributions Study conception and design were performed by ME and TFB. Material preparation and data collection were performed by MRSS, KMS and ME. Data analysis and interpretation were performed by MRSS, GAK and ME. The first draft of the manuscript was written by MRSS and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding The Research Council of Norway (Grant Number 295013), Norwegian University of Science and Technology Biotechnology (Grant Number 81770928), and the liaison Committee between the Central Norway Regional Health Authority and the Norwegian University of Science and Technology (Grant Numbers 90368401 and 90265300).

Compliance with ethical standards

Conflict of interest All authors declare that they have no conflict of interest.

Ethical approval All procedures performed in-house in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee [Regional Committee for Medical and Health Research Ethics (REC) in mid Norway, identifier 2017/576] and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A (2018) Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 68(6):394–424
2. Barentsz JO, Richenberg J, Clements R, Choyke P, Verma S, Villeirs G, Rouviere O, Logager V, Futterer JJ (2012) ESUR prostate MR guidelines 2012. Eur Radiol 22(4):746–757

3. Weinreb JC, Barentsz JO, Choyke PL, Cornud F, Haider MA, Macura KJ, Margolis D, Schnall MD, Shtern F, Tempany CM, Thoeny HC, Verma S (2016) PI-RADS prostate imaging—reporting and data system: 2015, version 2. Eur Urol 69(1):16–40
4. Hoeks CM, Barentsz JO, Hambrock T, Yakar D, Somford DM, Heijmink SW, Scheenen TW, Vos PC, Huisman H, van Oort IM, Witjes JA, Heerschap A, Futterer JJ (2011) Prostate cancer: multiparametric MR imaging for detection, localization, and staging. Radiology 261(1):46–66
5. Wang S, Burtt K, Turkbey B, Choyke P, Summers RM (2014) Computer aided-diagnosis of prostate cancer on multiparametric MRI: a technical review of current research. Biomed Res Int 2014:789561
6. Simmons A, Tofts PS, Barker GJ, Arridge SR (1994) Sources of intensity nonuniformity in spin-echo images at 1.5-T. Magn Reson Med 32(1):121–128
7. Nyul LG, Udupa JK, Zhang X (2000) New variants of a method of MRI scale standardization. IEEE Trans Med Imaging 19(2):143–150
8. Loizou CP, Pantziaris M, Seimenis I, Pattichis CS (2009) Brain MR image normalization in texture analysis of multiple sclerosis. In: Proceedings of the 9th international conference on information technology and applications in biomedicine, Larnaca, p 131
9. Madabhushi A, Udupa JK (2006) New methods of MR image intensity standardization via generalized scale. Med Phys 33(9):3426–3434
10. Lemaitre G, Rastgoo M, Massich J, Vilanova JC, Walker PM, Freixenet J, Meyer-Baese A, Meriaudeau F, Marti R (2015) Normalization of T2W-MRI prostate images using Rician a priori. In: Proceedings of SPIE medical imaging, San Diego, p 978529
11. Schwier M, van Griethuysen J, Vangel MG, Pieper S, Peled S, Tempany C, Aerts H, Kikinis R, Fennessy FM, Fedorov A (2019) Repeatability of multiparametric prostate MRI radiomics features. Sci Rep 9(1):9441
12. Ge YL, Udupa JK, Nyul LG, Wei LG, Grossman RI (2000) Numerical tissue characterization in MS via standardization of the MR image intensity scale. J Magn Reson Imaging 12(5):715–721
13. Liu P, Wang SJ, Turkbey B, Grant K, Pinto P, Choyke P, Wood BJ, Summers RM (2013) A prostate cancer computer-aided diagnosis system using multimodal magnetic resonance imaging and targeted biopsy labels. In: Proceedings of SPIE medical imaging, Lake Buena Vista, p 86701G
14. Artan Y, Haider MA, Langer DL, van der Kwast TH, Evans AJ, Yang YY, Wernick MN, Trachtenberg J, Yetik IS (2010) Prostate cancer localization with multispectral MRI using cost-sensitive support vector machines and conditional random fields. IEEE Trans Image Process 19(9):2444–2455
15. Ozer S, Langer DL, Liu X, Haider MA, van der Kwast TH, Evans AJ, Yang YY, Wernick MN, Yetik IS (2010) Supervised and unsupervised methods for prostate cancer segmentation with multispectral MRI. Med Phys 37(4):1873–1883
16. Lv DJ, Guo XM, Wang XY, Zhang J, Fang J (2009) Computerized characterization of prostate cancer by fractal analysis in MR images. J Magn Reson Imaging 30(1):161–168
17. Peng YH, Jiang YL, Oto A (2014) Reference-tissue correction of T-2-weighted signal intensity for prostate cancer detection. In: Proceedings of SPIE medical imaging, San Diego, p 903508
18. Leung KK, Clarkson MJ, Bartlett JW, Clegg S, Jack CR, Weiner MW, Fox NC, Ourselin S, AsDN I (2010) Robust atrophy rate measurement in Alzheimer's disease using multi-site serial MRI: tissue-specific intensity normalization and parameter selection. Neuroimage 50(2):516–523
19. Stoilescu L, Maas MC, Huisman HJ (2017) Feasibility of multi-reference tissue normalization of T2-weighted prostate MRI. In: Proceedings of the 34th annual scientific meeting, European Society for Magnetic Resonance in Medicine & Biology, Barcelona, p 353
20. Niaf E, Rouviere O, Lartizien C (2011) Computer-aided diagnosis for prostate cancer detection in the peripheral zone via multisequence MRI. In: Proceedings of SPIE medical imaging, Lake Buena Vista, p 79633P
21. Niaf E, Rouviere O, Mege-Lechevallier F, Bratan F, Lartizien C (2012) Computer-aided diagnosis of prostate cancer in the peripheral zone using multiparametric MRI. Phys Med Biol 57(12):3833–3851
22. Engelhard K, Hollenbach HP, Deimling M, Kreckel M, Riedl C (2000) Combination of signal intensity measurements of lesions in the peripheral zone of prostate with MRI and serum PSA level for differentiating benign disease from prostate cancer. Eur Radiol 10(12):1947–1953
23. Dikaios N, Alkalbani J, Abd-Alazeez M, Sidhu HS, Kirkham A, Ahmed HU, Emberton M, Freeman A, Halligan S, Taylor S, Atkinson D, Punwani S (2015) Zone-specific logistic regression models improve classification of prostate cancer on multi-parametric MRI. Eur Radiol 25(9):2727–2737
24. Dikaios N, Alkalbani J, Sidhu HS, Fujiwara T, Abd-Alazeez M, Kirkham A, Allen C, Ahmed H, Emberton M, Freeman A, Halligan S, Taylor S, Atkinson D, Punwani S (2015) Logistic regression model for diagnosis of transition zone prostate cancer on multi-parametric MRI. Eur Radiol 25(2):523–532
25. Dollar P, Appel R, Belongie S, Perona P (2014) Fast feature pyramids for object detection. IEEE Trans Pattern Anal Mach Intell 36(8):1532–1545
26. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, van der Laak J, van Ginneken B, Sanchez CI (2017) A survey on deep learning in medical image analysis. Med Image Anal 42:60–88
27. Litjens G, Toth R, van de Ven W, Hoeks C, Kerkstra S, van Ginneken B, Vincent G, Guillard G, Birbeck N, Zhang J, Strand R, Malmberg F, Ou Y, Davatzikos C, Kirschner M, Jung F, Yuan J, Qiu W, Gao Q, Edwards PE, Maan B, van der Heijden F, Ghose S, Mitra J, Dowling J, Barratt D, Huisman H, Madabhushi A (2014) Evaluation of prostate segmentation algorithms for MRI: the PROMISE12 challenge. Med Image Anal 18(2):359–373
28. Armato SG 3rd, Huisman H, Drukker K, Hadjiiski L, Kirby JS, Petrick N, Redmond G, Giger ML, Cha K, Mamonov A, Kalpathy-Cramer J, Farahani K (2018) PROSTATEx Challenges for computerized classification of prostate lesions from multiparametric magnetic resonance images. J Med Imaging (Bellingham) 5(4):044501
29. Tustison NJ, Avants BB, Cook PA, Zheng YJ, Egan A, Yushkevich PA, Gee JC (2010) N4ITK: improved N3 bias correction. IEEE Trans Med Imaging 29(6):1310–1320
30. Otsu N (1979) Threshold selection method from gray-level histograms. IEEE Trans Sys Man Cybern 9(1):62–66
31. Bojorquez JZ, Bricq S, Brunotte F, Walker PM, Lalande A (2016) A novel alternative to classify tissues from T 1 and T 2 relaxation times for prostate MRI. Magn Reson Mater Phys 29(5):777–788
32. Delong ER, Delong DM, Clarkepearson DI (1988) Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 44(3):837–845
33. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate—a practical and powerful approach to multiple testing. J R Stat Soc B 57(1):289–300
34. Stoyanova R, Takhar M, Tschudi Y, Ford JC, Solorzano G, Erho N, Balagurunathan Y, Punnen S, Davicioni E, Gillies RJ, Pollack A (2016) Prostate cancer radiomics and the promise of radiogenomics. Transl Cancer Res 5(4):432–447

35. Vos PC, Hambrock T, Barenstz JO, Huisman HJ (2010) Computer-assisted analysis of peripheral zone prostate lesions using T2-weighted and dynamic contrast enhanced T1-weighted MRI. Phys Med Biol 55(6):1719–1734

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Supplementary material to:

Automated reference tissue normalization of T2-weighted MR images of the prostate using object recognition

By: Mohammed R. S. Sunoqrot, MSc,[1] Gabriel A. Nketiah, PhD,[1,2] Kirsten M. Selnæs, PhD,[1,2] Tone F. Bathen, PhD,[1,2] and Mattijs Elschot, PhD,[1,2]

[1]*Department of Circulation and Medical Imaging, NTNU - Norwegian University of Science and Technology, 7030 Trondheim, Norway.*

[2]*Department of Radiology and Nuclear Medicine, St. Olavs Hospital, Trondheim University Hospital, 7030 Trondheim, Norway.*

*** Corresponding author:**

Name: Mohammed R. S. Sunoqrot

Email: mohammed.sunoqrot@ntnu.no

Magnetic Resonance Materials in Physics, Biology and Medicine

Automated reference tissue normalization of T2-weighted MR images of the prostate using object recognition

Mohammed R. S. Sunoqrot, Gabriel A. Nketiah, Kirsten M. Selnæs, Tone F. Bathen and Mattijs Elschot

Online Resource 1

Optimization of the AutoRef pre- and post-processing settings on the validation set

We optimized the following pre- and post-processing parameters on the validation set: type of scaling, number of the detector training stages, radius of the morphological opening structure, number of the evaluated slices, low (muscle) and high (fat) percentiles used to calculate the reference intensity values and focus regions (inferior-superior slices and posterior-anterior rows) for fat and muscle. The table below lists the effect on the median histogram intersection (HI) when changing one parameter at the time, while keeping the other parameters set at the ***bold*** *values. The* ***bold*** *set of parameters was considered our optimal set as no further improvements of more than 1% could be achieved by additional tuning of the parameters (i.e. Change in HI median < 1% for all parameter values).*

Settings	HI median	Change in HI median (%)	Selected as an optimal
		Scaling	
99th Percentile	0.76997959	**0**	*X*
Median	0.77183996	0.241613039	
Max	0.76662009	-0.436310938	
None	0.75950878	-1.359881374	
		Number of detector Stages	
1	0.77192269	0.252357431	
2	0.76997959	**0**	*X*
3	0.76997959	0	
4	0.76997959	0	
5	0.76997959	0	
		Morphological opening structure radius	
1	0.76997959	**0**	*X*
2	0.77002793	0.006277583	
3	0.7677967	-0.283500407	
4	0.76807893	-0.246846188	
5	0.7675968	-0.309461357	
		Number of evaluated Slices	
1	0.76806568	-0.248566275	
2	0.77272643	0.356741256	
3	0.76997959	**0**	*X*
4	0.76859372	-0.179988324	
5	0.77103737	0.137376853	

Low percentile (Muscle)			
25	0.7736801	0.48059794	
20	0.77221788	0.290693982	
15	0.77017365	0.025203469	
10	0.76997959	**0**	***X***
5	0.76839093	-0.206325746	
High percentile (Fat)			
75	0.76036739	-1.044199148	
80	0.76423774	-0.540504432	
85	0.76676418	-0.211708731	
90	0.76839093	**0**	***X***
95	0.76178638	-0.859529925	
Focus Region Slices (Fat)			
All [0-100%]	0.76997959	0	
Lower Half [0-50%]	0.76997959	0	
lower 75% [0-75%]	0.76997959	**0**	***X***
Middle [25%-75%]	0.76964891	-0.042946643	
Upper Half [50%-100%]	0.74567085	-3.157063397	
Upper 75% [25%-100%]	0.76964891	-0.042946643	
Focus Region Rows (Fat)			
All [0-100%]	0.76968231	-0.03860894	
Lower Half [0-50%]	0.76997959	**0**	***X***
lower 75% [0-75%]	0.76968231	-0.03860894	
Middle [25%-75%]	0.75187963	-2.3507062	
Upper 75% [25%-100%]	0.75562408	-1.864402074	
Focus Region Slices (Muscle)			
All [0-100%]	0.77053318	0.071896978	
Lower Half [0-50%]	0.7737537	0.490156996	
lower 75% [0-75%]	0.76997959	0	
Middle [25%-75%]	0.76997959	**0**	***X***
Upper Half [50%-100%]	0.7679594	-0.262369737	
Upper 75% [25%-100%]	0.77053318	0.071896978	
Focus Region Rows (Muscle)			
All [0-100%]	0.76913026	-0.110305212	
Lower Half [0-50%]	0.76709563	-0.374550451	
lower 75% [0-75%]	0.76913026	-0.110305212	
Middle [25%-75%]	0.76997959	**0**	***X***
Upper 75% [25%-100%]	0.76912854	-0.110528935	

Magnetic Resonance Materials in Physics, Biology and Medicine

Automated reference tissue normalization of T2-weighted MR images of the prostate using object recognition

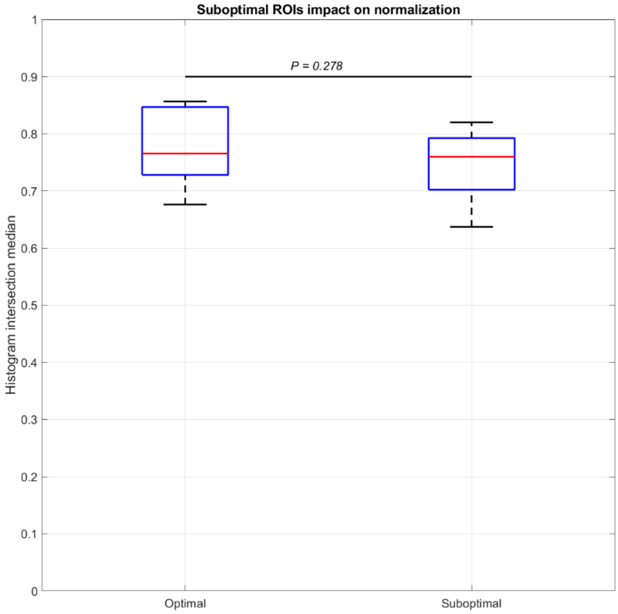
Mohammed R. S. Sunoqrot, Gabriel A. Nketiah, Kirsten M. Selnæs, Tone F. Bathen and Mattijs Elschot

Online Resource 2

All suboptimal automatically extracted ROIs using AutoRef and their impact on the normalization

The suboptimal ROIs impact on the normalization:

To measure the suboptimal ROIs impact on the normalization, we compared the medians of the histogram intersections of all the cases with suboptimal ROIs with an equivalent number (overall and per dataset) of randomly selected cases with optimal ROIs. For each case, the median of the histogram intersections with the rest of its dataset cases, excluding those with suboptimal ROIs, was taken. Wilcoxon signed rank test was used to assess the statistical difference and *p*-values less than 0.05 were considered statistically significant. The test showed no significant difference between the medians (*p*=0.278).



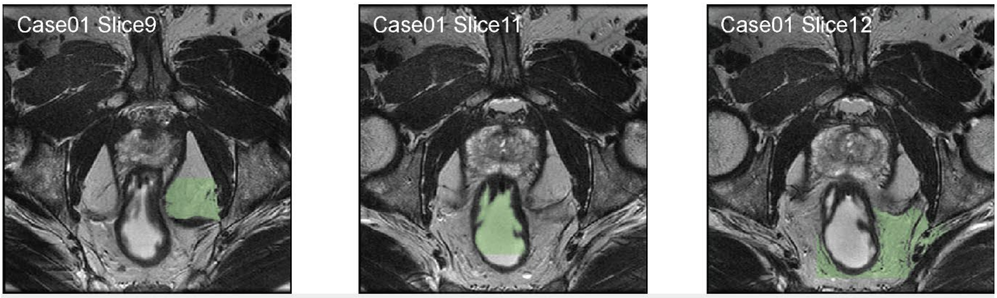
The suboptimal automatically extracted ROIs using AutoRef:

Below, each row represents a case, while the columns are the 3 detected regions-of-interest (ROIs) for that case. The fat ROIs are in green and the muscle ROIs are in red. Under each case of PROMISE12 test subset, the absolute relative difference of the reference intensity values between the manual and automated approach has been given.

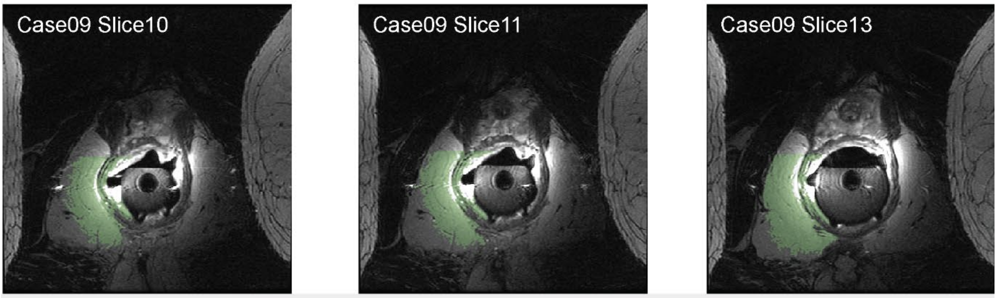
Criteria:

A case was considered a suboptimal when any of its extracted ROIs failed to detect the fat or muscle tissue, or covered additional regions not belonging to fat or muscle.

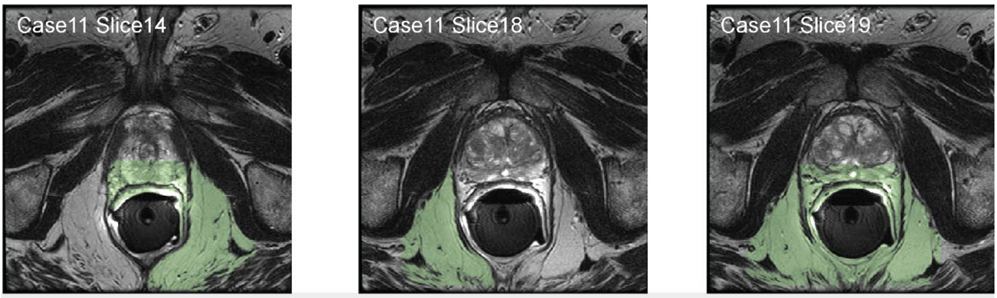
PROMISE12 test subset



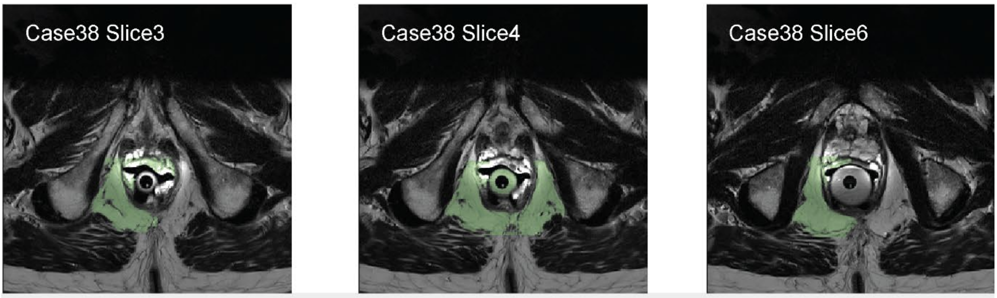
The absolute relative difference = 16.37%



The absolute relative difference = 29.46%

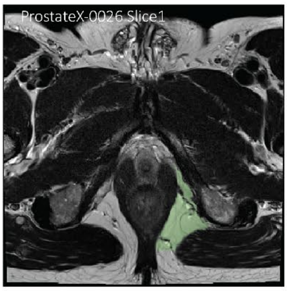


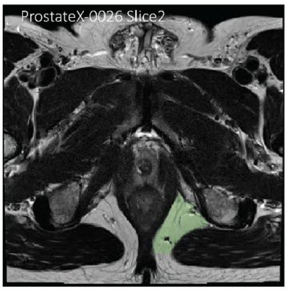
The absolute relative difference = 6.96%

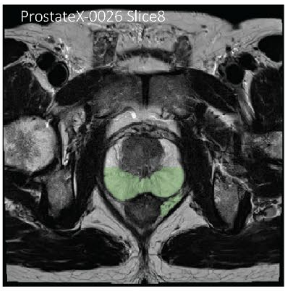


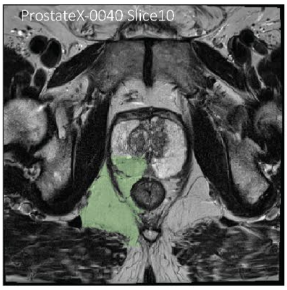
The absolute relative difference = 39.86%

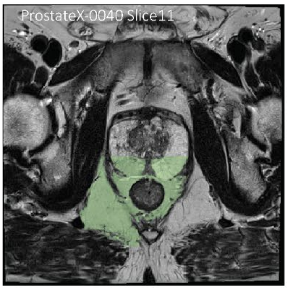
PROSTATEx dataset

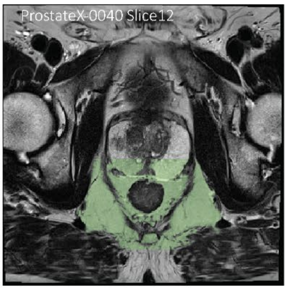


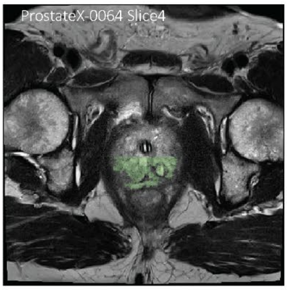


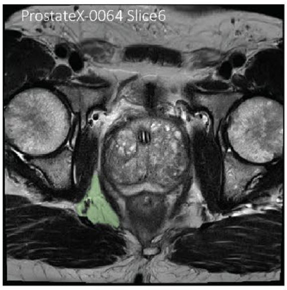


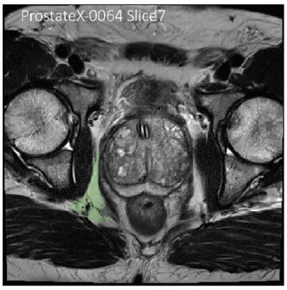


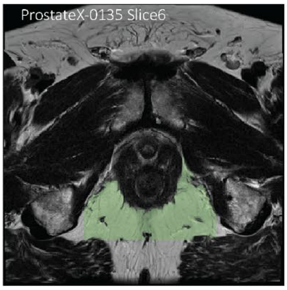


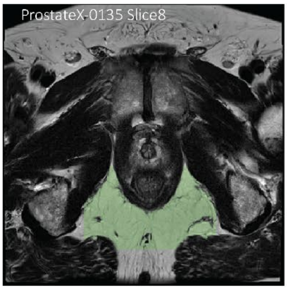


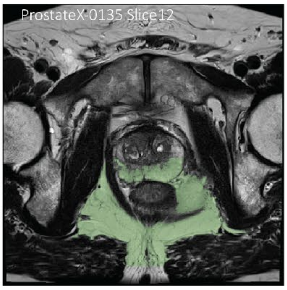


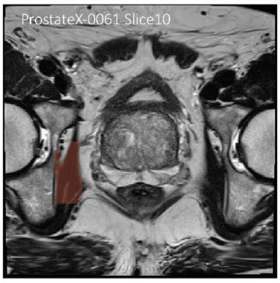


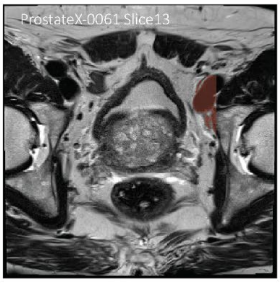


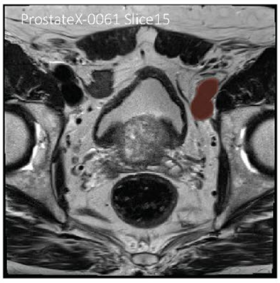


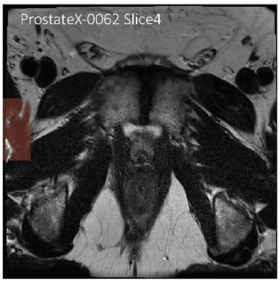


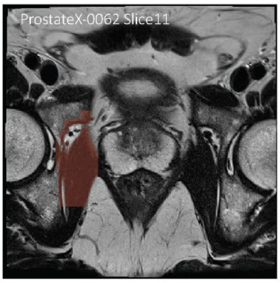


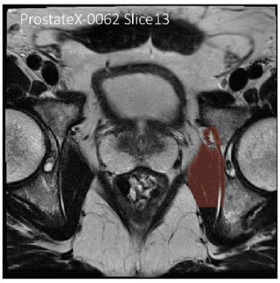

ProstateX-0061 Slice10

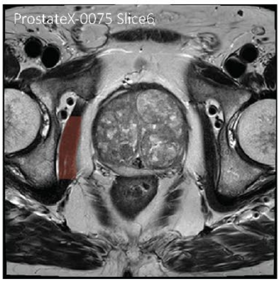

ProstateX-0061 Slice13

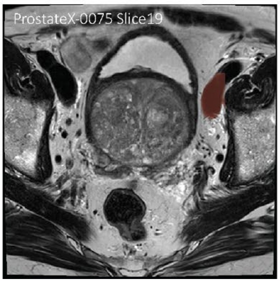

ProstateX-0061 Slice15

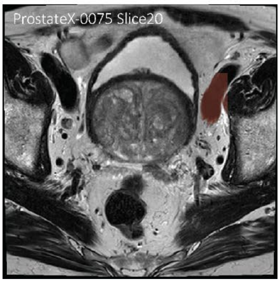

ProstateX-0062 Slice4


ProstateX-0062 Slice11


ProstateX-0062 Slice13


ProstateX-0075 Slice6


ProstateX-0075 Slice19


ProstateX-0075 Slice20

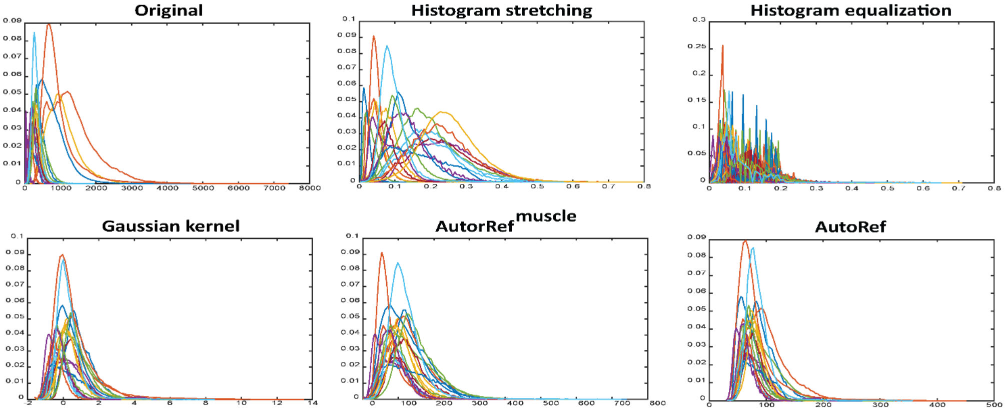
Magnetic Resonance Materials in Physics, Biology and Medicine

Automated reference tissue normalization of T2-weighted MR images of the prostate using object recognition

Mohammed R. S. Sunoqrot, Gabriel A. Nketiah, Kirsten M. Selnæs, Tone F. Bathen and Mattijs Elschot

Online Resource 3

A visual representation of the image intensity histograms from the original and normalized images of PROMISE12 test subset. AutoRef is our proposed method



Paper II

diagnostics

Article

A Quality Control System for Automated Prostate Segmentation on T2-Weighted MRI

Mohammed R. S. Sunoqrot [1,*], Kirsten M. Selnæs [1,2], Elise Sandsmark [2], Gabriel A. Nketiah [1,2], Olmo Zavala-Romero [3,4], Radka Stoyanova [3], Tone F. Bathen [1,2,†] and Mattijs Elschot [1,2,†]

[1] Department of Circulation and Medical Imaging, NTNU—Norwegian University of Science and Technology, 7030 Trondheim, Norway; Kirsten.Margrete.Selnes@stolav.no (K.M.S.); gabriel.a.nketiah@ntnu.no (G.A.N.); tone.f.bathen@ntnu.no (T.F.B.); mattijs.elschot@ntnu.no (M.E.)
[2] Department of Radiology and Nuclear Medicine, St. Olavs Hospital, Trondheim University Hospital, 7030 Trondheim, Norway; elise.sandsmark@stolav.no
[3] Department of Radiation Oncology, University of Miami Miller School of Medicine, Miami, FL 33136, USA; ozavala@coaps.fsu.edu (O.Z.-R.); rstoyanova@med.miami.edu (R.S.)
[4] Center for Ocean-Atmospheric Prediction Studies, Florida State University, Tallahassee, FL 32306, USA
* Correspondence: mohammed.sunoqrot@ntnu.no
† These authors contributed equally to this work.

Received: 31 August 2020; Accepted: 16 September 2020; Published: 18 September 2020

Abstract: Computer-aided detection and diagnosis (CAD) systems have the potential to improve robustness and efficiency compared to traditional radiological reading of magnetic resonance imaging (MRI). Fully automated segmentation of the prostate is a crucial step of CAD for prostate cancer, but visual inspection is still required to detect poorly segmented cases. The aim of this work was therefore to establish a fully automated quality control (QC) system for prostate segmentation based on T2-weighted MRI. Four different deep learning-based segmentation methods were used to segment the prostate for 585 patients. First order, shape and textural radiomics features were extracted from the segmented prostate masks. A reference quality score (QS) was calculated for each automated segmentation in comparison to a manual segmentation. A least absolute shrinkage and selection operator (LASSO) was trained and optimized on a randomly assigned training dataset (N = 1756, 439 cases from each segmentation method) to build a generalizable linear regression model based on the radiomics features that best estimated the reference QS. Subsequently, the model was used to estimate the QSs for an independent testing dataset (N = 584, 146 cases from each segmentation method). The mean ± standard deviation absolute error between the estimated and reference QSs was 5.47 ± 6.33 on a scale from 0 to 100. In addition, we found a strong correlation between the estimated and reference QSs (rho = 0.70). In conclusion, we developed an automated QC system that may be helpful for evaluating the quality of automated prostate segmentations.

Keywords: prostate; segmentation; deep learning; radiomics; quality control; computer-aided detection and diagnosis; MRI; machine learning

1. Introduction

Prostate cancer is one of the most commonly diagnosed cancers among men worldwide [1]. Precise diagnosis is essential for management of the disease, where early detection and staging can increase the survival rate [2]. The current diagnostic process includes measuring elevated prostate-specific antigen (PSA) in the blood followed by prostate biopsy sampling and histopathology analysis. The addition of multiparametric magnetic resonance imaging (mpMRI) and the establishment of international guidelines for the image acquisition and interpretation have improved the diagnostic precision for prostate cancer [3,4]. However, the traditional, qualitative radiological interpretation of the images has a number of limitations, such as high inter-observer variability [5], its time-consuming nature and a lack of scalability of the manual data handling approach with increasing demand [6,7].

Automated computer-aided detection and diagnosis (CAD) systems, which exploit the quantitative information in MR images, are providing promising solutions to overcome these limitations of qualitative image interpretation and support clinical decision making [6,8]. Typically, the segmentation of the organ of interest, in this case the prostate gland, constitutes one of the first important steps in a CAD system workflow [7,9]. This step helps remove irrelevant image information and facilitates subsequent extraction of quantitative image features (radiomics) from sub-regions/volumes such as tumors for further analysis or diagnosis. However, manual segmentation of the prostate, which is traditionally performed on T2-weighted (T2W) MR images by radiologists, is a time-consuming task. Fortunately, recently developed segmentation algorithms have shown great promise to fully automate this step [10–15], which would save valuable time and could facilitate the integration of CAD systems in clinical practice.

Deep learning-based methods seem to be the most promising for this purpose, as they outperform the more traditional methods in the PROMISE12 prostate segmentation grand challenge [16] (https://promise12.grand-challenge.org/evaluation/results). Interestingly, the top-performing methods in this challenge scored better—on average—than a non-expert second reader. Nevertheless, none of the proposed methods is perfect. Occasionally, each of the proposed segmentation methods results in a few cases with unpredictable, suboptimal contours. Time-consuming manual verification of the contours by radiologists is thus still a necessary step, which limits the implementation of automated prostate segmentation algorithms in clinical practice. A quality control (QC) system that automatically provides an assessment of the segmentation quality could help overcome this limitation and standardize decisions about segmentation quality.

The aim of this study was to develop a fully automated QC system that generates a quality score for assessing the accuracy of automated prostate segmentations on T2W MR images. We trained, optimized and tested the proposed QC system using two data cohorts and four different deep learning-based segmentation algorithms. We explored the importance of the radiomics features the system is based on and compared a generalizable model with models trained on specific combinations of dataset and segmentation algorithm. Finally, we show that the quality of the segmentations can be successfully estimated by our QC system.

2. Materials and Methods

We propose a novel QC system, which is designed to automatically score the quality of prostate segmentations on T2W MR images. Briefly, the inputs to the QC system are the T2W MR image and the corresponding deep learning-based prostate segmentation. Radiomics features are extracted from the segmented prostate image volume and fed into a least absolute shrinkage and selection operator (LASSO) to build a linear regression model [17], which is trained to generate an estimated quality score (eQS). Reference quality scores (rQSs) based on manual segmentations from experts are then used to assess the performance of the QC system.

2.1. Dataset

In this study, the PROMISE12 grand challenge [16] training dataset (N = 50) was only used to train and validate four different deep learning-based networks to segment three-dimensional (3D) prostate volumes on T2W MR images. This dataset consists of multi-center and multi-vendor transverse T2W MR images obtained with different acquisition protocols, field strengths and coils. Each of the trained networks was subsequently used to segment T2W MR images from the PROSTATEx challenges [18] (N = 346; seven cases excluded due to technical errors) and a dataset of in-house collected T2W MR images (N = 246), resulting in a combined dataset (N = 585). The combined dataset was shuffled and randomly split, in a controlled way, to ensure similar data distribution, into a training dataset (75%, N = 439) and a testing dataset (25%, N = 146) to respectively train/optimize and test the proposed QC system.

The in-house collected dataset was obtained from St. Olavs Hospital, Trondheim University Hospital, Trondheim, Norway between March 2015 and December 2017. It consists of pre-biopsy 3T images from 246 patients (median age = 65; range: 44–76 years). T2W imaging was performed on a Magnetom Skyra 3T MRI system (Siemens Healthineers, Erlangen, Germany) with a turbo spin-echo sequence (repetition time/echo time = 4450–9520/101–108 ms, 320 × 320–384 × 384 matrix size, 26–36 slices, 3 mm slice thickness and 0.5 × 0.5–0.6 × 0.6 mm^2 in plane resolution).

The Regional Committee for Medical and Health Research Ethics (REC Mid Norway) approved the use of the in-house collected dataset (identifier 2017/576; 5 May 2017) and granted permission for passive consent to be used. The two other datasets (PROMISE12 and PROSTATEx) were publicly available and details can be found in [16,18].

2.2. Prostate Segmentation

For each dataset, manual segmentations of the prostate gland without seminal vesicles were used as the gold standard. The PROMISE12 training dataset segmentations, used for training the segmentation algorithms, were publicly available [16]. The segmentation for the PROSTATEx dataset was performed by imaging experts with more than 25 years' combined expertise in prostate imaging and reviewed by radiation oncologists at Miller School of Medicine, Miami, FL, USA. The in-house collected dataset segmentation was performed by a radiology resident (E.S.) at St. Olavs Hospital, Trondheim University Hospital, Trondheim, Norway, under the supervision of a radiologist with more than 10 years' experience in prostate imaging. The manual segmentations of the PROSTATEx and in-house collected dataset were used to calculate the rQSs (see 2.3. Reference Quality Scores).

The deep learning-based prostate segmentation was performed using four different convolutional neural networks (CNNs), which are all variants of the famous U-Net with skip connections [15], here further referred to as U-Net [19], V-Net [10], nnU-Net-2D [11] and nnU-Net-3D [11]. U-Net and nnU-Net-2D perform the segmentation on a 2D slice-by-slice basis, whereas V-Net and nnU-Net-3D perform the segmentation on a 3D volume basis. Prior to segmentation, all images were pre-processed in accordance with the corresponding segmentation method. The segmentation pre-processing and the network training, validation and testing were performed on a single NVIDIA Tesla P100 PCIe 16 GB GPU in Ubuntu 16.04.6 LTS system. U-Net was implemented with the Keras API (version 2.3.0; https://keras.io) backboned with TensorFlow (version 1.9.0; https://www.tensorflow.org) using Python (version 2.7.12; Python Software Foundation, Wilmington, DE, USA). V-Net, nnU-Net-2D and nnU-Net-3D were implemented with PyTorch (version 1.4.0; https://pytorch.org) using Python (version 3.6.9).

2.3. Reference Quality Scores

To assess the true quality of the automated segmentations, rQSs were calculated in accordance with Litjens et al. [16]. Briefly, the rQS is a combination of the dice similarity coefficient (DSC) [20], the absolute relative volume difference (aRVD) [21], the 95% Hausdorff distance (95HD) [22] and the average symmetric surface distance (ASD) [21], separately obtained from the whole prostate, apex and base by comparing the automated segmentations with the manual segmentations (gold standard). Here, we defined the apex and base of the prostate to be the inferior and superior third parts of the mask-containing slices, respectively. However, before these 12 metrics can be combined in a single rQS, they need to be transformed to a common scale [16,21]. To do this, a second observer (M.R.S.S., three years of experience with prostate imaging) manually segmented 50 randomly selected cases from the combined dataset. These cases were used to calculate, for each metric, a linear function that maps the metric on a scale from 0 to 100, with the average performance of the second observer fixed at 85. These linear functions were subsequently applied to the 12 metrics calculated for each automated segmentation and the resulting 12 scores were averaged to obtain a single rQS for each segmentation. Details are provided in Appendix A.

2.4. Quality Control System

Figure 1 gives an overview of the proposed QC system. After preprocessing the T2W images, the LASSO model was trained and optimized on the training dataset, and tested on the independent testing dataset. All steps were implemented in MATLAB R2019b (MathWorks, Natick, MA, USA), except for the feature extraction which was performed using Python (version 3.7.3). The proposed system will be made available on GitHub at https://github.com/ntnu-mr-cancer/SegmentationQualityControl. Figure 2 shows an example of how the proposed QC system can be integrated in the image analysis pipeline.

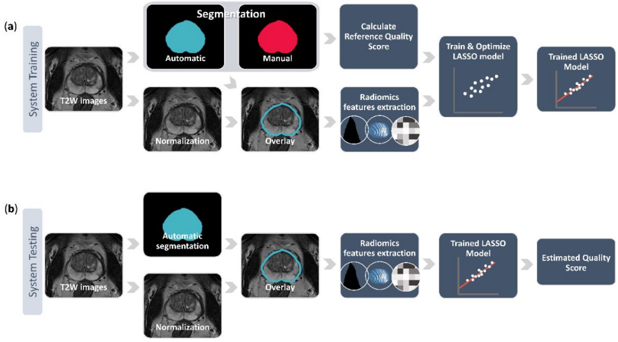
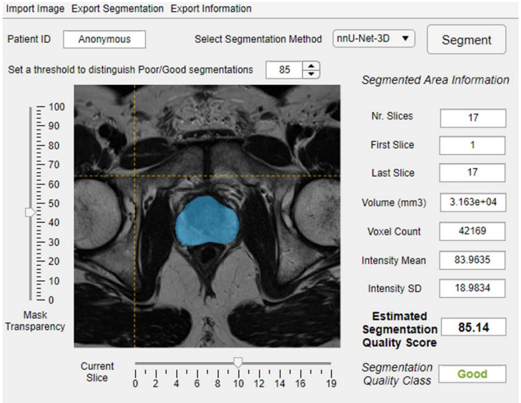


Figure 1. The pipeline of training (**a**) and testing (**b**) the proposed quality control system. The system training starts from the normalized T2-weighted (T2W) image stack with its corresponding manual prostate segmentation and automated segmentation delivered by a deep learning-based segmentation method. These two segmentations are used to calculate the reference quality score and the automated segmentation is also overlaid on the normalized image stack to extract various radiomics features. The reference quality score and the features are then fed into a least absolute shrinkage and selection operator (LASSO) to train and optimize a linear regression model that predicts the quality scores based on the imaging features. During the system testing, the trained model uses the radiomics features extracted from the overlaid automated segmentation on the normalized image stack to estimate a quality score for a previously unseen case.



Patient ID: Anonymous

Select Segmentation Method: nnU-Net-3D ▾

Set a threshold to distinguish Poor/Good segmentations: 85

Segmented Area Information

Nr. Slices	17
First Slice	1
Last Slice	17
Volume (mm3)	3.163e+04
Voxel Count	42169
Intensity Mean	83.9635
Intensity SD	18.9834
Estimated Segmentation Quality Score	**85.14**
Segmentation Quality Class	Good

Figure 2. An example of integrating the proposed quality control system within image analysis software.

2.4.1. Data Preparation

All T2W images were N4 bias field corrected [23] and intensity normalized using the AutoRef method [24]. In an attempt to develop a generalizable QC model, the segmentations generated by the four CNNs were combined in one dataset, producing a system training dataset of n = 1756 images (439 images from each CNN) and a system testing dataset of n = 584 images (146 images from each CNN) with corresponding segmentations. The dataset was split on the patient level, so all four segmentations belonging to one patient ended up in either the system training dataset or the system testing dataset.

Feature extraction from the preprocessed T2W images was performed using the automated prostate segmentations as the region of interest. All the features were extracted using Pyradiomics (version 2.2.0; an open-source Python package) [25]. Discretization of image intensity was performed using the fixed bin size approach, as recommended by Pyradiomics. The bin width was set to 64 in correspondence with the relatively large volume of interest. The features (N = 107) consisted of first-order features (N = 18), shape features (N = 14, performed on prostate 3D volume) and texture features (N = 75, 24 features from the gray level co-occurrence matrix (GLCM; in 3D along 13 directions (26-connectivity) and 1 pixel distance) [26], 16 features from the gray level run length matrix (GLRLM; in 3D along 13 directions) [27], 16 features from the gray level size zone matrix (GLSZM; in 3D along 13 directions) [28], 14 features from the gray level dependence matrix (GLDM; 1 pixel distance) [29] and 5 features from the neighboring gray tone difference matrix (NGTDM; 1 pixel distance) [30]). The average of the GLCM, GLRLM and GLSZM features across the direction was used. A complete list of the extracted features is given in Table S1. The features were extracted from the 3D volume of the whole prostate, apex and base parts of the prostate, separately, giving a total of 321 features per case.

2.4.2. Model Training, Optimizing and Testing

A least absolute shrinkage and selection operator (LASSO) [17] was used to build a linear regression model. The model was trained using the extracted features (N = 321) as predictors and the rQSs as responses. The LASSO, by nature, performs feature selection to enhance the model accuracy and interpretability [31]. How many features are selected depends on the regularization parameter lambda, which needs to be optimized. We employed a 5-fold cross-validation scheme to find the optimal lambda, here defined as the model returning the lowest mean squared errors between the eQS and rQS while satisfying a non-biased distribution as visualized by Bland–Altman plots [32].

The optimized model was tested and evaluated on the system testing dataset, returning an eQS for each segmentation based on features extracted from the deep learning-based prostate segmentation in the T2W MR image. If the returned eQS was > 100 it was set to 100 and if it was < 0 it was set to 0. The mean absolute error (MAE) and Spearman's rank test between eQSs and rQSs were used to evaluate the performance of the QC system. This was done on all the cases of the system testing dataset (General model), as well as separately for each of the eight combinations of dataset and segmentation method (sub-results from the General model; e.g., PROSTATEx—U-Net). The sub-results from the General model were also compared to the performance of (non-generalizable) models specifically trained on each combination of dataset and segmentation method. The manual and automated segmentations belonging to outliers of the tested General model were visually inspected by a researcher with three years of experience with prostate imaging (M.R.S.S.).

3. Results

3.1. Reference Quality Scores

The rQSs of the system training and testing dataset segmentations are presented in Figure 3. The maximum, mean ± standard deviation and minimum rQS of the combined dataset were 98.65, 82.26 ± 12.19 and 34.24, respectively, for the system training dataset and 98.95, 82.51 ± 12.22 and 26.24, respectively, for the system testing dataset. Figure 3 shows that the distribution of rQSs varies both between datasets and among the segmentation methods, indicating that the performance of automated prostate segmentation depends on both the dataset and the method used.

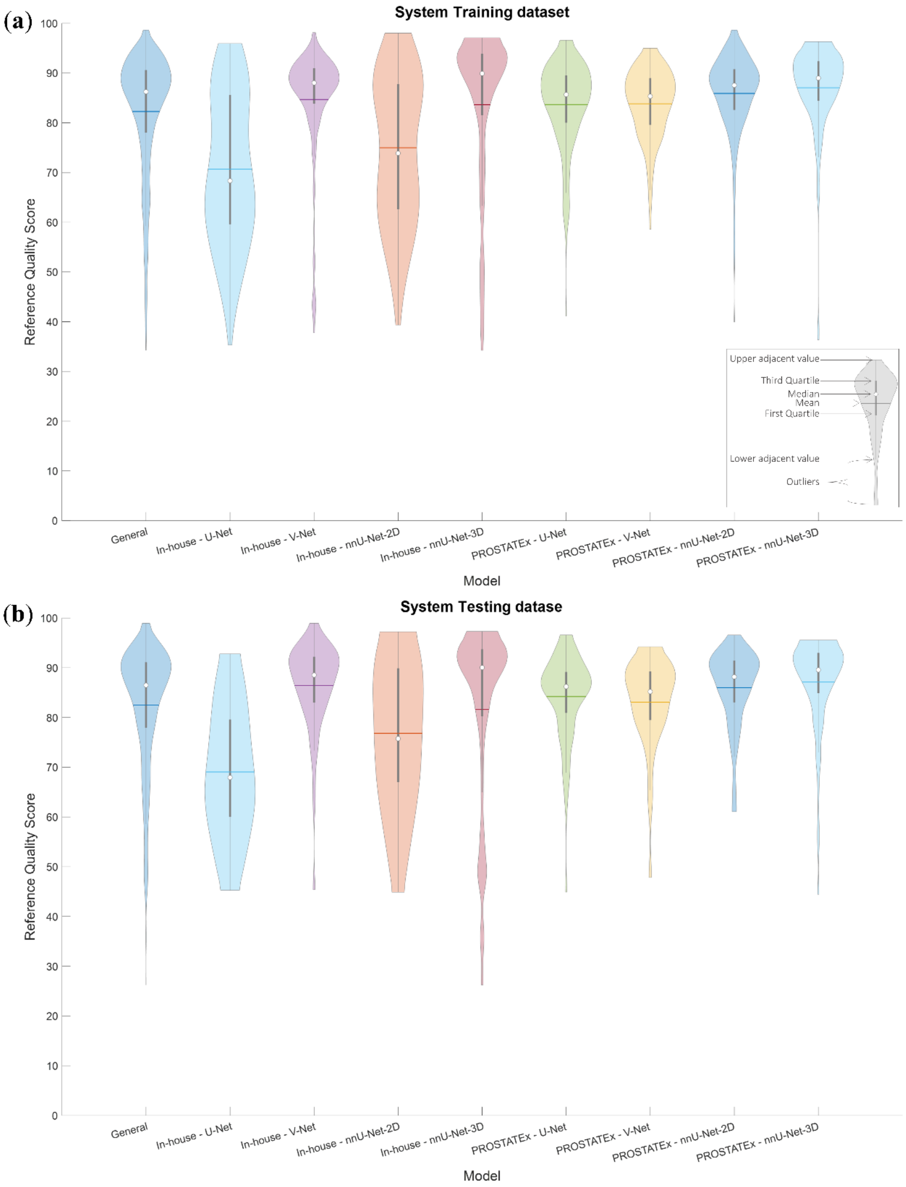


Figure 3. Violin plots visualizing the distribution of the reference quality scores for the system training (**a**) and testing (**b**) datasets, both combined and for each combination of dataset and segmentation method.

3.2. Training and Optimization

The maximum, mean ± standard deviation and minimum eQS of the General model were 98.04, 82.26 ± 10.71 and 28.06, respectively, for the system training dataset.

The optimal lambda was found to be 0.01, which resulted in the selection of 142 out of 321 radiomics features in the trained General LASSO model. Figure 4 shows the distribution of the selected features. Overall, 46.30%, 76.19%, 45.83%, 20.83%, 41.67%, 33.34% and 66.67% of the extracted first order, shape, GLCM, GLRLM, GLSZM, GLDM and NGTDM features were selected, respectively. Further details of the trained model are provided in Table S2. The details of the eight non-generalizable models are provided in Tables S3–S10.

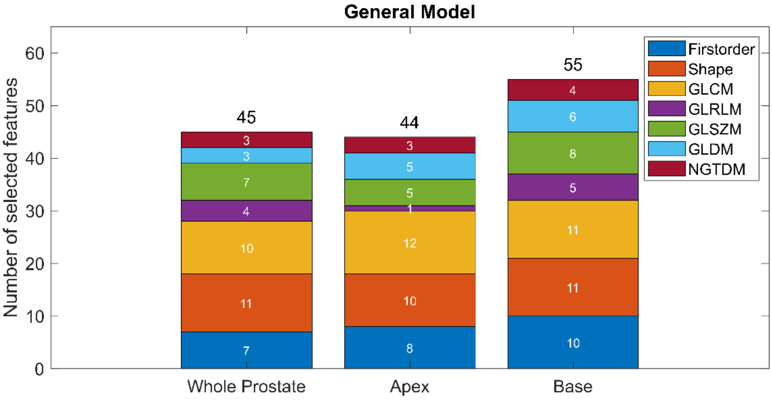


Figure 4. The distribution of the selected features in the optimized General model.

Figure 5 shows the overlap between the selected features in the PROSTATEx and in-house datasets of non-generalizable models trained on data processed with the same segmentation method. To account for the high co-linearity between features, overlap was defined as the selection of the same feature or a highly correlated feature (rho > 0.9). For each segmentation method, we found a high number of overlapping features (directly or highly correlated), indicating that the models extracted similar features irrespective of dataset.

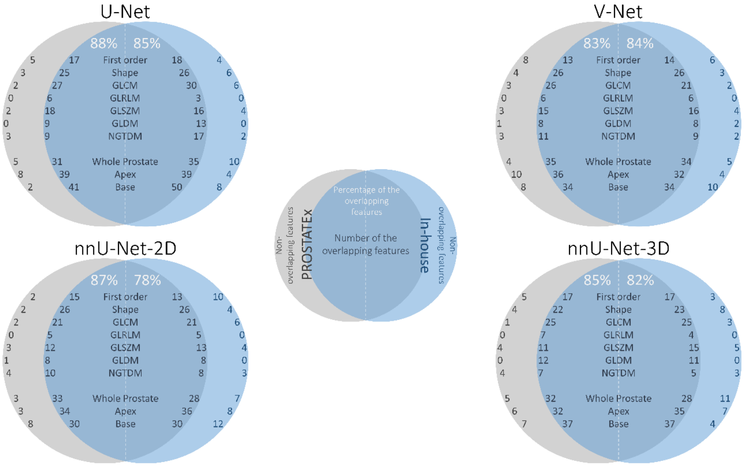


Figure 5. The overlap between the features in the PROSTATEx (gray) and in-house (blue) datasets of the same segmentation method (e.g., overlap between the "PROSTATEx—U-Net model" and "In-house—U-Net model"). The intersection area presents the overlapping features, whereas the areas out of the intersection present the set of features unique to each dataset.

3.3. Testing

For the system testing dataset, the maximum, mean ± standard deviation and minimum eQS of the General model were 97.60, 82.03 ± 11.02 and 0.00, respectively.

The performance of the tested models is presented in Table 1. Table 2 presents sub-results from the tested General model, for direct comparison with the non-generalizable models. Sub-results from the General model resulted in lower MAE in 7/8 cases than their non-generalizable counterparts, indicating that the overall performance of the General model is better than the non-generalizable models. Nevertheless, it should be noted that the sub-results vary considerably. This is especially apparent from the difference in slope (ideally 1), intercept (ideally 0) and rho (ideally 1) between results from the PROSTATEx and in-house datasets.

Table 1. The performance evaluation of the separately tested models.

Model	N	MAE ± SD	IQR	Slope	Intercept	Rho	Correlation p-Value
General	584	5.37 ± 11.02	9.32	0.72	22.40	0.70	<0.001
PROSTATEx—U-Net	89	5.48 ± 9.04	7.20	0.67	27.83	0.49	<0.001
PROSTATEx—V-Net	89	5.91 ± 8.21	6.80	0.40	50.43	0.43	<0.001
PROSTATEx—nnU-Net-2D	89	5.14 ± 6.04	5.96	0.40	51.25	0.41	<0.001
PROSTATEx—nnU-Net-3D	89	5.89 ± 7.79	5.64	0.47	44.97	0.40	<0.001
In-house—U-Net	57	9.55 ± 17.24	22.95	0.86	7.92	0.70	<0.001
In-house—V-Net	57	6.58 ± 13.01	12.33	1.07	−9.55	0.55	<0.001
In-house—nnU-Net-2D	57	8.18 ± 14.2	21.26	0.71	21.99	0.67	<0.001
In-house—nnU-Net-3D	57	8.35 ± 19.02	14.78	0.75	20.73	0.48	<0.001

N: Number of segmentations; MAE: Mean absolute error; SD: Standard deviation of the absolute error; IQR: Interquartile range.

Table 2. Sub-results from the tested General model performance evaluation.

Sub-Results Combination	N	MAE ± SD	IQR	Slope	Intercept	Rho	Correlation *p*-Value
PROSTATEx—U-Net	89	5.24 ± 5.28	6.20	0.36	52.69	0.50	<0.001
PROSTATEx—V-Net	89	5.50 ± 4.67	5.33	0.27	61.28	0.38	<0.001
PROSTATEx—nnU-Net-2D	89	5.41 ± 4.46	5.37	0.26	62.80	0.43	<0.001
PROSTATEx—nnU-Net-3D	89	4.85 ± 5.76	6.12	0.35	57.17	0.50	<0.001
In-house—U-Net	57	7.27 ± 12.61	19.84	0.73	17.59	0.76	<0.001
In-house—V-Net	57	4.39 ± 6.64	6.47	0.59	34.65	0.70	<0.001
In-house—nnU-Net-2D	57	4.84 ± 12.4	17.78	0.78	16.90	0.87	<0.001
In-house—nnU-Net-3D	57	5.76 ± 20.79	10.17	1.02	−3.50	0.74	<0.001

N: Number of segmentations; MAE: Mean absolute error; SD: Standard deviation of the absolute error; IQR: Interquartile range.

Figure 6a shows the linear fit of the eQSs for the General model with examples of segmentations. The segmentations of the cases outside of the 95% prediction interval were visually inspected. We subjectively judged the eQS to be extremely overestimated in 2/9 segmentations that were over the 95% prediction interval, and extremely underestimated in 3/18 segmentations that were under the 95% prediction interval. The rest of the visually inspected segmentations were judged to have an eQS that acceptably represented the quality of the automated segmentation. All of the segmentations over the 95% prediction interval belonged to the PROSTATEx dataset, and all of the segmentations under the 95% prediction interval belonged to the in-house dataset. Interestingly, in 8/27 segmentations, the discrepancy between the eQS and rQS was judged to result from a sub-optimal manual segmentation. Examples of over- and underestimated segmentations are shown in Figure 6a. Figure 6b shows the difference between the eQSs and rQSs of the General model. The mean difference was −0.48, with a tendency for overestimating cases with a low rQS and underestimating cases with a high rQS.

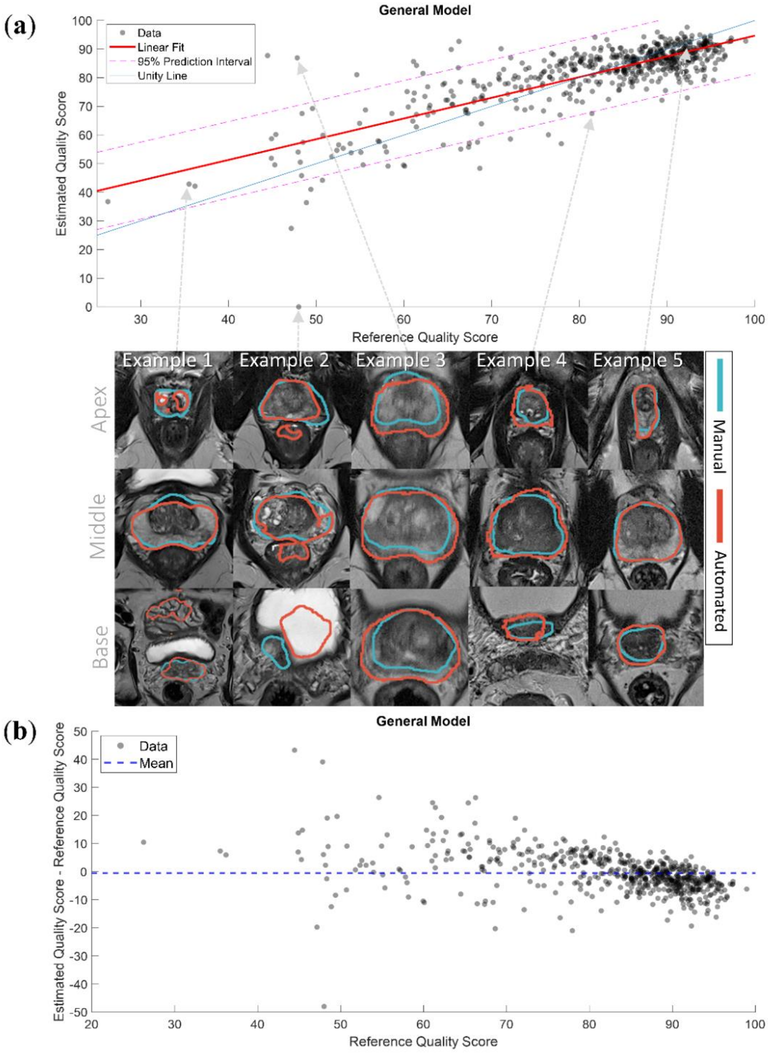


Figure 6. (**a**) The linear fit of the estimated quality scores with 95% prediction interval of the General model with examples of segmentations. Example 1 presents a case where the model accurately predicted the quality score (QS) of a low-quality automated segmentation; Example 2 presents a case where the model extremely underestimated the QS of a low-quality automated segmentation, the automated segmentation here covered parts of the rectum and the bladder; Example 3 presents a case where the model extremely overestimated the QS of a low-quality automated segmentation, the manual segmentation here misses the peripheral zone; Example 4 presents a case where the model slightly underestimated the QS of a high-quality automated segmentation, the automated segmentation here was slightly over segmented; Example 5 presents a case where the model accurately predicted the QS of a high-quality automated segmentation; (**b**) the difference between the estimated and reference quality scores of the General model.

The linear fits of the eight non-generalizable models and the sub-results from the General model are presented in Figure 7. It can be appreciated that the slopes and intercepts of the models/sub-results associated with the in-house dataset were better than those associated with the PROSTATEx dataset.

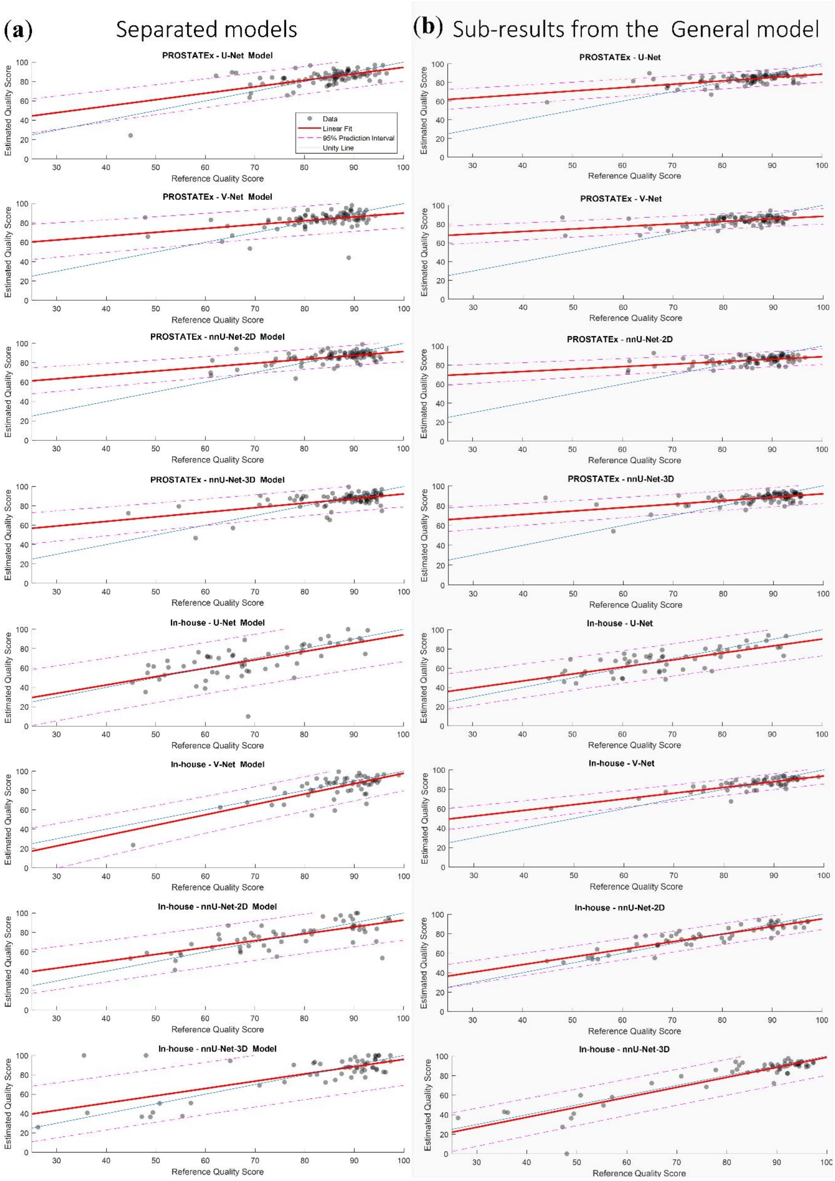


Figure 7. The linear fit of the estimated quality scores with 95% prediction interval of the eight non-generalizable models (**a**) and the sub-results from the General model (**b**).

4. Discussion

Automated segmentation of the prostate is a crucial step in the CAD of prostate cancer, but quality control and possibly adjustment by a trained radiologist is still required. In this work, we present a fully automated QC system that aims to present the user with an estimated score indicative of the segmentation quality. This system could function as a safety net that saves time and costs, standardizes the decision about the segmentation accuracy and thus facilitate the clinical implementation of automated prostate segmentation algorithms. The system could be specifically useful for clinical applications that are sensitive to errors in segmentation, such as MRI–ultrasound fusion for targeted prostate biopsies, which is currently becoming a clinical standard procedure [33], and prostate-targeted MR-guided radiotherapy, which has been implemented in the treatment of prostate cancer patients during the last few years [34].

Our results indicate that the proposed QC system could be helpful for this purpose. Overall, the General model had better performance than the non-generalizable models. We found a strong correlation between the rQSs and eQSs (rho = 0.70) and MAE values less than the standard deviation between the experts and the second observer segmentations (5.37 vs. 7.76), implying that errors were in an acceptable range. In addition, the mean of the differences between the eQSs and rQSs was low (mean = −0.48). Despite the overall good performance, some of the eQSs of the segmentations were over- or underestimated. This can be partly explained by the fact that the rQSs, used as input for training the model, were imbalanced and skewed towards high scores. This probably had an effect on the model performance, leading to a higher number of over- and underestimated segmentations around the low rQSs. Indeed, the non-generalizable models that had the most balanced distribution of rQSs in the training dataset (e.g., "In-house—U-Net" and "In-house—nnU-Net_2D") performed better than the other models.

T2W MRI clearly depicts the borders and anatomy of the prostate gland, and thus constitutes an excellent starting point for both prostate segmentation algorithms and the proposed QC system. In this work, we implemented four deep learning-based segmentation methods using two different T2W MRI datasets. Combining these datasets made the proposed system more generalizable and robust, and it is thus potentially applicable to other segmentation methods and datasets. However, we also showed that the model did not perform well for all combinations of dataset and segmentation method. Consequently, the proposed QC system should be carefully tested and evaluated on new data and methods before application.

The first-order, shape and texture features were investigated because they describe distinct characteristics of the volume of interest. Our QC system was trained to find common features and assess the segmentation quality among the investigated cases. We selected the LASSO model due to its model interpretability advantage [31] and its good performance in multiple radiomics studies [35–37]. To calculate the rQS, we chose to use the established PROMISE12 challenge evaluation metric [16] as it imparts a comprehensive overview of the segmentation accuracy, and shows interest in the prostate apex and base segmentations, which are the most difficult parts of the prostate gland to segment. It is paramount to segment these two sections correctly in some of the clinical applications and procedures, e.g., in MRI–ultrasound fusion for targeted prostate biopsy [33]. Similar to Litjens et al. [16], the average performance of the second observer was fixed at 85 during the rQS calculation due to the relatively good correspondence between the second observer segmentations and the gold standard.

To develop a flexible system, we chose to train a regression model instead of a classifier. A classifier would require a fixed threshold to distinguish the good and poor rQSs, which is challenging and depends on the targeted clinical application. Moreover, a fixed threshold may restrict the system's generalizability. Depending on the desired application and corresponding acceptable segmentation error margin, a threshold can be set to distinguish poor from acceptable segmentations (e.g., Figure 2). The QC system could thus save time for radiologists, as many of the segmentations can be used without further manual verification; the total computational time to generate an automated segmentation and corresponding eQS was less than one minute per case on the described computing system, which is drastically less than the time required by a radiologist to do the same task. In addition, the system could standardize the decision about the segmentation accuracy and build confidence using the deep learning-based algorithms.

All the different types of features available in Pyradiomics were used in the trained General model. However, shape features were found to be the most important, since approximately 76% of them were selected. This finding is in accordance with the way the CNNs work, gradually moving from shape-based to texture-based features through the layers [38]. Interestingly, compared to features extracted from the apex and the whole prostate, a higher number of features was selected from the base of the prostate for the model training. This potentially reflects how difficult it is to segment—both manually and automatically—the prostate base due to the variability between patients [16].

It could be noticed from Figure 3 that the rQS distributions are wider and more balanced in the case of the in-house dataset. This has a positive effect on the performance of the non-generalizable models as well as the General model's sub-results associated with the in-house dataset. This is especially noticeable from the low eQSs, which are closer to the unity line than those associated with the PROSTATEx dataset. The high overlap between the features of the non-generalizable models trained on the PROSTATEx and the in-house datasets indicates that the performance difference is probably due to the input data, and not caused by differences in the selected dataset features. For future work, the model performance could potentially be enhanced by increasing the number of low rQSs by stopping the CNN training early, i.e., before finishing the recommended number of iterations.

Despite the acceptable performance of the QC system General model, there were some outliers, here defined as the eQSs outside the 95% prediction interval limits. Visual inspection revealed that the eQSs of the segmentations actually accurately represented the quality of the automated segmentations for most of these over- and underestimated cases. It was found that 8/9 segmentations over the 95% prediction interval outliers belonged to the four CNN segmentations of two patients. The manual segmentation of one of these patients was missing the contour in some of the slices in the apex and the base, and not properly covering the peripheral zone in the middle part of the prostate gland (see Example 3 in Figure 6a). The manual segmentation of the other patient did not include the peripheral zone in all slices from base to middle prostate. The automated segmentations associated with an underestimated eQS included in many cases small areas outside of the regions of interest (see Example 2 in Figure 6a). The visual inspection also revealed that all of the overestimated cases belonged to the PROSTATEx dataset and all of the underestimated cases belonged to the in-house dataset, which might be explained by the distribution of the rQSs used in training the system.

In this study, we propose a QC system that estimates the quality of automated prostate segmentations based on the shape of the segmentation mask, and the histogram intensity and texture of the underlying T2W image. Another interesting approach, which requires an additional step, was recently proposed by Valindria et al. [39]. In their reverse classification accuracy method, a segmentation model was built from the segmentation mask and corresponding image of a single new case (lacking ground truth). Subsequently, this model was applied to all images of a database with corresponding expert segmentations. Under the assumption that the same segmentation model should work for at least one of these images, the best segmentation accuracy (DSC) is assumed to reflect the accuracy of the newly segmented case. Robinson et al. [40] showed that this approach works well for QC of segmentation of the heart in cardiovascular MRI. Yet another interesting approach was recently presented by Roy et al. [41], in which a structure-wise uncertainty estimate was intrinsically included in a CNN algorithm for brain segmentation on T1 MR images. This approach keeps the drop-out layers of the CNN active during test time, to produce multiple segmentation variants from which uncertainty measures can be calculated. One disadvantage compared to our system is that Roy et al.'s approach cannot be easily generalized to other segmentation methods. Moreover, unlike our system, both of the aforementioned approaches used only volume-based segmentation accuracy metrics and did not take boundary-based metrics into consideration. To the best of our knowledge, these methods have not yet been tested for prostate segmentation. Although they are more complex, it will be interesting to compare them with our system in future work.

Our study has limitations. The number of cases with a low rQS was relatively small; a more balanced dataset would probably have led to a more robust system over all datasets and segmentation methods and would have given better insight into the system's potential. In addition, there are other radiomics features such as wavelet transformation-based texture features, which were not include in our model. These features could potentially enhance the performance of the system at the cost of generating a more complex model and expanding the computational time. For these reasons, they have not been used in this study, but their additional value will be investigated in future work.

5. Conclusions

We propose a QC system for estimating the quality of automated segmentation of the prostate in T2W MR images, which could be an important step towards the clinical implementation of computer-aided detection and diagnosis of prostate cancer. The performance of the generalizable model is acceptable in regard to estimating the segmentation quality scores, but varies between datasets and segmentation methods. The system is transparent and could save considerable time and standardize decision-making in clinical practice, albeit careful implementation and testing is required.

Supplementary Materials: The following are available online at http://www.mdpi.com/2075-4418/10/9/714/s1, Table S1: List of the extracted radiomics features. The shape features were extracted from the three-dimensional volume, Table S2: The trained General model intercept and coefficients, Table S3–S10: The trained non-generalizable models' intercepts and coefficients.

Author Contributions: Conceptualization, M.R.S.S., T.F.B. and M.E.; Data curation, M.R.S.S., K.M.S., E.S., G.A.N., O.Z.-R. and R.S.; Formal analysis, M.R.S.S. and M.E.; Investigation, M.R.S.S., K.M.S., G.A.N., T.F.B. and M.E.; Methodology, M.R.S.S., O.Z.-R., R.S., T.F.B. and M.E.; Project administration, T.F.B. and M.E.; Supervision, K.M.S., T.F.B. and M.E.; Writing—original draft, M.R.S.S.; Writing—review and editing, M.R.S.S., K.M.S., E.S., G.A.N., O.Z.-R., R.S., T.F.B. and M.E. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Norwegian University of Science and Technology (NTNU) Biotechnology (grant number 81770928; M.R.S.S.), the Research Council of Norway (grant number 295013; T.F.B.) and the liaison committee between the Central Norway Regional Health Authority and the Norwegian University of Science and Technology (grant numbers 90368401; G.A.N. and 90265300; M.E.).

Acknowledgments: We would like to thank SPIE, the AAPM, the NCI and Radboud University Nijmegen (Nijmegen, Netherlands) in addition to the organizers of the PROSTATEx and PROMISE12 challenges for making their datasets available. In addition, we would like to thank Inom Mirzaev from Ohio State University (Ohio, USA), Fausto Milletari from the Technical University of Munich (Munich, Germany) and Fabian Isensee from the German Cancer Research Center (Heidelberg, Germany) for making their segmentation methods publicly available. We would like to thank Sverre Langørgen from St. Olavs Hospital, Trondheim University Hospital (Trondheim, Norway), for supervising the prostate delineation process for the in-house collected dataset.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Appendix A

The metrics we used to calculate the rQSs, as mentioned in Section 2.3. Reference Quality Scores, are defined in Equations (A1), (A2), (A4) and (A5). Equation (A3) was required to enable the calculation of Equations (A4) and (A5). The linear mapping function that we used is defined in Equation (A6). All calculations are in accordance with Litjens et al. [16].

$$\mathrm{DSC}\ (X,Y) = 2|X \cap Y|/(|X| + |Y|) \tag{A1}$$

where X and Y represent the reference and automated segmentation voxels, respectively.

$$\mathrm{aRVD}\ (X,Y) = |(Y/X - 1) \times 100| \tag{A2}$$

to calculate 95HD and ASD, the Euclidean distance of the surface point sets (d_H) from the reference (Xs) and automated (Ys) segmentations was measured using Equation (A3):

$$d_H\ (Xs, Ys) = \max_{x \in X_s}(\min_{y \in Y_s} d(x,y)) \tag{A3}$$

where d is the Euclidean distance operator.

$$95\mathrm{HD} = \max\ (P_{95}\ (d_H\ (Xs,Ys)), P_{95}\ (d_H\ (Ys,Xs))) \tag{A4}$$

where P_{95} represents the 95th percentile of d_H.

$$\mathrm{ASD}\ (\mathrm{Xs}, \mathrm{Ys}) = \left(\sum_{x \in X_s} \min_{y \in Y_s} d(x, y) + \sum_{y \in Y_s} \min_{x \in X_s} d(y, x)\right) / \left(N_{Xs} + N_{Ys}\right) \quad (A5)$$

where N_{Xs} and N_{Ys} represent the number of surface points of the reference and automated segmentations, respectively.

$$\text{metric score (Z)} = \max\ (aZ + b, 0) \quad (A6)$$

where Z is the average unmapped metric value. The variables a and b were determined by solving two equations through setting the metric score to equal 85, representing the average performance of the second observer, and a perfect score to equal 100.

References

1. Bray, F.; Ferlay, J.; Soerjomataram, I.; Siegel, R.L.; Torre, L.A.; Jemal, A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **2018**, *68*, 394–424. [CrossRef] [PubMed]
2. Siegel, R.; Ma, J.; Zou, Z.; Jemal, A. Cancer statistics, 2014. *CA Cancer J. Clin.* **2014**, *64*, 9–29. [CrossRef] [PubMed]
3. Barentsz, J.O.; Richenberg, J.; Clements, R.; Choyke, P.; Verma, S.; Villeirs, G.; Rouviere, O.; Logager, V.; Futterer, J.J. ESUR prostate MR guidelines 2012. *Eur. Radiol.* **2012**, *22*, 746–757. [CrossRef]
4. Weinreb, J.C.; Barentsz, J.O.; Choyke, P.L.; Cornud, F.; Haider, M.A.; Macura, K.J.; Margolis, D.; Schnall, M.D.; Shtern, F.; Tempany, C.M.; et al. PI-RADS Prostate Imaging-Reporting and Data System: 2015, Version 2. *Eur. Urol.* **2016**, *69*, 16–40. [CrossRef] [PubMed]
5. Ruprecht, O.; Weisser, P.; Bodelle, B.; Ackermann, H.; Vogl, T.J. MRI of the prostate: Interobserver agreement compared with histopathologic outcome after radical prostatectomy. *Eur. J. Radiol.* **2012**, *81*, 456–460. [CrossRef] [PubMed]
6. Litjens, G.; Debats, O.; Barentsz, J.; Karssemeijer, N.; Huisman, H. Computer-aided detection of prostate cancer in MRI. *IEEE Trans. Med. Imaging* **2014**, *33*, 1083–1092. [CrossRef]
7. Wang, S.; Burtt, K.; Turkbey, B.; Choyke, P.; Summers, R.M. Computer aided-diagnosis of prostate cancer on multiparametric MRI: A technical review of current research. *BioMed Res. Int.* **2014**, *2014*, 789561. [CrossRef]
8. Hambrock, T.; Vos, P.C.; Hulsbergen-van de Kaa, C.A.; Barentsz, J.O.; Huisman, H.J. Prostate cancer: Computer-aided diagnosis with multiparametric 3-T MR imaging—Effect on observer performance. *Radiology* **2013**, *266*, 521–530. [CrossRef]
9. Lemaitre, G.; Marti, R.; Freixenet, J.; Vilanova, J.C.; Walker, P.M.; Meriaudeau, F. Computer-Aided Detection and diagnosis for prostate cancer based on mono and multi-parametric MRI: A review. *Comput. Biol. Med.* **2015**, *60*, 8–31. [CrossRef]
10. Milletari, F.; Navab, N.; Ahmadi, S.A. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In Proceedings of the 4th International Conference on 3D Vision (3DV), Stanford, CA, USA, 25–28 October 2016; pp. 565–571.
11. Isensee, F.; Jäger, P.F.; Kohl, S.A.A.; Petersen, J.; Maier-Hein, K.H. Automated Design of Deep Learning Methods for Biomedical Image Segmentation. *arXiv* **2019**, arXiv:1904.08128.
12. Zavala-Romero, O.; Breto, A.L.; Xu, I.S.R.; Chang, Y.C.C.; Gautney, N.; Pra, A.D.; Abramowitz, M.C.; Pollack, A.; Stoyanova, R. Segmentation of prostate and prostate zones using deep learning A multi-MRI vendor analysis. *Strahlenther. Onkol.* **2020**. [CrossRef]
13. Wang, B.; Lei, Y.; Tian, S.; Wang, T.; Liu, Y.; Patel, P.; Jani, A.B.; Mao, H.; Curran, W.J.; Liu, T.; et al. Deeply supervised 3D fully convolutional networks with group dilated convolution for automatic MRI prostate segmentation. *Med. Phys.* **2019**, *46*, 1707–1718. [CrossRef] [PubMed]
14. Yan, K.; Wang, X.; Kim, J.; Khadra, M.; Fulham, M.; Feng, D. A propagation-DNN: Deep combination learning of multi-level features for MR prostate segmentation. *Comput. Methods Programs Biomed.* **2019**, *170*, 11–21. [CrossRef] [PubMed]

15. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. Medical Image Computing and Computer-Assisted Intervention. In *International Conference on Medical image Computing and Computer-Assisted Intervention*; Springer: Cham, Switzerland, 2015; pp. 234–241.
16. Litjens, G.; Toth, R.; van de Ven, W.; Hoeks, C.; Kerkstra, S.; van Ginneken, B.; Vincent, G.; Guillard, G.; Birbeck, N.; Zhang, J.; et al. Evaluation of prostate segmentation algorithms for MRI: The PROMISE12 challenge. *Med. Image Anal.* **2014**, *18*, 359–373. [CrossRef] [PubMed]
17. Tibshirani, R. Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. B* **1996**, *58*, 267–288. [CrossRef]
18. Armato, S.G., 3rd; Huisman, H.; Drukker, K.; Hadjiiski, L.; Kirby, J.S.; Petrick, N.; Redmond, G.; Giger, M.L.; Cha, K.; Mamonov, A.; et al. PROSTATEx Challenges for computerized classification of prostate lesions from multiparametric magnetic resonance images. *J. Med. Imaging* **2018**, *5*, 044501. [CrossRef] [PubMed]
19. Mirzaev, I. Fully Convolutional Neural Network with Residual Connections for Automatic Segmentation of Prostate Structures from MR Images. Available online: https://grand-challenge-public.s3.amazonaws.com/evaluation-supplementary/40/d70ba7d1-bc95-439e-a81e-7f1a4ed5fda0/18_MBIOS.pdf (accessed on 28 August 2020).
20. Klein, S.; van der Heide, U.A.; Lips, I.M.; van Vulpen, M.; Staring, M.; Pluim, J.P.W. Automatic segmentation of the prostate in 3D MR images by atlas matching using localized mutual information. *Med. Phys.* **2008**, *35*, 1407–1417. [CrossRef] [PubMed]
21. Heimann, T.; van Ginneken, B.; Styner, M.A.; Arzhaeva, Y.; Aurich, V.; Bauer, C.; Beck, A.; Becker, C.; Beichel, R.; Bekes, G.; et al. Comparison and evaluation of methods for liver segmentation from CT datasets. *IEEE Trans. Med. Imaging* **2009**, *28*, 1251–1265. [CrossRef] [PubMed]
22. Chandra, S.S.; Dowling, J.A.; Shen, K.K.; Raniga, P.; Pluim, J.P.; Greer, P.B.; Salvado, O.; Fripp, J. Patient specific prostate segmentation in 3-d magnetic resonance images. *IEEE Trans. Med. Imaging* **2012**, *31*, 1955–1964. [CrossRef] [PubMed]
23. Tustison, N.J.; Avants, B.B.; Cook, P.A.; Zheng, Y.; Egan, A.; Yushkevich, P.A.; Gee, J.C. N4ITK: Improved N3 bias correction. *IEEE Trans. Med. Imaging* **2010**, *29*, 1310–1320. [CrossRef]
24. Sunoqrot, M.R.S.; Nketiah, G.A.; Selnaes, K.M.; Bathen, T.F.; Elschot, M. Automated reference tissue normalization of T2-weighted MR images of the prostate using object recognition. *Magn. Reson. Mater. Phys. Biol. Med.* **2020**. [CrossRef] [PubMed]
25. Van Griethuysen, J.J.M.; Fedorov, A.; Parmar, C.; Hosny, A.; Aucoin, N.; Narayan, V.; Beets-Tan, R.G.H.; Fillion-Robin, J.C.; Pieper, S.; Aerts, H.J.W.L. Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer Res.* **2017**, *77*, E104–E107. [CrossRef] [PubMed]
26. Haralick, R.M.; Shanmugam, K.; Dinstein, I. Textural Features for Image Classification. *IEEE Trans. Syst. Man Cybern.* **1973**, *SMC-3*, 610–621. [CrossRef]
27. Chu, A.; Sehgal, C.M.; Greenleaf, J.F. Use of Gray Value Distribution of Run Lengths for Texture Analysis. *Pattern Recogn. Lett.* **1990**, *11*, 415–419. [CrossRef]
28. Thibault, G.; Fertil, B.; Navarro, C.; Pereira, S.; Cau, P.; Levy, N.; Sequeira, J.; Mari, J. Texture Indexes and Gray Level Size Zone Matrix. Application to Cell Nuclei Classification. In Proceedings of the 10th International Conference on Pattern Recognition and Information Processing, Minsk, Belarus, Minsk, Belarus, 19–21 May 2009; pp. 140–145.
29. Sun, C.J.; Wee, W.G. Neighboring Gray Level Dependence Matrix for Texture Classification. *Comput. Vis. Graph. Image Process.* **1983**, *23*, 341–352. [CrossRef]
30. Amadasun, M.; King, R. Textural Features Corresponding to Textural Properties. *IEEE Trans. Syst. Man Cybern.* **1989**, *19*, 1264–1274. [CrossRef]
31. Friedman, J.; Hastie, T.; Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.* **2010**, *33*, 1–22. [CrossRef] [PubMed]
32. Bland, J.M.; Altman, D.G. Statistical Methods for Assessing Agreement between Two Methods of Clinical Measurement. *Lancet* **1986**, *1*, 307–310. [CrossRef]
33. Jayadevan, R.; Zhou, S.; Priester, A.M.; Delfin, M.; Marks, L.S. Use of MRI-ultrasound Fusion to Achieve Targeted Prostate Biopsy. *J. Vis. Exp.* **2019**, *146*, e59231. [CrossRef] [PubMed]
34. Salembier, C.; Villeirs, G.; De Bari, B.; Hoskin, P.; Pieters, B.R.; Van Vulpen, M.; Khoo, V.; Henry, A.; Bossi, A.; De Meerleer, G.; et al. ESTRO ACROP consensus guideline on CT- and MRI-based target volume delineation for primary radiation therapy of localized prostate cancer. *Radiother. Oncol.* **2018**, *127*, 49–61. [CrossRef]

35. Ma, S.; Xie, H.H.; Wang, H.H.; Han, C.; Yang, J.J.; Lin, Z.Y.; Li, Y.F.; He, Q.; Wang, R.; Cui, Y.P.; et al. MRI-Based Radiomics Signature for the Preoperative Prediction of Extracapsular Extension of Prostate Cancer. *J. Magn. Reson. Imaging* **2019**, *50*, 1914–1925. [CrossRef] [PubMed]
36. Min, X.D.; Li, M.; Dong, D.; Feng, Z.Y.; Zhang, P.P.; Ke, Z.; You, H.J.; Han, F.F.; Ma, H.; Tian, J.; et al. Multi-parametric MRI-based radiomics signature for discriminating between clinically significant and insignificant prostate cancer: Cross-validation of a machine learning method. *Eur. J. Radiol.* **2019**, *115*, 16–21. [CrossRef] [PubMed]
37. Xu, L.; Zhang, G.; Zhao, L.; Mao, L.; Li, X.; Yan, W.; Xiao, Y.; Lei, J.; Sun, H.; Jin, Z. Radiomics Based on Multiparametric Magnetic Resonance Imaging to Predict Extraprostatic Extension of Prostate Cancer. *Front. Oncol.* **2020**, *10*, 940. [CrossRef] [PubMed]
38. Zeiler, M.D.; Fergus, R. Visualizing and Understanding Convolutional Networks. *Lect. Notes Comput. Sci.* **2014**, *8689*, 818–833.
39. Valindria, V.V.; Lavdas, I.; Bai, W.; Kamnitsas, K.; Aboagye, E.O.; Rockall, A.G.; Rueckert, D.; Glocker, B. Reverse Classification Accuracy: Predicting Segmentation Performance in the Absence of Ground Truth. *IEEE Trans. Med. Imaging* **2017**, *36*, 1597–1606. [CrossRef] [PubMed]
40. Robinson, R.; Valindria, V.V.; Bai, W.; Oktay, O.; Kainz, B.; Suzuki, H.; Sanghvi, M.M.; Aung, N.; Paiva, J.M.; Zemrak, F.; et al. Automated quality control in image segmentation: Application to the UK Biobank cardiovascular magnetic resonance imaging study. *J. Cardiovasc. Magn. Reson.* **2019**, *21*, 18. [CrossRef]
41. Roy, A.G.; Conjeti, S.; Navab, N.; Wachinger, C.; Alzheimer's Disease Neuroimaging Initiative. Bayesian QuickNAT: Model uncertainty in deep whole-brain segmentation for structure-wise quality control. *NeuroImage* **2019**, *195*, 11–22. [CrossRef]

© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).

Supplementary material to:

A Quality Control System for Automated Prostate Segmentation on T2-Weighted MRI

By: Mohammed R. S. Sunoqrot [1,*], Kirsten M. Selnæs [1,2], Elise Sandsmark [2], Gabriel A. Nketiah [1,2], Olmo Zavala-Romero [3,4], Radka Stoyanova [3], Tone F. Bathen [1,2,†] and Mattijs Elschot [1,2,†]

[1] Department of Circulation and Medical Imaging, NTNU—Norwegian University of Science and Technology, 7030 Trondheim, Norway.

[2] Department of Radiology and Nuclear Medicine, St. Olavs Hospital, Trondheim University Hospital, 7030 Trondheim, Norway.

[3] Department of Radiation Oncology, University of Miami Miller School of Medicine, Miami, FL 33136, USA.

[4] Center for Ocean-Atmospheric Prediction Studies, Florida State University, Tallahassee, FL 32306, USA.

[†] These authors contributed equally to this work.

*** Corresponding author:**

Name: Mohammed R. S. Sunoqrot

Email: mohammed.sunoqrot@ntnu.no

Table S1. List of the extracted Radiomics features. The shape features were extracted from the 3-dimensional volume.

First order	Shape
10Percentile	Elongation
90Percentile	Flatness
Energy	Least Axis Length
Entropy	Major Axis Length
Interquartile Range	Maximum 2D Diameter Column
Kurtosis	Maximum 2D DiameterRow
Maximum	Maximum 2D DiameterSlice
Mean Absolute Deviation	Maximum 3D Diameter
Mean	Mesh Volume
Median	Minor Axis Length
Minimum	Sphericity
Range	Surface Area
Robust Mean Absolute Deviation	Surface Volume Ratio
Root Mean Squared	Voxel Volume
Skewness	
Total Energy	
Uniformity	
Variance	

GLCM	GLRLM
Autocorrelation	Gray Level Non Uniformity
Cluster Prominence	Gray Level Non Uniformity Normalized
Cluster Shade	Gray Level Variance
Cluster Tendency	High Gray Level Run Emphasis
Contrast	Long Run Emphasis
Correlation	Long Run High Gray Level Emphasis
Difference Average	Long Run Low Gray Level Emphasis
Difference Entropy	Low Gray Level Run Emphasis
Difference Variance	Run Entropy
Inverse Difference	Run Length Non Uniformity
Inverse Difference Moment	Run Length Non Uniformity Normalized
Inverse Difference Moment Normalized	Run Percentage
Inverse Difference Normalized	Run Variance
Informational Measure of Correlation1	Short Run Emphasis
Informational Measure of Correlation2	Short Run High Gray Level Emphasis
Inverse Variance	Short Run Low Gray Level Emphasis
Joint Average	
Joint Energy	
Joint Entropy	
Maximal Correlation Coefficient	
Maximum Probability	
Sum Average	
Sum Entropy	
Sum Squares	

GLSZM	GLDM
Gray Level Non Uniformity	Dependence Entropy
Gray Level Non Uniformity Normalized	Dependence Non Uniformity
Gray Level Variance	Dependence Non Uniformity Normalized
High Gray Level Zone Emphasis	Dependence Variance
Large Area Emphasis	Gray Level Non Uniformity
Large Area High Gray Level Emphasis	Gray Level Variance
Large Area Low Gray Level Emphasis	High Gray Level Emphasis
Low Gray Level Zone Emphasis	Large Dependence Emphasis
Size Zone Non Uniformity	Large Dependence High Gray Level Emphasis
Size Zone Non Uniformity Normalized	Large Dependence Low Gray Level Emphasis
Small Area Emphasis	Low Gray Level Emphasis
Small Area High Gray Level Emphasis	Small Dependence Emphasis
Small Area Low Gray Level Emphasis	Small Dependence High Gray Level Emphasis
Zone Entropy	Small Dependence Low Gray Level Emphasis
Zone Percentage	
Zone Variance	

NGTDM	
Busyness	
Coarseness	
Complexity	
Contrast	
Strength	

Table S2. The trained General model intercept and coefficients.

Feature	Coefficient
Intercept	-589.649
firstorder_10Percentile_WP	-0.146612531
firstorder_90Percentile_WP	0.314
firstorder_Kurtosis_WP	-0.117812674
firstorder_Maximum_WP	-0.084492677
firstorder_Skewness_WP	0.593
firstorder_TotalEnergy_WP	1.02358E-08
firstorder_Uniformity_WP	230.905
shape_Elongation_WP	5.983
shape_Flatness_WP	4.539
shape_MajorAxisLength_WP	-0.06845567
shape_Maximum2DDiameterColumn_WP	-0.230560053
shape_Maximum2DDiameterRow_WP	-0.197985328
shape_Maximum2DDiameterSlice_WP	0.097
shape_Maximum3DDiameter_WP	-0.017992308
shape_Sphericity_WP	60.282
shape_SurfaceArea_WP	0.000246851
shape_SurfaceVolumeRatio_WP	14.199
shape_VoxelVolume_WP	1.89481E-05
glcm_Autocorrelation_WP	0.012
glcm_ClusterProminence_WP	2.54466E-06
glcm_ClusterShade_WP	-0.000307025

glcm_Correlation_WP	-6.007937184
glcm_Idmn_WP	399.583
glcm_Idn_WP	39.058
glcm_Imc2_WP	-2.099325012
glcm_MCC_WP	1.038
glcm_MaximumProbability_WP	399.127
glcm_SumSquares_WP	-0.009626445
glrlm_GrayLevelNonUniformity_WP	8.3781E-05
glrlm_RunEntropy_WP	-3.564375279
glrlm_ShortRunEmphasis_WP	169.42
glrlm_ShortRunLowGrayLevelEmphasis_WP	-94.32961035
glszm_GrayLevelNonUniformityNormalized_WP	-6.943523116
glszm_HighGrayLevelZoneEmphasis_WP	-0.000191147
glszm_LargeAreaEmphasis_WP	-7.997E-05
glszm_LargeAreaHighGrayLevelEmphasis_WP	-1.77614E-07
glszm_LowGrayLevelZoneEmphasis_WP	-826.0651781
glszm_SizeZoneNonUniformity_WP	0.00055125
glszm_SmallAreaEmphasis_WP	13.466
ngtdm_Coarseness_WP	-1711.077441
ngtdm_Complexity_WP	-1.5201E-05
ngtdm_Contrast_WP	-41.04467909
gldm_LargeDependenceHighGrayLevelEmphasis_WP	-0.001140483
gldm_LargeDependenceLowGrayLevelEmphasis_WP	2.304
gldm_SmallDependenceLowGrayLevelEmphasis_WP	4037.305
firstorder_10Percentile_Apex	0.07
firstorder_90Percentile_Apex	0.005208716
firstorder_InterquartileRange_Apex	-0.167191767
firstorder_Kurtosis_Apex	-1.413813301
firstorder_Maximum_Apex	0.002036756
firstorder_Range_Apex	0.022
firstorder_TotalEnergy_Apex	3.86153E-09
firstorder_Variance_Apex	-0.003263068
shape_Flatness_Apex	4.344
shape_LeastAxisLength_Apex	-0.25878969
shape_MajorAxisLength_Apex	-0.181309904
shape_Maximum2DDiameterColumn_Apex	0.144
shape_Maximum2DDiameterRow_Apex	-0.05676138
shape_Maximum2DDiameterSlice_Apex	-0.160179193
shape_Maximum3DDiameter_Apex	0.128
shape_MinorAxisLength_Apex	-0.078853106
shape_Sphericity_Apex	-18.44345633
shape_SurfaceVolumeRatio_Apex	-8.651565013
glcm_Autocorrelation_Apex	0.002396675
glcm_ClusterProminence_Apex	1.58546E-05
glcm_ClusterShade_Apex	-0.000233358
glcm_Contrast_Apex	0.024
glcm_DifferenceEntropy_Apex	-3.181807907
glcm_Imc2_Apex	-24.38850896
glcm_InverseVariance_Apex	-4.997607902
glcm_JointAverage_Apex	0.257
glcm_MCC_Apex	-5.063286745

glcm_MaximumProbability_Apex	274.934
glcm_SumAverage_Apex	0.000565326
glcm_SumEntropy_Apex	8.527
glrlm_GrayLevelNonUniformityNormalized_Apex	29.468
glszm_LowGrayLevelZoneEmphasis_Apex	-74.12210353
glszm_SizeZoneNonUniformity_Apex	-0.001095124
glszm_SmallAreaEmphasis_Apex	-21.1662332
glszm_SmallAreaHighGrayLevelEmphasis_Apex	0.000109585
glszm_SmallAreaLowGrayLevelEmphasis_Apex	-646.1424857
ngtdm_Coarseness_Apex	-100.5948909
ngtdm_Complexity_Apex	-0.0004838
ngtdm_Contrast_Apex	-9.923565475
gldm_DependenceNonUniformityNormalized_Apex	-8.445707876
gldm_DependenceVariance_Apex	1.657
gldm_GrayLevelNonUniformity_Apex	0.002130354
gldm_LargeDependenceLowGrayLevelEmphasis_Apex	11.498
gldm_SmallDependenceLowGrayLevelEmphasis_Apex	1392.455
firstorder_10Percentile_Base	-0.04819649
firstorder_InterquartileRange_Base	0.023
firstorder_Kurtosis_Base	0.099
firstorder_Median_Base	0.093
firstorder_Minimum_Base	0.067
firstorder_Range_Base	-0.024171931
firstorder_RobustMeanAbsoluteDeviation_Base	0.005885974
firstorder_Skewness_Base	0.311
firstorder_TotalEnergy_Base	-5.06827E-08
firstorder_Variance_Base	-0.000494672
shape_Elongation_Base	7.267
shape_Flatness_Base	-7.148647697
shape_MajorAxisLength_Base	0.015
shape_Maximum2DDiameterColumn_Base	0.088
shape_Maximum2DDiameterRow_Base	-0.029136784
shape_Maximum2DDiameterSlice_Base	0.145
shape_Maximum3DDiameter_Base	-0.15695807
shape_MinorAxisLength_Base	-0.299508723
shape_Sphericity_Base	7.676
shape_SurfaceArea_Base	0.001355521
shape_SurfaceVolumeRatio_Base	-7.966227392
glcm_ClusterProminence_Base	2.78103E-06
glcm_ClusterShade_Base	0.000147965
glcm_Contrast_Base	0.024
glcm_Correlation_Base	-4.439963012
glcm_DifferenceVariance_Base	-0.091992981
glcm_Imc1_Base	7.974
glcm_InverseVariance_Base	3.67
glcm_JointEntropy_Base	1.319
glcm_MCC_Base	-4.742227434
glcm_MaximumProbability_Base	27.757
glcm_SumSquares_Base	7.24288E-07
glrlm_GrayLevelNonUniformity_Base	0.000624266
glrlm_GrayLevelNonUniformityNormalized_Base	162.796

glrlm_GrayLevelVariance_Base	0.024
glrlm_RunLengthNonUniformityNormalized_Base	-2.037792585
glrlm_RunPercentage_Base	-43.98514778
glszm_GrayLevelVariance_Base	0.002715081
glszm_HighGrayLevelZoneEmphasis_Base	-0.005185856
glszm_LargeAreaEmphasis_Base	0.001428885
glszm_LargeAreaHighGrayLevelEmphasis_Base	-5.14652E-06
glszm_LargeAreaLowGrayLevelEmphasis_Base	-0.024674203
glszm_SizeZoneNonUniformity_Base	-0.000780204
glszm_SizeZoneNonUniformityNormalized_Base	31.825
glszm_SmallAreaHighGrayLevelEmphasis_Base	-0.000885754
ngtdm_Coarseness_Base	-258.3717435
ngtdm_Complexity_Base	-0.000762292
ngtdm_Contrast_Base	-29.49518732
ngtdm_Strength_Base	0.014
gldm_DependenceEntropy_Base	11.801
gldm_DependenceNonUniformityNormalized_Base	0.814
gldm_DependenceVariance_Base	-1.555376005
gldm_GrayLevelVariance_Base	0.017
gldm_LargeDependenceHighGrayLevelEmphasis_Base	-4.40462E-05
gldm_SmallDependenceLowGrayLevelEmphasis_Base	-664.1445652

The Radiomics features named as: Feature type_Feature name_Extraction area. WP: whole prostate.

Table S3. The trained PROSTATEx - U-Net model intercept and coefficients.

Feature	Coefficient
Intercept	-321.976
firstorder_10Percentile_WP	-0.361109955
firstorder_InterquartileRange_WP	0.169
firstorder_Kurtosis_WP	2.423
firstorder_Minimum_WP	-0.156260053
firstorder_Range_WP	0.039
firstorder_RobustMeanAbsoluteDeviation_WP	1.557
firstorder_TotalEnergy_WP	2.22304E-08
firstorder_Variance_WP	-0.046085079
shape_Elongation_WP	1.968
shape_MajorAxisLength_WP	-0.250079909
shape_Maximum2DDiameterColumn_WP	-0.257196182
shape_Maximum2DDiameterRow_WP	-0.145420583
shape_Maximum2DDiameterSlice_WP	-0.01886947
shape_Maximum3DDiameter_WP	-0.130069698
shape_MeshVolume_WP	0.000120006
shape_MinorAxisLength_WP	0.396
glcm_ClusterProminence_WP	2.71164E-06
glcm_ClusterShade_WP	0.000646111
glcm_Correlation_WP	-30.71256449
glcm_Idmn_WP	76.422
glcm_Imc1_WP	64.115
glcm_Imc2_WP	-42.31670998
glcm_MCC_WP	16.528
glcm_MaximumProbability_WP	-1094.72248
glszm_GrayLevelNonUniformityNormalized_WP	780.396

glszm_SizeZoneNonUniformity_WP	0.000570052
glszm_SmallAreaLowGrayLevelEmphasis_WP	-2417.522262
glszm_ZoneEntropy_WP	-8.843609049
glszm_ZonePercentage_WP	-7.43438445
ngtdm_Coarseness_WP	8643.934
ngtdm_Complexity_WP	-0.001616607
ngtdm_Strength_WP	4.063
gldm_DependenceEntropy_WP	41.577
gldm_DependenceNonUniformityNormalized_WP	-1.37948949
gldm_LargeDependenceHighGrayLevelEmphasis_WP	-0.001317192
gldm_LargeDependenceLowGrayLevelEmphasis_WP	64.421
firstorder_90Percentile_Apex	-0.107347033
firstorder_InterquartileRange_Apex	0.329
firstorder_Kurtosis_Apex	0.336
firstorder_Maximum_Apex	-0.000548995
firstorder_Median_Apex	-0.13347353
firstorder_Minimum_Apex	-0.024977611
firstorder_Variance_Apex	0.002616362
shape_Elongation_Apex	5.729
shape_LeastAxisLength_Apex	-0.410394504
shape_MajorAxisLength_Apex	-0.359425987
shape_Maximum2DDiameterColumn_Apex	0.279
shape_Maximum2DDiameterRow_Apex	0.036
shape_Maximum2DDiameterSlice_Apex	0.207
shape_MinorAxisLength_Apex	-0.384513413
shape_Sphericity_Apex	3.807
shape_SurfaceVolumeRatio_Apex	-44.062383
glcm_ClusterShade_Apex	-0.000321166
glcm_Correlation_Apex	6.09
glcm_Idn_Apex	-15.19375392
glcm_Imc2_Apex	-5.01588654
glcm_JointAverage_Apex	1.1824E-05
glcm_JointEnergy_Apex	-3038.847586
glcm_JointEntropy_Apex	-19.36122323
glcm_MCC_Apex	-11.42593655
glcm_MaximumProbability_Apex	2787.13
glcm_SumAverage_Apex	0.503
glcm_SumEntropy_Apex	44.621
glrlm_GrayLevelNonUniformity_Apex	-0.004539868
glrlm_LongRunLowGrayLevelEmphasis_Apex	-2134.126787
glrlm_ShortRunLowGrayLevelEmphasis_Apex	-665.1488514
glszm_GrayLevelNonUniformityNormalized_Apex	-691.5931433
glszm_GrayLevelVariance_Apex	0.013
glszm_SizeZoneNonUniformity_Apex	-0.000650814
glszm_SizeZoneNonUniformityNormalized_Apex	-54.0932208
glszm_SmallAreaEmphasis_Apex	-35.20923677
glszm_SmallAreaLowGrayLevelEmphasis_Apex	2823.613
glszm_ZoneEntropy_Apex	2.863
glszm_ZonePercentage_Apex	59.685
glszm_ZoneVariance_Apex	-0.006606771
ngtdm_Coarseness_Apex	1438.473

ngtdm_Complexity_Apex	-0.000288665
ngtdm_Strength_Apex	-3.219797197
gldm_DependenceEntropy_Apex	-36.05944425
gldm_DependenceVariance_Apex	8.133
gldm_HighGrayLevelEmphasis_Apex	-0.008487508
gldm_LargeDependenceHighGrayLevelEmphasis_Apex	-0.00071812
gldm_LargeDependenceLowGrayLevelEmphasis_Apex	81.651
firstorder_10Percentile_Base	0.204
firstorder_90Percentile_Base	0.292
firstorder_Kurtosis_Base	0.345
firstorder_Maximum_Base	-0.175831002
firstorder_RobustMeanAbsoluteDeviation_Base	1.058
firstorder_Skewness_Base	2.1
firstorder_TotalEnergy_Base	-1.01867E-07
shape_Elongation_Base	19.307
shape_Flatness_Base	-2.189503192
shape_LeastAxisLength_Base	0.142
shape_MajorAxisLength_Base	0.191
shape_Maximum2DDiameterColumn_Base	0.198
shape_Maximum2DDiameterRow_Base	-0.052203795
shape_Maximum3DDiameter_Base	0.025
shape_MinorAxisLength_Base	-0.395836828
shape_Sphericity_Base	-42.1834581
shape_SurfaceArea_Base	-0.001978519
shape_SurfaceVolumeRatio_Base	-30.98683539
glcm_Autocorrelation_Base	0.02
glcm_ClusterProminence_Base	1.6374E-05
glcm_ClusterShade_Base	-0.000524944
glcm_ClusterTendency_Base	-0.033890673
glcm_Idm_Base	-78.12670175
glcm_Idmn_Base	423.115
glcm_JointEnergy_Base	181.847
glcm_MCC_Base	10.921
glcm_MaximumProbability_Base	-41.13260693
glcm_SumSquares_Base	-0.05457626
glrlm_GrayLevelNonUniformityNormalized_Base	-145.1853283
glrlm_RunVariance_Base	-39.09025331
glrlm_ShortRunLowGrayLevelEmphasis_Base	68.658
glszm_LargeAreaHighGrayLevelEmphasis_Base	2.40474E-05
glszm_LargeAreaLowGrayLevelEmphasis_Base	0.387
glszm_LowGrayLevelZoneEmphasis_Base	1340.043
glszm_SizeZoneNonUniformityNormalized_Base	-7.925389028
glszm_SmallAreaLowGrayLevelEmphasis_Base	-900.8285969
glszm_ZoneVariance_Base	0.015
ngtdm_Busyness_Base	1.77
ngtdm_Complexity_Base	0.000728423
ngtdm_Strength_Base	-1.767239323
gldm_GrayLevelVariance_Base	-0.072416517
gldm_LargeDependenceLowGrayLevelEmphasis_Base	-64.18091988
gldm_SmallDependenceHighGrayLevelEmphasis_Base	-0.052635022

The Radiomics features named as: Feature type_Feature name_Extraction area. WP: whole prostate.

Table S4. The trained PROSTATEx - V-Net model intercept and coefficients.

Feature	Coefficient
Intercept	382.283
firstorder_90Percentile_WP	0.104
firstorder_InterquartileRange_WP	-0.12899759
firstorder_Median_WP	-0.184649075
firstorder_Minimum_WP	-0.258722838
firstorder_Skewness_WP	-0.499665432
firstorder_TotalEnergy_WP	1.14172E-08
firstorder_Uniformity_WP	734.318
firstorder_Variance_WP	-0.003975761
shape_Elongation_WP	14.085
shape_Flatness_WP	-26.24663162
shape_LeastAxisLength_WP	-0.188759158
shape_Maximum2DDiameterColumn_WP	-0.350192943
shape_Maximum2DDiameterRow_WP	-0.275736401
shape_Maximum2DDiameterSlice_WP	0.178
shape_Maximum3DDiameter_WP	0.103
shape_Sphericity_WP	23.296
shape_SurfaceVolumeRatio_WP	-187.1852196
glcm_Autocorrelation_WP	0.026
glcm_ClusterProminence_WP	-2.64673E-06
glcm_ClusterShade_WP	-0.000443921
glcm_Correlation_WP	-7.74557482
glcm_DifferenceVariance_WP	-0.205850712
glcm_Imc1_WP	49.886
glcm_MCC_WP	36.646
glcm_MaximumProbability_WP	572.011
glrlm_GrayLevelNonUniformity_WP	-0.001540021
glrlm_RunVariance_WP	-80.71373963
glszm_LargeAreaLowGrayLevelEmphasis_WP	0.032
glszm_SizeZoneNonUniformity_WP	0.000386128
glszm_SizeZoneNonUniformityNormalized_WP	-4.383540801
glszm_SmallAreaHighGrayLevelEmphasis_WP	0.019
glszm_SmallAreaLowGrayLevelEmphasis_WP	0.022
glszm_ZoneEntropy_WP	0.207
ngtdm_Coarseness_WP	31851.18
ngtdm_Complexity_WP	-0.000796818
ngtdm_Strength_WP	7.652
gldm_DependenceEntropy_WP	-3.524423106
gldm_LowGrayLevelEmphasis_WP	202.878
gldm_SmallDependenceLowGrayLevelEmphasis_WP	2828.786
firstorder_10Percentile_Apex	-0.176567555
firstorder_InterquartileRange_Apex	-0.139750995
firstorder_Kurtosis_Apex	-0.269244183
firstorder_Minimum_Apex	0.088
firstorder_Skewness_Apex	7.323
firstorder_Variance_Apex	0.011
shape_Elongation_Apex	-15.28827643
shape_Flatness_Apex	-9.750282188
shape_MajorAxisLength_Apex	-1.049820093

shape_Maximum2DDiameterColumn_Apex	-0.062538105
shape_Maximum2DDiameterRow_Apex	-0.011840197
shape_Maximum2DDiameterSlice_Apex	0.802
shape_Maximum3DDiameter_Apex	-0.797852217
shape_MeshVolume_Apex	0.000260848
shape_MinorAxisLength_Apex	0.241
shape_Sphericity_Apex	-69.6374527
shape_VoxelVolume_Apex	0.000124394
glcm_ClusterProminence_Apex	1.57165E-05
glcm_ClusterShade_Apex	-0.002219026
glcm_Correlation_Apex	12.85
glcm_DifferenceEntropy_Apex	-4.09303929
glcm_DifferenceVariance_Apex	-0.032803925
glcm_Imc1_Apex	-11.12989885
glcm_Imc2_Apex	-13.6901273
glcm_JointEnergy_Apex	-6990.37425
glcm_MCC_Apex	-31.9032659
glcm_MaximumProbability_Apex	1533.047
glcm_SumAverage_Apex	0.089
glcm_SumEntropy_Apex	13.538
glrlm_RunLengthNonUniformityNormalized_Apex	-92.79630735
glrlm_ShortRunEmphasis_Apex	-61.55064738
glszm_GrayLevelNonUniformityNormalized_Apex	101.119
glszm_LargeAreaHighGrayLevelEmphasis_Apex	-0.000181693
glszm_LargeAreaLowGrayLevelEmphasis_Apex	5.608
glszm_SmallAreaHighGrayLevelEmphasis_Apex	-0.014196094
glszm_ZoneVariance_Apex	-0.049201918
ngtdm_Busyness_Apex	2.637
ngtdm_Complexity_Apex	0.000669415
ngtdm_Contrast_Apex	-4.427026509
gldm_DependenceEntropy_Apex	1.031
gldm_DependenceNonUniformityNormalized_Apex	92.378
gldm_LargeDependenceEmphasis_Apex	3.355
gldm_LargeDependenceHighGrayLevelEmphasis_Apex	0.001035733
gldm_LargeDependenceLowGrayLevelEmphasis_Apex	-50.79463837
gldm_LowGrayLevelEmphasis_Apex	-157.2508035
gldm_SmallDependenceLowGrayLevelEmphasis_Apex	-2099.038966
firstorder_10Percentile_Base	-0.587515943
firstorder_90Percentile_Base	0.286
firstorder_Energy_Base	9.70197E-09
firstorder_InterquartileRange_Base	-0.393165233
firstorder_Kurtosis_Base	1.857
firstorder_Skewness_Base	-9.12033662
firstorder_TotalEnergy_Base	2.81459E-08
shape_Elongation_Base	18.966
shape_LeastAxisLength_Base	-0.177479526
shape_MajorAxisLength_Base	0.15
shape_Maximum2DDiameterColumn_Base	0.125
shape_Maximum2DDiameterRow_Base	0.037
shape_Maximum3DDiameter_Base	-0.245084468
shape_MinorAxisLength_Base	-0.658126709

shape_Sphericity_Base	-19.69112292
shape_SurfaceArea_Base	7.8765E-05
shape_SurfaceVolumeRatio_Base	-11.58048961
glcm_ClusterShade_Base	-0.000434701
glcm_Correlation_Base	1.614
glcm_DifferenceVariance_Base	-0.4222514
glcm_Idm_Base	-50.69111048
glcm_Imc2_Base	54.678
glcm_JointEnergy_Base	451.315
glcm_MCC_Base	-8.432719431
glcm_MaximumProbability_Base	1006.898
glcm_SumEntropy_Base	-5.542663873
glrlm_RunLengthNonUniformityNormalized_Base	-0.493380462
glrlm_RunVariance_Base	-37.80154633
glszm_GrayLevelNonUniformity_Base	-0.000570521
glszm_GrayLevelNonUniformityNormalized_Base	-717.5100996
glszm_LargeAreaEmphasis_Base	0.005691618
glszm_LargeAreaLowGrayLevelEmphasis_Base	0.003134206
glszm_LowGrayLevelZoneEmphasis_Base	-2803.662984
glszm_SmallAreaHighGrayLevelEmphasis_Base	-0.053575609
glszm_ZoneVariance_Base	3.25766E-07
ngtdm_Coarseness_Base	-13159.97686
ngtdm_Complexity_Base	0.001141479
ngtdm_Contrast_Base	58.232
gldm_DependenceEntropy_Base	-6.324520776
gldm_DependenceNonUniformityNormalized_Base	112.956
gldm_GrayLevelVariance_Base	0.055
gldm_SmallDependenceHighGrayLevelEmphasis_Base	-0.044455531

The Radiomics features named as: Feature type_Feature name_Extraction area. WP: whole prostate.

Table S5. The trained PROSTATEx - nnU-Net-2D model intercept and coefficients.

Feature	**Coefficient**
Intercept	796.981
firstorder_10Percentile_WP	0.29
firstorder_Energy_WP	1.19916E-10
firstorder_InterquartileRange_WP	0.636
firstorder_Maximum_WP	-0.109250329
firstorder_Minimum_WP	-0.694445779
shape_Elongation_WP	-24.70301357
shape_Flatness_WP	-14.17148864
shape_MajorAxisLength_WP	-0.718393217
shape_Maximum2DDiameterColumn_WP	-0.202531101
shape_Maximum2DDiameterRow_WP	-0.122970008
shape_Maximum2DDiameterSlice_WP	0.092
shape_Maximum3DDiameter_WP	-0.208910964
shape_SurfaceArea_WP	0.00044424
shape_SurfaceVolumeRatio_WP	-193.03525
glcm_ClusterProminence_WP	2.61885E-05
glcm_ClusterShade_WP	0.000594259
glcm_Correlation_WP	5.482
glcm_DifferenceVariance_WP	-0.236450592

glcm_JointEnergy_WP	866.556
glcm_MCC_WP	-5.194276409
glcm_SumSquares_WP	-0.085176649
glrlm_GrayLevelNonUniformity_WP	0.001755906
glrlm_GrayLevelNonUniformityNormalized_WP	162.861
glszm_GrayLevelNonUniformityNormalized_WP	-933.9826575
glszm_GrayLevelVariance_WP	0.026
glszm_HighGrayLevelZoneEmphasis_WP	0.001439849
glszm_SizeZoneNonUniformity_WP	0.000331595
glszm_SmallAreaLowGrayLevelEmphasis_WP	-1675.145157
glszm_ZoneEntropy_WP	-3.125394281
ngtdm_Coarseness_WP	2096.677
ngtdm_Complexity_WP	-0.000475403
ngtdm_Contrast_WP	-121.8550695
gldm_DependenceEntropy_WP	-15.70114408
gldm_DependenceNonUniformityNormalized_WP	178.047
gldm_LargeDependenceLowGrayLevelEmphasis_WP	19.94
gldm_LowGrayLevelEmphasis_WP	2135.631
firstorder_Kurtosis_Apex	-2.711844451
firstorder_Minimum_Apex	0.129
firstorder_Range_Apex	-0.07191723
firstorder_RobustMeanAbsoluteDeviation_Apex	-0.1420014
firstorder_Skewness_Apex	6.353
firstorder_TotalEnergy_Apex	4.13619E-08
shape_Elongation_Apex	3.668
shape_Flatness_Apex	-7.912707405
shape_LeastAxisLength_Apex	-0.10926489
shape_MajorAxisLength_Apex	-0.43237436
shape_Maximum2DDiameterColumn_Apex	0.189
shape_Maximum2DDiameterRow_Apex	-0.073192134
shape_Maximum2DDiameterSlice_Apex	0.178
shape_Maximum3DDiameter_Apex	-0.187472525
shape_SurfaceArea_Apex	0.000530272
glcm_ClusterProminence_Apex	2.77501E-05
glcm_ClusterShade_Apex	-0.000451384
glcm_Contrast_Apex	0.254
glcm_Correlation_Apex	-1.619781061
glcm_DifferenceEntropy_Apex	-17.07934278
glcm_DifferenceVariance_Apex	0.143
glcm_Imc2_Apex	-23.14455426
glcm_JointEnergy_Apex	-6212.219438
glcm_JointEntropy_Apex	-6.518581485
glcm_MaximumProbability_Apex	-120.8394487
glrlm_GrayLevelVariance_Apex	-0.036884918
glszm_GrayLevelVariance_Apex	-0.179393287
glszm_HighGrayLevelZoneEmphasis_Apex	0.021
glszm_LowGrayLevelZoneEmphasis_Apex	-267.8410649
glszm_SizeZoneNonUniformity_Apex	-0.001044345
glszm_ZoneVariance_Apex	-0.006817436
ngtdm_Complexity_Apex	-0.001420348
ngtdm_Contrast_Apex	-17.76628766

gldm_DependenceNonUniformityNormalized_Apex	-104.3249813
gldm_GrayLevelNonUniformity_Apex	-0.002647112
gldm_LargeDependenceLowGrayLevelEmphasis_Apex	0.441
gldm_SmallDependenceLowGrayLevelEmphasis_Apex	3108.422
firstorder_Energy_Base	2.51608E-08
firstorder_Entropy_Base	-2.704729568
firstorder_Kurtosis_Base	-0.419054007
firstorder_Skewness_Base	1.567
firstorder_Uniformity_Base	82.544
firstorder_Variance_Base	0.001343718
shape_Elongation_Base	-5.484035842
shape_Flatness_Base	-18.17738825
shape_LeastAxisLength_Base	0.996
shape_MajorAxisLength_Base	-0.311083432
shape_Maximum2DDiameterColumn_Base	0.425
shape_Maximum2DDiameterRow_Base	0.102
shape_Maximum2DDiameterSlice_Base	-0.375341977
shape_Maximum3DDiameter_Base	0.353
shape_SurfaceArea_Base	-0.001191261
shape_SurfaceVolumeRatio_Base	19.444
glcm_ClusterProminence_Base	-2.03688E-06
glcm_Correlation_Base	-5.390434191
glcm_InverseVariance_Base	37.919
glcm_JointAverage_Base	1.506
glcm_JointEntropy_Base	3.562
glcm_SumAverage_Base	0.001159842
glrlm_GrayLevelNonUniformityNormalized_Base	128.364
glrlm_RunLengthNonUniformityNormalized_Base	-388.4055959
glszm_LargeAreaHighGrayLevelEmphasis_Base	-3.98377E-05
glszm_LowGrayLevelZoneEmphasis_Base	-749.5529367
glszm_SmallAreaHighGrayLevelEmphasis_Base	-0.007953134
glszm_ZoneEntropy_Base	2.794
ngtdm_Busyness_Base	-4.202338416
ngtdm_Complexity_Base	-0.001070331
ngtdm_Contrast_Base	23.444
ngtdm_Strength_Base	-1.538993176
gldm_DependenceNonUniformityNormalized_Base	135.232
gldm_HighGrayLevelEmphasis_Base	-0.01486939
gldm_LargeDependenceEmphasis_Base	0.262
gldm_LargeDependenceHighGrayLevelEmphasis_Base	-0.001878785
gldm_LargeDependenceLowGrayLevelEmphasis_Base	-38.97542234
gldm_SmallDependenceLowGrayLevelEmphasis_Base	4057.745

The Radiomics features named as: Feature type_Feature name_Extraction area. WP: whole prostate.

Table S6. The trained PROSTATEx - nnU-Net-3D model intercept and coefficients.

Feature	Coefficient
Intercept	1309.623
firstorder_10Percentile_WP	-0.721142033
firstorder_Energy_WP	1.11777E-08
firstorder_InterquartileRange_WP	0.466
firstorder_Kurtosis_WP	-1.807098348

firstorder_Minimum_WP	-0.414416705
firstorder_RobustMeanAbsoluteDeviation_WP	0.343
firstorder_Skewness_WP	2.235
firstorder_TotalEnergy_WP	2.76743E-08
shape_Elongation_WP	-14.90949407
shape_Flatness_WP	5.777
shape_Maximum2DDiameterColumn_WP	-0.02988986
shape_Maximum3DDiameter_WP	-0.108640637
shape_MeshVolume_WP	0.000129798
shape_SurfaceArea_WP	-0.005093056
shape_SurfaceVolumeRatio_WP	-47.40701402
shape_VoxelVolume_WP	3.04889E-05
glcm_ClusterShade_WP	0.001115036
glcm_Contrast_WP	-0.362129542
glcm_DifferenceVariance_WP	-0.116939892
glcm_Imc1_WP	117.02
glcm_JointAverage_WP	0.944
glcm_MCC_WP	-10.81140584
glcm_MaximumProbability_WP	50.582
glcm_SumAverage_WP	0.045
glrlm_GrayLevelNonUniformity_WP	0.001113064
glrlm_GrayLevelVariance_WP	-0.25316124
glszm_GrayLevelVariance_WP	-0.023530978
glszm_LargeAreaHighGrayLevelEmphasis_WP	-1.76798E-06
glszm_LowGrayLevelZoneEmphasis_WP	2739.237
glszm_SmallAreaEmphasis_WP	82.808
glszm_SmallAreaHighGrayLevelEmphasis_WP	0.021
ngtdm_Busyness_WP	1.076
ngtdm_Coarseness_WP	-0.939921946
ngtdm_Complexity_WP	0.00427186
ngtdm_Strength_WP	-4.814057849
gldm_DependenceVariance_WP	-2.537048333
gldm_SmallDependenceLowGrayLevelEmphasis_WP	-4212.588098
firstorder_10Percentile_Apex	-0.466481814
firstorder_InterquartileRange_Apex	-0.114658415
firstorder_Kurtosis_Apex	-1.75226954
firstorder_Median_Apex	0.33
firstorder_Minimum_Apex	0.395
firstorder_Range_Apex	-0.065000378
firstorder_Skewness_Apex	8.473
shape_Elongation_Apex	1.819
shape_Flatness_Apex	-25.91600339
shape_LeastAxisLength_Apex	0.661
shape_MajorAxisLength_Apex	-0.439242302
shape_Maximum2DDiameterColumn_Apex	0.07
shape_Maximum2DDiameterRow_Apex	-0.177117185
shape_Maximum2DDiameterSlice_Apex	0.144
shape_Maximum3DDiameter_Apex	-0.068137016
glcm_ClusterProminence_Apex	1.65745E-05
glcm_ClusterShade_Apex	-0.000651419
glcm_Correlation_Apex	-28.7185583

glcm_DifferenceVariance_Apex	-0.124905076
glcm_Imc1_Apex	-38.63027152
glcm_Imc2_Apex	3.105
glcm_InverseVariance_Apex	-222.5726206
glcm_JointEnergy_Apex	615.764
glcm_MCC_Apex	1.043
glrlm_LongRunLowGrayLevelEmphasis_Apex	370.289
glrlm_ShortRunEmphasis_Apex	-364.7467476
glszm_GrayLevelVariance_Apex	0.014
glszm_LargeAreaHighGrayLevelEmphasis_Apex	-0.000126603
glszm_SmallAreaEmphasis_Apex	-61.04673102
glszm_SmallAreaLowGrayLevelEmphasis_Apex	-1246.855983
ngtdm_Busyness_Apex	0.111
ngtdm_Complexity_Apex	-0.001052067
ngtdm_Contrast_Apex	2.963
gldm_DependenceEntropy_Apex	5.341
gldm_DependenceNonUniformity_Apex	-0.000773742
gldm_DependenceVariance_Apex	4.539
gldm_LargeDependenceHighGrayLevelEmphasis_Apex	0.00149014
gldm_SmallDependenceHighGrayLevelEmphasis_Apex	-0.00868048
firstorder_10Percentile_Base	1.494
firstorder_InterquartileRange_Base	-0.297270611
firstorder_Kurtosis_Base	0.841
firstorder_Median_Base	-1.129276822
firstorder_TotalEnergy_Base	9.28406E-09
firstorder_Uniformity_Base	-967.8388438
firstorder_Variance_Base	0.016
shape_Elongation_Base	-11.1824233
shape_Flatness_Base	-1.92091796
shape_MajorAxisLength_Base	-0.07769645
shape_Maximum2DDiameterColumn_Base	0.035
shape_Maximum2DDiameterRow_Base	0.177
shape_Maximum2DDiameterSlice_Base	0.333
shape_Maximum3DDiameter_Base	-0.454223481
shape_MinorAxisLength_Base	-0.041608717
shape_Sphericity_Base	-54.66939797
shape_SurfaceVolumeRatio_Base	-48.9220416
glcm_ClusterShade_Base	-0.0007041
glcm_ClusterTendency_Base	-0.003726705
glcm_Correlation_Base	27.166
glcm_DifferenceVariance_Base	0.191
glcm_Imc1_Base	16.911
glcm_JointAverage_Base	2.628
glcm_MCC_Base	5.967
glcm_MaximumProbability_Base	-591.1130396
glcm_SumEntropy_Base	-26.76138121
glrlm_RunLengthNonUniformity_Base	-0.000192395
glrlm_RunLengthNonUniformityNormalized_Base	-675.9008458
glrlm_RunVariance_Base	-172.7504116
glszm_GrayLevelVariance_Base	3.21173E-06
glszm_LargeAreaLowGrayLevelEmphasis_Base	1.932

glszm_SizeZoneNonUniformityNormalized_Base	71.564
glszm_SmallAreaHighGrayLevelEmphasis_Base	-0.040155764
glszm_SmallAreaLowGrayLevelEmphasis_Base	139.748
glszm_ZoneEntropy_Base	4.103
ngtdm_Busyness_Base	1.303
ngtdm_Coarseness_Base	-364.9687301
ngtdm_Complexity_Base	-0.000170833
ngtdm_Contrast_Base	49.398
ngtdm_Strength_Base	2.141
gldm_DependenceNonUniformity_Base	-0.00106838
gldm_DependenceVariance_Base	-6.512468818
gldm_LargeDependenceHighGrayLevelEmphasis_Base	-0.002266611
gldm_SmallDependenceLowGrayLevelEmphasis_Base	263.478

The Radiomics features named as: Feature type_Feature name_Extraction area. WP: whole prostate.

Table S7. The trained In-house - U-Net model intercept and coefficients.

Feature	Coefficient
Intercept	58.793
firstorder_10Percentile_WP	-0.409826323
firstorder_InterquartileRange_WP	0.105
firstorder_Kurtosis_WP	0.115
firstorder_Minimum_WP	0.642
firstorder_Range_WP	-0.010042831
firstorder_Skewness_WP	-1.126762793
firstorder_Variance_WP	0.021
shape_Elongation_WP	5.546
shape_LeastAxisLength_WP	-0.588233617
shape_MajorAxisLength_WP	-0.009761257
shape_Maximum2DDiameterColumn_WP	-0.287024615
shape_Maximum2DDiameterRow_WP	-0.084692949
shape_Maximum3DDiameter_WP	-0.35243464
shape_MeshVolume_WP	0.000124765
shape_Sphericity_WP	48.425
shape_SurfaceVolumeRatio_WP	75.795
shape_VoxelVolume_WP	0.0001127
glcm_Autocorrelation_WP	0.065
glcm_ClusterShade_WP	-0.000371266
glcm_ClusterTendency_WP	-0.039895043
glcm_JointAverage_WP	0.00043459
glcm_JointEnergy_WP	-3086.280781
glcm_MCC_WP	-0.682031328
glcm_MaximumProbability_WP	1315.363
glcm_SumAverage_WP	0.011
glcm_SumSquares_WP	-0.209410018
glrlm_RunEntropy_WP	-2.523579447
glszm_GrayLevelNonUniformityNormalized_WP	-1786.604709
glszm_LargeAreaHighGrayLevelEmphasis_WP	-4.08143E-07
glszm_LargeAreaLowGrayLevelEmphasis_WP	-0.076981146
glszm_LowGrayLevelZoneEmphasis_WP	169.134
glszm_SmallAreaHighGrayLevelEmphasis_WP	-0.05355589
glszm_SmallAreaLowGrayLevelEmphasis_WP	120.991

glszm_ZoneEntropy_WP	-2.334812754
glszm_ZonePercentage_WP	-24.33033128
ngtdm_Busyness_WP	1.616
ngtdm_Coarseness_WP	-31168.57073
ngtdm_Complexity_WP	-0.000218006
ngtdm_Contrast_WP	17.704
gldm_DependenceNonUniformity_WP	0.000725363
gldm_DependenceNonUniformityNormalized_WP	-202.4339009
gldm_DependenceVariance_WP	-5.690421242
gldm_LargeDependenceHighGrayLevelEmphasis_WP	-0.001691745
gldm_SmallDependenceLowGrayLevelEmphasis_WP	-2640.112347
firstorder_10Percentile_Apex	1.024
firstorder_Energy_Apex	-5.20773E-08
firstorder_Kurtosis_Apex	-2.783838802
firstorder_Maximum_Apex	-0.152991221
firstorder_Minimum_Apex	-0.078605903
firstorder_Skewness_Apex	4.895
shape_Elongation_Apex	-23.91431753
shape_Flatness_Apex	44.268
shape_LeastAxisLength_Apex	-1.042189536
shape_MajorAxisLength_Apex	-0.180133877
shape_Maximum2DDiameterColumn_Apex	0.037
shape_Maximum3DDiameter_Apex	0.162
shape_MinorAxisLength_Apex	0.638
shape_Sphericity_Apex	-1.663816303
shape_SurfaceArea_Apex	-9.66892E-05
shape_SurfaceVolumeRatio_Apex	-3.74923741
glcm_ClusterProminence_Apex	1.20644E-05
glcm_Contrast_Apex	0.237
glcm_Correlation_Apex	3.756
glcm_Idm_Apex	94.084
glcm_Imc1_Apex	191.307
glcm_Imc2_Apex	-19.80854251
glcm_JointAverage_Apex	-0.165133348
glcm_MCC_Apex	4.809
glcm_MaximumProbability_Apex	-592.7530633
glcm_SumAverage_Apex	-0.164328998
glrlm_RunVariance_Apex	255.402
glszm_GrayLevelNonUniformity_Apex	-0.007447548
glszm_GrayLevelNonUniformityNormalized_Apex	-335.9868087
glszm_GrayLevelVariance_Apex	0.054
glszm_LargeAreaEmphasis_Apex	0.006333208
glszm_SmallAreaLowGrayLevelEmphasis_Apex	-582.8753445
glszm_ZoneEntropy_Apex	32.543
ngtdm_Busyness_Apex	-4.009037279
ngtdm_Complexity_Apex	-0.001872825
ngtdm_Contrast_Apex	-76.42068449
ngtdm_Strength_Apex	2.212
gldm_DependenceEntropy_Apex	-11.90863505
gldm_DependenceNonUniformityNormalized_Apex	93.417
gldm_LargeDependenceEmphasis_Apex	1.168

gldm_LargeDependenceHighGrayLevelEmphasis_Apex	-0.001201862
gldm_LargeDependenceLowGrayLevelEmphasis_Apex	-9.513155846
gldm_LowGrayLevelEmphasis_Apex	1653.482
firstorder_90Percentile_Base	-0.156919732
firstorder_Energy_Base	-5.6035E-08
firstorder_InterquartileRange_Base	-0.040518043
firstorder_Kurtosis_Base	1.093
firstorder_Maximum_Base	-0.042287615
firstorder_Minimum_Base	-0.174962769
firstorder_Skewness_Base	-4.199818513
firstorder_Uniformity_Base	-218.3511658
firstorder_Variance_Base	0.012
shape_Elongation_Base	-27.82042158
shape_Flatness_Base	33.905
shape_LeastAxisLength_Base	-0.561529979
shape_MajorAxisLength_Base	0.014
shape_Maximum2DDiameterColumn_Base	-0.164212242
shape_Maximum2DDiameterRow_Base	-0.122066823
shape_Maximum2DDiameterSlice_Base	0.164
shape_Maximum3DDiameter_Base	-0.012223014
shape_MinorAxisLength_Base	0.532
shape_Sphericity_Base	25.803
shape_SurfaceArea_Base	0.002999379
shape_SurfaceVolumeRatio_Base	-26.29355783
glcm_ClusterProminence_Base	9.642E-06
glcm_ClusterShade_Base	-0.000269434
glcm_Contrast_Base	0.045
glcm_Correlation_Base	-20.08108148
glcm_DifferenceEntropy_Base	-11.68204478
glcm_DifferenceVariance_Base	-0.164338494
glcm_Id_Base	-30.79271011
glcm_Idm_Base	-60.74166585
glcm_Imc1_Base	63.589
glcm_Imc2_Base	31.805
glcm_InverseVariance_Base	90.95
glcm_JointAverage_Base	-0.251969859
glcm_JointEnergy_Base	3.817
glcm_JointEntropy_Base	1.812
glcm_MCC_Base	-25.5010576
glcm_SumAverage_Base	-0.005675909
glcm_SumSquares_Base	0.15
glrlm_LongRunEmphasis_Base	-19.93426572
glszm_GrayLevelVariance_Base	0.066
glszm_LargeAreaHighGrayLevelEmphasis_Base	3.3941E-05
glszm_LargeAreaLowGrayLevelEmphasis_Base	-0.024327213
glszm_LowGrayLevelZoneEmphasis_Base	-139.2555514
glszm_SmallAreaEmphasis_Base	44.033
glszm_ZoneEntropy_Base	13.789
ngtdm_Busyness_Base	0.17
ngtdm_Coarseness_Base	1130.563
ngtdm_Complexity_Base	-0.001859765

ngtdm_Contrast_Base	-20.48422515
ngtdm_Strength_Base	0.192
gldm_DependenceNonUniformity_Base	-0.003340815
gldm_DependenceNonUniformityNormalized_Base	57.677
gldm_DependenceVariance_Base	1.992
gldm_GrayLevelNonUniformity_Base	-0.002678311
gldm_GrayLevelVariance_Base	-0.16360058
gldm_LargeDependenceHighGrayLevelEmphasis_Base	-0.001055181
gldm_LargeDependenceLowGrayLevelEmphasis_Base	6.17
gldm_SmallDependenceLowGrayLevelEmphasis_Base	-1014.191344

The Radiomics features named as: Feature type_Feature name_Extraction area. WP: whole prostate.

Table S8. The trained In-house - V-Net model intercept and coefficients.

Feature	Coefficient
Intercept	-665.676
firstorder_10Percentile_WP	0.607
firstorder_Energy_WP	-2.47051E-08
firstorder_Kurtosis_WP	-3.202042156
firstorder_MeanAbsoluteDeviation_WP	0.823
firstorder_Minimum_WP	-0.060290664
firstorder_Skewness_WP	6.37
firstorder_Uniformity_WP	964.467
firstorder_Variance_WP	0.022
shape_Elongation_WP	2.541
shape_Flatness_WP	-11.00444747
shape_LeastAxisLength_WP	-0.185191432
shape_MajorAxisLength_WP	-1.024911096
shape_Maximum2DDiameterColumn_WP	0.217
shape_Maximum2DDiameterRow_WP	0.035
shape_Maximum3DDiameter_WP	-0.090905338
shape_Sphericity_WP	161.442
glcm_ClusterShade_WP	-0.003112853
glcm_DifferenceVariance_WP	-0.185309313
glcm_JointAverage_WP	0.219
glcm_MCC_WP	37.95
glcm_MaximumProbability_WP	867.141
glcm_SumAverage_WP	0.008326982
glrlm_GrayLevelNonUniformityNormalized_WP	812.002
glszm_GrayLevelVariance_WP	-0.001948064
glszm_LargeAreaHighGrayLevelEmphasis_WP	-7.69233E-06
glszm_LargeAreaLowGrayLevelEmphasis_WP	0.033
glszm_LowGrayLevelZoneEmphasis_WP	-988.8339253
glszm_SizeZoneNonUniformity_WP	0.002356206
glszm_SizeZoneNonUniformityNormalized_WP	38.939
glszm_SmallAreaLowGrayLevelEmphasis_WP	-2355.458998
glszm_ZoneEntropy_WP	11.398
ngtdm_Busyness_WP	-0.169001974
ngtdm_Complexity_WP	0.002545098
ngtdm_Strength_WP	-4.096543889
gldm_DependenceEntropy_WP	20.5
gldm_DependenceVariance_WP	-1.415300226

gldm_GrayLevelNonUniformity_WP	-0.001548627
gldm_LargeDependenceLowGrayLevelEmphasis_WP	-22.42378295
gldm_SmallDependenceLowGrayLevelEmphasis_WP	9605.562
firstorder_Kurtosis_Apex	-1.468692815
firstorder_Mean_Apex	-0.102365267
firstorder_Minimum_Apex	1.103
firstorder_Range_Apex	-0.040394544
firstorder_TotalEnergy_Apex	8.33531E-09
firstorder_Uniformity_Apex	-94.07051406
shape_Flatness_Apex	12.813
shape_LeastAxisLength_Apex	-1.13644107
shape_MajorAxisLength_Apex	0.061
shape_Maximum2DDiameterColumn_Apex	0.058
shape_Maximum2DDiameterRow_Apex	-0.00670403
shape_Maximum2DDiameterSlice_Apex	-0.126244377
shape_Maximum3DDiameter_Apex	-0.170450804
shape_MinorAxisLength_Apex	-0.606954253
shape_Sphericity_Apex	-115.8931811
shape_SurfaceVolumeRatio_Apex	-47.23503452
glcm_ClusterProminence_Apex	-1.02783E-05
glcm_ClusterShade_Apex	0.000543035
glcm_Correlation_Apex	-5.541639536
glcm_Idm_Apex	-30.61394688
glcm_Imc1_Apex	76.203
glcm_MCC_Apex	10.11
glrlm_GrayLevelNonUniformity_Apex	0.008073803
glrlm_LongRunLowGrayLevelEmphasis_Apex	59.78
glszm_GrayLevelVariance_Apex	-0.022018761
glszm_HighGrayLevelZoneEmphasis_Apex	0.027
glszm_SizeZoneNonUniformity_Apex	-0.003308016
glszm_SizeZoneNonUniformityNormalized_Apex	-31.49718966
glszm_SmallAreaLowGrayLevelEmphasis_Apex	-980.0185665
ngtdm_Busyness_Apex	6.393
ngtdm_Coarseness_Apex	-12137.96129
ngtdm_Complexity_Apex	-0.000147367
ngtdm_Strength_Apex	7.118
gldm_GrayLevelNonUniformity_Apex	0.007816823
gldm_LargeDependenceHighGrayLevelEmphasis_Apex	-0.000363938
gldm_LargeDependenceLowGrayLevelEmphasis_Apex	17.551
firstorder_90Percentile_Base	0.311
firstorder_Energy_Base	-5.24322E-08
firstorder_InterquartileRange_Base	-0.409002608
firstorder_Maximum_Base	-0.15112896
firstorder_Minimum_Base	-0.358450791
firstorder_Skewness_Base	-0.554905249
shape_Elongation_Base	2.538
shape_Flatness_Base	35.184
shape_LeastAxisLength_Base	-0.375346084
shape_MajorAxisLength_Base	0.373
shape_Maximum2DDiameterColumn_Base	-0.204313436
shape_Maximum2DDiameterRow_Base	-0.418496704

shape_Maximum2DDiameterSlice_Base	0.048
shape_Maximum3DDiameter_Base	0.025
shape_Sphericity_Base	-39.96665213
shape_SurfaceArea_Base	0.003922066
shape_SurfaceVolumeRatio_Base	-14.51836726
glcm_ClusterProminence_Base	4.57063E-06
glcm_ClusterTendency_Base	0.012
glcm_Correlation_Base	-5.153969245
glcm_DifferenceVariance_Base	0.176
glcm_Idn_Base	404.812
glcm_Imc1_Base	-6.685165756
glcm_Imc2_Base	1.258
glcm_JointAverage_Base	0.925
glcm_JointEnergy_Base	-1718.336606
glcm_MCC_Base	2.527
glcm_MaximumProbability_Base	-401.0091738
glrlm_GrayLevelNonUniformity_Base	0.004031616
glrlm_GrayLevelNonUniformityNormalized_Base	99.731
glrlm_ShortRunLowGrayLevelEmphasis_Base	366.209
glszm_GrayLevelNonUniformity_Base	0.00597847
glszm_HighGrayLevelZoneEmphasis_Base	-0.04705148
glszm_LargeAreaEmphasis_Base	0.002089843
glszm_LargeAreaHighGrayLevelEmphasis_Base	-1.61651E-05
glszm_LowGrayLevelZoneEmphasis_Base	36.704
glszm_SmallAreaEmphasis_Base	98.631
glszm_ZoneVariance_Base	0.003766068
ngtdm_Busyness_Base	0.447
ngtdm_Complexity_Base	-0.00019638
ngtdm_Strength_Base	1.893
gldm_DependenceNonUniformityNormalized_Base	-43.49194441
gldm_LargeDependenceHighGrayLevelEmphasis_Base	0.000813639
gldm_SmallDependenceLowGrayLevelEmphasis_Base	-1314.284524

The Radiomics features named as: Feature type_Feature name_Extraction area. WP: whole prostate.

Table S9. The trained In-house - nnU-Net-2D model intercept and coefficients.

Feature	Coefficient
Intercept	-75.803
firstorder_10Percentile_WP	-0.952585663
firstorder_InterquartileRange_WP	-0.804508418
firstorder_Kurtosis_WP	1.149
firstorder_Maximum_WP	0.239
firstorder_Minimum_WP	1.595
firstorder_Skewness_WP	-5.171647096
firstorder_Variance_WP	0.016
shape_Elongation_WP	-14.52052431
shape_Flatness_WP	-9.36098304
shape_MajorAxisLength_WP	-0.448750298
shape_Maximum2DDiameterColumn_WP	-0.355542567
shape_Maximum2DDiameterRow_WP	-0.230861438
shape_Maximum2DDiameterSlice_WP	-0.13734098
shape_Sphericity_WP	80.22

shape_SurfaceArea_WP	0.00059422
shape_SurfaceVolumeRatio_WP	182.034
shape_VoxelVolume_WP	0.000197024
glcm_ClusterShade_WP	-0.0002725
glcm_Idmn_WP	133.637
glcm_Imc1_WP	-160.1112042
glcm_Imc2_WP	-52.75569003
glcm_InverseVariance_WP	-83.77687998
glcm_MCC_WP	-25.98528238
glcm_MaximumProbability_WP	104.238
glszm_GrayLevelVariance_WP	0.198
glszm_LargeAreaLowGrayLevelEmphasis_WP	0.082
glszm_LowGrayLevelZoneEmphasis_WP	-109.501208
glszm_SizeZoneNonUniformityNormalized_WP	2.242
ngtdm_Busyness_WP	-1.679498807
ngtdm_Complexity_WP	9.96145E-05
ngtdm_Contrast_WP	-58.94837564
ngtdm_Strength_WP	-40.34304987
gldm_DependenceNonUniformity_WP	0.001733947
gldm_LargeDependenceHighGrayLevelEmphasis_WP	0.002343983
gldm_LargeDependenceLowGrayLevelEmphasis_WP	32.69
firstorder_Energy_Apex	-2.25125E-08
firstorder_Kurtosis_Apex	-4.694483056
firstorder_Mean_Apex	0.081
firstorder_Median_Apex	0.2
firstorder_Minimum_Apex	0.026
firstorder_RobustMeanAbsoluteDeviation_Apex	-0.612443446
firstorder_Skewness_Apex	2.87
firstorder_Variance_Apex	-0.014627929
shape_Elongation_Apex	-1.486408625
shape_Flatness_Apex	-5.474652002
shape_LeastAxisLength_Apex	0.322
shape_MajorAxisLength_Apex	0.356
shape_Maximum2DDiameterColumn_Apex	0.062
shape_Maximum2DDiameterRow_Apex	-0.120393147
shape_Maximum2DDiameterSlice_Apex	-0.306718435
shape_Maximum3DDiameter_Apex	0.118
shape_Sphericity_Apex	-25.47544312
shape_SurfaceArea_Apex	-0.00115801
glcm_Autocorrelation_Apex	0.00212524
glcm_ClusterProminence_Apex	1.35505E-05
glcm_ClusterShade_Apex	0.00023352
glcm_Correlation_Apex	-31.78202127
glcm_DifferenceVariance_Apex	0.004458033
glcm_Imc2_Apex	-19.37831006
glcm_InverseVariance_Apex	-75.37312521
glcm_JointEnergy_Apex	-650.9100976
glcm_MCC_Apex	30.236
glcm_MaximumProbability_Apex	1369.891
glcm_SumEntropy_Apex	23.394
glrlm_GrayLevelNonUniformity_Apex	-0.008937945

glrlm_LongRunEmphasis_Apex	-78.29580954
glrlm_ShortRunEmphasis_Apex	340.148
glszm_GrayLevelNonUniformity_Apex	-3.15839E-07
glszm_LargeAreaHighGrayLevelEmphasis_Apex	8.70944E-05
glszm_LowGrayLevelZoneEmphasis_Apex	-892.9189737
glszm_SizeZoneNonUniformity_Apex	0.000268337
glszm_SmallAreaEmphasis_Apex	68.688
glszm_ZoneEntropy_Apex	-1.497017078
glszm_ZonePercentage_Apex	-298.2110028
ngtdm_Contrast_Apex	5.488
gldm_DependenceEntropy_Apex	-22.24853234
gldm_LargeDependenceEmphasis_Apex	-1.350096236
gldm_LargeDependenceHighGrayLevelEmphasis_Apex	-0.001367631
gldm_LowGrayLevelEmphasis_Apex	1109.321
firstorder_90Percentile_Base	0.038
firstorder_InterquartileRange_Base	0.056
firstorder_Maximum_Base	-0.068806763
firstorder_Median_Base	-0.167674187
firstorder_Minimum_Base	-0.108160733
firstorder_Skewness_Base	-3.71693444
firstorder_TotalEnergy_Base	4.37878E-08
firstorder_Variance_Base	0.009403888
shape_Elongation_Base	41.712
shape_Flatness_Base	-15.3874693
shape_LeastAxisLength_Base	0.022
shape_MajorAxisLength_Base	0.467
shape_Maximum2DDiameterColumn_Base	0.034
shape_Maximum2DDiameterRow_Base	-0.106613069
shape_Maximum2DDiameterSlice_Base	0.362
shape_MinorAxisLength_Base	-0.895357483
shape_Sphericity_Base	1.776
shape_SurfaceArea_Base	0.001371277
glcm_Autocorrelation_Base	-0.018914542
glcm_ClusterShade_Base	0.000367885
glcm_Correlation_Base	4.784
glcm_DifferenceVariance_Base	-0.13629786
glcm_Imc1_Base	36.779
glcm_Imc2_Base	-27.53258125
glcm_JointEnergy_Base	491.857
glcm_JointEntropy_Base	3.118
glcm_MCC_Base	10.865
glrlm_RunEntropy_Base	15.824
glrlm_RunLengthNonUniformity_Base	-0.000408762
glszm_HighGrayLevelZoneEmphasis_Base	0.026
glszm_LargeAreaEmphasis_Base	-0.034765993
glszm_SizeZoneNonUniformity_Base	-0.003023618
glszm_SmallAreaLowGrayLevelEmphasis_Base	1459.157
glszm_ZoneEntropy_Base	-15.84480995
glszm_ZonePercentage_Base	-79.29353486
ngtdm_Coarseness_Base	2348.112
ngtdm_Complexity_Base	0.001814881

ngtdm_Contrast_Base	31.051
gldm_GrayLevelVariance_Base	-0.159526151
gldm_LargeDependenceEmphasis_Base	-0.042181185
gldm_LowGrayLevelEmphasis_Base	-834.1902871
gldm_SmallDependenceLowGrayLevelEmphasis_Base	-408.9185167

The Radiomics features named as: Feature type_Feature name_Extraction area. WP: whole prostate.

Table S10. The trained In-house - nnU-Net-3D model intercept and coefficients.

Feature	Coefficient
Intercept	-149.82
firstorder_Energy_WP	-2.47222E-08
firstorder_InterquartileRange_WP	-0.106069147
firstorder_Kurtosis_WP	-1.647215782
firstorder_Mean_WP	1.313
firstorder_Minimum_WP	1.037
firstorder_Range_WP	-0.165230167
firstorder_Skewness_WP	25.027
shape_Elongation_WP	0.561
shape_Flatness_WP	12.073
shape_LeastAxisLength_WP	-0.288934047
shape_MajorAxisLength_WP	-0.02937799
shape_Maximum2DDiameterRow_WP	0.013
shape_Maximum2DDiameterSlice_WP	0.062
shape_Maximum3DDiameter_WP	-0.106485299
shape_MinorAxisLength_WP	0.123
shape_Sphericity_WP	100.511
glcm_ClusterShade_WP	-0.003130491
glcm_ClusterTendency_WP	0.04
glcm_Contrast_WP	-0.000125038
glcm_Correlation_WP	0.053
glcm_DifferenceVariance_WP	-0.198911163
glcm_Imc2_WP	-1.265234243
glcm_InverseVariance_WP	-20.32500725
glcm_MaximumProbability_WP	717.81
glrlm_ShortRunEmphasis_WP	345.005
glszm_GrayLevelNonUniformity_WP	0.009552618
glszm_LargeAreaHighGrayLevelEmphasis_WP	-2.47514E-06
glszm_LargeAreaLowGrayLevelEmphasis_WP	-0.387610541
glszm_LowGrayLevelZoneEmphasis_WP	-459.4574471
glszm_SizeZoneNonUniformity_WP	0.000138626
glszm_SmallAreaHighGrayLevelEmphasis_WP	-0.021041159
glszm_SmallAreaLowGrayLevelEmphasis_WP	-1052.318132
glszm_ZoneEntropy_WP	-12.53775801
ngtdm_Busyness_WP	1.568
ngtdm_Coarseness_WP	-26321.58435
ngtdm_Complexity_WP	-0.001319323
ngtdm_Strength_WP	14.672
gldm_DependenceEntropy_WP	18.6
gldm_LargeDependenceLowGrayLevelEmphasis_WP	69.318
firstorder_10Percentile_Apex	-0.157698579
firstorder_InterquartileRange_Apex	-0.299308216

firstorder_Kurtosis_Apex	-4.023784202
firstorder_Minimum_Apex	-0.86249131
firstorder_Range_Apex	0.109
firstorder_TotalEnergy_Apex	8.38876E-08
firstorder_Variance_Apex	-0.012439016
shape_Elongation_Apex	6.766
shape_LeastAxisLength_Apex	0.205
shape_MajorAxisLength_Apex	0.173
shape_Maximum2DDiameterColumn_Apex	0.068
shape_Maximum2DDiameterRow_Apex	-0.05706543
shape_Maximum2DDiameterSlice_Apex	-0.269497398
shape_Maximum3DDiameter_Apex	0.136
shape_Sphericity_Apex	0.508
shape_SurfaceArea_Apex	-0.00020118
shape_SurfaceVolumeRatio_Apex	-9.955860422
glcm_Autocorrelation_Apex	0.048
glcm_ClusterProminence_Apex	1.80532E-06
glcm_ClusterShade_Apex	0.000398695
glcm_Correlation_Apex	-21.41000414
glcm_DifferenceEntropy_Apex	-5.340768914
glcm_DifferenceVariance_Apex	-0.009513745
glcm_Idm_Apex	193.115
glcm_Imc2_Apex	-5.607008112
glcm_MCC_Apex	8.623
glcm_MaximumProbability_Apex	9.281
glcm_SumAverage_Apex	0.103
glrlm_ShortRunLowGrayLevelEmphasis_Apex	599.045
glszm_GrayLevelVariance_Apex	0.179
glszm_LargeAreaEmphasis_Apex	0.085
glszm_LargeAreaLowGrayLevelEmphasis_Apex	-0.671634403
glszm_SmallAreaEmphasis_Apex	-110.3198284
glszm_SmallAreaLowGrayLevelEmphasis_Apex	-217.001015
glszm_ZoneEntropy_Apex	-26.84580904
glszm_ZoneVariance_Apex	0.015
ngtdm_Busyness_Apex	1.381
ngtdm_Coarseness_Apex	-2782.771041
ngtdm_Complexity_Apex	0.000275014
gldm_GrayLevelNonUniformity_Apex	-0.011243202
gldm_LargeDependenceHighGrayLevelEmphasis_Apex	-0.005852267
gldm_LargeDependenceLowGrayLevelEmphasis_Apex	-4.484751303
firstorder_10Percentile_Base	-0.221789269
firstorder_Energy_Base	-2.31411E-08
firstorder_InterquartileRange_Base	-0.191553975
firstorder_Kurtosis_Base	0.28
firstorder_Minimum_Base	-0.10033044
firstorder_Uniformity_Base	511.824
shape_Elongation_Base	9.012
shape_Flatness_Base	-22.76796896
shape_LeastAxisLength_Base	-0.491363729
shape_MajorAxisLength_Base	-0.023559668
shape_Maximum2DDiameterColumn_Base	0.046

shape_Maximum2DDiameterRow_Base	-0.017964975
shape_Maximum2DDiameterSlice_Base	0.095
shape_Maximum3DDiameter_Base	-0.198770216
shape_MinorAxisLength_Base	-0.457096178
shape_Sphericity_Base	-2.336855096
shape_SurfaceArea_Base	0.003600587
shape_SurfaceVolumeRatio_Base	-4.416308449
glcm_ClusterProminence_Base	-3.5651E-06
glcm_Contrast_Base	-0.171013698
glcm_Correlation_Base	0.145
glcm_Imc1_Base	35.684
glcm_Imc2_Base	-7.574759687
glcm_JointAverage_Base	0.861
glcm_JointEnergy_Base	-958.2169646
glcm_MCC_Base	7.3
glcm_MaximumProbability_Base	-734.7529329
glrlm_RunVariance_Base	-154.8836307
glrlm_ShortRunHighGrayLevelEmphasis_Base	-0.021418708
glszm_GrayLevelNonUniformityNormalized_Base	-628.9855369
glszm_LargeAreaHighGrayLevelEmphasis_Base	0.000268432
glszm_LargeAreaLowGrayLevelEmphasis_Base	-2.395007695
glszm_SizeZoneNonUniformity_Base	-0.002338366
glszm_SmallAreaEmphasis_Base	36.437
ngtdm_Busyness_Base	-0.888311909
ngtdm_Complexity_Base	6.33956E-05
ngtdm_Contrast_Base	50.932
ngtdm_Strength_Base	-0.323326191
gldm_DependenceNonUniformityNormalized_Base	-34.53125393
gldm_LargeDependenceLowGrayLevelEmphasis_Base	37.811
gldm_LowGrayLevelEmphasis_Base	371.773

The Radiomics features named as: Feature type_Feature name_Extraction area. WP: whole prostate.

Paper III

The reproducibility of deep learning-based segmentation of the prostate gland and zones on T2-weighted MR images

Mohammed R. S. Sunoqrot[1,*], Kirsten M. Selnæs[1,2], Elise Sandsmark[2], Sverre Langørgen[2], Helena Bertilsson[3,4], Tone F. Bathen[1,2] and Mattijs Elschot[1,2]

[1]Department of Circulation and Medical Imaging, NTNU, Norwegian University of Science and Technology, Trondheim, 7030, Norway

[2]Department of Radiology and Nuclear Medicine, St. Olavs Hospital, Trondheim University Hospital, Trondheim, 7030, Norway

[3]Department of Cancer Research and Molecular Medicine, NTNU, Norwegian University of Science and Technology, Trondheim, 7030, Norway

[4]Department of Urology, St. Olavs Hospital, Trondheim University Hospital, Trondheim, 7030, Norway

[*]mohammed.sunoqrot@ntnu.no

* Corresponding author contact info:

Mohammed R. S. Sunoqrot

Email: mohammed.sunoqrot@ntnu.no

Address:

NTNU, MR Centre
Olav Kyrresgate 9 MTFS, 3rd floor, south
7030 Trondheim Norway

Abstract

Volume of interest segmentation is an essential step in computer-aided detection and diagnosis (CAD) systems. Deep learning (DL)-based methods provide good performance for prostate segmentation, but little is known about the reproducibility of these methods. In this work, an in-house collected dataset from 244 patients was used to investigate the intra-patient reproducibility of 14 shape features for DL-based segmentation methods of the whole prostate gland (WP), peripheral zone (PZ) and the remaining prostate zones (non-PZ) on T2-weighted (T2W) magnetic resonance (MR) images compared to manual segmentations. The DL-based segmentation was performed using three different convolutional neural networks (CNNs): V-Net, nnU-Net-2D and nnU-Net-3D. The two-way random, single score intra-class correlation coefficient (ICC) was used to measure the inter-scan reproducibility of each feature for each CNN and the manual segmentation. We found that the reproducibility of the investigated methods is comparable to manual for all CNNs (14/14 features), except for V-Net in PZ (7/14 features). The ICC score for segmentation volume was found to be 0.888, 0.607, 0.819 and 0.903 in PZ; 0.988, 0.967, 0.986 and 0.983 in non-PZ; and 0.982, 0.975, 0.973 and 0.984 in WP for manual, V-Net, nnU-Net-2D and nnU-Net-3D, respectively. The results of this work show the feasibility of embedding DL-based segmentation in CAD systems based on multiple T2W MR scans of the prostate, which is an important step towards the clinical implementation.

Introduction

Prostate cancer is the most detected cancer in men and the second most common cause of cancer related death for men worldwide[1]. An early diagnosis of prostate cancer is essential for a better disease management[2]. Following reasonable suspicion of prostate cancer, based on elevated prostate-specific antigen (PSA) levels in blood and a digital rectal examination, the patient, in many countries, is likely to be referred to a pre-biopsy magnetic resonance imaging (MRI) to guide the collection of biopsies[3]. To improve the diagnostic process, the use of multi-parametric MRI (mpMRI) has been established through international guidelines[4-6]. mpMRI has also been employed in active surveillance programs to follow up the patients with indolent lesions[7], prostate cancer risk calculators[8] and treatment response monitoring[6,9]. Currently, the mpMR images are interpreted qualitatively by a radiologist, which is expensive, time-consuming[10] and reader opinion-dependent[11,12]. The resulting vulnerability to inter and intra-observer variability is problematic for clinical applications based on multiple scans in time, such as with active surveillance and response monitoring, where reproducibility of results is paramount. Automated computer-aided detection and diagnosis (CAD) systems have the potential to overcome the limitations of the traditional radiological reading by implementing quantitative models to automate, standardize and support reproducible interpretation of radiological images[10,13-15].

Segmentation is an essential step for prostate CAD systems[14,15]. It helps locate the volume of interest (VOI), enabling subsequent extraction of quantitative features for radiomics-based approaches.

Accurate segmentation is paramount as the following steps of a CAD system are dependent on it. Traditionally, the VOI segmentation is performed manually by a radiologist on T2-weighted (T2W) MR images. However, deep learning (DL)-based segmentation methods have shown promising performance[16-20]. Importantly, the inter-observer variability between DL-based segmentation methods and expert radiologists has been shown to be approximately equal to that between expert radiologists[21]. However, little is known about the reproducibility of DL-based segmentation methods for clinical MRI scans. To investigate the reproducibility of DL-based segmentation, radiomics shape features can be used. Shape features like prostate volume are already part of today's clinical risk calculators for prostate cancer[8] and will likely play an important role in future radiomics-based clinical applications. In addition, these features show high reproducibility between mpMRI scans during a short time interval[22].

In this work, we investigated the reproducibility of DL-based segmentations of the whole prostate gland (WP), peripheral zone (PZ) and the remaining prostate zones (non-PZ; central, transition and anterior fibro-muscular zones, combined) by comparing radiomics shape features from T2W MR images acquired with short time intervals.

Methods

Dataset

In this study, we used an in-house collected mpMRI dataset from 244 patients (median age = 65; range: 44 – 76 years) examined at St. Olavs Hospital, Trondheim University Hospital, Trondheim, Norway between March 2015 and December 2017 due to suspicion of prostate cancer. All methods were carried out in accordance with the relevant guidelines and regulations. The study was approved by the institutional review board and The Regional Committee for Medical and Health Research Ethics (REC Central Norway, identifiers 2013/1869 and 2017/576). All patients signed informed consent prior to the initiation of the study. The dataset (T2W images) was split into a training set (N = 182), to train the DL-based segmentation networks, and an investigation set (N = 62), to investigate the reproducibility of shape features extracted from the segmented prostate masks.

The investigation set was acquired at two different time points: first, at the initial visit for detection of prostate cancer (scan 1), and second, during an MR-guided biopsy procedure (scan 2). The interval between scans ranged from 1 – 71 (median = 7) days.

T2W MRI was performed on a Magnetom Skyra 3 T MRI system (Siemens Healthineers, Erlangen, Germany) with a turbo spin-echo sequence. The scanning parameters details are given in Table 1.

Table 1. Summary of MRI scanning parameters.

	Investigation set		Training set
	Scan 1	**Scan 2**	
Repetition time (ms)	4800 – 8921	5660 – 7740	4450 – 9520
Echo time (ms)	101 – 104	101 – 104	101 – 108
Flip angle (degree)	152 – 160	152 – 160	145 – 160
Number of averages	3	3 – 6	1 – 3
Matrix size	320×320 – 384×384	320×320 – 384×384	320×320 – 384×384
Slices	24 – 30	17 – 24	24 – 34
Slice thickness (mm)	3	3	3 – 3.5
In plane resolution (mm^2)	0.5×0.5 – 0.6×0.6	0.5×0.5 – 0.6×0.6	0.5×0.5 – 0.6×0.6

Prostate Segmentation

Manual segmentation of PZ and non-PZ for the in-house collected dataset was performed using ITK-SNAP[23] by a radiology resident (E.S.) at St. Olavs Hospital, Trondheim University Hospital, Trondheim, Norway, under the supervision of a radiologist (S.L.) with more than 10 years′ experience in prostate imaging. PZ and non-PZ masks were used to generate the WP masks by merging.

The DL-based segmentation was performed using three different convolutional neural networks (CNNs): V-Net[19], nnU-Net-2D[20] and nnU-Net-3D[20]. nnU-Net-2D performed the segmentation on a 2D slice-by-slice basis, whereas V-Net and nnU-Net-3D performed the segmentation on a 3D volume basis. Prior to segmentation, all images were pre-processed in accordance with the corresponding segmentation method. The segmentation pre-processing, training, and testing were performed on a single NVIDIA Tesla P100 PCIe 16 GB GPU in Ubuntu 18.04.4 LTS. V-Net was implemented with PyTorch[24] (version 1.4.0) using Python (version 3.6.9; Python Software Foundation, Wilmington, DE, USA) to generate two models for WP and PZ which were used to generate non-PZ masks by subtraction. nnU-Net-2D and nnU-Net-3D were implemented with PyTorch (version 1.7.0) using Python (version 3.6.10) to generate both PZ and non-PZ, which were used to generate the WP masks by merging. The DL-based segmentations were post-processed to only keep the largest 3D connected component using a pixel connectivity of 26.

Feature extraction

Shape features were extracted from the 3D segmentation masks (Manual or DL-based) of PZ, non-PZ and WP using Pyradiomics[25] (version 3.0; an open-source Python package). The following 14 shape features were extracted: Elongation, Flatness, Least Axis Length, Major Axis Length, Maximum 2D diameter (Column), Maximum 2D diameter (Row), Maximum 2D diameter (Slice), Maximum 3D diameter, Mesh Volume, Minor Axis Length, Sphericity, Surface Area, Surface Area to Volume ratio and Voxel Volume. A detailed description of the features can be found at [26].

Investigation of reproducibility

Reproducibility is defined as the "variability in measurements made on the same subject, but under changing conditions"[27]. The variability and reproducibility are inversely related, i.e. the higher the variability, the lower the reproducibility. In this work, scan 1 and scan 2 were performed on the same patients, but at different time points and using different scanning procedures.

To investigate the reproducibility, all extracted features from the two scans of 62 patients' scans using the manual and post-processed DL-based segmentations were included. The reproducibility for each of the 14 shape features was investigated, separately, for each of the CNNs and compared to that of the corresponding feature from the manual segmentations. Furthermore, the DL-based segmentation performance and segmentation volume (Voxel Volume feature) in scan 1 and scan 2 were compared to those of manual segmentations.

In addition, the reproducibility results were compared to the corresponding results where (1) the post-processing step was excluded and (2) patients with a poor segmentation quality score were excluded. To enable the last comparison, our previously proposed automated segmentation quality control system (SQCS)[28] was implemented and the patients with a quality score less than 85 for scan 1 or/and scan 2 were excluded from further analysis. As per [28], the SQCS was implemented using pre-processed T2W images and WP segmentations.

Statistical analysis

The dice similarity coefficient (DSC)[29] between manual and DL-based segmentations was calculated as a metric of segmentation performance.

The two-way random, single score intra-class correlation coefficient (ICC)[30,31] was used to measure the inter-scan reproducibility of each feature for each CNN and the manual segmentations. Statistical significance between features from manual segmentation and each CNN, and between features from including and excluding the post-processing step was based on overlapping 95% confidence intervals (CI)[32].

The paired Wilcoxon signed rank test[33] followed by Benjamini-Hochberg correction for multiple testing[34] was used to assess the differences in DSC, the ICC values between VOIs and segmentation volume between networks and scans.

The Bland-Altman analysis[35] and Spearman's rank test[33] were performed to assess the correlation between the segmentation volumes for scan 1 and scan 2, and between the segmentation volumes of each of the CNNs and the manual segmentations in scan 1 and scan 2.

To assess the difference in feature reproducibility before and after implementing the SQCS, a permutation test[33] with 1000 runs was performed for each CNN. In each of these 1000 runs, the ICC value was calculated after randomly excluding the same number of cases as excluded by the SQCS. The improvement in ICC after applying the SQCS was considered significant if less than 50/1000 randomly permuted values were higher or equal to the ICC after the SQCS implementation.

MATLAB R2019b (Mathworks, Natick, MA, USA) was used for statistical analysis.

Results

An example case segmented with the three investigated CNNs is shown in Figure 1.

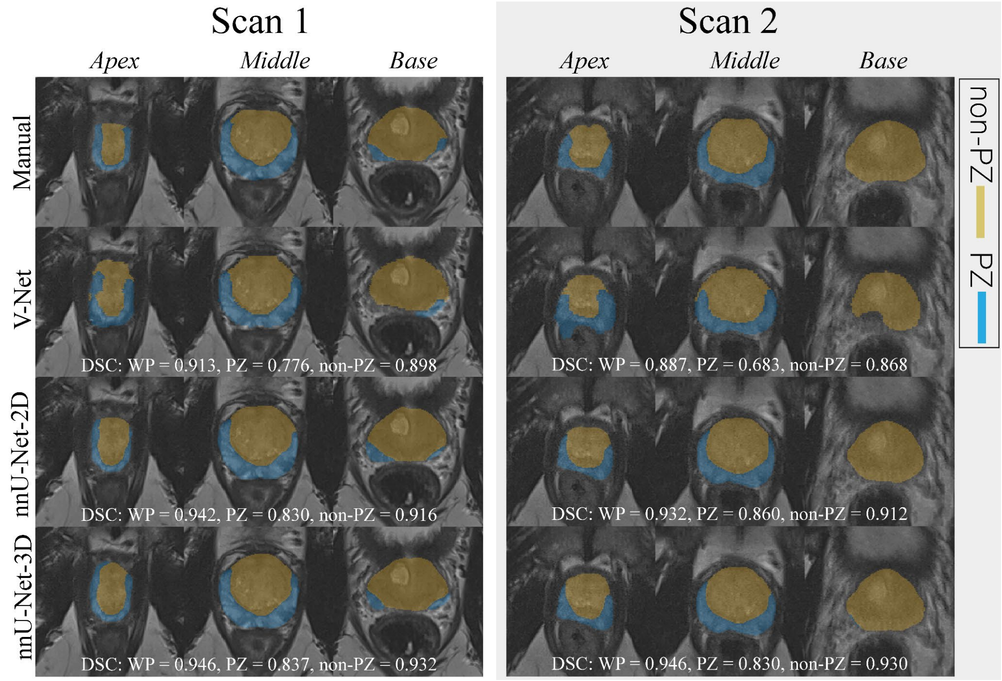

Scan 1
Scan 2
Apex
Middle
Base
Apex
Middle
Base
Manual
V-Net
nnU-Net-2D
nnU-Net-3D
non-PZ
PZ
DSC: WP = 0.913, PZ = 0.776, non-PZ = 0.898
DSC: WP = 0.887, PZ = 0.683, non-PZ = 0.868
DSC: WP = 0.942, PZ = 0.830, non-PZ = 0.916
DSC: WP = 0.932, PZ = 0.860, non-PZ = 0.912
DSC: WP = 0.946, PZ = 0.837, non-PZ = 0.932
DSC: WP = 0.946, PZ = 0.830, non-PZ = 0.930

Figure 1. The middle slice for the whole prostate, apex and base of a randomly selected case segmented (peripheral zone (PZ) and the remaining prostate zones (non-PZ)) by different approaches for scan 1 and 2. For each network, the dice similarity coefficient (DSC) of the 3D segmented volume is reported for the whole prostate gland (WP), PZ and non-PZ.

Figure 2 shows the performance of the investigated CNNs segmentations. The median DSCs were 0.781, 0.821 and 0.825 in PZ; 0.871, 0.916 and 0.917 in non-PZ; and 0.909, 0.937 and 0.940 in WP for V-Net, nnU-Net-2D and nnU-Net-3D, respectively, in scan 1, and 0.714, 0.788 and 0.798 in

PZ; 0.853, 0.896 and 0.904 in non-PZ; and 0.893, 0.917 and 0.929 in WP for V-Net, nnU-Net-2D and nnU-Net-3D, respectively, in scan 2. Median of DSC difference between the scans (scan 2 - scan 1) was -9.49%, -4.06% and -3.65% in PZ; -3.12%, -1.80% and -1.08% in non-PZ; and -1.98%, -1.95% and -1.39% in WP for V-Net, nnU-Net-2D and nnU-Net-3D, respectively. V-Net performed significantly lower (p<0.001) than nnU-Net-2D and nnU-Net-3D in both of the scans and all of VOIs. nnU-Net-3D performed significantly higher (p<0.01) than nnU-Net-2D in scan 2 for all of VOIs. In addition, each of the CNNs performed significantly lower (p<0.001) in scan 2 compared to scan 1.

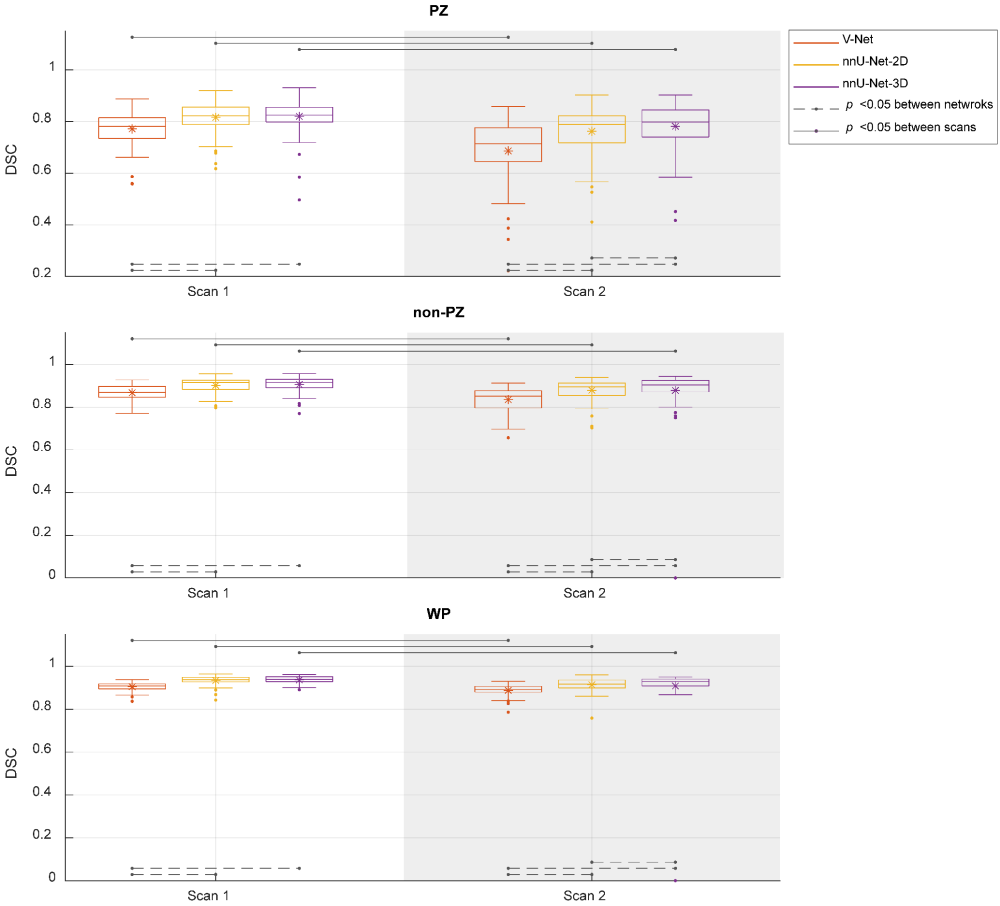


Figure 2. The performance (dice similarity coefficient - DSC) of the segmentation methods for the whole prostate gland (WP), peripheral zone (PZ) and the remaining prostate zones (non-PZ). The Manual segmentations were considered as reference.

Figure 3 shows the ICCs from the extracted shape features from scan 1 and scan 2, where the segmentation post-processing step was included and the segmentation quality control system was not implemented, demonstrating that the reproducibility of DL-based segmentation is comparable to manual segmentation for all networks (14/14 features), except for V-Net in PZ (7/14 features). In both manual and DL-based segmentations, Elongation, Flatness and Sphericity had a remarkably lower ICC than the other features in WP and non-PZ. nnU-Net-3D showed higher reproducibility than the rest of the CNNs with a median difference in ICC equal to 54.03% and 9.06% in PZ; 3.95% and 0.38% in non-PZ; and 0.95% and 1.09% in WP with V-Net and nnU-Net-2D, respectively. Additionally, in most cases feature reproducibility in the non-PZ and WP was higher than in the PZ. V-Net had significantly higher (p<0.01) ICCs in non-PZ and WP compared to PZ.

Comparing reproducibility when including (Figure 3) and excluding (Figure A1) the segmentation post-processing step, while SQCS was not implemented in any of them, shows that the reproducibility is remarkably enhanced when including the segmentation post-processing step. The ICC after including the segmentation post-processing step was significantly better in (4/14) features for V-Net in non-PZ; (14/14), (12/14) and (13/14) features for nnU-Net-2D in PZ, non-PZ and WP, respectively; and (13/14), (14/14) and (13/14) features for nnU-Net-3D in PZ, non-PZ and WP, respectively.

Similarly, the reproducibility was increased with the SQCS implementation (Figure A2) compared to no implementation (Figure 3); the segmentation post-processing step was included in both cases. After implementing the SQCS, 10, 11 and 6 patient's segmentations were excluded from V-Net, nnU-Net-2D and nnU-Net-3D, respectively. The ICC after implementing the SQCS was significantly better in (3/14), (2/14) and (3/14) features for V-Net in PZ, non-PZ and WP, respectively; in (7/14) and (2/14) features for nnU-Net-2D in non-PZ and WP, respectively; and in (1/14) and (5/14) features for nnU-Net-3D in non-PZ and WP, respectively.

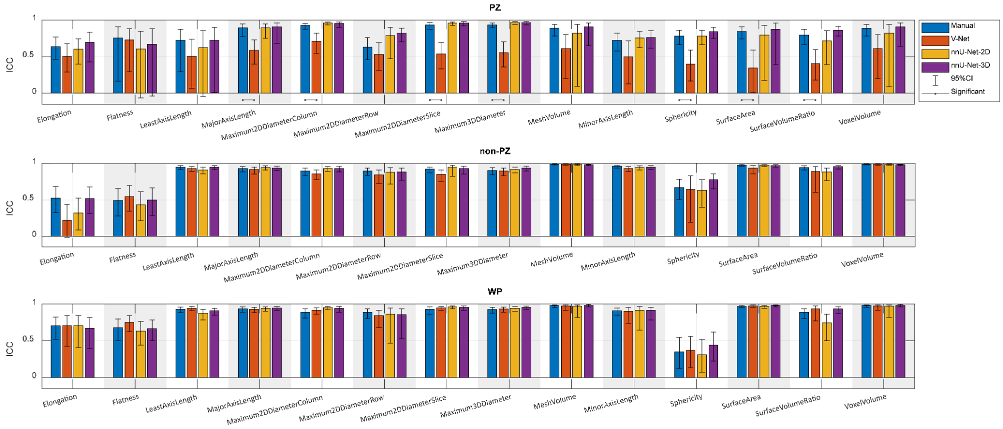


Figure 3. The single score intra-class correlation coefficient (ICC) with the 95% confidence interval (95%CI) of the shape features extracted from the whole prostate gland (WP), peripheral zone (PZ) and the remaining prostate zones (non-PZ) for the investigated methods, where the segmentation post-processing step was included and the segmentation quality control system was not implemented.

The segmented volume (Voxel Volume feature) was further investigated, as it is an important and in-use biomarker for multiple clinical applications[36-38]. Its ICC score, when the segmentation post-processing step was included and the SQCS was not implemented, was 0.888, 0.607, 0.819 and 0.903 in PZ; 0.988, 0.967, 0.986 and 0.983 in non-PZ; and 0.982, 0.975, 0.973 and 0.984 in WP for manual, V-Net, nnU-Net-2D and nnU-Net-3D, respectively. Figure 4 shows that the segmented volume was significantly lower in scan 2 compared to scan 1 for all the methods in PZ and WP (p<0.001) and for nnU-Net-2D in non-PZ (p=0.003). Bland-Altman analysis shows a similar bias for manual and DL-based methods (Figure A3). Median of volume difference between the scans (scan 2 - scan 1) was -4.33%, -3.58%, -5.80% and -3.32% in WP for manual, V-Net, nnU-Net-2D and nnU-Net-3D, respectively. It also shows a small bias between the volumes of the DL-based and manual segmentations in scan 1 (Figure A4) and scan 2 (Figure A5). It was noticed that PZ has higher bias between scans and methods than non-PZ and WP. V-Net has also showed a slightly higher bias between scans and methods than nnU-Net-2D and nnU-Net-3D.

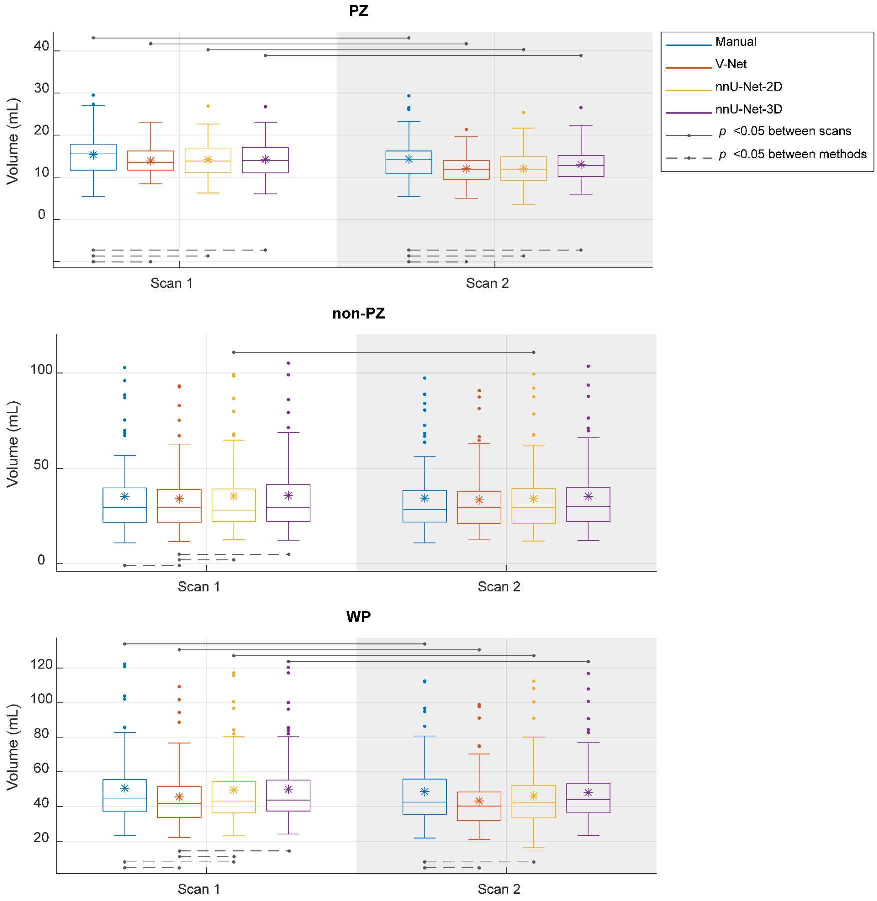


Figure 4. The segmented volume of the whole prostate gland (WP), peripheral zone (PZ) and the remaining prostate zones (non-PZ) from the investigated methods in scan 1 and scan 2.

Discussion

VOI segmentation is an essential step in CAD systems. DL-based methods provide good performance for prostate segmentation, but little is known about their reproducibility. The reproducibility of radiomics shape features can be used as an indicator of the segmentation reproducibility. Therefore, in this paper, we investigated the reproducibility of the shape features extracted from DL-based segmentations of the WP, PZ and non-PZ on T2W MR images acquired with short time intervals (median = 7 days), and compared them to those of manual segmentations. Prostate gland volume is proportionally related to benign enlargement[39] and inversely related to prostate cancer[40]. Both of those conditions usually require long time to develop, thus no significant change in prostate gland volume is expected during a short time interval. Shape features like prostate volume, used to measure the PSA-density (PSA level/prostate volume)[41], are already part of today's clinical risk calculators for prostate cancer[8] and will likely play an important role in future radiomics-based clinical applications. For clinical applications based on multiple scans in time, like active surveillance, it is key that extracted features are both accurate and reproducible.

The DSC values were in line with those expected from the literature[19,20], indicating that the trained networks have state-of-the-art performance. nnU-Net-3D had the best overall segmentation accuracy, while V-Net showed the lowest segmentation accuracy comparable to the manual segmentations. This work extends previous studies showing the excellent performance of nnU-Net, specifically the 3D volume basis model, on a wide variety of medical image segmentation tasks[20]. The DSC values were slightly lower in scan 2 compared to scan 1. This is probably due to the nature of the segmentation training set, which consisted of cases acquired with a scan protocol similar to that of scan 1.

Based on ICCs of the shape features, nnU-Net-2D and nnU-Net-3D were shown to have comparable reproducibility to manual segmentations in all VOIs. WP and non-PZ shown higher ICCs compared to PZ, which was expected due to the low PZ segmentation performance. nnU-Net-3D provided higher ICCs compared to the other CNNs, which was expected as it had the highest segmentation performance among CNNs. Overall, the results show that DL-based segmentation methods can generate highly intra-patient reproducible masks for T2W images of the prostate. Good reproducibility gives potential for picking up changes in the prostate when they appear, which is an important step towards the clinical implementation of prostate CAD systems based on multiple T2W MRI scans.

Including a post-processing step to the segmentation, where only the largest connected component in 3D volume is kept, was shown to remarkably enhance the features reproducibility. Similarly, the implementation of the SQCS significantly increased the reproducibility. Therefore, the implementation of these two post-processing steps in a CAD system pipeline is recommended to assure highly reproducible shape features. In clinical applications, the cases with low segmentation quality score, predicted by the SQCS, should be either referred to a radiologist for manual intervention or re-segmented using another CNN.

One possible explanation for the lower ICC of Elongation, Flatness and Sphericity in WP and non-PZ is that the prostate gland in scan 2 was potentially compressed due to a guiding probe for the biopsy needle inside the patient's rectum during the image acquisition. Moreover, the patients were scanned in prone position during scan 2, in contrast to scan 1, where they were scanned in supine position. The probe and the prone position would indeed not alter the volume of the prostate but might deform its shape slightly. In their study, Osman et al.[42] have investigated the endorectal coil effect on the WP volume and shape during prostate T2W MRI and concluded that despite shape deformation, there is no significant change in the WP volume between including and excluding the endorectal coil. Although the needle guiding probe differs from the endorectal coil, its impact may be expected to be similar. In addition, the prostate gland might deform between scans due to other factors e.g., different bladder/bowel loading, which were not taken into account in this study. The shape deformation may have had an impact on the decision of including or excluding a slice from the segmentation. We noticed that, overall, scan 2 had a lower number of segmented slices than scan 1. Median of the segmented slices number was 14, 14, 14 and 14 in WP for manual, V-Net, nnU-Net-2D and nnU-Net-3D, respectively, in scan 1 and 13, 13.5, 13 and 14 in WP for manual, V-Net, nnU-Net-2D and nnU-Net-3D, respectively, in scan 2. Although the difference between the numbers is small (≈1 slice), it will influence the segmented volumes, which were indeed slightly lower in scan 2 than in scan 1.

The reproducibility of the segmented volume might be the most important among the 14 investigated features. WP volume is used by the radiologist to measure the PSA-density, which is part of today's clinical risk calculators[8] and can be used as a biomarker to evaluate prostate cancer progression and the need for re-biopsy [38]. An alternative biomarker to the traditional PSA-density is the zonal adjusted PSA-density, which depends on the segmented volume from various prostate gland zones, i.e. non-PZ volume[43,44]. Our study shows that the segmented volume feature is highly reproducibility, and in agreement with manual volumes on both zonal and whole prostate-level.

In their work, Schwier et al.[22] used manual segmentations to assess the reproducibility of radiomics features on prostate T2W MR images. Their focus was mainly on the reproducibility of the radiomics textural features under different settings, but they have also included results on some of the shape features reproducibility. Although there is some similarity between their work and ours, our work focused on the reproducibility of DL-based segmentations. Like in our work, Schwier et al. showed that the reproducibility of shape features is high. Furthermore, they showed that the segmented volume reproducibility is higher in WP than in PZ, which was also in line with our findings. The high ICC values found in this work suggest that all the shape features, except for Elongation, Flatness and Sphericity, extracted using DL-based segmentation methods, can be used in clinical applications based on multiple scans without being concerned about their reproducibility.

In this work, we used a dataset from prostate cancer patients referred and scanned according to prevailing guidelines. Consequently, the results represent the reproducibility of the DL-based segmentations in a real clinical setting. Nevertheless, our study has some limitations. The patient cohort was relatively small and it was obtained from a single centre. Conducting a multicentre study in the future might give additional insight on the reproducibility of DL-based segmentation across institutions. Moreover, the manual segmentations in this study have been performed by one reader. A set of manual segmentations, where multiple readers included, will facilitate additional comparisons, which might provide us with more information, but this can be considered for a future work.

Conclusion

We investigated the reproducibility of the shape features extracted from DL-based segmentations of the prostate gland and zones on T2W MR images acquired with short time intervals. The reproducibility of the best-performing DL-based prostate segmentation methods is comparable to that of manual segmentations, which is important for clinical applications based on multiple scans in time.

Data availability

The dataset used in this study are not publicly available due to specific institutional requirements governing privacy protection.

Acknowledgements

We would like to thank Fausto Milletari from the Technical University of Munich (Munich, Germany) and Fabian Isensee from the German Cancer Research Center (Heidelberg, Germany) for making their segmentation methods publicly available. This work was supported by the Norwegian University of Science and Technology (NTNU) Biotechnology (grant number 81770928), the Research Council of Norway (grant number 295013) and Central Norway Regional Health Authority (grant number 17/38297).

Author contributions

M.R.S.S. designed the study, performed experiments, analysed data and wrote the manuscript. T.F.B and M.E. designed the study, supervised work and edited the manuscript. K.M.S, E.S., S.L. and H.B. collected data, gave conceptual advice and edited the manuscript.

Competing interests

The authors declare no competing interests.

References

1 Bray, F. *et al.* Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* **68**, 394-424, doi:10.3322/caac.21492 (2018).

2 Mottet, N. *et al.* EAU-ESTRO-SIOG Guidelines on Prostate Cancer. Part 1: Screening, Diagnosis, and Local Treatment with Curative Intent. *Eur Urol* **71**, 618-629, doi:10.1016/j.eururo.2016.08.003 (2017).

3 Ahdoot, M. *et al.* MRI-Targeted, Systematic, and Combined Biopsy for Prostate Cancer Diagnosis. *N Engl J Med* **382**, 917-928, doi:10.1056/NEJMoa1910038 (2020).

4 Barentsz, J. O. *et al.* ESUR prostate MR guidelines 2012. *Eur Radiol* **22**, 746-757, doi:10.1007/s00330-011-2377-y (2012).

5 Turkbey, B. *et al.* Prostate Imaging Reporting and Data System Version 2.1: 2019 Update of Prostate Imaging Reporting and Data System Version 2. *Eur Urol* **76**, 340-351, doi:10.1016/j.eururo.2019.02.033 (2019).

6 Weinreb, J. C. *et al.* PI-RADS Prostate Imaging - Reporting and Data System: 2015, Version 2. *Eur Urol* **69**, 16-40, doi:10.1016/j.eururo.2015.08.052 (2016).

7 Fascelli, M. *et al.* The role of MRI in active surveillance for prostate cancer. *Curr Urol Rep* **16**, 42, doi:10.1007/s11934-015-0507-9 (2015).

8 Alberts, A. R. *et al.* Prediction of High-grade Prostate Cancer Following Multiparametric Magnetic Resonance Imaging: Improving the Rotterdam European Randomized Study of Screening for Prostate Cancer Risk Calculators. *Eur Urol* **75**, 310-318, doi:10.1016/j.eururo.2018.07.031 (2019).

9 Patel, P., Mathew, M. S., Trilisky, I. & Oto, A. Multiparametric MR Imaging of the Prostate after Treatment of Prostate Cancer. *Radiographics* **38**, 437-449, doi:10.1148/rg.2018170147 (2018).

10 Litjens, G., Debats, O., Barentsz, J., Karssemeijer, N. & Huisman, H. Computer-aided detection of prostate cancer in MRI. *IEEE Trans Med Imaging* **33**, 1083-1092, doi:10.1109/TMI.2014.2303821 (2014).

11 Girometti, R. *et al.* Interreader agreement of PI-RADS v. 2 in assessing prostate cancer with multiparametric MRI: A study using whole-mount histology as the standard of reference. *J Magn Reson Imaging* **49**, 546-555, doi:10.1002/jmri.26220 (2019).

12 Ruprecht, O., Weisser, P., Bodelle, B., Ackermann, H. & Vogl, T. J. MRI of the prostate: interobserver agreement compared with histopathologic outcome after radical prostatectomy. *Eur J Radiol* **81**, 456-460, doi:10.1016/j.ejrad.2010.12.076 (2012).

13 Liu, L., Tian, Z., Zhang, Z. & Fei, B. Computer-aided Detection of Prostate Cancer with MRI: Technology and Applications. *Acad Radiol* **23**, 1024-1046, doi:10.1016/j.acra.2016.03.010 (2016).

14 Wang, S., Burtt, K., Turkbey, B., Choyke, P. & Summers, R. M. Computer aided-diagnosis of prostate cancer on multiparametric MRI: a technical review of current research. *Biomed Res Int* **2014**, 789561, doi:10.1155/2014/789561 (2014).

15 Lemaitre, G. *et al.* Computer-Aided Detection and diagnosis for prostate cancer based on mono and multi-parametric MRI: a review. *Comput Biol Med* **60**, 8-31, doi:10.1016/j.compbiomed.2015.02.009 (2015).

16 Zavala-Romero, O. *et al.* Segmentation of prostate and prostate zones using deep learning : A multi-MRI vendor analysis. *Strahlenther Onkol* **196**, 932-942, doi:10.1007/s00066-020-01607-x (2020).

17 Wang, B. *et al.* Deeply supervised 3D fully convolutional networks with group dilated convolution for automatic MRI prostate segmentation. *Med Phys* **46**, 1707-1718, doi:10.1002/mp.13416 (2019).

18 Khan, Z., Yahya, N., Alsaih, K., Ali, S. S. A. & Meriaudeau, F. Evaluation of Deep Neural Networks for Semantic Segmentation of Prostate in T2W MRI. *Sensors (Basel)* **20**, doi:10.3390/s20113183 (2020).

19 Milletari, F., Navab, N. & Ahmadi, S. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. *2016 Fourth International Conference on 3D Vision (3DV), Stanford, CA*, 565-571, doi:10.1109/3DV.2016.79 (2016).

20 Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J. & Maier-Hein, K. H. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods*, doi:10.1038/s41592-020-01008-z (2020).

21 Schelb, P. *et al.* Classification of Cancer at Prostate MRI: Deep Learning versus Clinical PI-RADS Assessment. *Radiology* **293**, 607-617, doi:10.1148/radiol.2019190938 (2019).

22 Schwier, M. *et al.* Repeatability of Multiparametric Prostate MRI Radiomics Features. *Sci Rep* **9**, 9441, doi:10.1038/s41598-019-45766-z (2019).

23 Yushkevich, P. A. *et al.* User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. *Neuroimage* **31**, 1116-1128, doi:10.1016/j.neuroimage.2006.01.015 (2006).

24 Paszke, A. *et al.* PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Adv Neur In* **32** (2019).

25 van Griethuysen, J. J. M. *et al.* Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer Res* **77**, e104-e107, doi:10.1158/0008-5472.CAN-17-0339 (2017).

26 Pyradiomics community. *Shape Features (3D)*, <https://pyradiomics.readthedocs.io/en/v3.0/features.html#module-radiomics.shape> (2020).

27 Wang, Y. *et al.* Quantitative MRI: Defining repeatability, reproducibility and accuracy for prostate cancer imaging biomarker development. *Magn Reson Imaging*, doi:10.1016/j.mri.2020.12.018 (2020).

28 Sunoqrot, M. R. S. *et al.* A Quality Control System for Automated Prostate Segmentation on T2-Weighted MRI. *Diagnostics (Basel)* **10**, doi:10.3390/diagnostics10090714 (2020).

29 Klein, S. *et al.* Automatic segmentation of the prostate in 3D MR images by atlas matching using localized mutual information. *Med Phys* **35**, 1407-1417, doi:10.1118/1.2842076 (2008).

30 Shrout, P. E. & Fleiss, J. L. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull* **86**, 420-428, doi:10.1037//0033-2909.86.2.420 (1979).

31 McGraw, K. O. & Wong, S. P. Forming inferences about some intraclass correlation coefficients. *Psychological Methods* **1**, 30-46, doi:10.1037/1082-989X.1.1.30 (1996).

32 Stolarova, M., Wolf, C., Rinker, T. & Brielmann, A. How to assess and compare inter-rater reliability, agreement and correlation of ratings: an exemplary analysis of mother-father and parent-teacher expressive vocabulary rating pairs. *Frontiers in Psychology* **5**, doi:10.3389/fpsyg.2014.00509 (2014).

33 Gibbons, J. D. & Chakraborti, S. *Nonparametric statistical inference*. 5th edn, (Taylor & Francis, 2011).

34 Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)* **57**, 289-300, doi:https://doi.org/10.1111/j.2517-6161.1995.tb02031.x (1995).

35 Bland, J. M. & Altman, D. G. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* **1**, 307-310 (1986).

36 Cary, K. C. & Cooperberg, M. R. Biomarkers in prostate cancer surveillance and screening: past, present, and future. *Ther Adv Urol* **5**, 318-329, doi:10.1177/1756287213495915 (2013).

37 Cannon, G. W. & Getzenberg, R. H. Biomarkers for benign prostatic hyperplasia progression. *Curr Urol Rep* **9**, 279-283, doi:10.1007/s11934-008-0049-5 (2008).

38 Nordstrom, T., Akre, O., Aly, M., Gronberg, H. & Eklund, M. Prostate-specific antigen (PSA) density in the diagnostic algorithm of prostate cancer. *Prostate Cancer Prostatic Dis* **21**, 57-63, doi:10.1038/s41391-017-0024-7 (2018).

39 Loeb, S. *et al.* Prostate volume changes over time: results from the Baltimore Longitudinal Study of Aging. *J Urol* **182**, 1458-1462, doi:10.1016/j.juro.2009.06.047 (2009).

40 Al-Khalil, S., Ibilibor, C., Cammack, J. T. & de Riese, W. Association of prostate volume with incidence and aggressiveness of prostate cancer. *Res Rep Urol* **8**, 201-205, doi:10.2147/RRU.S117963 (2016).

41 Benson, M. C. *et al.* Prostate specific antigen density: a means of distinguishing benign prostatic hypertrophy and prostate cancer. *J Urol* **147**, 815-816, doi:10.1016/s0022-5347(17)37393-7 (1992).

42 Osman, M. *et al.* Whole prostate volume and shape changes with the use of an inflatable and flexible endorectal coil. *Radiol Res Pract* **2014**, 903747, doi:10.1155/2014/903747 (2014).

43 Chang, T. H. *et al.* Zonal adjusted PSA density improves prostate cancer detection rates compared with PSA in Taiwanese males with PSA < 20 ng/ml. *BMC Urol* **20**, 151, doi:10.1186/s12894-020-00717-z (2020).

44 Kalish, J., Cooner, W. H. & Graham, S. D., Jr. Serum PSA adjusted for volume of transition zone (PSAT) is more accurate than PSA adjusted for total gland volume (PSAD) in detecting adenocarcinoma of the prostate. *Urology* **43**, 601-606, doi:10.1016/0090-4295(94)90170-8 (1994).

Supplementary material to:

The reproducibility of deep learning-based segmentation of the prostate gland and zones on T2-weighted MR images

By: Mohammed R. S. Sunoqrot[1,*], Kirsten M. Selnæs[1,2], Elise Sandsmark[2], Sverre Langørgen[2], Helena Bertilsson[3,4], Tone F. Bathen[1,2] and Mattijs Elschot[1,2]

[1] Department of Circulation and Medical Imaging, NTNU, Norwegian University of Science and Technology, Trondheim, 7030, Norway.

[2] Department of Radiology and Nuclear Medicine, St. Olavs Hospital, Trondheim University Hospital, Trondheim, 7030, Norway.

[3] Department of Cancer Research and Molecular Medicine, NTNU, Norwegian University of Science and Technology, Trondheim, 7030, Norway.

[4] Department of Urology, St. Olavs Hospital, Trondheim University Hospital, Trondheim, 7030, Norway.

*** Corresponding author:**

Name: Mohammed R. S. Sunoqrot

Email: mohammed.sunoqrot@ntnu.no

Appendix

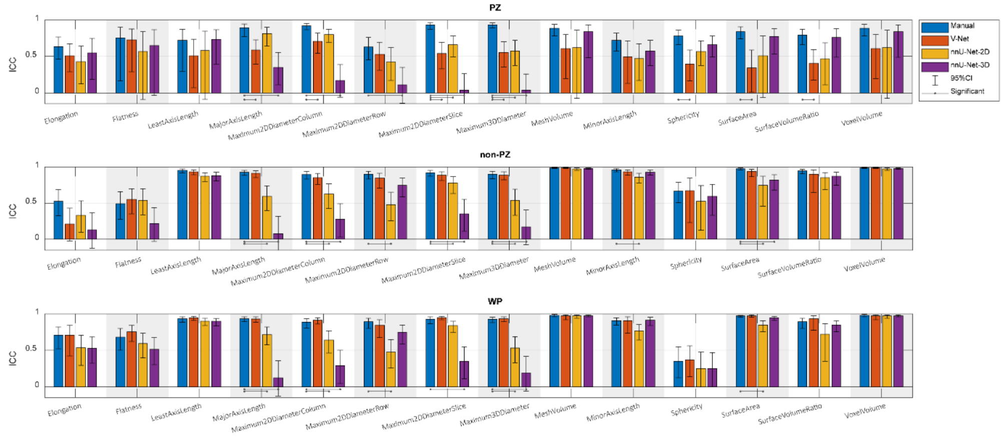


Figure A1. The single score intra-class correlation coefficient (ICC) with its 95% confidence interval (95%CI) of the shape features extracted from the whole prostate gland (WP), peripheral zone (PZ) and the remaining prostate zones (non-PZ) for the investigated methods, where the segmentation post-processing step was skipped and the segmentation quality control system was not implemented.

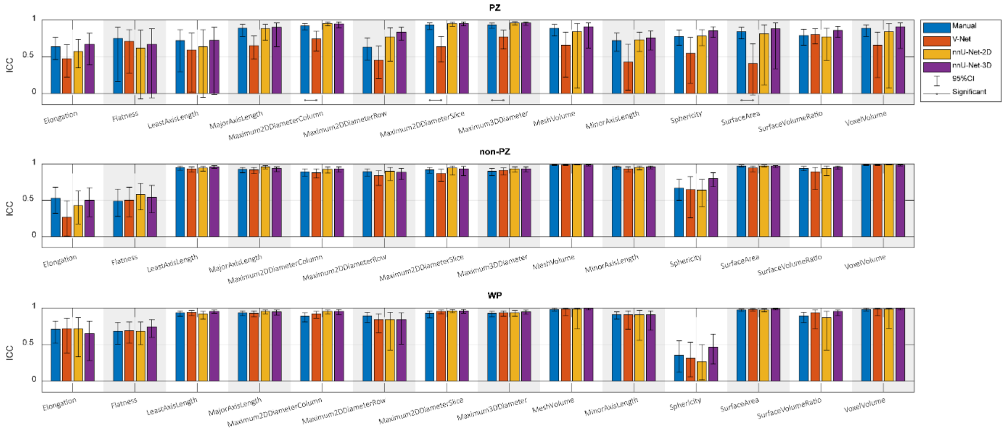


Figure A2. The single score intra-class correlation coefficient (ICC) with its 95% confidence interval (95%CI) of the shape features extracted from the whole prostate gland (WP), peripheral zone (PZ) and the remaining prostate zones (non-PZ) for the investigated methods, where the segmentation post-processing step was included and the segmentation quality control system was implemented. The patients with a quality score less than 85 for scan 1 or/and scan 2 were excluded.

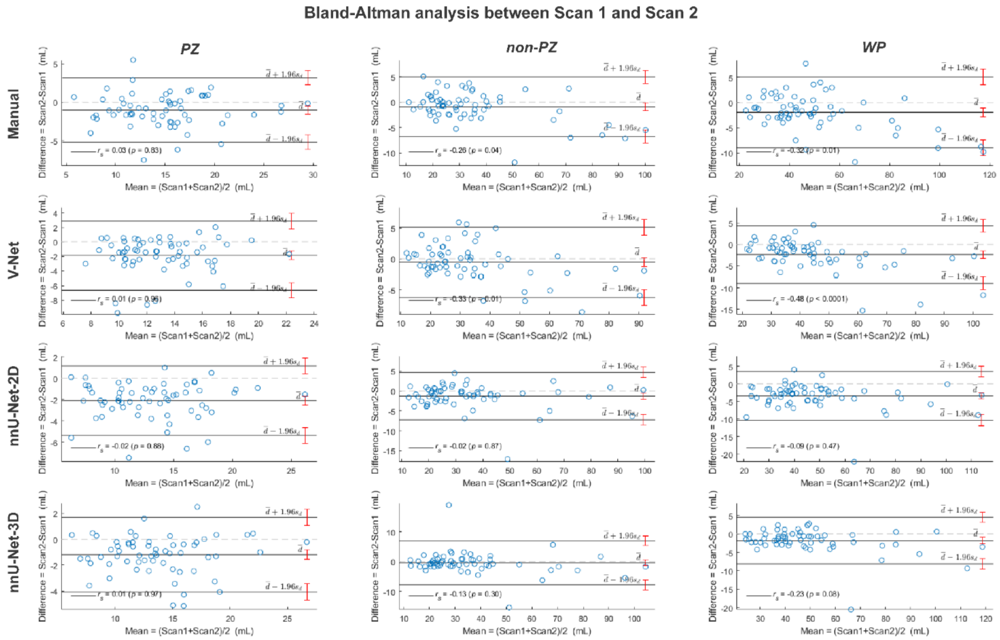


Figure A3. The Bland-Altman plots for the agreement between scan 1 and scan 2 volumes from the whole prostate gland (WP), peripheral zone (PZ) and the remaining prostate zones (non-PZ) for the investigated methods.

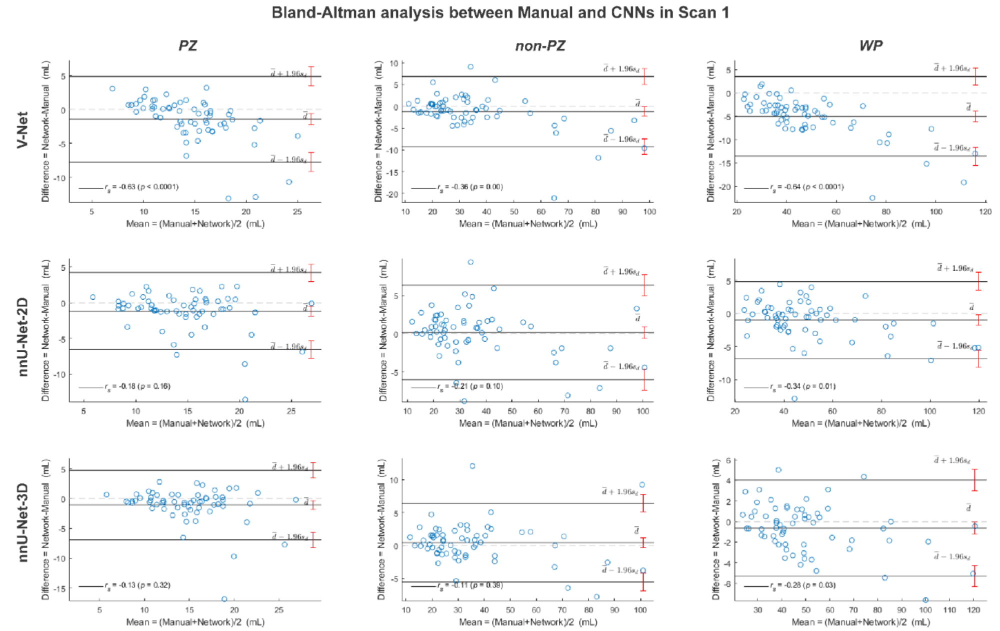


Figure A4. The Bland-Altman plots for the agreement between manual and rest of the investigated methods volumes from the whole prostate gland (WP), peripheral zone (PZ) and the remaining prostate zones (non-PZ) in scan 1.

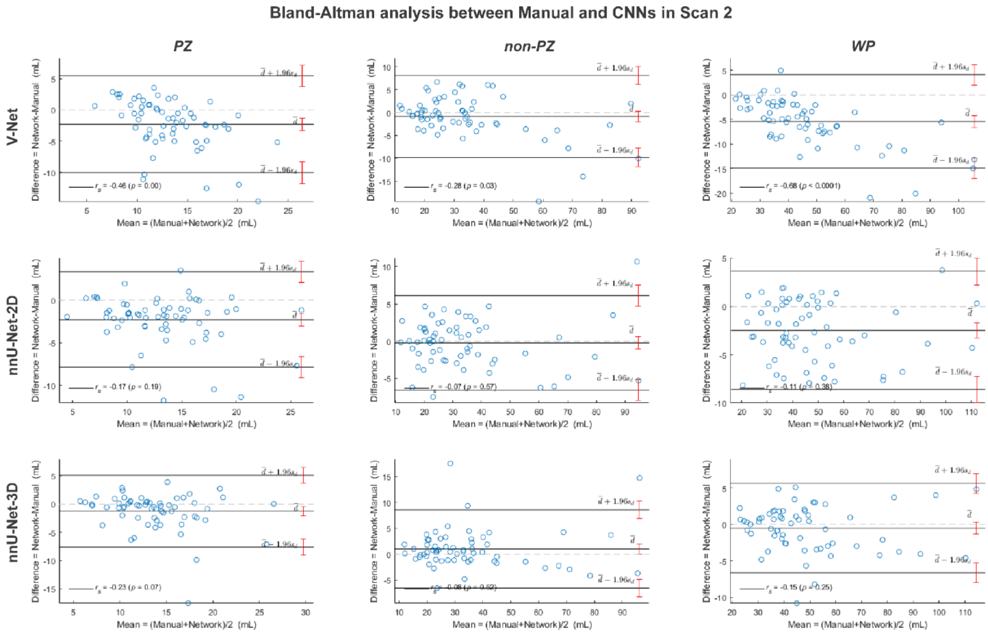


Figure A5. The Bland-Altman plots for the agreement between manual and rest of the investigated methods volumes from the whole prostate gland (WP), peripheral zone (PZ) and the remaining prostate zones (non-PZ) in scan 2.

ISBN 978-82-326-6531-0 (printed ver.)
ISBN 978-82-326-6417-7 (electronic ver.)
ISSN 1503-8181 (printed ver.)
ISSN 2703-8084 (online ver.)

NTNU
Norwegian University of
Science and Technology