Aus dem Deutschen Krebsforschungszentrum (DKFZ) Heidelberg
(Vorstand: Prof. Dr. med. Michael Baumann, Ursula Weyrich)
Abteilung Medizinische Bildverarbeitung (MIC)
(Leiter: Prof. Dr. rer. nat. Klaus Hermann Maier-Hein)
in Zusammenarbeit mit der Abteilung Radiologie
(Leiter: Prof. Dr. med. Dipl.-Phys. Heinz-Peter Schlemmer)

Data-Centric Artificial Intelligence for Enhanced Prostate Cancer Diagnosis on Magnetic Resonance Images

Inauguraldissertation
zur Erlangung des Doctor scientiarum humanarum (Dr. sc. hum.)
an der
Medizinischen Fakultät Heidelberg
der
Ruprecht-Karls-Universität

vorgelegt von
Bálint Kovács
aus
Békéscsaba, Ungarn

2025

Dekan: Prof. Dr. rer. nat. Michael Boutros
Doktorvater: Prof. Dr. rer. nat. Klaus Hermann Maier-Hein

To my wife Dora, she knows why.
To our daughter Mira, hoping that she will overperform me soon.

Contents

Acronyms **i**

List of Figures **iv**

List of Tables **viii**

1 Introduction **1**
1.1 Motivation 1
1.2 Background 5
1.2.1 Diagnostic Pathway of Prostate Cancer 6
1.2.2 Multi-parametric Prostate MRI 16
1.2.3 Multi-Modal Medical Image Analysis with Convolutional Neural Networks 27
1.3 Related Work 44
1.3.1 Research on Soft Tissue Deformations of the Prostate 45
1.3.2 Research on Multi-Modal Misalignments in Prostate MRI 46
1.4 Objectives and Contributions 47
1.4.1 Research Objectives 47
1.4.2 Summary of Main Contributions 50
1.5 Outline 50

2 Materials and Methodology **53**
2.1 Characteristics of Datasets 53
2.1.1 In-House Data 53
2.1.2 PROSTATEx Dataset 56
2.2 Soft Tissue Deformations of the Prostate (Objective #1) 57

2.2.1 Study Cohort . . . 57
2.2.2 Biomechanical Model Creation . . . 58
2.2.3 Experimental Setting . . . 60
2.2.4 AI Model Developement . . . 62
2.3 Multi-Modal Misalignments in Prostate MRI (Objective #2) . . . 68
2.3.1 Study Cohort . . . 68
2.3.2 Registration Techniques Used for Model Training . . . 69
2.3.3 Misalignment Augmentation: A Data-Centric Alternative to Multi-Modal Registration . . . 71
2.3.4 Evaluation Concept for Strategies Addressing Spatial Alignment Errors . . . 72
2.3.5 Experimental Setting . . . 74
2.3.6 AI Model Developement . . . 76

3 Results **81**
3.1 Soft Tissue Deformations of the Prostate (Objective #1) . . . 81
3.1.1 Highly realistic Images Passing the Turing Test - Qualitative Results 81
3.1.2 High Applicability . . . 84
3.1.3 Approaching Towards Radiologists' Performance - Quantitative Results . . . 85
3.2 Multi-Modal Misalignments in Prostate MRI (Objective #2) . . . 90
3.2.1 Maximized Robustness Achieved Through the Combination of Registration and Misalignment Augmentation . . . 90
3.2.2 Achieving Radiologist-Level Diagnostic Performance by Addressing Multi-Modal Misalignments . . . 92
3.2.3 Visualization of Detected Punctate Lesion . . . 94
3.2.4 B-spline Registration on Par with Method Using Human Ground Truth Segmentation . . . 96

4 Discussion **99**
4.1 Soft Tissue Deformations of the Prostate (Objective #1) . . . 100
4.1.1 Diagnostic Benefit of Simulating Physiological Deformations . . . 100
4.1.2 Localized Performance Gains from Targeted Organ Deformations . 101
4.1.3 Importance of Visual Realism and Label Preservation . . . 102
4.1.4 High Applicability with Practical Considerations . . . 103
4.1.5 Extending Anatomy-Informed Transformations Beyond Organs . . 104
4.1.6 Bridging the Gap in Spatial Transformations for Online Data Augmentation . . . 105

4.2 Multi-Modal Misalignments in Prostate MRI (Objective #2) 107
4.2.1 The Importance of Ground Truth Consistency Across Image Modalities for Diagnostic Performance 108
4.2.2 Enhanced Robustness Against Alignment Errors Gained Through Model Training 108
4.2.3 Synergistic Gains Through the Combination of Registration and Misalignment Augmentation 109
4.2.4 Potential Improvements in Small Lesion Detection Sensitivity . . . 110
4.2.5 Multi-Modal Alignment Errors Beyond Prostate MRI 111
4.2.6 Conceptual Summary of Multi-Modal Alignment Error Handling Strategies 112
4.3 Future of Application-specific Augmentations in Medicine 114
4.4 Conclusion 116

5 Summary **119**

6 Zusammenfassung **121**

Bibliography **123**

Own Contributions and Publications **153**
Own share in data acquisition and data analysis 153
Own Publications 156

Appendix **157**

A Dataset Properties **159**
A.1 Demographic Tables for In-House Datasets 159
A.2 MRI Protocol 161

B Dokumentation zur Verwendung KI-basierter elektronischer Hilfsmittel **163**

Acknowledgement **165**

Eidesstattliche Versicherung (Affidavit) in German **169**

Angabe zu verwendeten KI-basierter Elektronischer Hilfsmittel **171**

Acronyms

ADC	apparent diffusion coefficient
AI	artificial intelligence
ANN	artificial neural network
AUROC	area under the receiver operating characteristic curve
pAUROC	partial area under the receiver operating characteristic curve
avgFPs/scan	average number of false positives per scan
BPH	benign prostatic hyperplasia
CA	contrast agent
CAD	computer-aided diagnosis
CNN	convolutional neural network
CPU	central processing unit
CT	computed tomography
DA	data augmentation
DCE	dynamic contrast-enhanced imaging
DL	deep learning
DNN	deep neural network
DRE	digital rectal examination
DWI	diffusion-weighted imaging
EPI	echo-planar imaging
FE	finite element
FID	free induction decay

FoV	field of view
FROC	free-response receiver operating characteristic
FSE	fast-spin-echo
GGG	Gleason Grade Group
GPU	graphical processing unit
GS	Gleason Score
GRE	gradient echo
GT	ground truth
ISUP	International Society of Urological Pathology Consensus
IoU	Intersection over Union
MI	mutual information
MLP	multi-layer perceptron
MRI	magnetic resonance imaging
bpMRI	bi-parametric MRI
mpMRI	multi-parametric MRI
M_Z	longitudinal magnetization
M_T	transverse magnetization
PCa	prostate cancer
csPCa	clinically significant PCa
PI-RADS	Prostate Imaging Reporting and Data System
PSA	prostate-specific antigen
PZ	peripheral prostate zone
RF	radiofrequency
ROC	receiver operating characteristic
RoI	Region of Interest
SE	spin echo
SNR	signal-to-noise ratio
T_E	echo time
T_R	repetition time
TRUS	transrectal ultrasound

TSE	turbo-spin-echo
TZ	transitional prostate zone
T1w	T1-weighted
T2w	T2-weighted
US	ultrasound
5fCV	5-fold cross-validation

List of Figures

1.1 Statistics on the most common cancer type by incidence in men. 6
1.2 The diagnostic pathway of prostate cancer. 7
1.3 Prostate Imaging Reporting and Data System (PI-RADS) assessment utilizing MRI modalities in a prostate zone-specific decision tree. 10
1.4 Visualization of the five Gleason patterns used in the Gleason grading system for prostate cancer assessment. 14
1.5 The spin echo (SE) pulse sequence used for T2-weighted imaging. 17
1.6 Fast Spin Echo (FSE) – also known as Turbo Spin Echo (TSE) – pulse sequence. 18
1.7 Diffusion-Weighted Imaging (DWI) integrated into the Spin-Echo (SE) pulse sequence. 20
1.8 Echo-Planar Imaging (EPI) pulse sequence 22
1.9 Saturation recovery pulse sequence used for indirect measurement of T_1 relaxation time. 23
1.10 Pulse sequence with short repetition time (T_R) as used in DCE imaging. . 24
1.11 Schematic overview of the image registration pipeline. 30
1.12 U-Net architecture for semantic segmentation. 41
1.13 Example of receiver operating characteristic curves for multiple classifier models. 42
1.14 Example free-response receiver operating characteristic (FROC) curves . . 43

2.1 Anatomy-informed transformation pipeline. 59
2.2 Integration of the anatomy-informed transformation into the nnU-Net training pipeline. 64
2.3 The generated U-Net topology for the anatomy-informed training configuration. 66

2.4 Illustration of the three registration techniques used to generate distinct multi-modal datasets for model training. 70
2.5 Overview of the evaluation concept for strategies addressing spatial alignment errors in prostate MRI. 73
2.6 Overview of experimental configurations using strategies addressing alignment error for assessing the impact on prostate cancer diagnosis. 75
2.7 U-Net topology generated by nnU-Net for training configurations that incorporate strategies addressing multi-modal alignment errors. 79

3.1 Example of anatomy-informed transformation simulating bladder shape changes on a prostate MRI scan. 82
3.2 Example of anatomy-informed transformation simulating rectal shape changes on a prostate MRI scan. 83
3.3 Patient-level receiver operating characteristic (ROC) curves comparing model performance across augmentation strategies on the independent test set. 86
3.4 Predictive performance comparison of trained models and radiologists on the receiver operating characteristic curve. 93
3.5 Visualization of manual annotations and segmentation predictions of the convolutional neural network (CNN) for two distinct and clinically relevant prostate cancer lesions 95
3.6 Probability density functions of Dice scores between manual lesion annotations in T2-weighted (T2w) images and apparent diffusion coefficient (ADC) maps across different registration strategies. 96
3.7 Impact of the different registration strategies on patient-level diagnostic performance (receiver operating characteristic (ROC) analysis). 97

4.1 Lesion growth simulation using the anatomy-informed transformation . . 104
4.2 Comparison of commonly used spatial transformation strategies in terms of anatomical realism, degree of control, and their ability to introduce lesion shape variability. 105
4.3 Conceptual visualization of the spatial overlap of multi-modal ground truth information under different strategies. 112

List of Tables

1.1 An overview of the relationship between Gleason patterns, Gleason scores (GS), and Gleason Grade Groups (GGG). 15
1.2 Recommended technical parameters for MRI sequences according to PI-RADS v2.1 guidelines. 26
1.3 Categories of medical image registration based on subject scope and modality type. 29

2.1 Distribution of patients with and without clinically significant prostate cancer (csPCa) across the training and test sets. 57
2.2 Parameters used in the implementation of the anatomy-informed transformation. 64
2.3 Cohort composition across training and test sets, stratified by both the prevalence of clinically significant prostate cancer and dataset origin. . . . 69

3.1 Perceived realism of anatomy-informed deformations during the Turing test. 84
3.2 Patient-level diagnostic performance of nnU-Net models trained with different data augmentation schemes. 85
3.3 Specificity of nnU-Net models trained with different augmentation strategies. 87
3.4 Lesion-level performance of nnU-Net models trained with different augmentation strategies. 88
3.5 Number of detected clinically significant prostate cancer lesions stratified by prostate zone and augmentation strategy. 89
3.6 Patient-level area under the receiver operating characteristic curve results on the independent test set. 91
3.7 Patient-level area under the receiver operating characteristic curve results on the independent test set stratified by dataset. 92

4.1 Segmentation performance (five-fold cross-validation) of the winning solution of AutoPET3 Challenge for the baseline model trained with and without misalignment augmentation. Results from the study by Rokuss et al. (2024) are reproduced under the Creative Commons Attribution 4.0 License. 111

A.1 Demographic and clinical characteristics of the in-house cohort #1. Table adapted from our previously published work Kovacs et al. (2023b), reused under the Creative Commons Attribution 4.0 License. 159
A.2 Demographic and clinical characteristics of the in-house cohort #2. The cohort is used for our study Kovacs et al. (2023a) and values have been already partially published. 160
A.3 Detailed sequence parameters for biparametric MRI of the in-house cohort #2 utilized for training of the deep learning system. Ranges represent the 5th and 95th percentile. 161

Chapter 1

Introduction

1.1 Motivation

The introduction of magnetic resonance imaging (MRI) into the prostate cancer (PCa) diagnostic pathways has been a major development (Lomas and Ahmed, 2020; Turkbey and Choyke, 2018). Its capability to localize suspicious lesions through the entire prostate gland paved the way for more accurate biopsy, ultimately leading to increased sensitivity and reduced false negative rate for the detection of clinically significant PCa (csPCa) (Ahdoot et al., 2020; Siddiqui et al., 2015; Ahmed et al., 2017; Valerio et al., 2015). Biopsy risks – such as infection, pain, and bleeding – are also reduced, either due to the reduced number of biopsy cores required (Valerio et al., 2015), or because the biopsy can be omitted in the case of a negative MRI (Ahmed et al., 2017).

The interpretation of MRI images has been standardized by the Prostate Imaging Reporting and Data System (PI-RADS), currently in its second version (Weinreb et al., 2016; Turkbey et al., 2019). PI-RADS has been widely adopted because it is not only a standardized scoring system for the recommendation of subsequent biopsy, but it also contains recommendations on technical specifications of image acquisition (Engels et al., 2020), provides a framework for education of radiologists (Israël et al., 2020), and fosters active communication between radiologists and urologists (Venderink et al., 2020; Immerzeel et al., 2022).

Although PI-RADS has become a widely-accepted standard (Barentsz et al., 2016), interreader variability still remains high (Westphalen et al., 2020). Furthermore, the diagnostic accuracy for lesions scored as intermediate and high risk (PI-RADS 3 and 4) is still not ideal, along with a considerably high false positive rate due to image features related to benign hyperplasia, inflammation, prior trauma, and infection (Turkbey and Choyke, 2018; Panebianco et al., 2018). In addition to difficulties in the stratification between benign and malignant tissue conditions, there are practical challenges, too. The diagnostic performance is dependent on radiologist expertise (Kasel-Seibert et al., 2016) and highly influenced by the MRI image quality, which can vary significantly between different imaging settings (De Rooij et al., 2024; Giganti et al., 2020; Woernle et al., 2024). Inappropriate magnetic field gradient strength, patient motion, and strong transitions in the material properties (referred to as tissue susceptibility, like metal-tissue and air-tissue

transitions) can lead to imaging artifacts. All of these conditions contribute to a highly complex and time-consuming process containing subjective and expertise-dependent factors.

Addressing these challenges, artificial intelligence (AI)-based computer-aided diagnosis (CAD) systems recently gained huge interest in the interpretation of prostate MRI sequences (Bhattacharya et al., 2022; Twilt et al., 2021). Deep learning (DL)-based convolutional neural networks (CNNs) have been proven to be a powerful technology in the field of medical image analysis by being useful in many complex clinical problems (Zhou et al., 2020). Approaches based on semantic segmentation are currently the most popular in the biomedical image analysis community due to their inherent interpretability and clinical value (Maier-Hein et al., 2018; De Fauw et al., 2018; Bernard et al., 2018; Nikolov et al., 2021). The self-configuring framework nnU-Net (Isensee et al., 2021) has made semantic segmentation even more applicable while providing a state-of-the-art performance (Ma, 2021).

Currently, the most successful CAD systems for PCa diagnosis use the approach of semantic segmentation to predict a heatmap image with high values correlating to the presence of malignant PCa lesions (Alkadi et al., 2019; Arif et al., 2020; Bosma et al., 2021a,b, 2023; Cao et al., 2019; De Vente et al., 2020; Hosseinzadeh et al., 2021; Kohl et al., 2017; Netzer et al., 2023, 2021; Saha et al., 2021a,b; Sanyal et al., 2020; Schelb et al., 2019, 2021). Supporting radiologists with AI in reading prostate MRI scans has been shown not only to increase diagnostic accuracy, but also to reduce reading time and more importantly inter-rater variability (Winkel et al., 2021; Penzkofer, 2024). These systems, when operating alone, have also been shown to perform comparably to radiologists (Cao et al., 2019; Netzer et al., 2021; Schelb et al., 2021, 2019) or even outperform radiologists with a certain level of expertise on large- and international scale (Saha et al., 2024).

Despite the outstanding progress in computer-aided PCa diagnosis in recent years, there is still significant room for improvement before these systems can be confidently established as superior to expert radiologists. Radiologists possess a distinct advantage due to their specialized training, which not only focuses on analyzing rare and special cases but also incorporates an awareness of challenging circumstances that can influence, distort, or limit the diagnostic information obtained during imaging (Rosenkrantz, 2016; Vilanova et al., 2018). This advantage is particularly evident considering that state-of-the-art CAD systems for PCa diagnosis use relatively small datasets, which typically consist of some hundreds or some thousands of exams, in contrast to the hundreds of thousands or millions of images available in the natural image processing domain (Deng et al., 2009; Lin et al., 2014). However, these studies have predominantly focused on improving predictive performance through the application of well-established model-centric techniques – such as optimized training procedures, architectural modifications,

and post-processing strategies – while paying comparatively less attention to the intrinsic characteristics of the data itself. Consequently, these systems are lagging behind in the diagnostic performance of highly skilled radiologists. To bridge this gap, addressing data-related challenges – such as patient-specific factors and variability in imaging conditions – into AI model training could enhance the robustness and generalizability of these systems.

Data-centric Challenge #1: Functional soft tissue deformations

One important data-related challenge that radiologists account for but is highly overlooked in AI model training is soft tissue deformations. The geometrical appearance of the prostate is constantly altered by muscle contractions, respiration, and primarily the physiological function of the directly adjacent organs, the bladder and the rectum (Boubaker and Ganghoffer, 2017). Among these sources, the influence of the rectum on the prostate is especially prominent due to its large motion (Boubaker and Ganghoffer, 2017) and the fact that around 70 % of the lesions are located in the adjacent peripheral prostate zone (PZ) (Ali et al., 2022). Furthermore, several studies have demonstrated a detrimental effect of rectal filling and distension with increased distortion and motion artifacts, ultimately leading to a decrease in image quality (Arnoldner et al., 2022; Caglic et al., 2017; Plodeck et al., 2020) These anatomical alterations further increase the already high variability in prostate lesion size and shape originating from the inherent tumor characteristics (Turkbey and Choyke, 2018), both of which significantly influence the assessment of prostate lesions on MRI according to PI-RADS (Chesnais et al., 2013; Turkbey et al., 2019).
However, each exam in the training datasets provides only one snapshot of all possible prostate morphologies arising from continuous functional soft tissue deformations, which cannot be identically captured in any repeat or subsequent examinations. Given prostate MRI datasets are limited in size (Bhattacharya et al., 2022; Twilt et al., 2021), such morphological changes can be underrepresented. That limits AI model ability to generalize across the full spectrum of lesion characteristics. To address this limitation, data augmentation (DA) (Shorten and Khoshgoftaar, 2019) is essential in the success of data-hungry DL-based approaches in medical image analysis. More importantly, DA provides an opportunity to introduce inductive bias into model training, such as incorporating soft tissue deformations to increase morphological diversity. Despite this potential, the DA schemes employed in state-of-the-art approaches for PCa diagnosis rely on simplistic spatial transformations – such as translation, rotation, and scaling – that affect the images globally, but introduce no local tissue deformations, thereby leaving the morphological diversity of the data unchanged.

Data-centric Challenge #2: Image misalignments between MRI sequences

Another data-related challenge where radiologists' cognitive ability plays an especially crucial role in accurate diagnosis is image misalignments between MRI modalities*. However, it has gotten less attention in AI model development.
Following PI-RADS, standard prostate MRI imaging employs a multi-parametric MRI (mpMRI) protocol. This involves capturing multiple modalities such as T2w, DWI, and DCE. Each MRI modality provides complementary information about the prostate tissue that is essential for a comprehensive diagnosis (Weinreb et al., 2016). However, performing mpMRI inevitably results in misalignments between the sequences. The scan itself takes at least 20 min, but typically 30–45 min (Alkadi et al., 2019; Giganti et al., 2022; Hötker et al., 2022; Kuhl et al., 2017). This especially long examination time is sufficient for global body position changes due to patient movements, as well as various involuntary soft tissue deformations discussed in **Data-centric Challenge #1**, even for slow processes like bladder filling. Moreover, geometric differences arise between MRI sequences due to shifted tissue boundaries caused by unique sequence contrast characteristics (Brown et al., 2014) and sequence-specific image artifacts, particularly between the T2w and DWI sequences (De Rooij et al., 2024; Giganti et al., 2020; Woernle et al., 2024).
Radiologists cognitively compensate for these misalignments by matching information across sequences, but it remains unclear whether AI models have this capability. Clinical segmentations used as lesion ground truth on one MRI sequence will be unintentionally inaccurate on other sequences due to multi-modal image misalignments, particularly for smaller punctate lesions. As a consequence, these geometrical inaccuracies between the ground truth and certain sequences can reduce CAD performance, particularly approaches relying on semantic segmentation, which are dependent on spatial information. In multi-modal approaches, co-registration of the image modal-

*To ensure consistency and clarity in this dissertation for readers from all relevant fields, **I use the term MRI modality to refer to MRI contrast mechanisms**, which describe different MRI techniques based on their underlying contrast properties, such as T2-weighted (T2w) imaging for spin relaxation, diffusion-weighted imaging (DWI) for diffusion, and dynamic contrast-enhanced imaging (DCE) imaging for perfusion. Although the term modality traditionally refers to distinct imaging techniques (e.g., CT, MRI, PET), different MRI contrast mechanisms provide complementary tissue characteristics (Brown et al., 2014; Kirimtat et al., 2020; Yang et al., 2020) that are particularly relevant for PCa diagnosis and assessment. This usage **aligns with image co-registration terminology**, which is a key aspect of this dissertation, where different MRI contrast mechanisms have to be considered as multi-modal images from a technical perspective. Additionally, I use the term MRI sequence to refer to the specific pulse sequence (e.g., free induction decay (FID), saturation recovery, spin echo (SE), gradient echo (GRE)) used to achieve a particular contrast characteristic. Following this terminology, details related to sequence design and imaging contrast based on tissue properties can be clearly distinguished, reducing ambiguity.

ities is typically performed as a preprocessing step to achieve better alignment of all sequences with the ground truth segmentation. However, there is no consensus among research groups on the necessity of registration, and its influence on diagnostic performance has not yet been explored. Moreover, image registration is inherently imperfect, especially in prostate MRI, due to factors such as highly anisotropic voxel dimensions, substantial local soft tissue deformations, and susceptibility artifacts. As a result, residual misalignments persist even after registration. Strategies to handle those remaining alignment errors, as well as strategies alternative to registration, have not been explored so far.

Addressing these complex data-related challenges, particularly soft tissue deformations and image misalignments between MRI sequences during AI model training for PCa diagnosis could increase diagnostic performance and narrow the gap to expert radiologists. This work underscores the importance of moving beyond traditional model-centric strategies toward a more holistic integration of data-centric approaches in AI model development.

1.2 Background

This section introduces the fundamental knowledge necessary to understand the complexity of the problems addressed in this thesis. First, it outlines the complex diagnostic pathway of PCa in Section 1.2.1, detailing each examination or intervention step and the clinical information derived from it. As the technical methods developed in this thesis are applied to prostate MRI images, Section 1.2.2 further describes the mpMRI sequences used, providing insight into the corresponding imaging contrasts and their implications for multi-modal alignment errors and soft tissue deformations. Finally, Section 1.2.3 introduces the fundamentals of current state-of-the-art prostate MRI image analysis by presenting the deep learning pipeline for multi-modal image analysis using CNNs.

1.2.1 Diagnostic Pathway of Prostate Cancer

Prostate cancer (PCa) is one of the most common cancers affecting men worldwide. According to GLOBOCAN (Bray et al., 2024), PCa was the second most frequently diagnosed cancer with an incidence of 14.2 %, the fifth leading cause of cancer-related mortality with 7.3 % of cancer deaths, and the most frequently diagnosed cancer in 118 out of 185 countries (see Fig. 1.1) among men in 2022.

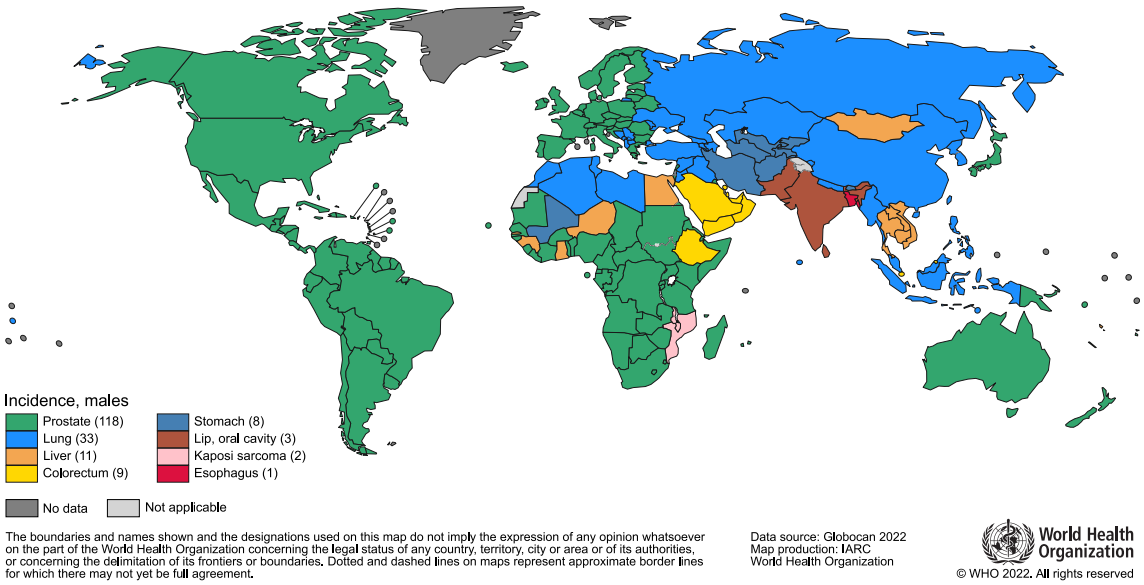

Figure 1.1: Statistics on the most common cancer type by incidence in men, based on data from the GLOBOCAN 2022 report. In 118 out of 185 countries indicated in ■ green, prostate cancer had the highest incidence rate among all cancer types. The image is adapted from the original publication (Bray et al., 2024) with permission from the publisher John Wiley and Sons.

Given its high incidence, the diagnostic pathway for PCa remains a focus for continuous improvement to enhance early detection and clinical outcomes. The current diagnostic pathway of PCa is complex and involves multiple steps, including standard urological first-line examinations, radiological imaging and assessment, image-guided urological interventions, and histopathological evaluation. This subsection introduces all the steps of the current diagnostic pathway, as depicted in Fig. 1.2.

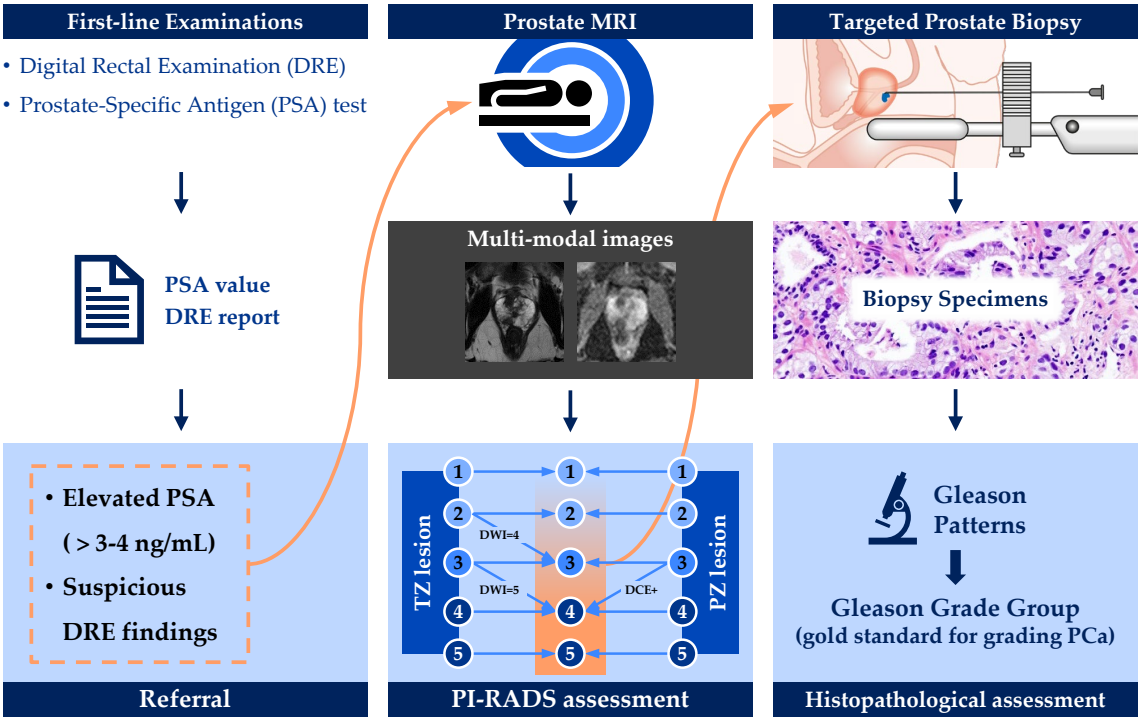

Figure 1.2: The diagnostic pathway of prostate cancer. The figure illustrates the diagnostic workflow across three main stages, shown from left to right: (1) Early detection of prostate cancer includes digital rectal examination (DRE) and prostate-specific antigen (PSA) testing during routine urological evaluation. If either DRE findings are suspicious or PSA levels are elevated, the patient is referred for further examination. (2) Prostate MRI, along with Prostate Imaging Reporting and Data System (PI-RADS) assessment, has become a widely accepted standard for evaluating the presence of potential prostate cancer lesions prior to biopsy. Each suspicious lesion is scored by prioritizing specific MRI modalities based on its zonal location. Patients with lesions scored as PI-RADS ≥ 4 are referred for biopsy, while those with lower scores undergo discussion regarding further surveillance or PSA-based follow-up biopsy. (3) Suspicious lesions identified on multiparametric MRI are directly targeted by the biopsy needle. The extracted tissue samples undergo histopathological assessment using the Gleason Group Grading system, which currently holds the highest prognostic value in clinical practice and is considered the gold standard for grading prostate cancer aggressiveness. The MRI scanner is adapted CC0 1.0, the histopathology image is adapted CC BY-SA 3.0, and the trans-rectal biopsy figure from *Cancer Research UK uploader* is adapted CC BY-SA 4.0, all via Wikimedia Commons. The MRI images are taken from our previously published work Kovacs et al. (2023b), reused under the Creative Commons Attribution 4.0 License.

First-line Examinations

Digital rectal examination (DRE) is a routine procedure performed during a urological visit. The clinician physically examines the prostate gland by inserting a finger into the rectum, assessing for physical abnormalities such as gland enlargement, signs of induration, or contour irregularities. The findings from DRE can indicate whether further diagnostic evaluation is necessary, but its prognostic value for PCa lacks accuracy, and its overall benefit is questionable (Bouras, 2024; Shish and Zabell, 2024; Ying et al., 2023).

Screening for PCa underwent a major change with the introduction of prostate-specific antigen (PSA) testing as a first-line examination. PSA is a protein primarily expressed in prostatic tissues, and its serum levels correlate to some extent with the presence of malignant tissue in the prostate (Carvalhal et al., 2010; Lojanapiwat et al., 2014; Vickers et al., 2010).

Both suspicious findings from DRE or an elevated PSA level (typically $\geq$ 3–4 ng/mL) reffer for further examination. First-line examinations contribute to a cost-effective process improving early PCa detection and thereby reducing the incidence of advanced-stage diagnoses and mortality. However, their specificity is limited, leading to a risk of overdiagnosis (Banerjee et al., 2016; Ilic et al., 2018), particularly for patients with benign hyperplasia or inflamatic conditions, as well as in cases of tumors that would never have become symptomatic, resulting in unnecessary treatment and side effects such as incontinence and impotence following surgery.

Multiparametric Prostate MRI and PI-RADS

Before the introduction of MRI into the diagnostic pathway, patients with suspected PCa following first-line examinations underwent systematic transrectal ultrasound (TRUS)-guided biopsy. This procedure involves obtaining prostate tissue samples transperineal or transrectal, using a biopsy needle, guided by ultrasound (US), following a predefined pattern and a fixed number of biopsy cores. However, systematic TRUS-guided biopsy alone is suboptimal due to frequent false-negative findings due to the following aspects: First, systematic biopsy has difficulties to sample from the entire prostate gland by design, particularly the anterior apex (Moussa et al., 2010). More importantly, the tissue contrast of the systematic TRUS-guided biopsy is insufficient to detect all csPCa lesions, especially small tumors (Ahmed et al., 2017). Increasing the number of biopsy cores improves the ability to detect smaller lesions by enhancing spatial sampling density. However, this also simultaneously increases the risk of complications, including infections (e.g., sepsis), antibiotic resistance, reduced sexual function, and urinary dysfunctions such as urinary retention (Miah et al., 2018; Loeb et al., 2013). Additionally, this approach ultimately contributes to overdiagnosis of insignificant PCa. To overcome these limitations, MRI has

been proposed as a noninvasive imaging modality prior to urological interventions (see Fig. 1.2).

MRI is becoming a standard pre-biopsy examination to assess the presence of potential PCa lesions for subsequent TRUS-guided biopsy (Barentsz et al., 2016). The technical specifications and interpretation of prostate MRI have been standardized by the Prostate Imaging Reporting and Data System (PI-RADS) (Weinreb et al., 2016; Turkbey et al., 2019). Due to the high heterogeneity of csPCa, and to distinguish it from clinically insignificant PCa and benign pathological tissues, PI-RADS recommends acquiring prostate MRI in a multiparametric (multi-modal*) fashion. mpMRI integrates multiple imaging contrasts, including:

- high-resolution T2-weighted (T2w) imaging, rich in morphological detail, for anatomical assessment,
- functional Diffusion-Weighted Imaging (DWI) and its derived Apparent Diffusion Coefficient (ADC) map, for physiological assessment (diffusion tissue properties),
- Dynamic Contrast-Enhanced (DCE) MRI, for perfusion assessment.

The technical specifications and sequence properties of these modalities are discussed in detail in Section 1.2.2.

The MRI modalities are used in PI-RADS assessment through a prostate zone-specific decision tree, which determines the overall PI-RADS score (see Fig. 1.3). Each detected lesion is graded on a 5-point Likert scale, reflecting the risk of csPCa based on its zonal location:

- For transitional prostate zone (TZ) lesions, the primary imaging modality is T2w imaging, but the current PI-RADS v2.1 (Turkbey et al., 2019) also recommends DWI assessment:
  - If a lesion is assigned a score of 2, it is upgraded to a 3 if the DWI score is higher than 4.
  - If a lesion is assigned a score of 3, it is upgraded to a 4 if the DWI score is 5.
- For peripheral zone (PZ) lesions, the DWI image is the primary modality. Additionally, if a lesion receives a score of 3, it is upgraded if the DCE image is suspicious. However, the role of DCE imaging in PI-RADS remains under debate, with increasing support for biparametric abbreviated protocols that exclude contrast-enhanced imaging (Gatti et al., 2019; Kuhl et al., 2017; Liang et al., 2020; Penzkofer, 2024; Tavakoli et al., 2023; Twilt et al., 2025; Zawaideh et al., 2020).

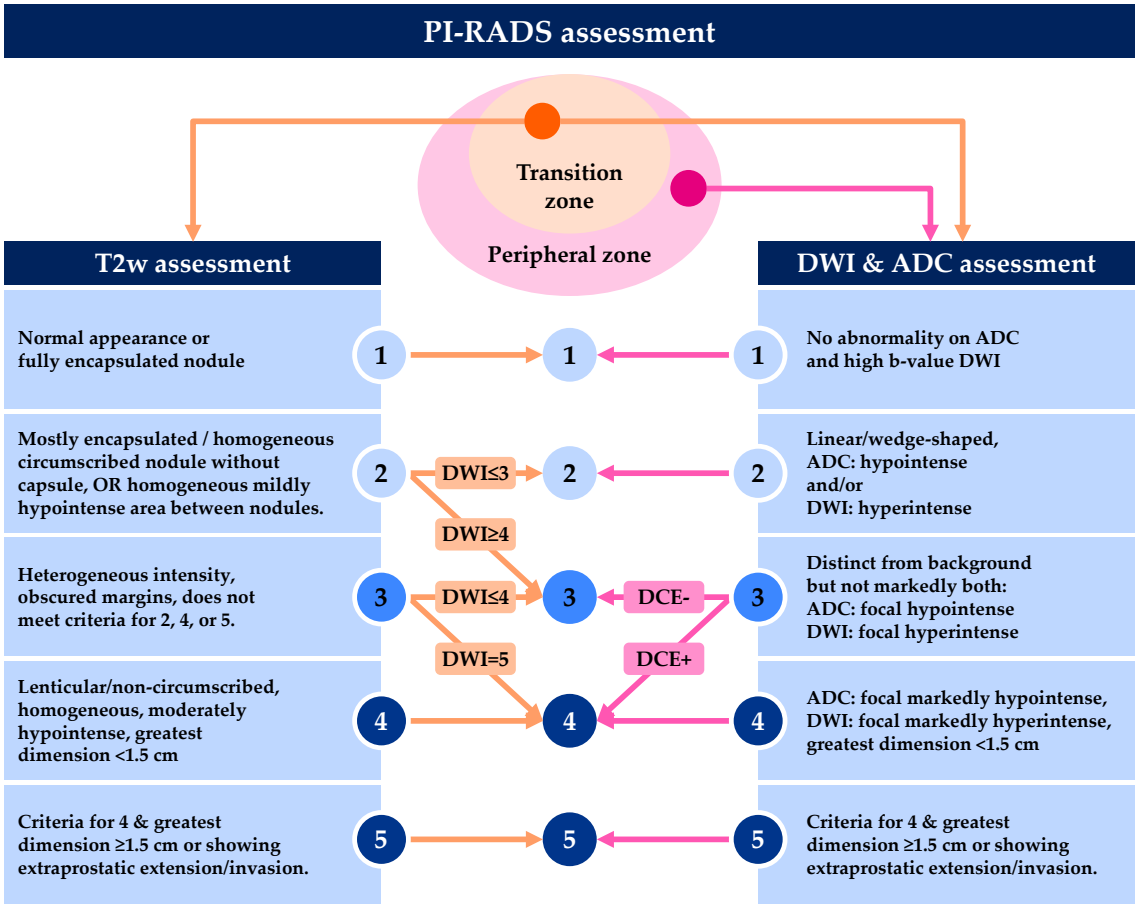

Figure 1.3: Prostate Imaging Reporting and Data System (PI-RADS) assessment utilizing MRI modalities in a prostate zone-specific decision tree. Lesions are graded on a 5-point Likert scale based on their risk of csPCa and zonal location. For transition zone (TZ) lesions, T2-weighted (T2w) imaging is the primary modality, but PI-RADS v2.1 (Turkbey et al., 2019) also incorporates diffusion-weighted imaging (DWI) assessment, allowing score upgrades based on DWI findings. For peripheral zone (PZ) lesions, DWI is the dominant modality, with potential upgrades based on dynamic contrast-enhanced imaging (DCE). However, the role of DCE remains debated, with growing support for biparametric protocols that exclude contrast-enhanced imaging (Gatti et al., 2019; Kuhl et al., 2017; Liang et al., 2020; Tavakoli et al., 2023; Twilt et al., 2025). The figure is created by following the current PI-RADS v2.1 guidelines (Turkbey et al., 2019)

After grading all suspicious lesions, the lesion with the highest PI-RADS score determines the final PI-RADS score for the patient and thereby the necessity of subsequent biopsy, which varies on a case-by-case basis and across different clinical practices. For patients at low-risk for PCa – typically with PI-RADS 1 and 2 – MRI has contributed to reducing overtreatment and biopsy-related risks – such as infection, pain, and bleeding – by ruling out PCa instead of immediate systematic TRUS-guided biopsy (Ahmed et al., 2017). In contrast, patients with high-risk lesions (typically PI-RADS 4 and 5) are referred for biopsy. Ambiguous PI-RADS 3 cases require more detailed multidisciplinary risk stratification by radiologists and urologists where the decision highly relies on the PSA-density (Schoots and Padhani, 2020). This can also change decision-making in low-risk PI-RADS 1 and 2 cases.

It is important to note that the diagnostic accuracy on MRI for intermediate- and high-risk lesions (PI-RADS 3 and 4) remained low, with a considerably high false-positive rate due to imaging features associated with benign prostatic hyperplasia, inflammation, prior trauma, and infection (Panebianco et al., 2018; Turkbey and Choyke, 2018). Therefore, MRI alone is insufficient for a definitive diagnosis, but it highly influences the success of the subsequent diagnostic step, the biopsy procedure. mpMRI has enabled precise localization of suspicious lesions throughout the entire prostate gland, paving the way for targeted lesion biopsy. The introduction of targeted biopsy has led to an overall improvement in diagnostic performance (Hugosson et al., 2022), which is discussed in the following subsection.

Although the treatment of csPCa is beyond the scope of this dissertation, it is important to note that prostate MRI also plays a crucial role in emerging PCa treatments, including:

- guidance for focal therapy (Ghai et al., 2024),
- treatment planning for radiation therapy (Kerkmeijer et al., 2021),
- surgical planning for radical prostatectomy (Marenco et al., 2019; Shirk et al., 2022).

Targeted Biopsy

Suspicious lesions identified on mpMRI by radiologists can be directly targeted by the biopsy needle, either as an alternative to or in combination with systematic biopsy, which involves sampling the entire prostate gland, as discussed in the previous subsection. The ability to perform targeted biopsy has led to increased sensitivity and a significantly reduced false-negative rate in detecting csPCa (Ahdoot et al., 2020; Ahmed et al., 2017; Hugosson et al., 2022; Puech et al., 2013; Siddiqui et al., 2015; Valerio et al., 2015). Additionally, it has resulted in a reduction in the number of biopsy cores required (Valerio et al., 2015), thereby lowering biopsy-related risks.

To guide targeted biopsy, three main techniques are used in clinical practice, listed in order of increasing complexity:

1. **Cognitive targeted biopsy**: The urologist cognitively registers and aligns the information from the preoperative mpMRI with the US image to target the suspicious lesions, which may not be directly visible on the US image (Ouzzane et al., 2011; Puech et al., 2013). Due to its reliance on the urologist's interpretation, this technique is also referred to as visual estimation. The main advantage of this technique is its simplicity, as it does not require additional equipment. However, its effectiveness depends on the urologist's experience and ability to interpret MRI images accurately.

2. **MRI-TRUS fusion targeted biopsy**: This approach utilizes computer-aided guidance, where the prostate and suspicious lesions delineated on MRI are registered in real-time to the US image with additional guidance provided for biopsy needle placement (Pinto et al., 2011; Puech et al., 2013). Although this method incorporates automated guidance, the accuracy of the intervention depends on the quality of MRI lesion delineation and, more importantly, on the precision of the registration algorithm. Due to strong localized soft tissue deformations caused by pressure from the US probe, the method is prone to registration errors, which represents its main limitation.

3. **In-bore MRI targeted biopsy**: This technique provides direct MRI guidance, allowing lesions to be targeted within the MRI scanner (Beyersdorff et al., 2005; Woodrum et al., 2016). It is considered the most accurate approach, as it eliminates cognitive errors and MRI-TRUS registration inaccuracies, ensuring precise needle positioning. However, this method is also the most resource-intensive, requiring long intervention time, MRI-compatible (non-ferromagnetic) biopsy equipment, access to an additional MRI scanner dedicated to interventions separate from diagnostic imaging, and synchronized collaboration between radiologists and urologists. Due to its complexity, in-bore targeted biopsy is not widely adopted in clinical practice.

Although these approaches differ significantly in their clinical and technical procedures, there is currently no strong evidence that they significantly differ in diagnostic performance (Monda et al., 2018; Puech et al., 2013; Wegelin et al., 2017; Wysock et al., 2014). The choice of technique should be carefully considered by each clinic based on available resources, expertise, and learning curves (Kasabwala et al., 2019; Meng et al., 2018), while also assessing both diagnostic performance and cost-efficiency, including the time required for training and implementation.

Histopathologic Assesment

The tissue samples extracted during the biopsy procedure (or radical prostatectomy) undergo histopathological assessment under a microscope, where a histopathologist grades them to evaluate PCa aggressiveness. The grading process follows the Gleason grading system (Gleason and Mellinger, 1974), which is based on morphological assessment and evaluates the degree of cellular differentiation on a 5-point scale, each representing a distinct Gleason pattern. Normal cells exhibit uniform shape and well-differentiated glandular structures, whereas malignant tissue is characterized by heterogeneous cellular morphology, poor or absent gland formation, and stromal invasion. The Gleason patterns are illustrated in Fig. 1.4.

The Gleason Score (GS) is composed of two Gleason patterns (primary + secondary) identified in the histopathological sample (Epstein, 2010; Epstein et al., 2016b; Humphrey, 2004):

- **Prostatectomy specimens**: The most prevalent Gleason pattern is reported as the primary, while the second most common pattern is reported as the secondary. This approach provides better prognostic discrimination by reflecting proportional composition.
- **Biopsy specimens:** The most prevalent Gleason pattern is also reported as the primary. However, the worst (highest-grade) pattern observed is reported as the secondary, being present in any amount. This adjustment compensates for the more limited sampling in biopsies compared to prostatectomy specimens.

This dual distinction is crucial, as seen in cases such as the score of 7 (3+4=7a vs. 4+3=7b), where a higher proportion of pattern 4 indicates a worse prognosis. Although the GS ranges from 2 to 10, in clinical practice, the lowest assigned score for malignancy is 6. Furthermore, scores of 9 and 10 do not show significant prognostic differences. To improve risk stratification, a Gleason Grade Group (GGG) with a 5-point scale has been proposed to reduce patient anxiety and overtreatment (Epstein et al., 2016b). In modern urology, GGG $\geq$2 lesions are commonly classified as csPCa, while GGG 1 is generally considered as clinically insignificant (Hugosson et al., 2022), although multiple definitions exist. An overview of the relationship between Gleason patterns, GSs, and GGGs is provided in Tab. 1.1.

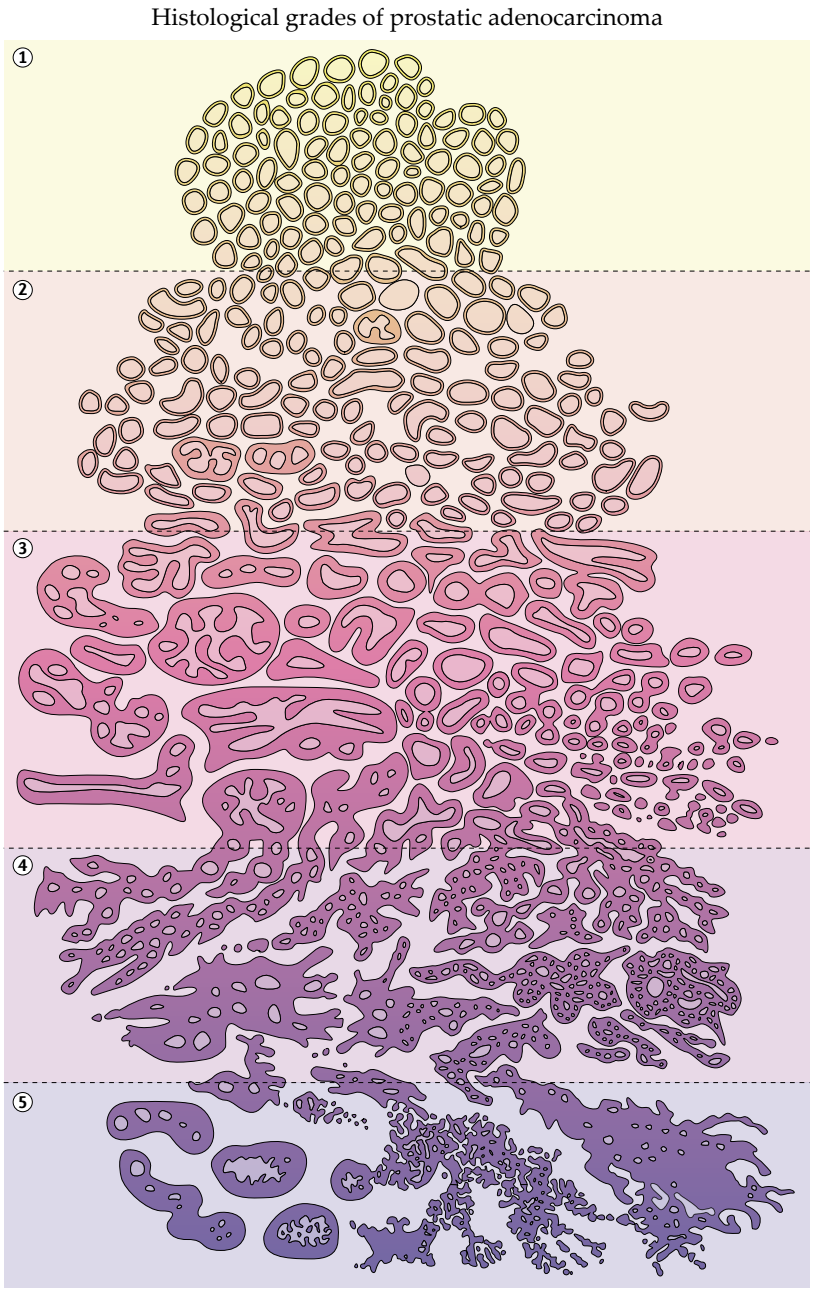

Figure 1.4: Visualization of the five Gleason patterns used in the Gleason grading system for prostate cancer assessment. The images illustrate the increasing loss of glandular differentiation from well-formed glands (pattern 1) to poor or absent gland formation (pattern 5). The image is adapted from Ali et al. (2022) with permission from the publisher Springer Nature.

Table 1.1: An overview of the relationship between Gleason patterns, Gleason scores (GS), and Gleason Grade Groups (GGG). The table illustrates how the most frequent and second-most frequent Gleason patterns determine the GS, which is further categorized into GGG for risk stratification. The classification differentiates clinically insignificant (GGG 1) from clinically significant prostate cancer (PCa) cases (GGG $\geq$ 2).

Gleason Pattern biopsy: most prevalent + worst prostatectomy: 1st + 2nd most frequent	3+3	3+4	4+3	4+4	4+5	5+4	5+5
Gleason Score	6	7a	7b	8	9a	9b	10
Gleason Grade Group	1	2	3	4	5		
Clinical significance	clinically insign.	clinically significant prostate cancer (csPCa)					

The GGG, based on Gleason patterns, currently holds the highest prognostic value in clinical practice and is considered the gold standard for PCa grading (He et al., 2017), with regular updates from the International Society of Urological Pathology Consensus (ISUP) (Epstein et al., 2005, 2016a; Van Leenders et al., 2020).

Given that GGG is primarily linked to morphological features correlating with tissue density, it also directly influences tissue diffusion properties. As a result, apparent diffusion coefficient (ADC) map values have strong prognostic significance, demonstrating a high correlation with underlying histological features such as cell density and glandular structure (Gibbs et al., 2009). This also explains why DWI and ADC assessment play a crucial role in the diagnostic performance of mpMRI, as they are recommended for lesion evaluation in both prostate zones (see Fig. 1.3). On the other hand, since the GGG is based exclusively on morphological assessment, there is growing interest in further major refinements. One limitation of the grading system is that it is not zone-specific, despite evidence that zonal tumor involvement significantly influences clinical outcomes (Ali et al., 2022). Due to the physical and functional differences between prostate zones, PZ tumors tend to have a worse prognosis than TZ tumors, particularly those near the ejaculatory ducts, which are highly aggressive, making them more prone to extracapsular extension and seminal vesicle invasion (Vargas et al., 2012). Additionally, functional factors such as stromal reactivity are not incorporated into the current system. In cases where tissue exhibits a strong stromogenic response, the risk should be considered higher than what the GGG alone suggests, as incorporating stromal activity has been shown to improve risk stratification (Frankenstein et al., 2020).

1.2.2 Multi-parametric Prostate MRI

Prostate MRI is recommended to be acquired in a multi-parametric fashion, following the PI-RADS guidelines. It includes three major imaging contrasts: T2w, DWI, and DCE images. The typical sequence properties required to achieve the desired imaging contrasts are described in the following subsections.

T2-weighted Imaging (T2w)

T2w imaging is one of the most widely used MRI modalities*, employing various pulse sequences primarily for morphological assessment due to its strong contrast between key tissue types such as muscle, fat, and water. As recommended by PI-RADS, it is also a core component of prostate mpMRI evaluation.

T_2 is the characteristic time describing the exponential magnitude decay of the transverse magnetization (M_T) due to magnetic moment interactions of neighboring spins. Hence, it is also referred to as the spin-spin relaxation time. The actual time T_2^* characterizing the exponential decay of the M_T signal after a 90° radiofrequency (RF) pulse – called free induction decay (FID) – is however affected by the spin dephasing effect of magnetic field inhomogeneities characterized by a time-invariant decay constant of T_2':

$$1/T_2^* = 1/T_2 + 1/T_2'. \tag{1.1}$$

As a result, the characteristic time of T_2^* is not only affected by the time-variant spin-spin interactions, but also the magnetic field imperfections of the MRI scanner and local differences of magnetic tissue properties, in other words, magnetic susceptibility.

To mitigate the dependence of T2w contrast on magnetic field inhomogeneities, the spin echo (SE) pulse sequence (see Fig. 1.5) is employed also serving as the foundation for prostate T2w imaging. After a 90° RF pulse, a spin ensemble is rotated into the M_T plane, where it begins to decay (FID) in amplitude while experiencing dephasing due to the magnetic field inhomogeneities. Introducing a 180° RF pulse after the 90° RF pulse rotates the spins around either the x- or y-axis in the M_T plane, causing them to reverse their previous rotation direction. The phase shift accumulated by faster spins relative to slower spins before the 180° RF pulse is also reversed. Consequently, over time, the faster spins will catch up with the slower spins and the spins will be rephased again after a time interval equal to the time difference between the RF pulses. This rephasing results in a M_T signal, known as the "echo". The time interval between the 90° RF pulse and the echo is called the echo time (T_E) and T_2 characterizes the exponential decay curve fitted to the maxima of the FID and the echo signal:

$$M_T(t) = M_{T0} \cdot e^{-t/T_2}. \tag{1.2}$$

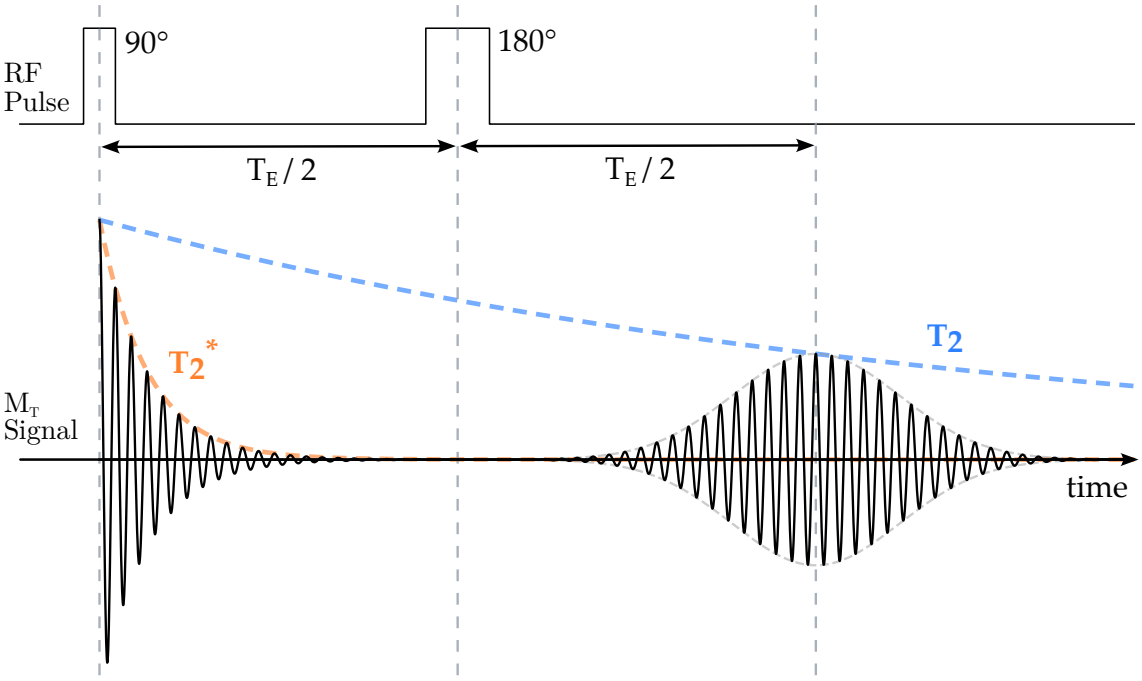

Figure 1.5: The spin echo (SE) pulse sequence used for T2-weighted imaging. After an initial 90° radiofrequency (RF) excitation, the transverse magnetization (M_T) undergoes free induction decay (FID) with a characteristic decay constant of T_2^*, influenced by both spin-spin interactions and magnetic field inhomogeneities. A 180° RF pulse applied at $T_E/2$ rephases the spins, generating a SE at time T_E. The echo follows the true T_2 decay, mitigating the effect of field inhomogeneities (T_2'), as described by the relation $1/T_2^* = 1/T_2 + 1/T_2'$. Figure is created based on Brown et al. (2014).

A single 90° RF excitation allows sampling of one k-space line per slice. Since T2w sequences require a long T_E, they result in prolonged image acquisition times, which is particularly critical in body regions prone to motion and frequent soft tissue deformations, such as the prostate. Prostate gland motion or deformations caused by bowel peristalsis, rectal distension, bladder distension, respiration, muscle contraction, or patient movement can easily lead to blurred images or ghosting artifacts (Engels et al., 2020). A straightforward approach to reduce the acquisition time and thereby minimize motion effects would be to shorten the time interval between consecutive 90° RF pulses, known as the repetition time (T_R). However, an excessively short T_R leads to a M_T signal loss – this phenomenon is utilized in T1-weighted MRI (see Section 1.2.2 in *Dynamic Contrast Enhanced Imaging*) – thereby reducing the signal-to-noise ratio (SNR).

One solution to reduce the overall acquisition time without loss of SNR is to collect multiple echoes during the T_R instead of just one. This technique, referred to as fast-spin-echo (FSE) – or alternatively turbo-spin-echo (TSE) – where a train of 180° RF pulses is applied after the initial 90° RF excitation (see Fig. 1.6), is recommended by PI-RADS for T2w imaging. After the first echo, the 180° RF pulse can be repeated multiple times with a repetition period of T_E, with each pulse rephasing the spin ensemble and generating an echo after $T_E/2$. This pulse sequence enables the sampling of multiple k-space lines for a single 90° RF excitation. However, the signal decay characterized by T_2 is independent of the multiple SEs. Due to magnetic field inhomogeneities and spin motion, spin refocusing cannot be perfect. Therefore, the M_T can only be recovered up to the exponentially decreasing envelope characterized by T_2. Over time, the echo signal diminishes to the noise level, providing no meaningful signal. This ultimately limits the number of echoes that can be collected during a single FSE pulse sequence.

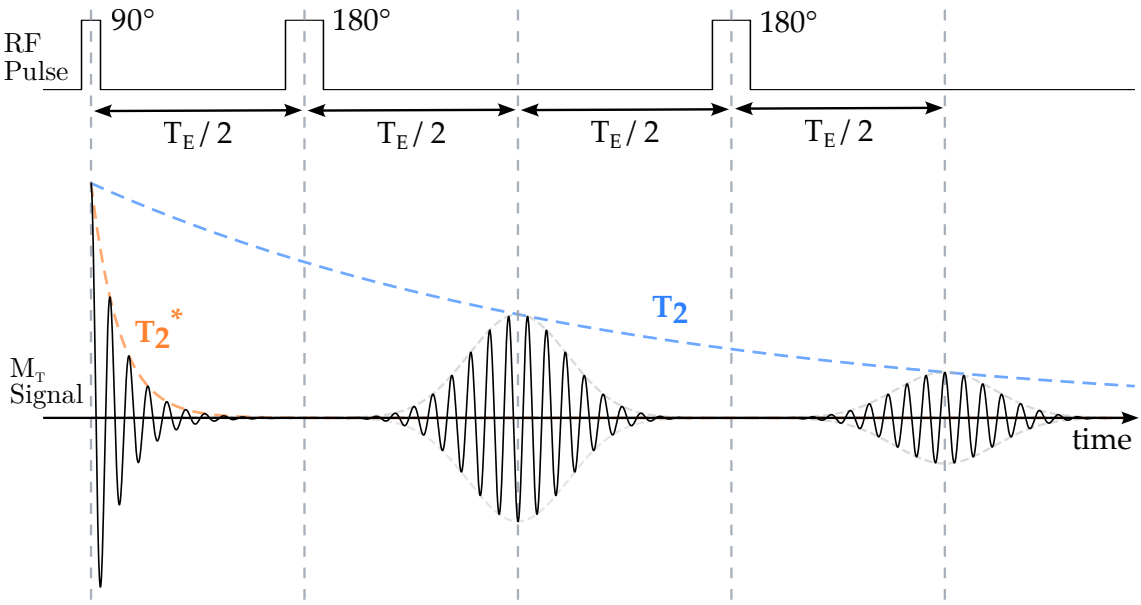

Figure 1.6: Fast Spin Echo (FSE) – also known as Turbo Spin Echo (TSE) – pulse sequence recommended by PI-RADS for prostate T2w imaging. After an initial 90° RF excitation, a train of 180° RF pulses is applied at intervals of T_E to repeatedly rephase the spin ensemble, generating multiple spin echoes. This allows for the acquisition of multiple k-space lines within a single T_R, reducing overall acquisition time without loss of SNR. Due to imperfect spin refocusing from magnetic field inhomogeneities and spin motion, the signal gradually diminishes to the noise level over time, limiting the number of usable echoes and thereby restricting the effective echo train length. Figure is created based on Brown et al. (2014).

Diffusion-Weighted Imaging (DWI)

Random movement of particles in a medium over time is referred to as Brownian motion. It is quantified by the diffusion coefficient D, and reflects the area traveled by a particle per unit of time and is measured in [$mm^2\,s^{-1}$]. Molecular diffusion within tissues is a property that can create significant contrast between healthy and densely packed malignant tissue. DWI, an imaging modality designed to visualize the restriction of this molecular-level process (Dietrich et al., 2010; Hagmann et al., 2006), is widely used in cancer diagnostics, including for brain (Kono et al., 2001), breast (Woodhams et al., 2011), and prostate cancer (Shimofusa et al., 2005).

Prostate DWI is based on the incorporation of a dephasing and rephasing mechanism of spins into the SE pulse sequence (described in Section 1.2.2 in T2-weighted imaging), see Fig. 1.7. To generate contrast between moving and static spins, additional spatially-dependent gradient fields $\Delta B\,(x)$ are applied on either side of the 180° RF pulse. The 1st gradient pulse dephases both moving and static spins by an angle of $\Delta\phi\,(x)$, where the extent of phase shift is proportional to the additional magnetic field strength along the gradient direction. The 2nd gradient pulse is applied after the 180° RF pulse with the same polarity for the same duration but with a phase flip. As a result, stationary spins are rephased by experiencing an equal phase shift in the opposite direction, ultimately restoring the same M_T as if the two gradient pulses had not been applied. In contrast, moving spins experience a phase shift with a different amplitude during the second gradient pulse due to their repositioning. Consequently, their phase shift is not completely rephased, and volumes with high molecular diffusion contain spins with unsynchronized phase angles. This leads to a drop in M_T, resulting in signal attenuation.

Properties of the two gradient pulses – namely their amplitude (g_D, typically 20–40 $mT\,m^{-1}$), duration of (δ, typically 20–40 ms), and time interval between their onset (Δ) – determine the most characteristic parameter of the DWI sequence, the so-called diffusion weighting of b, or with other words the b-value of the DWI pulse sequence:

$$b = \gamma^2 \cdot {g_D}^2 \cdot \delta^2 \cdot (\Delta - \delta/3)\,, \tag{1.3}$$

which is expressed in [$s\,mm^{-2}$] and is typically set to values in the range of 50–1500 $s\,mm^{-2}$. The greater the diffusion weighting, the higher the contrast between tissues with different molecular diffusion, but this also results in a lower SNR. The long gradient pulse duration of δ required for achieving proper diffusion-weighted contrast between tissues with different molecular diffusion results in a long T_E, inherently making DWI images T2-weighted. As a result, tissues characterized by long T2 relaxation times can be mistaken for diffusion restriction, as they also exhibit high signal intensity on DWI images. This phenomenon is called "T2-shinethrough", making the differentiation betweenbenign prostatic hyper-

plasia (BPH) with long T2 relaxation time from cancer tissue with restricted diffusion challenging.

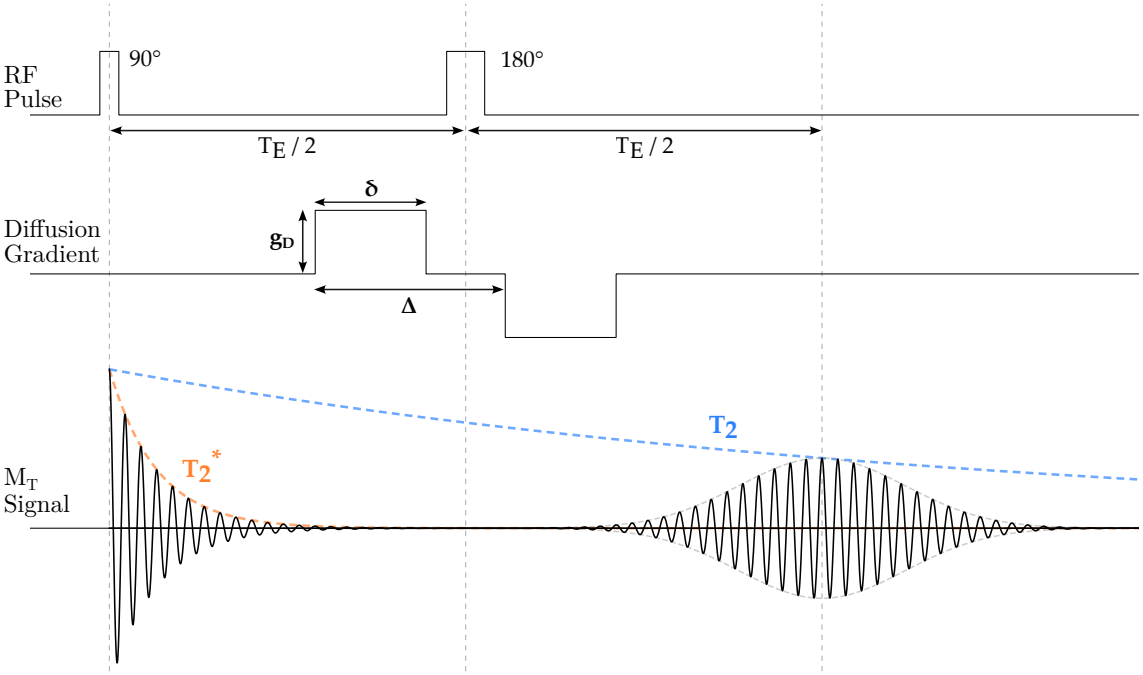

Figure 1.7: Diffusion-Weighted Imaging (DWI) integrated into the Spin-Echo (SE) pulse sequence. To generate diffusion contrast, additional dephasing and rephasing gradient pulses are applied on either side of the 180° radiofrequency (RF) pulse. The first gradient pulse dephases both static and moving spins, while the phase-reversed second pulse affects them differently. Stationary spins fully refocus, restoring their original transverse magnetization (M_T), whereas moving spins undergo incomplete rephasing, leading to signal attenuation proportional to molecular diffusion. The diffusion weighting is characterized by the b-value, determined by the gradient amplitude (g_D), pulse duration (δ), and onset difference (Δ). Higher b-values enhance diffusion contrast but reduce SNR. Additionally, the long echo time (T_E) required for diffusion encoding inherently results in T2w signal contamination, leading to the "T2-shinethrough" effect, which can make differentiation between benign and malignant tissues challenging. The figure illustrates the RF and gradient pulse sequence, as well as the phase evolution of spins in stationary and diffusing conditions.

To differentiate the diffusion-weighted from the T2w component on the image (overcoming the effect of T2-shinethrough), signal attenuation in the function of multiple b-values is measured. Signal attenuation can be modeled by an exponential relationship with the attenuation coefficient b along with the diffusion coefficient D:

$$I_{\mathrm{b}} = I_{\mathrm{b}=0} \cdot e^{-b \cdot D}, \tag{1.4}$$

where I_{b} is the measured signal for a specific b-value, and $I_{\mathrm{b}=0}$ is the signal measured without gradient pulses. By fitting an exponential curve to the measured signal intensities I_{b} across multiple b-values, the diffusion coefficient D can be calculated. Increasing the number of measurements across different b-values improves the accuracy of the calculation but comes at the cost of longer acquisition times. The diffusion coefficient D is defined for an unrestricted environment without restricting structures. However, in biological tissues, the presence of cells, blood vessels, and other structures violates this condition to varying degrees. Consequently, the measured signal within a given volume reflects not only molecular diffusion but also the influence of the microstructural tissue characteristics and is therefore referred to as the apparent diffusion coefficient ADC. Calculating the ADC for each voxel provides a parametric map that is highly independent of the magnetic field strength, unlike T1- and T2-weighted MRI sequences. The ADC map is ultimately derived from a diffusion model, meaning the measured values are semi-quantitative and influenced by acquisition parameters, such as number of b-values and their corresponding gradient strengths. While ADC values are not as absolute or directly comparable as Hounsfield units in computer tomography, they strongly correlate with the physical process of diffusion. This correlation allows for quantitative comparisons between ADC maps with similar acquisition protocols.

ADC maps are particularly sensitive to involuntary patient movements, as motion artifacts can accumulate across DWI images acquired with different b-values. Therefore, fast image acquisition is critical to reduce these artifacts and minimize errors in the ADC map calculation. To mitigate both intra- and inter-image motion artifacts, SE echo-planar imaging (EPI) is recommended by PI-RADS as a fast imaging pulse sequence for DWI. Similarly to the FSE pulse sequence used in T2w imaging with the same motivation (see Section 1.2.2 T2-weighted imaging), multiple lines in k-space are acquired after a single 90° RF pulse during the EPI pulse sequence. Unlike FSE, where multiple k-space lines are acquired using a train of SEs, EPI collects multiple k-space lines from a single SE by rapidly switching the readout gradients. This necessitates extremely fast gradient switching, placing the highest technical demands on MRI technology. At the same time, EPI has the advantage of being able to acquire a full image slice in under 90 ms. Fig. 1.8 illustrates an example for an EPI pulse sequence and the corresponding k-space readout.

Unlike FSE, where multiple k-space lines are acquired using a train of spin echoes, EPI collects multiple k-space lines following a single spin echo by rapidly switching the readout gradients. This requires extremely fast gradient switching, placing high technical demands on MRI hardware.

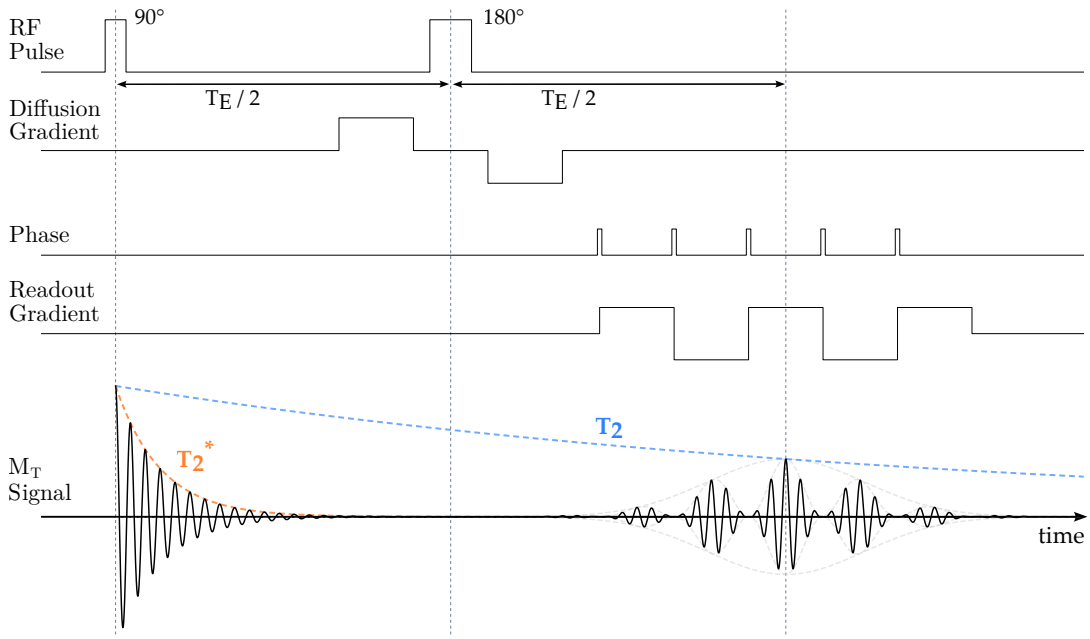

Figure 1.8: Echo-Planar Imaging (EPI) pulse sequence recommended by PI-RADS for prostate Diffusion-Weighted Imaging (DWI). EPI enables rapid image acquisition by collecting multiple k-space lines after a single 90° RF excitation. Unlike FSE, where multiple k-space lines are acquired using a train of SEs, EPI collects multiple k-space lines from a single SE by rapidly switching the readout gradients. This approach minimizes motion artifacts in DWI and allows full-slice acquisition in under 90 ms, but imposes high technical demands on MRI hardware. Figure is created based on Brown et al. (2014).

Dynamic Contrast-Enhanced Imaging (DCE)

DCE is a technique to visualize physiological properties related to tissue perfusion through the administration of a contrast agent (CA), usually a gadolinium-based CA (Brown et al., 2014). Characteristically for DCE sequences, a tain of images are acquired prior, during, and post to the intravenous CA injection, thereby providing the dynamic information. This type of MRI prioritizes rapid image acquisition where reducing the T_R is necessary to be able to capture the fast physiological process of perfusion.

After a 90° RF pulse, the longitudinal magnetization (M_Z) restores to its thermody-

namic equilibrium M_{Z0} parallel to the static magnetic field B_0 due to the interactions of the spins with their surroundings (spin-lattice relaxation) following

$$M_Z(t) = M_{Z0} \cdot \left(1 - e^{-t/T_1}\right) \tag{1.5}$$

characterized by T_1 relaxation time. However, if the subsequent 90° RF pulse approaches before the restoration of the M_Z, only the M_Z komponent of the current magnetization will be flipped into the transverse x-y plane, resulted in decrease in the measured M_T signal – has been already mentioned as a main motivation for FSE and EPI sequences in Section 1.2.2 in *T2-weighted imaging* and in *Diffusion-weighted Imaging*, respectively – which will be in correlation to the T_1 relaxation time. Fig. 1.9 shows the saturation recovery pulse sequence that allows the indirect measurement of the T_1 relaxation time.

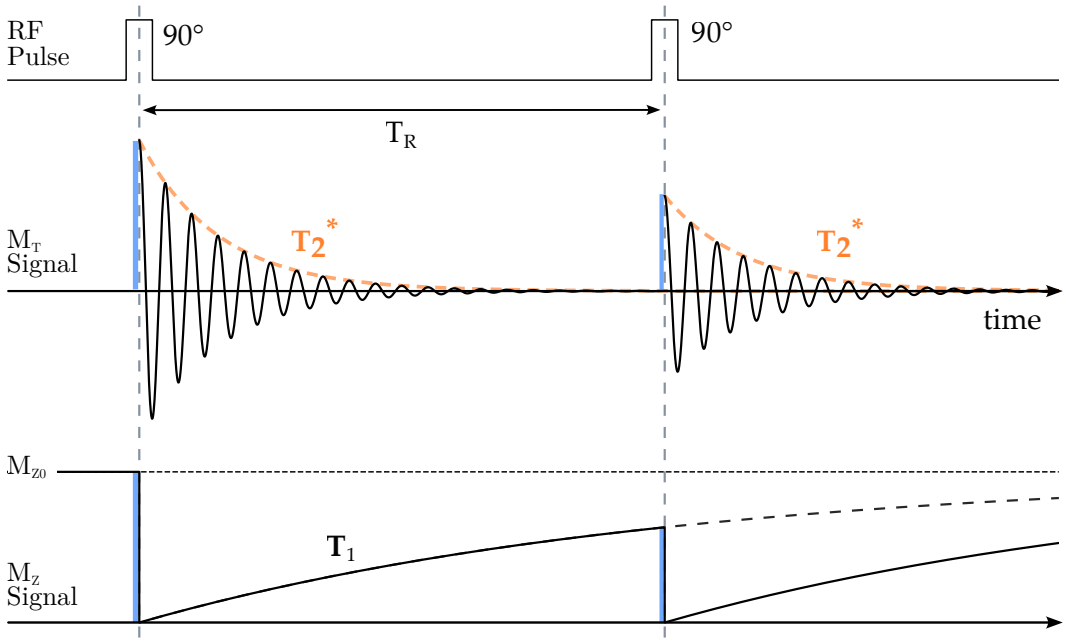

Figure 1.9: Saturation recovery pulse sequence used for indirect measurement of T_1 relaxation time. Following a 90° RF pulse, the longitudinal magnetization (M_Z) recovers toward its equilibrium value (M_{Z0}) following the function $M_Z(t) = M_{Z0} \cdot (1 - e^{-t/T_1})$. If a subsequent 90° RF pulse is applied before full M_Z recovery, only the remaining longitudinal magnetization is flipped into the transverse plane, leading to a reduced transverse magnetization signal (M_T). This M_T attenuation depends on T_1 relaxation, forming the basis for saturation recovery experiments. The figure illustrates the RF excitation pulses, and the corresponding evolution of M_Z and M_T. Figure is created based on Brown et al. (2014).

A pulse sequence with $T_R < T_1$ such that DCE also uses lead to an image contrast characteristics where the T_1 tissue properties dominate, ultimately leading to a T1-weighted (T1w) image. Fig. 1.10 shows an example for a pulse sequence with short T_R for two tissues with differences in their T_1 relaxation time.

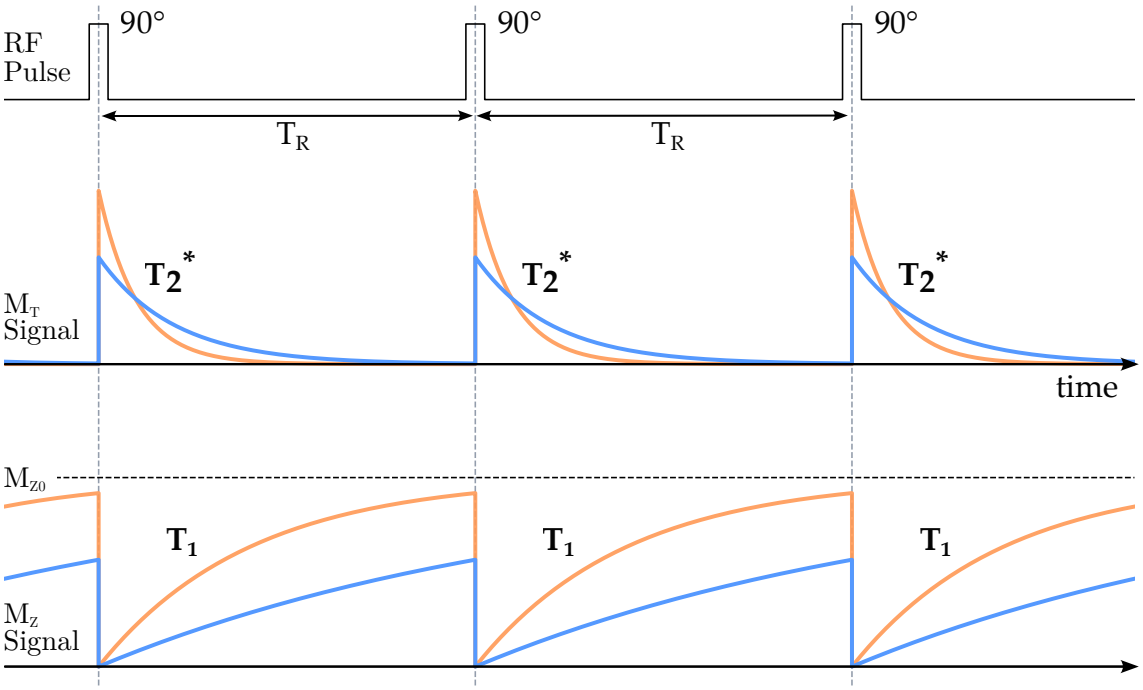

Figure 1.10: Pulse sequence with short repetition time (T_R) as used in DCE imaging. Repeated 90° RF pulses are applied before full longitudinal magnetization (M_Z) recovery, leading to signal differences between tissues with varying T_1 relaxation times. The curves illustrate the recovery behavior of the same tissue with (orange) and without contrast agent (blue), respectively, ultimately generating T1w image contrast.

The possibility to minimize the T_R as much as possible is allowed by an essential group of MRI sequences, namely the fast gradient echo (GRE) sequences (Markl and Leupold, 2012), where $T_R \ll T_1$ and $< T_2$. The mechanism of echo generation of fast GRE sequences is different from SE sequences. Instead of a 180° refocusing RF pulse, GRE sequences use the process of gradient reversal using the signal from the FID, characterized by T_2^*. First, an inverse gradient field is applied to dephase the M_T directly following the 90° excitation RF pulse. Rapid reversal of the gradient rephases the spins again, generating an echo with the magnitude up to the exponentially decreasing envelope characterized by T_2^*. On the other hand, the gradient corrects only the generated phase shift but the magnetic field inhomogeneity and susceptibility effects not. Therefore the effect of T_2' will be significant on the M_T measurements since the signal originates from the FID and images originating from GRE sequences are generally more sensitive for artifacts compared to SE sequences (see Section 1.2.2 in *T2-weighted imaging*).

Clinically used GRE sequences demand not only high imaging speed but also high T_1 contrast. However, in the case of fast GRE sequences, not only the M_Z, but neither the M_T can fully decay before the next 90° RF pulse due to the short T_R and therefore T_2 also affects the measured signal. To dephase the remaining M_T and thereby disrupt T_2 coherences in the measured signal, the so-called RF-spoiling technique is applied. During the train of RF pulses, their phase (tipping direction in the x-y plane) is quadratically implemented, which suppresses the formation of SE, thereby ensuring the T1-weighting.

Contrast enhancement with respect to perfusion tissue properties using fast GRE sequences is usually made by using CA. A paramagnetic material is used in a solution as a CA that is injected into the blood circulation. The CA in the blood helps the energy transfer between the spins, resulting in faster T_1 relaxation and thereby enhanced signal related to bloodstream. The most widely used paramagnetic material for this purpose is Gadolinium, such as for prostate DCE. Thanks to the rapid image acquisition provided by the fast GRE pulse sequence, differences in the CA uptake over time can highlight the contrast between malignant and benign tissue conditions.

Technical Parameters of MRI Sequences Recommended by PI-RADS

The PI-RADS guidelines include detailed technical specifications for MRI acquisition protocols to promote standardization across imaging centers and ensure adequate image quality for consistent interpretation. According to the latest version, all sequences are preferred to be acquired using a 3 T magnetic field and a slice thickness of 3 mm to maintain sufficient SNR. Additional sequence-specific parameters – T_E, T_R, field of view (FoV), and in-plane resolution – are summarized in Tab. 1.2.

Table 1.2: Recommended technical parameters for T2-weighted (T2w), diffusion-weighted imaging (DWI), and dynamic contrast-enhanced (DCE) MRI sequences according to PI-RADS v2.1 guidelines. These specifications ensure sufficient image quality and spatial resolution for reliable prostate lesion assessment. The table is created based on Weinreb et al. (2016); Turkbey et al. (2019).

Parameter	T2w	DWI	DCE
Sequence	Fast Spin Echo	Echo-Planar Imaging	Fast Gradient Echo
TE	–	$\leq$90 ms	$\leq$5 ms
TR	–	$\geq$3000 ms	$\leq$100 ms
FoV	12–20 cm (entire prostate gland & seminal vesicles)	16–22 cm	entire prostate gland & seminal vesicles
In-plane resolution (phase$\times$freq.)	$\leq$ 0.7mm$\times$ $\leq$ 0.4mm	$\leq$ 2.5mm for both	$\leq$ 2mm for both
b-values for ADC map	–	50–100 s mm^{-2} 800–1000 s mm^{-2}	–
high b-value	–	$\geq$ 1400s mm^{-2}	–
Temp. resolution	–	–	$\leq$ 15s
Total time	–	–	$\geq$ 2min
Dose	–	–	0.1 mmol kg^{-1} GBCA or equivalent
Injection rate	–	–	2–3 cm^3 s^{-1}

1.2.3 Multi-Modal Medical Image Analysis with Convolutional Neural Networks

The rapid advancements in AI over the past decade – particularly in deep neural network (DNN) architectures, training strategies, hardware accelerators (e.g., GPUs and TPUs), and software frameworks such as PyTorch (Paszke et al., 2019) and TensorFlow (Abadi et al., 2016) – have significantly increased the computational power and practical applicability of AI systems. These technical developments, combined with improved access to large, structured datasets, have driven major progress in image processing, including radiological medical image analysis.

Simultaneously, the diagnosis of many diseases has become both more accurate and complex due to advancements in medical imaging techniques (Azam et al., 2022). Similarly to prostate cancer, multi-modal imaging is also essential for the comprehensive assessment of many disorders. For example, glioblastoma characterization relies on combining contrast-enhanced T1w imaging to delineate the contrast-enhancing tumor core and T2-weighted as well as fluid-attenuated inversion recovery (FLAIR) imaging to identify the surrounding edema and non-enhancing infiltrative tumor regions (Shukla et al., 2017). Similarly, the evaluation of adnexal lesions (Sadowski et al., 2022) and breast cancer (Wekking et al., 2023) often involves mpMRI protocols that integrate T2w, DWI, ADC maps and DCE imaging. Hepatocellular carcinoma assessment also benefits from fusing various imaging modalities involving multiphase contrast-enhanced computed tomography (CT) and/or multiple MRI images, including T2w, unenhanced and multiphase contrast-enhanced T1w, and optionally DWI (Elmohr et al., 2021; Mitchell et al., 2015).

Multi-modal image analysis requires the fusion of complementary information, a task that is not only cognitively demanding for radiologists but also computationally challenging for DNN-based systems. Compared to mono-modal approaches, multi-modal analysis introduces additional complexities, including image alignment, modality-specific preprocessing, and increased computational demands. In this subsection, I outline the key steps of semantic segmentation in medical image analysis, with a particular focus on data-centric strategies and modality-specific considerations for training robust medical AI models.

Multi-modal Image Preprocessing

Image preprocessing is a step that prepares images before feeding them into artificial neural networks (ANNs). The primary goal is to resolve issues that would otherwise lead to inefficient or suboptimal learning during model training. In multi-modal applications, this is especially important, as different modalities often vary significantly in contrast and alignment. Two of the most important preprocessing techniques are image co-registration and image normalization, both of which are discussed in the following.

Image co-registration is a fundamental preprocessing step for medical image analysis tasks involving multiple images and is a vast field within medical computing (Cao et al., 2020; Hill et al., 2001). It is particularly critical for tasks where spatial alignment impacts diagnostic or computational accuracy. While this dissertation focuses on multi-modal prostate MRI images, it does not aim to develop novel registration methods. Instead, only the key concepts are briefly reviewed to convey the complexity of the problem and to provide essential context for its role in this dissertation.

Registration aims to calculate a displacement field $\mathbf{u}(\mathbf{x}) = [u(\mathbf{x}), v(\mathbf{x}), w(\mathbf{x})]$, where $\mathbf{x} = (x, y, z)$, that aligns a source image – referred to as the moving image $I_M(\mathbf{x})$ – with a reference image – referred to as the fixed image $I_F(\mathbf{x})$ – such that:

$$I_M(\mathbf{x} + \mathbf{u}(\mathbf{x})) \approx I_F(\mathbf{x}) \tag{1.6}$$

This mathematical problem is formulated as an optimization problem, where a dissimilarity metric $\mathcal{D}$ between the transformed moving image $I_M(\mathbf{x} + \mathbf{u}(\mathbf{x}))$ and the fixed image I_F is minimized. Additionally, to ensure anatomically plausible transformations, the regularization term $\lambda \cdot \mathcal{R}(\mathbf{u}(\mathbf{x}))$ penalizes implausible deformation fields controlled by the regularization parameter λ. As a result, the objective function for the optimization can be written as:

$$\hat{\mathbf{u}} = \arg\min_{\mathbf{u}} \left[\underbrace{\mathcal{D}\big(I_F(\mathbf{x}),\, I_M(\mathbf{x} + \mathbf{u}(\mathbf{x}))\big)}_{\text{Similarity term}} + \underbrace{\lambda \cdot \mathcal{R}\big(\mathbf{u}(\mathbf{x})\big)}_{\text{Regularization term}} \right]. \tag{1.7}$$

Depending on the imaging context and modalities involved, registration tasks can be categorized into four primary applications, as summarized in Tab. 1.3.

Table 1.3: Categories of medical image registration based on subject scope and modality type.

	Intra-Subject	**Inter-Subject**
Intra-modality	Monitoring disease progression and treatment response	Atlas creation & atlas-based segmentation
Inter-modality	**Image fusion**	Pathology research

Intra-subject, intra-modality registration is widely used in longitudinal studies, such as active surveillance or treatment monitoring, to track disease progression or therapeutic response (Chan et al., 2001; Galbán et al., 2009; Hering et al., 2021; Moraal et al., 2009). Inter-subject, intra-modality registration is commonly employed for atlas creation and atlas-based segmentation (Iglesias and Sabuncu, 2015), primarily to facilitate annotation transfer across patients. Inter-subject, inter-modality registration is less common and mainly applied in research settings, particularly for studying disease-specific patterns across populations (Bilello et al., 2016; Youssofzadeh et al., 2017). Intra-subject, inter-modality registration, referred to as image fusion, is one of the most challenging yet diagnostically and interventionally valuable tasks. It involves aligning images of the same patient acquired with different modalities, enabling the fusion of complementary anatomical and functional information. However, multi-modal image registration remains highly challenging due to the nonlinear and inconsistent intensity relationships across modalities. These challenges arise from:

- differing intensity values for the same tissue type,
- variable imaging contrast and tissue boundaries,
- and heterogeneous noise and imaging artifact characteristics.

From this point onward, I focus on image fusion, as multi-parametric prostate MRI is a typical example of image fusion, combining anatomical and functional imaging for comprehensive tissue characterization.

After introducing the fundamental categories of registration, the following points summarize the methodological background of each component involved in a typical image-based registration workflow, with a particular focus on aspects relevant to multi-modal prostate MRI. A schematic overview is provided in Fig. 1.11.

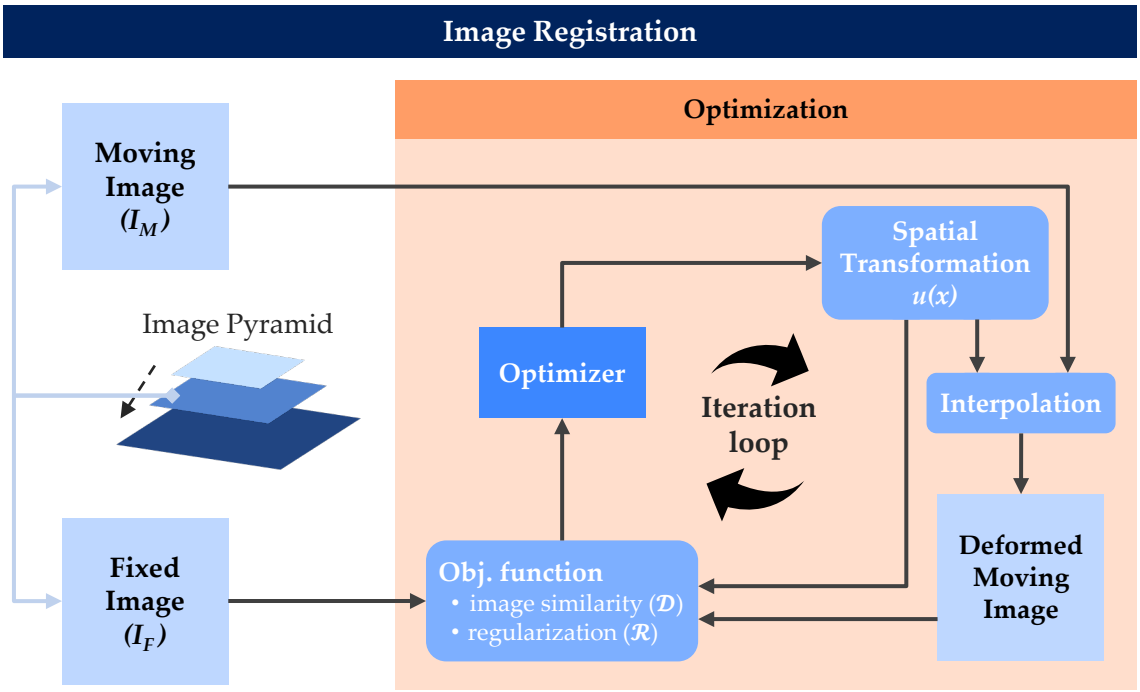

Figure 1.11: Schematic overview of the image registration pipeline. A moving image is iteratively aligned to a fixed image using an optimization loop guided by an objective function composed of an image similarity metric and a regularization term. The optimizer updates the transformation parameters, which are applied to the moving image through spatial transformation and interpolation. Registration is typically performed hierarchically using an image pyramid, progressing from coarse to fine resolutions. At each level, the transformation model also increases in complexity: starting with linear (rigid/affine) transformations for global alignment and advancing to deformable models for refined local alignment. This hierarchical strategy improves convergence and increases the likelihood of successful registration.

- **Image Transformation and Interpolation:** Image registration transformations are generally categorized into two main types:

  - **Linear:** Includes rigid (rotation and translation) and/or affine (scaling and shearing) transformations assuming global alignment and preserving the geometric structure of the image.
  - **Deformable:** Employs deformation vector fields to establish voxel-to-voxel correspondences. This is especially common in medical imaging due to soft tissue deformations. However, deformable registration is a high-dimensional, ill-posed problem where unrealistic warping can occur. Therefore, a regularization term is typically added to the objective function to constrain the solution space (see next item: *Objective Function*).

  After transformation, the grid of the deformed moving image I_M must be interpolated to match the grid of the fixed image I_F. The choice of interpolation method balances accuracy and computational cost: linear interpolation is computationally efficient, while B-spline interpolation offers higher accuracy at the expense of increased computation time.

- **Objective Function:** The objective function in image registration typically consists of two components: a similarity metric and a regularization term, as outlined in Eq. 1.7.

  - **Multi-modal similarity metric:** For multi-modal images, defining similarity is especially challenging due to the nonlinear and inconsistent intensity relationships between modalities. Among existing metrics, mutual information (MI) (Collignon et al., 1995; Viola and Wells III, 1997) is currently the most widely used and is considered the state-of-the-art for multi-modal image registration. MI is based on the information content of two images and aims to maximize their overlaps, thereby reducing the overall information content. It is computed using Shannon–Wiener entropy (Shannon, 1948), defined as:

    $$H(I) = -\sum_i p_I(i) \log p_I(i), \tag{1.8}$$

    where $p_I(i)$ denotes the probability of intensity value i in image I. Entropy is high when an image has a uniform intensity distribution (i.e., is highly textured) and low when dominated by a few intensity values. To capture joint

information between two images, joint entropy is defined as:

$$H(I_M, I_F) = -\sum_i \sum_j p_{I_M I_F}(i,j) \log \left(p_{I_M I_F}(i,j) \right), \tag{1.9}$$

where $p_{I_M I_F}(i,j)$ is the joint probability of observing intensity i in image I_M and intensity j in image I_F. The goal during registration is to maximize the mutual dependence between images, which corresponds to minimizing joint entropy. However, this metric can produce misleading results when there is little to no foreground overlap between images. To address this, mutual information is formulated as:

$$\begin{aligned} MI(I_M, I_F) &= H(I_M) + H(I_F) - H(I_M, I_F) \\ &= \sum_i \sum_j p_{I_M I_F}(i,j) \log \frac{p_{I_M I_F}(i,j)}{p_{I_M}(i)\, p_{I_F}(j)}, \end{aligned} \tag{1.10}$$

which incorporates the individual entropy of each image within the overlapping region. To further reduce sensitivity to overlap size and image noise, normalized mutual information was introduced (Studholme et al., 1999):

$$NMI(I_M, I_F) = \frac{H(I_M) + H(I_F)}{H(I_M, I_F)}, \tag{1.11}$$

which normalizes the similarity score, improving robustness across varying overlap volumes and noise levels.

– **Deformation field regularization:** To prevent unrealistic image warping – such as excessive local stretching, compression, or folding ("crossing" deformation vectors) – a regularization term is added to the objective function. This term penalizes non-plausible transformations and stabilizes the optimization process (Ashburner and Friston, 1999; Rueckert et al., 1999).

- **Optimizer:** Registration is formulated as an optimization problem that iteratively searches for the optimal displacement field $\hat{\mathbf{u}}(\mathbf{x})$ (Klein et al., 2007). Starting from an initial estimate $\mathbf{u_0}(\mathbf{x})$ – typically the identity transform – the displacement field is updated iteratively as follows:

$$\mathbf{u_{k+1}}(\mathbf{x}) = \mathbf{u_k}(\mathbf{x}) + a_k \cdot \mathbf{d_k}(\mathbf{x}), \quad k \in \mathbb{N}, \tag{1.12}$$

where a_k denotes the step size (or learning rate), and $\mathbf{d_k}(\mathbf{x})$ represents the direction of the parameter update derived from the objective function (see Eq. 1.7) evaluated

at the current iteration k. The simplest update strategy is the steepest descent (gradient descent) method, where $\mathbf{d_k}(\mathbf{x})$ is set to the negative gradient of the objective function with respect to $\mathbf{u_k}(\mathbf{x})$. More advanced optimizers approximate second-order information (e.g., Newton's method) or incorporate momentum (e.g., ADAM) to accelerate convergence and improve stability.

- **Hierarchical image registration:** To improve convergence and increase the likelihood of successful registration across diverse cases, the optimization process is typically executed in a hierarchical fashion (Lester and Arridge, 1999). Rather than performing a single optimization pass, registration is carried out systematically at multiple levels of image resolution – commonly referred to as an image pyramid – and transformation complexity. The process begins with coarse settings, typically using downsampled images and simple linear transformations (rigid or affine), to achieve an initial global alignment. It then progresses to higher resolutions where more flexible nonlinear transformations are applied, allowing for refined, localized alignment.

Image Normalization is an essential preprocessing step that needs to be considered to each modality in multi-modal analysis. Medical images that represent relative signal intensities – such as MRI sequences including T1w, T2w, and DWI – can exhibit substantial variability in both intensity offset and scale across patients. Without normalization, these variations must be learned by the model, which unnecessarily increases training complexity. Additionally, using a fixed learning rate across modalities with differing intensity distributions can lead to suboptimal or inconsistent convergence. To address this, per-case normalization is commonly applied to relative-valued modalities, improving training stability and convergence.

Other medical image modalities represent absolute physical quantities and benefit from different normalization schemes. A prominent example is CT imaging, where Hounsfield Units (HU) quantify X-ray attenuation on a standardized scale from approximately -1000 to +3000. For such modalities, per-case normalization would compromise the physical consistency of the signal. Instead, global normalization – applied uniformly across the dataset – is preferred, ideally using a meaningful intensity window tailored to the application or target of interest.

A gray area in choosing the normalization scheme is the ADC map, derived from multiple DWI acquisitions with different b-values, as described in Section 1.2.2. While ADC values have physical units $[\mathrm{mm}^2\,\mathrm{s}^{-1}]$, they are highly dependent on the underlying diffusion model and acquisition protocol. In homogeneous or mono-center datasets, global normalization can enhance model performance. However, in multi-center or

heterogeneous cohorts, variation in the applied diffusion model and acquisition protocol settings can undermine the consistency of ADC values. Thus, normalization strategies must be chosen carefully based on the dataset composition.

Data Loading and Online Data Augmenation

The most important goal of AI model development is to maximize generalizability, ensuring that models perform reliably on unseen data, ultimately leading to an applicable model in real-life scenarios. Generalization refers to a model's ability to maintain high performance on an independent test set, as compared to the training (or also called as development) set. Overfitting occurs when a model memorizes the training data rather than learning general patterns or discriminative features. The risk of overfitting is greatly reduced when a sufficiently large and diverse training dataset is available, encompassing a wide variety of clinical scenarios and edge cases.

In the domain of natural image processing, access to large-scale datasets, often containing millions of annotated examples (Deng et al., 2009; Lin et al., 2014), and data diversity is generally not a bottleneck. As a result, model-centric strategies to improve model generalization and capacity are the primary focus. Architectural developments like the evolution from AlexNet (Krizhevsky et al., 2012) through VGG (Simonyan and Zisserman, 2014), Inception (Szegedy et al., 2016), and DenseNet (Huang et al., 2017), to modern Vision Transformers (Zhai et al., 2022) is an active field of research. In addition, large-scale pretraining strategies – such as contrastive learning (Chen et al., 2020; Caron et al., 2021; Oquab et al., 2023) and self-supervised learning (He et al., 2022) – have further improved generalization by leveraging information from unlabeled data, leading to the first foundation models (Bommasani et al., 2021).

In contrast, publicly available 3D medical imaging datasets are far more limited, especially those with high-quality expert annotations. The creation of large, labeled medical datasets is highly labor-intensive, requiring time-consuming annotation by clinical experts, which is typically not part of their standard clinical duties. Moreover, privacy regulations and ethical constraints restrict the availability and sharing of medical data, as it contains sensitive personal health information. Although there are existing datasets providing substantial data for general organ segmentation models, such as TotalSegmentator (Wasserthal et al., 2023; Akinci D'Antonoli et al., 2025), the rarity and heterogeneity of many diseases, particularly in oncology, result in imbalanced and underrepresented samples for specific diagnostic tasks. This makes overfitting an even greater challenge in medical AI applications.

Compared to model-centric solutions, data augmentation is a data-centric strategy that addresses the problem of overfitting from the root of the problem, the data set

itself (Shorten and Khoshgoftaar, 2019). DA plays a crucial role in the success of DL in medical image analysis by artificially increasing the size of training data. The core assumption of DA is that more information can be extracted from each training sample by simulating realistic scenarios that can occur during image acquisition, such as patient motion (e.g., rotations, translations), variability in acquisition parameters (e.g., gamma transformations), or the introduction of noise and imaging artifacts. Importantly, the model prediction has to be invariant to such context-dependent features that are unrelated to the underlying pathology. In this sense, DA also plays a critical role in enhancing model robustness. A necessary condition for effective DA is that the transformations must be label-preserving, meaning that they should not alter the underlying ground truth. A well-known counterexamples are 180° rotation and mirroring of MNIST digits (Deng, 2012), which swaps the labels of digits of '6' and '9', clearly violating the label-preservation requirement. In disease stratification, similar caution must be taken to ensure that augmentations do not distort the diagnostic labels.

Another challenge unique to 3D medical image analysis is the high dimensionality of the data – especially in multi-modal scenarios – which necessitates patch-wise training due to current memory constraints. This introduces further challenges, including class imbalance and loss of contextual information. For example, MRI images that represent relative values require contextual information to be meaningfully interpreted, as Region of Interest (RoI) intensities alone are insufficient. Even in CT, which provides absolute Hounsfield Units, contextual information is helpful since similar HU values can belong to multiple anatomical structures. Therefore, large patch sizes are desirable for capturing both target and contextual information.

Patch-wise data loading contributes even further to the already existing problem of medical datasets being inherently imbalanced. Patches extracted from a cancerous volume may only contain benign tissue, further shifting the class distribution during training. To address this, oversampling strategies are often used by increasing the sampling rate of either foreground voxels or underrepresented classes to ensure balanced learning. Such techniques help train models that are unbiased with respect to class prevalence, aligning model predictions with the true clinical relevance of the findings rather than the statistical distribution in the dataset.

Supervised Representation Learning with Deep Neural Networks

Pattern recognition and machine learning systems are fundamentally designed to extract relevant features from input signals, enabling a classification subsystem to assign these signals to predefined categories. Unlike traditional machine learning approaches, where feature extraction and classification are distinct processes, deep learning methods unify

these steps by automatically learning features and classification rules simultaneously through a general-purpose learning process called training.

One common realization of such systems is through multi-layer perceptrons (MLPs), a class of ANNs composed of layers of simple processing units, known as perceptrons or artificial neurons. MLPs are highly parameterized models that realize nonlinear functions, mathematically described as:

$$f_{\text{MLP}}^{\mathbf{w}} : \mathbf{X} \mapsto \mathbf{Y}, \tag{1.13}$$

where $\mathbf{w}$ denotes the set of learnable parameters (weights and biases), $\mathbf{X} \in \mathbb{R}^n$ represents the input signal, and $\mathbf{Y} \in \mathbb{R}^q$ corresponds to the predicted output, typically a 1-of-q vector representing the probability distribution over q possible classes.

The predominant training paradigm in medical image analysis is supervised learning, where the objective is to predict a target output $\mathbf{T}$ (ground truth label) from a given input image, typically reflecting clinical diagnostic information. Training an MLP can be formulated as an optimization problem, where the goal is to optimize the network parameters $\mathbf{W}$ by minimizing an objective function $J(\mathbf{W})$, defined as the sum of individual error terms between the predicted outputs $\mathbf{Y}_i$ and the corresponding ground truth labels $\mathbf{T}_i$ across all n training examples $(\mathbf{X}_i, \mathbf{T}_i),\ i \in 1, \ldots, n$:

$$\min_{\mathbf{w}} \{J(\mathbf{w})\} = \min_{\mathbf{w}} \sum_i err\left\{\mathbf{Y}_i, \mathbf{T}_i\right\} = \min_{\mathbf{w}} \sum_i err\left\{f_{\text{MLP}}^{\mathbf{w}}\left(\mathbf{X}_i\right), \mathbf{T}_i\right\}, \tag{1.14}$$

where $\text{err}(\cdot, \cdot)$ denotes an appropriate loss function depending on the task.

Optimization is typically performed using gradient-based methods, with the backpropagation algorithm enabling efficient computation of the gradient of the objective function $J(\mathbf{w})$ with respect to $\mathbf{w}$. The network parameters are updated iteratively according to:

$$\mathbf{w}^{(k+1)} = \mathbf{w}^{(k)} - \mu \cdot \frac{\partial J(\mathbf{w})}{\partial \mathbf{w}}, \tag{1.15}$$

where μ is the learning rate controlling the step size of each update. Careful tuning of the learning rate μ is essential, as both excessively large and overly small values can hinder optimal convergence. A learning rate that is too large can cause the optimization process to miss the minimum of the loss function, while a too small value can result in slow convergence or cause the optimizer to become trapped in suboptimal local minima. However, in data-driven deep learning, reaching the global minimum of the training loss does not necessarily translate into satisfactory or convincing real-world performance. The goal of the training process is to learn general data representations or patterns that support accurate predictions not only on the training set but also on previously unseen samples. Therefore, model training is typically performed using 5-fold cross-validation (5fCV) for

hyperparameter tuning, and the final performance is evaluated on a strictly held-out independent test set to assess the real-world performance of the system. Model ability to perform well on such unseen data is referred to as generalization. In contrast, the case in which a model performs well on the training set but poorly on the validation or test set is called overfitting. It occurs when the model learns features that are too specific to the training set, and can be considered as memorization, instead of learning general representations. This is why deep learning methods are often referred to as representation learning approaches, where the aim is to capture meaningful features from the training data.

Since datasets typically contain a large number of images and the computational memory is limited, calculating gradients for all samples in each optimization step is computationally infeasible, particularly in the case of large 3D radiological image volumes. Therefore, modern optimizers apply strategies that compute gradients over subsets of the data, thereby balancing gradient calculation accuracy and training efficiency:

1. **Stochastic gradient descent:** This optimization algorithm computes gradients using only a randomly selected subset, called a mini-batch, of the training data in each iteration, rather than the entire dataset. This approach not only accelerates the training process but also improves robustness against local minima in the abstract feature space, while slightly reducing the risk of overfitting (Bottou, 2010; Bottou et al., 2018).

$$\mathbf{w}^{(k+1)} = \mathbf{w}^{(k)} - \mu \cdot \frac{\partial}{\partial \mathbf{w}} \sum_{i \in S} err\left\{ f_{\mathrm{MLP}}^{\mathbf{w}}\left(\mathbf{X}_i\right), \mathbf{T}_i \right\} \text{ with } S \subseteq \{1, ..., n\} \tag{1.16}$$

2. **Gradient descent with momentum:** This method extends stochastic gradient descent by incorporating a momentum term, which makes each weight update dependent on the previous update, thereby increasing the consistency of the optimization steps. The change in weights $\Delta \mathbf{w}^{(k+1)}$ is computed as a linear combination of the current gradient and the previous update:

$$\Delta \mathbf{w}^{(k+1)} = \alpha \cdot \Delta \mathbf{w}^{(k)} - \mu \cdot \frac{\partial}{\partial \mathbf{w}} \sum_{i \in S} err\left\{ f_{\mathrm{MLP}}^{\mathbf{w}}\left(\mathbf{X}_i\right), \mathbf{T}_i \right\} \text{ with } S \subseteq \{1, ..., n\}, \tag{1.17}$$

$$\mathbf{w}^{(k+1)} = \mathbf{w}^{(k)} + \Delta \mathbf{w}^{(k+1)}. \tag{1.18}$$

   Adding a momentum term, consisting of the momentum coefficient α and previous parameter update $\Delta \mathbf{w}^{(k)}$, to the weight update not only accelerates convergence by speeding up optimization in flat regions of the loss surface, but also reduces "zig-zagging" in steep valleys. This behavior is particularly helpful because first-order

gradient methods lack second-order information (curvature), and momentum helps to compensate for this limitation. Optimizers such as Adam (Diederik and Jimmy, 2014) incorporate momentum to enhance the stability and efficiency of convergence by leveraging historical gradients.

Convolutional Neural Networks

Currently, the most successful CAD systems are based on neural network architectures that utilize the building blocks of convolutional layers, referred to as convolutional neural networks (CNN) (LeCun et al., 1989, 1998).

CNNs follow the data-driven deep learning paradigm, similar to traditional MLPs, in which multiple stacked layers of artificial neurons – referred to as perceptrons – automatically learn increasingly abstract and semantically meaningful image features. The introduction of CNNs has been a major milestone over traditional MLPs with fully connected layers, primarily by drastically reducing the number of learnable parameters through architectural modifications. The two core architectural components that define CNNs are convolutional and pooling layers:

1. **Convolutional Layers:** The core building block of a CNN is the convolutional layer. Its learnable parameters consist of filters (also called kernels), which extract features from the input through a discrete convolution operation defined as:

$$(l * w)(x) = \sum_{n=0}^{N-1} l(n) \cdot w(x-n) \tag{1.19}$$

where l denotes the input activation (layer input), w the convolutional kernel (filter), and x the spatial output position. The primary function of the convolutional layer is to detect patterns in local spatial neighborhoods that are commonly present across the dataset and to encode their presence into feature maps. Each feature map is obtained by sliding the filter across the spatial dimensions of the input, computing the convolution at each location, adding a bias term, and applying a nonlinear activation function, resulting in the final layer output:

$$l_{i+1}(x) = f_{act}\big((l_i * w)(x) + w_0\big), \tag{1.20}$$

where w_0 is the learnable bias term and $f_{act}(\cdot)$ denotes a nonlinear activation function, such as the Rectified Linear Unit (ReLU).
Several key principles guide the design of convolutional layers:

- **Local receptive fields:** Each neuron in a convolutional layer is connected only to a small spatial region (or neighborhood) of the input. This design enables the

network to learn local patterns, such as edges or textures, which are essential for image understanding.

- **Weight sharing:** A single set of weights – also known as a convolution kernel or filter – is applied uniformly across all spatial locations. This allows the network to detect the same feature regardless of its position in the image, and drastically reduces the number of trainable parameters. It acts as a strong inductive bias for spatial invariance.
- **Multi-channel feature representations:** Each convolutional layer produces multiple feature maps (channels), with each channel capturing a different type of pattern or characteristic at every spatial location. This enables the network to represent a rich variety of features within the same layer.

2. **Pooling Layers:** The outputs of convolutional layers are typically followed by downsampling operations like max pooling. Pooling layers reduce the spatial resolution of the feature maps while simultaneously expanding the receptive field of subsequent layers, thereby allowing the network to capture broader contextual information. By stacking multiple convolutional and pooling layers, the network gradually transforms simple local features into increasingly complex and abstract representations.

Medical Image Analysis Tasks

The hierarchical feature extraction enables CNNs to learn both semantic meaning and spatial context, making them particularly well-suited for a wide range of medical image analysis tasks, including classification, segmentation, and object detection.

- **Classification** is the most common form of diagnostic task, where the goal is to assign a single score or label to an entire image. This score reflects the presence or severity of abnormality. In this context, CNNs typically employ an encoder architecture that compresses the high-dimensional input image into a low-dimensional representation, which is then mapped to a final prediction – either a single value or a categorical probability distribution for multi-class classification. The most commonly used loss function for training classification networks is the cross-entropy loss, defined as:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} t_i^{(c)} \cdot \log(p_i^{(c)}), \tag{1.21}$$

where N is the number of exams in the dataset, C is the number of classes, $t_i^{(c)}$ is the target indicator (1 if class c is the correct class for exam i, else 0), and $p_i^{(c)}$ is

the predicted probability for class c on exam i. In radiology, whole-volume 3D classification remains challenging and is unpopular among both clinicians and machine learning engineers. This is primarily because the RoI usually occupies a small part of the volume, making it difficult for the network to extract meaningful features related to the relevant areas. Additionally, 3D volumes require patch-wise training due to hardware limitations (see *Data Loading and Augmenation* in Section 1.2.3), which often leads to unstable optimization and a strong dependence on large-scale datasets that is rarely available in the medical domain. For these reasons, CNN-based classifiers are more commonly applied to 2D modalities, such as breast cancer detection in mammography (McKinney et al., 2020), chest pathology classification in chest X-ray scans (Majkowska et al., 2020), and melanoma detection in dermatoscopic images (Esteva et al., 2017). Even so, classification networks typically produce only a single output per image with no inherent interpretability, often referred to as black-box predictions. This lack of transparency has motivated a growing research interest in explainable AI techniques (Hamm et al., 2023; Naz et al., 2023), which aim to improve model trust and interpretability. In the case of 3D imaging, proxy tasks like segmentation or detection are favored, as they offer voxel- or object-level information that can be derived to a patient-level classification.

- **Semantic segmentation** provides the most detailed spatial information by predicting class labels at the voxel level, combining both semantic and spatial understanding. In contrast to classification, which outputs a single label per image, semantic segmentation assigns a label to every voxel in the input volume. To achieve this, the U-Net architecture was introduced (Ronneberger et al., 2015), which combines deep semantic features extracted by an encoder with high-dimensional spatial features from earlier shallow layers. The spatial features are passed to the decoder via skip connections at each level of the network, enabling the preservation of localization information, that are lost during the encoding process. An illustration of the U-Net architecture is shown in Fig. 1.12. The most commonly used loss function for semantic segmentation is a combination of two terms that account for spatial accuracy. The first is the voxel-wise cross-entropy loss, which extends the standard classification loss to operate on individual voxels. The second is the Sørensen-Dice coefficient, often referred to as Dice loss, which quantifies the overlap between predicted and ground truth segmentations. It is defined as:

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2 \sum_i \hat{p}_i \cdot \hat{t}_i}{\sum_i \hat{p}_i + \sum_i \hat{t}_i}, \tag{1.22}$$

where $\hat{p}_i = \arg\max_c \, p_i^{(c)}$ and $\hat{t}_i = \arg\max_c \, t_i^{(c)}$ denote the predicted and ground

truth class labels for voxel i, obtained by applying the argmax operation over the class dimension c. Here, $p_i^{(c)}$ and $t_i^{(c)}$ represent the predicted probability and one-hot encoded ground truth label for class c, respectively. The advantage of incorporating Dice loss is that it provides a strong training signal even for small objects, making it particularly valuable in imbalanced medical datasets, where lesions or abnormalities occupy only a small fraction of the total 3D volume. This dense supervision contributes to more stable training, even for annotated data sets limited in size. Moreover, the interpretable predictions, due to their rich spatial detail, make semantic segmentation highly clinically valuable. As a result, semantic segmentation has become the most widely adopted task in 3D medical image analysis (Maier-Hein et al., 2018).

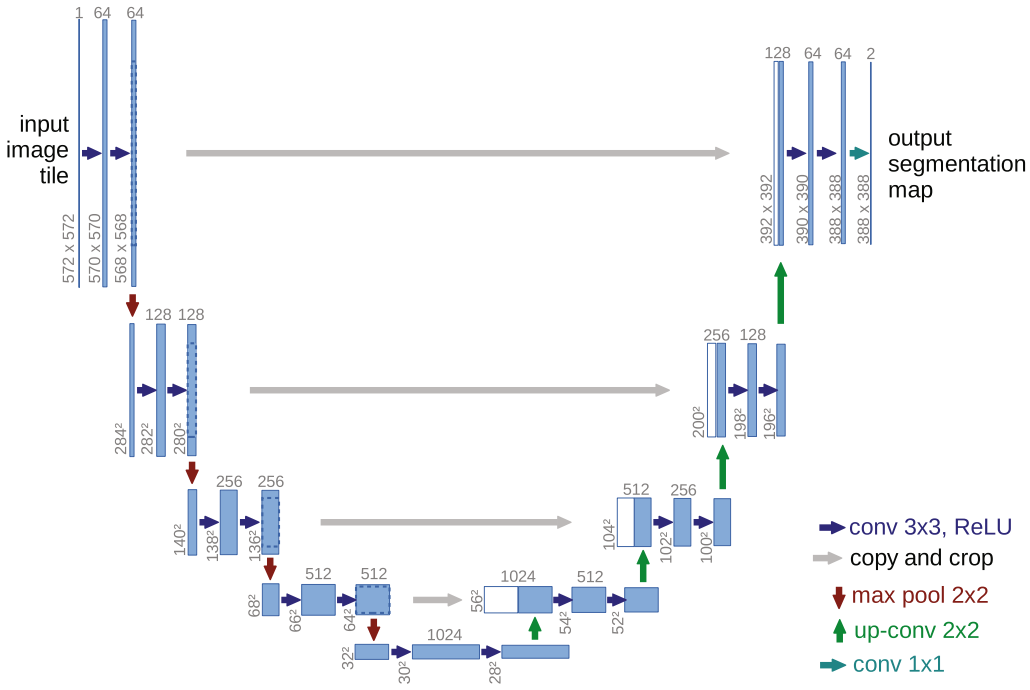

Figure 1.12: U-Net architecture for semantic segmentation. The network combines deep semantic features from the encoder with high-resolution spatial information from earlier layers via skip connections. These connections between corresponding encoder and decoder blocks help preserve fine localization details lost during downsampling. Figure taken from Ronneberger et al. (2015) with permission from the publisher Springer Nature.

Evaluation

In clinical practice, diagnostic decisions are ultimately made at the patient level, determining whether a patient has a particular disease in order to guide further therapeutic decisions. Accordingly, the performance of a CAD system is primarily assessed at the patient level using receiver operating characteristic receiver operating characteristic (ROC) analysis (Hanley and McNeil, 1982), which evaluates the trade-off between sensitivity and specificity across a range of classification thresholds, see Fig. 1.13. To summarize the performance across all thresholds, the area under the ROC curve area under the receiver operating characteristic curve (AUROC) is commonly reported. However, based on specific clinical conditions, performance at a particular operating point, defined by a selected threshold, is also frequently reported.

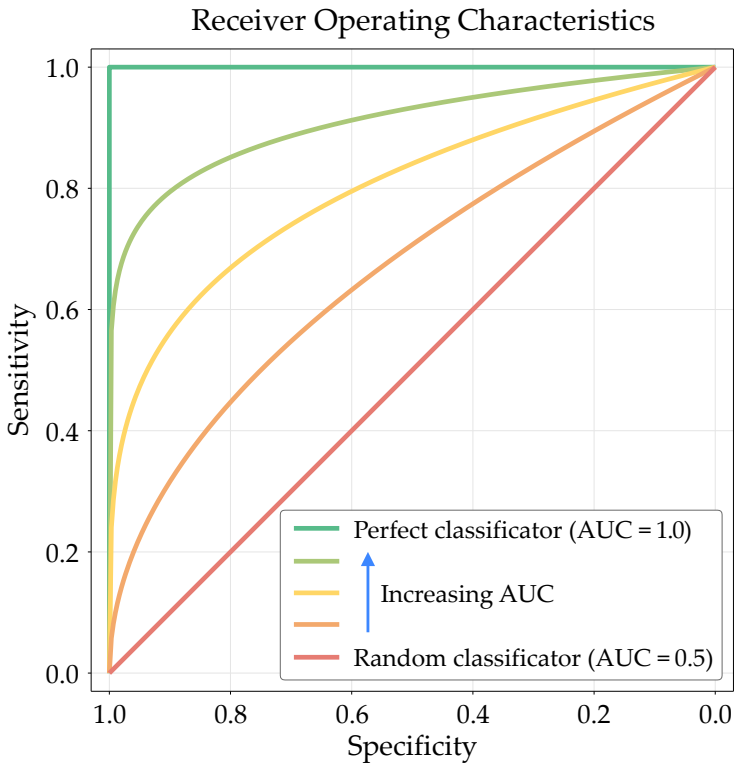

Figure 1.13: Example of receiver operating characteristic (ROC) curves for multiple classifier models, representing patient-level performance across classification thresholds.

Although AUROC is widely used for patient-level model evaluation, it does not account for the spatial correctness of predictions. This can be a limitation, especially for diseases

where subsequent clinical procedures rely on precise localization of the pathology. In such cases, relying solely on AUROC may result in suboptimal model selection.

Recent advances in location-dependent interventions – such as targeted biopsy (Lomas and Ahmed, 2020), transurethral ultrasound ablation (Ghai et al., 2024), or focal radiation therapy boosts (Kerkmeijer et al., 2021) in PCa – have increased the need for accurate tumor localization. As the effectiveness of these interventions depends on the accurate localization of lesions, object-level evaluation metrics are gaining importance alongside patient-level metrics. A clinically well-interpretable metric for lesion-level performance is free-response receiver operating characteristic (FROC) curve (Van Ginneken et al., 2010), which plots sensitivity against the average number of false positives per scan (avgFPs/scan), measured across multiple thresholds and for a predefined matching criterion (typically based on Intersection over Union (IoU)), see Fig. 1.14.

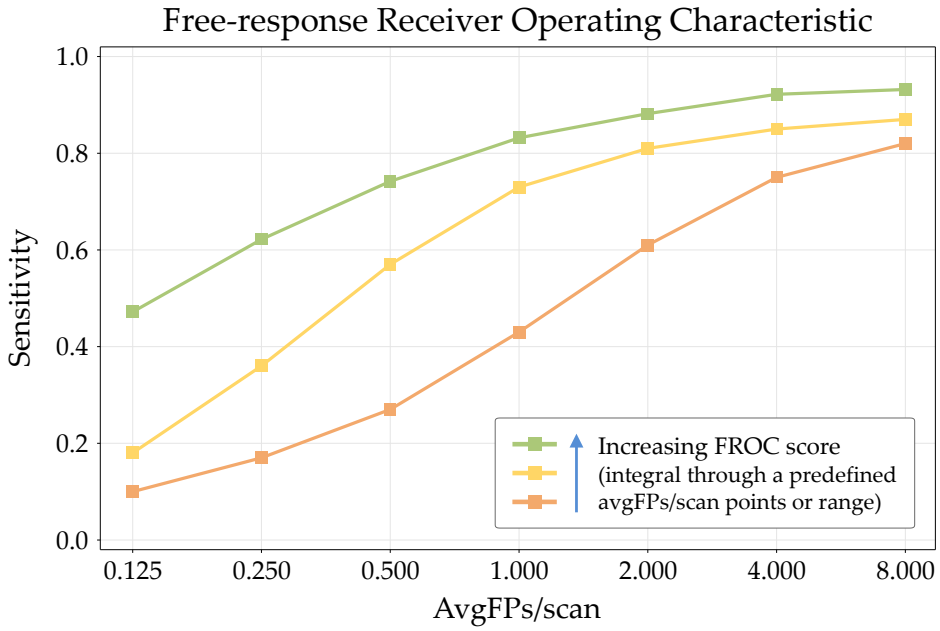

Figure 1.14: Example free-response receiver operating characteristic (FROC) curves showing lesion-level sensitivity across different average number of false positives per scan (avgFPs/scan) values.

Unlike ROC analysis, FROC places stricter requirements on what qualifies as a true positive by incorporating spatial overlap criteria. However, it also requires detailed object-level annotations from expert clinicians, even on the test set, which is an effort-intensive and non-trivial task limiting its widespread use despite its clinical relevance.

1.3 Related Work

Currently, methods based on semantic segmentation using CNN architectures are the most successful and widely adopted approaches for PCa diagnosis and interventions (Bhattacharya et al., 2022; Twilt et al., 2021). A significant milestone in biomedical image analysis was the introduction of the nnU-Net framework (Isensee et al., 2021), which also had a substantial impact on research in the field of PCa.

One of the key innovations of nnU-Net is the automatic configuration of design parameters for the vast variety of 3D medical imaging, addressing challenges discussed in Section 1.2.3 – such as normalization schemes for preprocessing, patch size for data loading, network topology aligned with the receptive field of convolutional perceptrons. These so-called rule-based parameters are derived from imaging properties, such as spatial dimensions and modality type. Additionally, empirical parameters – like resolution level or postprocessing strategies – are optimized considering the target-of-interest in a dataset to maximize performance.

In contrast to the model-centric AI trend, nnU-Net maintains fixed and straightforward training-related parameters, including the U-Net architecture, optimizer, learning rate, loss function, data augmentation scheme, and training/inference procedures. The central claim of the nnU-Net paper is that data-driven model configuration has a larger impact on performance than architecture variations. This was demonstrated by achieving state-of-the-art results on 33 out of 53 segmentation tasks and showing on-par performance on the remaining ones. The claim was further substantiated in follow-up studies through direct comparisons with alternative network architectures, including Mamba (Ma et al., 2024b) and Transformer-based models (He et al., 2023; Tang et al., 2022; Xie et al., 2021; Zhou et al., 2021), where nnU-Net consistently outperformed or matched their performance (Isensee et al., 2024). Since its original publication, nnU-Net has become a widely adopted and competitive baseline, dominating biomedical imaging challenges and establishing a new standard for reproducible, high-performance segmentation pipelines (Ma, 2021). Given its robust generalization capabilities and standardized implementation, this dissertation relies on the nnU-Net framework as the foundation for all experimental investigations.

The design choices implemented in nnU-Net are intended to generalize across a wide range of medical imaging tasks. As such, further performance gains may be achieved by incorporating solutions tailored to specific applications. However, this was explicitly outside the scope of the designers of the framework, as defining application-specific strategies for the vast diversity of medical imaging problems is infeasible. Consequently, challenges defined in Section 1.1 such as **Data-centric Challenge #1** and **Data-centric Challenge #2**, which are critical in the context of PCa diagnosis, remain unresolved. Despite the emphasis of nnU-Net on data-centric design, many current state-of-the-art

algorithms for PCa focus predominantly on model-centric solutions, often overlooking or only marginally addressing these application-specific challenges. In the remainder of this section, a selected group of research efforts that provide valuable starting points for addressing these challenges for medical image analysis are summarized.

1.3.1 Research on Soft Tissue Deformations of the Prostate

Biomechanics is a vast field of research dedicated to building models that simulate physiological processes in the human body. However, it is also one of the most complex areas, as modeling soft tissue deformation requires accounting for nonlinearity, time-dependency, and inhomogeneous as well as anisotropic tissue behavior (Payan and Ohayon, 2017). Constructing biomechanical models is a tedious and resource-intensive task involving multiple steps. This process is illustrated here using the example of pelvic organ modeling:

1. **Access to Organ Geometry:** This step requires accurate 3D geometry of the target and surrounding tissues. 3D radiological imaging modalities along with manual annotations or annotations generated using automated segmentation tools – such as nnU-Net trained on an annotated subset of data, or nnU-Net-based methods like TotalSegmentator (Wasserthal et al., 2023; Akinci D'Antonoli et al., 2025) – can be used to construct a 3D mesh.

2. **Incorporation of Biomechanical Properties and Boundary Conditions:** The estimation of soft tissue properties and their interactions is crucial. Prostate movement and deformation are primarily passive and result from muscle contractions and respiration, but primarily rectal and bladder filling/movements (Engels et al., 2020; Boubaker and Ganghoffer, 2017; Voyant et al., 2011). Rubod et al. (2012) provided comprehensive measurements of the mechanical properties of pelvic organs. They confirmed the hyperelastic and viscous behavior of the rectum and bladder, noting greater elasticity in the bladder. No significant location- or orientation-dependent differences were observed, except under high strain for the bladder. They also emphasized the importance of using human tissue for biomechanical modeling, demonstrating significant differences compared to animal-derived samples. Qasim et al. (2022) assumed isotropic linear elastic behavior using the Young's modulus (E) and Poisson's ratio (ν), showing similar values for the prostate and adjacent pelvic muscles (Hensel et al., 2007).

3. **Mesh Generation and Deformation Computation:** Biomechanical modeling is typically formulated as a partial differential equation problem and solved using the finite element (FE) method based on meshed organ geometries (Johnsen et al., 2015; Carter et al., 2005). FE-based prostate modeling has been used mainly in enhancing

registration for MRI-TRUS fusion biopsy (Khallaghi et al., 2015b; Qasim et al., 2022) and in image-guided radiotherapy for pelvic organs (Boubaker and Ganghoffer, 2017; Chai et al., 2011; Hensel et al., 2007).

However, FE-based biomechanical modeling is computationally expensive, often requiring minutes to hours per simulation and powerful hardware (Boubaker and Ganghoffer, 2017). To make such simulations more time-efficient, hybrid approaches have emerged. One direction combines biomechanical models with statistical shape modeling to construct statistical motion models (Hu et al., 2008, 2010; Khallaghi et al., 2015a), which have been used as offline DA for regularizing CNN-based image registration for MRI-TRUS fusion biopsy (Hu et al., 2018). More recently, data-driven motion modeling using deep neural networks (DNNs) has been explored (Romaguera et al., 2021), although these methods have not significantly reduced modeling complexity or computational cost.

Consequently, such deformation models do not scale yet to online DA. As a result, current state-of-the-art PCa CAD systems continue to rely on simple global transformations – such as translation, rotation, mirroring, and scaling (Aldoj et al., 2020; Bosma et al., 2023; Cao et al., 2019; Hamm et al., 2023; Hosseinzadeh et al., 2021; Netzer et al., 2021; Saha et al., 2021a) – leaving the variability of lesion morphology unchanged. Few studies have used random elastic deformations (Pellicer-Valero et al., 2022; Schelb et al., 2019), but the transformation lacks label-preserving properties (see Section 1.2.3), potentially risking the transformation of benign lesions like BPHs into malignant tumors.

1.3.2 Research on Multi-Modal Misalignments in Prostate MRI

Accurate ground truth is essential for training CNNs, particularly for semantic segmentation, which relies on spatial information on the finest level. However, the presence of misalignments within mpMRI – as elaborated in **Data-centric Challenge #2** – prevents spatially accurate correspondence between ground truth labels and all modalities. This ultimately limits the model ability to fuse complementary information across modalities. The literature reflects differing opinions on the significance of this issue:

- **Multi-modal misalignments are considered irrelevant in the context of prostate cancer detection:** A few studies argue that misalignments are negligible due to precautionary measures taken during MRI acquisition. Saha et al. (2021b) were on the opinion that the examination time was insufficient to produce significant spatial displacements across modalities. They highlighted the use of antispasmodic agents to reduce bowel motility and rectal catheters to minimize distension, claiming these precautions effectively mitigate deformation, supported by visual assessment. These views were adopted by Bosma et al. (2023) and Duran et al. (2022). Hosseinzadeh et al. (2021) did not address the problem.

- **Registration is needed:** The majority of studies argue that aligning T2w images with DWI images and ADC maps is essential to ensure accurate multi-modal information fusion. These studies employ various registration strategies using MI as the similarity metric:

  - **Linear registration:** Many works have applied rigid registration (Aldoj et al., 2020; Arif et al., 2020; Cao et al., 2019; De Vente et al., 2020; Kohl et al., 2017; Winkel et al., 2021). Sanyal et al. (2020) justified this choice by pointing out the potential risk of implausible image warping. This scenario is especially likely in regions with severe image artifacts, where similarity metrics may fail. Some studies extended rigid alignment with affine transformations for better alignment (Schelb et al., 2019).
  - **Deformable registration:** In contrast, Yang et al. (2017) emphasized that accurate registration is key for leveraging multi-modal information. They employed a hierarchical deformable registration approach. This approach has been adopted and extended by several other studies (Netzer et al., 2021, 2023; Pellicer-Valero et al., 2022; Schelb et al., 2021).

  However, these studies that emphasize the importance of spatial alignment did not evaluate whether the downstream task actually benefits from registration. Given that registration is an ill-posed problem (Section 1.2.3), it cannot eliminate all misalignments, particularly in mpMRI, where anisotropic voxel dimensions and severe local deformations resulting from imaging artifacts or large organ movements. As a result, critical spatial errors may persist even after registration.

1.4 Objectives and Contributions

1.4.1 Research Objectives

The primary objective of this dissertation is to enhance the performance and robustness of PCa diagnosis systems by incorporating data-centric solutions that have remained underexplored or have only been partially addressed. Specifically, the goal is to integrate application-specific knowledge – such as physiological processes and imaging-related characteristics – into AI model training. These are factors that radiologists are explicitly trained to recognize and account for during clinical decision-making. By embedding such domain knowledge into deep learning models, this work aims to narrow the performance gap between radiologists and CAD systems, ultimately leading to more accurate and clinically robust diagnostic systems.

This work addresses challenges associated with radiological prostate MRI assessment elaborated in Section 1.1, and targets the technical limitations discussed in Section 1.3. The proposed solutions are integrated and evaluated using a semantic segmentation framework, which relies heavily on spatial information and provides a standardized, state-of-the-art baseline for methodological development.

Objective 1: Soft Tissue Deformations

Given the influence of active pelvic soft tissue deformations on the captured MRI image and its interpretation identified as a **Data-centric Challenge #1**, and reviewing the current state of research on biomechanical modeling of the prostate in Subsection 1.3.1, I define the first research gap as follows:

> **Research Gap #1:** The high potential of soft tissue deformation models – specifically their ability to introduce substantial prostate and lesion shape variability as an inductive bias into training – remains underexplored in PCa diagnosis using deep learning-based AI systems.

To address **Research Gap #1**, the first part of the dissertation investigates the integration of biomechanical modeling into AI model training. The first research question addresses a core limitation of existing biomechanical models, namely their limited applicability for generating soft tissue deformations on the fly during model training:

> **Research Question 1:** Can a biomechanical model be constructed that mimics realistic soft tissue deformations and is suitable for online DA?

To reduce the computational cost, a lightweight mathematical model was designed using biomechanical properties of the pelvic organs informed by prior literature (Section 1.3.1). A Turing test involving expert radiologists was conducted to qualitatively validate the realism of the resulting deformations. Following this, the main performance-related question was posed:

> **Research Question 2:** Does increased lesion and organ morphological diversity by incorporating realistic soft tissue deformations into AI model training improve diagnostic performance?

To answer this, nnU-Net models were trained with and without the proposed deformations, with detailed evaluation at both patient- and lesion-level. In addition, a comparative

study was performed against random elastic deformations, which are easy to utilize but not guaranteed to be label-preserving, i.e., they may alter the underlying ground truth in unintended ways:

> **Research Question 3:** Is realism in augmentation crucial, or are generic random elastic deformations sufficient? Are they label-preserving in the context of prostate MRI?

This was tested by comparing model performance across both augmentation strategies.

Objective 2: Multi-Modal Misalignments

Given the importance of consistent ground truth labels across image modalities identified as a **Data-centric Challenge #2**, and based on the review of current approaches on multi-modal alignment errors reviewed in Subsection 1.3.2, I define the second research gap as follows:

> **Research Gap #2:** The problem of multi-modal alignment errors in PCa has remained largely unexplored. There is no consensus among research groups regarding the necessity or benefit of registration for improving CNN performance. Furthermore, alternative strategies to either replace registration or compensate for residual misalignments have not yet been investigated.

The second focus of the dissertation targets **Research Gap #2**, which was broken down into multiple research questions. The first addresses the lack of consensus in existing literature regarding registration:

> **Research Question 4:** Does registration improve CAD model performance for prostate MRI? Is that dependent on the type of registration?

To address this, nnU-Net pipelines were trained with various registration strategies and compared against a non-registered baseline. Moving beyond conventional registration, a research question addressing a new aspect of alignment errors is proposed:

> **Research Question 5:** Can an alternative strategy replace registration with a more practical and robust solution?

Instead of relying solely on alignment preprocessing, the proposed approach called *misalignment DA* aims to increase the model robustness against misalignments by simulating random alignment errors during training. This raised the final question:

Research Question 6: Does the proposed strategy complement registration, or can it serve as a replacement? Do they provide additive benefits when combined?

1.4.2 Summary of Main Contributions

In summary, this dissertation advances the use of data-centric model training strategies in PCa diagnosis on MRI. It contributes:

- A novel biomechanical model of pelvic soft tissue deformation, suitable for realistic, online data augmentation.
- An empirical evaluation of the benefit of realism in deformation-based augmentation.
- A comprehensive assessment of the necessity of registration.
- A new augmentation strategy for training models to be robust to multi-modal alignment errors.
- Evidence that these data-centric approaches enhance both patient-level and lesion-level performance.

1.5 Outline

The structure of this thesis is organized around the two main research objectives introduced in Section 1.4, focusing on 1) soft tissue deformations and 2) multi-modal alignment errors in prostate MRI. Each objective is addressed primarily through a dedicated research study, which is followed in parallel across Chapter 2, Chapter 3, and Chapter 4.

In Section 2.1, the characteristics of the clinical datasets used in both studies are described in detail, highlighting their properties supporting the experiments. Section 2.2 focuses on the development of a biomechanical model for simulating soft tissue deformations in the human pelvis, along with the qualitative and quantitative evaluation techniques used to assess the realism and practical benefits of the generated transformations. Section 2.3 presents the preparation of datasets with different registration techniques, the proposed misalignment data augmentation paradigm, and the experimental design used to systematically evaluate both strategies.

The corresponding results are presented in Section 3.1 and Section 3.2, respectively, covering both qualitative findings and quantitative diagnostic performance. These findings are then discussed in detail in Section 4.1 and Section 4.2, with interpretations specific to each research objective.

Finally, Chapter 4 also reflects on the generalizability of the proposed techniques beyond prostate MRI, and discusses the broader implications of incorporating application-specific knowledge into the data augmentation process. The potential future role of such domain-informed augmentation is considered in the context of emerging trends in medical image analysis.

Chapter 2

Materials and Methodology

2.1 Characteristics of Datasets

Although the Prostate Imaging Reporting and Data System (PI-RADS) guidelines recommend multi-parametric MRI (mpMRI), including dynamic contrast-enhanced imaging (DCE) imaging, recent studies have shown that bi-parametric MRI (bpMRI) – composed of T2-weighted (T2w) and diffusion-weighted imaging (DWI) sequences – can offer comparable diagnostic performance (Gatti et al., 2019; Kuhl et al., 2017; Liang et al., 2020; Tavakoli et al., 2023; Twilt et al., 2025; Zawaideh et al., 2020). While bpMRI may involve a slight trade-off in accuracy, it eliminates the need for contrast agent (CA) administration and allows for faster acquisition times (Van der Leest et al., 2019), making it an appealing choice not only in clinical workflows but also for emerging deep learning applications in prostate magnetic resonance imaging (MRI). Accordingly, this dissertation develops artificial intelligence (AI) systems based on images in the bi-parametric abbreviated MRI setting.

2.1.1 In-House Data

For this dissertation, an in-house dataset was provided by David Bonekamp from the Division of Radiology at the German Cancer Research Center (DKFZ) Heidelberg. This dataset is of particularly high quality, featuring carefully curated ground truth labels established through close collaboration with the Department of Urology and the Institute of Pathology at the University of Heidelberg Medical Center.

A key strength of this dataset is the use of extended systematic transperineal MRI-transrectal ultrasound (TRUS) fusion biopsies, which offer lesion detection sensitivity comparable to that of radical prostatectomy specimens Kuru et al. (2013). Unlike many publicly available datasets, which often provide annotations only on DWI images, which typically suffer from lower spatial accuracy, this dataset includes accurate T2w lesion segmentations. These high-precision segmentations make the dataset especially well-suited for the development and evaluation of data-centric approaches that rely on fine spatial information.

Patient Sample

This dissertation had access to a subset of institutional patients undergoing routine clinical care, all of whom received mpMRI followed by MRI-TRUS fusion transperineal targeted biopsies as well as extended systematic biopsies. The retrospective analysis of these study samples was approved by the ethics committee of the Medical Faculty Heidelberg, which waived the requirement for informed consent (institutional ethics approval number S-164/2019). All experiments were conducted in accordance with the Declaration of Helsinki (64th WMA General Assembly, Fortaleza, Brazil, October 2013) and relevant data privacy regulations.

For the studies included in this dissertation, access was granted to institutional data consisting of consecutive prostate MRI examinations acquired over different time periods, without reflecting any methodological considerations:

1. **In-House Cohort #1:** Exams acquired between January 2014 and December 2016
2. **In-House Cohort #2:** Exams acquired between January 2014 and December 2017

Despite the different time periods, all cohorts were subject to identical inclusion and exclusion criteria to ensure consistency across analyses. All men with suspected prostate cancer (PCa) or those enrolled in the institutional active surveillance program were included if the following conditions were met:

- MRI examination was performed using institutional scanners,
- biopsy was performed at the same institution,
- biopsy results were available.

Criteria for exclusion were:

- prior PCa therapy (prostatectomy, ablation or radiation therapy, anti-hormonal treatment),
- MRI scans with severe imaging artifacts,
- biopsy performed within 2 months prior MRI or more than 6 months post MRI,
- uncommon/rare pathological diagnoses with irregular/atypical imaging manifestations (e.g. smooth muscle cell tumor of uncertain malignant potential),
- exams in which clinically significant PCa (csPCa) was detected via systematic biopsy without MRI-visible lesion. The exclusion of these cases was necessary to enable both lesion-level evaluation and semantic segmentation network training requiring lesion ground truth segmentations.

Detailed demographic and clinical characteristics of the in-house cohorts are given in Appendix A.1.

MRI Protocol

mpMRI prostate examinations were performed on three different scanners from Siemens Healthineers: Magnetom Prisma and Biograph mMR (both 3T), and Magnetom Aera (1.5T). All examinations followed PI-RADS recommendations (Turkbey et al., 2019) and ESUR guidelines (Barentsz et al., 2012). Standard multichannel body coils and integrated spine-phased array coils were applied for image acquisition. As this dissertation uses biparametric abbreviated MRI protocol, only the acquisition parameters of T2w and DWI sequences are provided in Appendix A.2, stratified by scanner type.

PI-RADS and Histopathological Assessment

PI-RADS assessment of MRI-detected lesions was performed during clinical routine by board-certified radiologists. All exams were discussed in interdisciplinary conferences prior to biopsy.

All patients received extended systematic transperineal MRI-TRUS fusion biopsies according to the Ginsburg protocol and PI-RADS ≥ 3 lesions were assessed by targeted biopsies, a method that has demonstrated reliable ground-truth assessment with sensitivity comparable to radical prostatectomy specimen (Kuru et al., 2013). Before 03/2017, biopsies were guided by rigid (BiopSee by Medcom, Darmstadt) and afterward elastic (UroNav, Philips Invivo) software registration. Histological assessment was performed under the supervision of an experienced genitourinary pathologist with 20 years of experience. csPCa was defined as International Society of Urological Pathology (ISUP) Gleason Grade Group (GGG) ≥ 2.

Dataset Annotation for Segmentation Ground Truth

Existing csPCa segmentations were used as ground truth lesion annotations. These were retrospectively created on both T2w and DWI images by multiple in-house investigators, utilizing radiologists' reports and accompanying diagrams. T2w segmentations were performed with the aid of corresponding DWI images and ADC maps, leading to spatially highly accurate ground truth. All annotations were performed using the Medical Imaging Interaction Toolkit (MITK) (Wolf et al., 2005), under the supervision of a board-certified, fellowship-trained radiologist with 12 years of experience in prostate MRI (Prof. Dr. David Bonekamp).

To prepare the dataset for object-level evaluation and for assessing lesion-level overlap across modalities, I systematically reviewed the lesion annotations and identified

inaccuracies and inter-modality inconsistencies. These included missing slices, overlapping instance labels, and touching lesions. While these issues are negligible for semantic segmentation training and patient-level evaluation, they critically compromise the lesion-level performance assessment. Additionally, I detected mismatched lesion identifiers, missing slices, and incorrect or overlapping labels between lesion annotations on T2w images and ADC maps. These inconsistencies would have significantly affect the validity of lesion-level overlap across modalities used for evaluating image registration techniques. These inaccuracies and inconsistencies were corrected in collaboration with members of the Division of Radiology at DKFZ Heidelberg (Clara Meinzer, Nils Netzer, Cedric Weißer, Dr. Kevin S. Zhang, Prof. Dr. David Bonekamp).

Existing segmentations of the prostate gland on both T2w and DWI images, as well as rectum and bladder segmentations on T2w images, were previously generated using in-house nnU-Net models. These models were iteratively trained on a cohort that initially included a small subset of this patient population. The quality of all segmentations was confirmed by multiple in-house readers.

2.1.2 PROSTATEx Dataset

This dissertation also includes data from the public PROSTATEx dataset (Armato et al., 2018), which contains mpMRI examinations performed on two different 3 T MRI scanners (Siemens MAGNETOM Trio and Skyra), without the use of an endorectal coil.

Ground truth lesion annotations for 204 exams were available through existing in-house csPCa segmentations. These lesions were retrospectively segmented on both T2w and DWI images by a single in-house investigator, based on the publicly provided lesion coordinates and under the same supervision as described in Section 2.1.1. Additionally, prostate gland segmentations for both T2w and DWI images were included, generated using the same nnU-Net-based algorithm referenced in Section 2.1.1.

2.2 Soft Tissue Deformations of the Prostate (Objective #1)

The methods presented in this section have been primarily published in the following conference:

Balint Kovacs, Nils Netzer, Michael Baumgartner, Carolin Eith, Dimitrios Bounias, Clara Meinzer, Paul F. Jäger, Kevin S. Zhang, Ralf Floca, Adrian Schrader, Fabian Isensee, Regula Gnirs, Magdalena Görtz, Viktoria Schütz, Albrecht Stenzinger, Markus Hohenfellner, Heinz-Peter Schlemmer, Ivo Wolf, David Bonekamp, Klaus H. Maier-Hein *Anatomy-informed data augmentation for enhanced prostate cancer detection*. **International Conference on Medical Image Computing and Computer-Assisted Intervention - MICCAI 2023**. https://doi.org/10.1007/978-3-031-43990-2_50

2.2.1 Study Cohort

This study utilizes bpMRI exams from In-House Cohort #2 – described in Section 2.1.1 – comprising a total of 774 exams. The dataset was divided into a training set and a test set using an 80 % - 20 % split, resulting in 619 exams for training and 155 for testing. Stratification was performed based on the prevalence of csPCa, which was 36.3 % across the cohort to ensure a balanced representation of benign and malignant cases. An overview of the data split is provided in Tab. 2.1, while a detailed demographic breakdown is available in the supplementary material.

Table 2.1: Distribution of patients with and without clinically significant prostate cancer (csPCa) across the training and test sets. Stratified sampling ensured consistent csPCa prevalence in both subsets. Table adapted from our previously published work (Kovacs et al., 2023a), with permission from the publisher Springer Nature.

No. Exams	without csPCa	with csPCa	Sum
Training set	394	225	619
Test set	99	56	155
Sum	410	215	774

2.2.2 Biomechanical Model Creation

The first step toward enabling soft tissue deformation simulation for use in online data augmentation (DA) workflows is the development of a transformation model that effectively balances anatomical realism with computational efficiency. To achieve this, I constructed a lightweight biomechanical model inspired by the known mechanical properties of pelvic tissues, as discussed in Section 1.3.1.

In order to keep the model both conceptually simple and practically scalable, several simplifications were introduced:

- Soft tissue deformation of the prostate is a passive process resulting from active morphological changes in the rectum and bladder due to physiological functions such as filling and evacuation (Boubaker and Ganghoffer, 2017; Engels et al., 2020; Voyant et al., 2011). Therefore, the transformation is designed to simulate bladder and rectal shrinkage and dilation.

- Given the isotropic mechanical behavior of both the rectum and bladder (Rubod et al., 2012), the simulated transformation is applied isotropically and perpendicularly to the organ surface without prioritizing any specific direction.

- Due to the similar Young's modulus and Poisson's ratio of the prostate and surrounding muscle tissue (Qasim et al., 2022), these tissues are treated as a homogeneous material during the simulation of bladder and rectal deformation.

- Taking into account that deformation-preserving solid tissues – particularly the pelvic bones – are either distant from the prostate or absent from the imaging field of view, their influence on prostate deformation is considered negligible. As such, explicitly modeling boundary conditions would unnecessarily increase the complexity of the simulation.

- The nonlinear hyperelastic behavior of soft tissues (Rubod et al., 2012) is typically the most computationally intensive aspect of biomechanical models. To simplify this, the deformation is modeled as a displacement field that decays proportionally with distance from the rectal and bladder surfaces. Grid sampling and interpolation are then used to approximate the resulting deformation, achieving a similar effect with substantially reduced computational cost.

- Viscous tissue behavior – i.e., force- and time-dependent deformation – can be disregarded, as these dynamics are irrelevant in the context of static transformations for augmentation.

As a result, a soft tissue deformation originating from either the rectum or the bladder can be simulated using a transformation vector field V, defined as:

$$V = \nabla(G_\sigma * S_{organ}(x, y, z)) \cdot C(x, y, z), \tag{2.1}$$

where ∇ denotes the spatial gradient, G_σ is a Gaussian kernel with kernel size σ, and S_{organ} is the binary segmentation mask (indicator function) of the rectum or bladder. The scalar field of $C(x, y, z)$ controls the amplitude and direction of the resulting displacement field. Applying the vector field V as a displacement field to an MRI image $I(x, y, z)$ results in a spatially transformed image with simulated soft tissue deformation:

$$I^T(x, y, z) = I(x + V_x(x, y, z), y + V_y(x, y, z), z + V_z(x, y, z)). \tag{2.2}$$

This transformation is referred to as the anatomy-informed transformation. Fig. 2.1 illustrates the simulation pipeline, showing each step from the original anatomy to the deformed image.

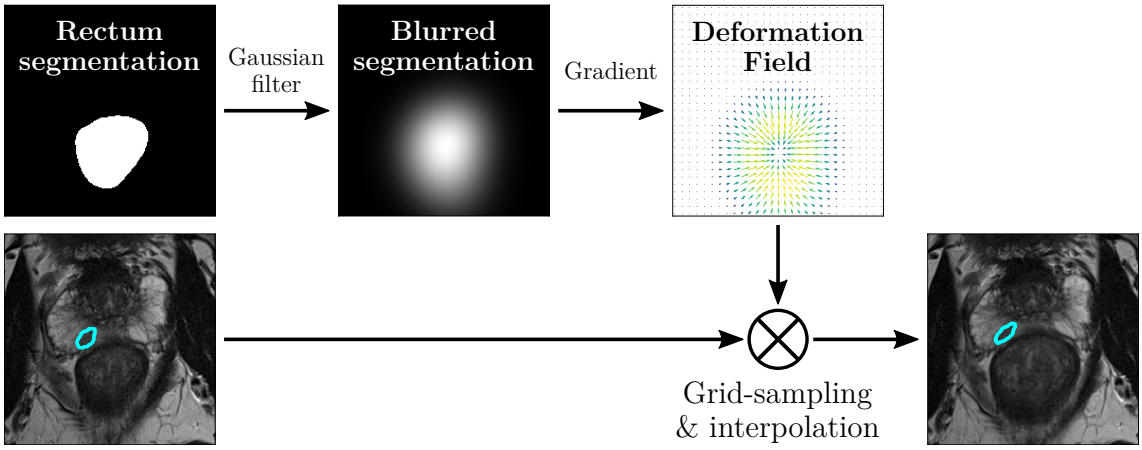

Figure 2.1: Anatomy-informed transformation pipeline demonsterated on a PROSTATEx exam (Armato et al., 2018). Soft tissue deformation is simulated by computing a deformation vector field from anatomical segmentations – such as rectal and/or bladder. The binary segmentation is first blurred using a Gaussian filter, and its gradient is then computed to generate a vector field. This field is applied as a displacement field to the input MRI image via grid sampling and interpolation, resulting in a transformed image that simulates anatomically plausible soft tissue deformation.

2.2.3 Experimental Setting

Qualitative Evaluation

First, the assumptions underlying the biomechanical model described in Section 2.2.2 were evaluated qualitatively, specifically assessing whether the applied rectal and bladder deformations produce visually realistic images. For this purpose, a controlled experiment is performed in which either the anatomy-informed transformation is applied either to the rectum or bladder, a random deformable transformation is applied, or no transformation is applied at all. These transformations are introduced in a randomly selected subset of exams. I then conduct a rigorous Turing test involving clinicians with varying levels of radiology expertise – a newly graduated physician (Carolin Eith) and two radiology residents specializing in prostate MRI (Clara Meinzer and Dr. Kevin S. Zhang, with 1.5 to 3 years of experience). Each clinician was asked to assess whether a given image was original or artificially modified, responding to the question: "Is this image an original MRI scan or has it been artificially altered? If you think it has been modified, please describe why." Responses were considered correct only if the artificial modification was identified with a valid justification, ensuring that cases where the decision was made for an incorrect reason were excluded.

Assessing Applicability

In this context, applicability is in relation to the computational complexity of the transformation, as it directly affects its computational time. Although an exact measurement of computation time is not feasible due to its dependence on image characteristics, transformation complexity, and hardware specifications, such precise timing is not essential in this context. In practice, data loading and augmentation are executed on the central processing unit (CPU) in parallel with model training steps (forward pass, loss computation, backpropagation, parameter update) that are handled on the graphical processing unit (GPU). Therefore, the critical question for applicability is whether the CPU can complete the transformation calculations in time before the GPU completes its training iteration. Therefore, I monitored GPU utilization during model training to detect any differences between training with and without the anatomy-informed transformation.

Quantitative Evaluation

To assess the clinical relevance of the anatomy-informed transformation, I quantitatively evaluate its impact on the tasks of patient-level PCa diagnosis and lesion-level PCa detection. Following prior works (Duran et al., 2022; Kohl et al., 2017; Netzer et al., 2021; Saha et al., 2021b; Sanyal et al., 2020), I derive diagnostic predictions through

semantic segmentation of malignant lesions. Semantic segmentation not only yields spatially detailed, interpretable outputs, but it is also sensitive to spatial changes, making it particularly suitable for evaluating spatial DA strategies.

To systematically compare the impact of the proposed anatomy-informed DA against commonly used techniques, I define the following three DA schemes:

1. **Basic augmentation strategy (baseline reference):**
   This refers to the standard nnU-Net pipeline (Isensee et al., 2021), which uses an extensive augmentation strategy including simple spatial transformations such as translation, rotation, and scaling. It serves as the baseline reference for model comparison.

2. **Random deformable extended strategy:**
   This strategy extends the basic augmentation pipeline by incorporating random elastic deformations, as optionally implemented in nnU-Net. While it has been used in the natural imaging domain (Simard et al., 2003), the effect of such transformations in the medical imaging domain remains uncertain due to the lack of guarantees for label preservation (Perez et al., 2018). I hypothesize that these deformations may introduce unrealistic anatomical changes – such as altering lesion characteristics – potentially degrading model performance.

3. **Anatomy-informed extended strategy (proposed method):**
   This strategy extends the basic augmentation pipeline by incorporating the proposed anatomy-informed transformation, which applies organ-specific deformations using the constructed biomechanical model. Two configurations are defined:

   (a) Deforming only the rectum, as rectal distension has the strongest impact on prostate morphology (Boubaker and Ganghoffer, 2017), particularly given that around 70 % of lesions are located in the adjacent peripheral prostate zone (PZ) (Ali et al., 2022).
   (b) Deforming both the bladder and the rectum, as bladder deformation also affects lesion appearance, though to a lesser extent.

To compare model performance with expert radiologists, I evaluate radiologist performance using clinical patient-level and lesion-level PI-RADS scores in conjunction with histopathological ground truth. PI-RADS scores are interpreted as prediction thresholds, while the histopathological findings serve as binary ground truth labels. This enables the computation of patient-level sensitivity and specificity, as well as lesion-level sensitivity and average number of false positives per scan (avgFPs/scan), across different PI-RADS thresholds. These thresholds can then be used as clinically meaningful operating points for model evaluation:

- **Patient-level evaluation:** Patient-level model performance is assessed using the partial area under the receiver operating characteristic curve (pAUROC), computed with a clinically meaningful sensitivity threshold of 78.75 % – corresponding to 90 % of the sensitivity achieved by radiologists at the PI-RADS ≥ 4 threshold. In addition, F_1-scores (the harmonic mean of precision and recall / sensitivity and positive predictive value) at the operating point corresponding to the sensitivity of PI-RADS ≥ 4 are reported to quantify performance further.

- **Lesion-level evaluation:** Free-response receiver operating characteristic (FROC) curves are calculated and the number of detected lesions at the avgFPs/scan of 0.32 per scan – corresponding to the radiologists' lesion-level sensitivity for PI-RADS ≥ 4. Predicted lesion instances are extracted by thresholding the softmax segmentation output at 0.5, followed by connected component analysis. A prediction is counted as a true positive if its Intersection over Union (IoU) with a ground truth lesion exceeds 0.1.

- **Zonal evaluation to test locality-specific effects of the transformation:** To assess whether localized soft-tissue deformations lead to zone-specific improvements in lesion detection, I evaluate the number of correctly detected lesions separately for the PZ and the transitional prostate zone (TZ). This comparison is made across nnU-Net configurations using the basic augmentation strategy and its extensions with different anatomy-informed transformations (rectum only, bladder only, and both rectum and bladder). My hypothesis is that targeted deformations can selectively enhance lesion detection sensitivity in the corresponding zonal region.

Bootstrapping with 1000 replications was applied to calculate p-values for the F1-scores and for the number of detected lesions using two-sided t-test determining statistical significance.

2.2.4 AI Model Developement

Input Images and Ground Truth Labels

The network receives a multi-channel input composed of the bpMRI images – T2w image, the DWI image with the highest b-value, and the corresponding apparent diffusion coefficient (ADC) map. For each exam, a multi-label ground truth is constructed. It includes semantic segmentations of csPCa lesions – derived from systematic biopsy-enhanced lesion ground truth annotations – as well as organ segmentations for the rectum and bladder to support the anatomy-informed DA.

Image Preprocessing

To ensure accurate spatial matching of ground truth segmentations across the multi-modal input channels, the MRI sequences were co-registered using B-spline deformable registration using mutual information as the image similarity metric, following the settings described by Netzer et al. (2021). The displacement field was computed by registering the T2w image to the DWI sequence with the lowest b-value, as these exhibit the most similar tissue contrast characteristics. The resulting transformation was then applied to the DWI images with the highest b-value and the corresponding ADC maps.

Following registration, the images were cropped around the anatomical regions of interest. Cropping was performed using offsets of ± 9 mm axial to the prostate (corresponding to ± 3 slices), and ± 11.25 mm in the axial plane around the rectum and bladder (equivalent to ± 32 voxels, matching the Gaussian kernel parameter described later in Section 2.2.4). This step ensures compatibility with the anatomy-informed augmentation pipeline and helps mitigate training instability, which can arise due to the limited number of csPCa cases and the relatively small lesion volumes in relation to the entire image (Sanyal et al., 2020).

Anatomy-informed Augmentation Parameters and Implementation

Some parameters of the anatomy-informed transformation were primarily determined based on visual plausibility. A Gaussian kernel size of 32 was found to effectively influence anatomically plausible regions around the rectum and bladder while maintaining sufficient spatial resolution in the deformation vector field. Given the rectum's tubular structure, oriented approximately along the superior-posterior axis, deformation was applied only within the axial plane. The amplitude parameter C for both the rectum and bladder was optimized using validation results from the 5-fold cross-validation (5fCV). During training, values were sampled from a uniform distribution centered at zero with bounds $C \in [300, 600, 900, 1200, 1500]$. The final selected values were $C_{\text{rectum}} = 1200$ and $C_{\text{bladder}} = 600$. To account for anisotropic voxel spacing $(0.3125, 0.3125, 3)$, the z-component of both the Gaussian kernel and the displacement field was scaled by the spacing ratio $0.3125/3$. The complete set of transformation parameters is provided in Tab. 2.2.

Table 2.2: Parameters used in the implementation of the anatomy-informed transformation. Anisotropic voxel spacing is taken into account, modifying the z-component of both the Gaussian kernel and the transformation amplitude. Rectal deformations are constrained to the axial plane, while bladder deformations are applied in all directions with z-scaling.

Parameter	Value
Voxel spacing (x,y,z)	(0.3125, 0.3125, 3.0) mm
Spacing ratio (SR)	$0.3125/3$
σ	$32 \cdot (1, 1, SR)$
C_{rectum}	$[-1200, 1200] \cdot (1, 1, 0)$
$C_{bladder}$	$[-600, 600] \cdot (1, 1, SR)$
probability	0.2

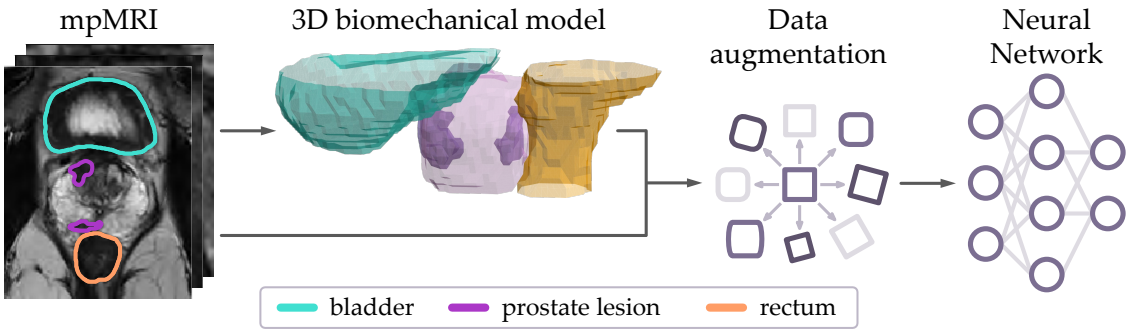

Figure 2.2: Integration of the anatomy-informed transformation into the nnU-Net training pipeline. During data loading, anatomical segmentations of the bladder and rectum are used to compute soft tissue deformations with a 20 % probability. The transformation is applied to both the image and label volumes. To ensure the model focuses on lesion segmentation, the anatomical labels are removed before the training step. Figure adapted from our previously published work (Kovacs et al., 2023a), with permission from the publisher Springer Nature.

The anatomy-informed transformation is integrated into model training as an online DA, as illustrated in Fig. 2.2. During training, the data loader provides MRI scans along with corresponding ground truth lesion annotations and anatomical segmentations of the rectum and bladder. The anatomy-informed transformation is applied for each organ with a probability of 20 %. When triggered, the transformation is computed using the provided organ segmentations and applied to both the input MRI images and corresponding labels.

To ensure that the network focuses solely on lesion segmentation, the anatomical labels (rectum and bladder) are removed from the ground truth prior to training. This avoids the risk of the model prioritizing larger anatomical structures over the relatively small lesion volumes, which would otherwise dilute the segmentation signal for csPCa.

As nnU-Net relies on the 3D medical image augmentation framework batchgenerators (Isensee et al., 2020), the anatomy-informed transformation was integrated directly into this framework following its conventions, and has been made publicly available: `https://github.com/MIC-DKFZ/batchgenerators` [accessed: 15 June 2025]. In addition, a customized nnU-Net trainer extended with the anatomy-informed augmentation has been released at `https://github.com/MIC-DKFZ/anatomy_informed_DA` [accessed: 15 June 2025].

nnU-Net rule-based Parameters of the Training Pipeline

Based on the dataset fingerprint, nnU-Net automatically extracted rule-based parameters for preprocessing and network configuration:

- **Normalization:** Z-score normalization was performed to the T2w and high b-value DWI images on a per-case basis due to their relative intensity nature. In contrast, ADC maps, considered to reflect absolute physical values and generated consistently within the same institution and scanner vendor, were globally normalized. nnU-Net used a global mean intensity of 1054.54 and a global standard deviation of 714.05 for the ADC maps.
- **Resampling:** All input modalities were linearly resampled to a common voxel spacing, calculated as the median spacing across the dataset: $[3.0, 0.3125, 0.3125]$ mm in the z-y-x axes.
- **Patch size and network topology:** Given the median image shape of $(21, 480, 311)$ in the z-y-x axes, nnU-Net determined the input patch size to be $(16 \times 384 \times 256)$. Based on this configuration, the network architecture was composed of two convolutional layers per stage, with pooling operations applied $[2, 6, 6]$ times along the respective axes. The network started with 32 feature maps, and a batch size of 2 was used. An overview of the resulting network configuration is illustrated in Fig. 2.3.

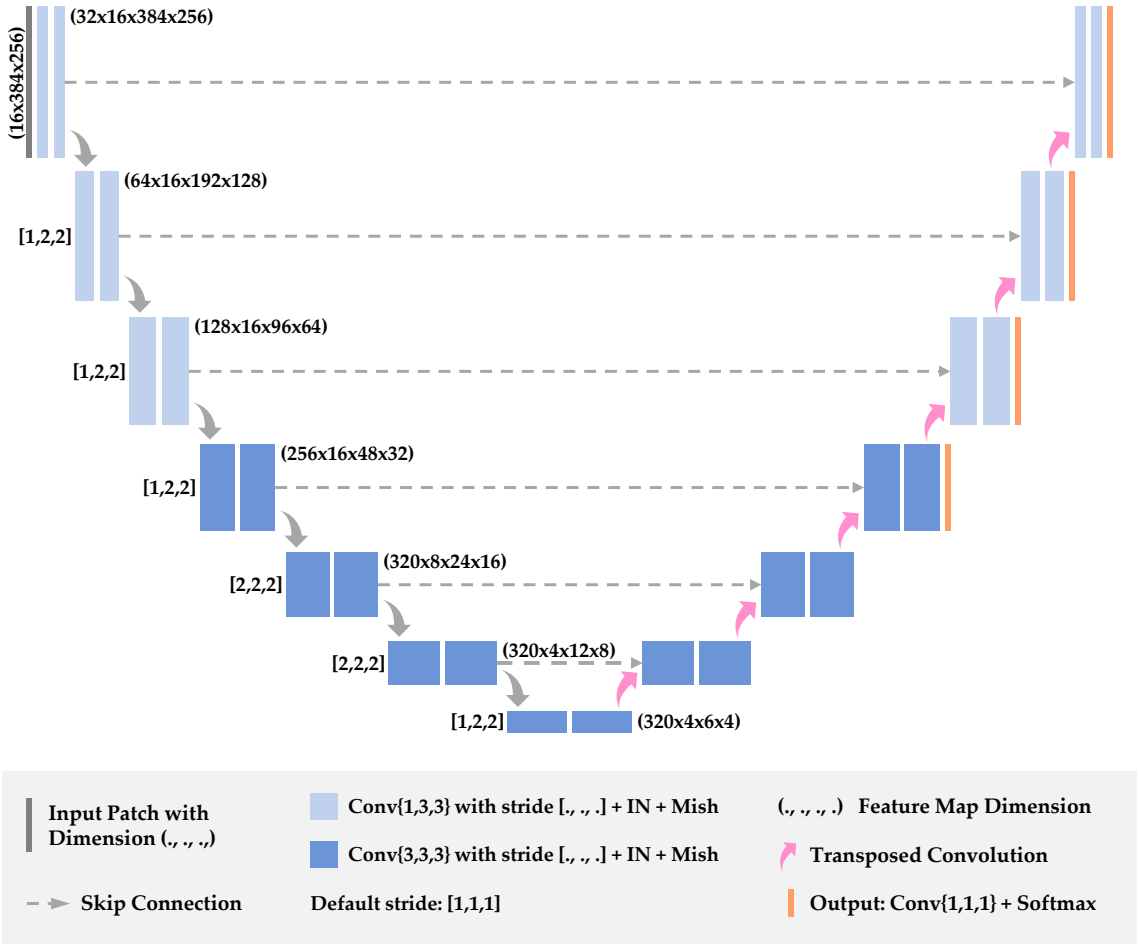

Figure 2.3: The generated U-Net topology for the anatomy-informed training configuration. Each encoder block consists of two convolutional layers (Conv) followed by instance normalization (IN) and Mish activation. Downsampling is performed via strided convolutions, while upsampling is achieved using transposed convolutions. The input patch size is $(16 \times 384 \times 256)$, with $[2, 6, 6]$ pooling operations per axis and an initial feature map count of 32. Final predictions are produced by a $1 \times 1 \times 1$ convolution followed by softmax activation, applied in a deep supervision manner across four decoder stages. The architecture is reconstructed based on nnU-Net configuration files following the convention used in Isensee et al. (2021) with permission from the publisher Springer Nature.

Network training

3D nnU-Net models were trained using a 5fCV strategy. Beyond the default nnU-Net configuration, a class-balanced data loader was employed to ensure training was not biased by the prevalence of csPCa in the dataset, thereby encouraging the network to focus more on learning discriminative image features. To reduce the risk of overfitting, the number of training epochs was limited to 350. Model training was performed using the Mish activation function, the Ranger optimizer, and a cosine annealing learning rate scheduler with an initial learning rate of 10^{-3}. The final models from the 5fCV folds were ensembled and evaluated on an independent test set.

2.3 Multi-Modal Misalignments in Prostate MRI (Objective #2)

The methods presented in this section have been primarily published in the following journal article:

Balint Kovacs, Nils Netzer, Michael Baumgartner, Adrian Schrader, Fabian Isensee, Cedric Weißer, Ivo Wolf, Magdalena Görtz, Paul F. Jaeger, Victoria Schütz, Ralf Floca, Regula Gnirs, Albrecht Stenzinger, Markus Hohenfellner, Heinz-Peter Schlemmer, David Bonekamp, Klaus H. Maier-Hein. **Nature Scientific Reports**. *Addressing image misalignments in multiparametric prostate MRI for enhanced computer-aided diagnosis of prostate cancer*. https://doi.org/10.1038/s41598-023-46747-z

2.3.1 Study Cohort

To ensure a sufficiently large and representative dataset for training and evaluation, I combined bpMRI exams from two cohorts: the public PROSTATEx cohort (204 exams) and the In-House Cohort #1 (421 exams), as described in Section 2.1.1. This resulted in a total of 625 exams.

Combining the two cohorts was considered feasible since both datasets consist of MRI scans acquired on the same vendor platform (Siemens), with similar ADC map calculations. More importantly, lesion segmentations were performed by in-house investigators, minimizing potential interrater variability across the two datasets.

As this study primarily focuses on methodological development rather than addressing the challenges of multi-center domain shift, efforts were made to minimize potential inter-cohort variability during training and evaluation. To achieve this, the combined dataset was split into training and test sets using an 80 %–20 % stratified split. Stratification was performed with respect to both csPCa prevalence – which was 34.4 % across the full dataset – and cohort origin, ensuring a balanced distribution of benign and malignant cases as well as cohort-specific consistency within each subset. An overview of the final cohort composition is provided in Tab. 2.3.

Table 2.3: Cohort composition across training and test sets, stratified by both the prevalence of clinically significant prostate cancer (csPCa) and dataset origin (PROSTATEx and In-House Cohort #1). Stratification by disease prevalence ensured a balanced distribution of benign and malignant cases, while cohort-based stratification helped minimize the influence of potential institutional domain shifts during methodological development and evaluation.

No. Exams (PROSTATEx+In-House)	without csPCa	with csPCa	Sum
Training set	327 (105+222)	169 (56+113)	496 (161+335)
Test set	83 (29+54)	46 (14+32)	129 (43+86)
Sum	410 (134+276)	215 (70+145)	625 (204+421)

2.3.2 Registration Techniques Used for Model Training

To enable a quantitative comparison of different registration strategies as a preprocessing step for model training, multiple multi-modal datasets were generated using the study cohort described in Section 2.3.1. An overview of these datasets is illustrated in Fig. 2.4.

Each dataset corresponds to a distinct registration technique, differing in complexity and clinical applicability:

1. **Dataset #1: Unregistered** This dataset contains the native images without any postprocessing for alignment. The input modalities (T2w, DWI with the highest b-value, and ADC map) remain unregistered. According to my hypothesis, this setting serves as the lower performance bound (see Section 2.3.5): the segmentation ground truth – typically derived from high-resolution T2w images – may not spatially align with the diffusion-derived modalities. This misalignment is particularly critical for small lesions, where minor spatial errors can result in major overlap loss. For larger lesions, the impact is expected to be less pronounced.

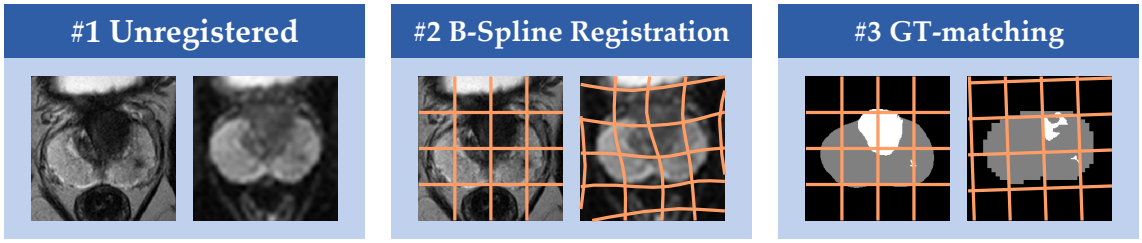

Figure 2.4: Illustration of the three registration techniques used to generate distinct multi-modal datasets for model training. Dataset #1 (Unregistered) includes native modalities reflecting the lower performance bound. Dataset #2 (B-spline Registration) applies deformable registration using mutual information to align modalities based on clinically available imaging, representing a practical and interpretable clinical setting. Dataset #3 (GT-matching) uses ground truth (GT) segmentations for alignment, aiming for maximal label consistency across modalities but relying on future knowledge, making it a non-clinical, reference-only setting. Figure parts adapted from our previously published work Kovacs et al. (2023b), reused under the Creative Commons Attribution 4.0 License.

2. **Dataset #2: Deformable B-spline Registered (Clinically Applicable)** This dataset represents the clinically interpretable setting where alignment errors are corrected using deformable registration based solely on routinely available imaging data. Registration follows the strategy of Netzer et al. (2021), which was optimized for the same in-house cohort. The method applies a B-spline deformable registration using mutual information (MI) as the similarity metric (Collignon et al., 1995; Viola and Wells III, 1997). As DWI images with the lowest b-values (0-50) show the closest appearance to T2w images, these are used to compute the transformation parameters, which are then applied to the corresponding DWI with the highest b-value and ADC maps. This results in high anatomical overlap across modalities. Since this registration is feasible in clinical practice and supports the direct interpretation of AI predictions, this dataset serves as the primary registered dataset for the main experiments, see Section 2.3.5.

3. **Dataset #3: Ground-Truth-Matched (Reference Only)** This dataset aims to provide a reference performance by using information unavailable during real-time clinical workflows. Similarly to the approach described in Sanyal et al. (2020), prostate segmentation masks are used for rigid registration, but I extend this by incorporating ground-truth lesion segmentations from both modalities. This ensures maximum spatial alignment of both the prostate and lesions across modalities, effectively eliminating ground truth inconsistencies caused by spatial misalignments between modal-

ities. Although this registration relies solely on rigid transformations, it directly addresses the challenge of inconsistent annotations due to spatial mismatches. While this method "cheats" by depending on information only available post-diagnosis thereby making it unsuitable for clinical deployment, it serves as a valuable reference for model performance to quantify the quality of the deformable image registration, see Section 2.3.5. I refer to this dataset as the ground-truth-matched (GT-matched) dataset.

2.3.3 Misalignment Augmentation: A Data-Centric Alternative to Multi-Modal Registration

The primary objective of image preprocessing is to resolve challenges that would otherwise be inefficient or suboptimal for AI models to learn directly during training. Registration is one such preprocessing step, typically applied before feeding multi-modal data into the model, with the goal of eliminating spatial misalignments that are assumed to be correctable. However, registration techniques – particularly in multi-modal medical imaging – are seldomly perfect and typically fail to fully correct local imaging distortions or anatomical inconsistencies.

While the primary motivation for data augmentation is to expand the training dataset and reduce the risk of overfitting Shorten and Khoshgoftaar (2019); LeCun et al. (1998), it also serves as a mechanism to introduce useful inductive biases that enhance model robustness. Specifically, augmentations help models generalize across transformations for which they are not inherently invariant. Spatial alignment errors between modalities – common in multi-parametric prostate MRI imaging – represent such a transformation, making them a potential candidate for such inductive biasing.

Building on this idea, I propose an alternative approach: instead of trying to resolve all alignment errors through registration, I simulate them. I introduce a technique called *misalignment augmentation*, a probabilistic data augmentation strategy that injects artificial alignment errors between image modalities during training. This approach can be easily integrated into any augmentation pipeline and operate on-the-fly. By simulating a range of plausible misalignment scenarios, the model is encouraged to learn representations that are invariant to those alignment errors.

The proposed technique serves as an alternative to complex registration pipelines, offering a simple yet effective strategy to improve robustness against alignment errors. Its effective utilization in training pipelines makes it a practical and lightweight alternative solution for handling multi-modal misalignments.

2.3.4 Evaluation Concept for Strategies Addressing Spatial Alignment Errors

Although registration methods have seen significant advances in recent years – particularly deep learning-based approaches Fu et al. (2020); Haskins et al. (2020) – they still rely on surrogate similarity metrics, which do not necessarily correlate with improved downstream clinical performance Rohlfing (2011). Therefore, instead of relying solely on such metrics, I evaluate strategies that address spatial alignment errors – including the registration techniques described in Section 2.3.2 and the proposed misalignment augmentation in Section 2.3.3 – directly on a clinically relevant downstream task: patient-level diagnosis of csPCa, as illustrated in Fig. 2.5.

I choose patient-wise whole image PCa diagnosis derived from semantic segmentation for the clinical downstream task (see Fig. 2.5c-e). Their predictions are not just strongly relying on spatial information, but semantic segmentation is clinically interpretable task by providing spatial localization De Fauw et al. (2018); Bernard et al. (2018); Nikolov et al. (2021). Additionally high quality lesion segmentation masks belonging to the T2w modalities were used due to their high spatial accuracy compared to the DWI annotations and because they are annotated by also taking information from the ADC maps into consideration, see detains in Dataset Annotation for Segmentation Ground Truth Section 2.1.1.

For assessing the performance of the trained models (Fig. 2.5g), the area under the receiver operating characteristic curve (AUROC) is used as a discrimination measure. Since the clinical diagnosis in case of PCa is based on the whole image, I am evaluating the results of the downstream task predictions as patient-wise AUC from the whole 3D images. For the patient-wise PCa prediction, the maximum value of the predicted lesion is taken (see Fig. 2.5h-i).

To be able to compare the performance of the trained models to radiologists' interpretation, I also calculate the radiologists' performance using the PI-RADS scores for the clinical index lesion as predictions and the maximum Gleason score of the systematic and targeted biopsy as the ground truth. According to PI-RADS, index lesions are scored on a Likert scale from 1 to 5 with higher scores indicating a higher risk of csPCa. The category of PI-RADS 3 has equivocal and PI-RADS 4 has high risk for csPCa Engels et al. (2020), which make these two categories the most informative area on the ROC curves. Thus, I calculate the sensitivity and specificity for PI-RADS ≥ 4 and PI-RADS ≥ 3. Calculating the performance of the radiologists during clinical practice provides a fixed reference point.

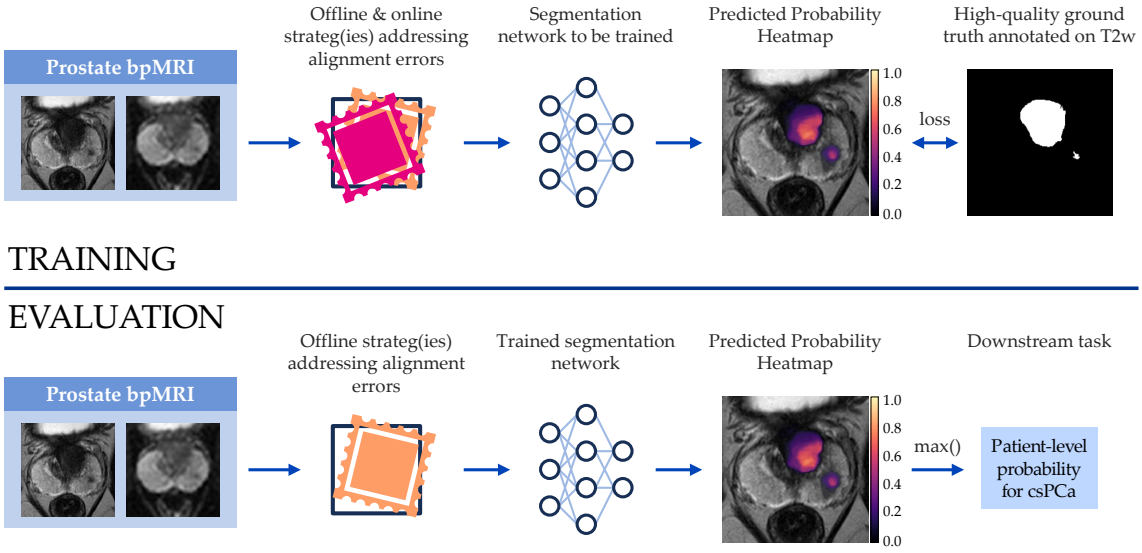

Figure 2.5: Overview of the evaluation concept for strategies addressing spatial alignment errors in prostate MRI. Rather than relying on surrogate alignment metrics commonly used for registration evaluation, the effectiveness of both offline registration (orange) and online misalignment augmentation (pink) strategies is assessed based on the clinical downstream task of prostate cancer (PCa) diagnosis derived from semantic segmentation of clinically significant PCa lesions. During training (top row), alignment correction strategies – including offline registration and online misalignment augmentation – are applied before segmentation network training. As the task of semantic segmentation relies heavily on spatial information, it serves as an ideal task to test the efficacy of spatial transformation strategies. Supervision is provided via high-quality lesion annotations on the T2-weighted (T2w) images, which offer a spatially accurate ground truth. During evaluation (bottom row), the trained network is applied to images preprocessed with offline registration. The resulting segmentation maps are used to compute patient-level malignancy probabilities. Final model performance is quantified using patient-level AUROC, calculated from the maximum value in the predicted probability heatmap. Figure parts adapted from our previously published work Kovacs et al. (2023b) under the Creative Commons Attribution 4.0 License.

2.3.5 Experimental Setting

Experiment #1: Assessing the Effect of Strategies Addressing Alignment Errors on Prostate Cancer Diagnosis

This experiment evaluates the necessity and effectiveness of strategies addressing multi-modal alignment errors – namely, image co-registration and the proposed misalignment augmentation – by quantifying their impact on the clinical downstream task of PCa diagnosis.

Combining these strategies results in four distinct model training configurations, illustrated in Fig. 2.6. These settings are:

1. **No Strategy for Alignment Errors / Baseline Training:** Standard nnU-Net training pipeline with its default augmentation strategy, without applying any registration as a preprocessing step (corresponding to Dataset #1). This configuration serves as the baseline and is hypothesized to represent the lower bound of performance.

2. **Training with Registered Images:** This setup includes deformable B-spline registration as a preprocessing step (corresponding to Dataset #2) to reduce large alignment errors. The hypothesis is that correcting these misalignments will improve the model's ability to fuse information across modalities, thereby enhancing segmentation performance.

3. **Training with Misalignment Augmentation:** This configuration augments the T2w image and its corresponding ground truth (GT) segmentation, while no registration is applied (Dataset #1). The goal is to simulate additional realistic alignment errors between the T2w image and the DWI/ADCmodalities. The applied transformations include:

   - Translation and rotation to mimic gland motion,
   - Scaling / squeezing in the dorsal-ventral direction to simulate acquisition-based geometric inconsistencies between T2w and DWI contrasts.

   The hypothesis is that this augmentation force the model to become partially invariant to the simulated alignment errors, resulting in a robust model that can achieve performance comparable to models trained with registration.

4. **Training with Both Strategies:** This combines registration (Dataset #2) with additional misalignment augmentation. The rationale is that registration reduces large misalignments, while augmentation accounts for remaining registration errors – potentially resulting in the highest diagnostic performance.

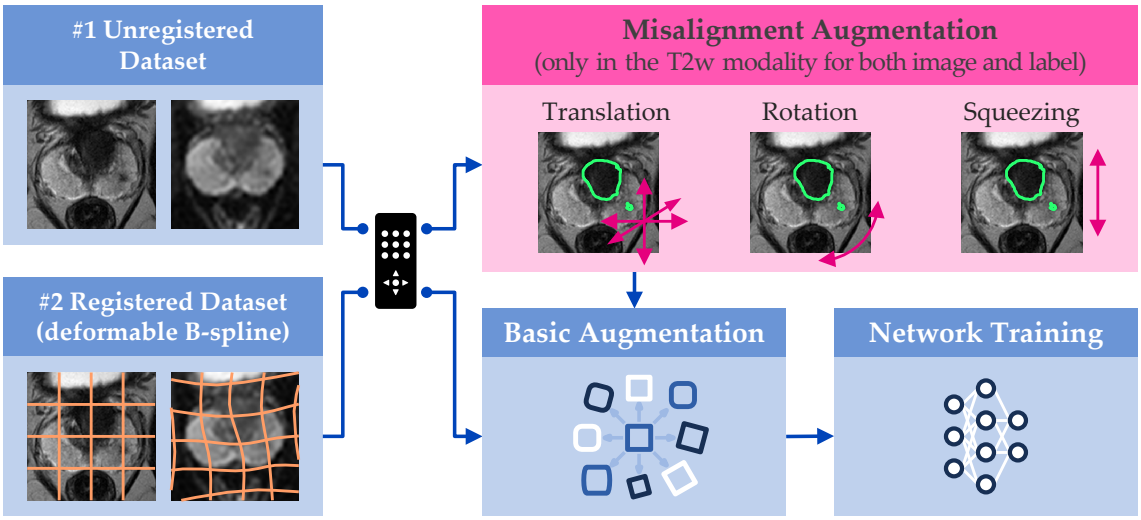

Figure 2.6: Overview of experimental configurations using strategies addressing alignment error for assessing the impact on prostate cancer diagnosis. Two input datasets are used: unregistered bpMRI images (Dataset #1) and deformably registered images (Dataset #2). Both can be fed into the standard nnU-Net training pipeline to assess the effect of registration. Misalignment augmentation can optionally be applied on top of either dataset, resulting in two additional training configurations: without registration, and with registration combined with misalignment augmentation. Misalignment transformations are applied only to the T2w modality (image and label), simulating realistic inter-modal alignment errors. This step is inserted before the standard nnU-Net augmentation pipeline to avoid introducing artifacts into downstream transformations. The experiment evaluates all four configurations to assess the individual and combined contributions of registration and misalignment augmentation to diagnostic performance. Figure parts adapted from our previously published work Kovacs et al. (2023b), reused under the Creative Commons Attribution 4.0 License.

Experiment #2 Assessing Registration Quality

This experiment evaluates the quality of the B-spline registration used in the main experiment (Section 2.3.5) by comparing datasets generated with different registration strategies. Two types of metrics are used for this assessment:

1. **Downstream clinical task performance:** Measured identically to Experiment #1, evaluating the impact of registration on prostate cancer diagnosis performance.
2. **Surrogate registration quality metric:** Dice score probability density functions are computed to quantify the spatial overlap between manually annotated lesions in the T2w and ADC images.

By analyzing both metrics, I aim to:

- Determine whether improved lesion-level alignment (higher T2w-ADC lesion Dice overlap) correlates with better diagnostic performance.
- Assess the suitability of the B-spline registration method used in the main experiment as a preprocessing step for multi-modal training.

2.3.6 AI Model Developement

Input Images and Ground Truth Labels

The network receives a multi-channel input consisting of bpMRI sequences: the T2w image, the DWI image with the highest b-value, and the corresponding ADC map. For high-quality lesion segmentation ground truth, clinical annotations on the T2w image were selected due to its superior spatial resolution compared to the ADC map and because the spatial information from the ADC map was also considered during the annotation process, enabling more precise delineation of lesion boundaries. The spatial reliability of the ground truth is further supported by the clinical systematic biopsy-enhanced lesion ground truth assesment. Both aspects are critical for evaluating strategies that aim to address multi-modal alignment errors.

Image Preprocessing

For model training, three preprocessed datasets were utilized, each corresponding to a different registration strategy: Dataset #1 contains native bpMRI images without registration, Dataset #2 includes deformably registered bpMRI using a B-spline transformations, and Dataset #3 applies rigid registration based on ground truth prostate and lesion segmentations (GT-matched). These preprocessing strategies are described in greater detail in Fig. 2.4.

To reduce potential training instability due to the limited number of csPCa cases and the relatively small lesion volumes compared to the full image size, all images were cropped around the anatomical region of interest – the prostate gland – following established practices (Sanyal et al. (2020); Netzer et al. (2021)). In addition to increasing class balance and improving model focus, this step also minimizes the influence of unrelated anatomical structures, which can otherwise introduce incorrect correlations in datasets with limited sample sizes.

Misalignment Augmentation Parameters and Implementation

The T2w sequence provides high-resolution structural information, whereas the DWI sequences have lower resolution and are more prone to distortion. Therefore, to simulate realistic multi-modal misalignments, the T2w modality – along with its corresponding lesion annotations – is displaced relative to the DWI sequences (DWI with high b-value and ADC map). To avoid introducing artifacts during subsequent augmentations, the misalignment augmentation is applied prior to any global transformations in the standard nnU-Net data augmentation pipeline.

Misalignments are generated using randomly sampled transformation parameters drawn from a uniform distribution, constrained by maximum amplitude thresholds in both positive and negative directions, following the conventions of the batchgenerators framework (Isensee et al., 2020). The following transformations are applied:

- Translation along the x, y, and z axes with maximum amplitudes of $(10, 10, 6)$ mm, respectively,
- Rotation in the axial (x-y) plane with a maximum amplitude of $15°$, avoiding z-axis rotations due to the anisotropic voxel spacing,
- Affine squeezing in the dorsal-ventral (z) direction with a maximum compression ratio of 0.1, reflecting natural imaging distortions commonly observed between T2w and DWI images caused by magnetic field inhomogeneities.

The augmentation was implemented directly into the batchgenerators medical data augmentation framework used by nnU-Net, in line with its implementation conventions. The code for misalignment transformations is publicly available at: `https://github.com/MIC-DKFZ/batchgenerators` [accessed: 15 June 2025]. A customized nnU-Net trainer that incorporates the misalignment augmentation strategy is also available at: `https://github.com/MIC-DKFZ/misalignment_DA` [accessed: 15 June 2025].

nnU-Net rule-based Parameters of the Training Pipeline

Based on the dataset fingerprint, nnU-Net automatically extracted rule-based parameters for preprocessing and network configuration:

- **Normalization:** Z-score normalization was performed to the T2w and high b-value DWI images on a per-case basis due to their relative intensity nature. Although the study cohort includes bpMRI exams from two institutions, all scans were acquired on scanners from the same vendor. A prior study Netzer et al. (2021), which also used the In-House and PROSTATEx datasets, suggested consistent ADC map calculations across centers. Therefore, the ADC maps were treated as absolute physical measurements and globally normalized. nnU-Net applied a global mean of 780.45 and a standard deviation of 240.66 for normalization of the ADC maps.

- **Resampling:** All input modalities were linearly resampled to a common voxel spacing, calculated as the median spacing across the dataset: $[3.0, 0.3125, 0.3125]$ mm in the z-y-x axes.

- **Patch size and network topology:** Given the median image shape of $(20, 260, 294)$ in the z-y-x axes, nnU-Net determined the input patch size to be $(20 \times 256 \times 320)$. Based on this configuration, the network architecture was composed of two convolutional layers per stage, with pooling operations applied $[2, 6, 6]$ times along the respective axes. The network started with 32 feature maps, and a batch size of 2 was used. An overview of the resulting network configuration is illustrated in Fig. 2.3.

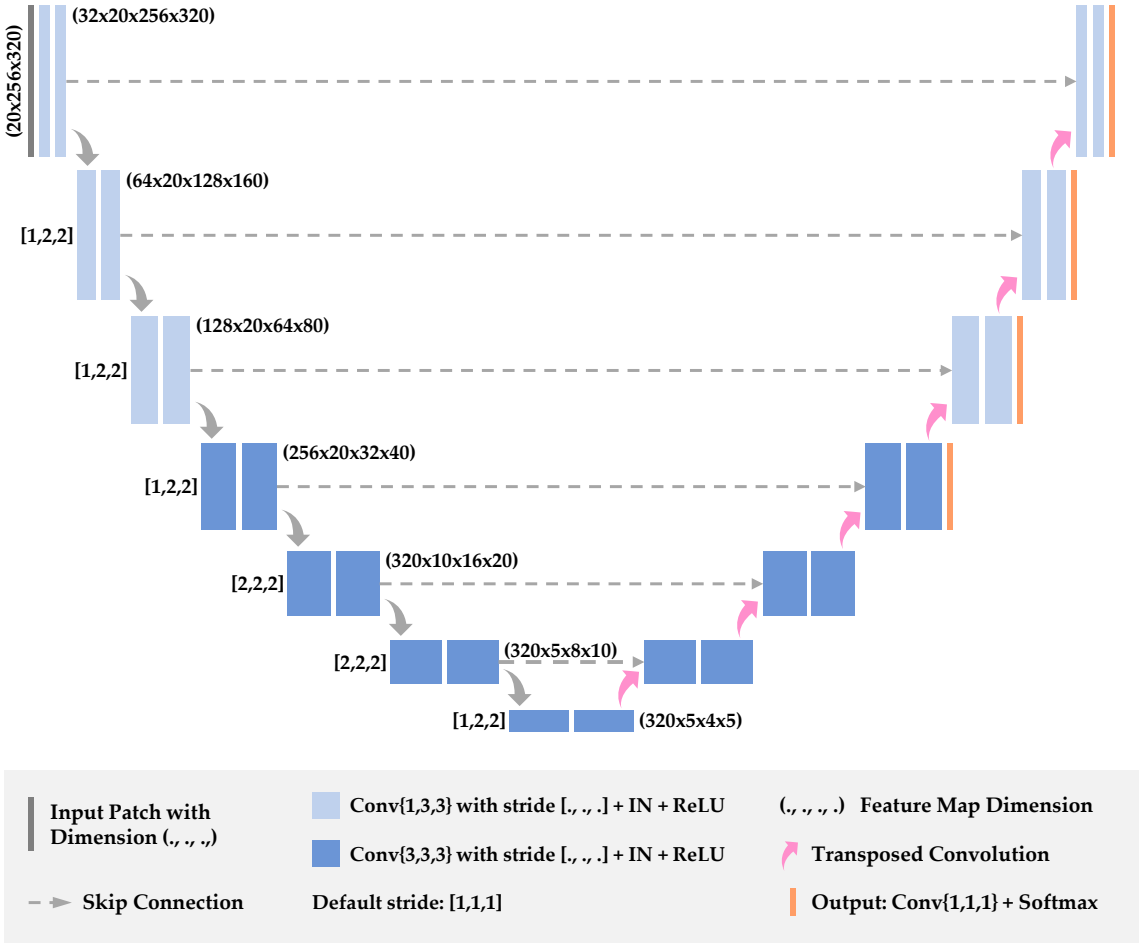

Figure 2.7: U-Net topology generated by nnU-Net for training configurations that incorporate strategies addressing multi-modal alignment errors. Each encoder block consists of two convolutional layers (Conv) followed by instance normalization (IN) and Mish activation. Downsampling is performed via strided convolutions, while upsampling is achieved using transposed convolutions. The input patch size is $(20 \times 256 \times 320)$, with $[2, 6, 6]$ pooling operations per axis and an initial feature map count of 32. Final predictions are produced by a $1 \times 1 \times 1$ convolution followed by softmax activation, applied in a deep supervision manner across four decoder stages. The architecture is reconstructed based on nnU-Net configuration files following the convention used in Isensee et al. (2021) with permission from the publisher Springer Nature.

Network training

For each training configuration (see Section 2.3.5) – defined by the type of preprocessing (see Section 2.3.2) and the presence or absence of misalignment augmentation (see Section 2.3.3) – a 3D nnU-Net ensemble was trained using 5fCV. Each fold produced an independently trained model, and final test predictions were obtained by ensembling outputs from all five models.

Model performance was optimized using early stopping based on cross-validation area under the receiver operating characteristic curve (AUROC). The probability of applying misalignment augmentation was tuned across the following values: $P = \{0.0, 0.1, 0.2, 0.4\}$.

Several modifications were introduced to the default nnU-Net training pipeline. Patient sampling was stratified to ensure balanced csPCa prevalence within each batch. A class-balanced data loader was used to mitigate bias introduced by class imbalance and encourage learning of discriminative features. The default Leaky ReLU activation was replaced by Mish for its smoother gradient properties, and the SGD optimizer was substituted with Ranger. Additionally, the learning rate schedule was switched from Poly to cosine annealing, with a reduced initial learning rate of 0.001 instead of the default 0.01.

Final model performance was evaluated on the independent test set using bootstrapping with 1000 replications to derive 95 % confidence intervals. To determine the statistical significance of differences between models, I applied the DeLong test (DeLong et al., 1988), considering results with $p < 0.05$ as statistically significant.

Chapter 3

Results

3.1 Soft Tissue Deformations of the Prostate (Objective #1)

The results presented in this section have been primarily published in the following conference:

Balint Kovacs, Nils Netzer, Michael Baumgartner, Carolin Eith, Dimitrios Bounias, Clara Meinzer, Paul F. Jäger, Kevin S. Zhang, Ralf Floca, Adrian Schrader, Fabian Isensee, Regula Gnirs, Magdalena Görtz, Viktoria Schütz, Albrecht Stenzinger, Markus Hohenfellner, Heinz-Peter Schlemmer, Ivo Wolf, David Bonekamp, Klaus H. Maier-Hein *Anatomy-informed data augmentation for enhanced prostate cancer detection*. **International Conference on Medical Image Computing and Computer-Assisted Intervention - MICCAI 2023**. https://doi.org/10.1007/978-3-031-43990-2_50

3.1.1 Highly realistic Images Passing the Turing Test - Qualitative Results

The anatomy-informed transformation produced highly realistic soft tissue deformations in the pelvic region. Fig. 3.1 and Fig. 3.2 show examples of transformations used during model training on a prostate magnetic resonance imaging (MRI) exam, simulating bladder and rectal physiological shape changes, respectively. In this case, the patient had clinically significant PCa (csPCa) lesions located in the anterior transitional prostate zone (TZ) and posterior peripheral prostate zone (PZ). The simulated deformations induced localized tissue shifts that affected lesion morphology only in prostate zones adjacent to the deformed organ – the bladder for TZ and the rectum for PZ.

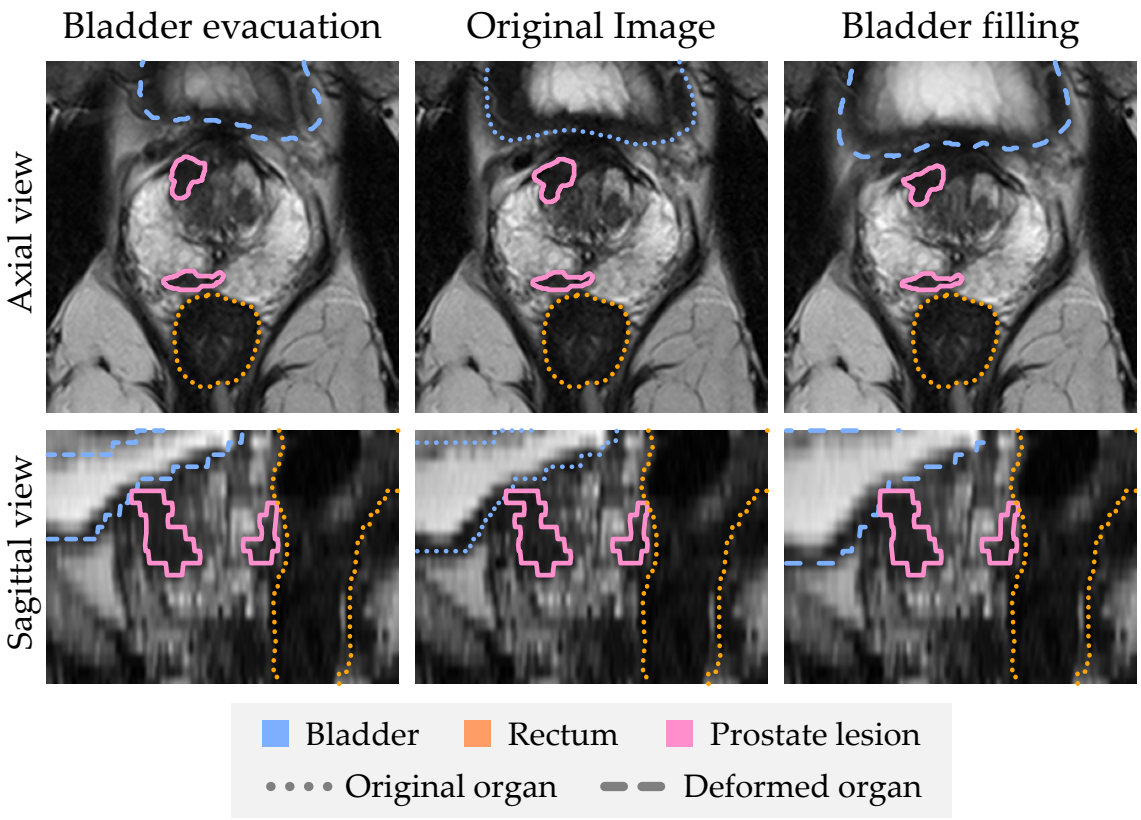

Figure 3.1: Example of anatomy-informed transformation simulating bladder shape changes on a prostate MRI scan. The central column shows the original image, while the left and right columns depict simulated bladder evacuation and filling, respectively. Malignant lesion contours (magenta), rectum (orange), and bladder (blue) boundaries are shown in axial and sagittal views. Dotted lines represent organ boundaries before deformation, while dashed lines indicate the simulated deformed shape. The simulated deformation induces localized tissue shifts that alter lesion morphology in the anterior transition zone (TZ) adjacent to the bladder. Figure adapted from our previously published work (Kovacs et al., 2023a), with permission from the publisher Springer Nature.

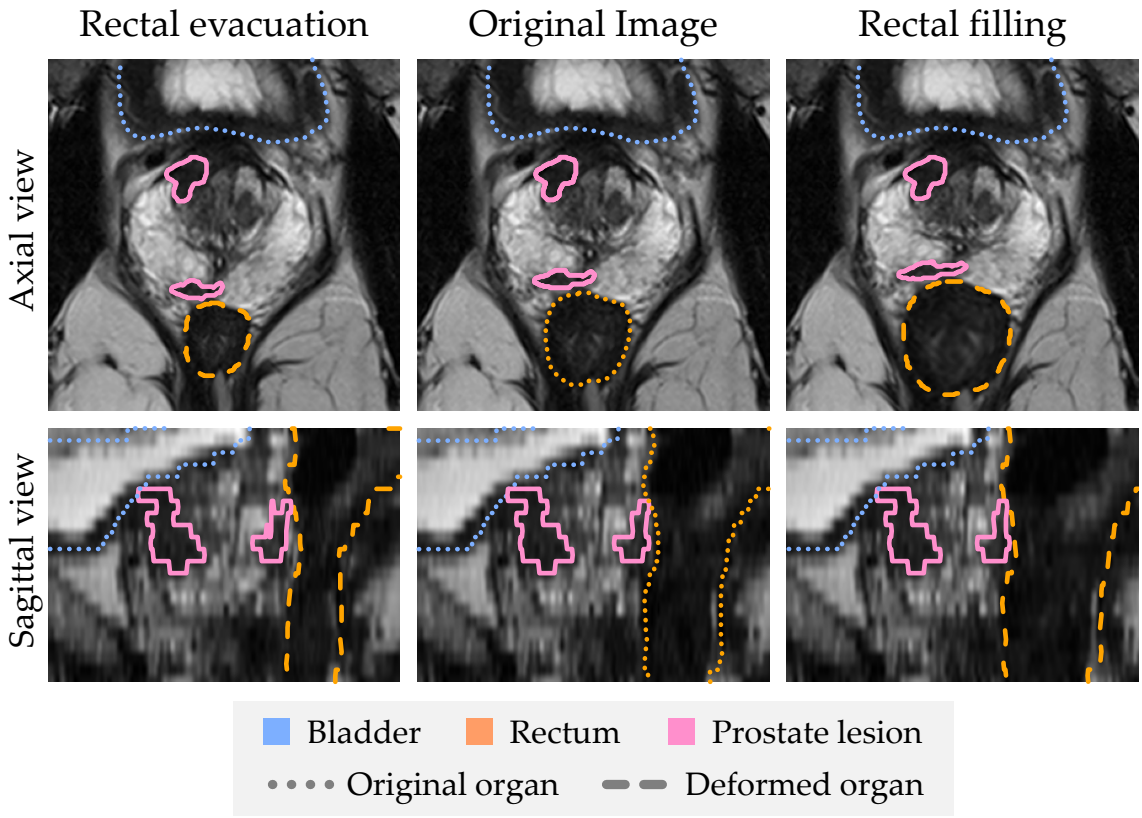

Figure 3.2: Example of anatomy-informed transformation simulating rectal shape changes on a prostate MRI scan. The central column shows the original image, while the left and right columns depict simulated rectal evacuation and filling, respectively. Malignant lesion contours (magenta), rectum (orange), and bladder (blue) boundaries are shown in axial and sagittal views. Dotted lines represent organ boundaries before deformation, while dashed lines indicate the simulated deformed shape. The simulated deformation induces localized tissue shifts that alter lesion morphology in the posterior peripheral zone (PZ) adjacent to the rectum. Figure adapted from our previously published work (Kovacs et al., 2023a), with permission from the publisher Springer Nature.

During the Turing test, 92 % of rectal and 93 % of bladder deformations produced by the anatomy-informed transformation were perceived as real by a freshly graduated clinician, demonstrating high visual plausibility. Resident radiologists with 1.5 - 3 years of experience in prostate MRI also classified 70 % of the bladder transformations as original. In contrast, they correctly identified 87.5 % of rectal deformations as synthetic, often citing subtle artifacts – not necessarily within the prostate itself (e.g., halo effects) – or relying on expert intuition. One participant commented: "It looks completely original, but something tells me it's artificial."

A summary of the results is shown in Tab. 3.1. Importantly, no original scan was consistently misclassified as synthetic by all participants, so no confusion matrix is reported. In contrast to the realism of the anatomy-informed deformations, random elastic deformations introduced obvious inconsistencies, making them easily identifiable in all cases.

Table 3.1: Perceived realism of anatomy-informed deformations during the Turing test. Percentage of the anatomy-informed rectal and bladder deformations classified as original by participants, based on visual assessment of the entire image, at two different levels of radiological expertise. Higher values reflect greater visual plausibility of the synthetic transformations. This table was constructed using the results previously published in Kovacs et al. (2023a), with permission from the publisher Springer Nature.

	Artificially deformed organ	
Level of Expertize	rectum	bladder
Freshly graduated	92 %	93 %
Radiologist residents	12.5 %	70 %

3.1.2 High Applicability

Although the anatomy-informed transformation includes additional steps – specifically blurring and gradient-based displacement computation – it builds on the same deformation vector field calculation used for standard spatial augmentations (e.g., rotation) in the Batchgenerators augmentation framework. As a result, the total transformation time remained in the µs range, ensuring that the augmentation pipeline remains lightweight and scalable. Despite this added complexity, the transformation introduced no observable overhead to training time. The GPU remained the primary bottleneck throughout training, indicating that the CPU was able to compute the anatomy-informed transformations in parallel without delay.

3.1.3 Approaching Towards Radiologists' Performance - Quantitative Results

Patient-level performance

Patient-level diagnostic performance across different augmentation strategies was quantitatively assessed using receiver operating characteristic (ROC) curves. The results are shown in Fig. 3.3.

To emphasize clinical relevance, model performance is compared to radiologists at the PI-RADS ≥ 4 threshold – a key decision point in prostate cancer diagnosis – as highlighted in Fig. 3.3. Anatomy-informed augmentation improves model sensitivity near this clinically meaningful operating point, narrowing the gap between artificial intelligence (AI) model and expert readers. By contrast, random elastic deformations degrade performance, shifting the ROC curve away from this region. The benefit of training with realistic soft tissue deformations is further supported by increases in both partial area under the receiver operating characteristic curve (pAUROC) and F_1-score, computed at the PI-RADS ≥ 4 diagnostic threshold. These results are summarized in Tab. 3.2.

Table 3.2: Patient-level diagnostic performance of nnU-Net models trained with different data augmentation schemes. Metrics are computed using the clinically relevant working point corresponding to the radiologists' sensitivity for PI-RADS ≥ 4. Anatomy-informed augmentations (rectum, and rectum & bladder) achieve the highest pAUROC and F_1-scores, outperforming the baseline strategy. In contrast, incorporating random elastic deformations into training results in a decline in patient-level performance. This table was constructed using the results previously published in Kovacs et al. (2023a), with permission from the publisher Springer Nature.

Augmentation scheme	pAUROC	F_1-score
1. basic (reference)	$44.33 \pm 11.65\,\%$	$57.31 \pm 3.14\,\%$
2. basic + random elastic	$38.94 \pm 14.38\,\%$	$56.98 \pm 3.08\,\%$
3.a) basic + Anatomy-inf. (rectum)	$59.92 \pm 13.27\,\%$	$61.64 \pm 3.61\,\%$
3.b) basic + Anatomy-inf. (rectum&bladder)	$53.27 \pm 13.42\,\%$	$62.42 \pm 3.84\,\%$
Radiologists' PI-RADS ≥ 4	–	64.44

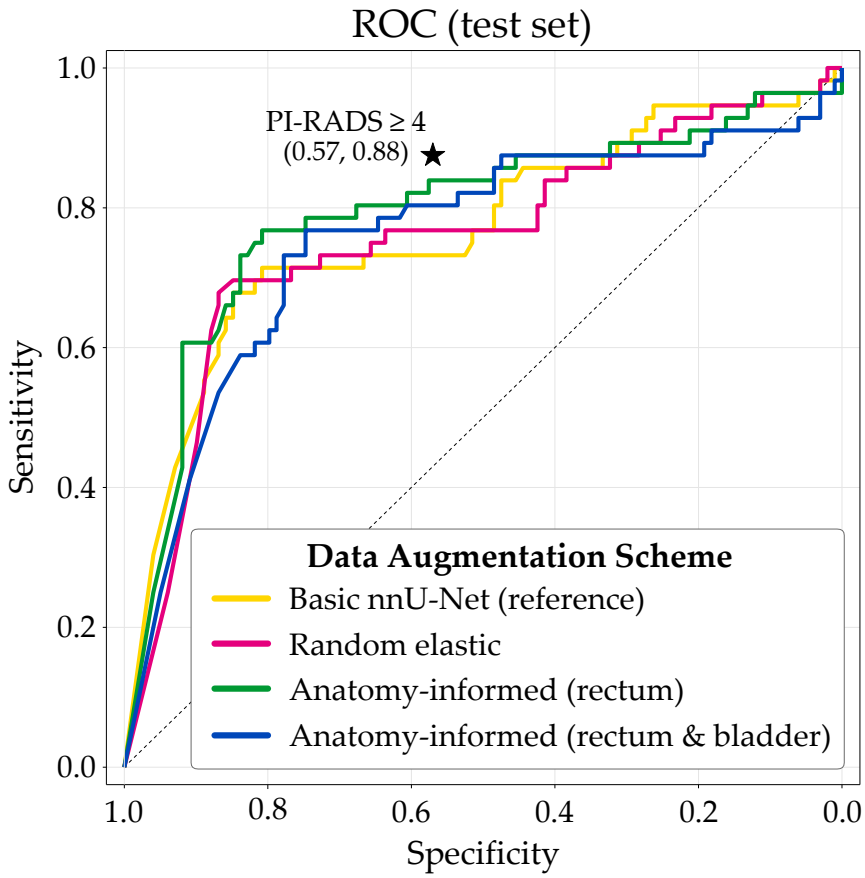

Figure 3.3: Patient-level receiver operating characteristic (ROC) curves comparing model performance across augmentation strategies on the independent test set. The diagnostic performance of radiologists at the clinical decision threshold of PI-RADS ≥ 4 is marked to indicate the most clinically informative reference point. Anatomy-informed augmentation improves model sensitivity near this decision boundary, effectively narrowing the performance gap between AI model and expert readers. In contrast, random elastic deformations degrade performance in this region, shifting the ROC curve away from the clinically relevant operating point. Figure adapted from our previously published work (Kovacs et al., 2023a), with permission from the publisher Springer Nature.

Extending the basic augmentation scheme with the proposed organ deformation significantly improved patient-level F_1-scores for both rectum-only (61.64, $p < 0.01$) and combined rectum & bladder deformations (62.42, $p < 0.01$), compared to the baseline strategy (57.31). Adding bladder deformation to rectum alone yielded a slight, non-significant improvement ($p = 0.31$). When calibrating sensitivity to match the radiologists' operating point (PI-RADS ≥ 4), the anatomy-informed DA strategies led to consistent gains in specificity. In contrast, random elastic deformations resulted in a drop in specificity, highlighting their limited practical benefit. The comparison of specificity values is summarized in Tab. 3.3.

Table 3.3: Specificity of nnU-Net models trained with different augmentation strategies, evaluated at a fixed sensitivity of 0.875 – corresponding to the radiologists' diagnostic threshold for PI-RADS ≥ 4. Anatomy-informed transformations improve specificity relative to the baseline, while random elastic augmentation slightly reduces it.

Augmentation scheme	Specificity
1. basic (reference)	0.333
2. basic + random elastic	0.323
3.a) basic + Anatomy-inf. (rectum)	0.455
3.b) basic + Anatomy-inf. (rectum&bladder)	0.475
Radiologists' PI-RADS ≥ 4	0.525

Lesion-level performance

Lesion-level diagnostic performance was evaluated using free-response receiver operating characteristic (FROC) analysis. In addition, the total number of correctly detected lesions was computed at the clinically relevant operating point corresponding to the radiologists' average false positives per scan (avgFPs/scan = 0.32) at the PI-RADS ≥ 4 threshold. The results are summarized in Tab. 3.4.

Table 3.4: Lesion-level performance of nnU-Net models trained with different augmentation strategies. Results are reported as FROC scores and the total number of correctly detected lesions. The number of detected lesions is computed at the clinically relevant working point corresponding to the radiologists' average false positives per scan (avgFPs/scan = 0.32) at the PI-RADS ≥ 4 threshold, on the independent test set consisting of 76 clinically significant prostate cancer lesions. Both random elastic and anatomy-informed deformations (rectum, rectum & bladder) led to improvements in lesion-level detection compared to the baseline. This table was constructed using the results previously published in Kovacs et al. (2023a), with permission from the publisher Springer Nature.

Augmentation scheme	FROC	Detected lesions
1. basic (reference)	58.14 ± 5.79 %	41
2. basic + random elastic	58.63 ± 5.42 %	45
3.a) basic + Anatomy-inf. (rectum)	59.55 ± 5.97 %	45
3.b) basic + Anatomy-inf. (rectum&bladder)	59.93 ± 5.53 %	46

Models trained with anatomy-informed augmentations showed the highest lesion detection rates. The augmentation strategy involving both rectum and bladder deformation achieved the best FROC score (59.93 ± 5.53%) and detected the most lesions (46 out of 76). Random elastic deformation also led to an increase in detected lesions (45), comparable to the proposed rectum-only strategy, but lower improvement in FROC score over the baseline.

Zonal performance

The impact of localized morphological variation in adjacent organs on zone-specific csPCa detection was evaluated by counting the number of detected lesions per prostate zone. The highest detection sensitivity was consistently achieved when the corresponding adjacent soft tissue was deformed during training: bladder for the TZ and rectum for the PZ. The results are summarized in Tab. 3.5.

Table 3.5: Number of detected clinically significant prostate cancer lesions stratified by prostate zone and augmentation strategy. Augmentation with bladder deformation improved detection in the transition zone (TZ), while rectal deformation enhanced sensitivity in the peripheral zone (PZ), highlighting the local effect of anatomy-informed transformations. Results are reported at the radiologist-level performance threshold (PI-RADS ≥ 4). Part of this table was constructed using the results previously published in Kovacs et al. (2023a) with permission from the publisher Springer Nature.

	Prostate zone		
Augmentation Strategy	Transition zone (18 TZ + 2 multi)	Peripheral zone (56 PZ + 2 multi)	Whole gland (76)
Basic (reference)	10	33	41
Anatomy-informed			
– rectum	11	**36**	45
– bladder	**12**	35	45
– rectum & bladder	**12**	**36**	**46**

In the TZ, augmentation strategies including bladder deformation (bladder only, or bladder & rectum) improved lesion detection by 2 cases compared to the baseline. This led to a statistically significant sensitivity increase of 10 % ($p < 0.01$). Similarly, in the PZ, applying rectal deformation (rectum only, or rectum & bladder) led to the detection of 3 additional lesions, corresponding to a 5.2 % sensitivity increase ($p < 0.01$). Importantly, rectal deformation did not significantly improve detection in the TZ, and bladder deformation resulted in a smaller, albeit significant, gain in PZ sensitivity.

Summary

At the selected patient- and lesion-level operating points, the model trained with the proposed anatomy-informed augmentation (rectum and bladder) achieved the best overall performance. It significantly outperformed the baseline model ($p < 0.01$), improving the F_1-score by 5.11 % and detecting 4 additional lesions (5.3 %) out of 76 in the independent test set. While both random elastic and anatomy-informed deformations (rectum, rectum & bladder) led to significant improvements in lesion-level detection, only the anatomy-informed strategies provided consistent and statistically significant gains at the patient-level. In contrast, random elastic deformations degraded patient-level performance despite their positive effect on lesion sensitivity.

3.2 Multi-Modal Misalignments in Prostate MRI (Objective #2)

The results presented in this section have been primarily published in the following journal article:

> **Balint Kovacs**, Nils Netzer, Michael Baumgartner, Adrian Schrader, Fabian Isensee, Cedric Weißer, Ivo Wolf, Magdalena Görtz, Paul F. Jaeger, Victoria Schütz, Ralf Floca, Regula Gnirs, Albrecht Stenzinger, Markus Hohenfellner, Heinz-Peter Schlemmer, David Bonekamp, Klaus H. Maier-Hein. **Nature Scientific Reports**. *Addressing image misalignments in multi-parametric prostate MRI for enhanced computer-aided diagnosis of prostate cancer*. https://doi.org/10.1038/s41598-023-46747-z

The results of the systematic analysis evaluating the impact of image co-registration and misalignment augmentation on an independent multicenter test set comprising 129 bi-parametric MRI (bpMRI) exams with biopsy-confirmed diagnoses are summarized. These findings highlight the importance of image registration for model performance and demonstrate the strong regularization effect introduced by misalignment augmentation. Moreover, they reveal the complementary benefits of combining both strategies. The detailed results are presented in the following subsections.

3.2.1 Maximized Robustness Achieved Through the Combination of Registration and Misalignment Augmentation

To determine the most robust configuration for handling multi-modal image alignment errors in prostate MRI, the effects of image registration and misalignment augmentation are systematically evaluated on the patient-level diagnostic performance, measured by the area under the receiver operating characteristic curve (AUROC). The results are summarized in Tab. 3.6.

Table 3.6: Patient-level area under the receiver operating characteristic curve (AUROC) results – with 95 % confidence intervals and associated p-values calculated using the DeLong test – on the independent test set. Both image registration and misalignment augmentation individually improved the AUROC compared to the unregistered baseline, although without reaching statistical significance. The combination of registration and misalignment augmentation achieved the highest AUROC value with statistically significant improvement over the baseline, indicating a complementary effect. Table adapted from our previously published work Kovacs et al. (2023b), reused under the Creative Commons Attribution 4.0 License.

AUROC test results w.r.t different strategies	Dataset without registration	Dataset with B-Spline registration
Default augmentations	75.93 % (67.49–83.68 %, reference)	79.11 % (70.95–86.93 %, $p = 0.31$)
Default augmentations + misalignment augm.	80.13 % (71.57–87.18 %, $p = 0.11$)	82.07 % (74.18–89.38 %, $p = 0.02$)

Addressing multi-modal misalignments either through image registration or misalignment augmentation individually resulted in an increase in the AUROC compared to training on unregistered data. However, these individual improvements did not reach statistical significance, with p-values of $p = 0.31$ for registration alone and $p = 0.11$ for misalignment augmentation alone. In contrast, combining registration with misalignment augmentation achieved the highest AUROC value, with a statistically significant improvement ($p = 0.02$) over the baseline (unregistered dataset without augmentation). This finding indicates that the two strategies have complementary effects on the performance increase, and the combined strategy offers the most robust solution for mitigating multi-modal alignment errors.

In addition, a stratified analysis was conducted to assess the performance of the proposed methods separately on the two datasets included in this study (PROSTATEx and the in-house cohort). The AUROC values were consistent across both datasets, indicating that the effectiveness of the proposed approach generalizes well across different imaging centers. Both registration and misalignment augmentation contributed to performance improvements individually on each dataset, suggesting that the observed gains result from the methods themselves rather than from any dataset-specific effect. Detailed AUROC results for each dataset are presented in Tab. 3.7.

Table 3.7: Patient-level area under the receiver operating characteristic curve (AUROC) results on the independent test set stratified by dataset (PROSTATEx and In-house cohort). Performance is reported separately for the unregistered and registered datasets, with and without the use of misalignment augmentation. Results demonstrate consistent improvements across both datasets, indicating that the proposed methods generalize well across different imaging centers. Table adapted from our previously published work Kovacs et al. (2023b), reused under the Creative Commons Attribution 4.0 License.

AUROC (test set)		Unregistered dataset	Registered dataset
Augmentation scheme w/o misalignments	PROSTATEx	78.33 %	80.54 %
	In-house	75.49 %	77.43 %
Augmentation scheme with misalignments	PROSTATEx	85.59 %	86.70 %
	In-house	77.78 %	79.98 %

3.2.2 Achieving Radiologist-Level Diagnostic Performance by Addressing Multi-Modal Misalignments

To emphasize the practical value of addressing multi-modal misalignments, the diagnostic performance of the trained models and the radiologists is compared on the test cohort. ROC curves were computed for each model configuration, alongside the radiologists' diagnostic performance using Prostate Imaging Reporting and Data System (PI-RADS) ≥ 3 and PI-RADS ≥ 4 thresholds – both of which represent clinically important decision points.

Fig. 3.4 illustrates the impact of image registration, misalignment augmentation, and their combination on model performance. The radiologists' operating points are also indicated, with specificity and sensitivity values of $(0.21, 0.98)$ for PI-RADS ≥ 3 and $(0.56, 0.91)$ for PI-RADS ≥ 4, respectively.

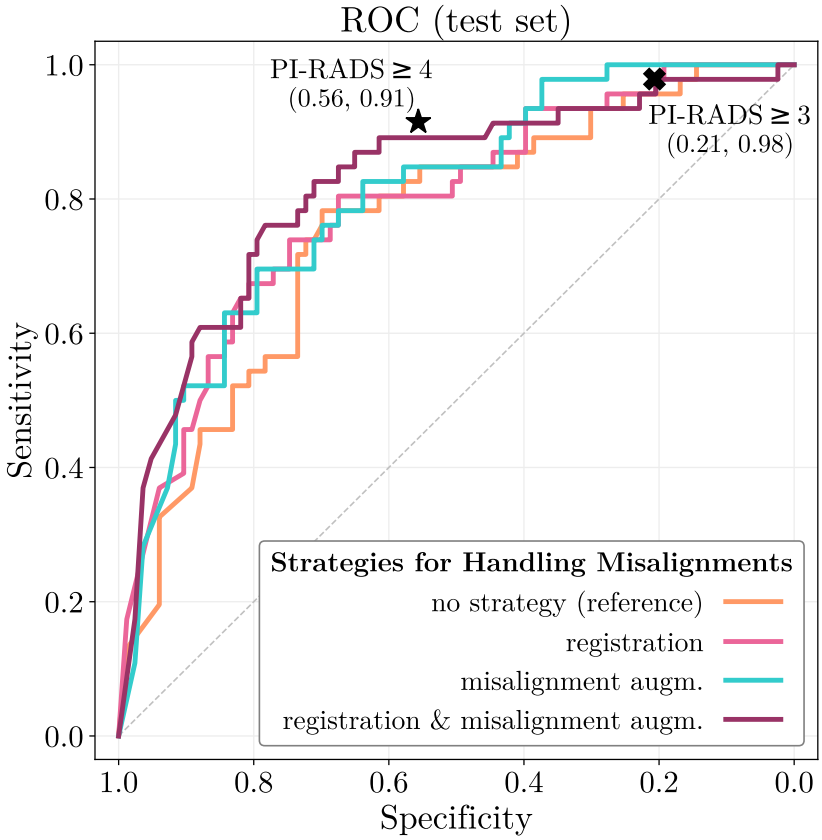

Figure 3.4: Predictive performance comparison of trained models and radiologists on the receiver operating characteristic (ROC) curve. The radiologists' diagnostic performance at the PI-RADS ≥ 3 and PI-RADS ≥ 4 thresholds is marked to highlight clinically important decision points. Applying either registration or misalignment augmentation individually improves model sensitivity compared to training on unregistered data, particularly at low and high specificity regions. However, only the combined use of registration and misalignment augmentation enables the model to closely match and partially exceed the radiologists' performance across the full range of clinically relevant operating points. Figure taken from our previously published work Kovacs et al. (2023b), reused under the Creative Commons Attribution 4.0 License.

Using either registration or misalignment augmentation individually improved sensitivity compared to the baseline model trained on unregistered data without augmentation, particularly in regions of low and high specificity. Registration alone enabled the model to closely match the radiologists' performance at the PI-RADS ≥ 3 threshold, while misalignment augmentation slightly exceeded it. However, neither approach individually enhanced sensitivity near the stricter PI-RADS ≥ 4 threshold. In contrast, combining registration with misalignment augmentation was the only strategy that closely matched both clinical PI-RADS performance points and achieved sensitivity gains across the widest range of specificity values.

3.2.3 Visualization of Detected Punctate Lesion

To qualitatively demonstrate the influence of registration, misalignment augmentation, and their combination on lesion segmentation, I present a visual comparison of the resulting model predictions. Fig. 3.5 displays clinical lesion ground truth (ground truth (GT)) annotations alongside the predicted probability maps for two highly distinct and clinically relevant lesion types: one larger lesion (with a PI-RADS score of 4 and Gleason Score (GS) 7a) located anteriorly in the TZ, and a small punctate lesion (also PI-RADS 4, GS 7a) located in the PZ at the left prostate base.

Both registration and misalignment augmentation led to a noticeable increase in the predicted lesion volumes, resulting in more complete coverage of the pathological regions. These improvements were particularly evident in the larger lesion, where the probability maps became more extensive and confident when either or both strategies were applied. Notably, the use of misalignment augmentation enabled the convolutional neural network (CNN) to more reliably detect the smaller punctate lesion, which was only marginally identified by models trained without this augmentation strategy.

Overall, this example highlights that addressing misalignments through augmentation strategies not only improves the segmentation confidence of prominent lesions but increases the sensitivity for detecting small punctate lesions too.

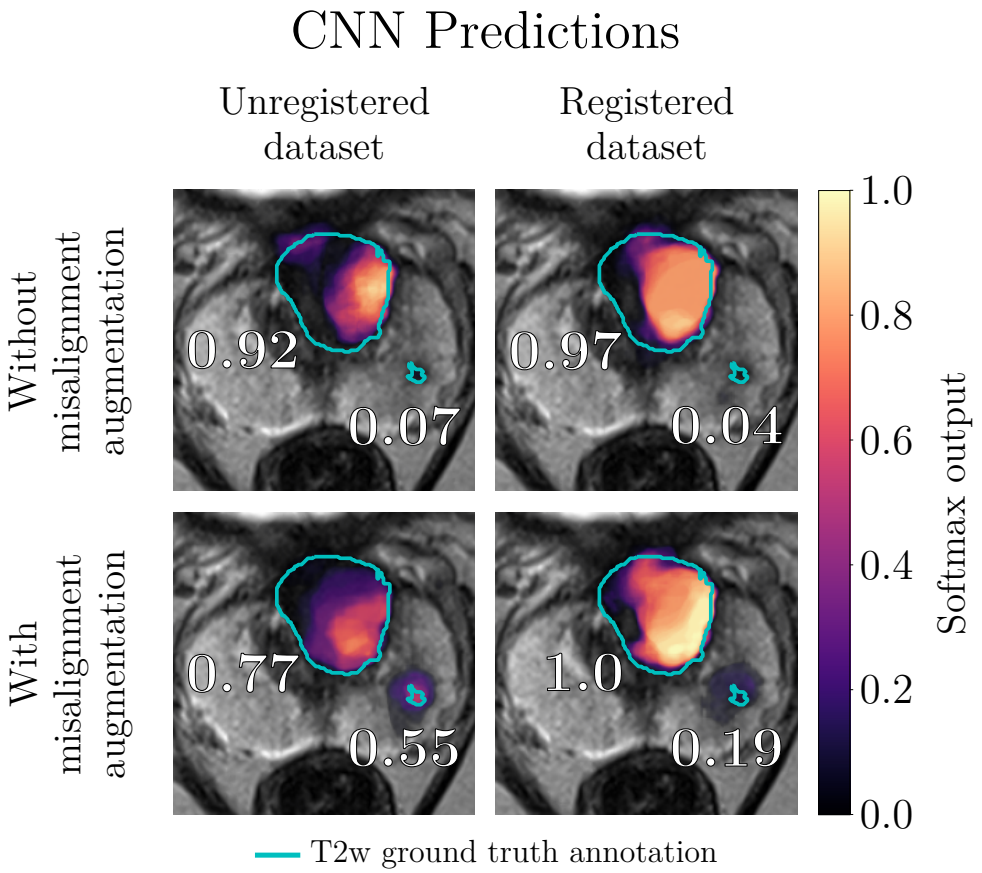

Figure 3.5: Visualization of manual annotations and segmentation predictions of the convolutional neural network (CNN) for two distinct and clinically relevant prostate cancer lesions: one large lesion (PI-RADS 4, GS 7a) located anteriorly in the TZ, and one small punctate lesion (PI-RADS 4, GS 7a) in the left prostate base. A representative axial T2-weighted (T2w) slice is shown, with manual delineations overlaid in cyan. Heatmaps depict the predicted softmax probabilities from different training configurations, using a color scale with gamma correction ($\gamma = 0.2$) for better visualization of low-probability regions. Top row: predictions without misalignment augmentation (unregistered and registered datasets). Bottom row: predictions with additional misalignment augmentation. All model settings correctly detected the large lesion, but the punctate lesion was only successfully highlighted in models trained with misalignment augmentation, emphasizing the clinical benefit of the proposed strategy. Figure taken from our previously published work Kovacs et al. (2023b), reused under the Creative Commons Attribution 4.0 License.

3.2.4 B-spline Registration on Par with Method Using Human Ground Truth Segmentation

To verify the quality and appropriateness of the B-spline registration method employed in the computer-aided diagnosis (CAD) pipeline (Dataset #2), I compared its performance against a reference registration based on ground truth matching (GT-matching, Dataset #3). This evaluation was performed using two complementary approaches: a lesion overlap metric and the impact on diagnostic performance assessed through ROC analysis.

Fig. 3.6 presents the probability density functions (PDFs) of the Dice similarity coefficient computed between manual lesion annotations in the T2-weighted (T2w) image and the apparent diffusion coefficient (ADC) map, evaluated across three settings: without registration (Dataset#1), with B-spline registration (Dataset#2), and with GT-matching registration ((Dataset#3)). Mean and standard deviation values for the Dice scores are also reported in the legend.

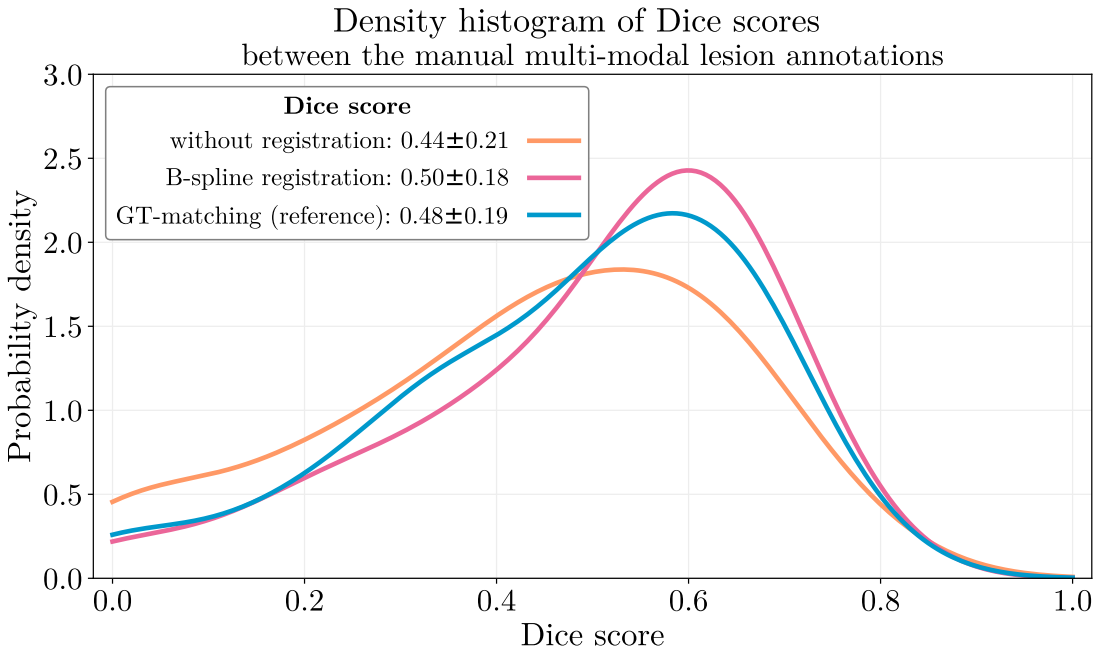

Figure 3.6: Probability density functions of Dice scores between manual lesion annotations in T2w images and ADC maps across different registration strategies. The B-spline registration achieves the highest mean Dice score, slightly outperforming the ground-truth-matching (GT-matching) reference method. Mean and standard deviation values are provided in the legend. Figure taken from our previously published work Kovacs et al. (2023b), reused under the Creative Commons Attribution 4.0 License.

Quantitatively, the mean Dice score of the unregistered dataset (0.44 ± 0.21) improved to 0.48 ± 0.19 with GT-matching registration and further to 0.50 ± 0.18 with the B-spline registration, indicating that the B-spline approach slightly outperforms the reference GT-matching strategy in terms of spatial alignment quality.

In addition to the overlap metric, the clinical relevance of registration quality was assesed by examining its influence on model diagnostic performance. Fig. 3.7 shows the ROC curves obtained under the three registration conditions.

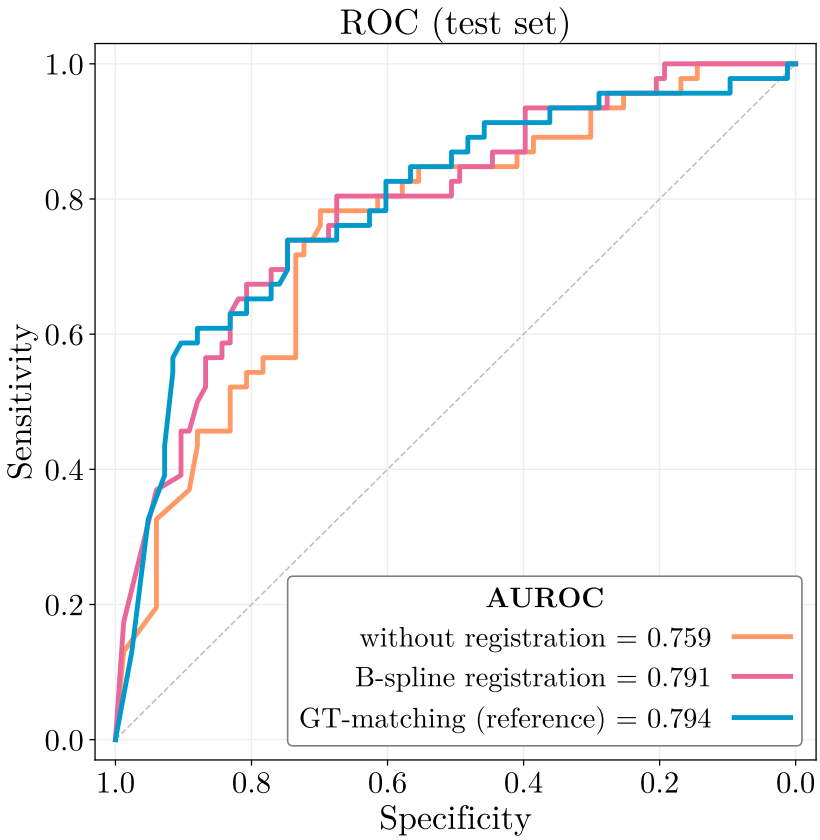

Figure 3.7: Impact of the different registration strategies on patient-level diagnostic performance (receiver operating characteristic (ROC) analysis). Both B-spline registration and ground-truth-matching (GT-matching) improved sensitivity across both low- and high-specificity regions compared to the unregistered dataset, leading to AUROC increases of 3.2 % and 3.5 %, respectively. Figure taken from our previously published work Kovacs et al. (2023b), reused under the Creative Commons Attribution 4.0 License.

Both registration approaches resulted in a notable improvement over the unregistered baseline: B-spline registration increased the AUROC by 3.2 %, while GT-matching achieved a slightly higher improvement of 3.5 %.

In summary, although the B-spline registration outperformed GT-matching in terms of Dice score distribution, this difference did not translate into a substantial advantage in AUROC. Both registration techniques were effective in improving the model's diagnostic performance relative to unregistered data.

Chapter 4

Discussion

In recent years, there has been growing recognition of the value of incorporating application-specific knowledge into artificial intelligence (AI) model training, a data-centric perspective that complements ongoing model-centric developments. This paradigm has yielded promising results across diverse tasks: incorporating anatomical priors, such as prostate zonal information for improved clinically significant PCa (csPCa) detection (Hosseinzadeh et al., 2021; Saha et al., 2021b); embedding clinically relevant factors into the loss function, such as topology-aware constraints for enhanced vessel and nerve segmentation (Shit et al., 2021; Kirchhoff et al., 2024) or lesion size-related risk term in breast cancer screening (Bounias et al., 2023). Additional innovations include the use of statistical shape and intensity models to generate anatomically plausible data augmentations, particularly in orthopedic imaging (Schmid et al., 2023).

Building on this concept, the present dissertation investigated data-centric training strategies to enhance AI-based diagnosis of prostate cancer using bi-parametric magnetic resonance imaging (MRI). The work specifically addressed two underexplored yet clinically relevant challenges:

1. accounting for large soft tissue deformations caused by physiological size variation in the rectum and bladder, and

2. handling multi-modal alignment errors that unavoidably arise between MRI sequences due to patient motion, differences in imaging contrast, or physiological changes.

The studies presented in this dissertation demonstrate how application-specific inductive bias via tailored data augmentation can guide AI models to be more robust to these real-world sources of variability. This approach not only improves diagnostic performance but also aligns model behavior more closely with that of expert radiologists, who naturally account for such contextual variations during clinical interpretation.

4.1 Soft Tissue Deformations of the Prostate (Objective #1)

The discussion presented in this section has been primarily published in the following conference:

Balint Kovacs, Nils Netzer, Michael Baumgartner, Carolin Eith, Dimitrios Bounias, Clara Meinzer, Paul F. Jäger, Kevin S. Zhang, Ralf Floca, Adrian Schrader, Fabian Isensee, Regula Gnirs, Magdalena Görtz, Viktoria Schütz, Albrecht Stenzinger, Markus Hohenfellner, Heinz-Peter Schlemmer, Ivo Wolf, David Bonekamp, Klaus H. Maier-Hein *Anatomy-informed data augmentation for enhanced prostate cancer detection*. **International Conference on Medical Image Computing and Computer-Assisted Intervention - MICCAI 2023**. https://doi.org/10.1007/978-3-031-43990-2_50

This study highlights the importance of accounting for variations in the functional state of adjacent organs during AI model training to enhance prostate cancer (PCa) diagnosis. To address this, I introduced the anatomy-informed transformation, a lightweight mathematical model for rectum and bladder deformations developed based on the biomechanical properties of the pelvic region. This model was integrated into the training pipeline of nnU-Net as an online data augmentation strategy to increase variability in the shape of lesions, the prostate, and adjacent organs.

To evaluate its effectiveness, I compared various augmentation strategies by assessing patient- and lesion-level PCa detection performance derived from semantic segmentation of malignant lesions. Given the strong dependence of segmentation tasks on spatial information, this setting provided an ideal test environment for evaluating the anatomy-informed transformation, specifically designed to model localized soft tissue deformations of the prostate.

4.1.1 Diagnostic Benefit of Simulating Physiological Deformations

Extending the standard yet extensive augmentation pipeline of nnU-Net, a widely adopted baseline in biomedical image segmentation, with the proposed anatomy-informed transformation led to improved diagnostic performance in PCa detection. By simulating physiologically realistic rectal and bladder size variations during model training, the model's sensitivity approached that of radiologists at the clinically relevant decision threshold of PI-RADS ≥ 4. These improvements were reflected in statistically significant gains ($p < 0.01$) at both the patient- and lesion-level: a 4.3 % increase in patient-level F_1-score when deforming only the rectum, and a 5.1 % increase when deforming both

organs; as well as a 5.3 % boost in lesion detection sensitivity (4 additional lesions) with rectal deformation, and 6.6 % (5 additional lesions) when both organs were deformed. These results were evaluated at the radiologists' clinical decision threshold of PI-RADS ≥ 4 for diagnostic sensitivity of 87.5 % and an average of 0.32 average number of false positives per scan (avgFPs/scan).

A potential explanation for this improvement lies in the fact that the anatomy-informed transformation – even if it is a single transformation – introduces meaningful morphological diversity into the training data. Soft tissue deformations, such as those caused by rectal distension or bladder filling, naturally occur due to physiological processes, but in clinical practice, only a single static snapshot can be captured per exam. The proposed augmentation strategy effectively enriches the training data with simulated physiological states that could have occurred at the same imaging time point, thereby improving the generalization ability and robustness of the model.

The best results were achieved when both the rectum and bladder were deformed during training, although this led to only a marginal improvement over rectal deformation alone. This modest gain can be attributed to the anatomical distribution of csPCa lesions: the majority are located in the peripheral prostate zone (PZ), which lies adjacent to the rectum. Therefore, rectal deformation alone likely introduces sufficient morphological variability for the majority of lesion locations. This aspect is discussed in more detail in the following subsection.

Answer for Research Question 2: Increased lesion morphological diversity led to enhanced diagnostic model performance.

4.1.2 Localized Performance Gains from Targeted Organ Deformations

Stratifying lesion detection performance by prostate zones revealed localized performance improvements resulting from organ-specific soft-tissue deformations during training. Detection sensitivity increased consistently in the prostate zones adjacent to the deformed organ: bladder deformations improved detection in the transitional prostate zone (TZ) (+10 % with 2 additional lesions for both bladder-only and bladder&rectum settings), while rectal deformations yielded the highest gains in the PZ (+5.2 % with 3 additional lesions for both rectum-only and bladder&rectum settings), both outperforming the standard augmentation scheme.

Interestingly, even bladder-only deformation led to improved detection in the PZ, albeit to a lesser extent than rectal deformation. While initially counterintuitive, this may be attributed to the anterior extension of the PZ, which brings parts of it into proximity with the bladder.

These findings support that the proposed anatomy-informed transformation operates in a spatially targeted manner, in contrast to traditional global transformations that lack anatomical specificity. This enables not only overall performance gains, but also selective improvement in detection sensitivity for anatomically localized regions of interest.

4.1.3 Importance of Visual Realism and Label Preservation

The anatomy-informed transformation was designed to simulate anatomically plausible deformations while preserving essential image features relevant for PCa diagnosis.

Most of the transformed MRI images successfully passed the Turing test when evaluated by a freshly graduated clinician with 3 years of expertise in prostate MRI, with 92 % of artificial rectum and 93 % of artificial bladder transformations perceived as authentic. Notably, some of these synthetic scans also passed the test when assessed by more experienced radiology residents: 12.5 % of the rectal and 70 % of the bladder deformations were classified to be original. The difference in perceived realism between rectum and bladder transformations may be attributed to their respective deformation amplitudes ($C_{rectum} = 1200$, $C_{bladder} = 600$), as more pronounced deformations are more likely to produce subtle visual artifacts. Furthermore, the phrasing of the evaluation question – "Is this image an original MRI scan or has it been artificially altered? If you believe it has been modified, please describe why." – prompted participants to assess the realism of the entire image. In some cases, the artificial nature of the transformation was identified based on visual clues located outside the prostate region, which is not directly relevant to the AI training objective. A more appropriate phrasing aligned with the training focus would have been: "Is the prostate in this image original or has it been artificially altered?" Nonetheless, participants often reported that the altered images were difficult to distinguish from real scans. In one example, a resident radiologist described an image as "visually perfect" yet correctly classified it as artificial, relying on their intuition based on expert domain knowledge.

To further underscore the importance of the label-preserving property of augmentations, I compared the performance of the anatomy-informed augmentation against both the standard and the random deformable augmentation strategies. While the random deformable transformations increased lesion shape variability – leading to a higher number of detected lesions – this came at the cost of decreased patient-level diagnostic performance. Although this result may seem counterintuitive, it aligns with the morphological characteristics of PCa. Given the amorphous nature of csPCa lesions, random elastic deformations may inadvertently transform benign prostatic hyperplasia (BPH) cases into malignancy-like patterns, resulting in implausible or even harmful warping that distorts key features and introduces misleading training examples. In contrast, the

anatomy-informed augmentation strategy not only preserved anatomical plausibility but also led to consistent improvements on both the patient and lesion level. These findings emphasize the importance of clinically meaningful augmentation design for achieving robust improvements in AI-based PCa diagnosis. A potential future direction to strengthen this conclusion would be the inclusion of BPH lesion segmentation, enabling a more detailed analysis of how augmentations impact benign versus malignant cases.

Answer for Research Question 3: The realism of the data augmentation is crucial due to the need for label-preservation.

4.1.4 High Applicability with Practical Considerations

While complex biomechanical models of the prostate have been explored for various applications (Hu et al., 2008, 2010; Khallaghi et al., 2015a,b; Qasim et al., 2022), their high computational demands have so far prevented their adoption in real-time data augmentation (DA) during training. To the best of my knowledge, the method proposed in this dissertation is the first to effectively leverage such modeling for real-time augmentation pipelines for deep learning. Its easy integration into standard DA frameworks and no increase in training time make the anatomy-informed strategy both practical and scalable.

One limitation of the approach is its reliance on the existence of segmentation masks, which introduces an additional annotation burden. However, recent advances in automatic segmentation – such as the availability of robust pre-trained models for organ segmentations like TotalSegmentator (Wasserthal et al., 2023; Akinci D'Antonoli et al., 2025) built on nnU-Net (Isensee et al., 2021) – can significantly reduce this manual effort.

A further consideration involves the parameterization of the transformation. Since the deformation is derived from the gradient field of a blurred segmentation mask, the amplitude parameter C remains a relative scalar. Its effect depends on both the Gaussian kernel size and the image resolution, making direct transferability across datasets not straightforward. For new applications, visually tuning these parameters is still necessary to ensure plausible deformations, introducing an additional effort when adapting the method to other domains.

Answer for Research Question 1: By simplifying complex biomechanical models, it is possible to construct a lightweight transformation that produces realistic soft tissue deformations and is suitable for online DA.

4.1.5 Extending Anatomy-Informed Transformations Beyond Organs

The proposed anatomy-informed transformation was originally named for its initial application in deforming organs based on anatomical segmentations. However, its applicability extends beyond organs and can be generalized to any structure for which segmentation masks are available.

In the study by Rokuss et al. (2025), I successfully adapted this transformation to simulate stochastic lesion growth. Unlike organs, which exhibit relatively consistent sizes across patients, lesions present a substantial variability in size. Therefore, applying the original transformation with fixed parameters would have been suboptimal, as it would either underrepresent growth in large lesions or overamplify changes in smaller ones. To address this, the transformation is applied iteratively on each lesion using smaller deformation amplitudes. Furthermore, to simulate stochastic progression, the deformation field was modulated by a smoothed random vector field. An illustrative example demonstrating both uniform and random lesion growth is shown in Fig. 4.1.

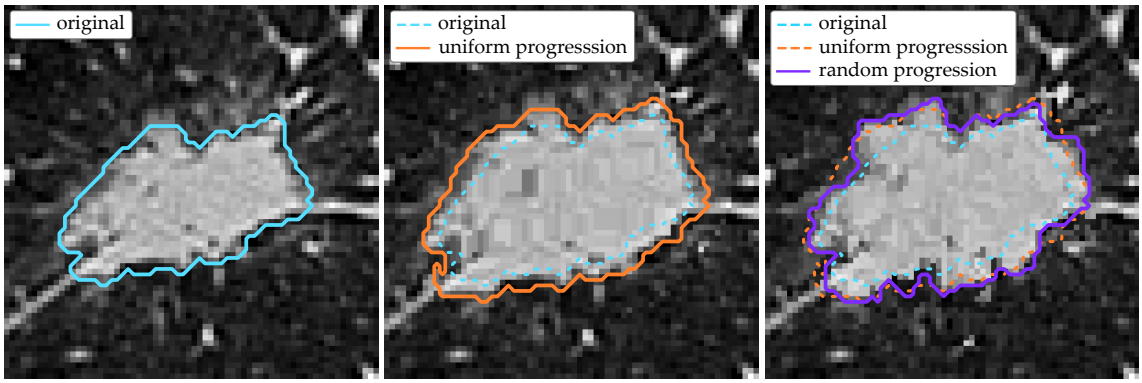

Figure 4.1: Lesion growth simulation using the anatomy-informed transformation applied to a lesion from the RIDER Lung CT dataset (Zhao et al., 2015). Both ■ uniform and ■ random simulated lesion growths are shown alongside the ■ original lesion outline for visual comparison.

The lesion tracking network proposed by Rokuss et al. (2025) exhibited performance improvements across all evaluated metrics when trained with the generated lesion shape variations. It is important to emphasize that the anatomy-informed transformation was one of the several contributing components in this framework including the integration of prior timepoint segmentations and the use of various visual prompts.

4.1.6 Bridging the Gap in Spatial Transformations for Online Data Augmentation

Based on the experimental results, the proposed anatomy-informed transformation can be positioned within the broader landscape of spatial transformations commonly used in deep learning-based medical image analysis.

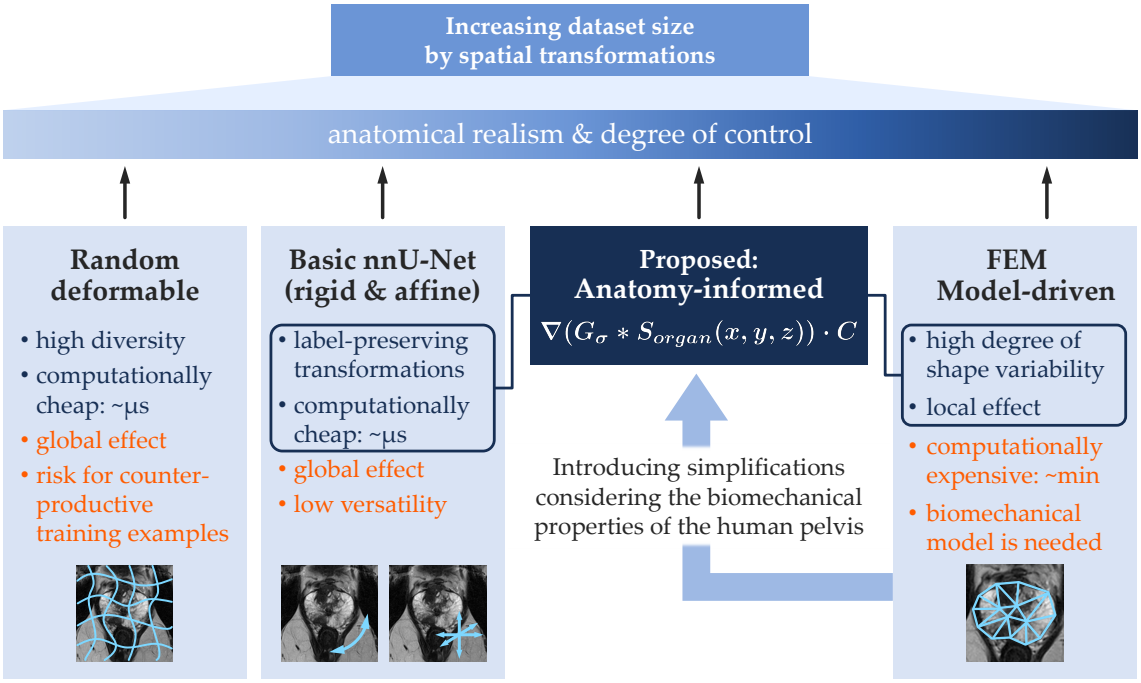

Figure 4.2: Comparison of commonly used spatial transformation strategies in terms of anatomical realism, degree of control, and their ability to introduce lesion shape variability. The proposed anatomy-informed transformation bridges the gap between simplistic methods (rigid/affine, random deformable) globally affecting the image, and complex biomechanical finite element modeling (FEM)-based models. It introduces anatomically plausible, localized deformations that significantly enhance lesion shape variability while preserving pathological training labels, a critical property for AI model training. With minimal computational overhead, it enables scalable, online data augmentation uniquely suited for medical imaging tasks. The MRI images are taken from our previously published work (Kovacs et al., 2023a), with permission from the publisher Springer Nature.

As illustrated in Fig. 4.2, this method bridges the gap between low-complexity global transformations and high-fidelity, physics-based deformable models.

- **Increased lesion shape variability with label-preserving capability:** The transformation substantially increases lesion shape variability by simulating plausible soft tissue deformations that reflect natural anatomical differences across patients, similar to those modeled by FEM-based approaches. Unlike random deformable transformations, which risk introducing label noise through unrealistic distortions, the anatomy-informed method maintains anatomical consistency and lesion integrity – a critical requirement for safe and effective data augmentation in AI-based model training.
- **Mid-level complexity:** The method offers a practical trade-off between traditional affine transformations (which require no domain information priors) and high-complexity finite element (FE) models. It leverages organ segmentations to construct deformation fields, avoiding the need for complex biomechanical simulation while still introducing meaningful variability.
- **Efficient computation:** Although the transformation includes additional steps – such as Gaussian blurring and gradient field calculation – it remains computationally efficient, on par with conventional augmentations – like affine or random deformable methods – and far faster than FEM-based transformations, which typically require several minutes.
- **Localized effect:** Unlike traditional augmentations that apply transformations uniformly across the entire image, the anatomy-informed method introduces spatially localized changes. These are constrained to the region surrounding the organ of interest, enabling anatomically targeted data augmentation that preserves the global spatial context.

This unique balance of anatomical fidelity, computational efficiency, and targeted augmentation makes the proposed method a novel and practical addition to the spectrum of online spatial data augmentation techniques used in deep learning pipelines.

4.2 Multi-Modal Misalignments in Prostate MRI (Objective #2)

The discussion presented in this section has been primarily published in the following journal article:

Balint Kovacs, Nils Netzer, Michael Baumgartner, Adrian Schrader, Fabian Isensee, Cedric Weißer, Ivo Wolf, Magdalena Görtz, Paul F. Jaeger, Victoria Schütz, Ralf Floca, Regula Gnirs, Albrecht Stenzinger, Markus Hohenfellner, Heinz-Peter Schlemmer, David Bonekamp, Klaus H Maier-Hein. **Nature Scientific Reports**. *Addressing image misalignments in multi-parametric prostate MRI for enhanced computer-aided diagnosis of prostate cancer*. https://doi.org/10.1038/s41598-023-46747-z

This study systematically investigated multiple strategies to address multi-modal misalignments in PCa diagnosis based on bi-parametric MRI (bpMRI). Rather than relying on surrogate alignment or image similarity metrics, the evaluation is based directly on the performance of a clinically meaningful downstream task – patient-level PCa diagnosis derived from a computer-aided diagnosis (CAD) system. As the clinical diagnosis was obtained through semantic segmentation of malignant lesions, a task inherently dependent on spatial consistency across modalities, the experimental setup was particularly suited for assessing the impact of the applied methods. I evaluated two complementary strategies with distinct objectives:

- the application of B-spline image registration to spatially align different MRI sequences and match ground-truth segmentations across modalities, and
- the introduction of misalignment augmentation, a novel data augmentation technique designed to simulate plausible spatial inconsistencies during training to improve model robustness.

The systematic analysis across an independent multi-centric test set comprising 129 biopsy-confirmed bpMRI exams demonstrated that both strategies contributed to performance improvements. Moreover, I highlighted the complementary nature of combining registration with misalignment augmentation, exceeding the effects of either strategy alone. This discussion reflects on the clinical relevance, applicability, and broader implications of these findings.

4.2.1 The Importance of Ground Truth Consistency Across Image Modalities for Diagnostic Performance

Aligning the prostate MRI image modalities through a B-Spline registration resulted in an improvement in patient-level diagnostic performance, as reflected by the increase in area under the receiver operating characteristic curve (AUROC) compared to the unregistered dataset. These findings suggest that enhancing anatomical correspondence across modalities enables the CAD system to better fuse complementary imaging contrast information, ultimately leading to more effective clinical decision-making.

Importantly, although improvements were observed, the differences in AUROC values across the used registration strategies did not fully reflect the more apparent differences seen in the lesion-wise Dice score distributions. This underscores an important limitation of relying on surrogate registration quality metrics, such as Dice scores, as they do not necessarily translate proportionally into improvements in downstream diagnostic performance – consistent with prior observations regarding the limitations of surrogate metrics for assessing registration quality (Rohlfing, 2011). These findings therefore emphasize the importance of evaluating registration strategies directly on clinical tasks, such as patient-level PCa diagnosis, rather than relying solely on surrogate alignment measures.

Despite the observed positive trend, the performance improvement achieved by registering the MRI sequences did not reach statistical significance. This observation may explain why a vast number of studies have employed registration as a standard preprocessing step (Aldoj et al., 2020; Arif et al., 2020; Cao et al., 2019; De Vente et al., 2020; Kohl et al., 2017; Winkel et al., 2021; Sanyal et al., 2020; Schelb et al., 2019; Yang et al., 2017; Netzer et al., 2021, 2023; Pellicer-Valero et al., 2022). While it indicates that registration systematically provides a meaningful benefit, it alone may not be sufficient to fully correct for all sources of misalignment in prostate MRI, potentially explaining why previous studies did not report baseline comparisons without registration.

Answer for Research Question 4: Registration improves model performance and its quality should be evaluated on the clinical downstream task.

4.2.2 Enhanced Robustness Against Alignment Errors Gained Through Model Training

Introducing artificial misalignments between modalities during convolutional neural network (CNN) training made the diagnostic task more challenging, encouraging the network to develop greater invariance to spatial inconsistencies between the input MRI modalities and the ground truth annotations. The results showed that applying misalign-

ment augmentation to unregistered data could partially compensate for the lack of explicit registration due to its strong regularization effect. This suggests that misalignment augmentation can be used as a highly applicable alternative strategy for addressing alignment errors, thereby reducing the reliance on complex non-rigid registration techniques.

Nonetheless, it is important to highlight that the robustness gained by misalignment augmentation is not without limitations. Its effectiveness likely depends on the degree of initial anatomical alignment between modalities, which exhibited an average lesion segmentation overlap of approximately 44 % in the cohort, as well as on the characteristics (type, amplitude, and probability) of the artificially introduced misalignments. Thus, careful parameter tuning is essential to adapt the augmentation strategy to the specific needs of a given clinical application. Moreover, an initial correction step, such as a simple affine registration to eliminate large spatial offsets, likely remains necessary in cases where the input modalities exhibit negligible anatomical overlap. Despite the observed performance improvements, the gains achieved through misalignment augmentation alone did not reach statistical significance, consistent with the findings for explicit registration.

Answer for Research Question 5: Misalignment data augmentation serves as a lightweight alternative to complex registration algorithms.

4.2.3 Synergistic Gains Through the Combination of Registration and Misalignment Augmentation

Although both image registration and misalignment augmentation independently address the same underlying issue of multi-modal misalignments, their combination led to a further improvement in patient-level diagnostic performance. Specifically, integrating both strategies produced a statistically significant increase in AUROC – achieving an additional 6.14 % performance gain ($p = 0.02$) – compared to the unregistered dataset baseline. This complementary effect suggests that registration and misalignment augmentation address distinct limitations and, when combined, reinforce each other.

A potential explanation for this synergistic behavior is that registration primarily ensures anatomical alignment across modalities by eliminating large spatial offsets, thereby allowing the network to focus on learning more complex and consistent multi-modal representations. In contrast, misalignment augmentation introduces controlled variability into the training process, forcing the CNN to develop greater robustness to residual alignment imperfections that inevitably persist even after preprocessing. Together, these strategies handle different aspects of the alignment problem: registration improves the overall data consistency, while augmentation enhances model robustness. By extending

each other's strengths, the combined approach leads to a more powerful and generalizable model capable of handling realistic clinical variability.

It is important to note, however, that the necessity of such explicit strategies may depend on dataset size. In very large datasets with extensive variability, a CNN might gradually learn to become invariant to modest misalignments, provided that the network architecture and capacity are sufficient. Nevertheless, in the context of typical medical imaging studies – where datasets are limited in both size and diversity, and spatial ground truth consistency is crucial – targeted strategies such as registration and augmentation remain essential for ensuring robust model performance.

Notably, the combined strategy not only achieved the highest AUROC on the independent test set but also shifted the receiver operating characteristic (ROC) curve significantly closer to the radiologists' performance threshold at Prostate Imaging Reporting and Data System (PI-RADS) ≥ 4 – a clinically critical decision point that was not reached by any of the other configurations tested.

Answer for Research Question 6: Registration and misalignment data augmentation are complementary strategies leading to a synergetic performance increase.

4.2.4 Potential Improvements in Small Lesion Detection Sensitivity

The qualitative results presented in Fig. 3.5 suggest that addressing alignment errors between modalities through misalignment augmentation leads to improved lesion segmentation coverage, not only for large lesions but also for small clinically significant lesions. Notably, misalignment augmentation enabled the network to more reliably detect a small punctate lesion that was only weakly identified by models trained without this augmentation strategy. This observation highlights the particular vulnerability of small lesions to subtle spatial inconsistencies: even minor misalignments, not necessarily visible to the human eye, can critically impair the ability of CNN-based models to detect such lesions.

Augmentation strategies that increase the alignment variability during training therefore can be particularly beneficial – a critical consideration for clinical applicability, especially in early disease stages where small punctate lesions may represent the only indicators of malignancy. Beyond improving sensitivity, robust detection of small lesions is essential for ensuring spatially accurate segmentation outputs – an important requirement in applications such as targeted biopsies or focal therapy planning.

Nevertheless, it is important to acknowledge that these findings are based on individual examples. To quantitatively prove the observed improvements, further experiments

using larger datasets with a higher number of small lesions, alongside stratified evaluation according to lesion size, would be necessary.

4.2.5 Multi-Modal Alignment Errors Beyond Prostate MRI

Alignment errors between image modalities are a general challenge in multi-modal medical image analysis wherever spatial accuracy is critical, not just in prostate multi-parametric MRI (mpMRI). Therefore, the proposed misalignment DA technique is relevant to other applications. In the studies by Kovacs et al. (2024) and Rokuss et al. (2024), I successfully utilized this strategy in a different domain: automated lesion segmentation in whole-body PET/CT images for the AutoPET3 challenge (Ingrisch et al., 2024; Gatidis and Kuestner, 2022; Jeblick, 2024).

Despite the hybrid PET/CT imaging system, patient and organ motion during image acquisition often introduce alignment errors, similarly to prostate mpMRI. Furthermore, due to attenuation correction during image reconstruction, these errors may be further amplified (Alessio et al., 2004; Hunter et al., 2016; Kaji et al., 2024), potentially limiting segmentation accuracy. To mitigate this, misalignment augmentation was utilized in both challenge submissions and shown to enhance segmentation performance. The impact of this augmentation in the winning solution by Rokuss et al. (2024) is summarized in Tab. 4.1.

Table 4.1: Segmentation performance (five-fold cross-validation) of the winning solution of AutoPET3 Challenge for the baseline model trained with and without misalignment augmentation. Results from the study by Rokuss et al. (2024) are reproduced under the Creative Commons Attribution 4.0 License.

	Dice↑				
Training Configuration	All	FDG	PSMA	FPvol↓	FNvol↓
nnU-Net (ResEnc L)	65.31	72.87	58.25	10.47	13.63
+ misalignment augmentation	**65.76**	**73.13**	**58.89**	**10.12**	**12.50**

It is important to highlight that the winning method by Rokuss et al. (2024) did not rely solely on misalignment augmentation. Additional improvements included modifications to the loss function, increased batch size, and incorporation of anatomical knowledge by jointly segmenting problem-relevant organs during training.

4.2.6 Conceptual Summary of Multi-Modal Alignment Error Handling Strategies

Based on the results obtained from applying different strategies to address multi-modal alignment errors between prostate MRI images, the effect of each approach can be conceptually illustrated on the spatial representation of lesion ground truth during model training, as shown in Fig. 4.3.

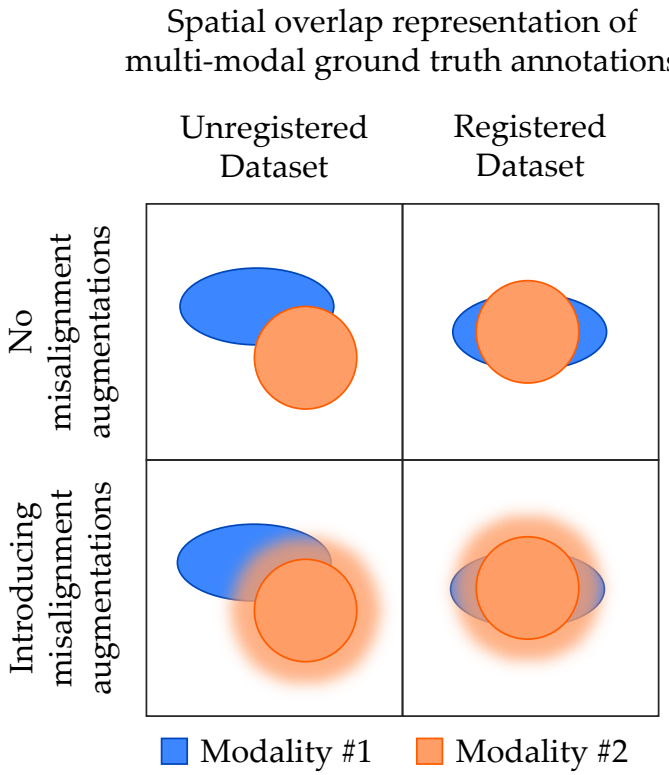

Figure 4.3: Conceptual visualization of the spatial overlap of multi-modal ground truth information under different strategies. Registration improves anatomical alignment between modalities, although residual errors typically persist. Misalignment augmentation enhances model robustness by introducing controlled variability during training. While augmentation alone can compensate for moderate alignment errors, combining it with registration has the potential to fully address residual misalignments, leading to improved learning of multi-modal representations.

As a summary:

- **Registration** aims to spatially align the different modalities, thereby the associated lesion ground truth across them. This alignment facilitates the integration of complementary information from the T2-weighted (T2w), diffusion-weighted imaging (DWI), and apparent diffusion coefficient (ADC) images, a process that mirrors the clinical PI-RADS interpretation workflow (Fig. 1.3), where radiologists rely on their cognitive abilities to mentally compensate for alignment differences. However, even after registration, residual errors typically remain.
- **Misalignment augmentation**, in contrast, impacts training differently by intentionally introducing potential alignment errors as an inductive bias during model training. This forces the model to become partially invariant to multi-modal misalignments to some extent, thereby leading to a more robust model.
- **Combining registration and misalignment augmentation** achieves the best model configuration, leveraging the advantages of both strategies: improved anatomical consistency through registration and enhanced robustness through augmentation against remaining registration errors.

Overall, the findings suggest that combining registration with misalignment augmentation provides an optimal strategy for training CNN-based CAD systems in prostate MRI. Furthermore, I advocate for incorporating misalignment augmentation into the standard set of training tools for multi-modal image analysis, especially in applications where imperfect ground truth alignment persists and cannot be fully corrected by preprocessing alone.

4.3 Future of Application-specific Augmentations in Medicine

DA plays currently a crucial role in increasing both the quantity and quality of training data, thereby enhancing the robustness and generalizability of AI models. This is especially critical in medical imaging, where datasets are often limited in size, highly imbalanced, and expensive to annotate. However, this dependency on augmentation may shift over time due to current trends in medical AI and data availability. Several key factors are shaping this evolving landscape:

- **Growing Availability of Clinical Data:** As more institutions adopt digital infrastructure and standardized protocols for AI applications, the volume and representativeness of clinical datasets are expected to increase. This growth in data volume and diversity could reduce the dependency on DA for many routine applications. Models trained on naturally diverse clinical samples are likely to generalize better across populations, potentially decreasing the relative importance of traditional augmentation strategies over time. Nevertheless, it is important to emphasize that even with expanding datasets, augmentation will likely remain essential in specific or newly emerging scenarios – such as diseases with low prevalence, rare subtypes of common conditions, or specialized imaging protocols – where applications continue to suffer from limited sample sizes and demographic imbalance. In such cases, targeted augmentation strategies, particularly those informed by domain knowledge, will probably still play an important role in providing useful guidance for AI model training.

- **Emergence of Federated Learning:** Federated learning is an emerging paradigm that enables the training of models across distributed datasets from multiple clinical centers without violating data protection regulations (Rieke et al., 2020; Roth et al., 2020). This approach offers access to otherwise privately held data, making it possible to train models on larger, more diverse, and multi-center datasets without the need for centralized aggregation (Pati et al., 2022). However, while federated learning increases both dataset size and diversity, it also introduces new challenges. The heterogeneity of multi-center datasets – originating from differences in imaging protocols, scanner hardware, and annotation standards – and unseen data quality can significantly affect model performance. In particular, inconsistent ground truth definitions and inter-rater variability often degrade model generalizability. Additionally, federated learning demands collaboration beyond model development, including legal alignment, data harmonization, shared research goals, and coordinated project management (Bujotzek et al., 2025). Despite these hurdles, federated learning remains the most promising collaborative training framework available.

When effectively implemented, it has been shown to match the performance of centralized approaches while preserving privacy and enabling access to a huge amount of clinical data (Sheller et al., 2020).

- **Rise of Foundation Models:** In recent years, foundation models resulting from large-scale pretraining have become a central focus of the AI community (Bommasani et al., 2021). These models represent a fundamentally different paradigm from application-specific solutions, aiming to operate in a purely data-driven manner by leveraging vast amounts of unlabeled data. A variety of self-supervised and contrastive learning strategies – such as DINO (Caron et al., 2021; Oquab et al., 2023), SimCLR (Chen et al., 2020), and masked autoencoders (He et al., 2022) – have been developed for this purpose, commonly based on high-capacity Vision Transformer architectures. While foundation models have demonstrated remarkable performance in the natural imaging domain, their adoption in medical imaging is still emerging. Initial efforts such as MedSAM (Ma et al., 2024a) have demonstrated their feasibility for medical image segmentation. More recently, the integration of large language models (Beyer et al., 2024; Liu et al., 2023; Touvron et al., 2023) into vision architectures has enabled multi-modal reasoning by incorporating clinical metadata and radiology reports. This approach is demonstrated by models such as BioMedParse (Zhao et al., 2025) and BioMedCLIP (Zhang et al., 2023). The strong performance of these generalist models, which have been pre-trained on large-scale and diverse datasets, suggests a potential shift toward representation and transfer learning, possibly reducing the reliance on application-specific augmentation strategies. However, their clinical deployment still faces several challenges. Most notably, their extreme data requirements pose a significant barrier in the medical domain, where data privacy limits access to large, diverse datasets. Furthermore, current implementations primarily rely on native 2D images (like chest X-rays, histopathology scans) and 2D slices extracted from 3D scans, which introduces domain imbalance and lack of 3D understanding. Moreover, foundation models typically require fine-tuning to adapt to domain-specific imaging characteristics, including contrast and geometrical properties. Consequently, robust end-to-end solutions based on training on 3D volumes remain absent. Overcoming these limitations will be critical for realizing the full potential of foundation models in clinical workflows.

While these developments may reduce the dependency on domain-specific solutions – including strategies that introduce application-specific biases through augmentation – their practical effectiveness in the medical domain remains an open question. Clinical data is inherently heterogeneous – ranging across differences in scanners, imaging protocols, patient populations, and disease presentations – and it continues to evolve with ongoing

technological innovation.

In such a dynamic and complex setting, application-specific strategies may still offer meaningful performance gains, particularly in subpopulations with rare or atypical conditions that are frequently the focus of clinical research. Carefully designed inductive biases may thus remain important tools for bridging the performance gap between expert clinicians and AI systems. Rather than fully replacing domain-specific solutions, the future of clinical AI may rely on thoughtfully combining broad generalization capabilities with application-specific refinements, achieving both robust overall performance and improved handling of specialized cases across diverse real-world settings.

4.4 Conclusion

This dissertation focused on incorporating domain-specific knowledge into the training of AI models for prostate cancer PCa diagnosis, aiming to enhance the performance of state-of-the-art methods, thereby narrowing the gap between automated systems and expert radiologists. Specifically, it successfully addressed two clinically relevant challenges that radiologists routinely account for during image interpretation but are often overlooked in standard AI pipelines:

- To increase lesion shape variability, a lightweight biomechanical model was developed to simulate realistic soft tissue deformations in the pelvic region, particularly variations in rectal and bladder size that influence prostate anatomy and lesion morphology. By integrating these simulated deformations as online data augmentation during model training, diagnostic accuracy improved significantly, approaching towards radiologist-level performance. Due to its computational efficiency, the method is highly compatible with modern deep learning workflows.

- To address multi-modal misalignments between prostate MRI sequences – which lead to inconsistencies in spatial alignment and ground truth representation – this work systematically evaluated two strategies: conventional image registration and a proposed novel approach called misalignment augmentation. While both techniques improved diagnostic performance individually, their combination resulted in a statistically significant synergistic effect, addressing complementary aspects of the misalignment problem, and resulted in an on par model performance with radiologists.

Given their effectiveness in enhancing model performance in PCa diagnosis – as well as their conceptual applicability to other domains – both techniques are recommended to use as blueprints for future research in medical AI.

In conclusion, this work demonstrates that even a single, well-designed augmentation strategy – when guided by domain-specific knowledge, as in the case of anatomy-informed and misalignment augmentation – can significantly improve model robustness and lead to enhanced diagnostic accuracy.

While emerging trends in medical AI – such as federated learning and foundation models – may reduce reliance on task-specific augmentations, the inherent challenges of the domain – including limited data availability, heterogeneity, and highly specialized cases – will likely continue to necessitate carefully engineered inductive biases. Informed by domain expertise, these targeted strategies could still remain powerful tools for aligning model behavior with clinical needs. The future of clinical AI may not lie in choosing between generalization and specificity, but rather in strategically combining the two: pairing large-scale, general-purpose learning with targeted, application-specific refinements to achieve both scalable performance and reliable clinical utility across diverse real-world contexts.

Chapter 5

Summary

In recent years, artificial intelligence (AI) has made a significant impact on prostate cancer diagnosis using magnetic resonance imaging (MRI), particularly through diagnostic systems based on deep learning approaches. Among these, convolutional neural networks trained for semantic segmentation of clinically significant lesions have gained attention due to their clinical value and inherent interpretability. Used as assistive tools, such systems have already been shown not only to increase diagnostic accuracy, but also to reduce both inter-rater variability and diagnostic time. Despite these advances, standalone AI models for prostate cancer diagnosis still underperform compared to expert radiologists. The reason for radiologists' superiority may lie in their clinical training to account for physiological and modality-specific image alterations using domain knowledge and cognitive reasoning, aspects that are currently overlooked in state-of-the-art computer-aided diagnosis systems.

To address this performance gap, this thesis advances prostate MRI interpretation by incorporating two real-world, yet often overlooked, challenges into AI model development: (1) frequent soft tissue deformations caused by physiological processes and (2) misalignment between multi-modal images. Both are forms of spatial variation to which segmentation networks are potentially sensitive. For each challenge, targeted, domain-informed strategies are proposed. These data-centric solutions are implemented as on-the-fly data augmentations during training, acting as inductive biases to improve model robustness against clinically relevant sources of image alterations.

Although biomechanical models based on finite element methods hold strong potential for increasing prostate and lesion shape variability during training by simulating realistic soft tissue deformations, their practical utility is limited due to computational complexity and the need for specialized modeling expertise. To make such deformations suitable for scalable online data augmentation, a lightweight model was developed by introducing simplified biomechanical assumptions. Incorporating these deformations into model training improved both patient-level diagnostic accuracy and lesion-level detection rates. Furthermore, the benefit of using anatomically realistic transformations was demonstrated in contrast to random elastic deformations, which are prone to distort image features and compromise the fidelity of ground truth labels for benign and malignant conditions.

Another clinical challenge addressed is the alignment errors between MRI imag-

ing modalities. While radiologists can cognitively compensate for such inconsistencies, computer-aided diagnosis systems rely on aligned ground truth representations across all image modalities. However, the literature lacks consensus on whether image co-registration is beneficial for model training. Furthermore, when registration is applied, its effect on model performance is rarely reported. To systematically investigate this, multiple registration strategies were evaluated alongside a novel approach: misalignment augmentation. Instead of aiming for perfect anatomical alignment, this method introduces synthetic alignment errors during training to make network predictions invariant to such errors. Both registration and misalignment augmentation independently improved performance. Moreover, combining the two approaches led to a synergistic effect, further improving performance due to their complementary behavior and yielding a statistically significant improvement that brought diagnostic performance on par with expert radiologists. Further results also highlighted that common surrogate registration metrics (e.g. Dice coefficient) do not necessarily correlate with clinical task performance, emphasising the importance of evaluating strategies based on their impact on clinically relevant questions.

The insights gained from the proposed data-centric strategies demonstrated their effectiveness, as reflected in the significant performance improvements observed on independent test sets. These findings underscore that incorporating domain knowledge into neural network training via data augmentation as an inductive bias can yield substantial benefits beyond those of generic state-of-the-art training pipelines. While the increasing availability of large-scale training data and the rise of generalist foundation models may reduce the reliance on such targeted solutions for routine applications, the inherent complexity of medical imaging suggests that domain-specific strategies will likely remain essential for enabling neural networks to address nuanced, clinically complex scenarios. This thesis makes a significant contribution to the field by demonstrating how clinically grounded, data-centric strategies can narrow the performance gap between automated systems and expert radiologists.

Chapter 6

Zusammenfassung

Künstliche Intelligenz (KI) hat in den letzten Jahren einen signifikanten Einfluss auf die Diagnose von Prostatakrebs mittels Magnetresonanztomographie (MRT) gehabt. Diagnostische Systeme auf Basis tiefer neuronaler Faltungsnetze, insbesondere solche, die auf die semantische Segmentierung klinisch signifikanter Läsionen trainiert wurden, sind aufgrund ihres klinischen Nutzens und ihrer inhärenten Interpretierbarkeit beliebt geworden. Als diagnostische Unterstützung eingesetzt, haben diese Systeme nicht nur die diagnostische Genauigkeit erhöht, sondern auch die Variabilität zwischen Beurteilenden sowie die Diagnosedauer reduziert. Trotz dieser Fortschritte bleiben KI-Systeme zur eigenständigen Prostatakrebsdiagnose hinter der Leistung erfahrener Radiolog:innen zurück. Der Grund dafür liegt möglicherweise in der klinischen Ausbildung der Radiolog:innen. Aufgrund ihres Fachwissens und kognitiven Denkens, können sie physiologische und bildmodalitätsspezifische Bildveränderungen berücksichtigen. Aspekte, die bislang in modernen rechnergestützten Diagnosesystemen wenig berücksichtigt wurden.

Um diese Lücke zu überbrücken, verbessert diese Dissertation die Interpretation von Prostata MRTs durch Einbeziehung zwei realer, jedoch häufig übersehener Herausforderungen in der Entwicklung von KI-Modellen: (1) häufige Weichteildeformationen durch physiologische Prozesse und (2) Fehlausrichtungen zwischen multimodalen Bildern. Beide stellen Formen räumlicher Variationen dar, auf die Segmentierungsnetzwerke potenziell empfindlich reagieren. Für jede dieser Herausforderungen werden gezielte, domänenspezifische Strategien entwickelt. Diese datenzentrierten Lösungen werden als Echtzeit-Datenaugmentierungen während des Trainings implementiert und dienen als induktive Bias, um die Robustheit des Modells gegenüber klinisch relevanten Bildveränderungen zu erhöhen.

Obwohl biomechanische Modelle basiert auf Finite-Elemente-Methoden großes Potenzial zur Erhöhung der anatomischen Variabilität von Prostata und Läsionen durch die Simulation realistischer Weichteildeformationen bieten, ist ihr praktischer Nutzen aufgrund der hohen Rechenkomplexität und des Bedarfs an spezialisierter Modellierungsexpertise begrenzt. Um solche Deformationen für eine skalierbare Echtzeit-Datenerweiterung nutzbar zu machen, wurde ein leichtgewichtiges Modell entwickelt, das auf vereinfachten biomechanischen Annahmen basiert. Die Integration dieser Deformationen in das Modelltraining verbesserte sowohl die diagnostische Genauigkeit auf Patientenebene als auch die

Erkennungsraten auf Läsionsebene. Darüber hinaus zeigte sich der Nutzen anatomisch realistischer Transformationen im Vergleich zu zufälligen elastischen Deformationen, die dazu neigen, Bildmerkmale zu verzerren und die Genauigkeit der manuellen Referenzannotationen für benigne und maligne Befunde zu beeinträchtigen.

Eine weitere klinische Herausforderung sind Fehlregistrierungen zwischen verschiedenen MRT-Bildmodalitäten. Während Radiolog:innen solche Inkonsistenzen kognitiv ausgleichen können, sind computergestützte Diagnosesysteme auf exakt ausgerichtete manuelle Referenzannotationen über alle Bildmodalitäten hinweg angewiesen. In der Fachliteratur besteht jedoch kein Konsens darüber, ob Bild-Koregistrierung tatsächlich vorteilhaft für das Modelltraining ist. Zudem wird deren Auswirkung auf die Modellleistung kaum berichtet. Um dies systematisch zu untersuchen, wurden mehrere Registrierungsstrategien entlang mit einem neuartigen Ansatz evaluiert: die Fehlregistrierung-Augmentierung. Anstatt eine perfekte anatomische Ausrichtung anzustreben, führt diese Methode während des Trainings synthetische Ausrichtungsfehler zwischen den MRT-Bildmodalitäten gezielt ein, um die Netzwerkvorhersagen gegenüber solchen Fehlern robust zu machen. Sowohl die Registrierung als auch die Fehlregistrierung-Augmentierung verbesserten jeweils unabhängig die diagnostische Leistung. Darüber hinaus führte die Kombination beider Ansätze zu einem synergetischen Effekt, der durch ihr komplementäres Verhalten zu einer signifikanten Leistungsverbesserung führte und die diagnostische Leistung auf das Niveau erfahrener Radiolog:innen brachte. Weitere Ergebnisse verdeutlichen zudem, dass herrkömmliche Ersatz-Registrierungsmetriken (z.B. Dice-Koeffizient) nicht unbedingt mit der klinischen Aufgabenleistung korrelieren, was die Bedeutung der Evaluierung von Strategien anhand ihrer Auswirkungen auf klinisch relevante Fragestellungen betont.

Die Erkenntnisse aus den vorgeschlagenen datenzentrierten Strategien zeigten ihre Wirksamkeit, was sich in den signifikanten Leistungsverbesserungen auf unabhängigen Testdatensätzen widerspiegelte. Diese Ergebnisse unterstreichen, dass die Einbeziehung von Domänenwissen in das Training neuronaler Netze durch Datenaugmentation als induktive Bias erhebliche Vorteile gegenüber herkömmlichen, modernen Trainingspipelines bieten kann. Zwar könnten die zunehmende Verfügbarkeit umfangreicher Trainingsdaten und der Aufstieg generalistischer Foundation-Modelle die Abhängigkeit von solch zielgerichteten Lösungen in Routineanwendungen verringern, doch die inhärente Komplexität der medizinischen Bildgebung lässt vermuten, dass domänenspezifische Strategien weiterhin essentiell bleiben, um neuronale Netze in die Lage zu versetzen, differenzierte und klinisch komplexe Szenarien zu bewältigen. Diese Arbeit leistet einen wichtigen Beitrag zum Fachgebiet, indem sie zeigt, wie klinisch fundierte, datenzentrierte Strategien die Leistungslücke zwischen automatisierten Systemen und erfahrenen Radiolog:innen verringern können.

Bibliography

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I. J., Harp, A., Irving, G., Isard, M., Jia, Y., Józefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D. G., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P. A., Vanhoucke, V., Vasudevan, V., Viégas, F. B., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2016). **Tensorflow: Large-scale machine learning on heterogeneous distributed systems**. *arXiv preprint arXiv:1603.04467*. unpublished findings.

Ahdoot, M., Wilbur, A. R., Reese, S. E., Lebastchi, A. H., Mehralivand, S., Gomella, P. T., Bloom, J., Gurram, S., Siddiqui, M., Pinsky, P., Parnes, H., Linehan, W. M., Merino, M., Choyke, P. L., Shih, J. H., Turkbey, B., Wood, B. J., and Pinto, P. A. (2020). **MRI-targeted, systematic, and combined biopsy for prostate cancer diagnosis**. *New England Journal of Medicine*, 382(10):917–928. doi: 10.1056/NEJMoa1910038.

Ahmed, H. U., El-Shater Bosaily, A., Brown, L. C., Gabe, R., Kaplan, R., Parmar, M. K., Collaco-Moraes, Y., Ward, K., Hindley, R. G., Freeman, A., Kirkham, A. P., Oldroyd, R., Parker, C., Emberton, M., and PROMIS study group. (2017). **Diagnostic accuracy of multi-parametric MRI and trus biopsy in prostate cancer (PROMIS): a paired validating confirmatory study**. *The Lancet*, 389(10071):815–822. doi: 10.1016/S0140-6736(16)32401-1.

Akinci D'Antonoli, T., Berger, L. K., Indrakanti, A. K., Vishwanathan, N., Weiss, J., Jung, M., Berkarda, Z., Rau, A., Reisert, M., Küstner, T., Walter, A., Merkle, E. M., Boll, D. T., Breit, H.-C., Nicoli, A. P., Segeroth, M., Cyriac, J., Yang, S., and Wasserthal, J. (2025). **TotalSegmentator MRI: Robust sequence-independent segmentation of multiple anatomic structures in MRI**. *Radiology*, 314(2):e241613.

Aldoj, N., Lukas, S., Dewey, M., and Penzkofer, T. (2020). **Semi-automatic classification of prostate cancer on multi-parametric MR imaging using a multi-channel 3D convolutional neural network**. *European radiology*, 30(2):1243–1253.

Alessio, A. M., Kinahan, P. E., Cheng, P. M., Vesselle, H., and Karp, J. S. (2004). **PET/CT scanner instrumentation, challenges, and solutions**. *Radiologic Clinics*, 42(6):1017–1032.

Ali, A., Du Feu, A., Oliveira, P., Choudhury, A., Bristow, R. G., and Baena, E. (2022). **Prostate zones and cancer: lost in transition?** *Nature Reviews Urology*, 19(2):101–115.

Alkadi, R., Taher, F., El-Baz, A., and Werghi, N. (2019). **A deep learning-based approach for the detection and localization of prostate cancer in T2 magnetic resonance images**. *Journal of digital imaging*, 32(5):793–807. doi: 10.1007/s10278-018-0160-1.

Arif, M., Schoots, I. G., Tovar, J. C., Bangma, C. H., Krestin, G. P., Roobol, M. J., Niessen, W., and Veenland, J. F. (2020). **Clinically significant prostate cancer detection and segmentation in low-risk patients using a convolutional neural network on multi-parametric MRI**. *European radiology*, 30(12):6582–6592. doi: 10.1007/s00330-020-07008-z.

Armato, S. G., Huisman, H., Drukker, K., Hadjiiski, L., Kirby, J. S., Petrick, N., Redmond, G., Giger, M. L., Cha, K., Mamonov, A., Kalpathy-Cramer, J., and Farahani, K. (2018). **PROSTATEx challenges for computerized classification of prostate lesions from multiparametric magnetic resonance images**. *Journal of Medical Imaging*, 5(4):044501–044501.

Arnoldner, M. A., Polanec, S. H., Lazar, M., Noori Khadjavi, S., Clauser, P., Pötsch, N., Schwarz-Nemec, U., Korn, S., Hübner, N., Shariat, S. F., Helbich, T. H., and Baltzer, P. A. (2022). **Rectal preparation significantly improves prostate imaging quality: assessment of the PI-QUAL score with visual grading characteristics**. *European Journal of Radiology*, 147:110145.

Ashburner, J. and Friston, K. J. (1999). **Nonlinear spatial normalization using basis functions**. *Human brain mapping*, 7(4):254–266.

Azam, M. A., Khan, K. B., Salahuddin, S., Rehman, E., Khan, S. A., Khan, M. A., Kadry, S., and Gandomi, A. H. (2022). **A review on multimodal medical image fusion: Compendious analysis of medical modalities, multimodal databases, fusion techniques and quality metrics**. *Computers in biology and medicine*, 144:105253.

Banerjee, B., Iqbal, B. M., Kumar, H., Kambale, T., and Bavikar, R. (2016). **Correlation between prostate specific antigen levels and various prostatic pathologies**. *Journal of Medical Society*, 30(3):172–175.

Barentsz, J. O., Richenberg, J., Clements, R., Choyke, P., Verma, S., Villeirs, G., Rouviere, O., Logager, V., and Fütterer, J. J. (2012). **Esur prostate MR guidelines 2012**. *European radiology*, 22:746–757.

Barentsz, J. O., Weinreb, J. C., Verma, S., Thoeny, H. C., Tempany, C. M., Shtern, F., Padhani, A. R., Margolis, D., Macura, K. J., Haider, M. A., Cornud, F., and Choyke, P. L. (2016). **Synopsis of the PI-RADS v2 guidelines for multiparametric prostate magnetic resonance imaging and recommendations for use**. *European urology*, 69(1):41.

Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.-A., Cetin, I., Lekadir, K., Camara, O., Gonzalez Ballester, M. A., Sanroma, G., Napel, S., Petersen, S., Tziritas, G., Grinias, E., Khened, M., Kollerathu, V. A., Krishnamurthi, G., Rohé, M.-M., Pennec, X., Sermesant, M., Isensee, F., Jäger, P., Maier-Hein, K. H., Full, P. M., Wolf, I., Engelhardt, S., Baumgartner, C. F., Koch, L. M., Wolterink, J. M., Išgum, I., Jang, Y., Hong, Y., Patravali, J., Jain, S., Humbert, O., and Jodoin, P.-M. (2018). **Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved?** *IEEE transactions on medical imaging*, 37(11):2514–2525. doi: 10.1109/TMI.2018.2837502.

Beyer, L., Steiner, A., Pinto, A. S., Kolesnikov, A., Wang, X., Salz, D., Neumann, M., Alabdulmohsin, I., Tschannen, M., Bugliarello, E., Unterthiner, T., Keysers, D., Koppula, S., Liu, F., Grycner, A., Gritsenko, A., Houlsby, N., Kumar, M., Rong, K., Eisenschlos, J., Kabra, R., Bauer, M., Bošnjak, M., Chen, X., Minderer, M., Voigtlaender, P., Bica, I., Balazevic, I., Puigcerver, J., Papalampidi, P., Henaff, O., Xiong, X., Soricut, R., Harmsen, J., and Zhai, X. (2024). **Paligemma: A versatile 3b vlm for transfer**. *arXiv preprint arXiv:2407.07726*. unpublished findings.

Beyersdorff, D., Winkel, A., Hamm, B., Lenk, S., Loening, S. A., and Taupitz, M. (2005). **MR imaging–guided prostate biopsy with a closed MR unit at 1.5 T: initial results**. *Radiology*, 234(2):576–581.

Bhattacharya, I., Khandwala, Y. S., Vesal, S., Shao, W., Yang, Q., Soerensen, S. J., Fan, R. E., Ghanouni, P., Kunder, C. A., Brooks, J. D., Hu, Y., Rusu, M., and Sonn, G. A. (2022). **A review of artificial intelligence in prostate cancer detection on imaging**. *Therapeutic advances in urology*, 14:17562872221128791.

Bilello, M., Akbari, H., Da, X., Pisapia, J. M., Mohan, S., Wolf, R. L., O'Rourke, D. M., Martinez-Lage, M., and Davatzikos, C. (2016). **Population-based MRI atlases of spatial distribution are specific to patient and tumor characteristics in glioblastoma**. *NeuroImage: Clinical*, 12:34–40.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., Donahue, C., Doumbouya, M., Durmus, E., Ermon, S., Etchemendy, J., Ethayarajh, K., Fei-Fei, L., Finn, C., Gale, T., Gillespie, L., Goel, K., Goodman, N., Grossman, S., Guha, N., Hashimoto, T., Henderson, P., Hewitt, J., Ho, D. E., Hong, J., Hsu, K., Huang, J., Icard, T., Jain, S., Jurafsky, D., Kalluri, P., Karamcheti, S., Keeling, G., Khani, F., Khattab, O., Koh, P. W., Krass, M., Krishna, R., Kuditipudi, R., Kumar, A., Ladhak, F., Lee, M., Lee, T., Leskovec, J., Levent, I., Li, X. L., Li, X., Ma, T., Malik, A., Manning, C. D., Mirchandani, S., Mitchell, E., Munyikwa, Z., Nair, S., Narayan, A., Narayanan, D., Newman, B., Nie, A., Niebles, J. C., Nilforoshan, H., Nyarko, J., Ogut, G., Orr, L., Papadimitriou, I., Park, J. S., Piech, C., Portelance, E., Potts, C., Raghunathan, A., Reich, R., Ren, H., Rong, F., Roohani, Y., Ruiz, C., Ryan, J., Ré, C., Sadigh, D., Sagawa, S., Santhanam, K., Shih, A., Srinivasan, K., Tamkin, A., Taori, R., Thomas, A. W., Tramèr, F., Wang, R. E., Wang, W., Wu, B., Wu, J., Wu, Y., Xie, S. M., Yasunaga, M., You, J., Zaharia, M., Zhang, M., Zhang, T., Zhang, X., Zhang, Y., Zheng, L., Zhou, K., and Liang, P. (2021). **On the opportunities and risks of foundation models**. *arXiv preprint arXiv:2108.07258*. unpublished findings.

Bosma, J. S., Saha, A., Hosseinzadeh, M., Slootweg, I., de Rooij, M., and Huisman, H. (2021)a. **Annotation-efficient cancer detection with report-guided lesion annotation for deep learning-based prostate cancer detection in bpMRI**. *arXiv preprint arXiv:2112.05151*. pre-published findings.

Bosma, J. S., Saha, A., Hosseinzadeh, M., Slootweg, I., de Rooij, M., and Huisman, H. (2021)b. **Report-guided automatic lesion annotation for deep learning-based prostate cancer detection in bpMRI**. *arXiv preprint arXiv:2112.05151*. pre-published findings.

Bosma, J. S., Saha, A., Hosseinzadeh, M., Slootweg, I., de Rooij, M., and Huisman, H. (2023). **Semisupervised learning with report-guided pseudo labels for deep learning–based prostate cancer detection using biparametric MRI**. *Radiology: Artificial Intelligence*, 5(5): e230031.

Bottou, L. (2010). **Large-scale machine learning with stochastic gradient descent**. In *Proceedings of COMPSTAT'2010: 19th International Conference on Computational StatisticsParis France, August 22-27, 2010 Keynote, Invited and Contributed Papers*, pages 177–186. Springer.

Bottou, L., Curtis, F. E., and Nocedal, J. (2018). **Optimization methods for large-scale machine learning**. *SIAM review*, 60(2):223–311.

Boubaker, M. B. and Ganghoffer, J.-F. (2017). **Chapter 14 - bladder/prostate/rectum: Biomechanical models of the mobility of pelvic organs in the context of prostate radiotherapy**. In Payan, Y. and Ohayon, J., editors, *Biomechanics of Living Organs*, volume 1 of *Translational Epigenetics*, pages 307–324. Academic Press, Oxford. doi: 10.1016/B978-0-12-804009-6.00014-6.

Bounias, D., Baumgartner, M., Neher, P., Kovacs, B., Floca, R., Jaeger, P. F., Kapsner, L., Eberle, J., Hadler, D., Laun, F., Ohlmeyer, S., Maier-Hein, K., and Bickelhaupt, S. (2023). **Risk-adjusted training and evaluation for medical object detection in breast cancer MRI**. In *ICML 3rd Workshop on Interpretable Machine Learning in Healthcare (IMLH)*.

Bouras, S. (2024). **Digital rectal exam in prostate cancer screening: a critical review of the erspc rotterdam study**. *African Journal of Urology*, 30(1):51.

Bray, F., Laversanne, M., Sung, H., Ferlay, J., Siegel, R. L., Soerjomataram, I., and Jemal, A. (2024). **Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries**. *CA: A Cancer Journal for Clinicians*, 74(3):229–263. doi: 10.3322/caac.21834.

Brown, R. W., Cheng, Y.-C. N., Haacke, E. M., Thompson, M. R., and Venkatesan, R. (2014). *Magnetic resonance imaging: physical principles and sequence design*. John Wiley & Sons, Hoboken, 2 edition.

Bujotzek, M. R., Akünal, U., Denner, S., Neher, P., Zenk, M., Frodl, E., Jaiswal, A., Kim, M., Krekiehn, N. R., Nickel, M., Ruppel, R., Both, M., Döllinger, F., Opitz, M., Persigehl, T., Kleesiek, J., Penzkofer, T., Maier-Hein, K., Bucher, A., and Braren, R. (2025). **Real-world federated learning in radiology: hurdles to overcome and benefits to gain**. *Journal of the American Medical Informatics Association*, 32(1):193–205.

Caglic, I., Hansen, N. L., Slough, R. A., Patterson, A. J., and Barrett, T. (2017). **Evaluating the effect of rectal distension on prostate multiparametric MRI image quality**. *European journal of radiology*, 90:174–180.

Cao, R., Bajgiran, A. M., Mirak, S. A., Shakeri, S., Zhong, X., Enzmann, D., Raman, S., and Sung, K. (2019). **Joint prostate cancer detection and gleason score prediction in mp-MRI via focalnet**. *IEEE transactions on medical imaging*, 38(11):2496–2506. doi: 10.1109/TMI.2019.2901928.

Cao, X., Fan, J., Dong, P., Ahmad, S., Yap, P.-T., and Shen, D. (2020). **Chapter 14 - image registration using machine and deep learning**. In Zhou, S. K., Rueckert, D., and Fichtinger, G., editors, *Handbook of Medical Image Computing and Computer Assisted*

Intervention, The Elsevier and MICCAI Society Book Series, pages 319–342. Academic Press. ISBN 978-0-12-816176-0. doi: 10.1016/B978-0-12-816176-0.00019-3.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. (2021). **Emerging properties in self-supervised vision transformers**. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660.

Carter, T. J., Sermesant, M., Cash, D. M., Barratt, D. C., Tanner, C., and Hawkes, D. J. (2005). **Application of soft tissue modelling to image-guided surgery**. *Medical engineering & physics*, 27(10):893–909.

Carvalhal, G. F., Daudi, S. N., Kan, D., Mondo, D., Roehl, K. A., Loeb, S., and Catalona, W. J. (2010). **Correlation between serum prostate-specific antigen and cancer volume in prostate glands of different sizes**. *Urology*, 76(5):1072–1076.

Chai, X., Van Herk, M., Van De Kamer, J. B., Hulshof, M. C., Remeijer, P., Lotz, H. T., and Bel, A. (2011). **Finite element based bladder modeling for image-guided radiotherapy of bladder cancer**. *Medical physics*, 38(1):142–150.

Chan, D., Fox, N., Jenkins, R., Scahill, R., Crum, W., and Rossor, M. (2001). **Rates of global and regional cerebral atrophy in AD and frontotemporal dementia**. *Neurology*, 57(10): 1756–1763.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020). **A simple framework for contrastive learning of visual representations**. In *International conference on machine learning*, pages 1597–1607. PmLR.

Chesnais, A., Niaf, E., Bratan, F., Mège-Lechevallier, F., Roche, S., Rabilloud, M., Colombel, M., and Rouvière, O. (2013). **Differentiation of transitional zone prostate cancer from benign hyperplasia nodules: evaluation of discriminant criteria at multiparametric MRI**. *Clinical radiology*, 68(6):e323–e330.

Collignon, A., Maes, F., Delaere, D., Vandermeulen, D., Suetens, P., and Marchal, G. (1995). **Automated multi-modality image registration based on information theory**. In *Information processing in medical imaging*, volume 3, pages 263–274. Citeseer.

De Fauw, J., Ledsam, J. R., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., Askham, H., Glorot, X., O'Donoghue, B., Visentin, D., van den Driessche, G., Lakshminarayanan, B., Meyer, C., Mackinder, F., Bouton, S., Ayoub, K., Chopra, R., King, D., Karthikesalingam, A., Hughes, C. O., Raine, R., Hughes, J., Sim, D. A., Egan, C., Tufail, A., Montgomery, H., Hassabis, D., Rees, G., Back, T., Khaw, P. T., Suleyman, M.,

Cornebise, J., Keane, P. A., and Ronneberger, O. (2018). **Clinically applicable deep learning for diagnosis and referral in retinal disease**. *Nature medicine*, 24(9):1342–1350.

De Rooij, M., Allen, C., Twilt, J. J., Thijssen, L. C., Asbach, P., Barrett, T., Brembilla, G., Emberton, M., Gupta, R. T., Haider, M. A., Kasivisvanathan, V., Løgager, V., Moore, C. M., Padhani, A. R., Panebianco, V., Puech, P., Purysko, A. S., Renard-Penna, R., Richenberg, J., Salomon, G., Sanguedolce, F., Schoots, I. G., Thöny, H. C., Turkbey, B., Villeirs, G., Walz, J., Barentsz, J., and Giganti, F. (2024). **PI-QUAL version 2: an update of a standardised scoring system for the assessment of image quality of prostate MRI**. *European Radiology*, pages 1–12.

De Vente, C., Vos, P., Hosseinzadeh, M., Pluim, J., and Veta, M. (2020). **Deep learning regression for prostate cancer detection and grading in bi-parametric MRI**. *IEEE Transactions on Biomedical Engineering*, 68(2):374–383. doi: 10.1109/TBME.2020.2993528.

DeLong, E. R., DeLong, D. M., and Clarke-Pearson, D. L. (1988). **Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach**. *Biometrics*, pages 837–845.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). **Imagenet: A large-scale hierarchical image database**. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Deng, L. (2012). **The mnist database of handwritten digit images for machine learning research [best of the web]**. *IEEE signal processing magazine*, 29(6):141–142.

Diederik, P. K. and Jimmy, B. (2014). **Adam: A method for stochastic optimization. iclr**. *arXiv preprint arXiv:1412.6980*, 5. unpublished findings.

Dietrich, O., Biffar, A., Baur-Melnyk, A., and Reiser, M. F. (2010). **Technical aspects of MR diffusion imaging of the body**. *European journal of radiology*, 76(3):314–322.

Duran, A., Dussert, G., Rouvière, O., Jaouen, T., Jodoin, P.-M., and Lartizien, C. (2022). **Prostattention-net: A deep attention model for prostate cancer segmentation by aggressiveness in MRI scans**. *Medical Image Analysis*, 77:102347.

Elmohr, M., Elsayes, K. M., and Chernyak, V. (2021). **Li-rads: review and updates**. *Clinical Liver Disease*, 17(3):108–112.

Engels, R. R., Israël, B., Padhani, A. R., and Barentsz, J. O. (2020). **Multiparametric magnetic resonance imaging for the detection of clinically significant prostate cancer: what urologists need to know. part 1: acquisition**. *European urology*, 77(4):457–468.

Epstein, J. I. (2010). **An update of the gleason grading system**. *The Journal of urology*, 183 (2):433–440.

Epstein, J. I., Allsbrook Jr, W. C., Amin, M. B., Egevad, L. L., and ISUP Grading Committee. (2005). **The 2005 international society of urological pathology (ISUP) consensus conference on Gleason grading of prostatic carcinoma**. *The American journal of surgical pathology*, 29(9):1228–1242.

Epstein, J. I., Egevad, L., Amin, M. B., Delahunt, B., Srigley, J. R., Humphrey, P. A., and Grading Committee. (2016)a. **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**. *The American journal of surgical pathology*, 40(2):244–252.

Epstein, J. I., Zelefsky, M. J., Sjoberg, D. D., Nelson, J. B., Egevad, L., Magi-Galluzzi, C., Vickers, A. J., Parwani, A. V., Reuter, V. E., Fine, S. W., Eastham, J. A., Wiklund, P., Han, M., Reddy, C. A., Ciezki, J. P., Nyberg, T., and Klein, E. A. (2016)b. **A contemporary prostate cancer grading system: a validated alternative to the Gleason score**. *European urology*, 69(3):428–435.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., and Thrun, S. (2017). **Dermatologist-level classification of skin cancer with deep neural networks**. *nature*, 542(7639):115–118.

Frankenstein, Z., Basanta, D., Franco, O. E., Gao, Y., Javier, R. A., Strand, D. W., Lee, M., Hayward, S. W., Ayala, G., and Anderson, A. R. (2020). **Stromal reactivity differentially drives tumour cell evolution and prostate cancer progression**. *Nature ecology & evolution*, 4(6):870–884.

Fu, Y., Lei, Y., Wang, T., Curran, W. J., Liu, T., and Yang, X. (2020). **Deep learning in medical image registration: a review**. *Physics in Medicine & Biology*, 65(20):20TR01. doi: 10.1088/1361-6560/ab843e.

Galbán, C. J., Chenevert, T. L., Meyer, C. R., Tsien, C., Lawrence, T. S., Hamstra, D. A., Junck, L., Sundgren, P. C., Johnson, T. D., Ross, D. J., Rehemtulla, A., and Ross, B. D. (2009). **The parametric response map is an imaging biomarker for early cancer treatment outcome**. *Nature medicine*, 15(5):572–576.

Gatidis, S. and Kuestner, T. (2022). **A whole-body FDG-PET/CT dataset with manually annotated tumor lesions (FDG-PET-CT-lesions)**. The Cancer Imaging Archive.

Gatti, M., Faletti, R., Calleris, G., Giglio, J., Berzovini, C., Gentile, F., Marra, G., Misischi, F., Molinaro, L., Bergamasco, L., Gontero, P., Papotti, M., and Fonio, P. (2019). **Prostate cancer detection with biparametric magnetic resonance imaging (bpMRI) by readers with different experience: performance and comparison with multiparametric (mpMRI)**. *Abdominal Radiology*, 44:1883–1893.

Ghai, S., Finelli, A., Corr, K., Lajkosz, K., McCluskey, S., Chan, R., Gertner, M., van der Kwast, T. H., Incze, P. F., Zlotta, A. R., Kucharczyk, W., and Perlis, N. (2024). **MRI-guided focused ultrasound focal therapy for intermediate-risk prostate cancer: final results from a 2-year phase ii clinical trial**. *Radiology*, 310(3):e231473.

Gibbs, P., Liney, G. P., Pickles, M. D., Zelhof, B., Rodrigues, G., and Turnbull, L. W. (2009). **Correlation of adc and t2 measurements with cell density in prostate cancer at 3.0 tesla**. *Investigative radiology*, 44(9):572–576.

Giganti, F., Allen, C., Emberton, M., Moore, C. M., Kasivisvanathan, V., and PRECISION Study Group. (2020). **Prostate imaging quality (PI-QUAL): a new quality control scoring system for multiparametric magnetic resonance imaging of the prostate from the PRECISION trial**. *European urology oncology*, 3(5):615–619.

Giganti, F., Kasivisvanathan, V., Kirkham, A., Punwani, S., Emberton, M., Moore, C. M., and Allen, C. (2022). **Prostate MRI quality: a critical review of the last 5 years and the role of the PI-QUAL score**. *The British Journal of Radiology*, 95(1131):20210415.

Gleason, D. F. and Mellinger, G. T. (1974). **Prediction of prognosis for prostatic adenocarcinoma by combined histological grading and clinical staging**. *The Journal of urology*, 111(1):58–64.

Hagmann, P., Jonasson, L., Maeder, P., Thiran, J.-P., Wedeen, V. J., and Meuli, R. (2006). **Understanding diffusion MR imaging techniques: from scalar diffusion-weighted imaging to diffusion tensor imaging and beyond**. *Radiographics*, 26(suppl_1):S205–S223.

Hamm, C. A., Baumgärtner, G. L., Biessmann, F., Beetz, N. L., Hartenstein, A., Savic, L. J., Froböse, K., Dräger, F., Schallenberg, S., Rudolph, M., Baur, A. D. J., Hamm, B., Haas, M., Hofbauer, S., Cash, H., and Penzkofer, T. (2023). **Interactive explainable deep learning model informs prostate cancer diagnosis at MRI**. *Radiology*, 307(4):e222276.

Hanley, J. A. and McNeil, B. J. (1982). **The meaning and use of the area under a receiver operating characteristic (ROC) curve.** *Radiology*, 143(1):29–36.

Haskins, G., Kruger, U., and Yan, P. (2020). **Deep learning in medical image registration: a survey**. *Machine Vision and Applications*, 31(1):1–18. doi: 10.1007/s00138-020-01060-x.

He, J., Albertsen, P. C., Moore, D., Rotter, D., Demissie, K., and Lu-Yao, G. (2017). **Validation of a contemporary five-tiered gleason grade grouping using population-based data**. *European urology*, 71(5):760–763.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. (2022). **Masked autoencoders are scalable vision learners**. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009.

He, Y., Nath, V., Yang, D., Tang, Y., Myronenko, A., and Xu, D. (2023). **Swinunetr-v2: Stronger swin transformers with stagewise convolutions for 3D medical image segmentation**. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 416–426. Springer.

Hensel, J. M., Ménard, C., Chung, P. W., Milosevic, M. F., Kirilova, A., Moseley, J. L., Haider, M. A., and Brock, K. K. (2007). **Development of multiorgan finite element-based prostate deformation model enabling registration of endorectal coil magnetic resonance imaging for radiotherapy planning**. *International Journal of Radiation Oncology* Biology* Physics*, 68(5):1522–1528.

Hering, A., Peisen, F., Amaral, T., Gatidis, S., Eigentler, T., Othman, A., and Moltz, J. H. (2021). **Whole-body soft-tissue lesion tracking and segmentation in longitudinal CT imaging studies**. In *Medical Imaging with Deep Learning*, pages 312–326. PMLR.

Hill, D. L., Batchelor, P. G., Holden, M., and Hawkes, D. J. (2001). **Medical image registration**. *Physics in medicine & biology*, 46(3):R1.

Hosseinzadeh, M., Saha, A., Brand, P., Slootweg, I., de Rooij, M., and Huisman, H. (2021). **Deep learning–assisted prostate cancer detection on bi-parametric MRI: minimum training data size requirements and effect of prior knowledge**. *European Radiology*, pages 1–11. doi: 10.1007/s00330-021-08320-y.

Hötker, A. M., Vargas, H. A., and Donati, O. F. (2022). **Abbreviated MR protocols in prostate MRI**. *Life*, 12(4):552.

Hu, Y., Morgan, D., Ahmed, H. U., Pendsé, D., Sahu, M., Allen, C., Emberton, M., Hawkes, D., and Barratt, D. (2008). **A statistical motion model based on biomechanical simulations for data fusion during image-guided prostate interventions**. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2008: 11th International*

Conference, New York, NY, USA, September 6-10, 2008, Proceedings, Part I 11, pages 737–744. Springer.

Hu, Y., van den Boom, R., Carter, T., Taylor, Z., Hawkes, D., Ahmed, H. U., Emberton, M., Allen, C., and Barratt, D. (2010). **A comparison of the accuracy of statistical models of prostate motion trained using data from biomechanical simulations**. *Progress in biophysics and molecular biology*, 103(2-3):262–272.

Hu, Y., Gibson, E., Ghavami, N., Bonmati, E., Moore, C. M., Emberton, M., Vercauteren, T., Noble, J. A., and Barratt, D. C. (2018). **Adversarial deformation regularization for training image registration neural networks**. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I*, pages 774–782. Springer.

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). **Densely connected convolutional networks**. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708.

Hugosson, J., Månsson, M., Wallström, J., Axcrona, U., Carlsson, S. V., Egevad, L., Geterud, K., Khatami, A., Kohestani, K., Pihl, C.-G., Socratous, A., Stranne, J., Godtman, R. A., and Hellström, M. (2022). **Prostate cancer screening with PSA and MRI followed by targeted biopsy only**. *New England Journal of Medicine*, 387(23):2126–2137.

Humphrey, P. A. (2004). **Gleason grading and prognostic factors in carcinoma of the prostate**. *Modern pathology*, 17(3):292–306.

Hunter, C. R., Klein, R., Beanlands, R. S., and deKemp, R. A. (2016). **Patient motion effects on the quantification of regional myocardial blood flow with dynamic pet imaging**. *Medical physics*, 43(4):1829–1840.

Iglesias, J. E. and Sabuncu, M. R. (2015). **Multi-atlas segmentation of biomedical images: a survey**. *Medical image analysis*, 24(1):205–219.

Ilic, D., Djulbegovic, M., Jung, J. H., Hwang, E. C., Zhou, Q., Cleves, A., Agoritsas, T., and Dahm, P. (2018). **Prostate cancer screening with prostate-specific antigen (PSA) test: a systematic review and meta-analysis**. *bmj*, 362.

Immerzeel, J., Israël, B., Bomers, J., Schoots, I. G., Van Basten, J.-P., Kurth, K.-H., de Reijke, T., Sedelaar, M., Debruyne, F., and Barentsz, J. (2022). **Multiparametric magnetic resonance imaging for the detection of clinically significant prostate cancer: what urologists need to know. part 4: transperineal magnetic resonance–ultrasound fusion guided biopsy using local anesthesia**. *European Urology*, 81(1):110–117.

Ingrisch, M., Dexl, J., Jeblick, K., Cyran, C., Gatidis, S., and Kuestner, T. (2024). **Automated lesion segmentation in whole-body PET/CT-multitracer multicenter generalization**. In *27th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI 2024).*, volume 10990932. doi: 10.5281/zenodo.

Isensee, F., Jäger, P., Wasserthal, J., Zimmerer, D., Petersen, J., Kohl, S., Schock, J., Klein, A., Roß, T., Wirkert, S., Neher, P., Dinkelacker, S., Köhler, G., and Maier-Hein, K. *Batchgenerators - a python framework for data augmentation*, (2020). URL `https://github.com/MIC-DKFZ/batchgenerators`. [visited on 15.06.2025].

Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J., and Maier-Hein, K. H. (2021). **nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation**. *Nature methods*, 18(2):203–211.

Isensee, F., Wald, T., Ulrich, C., Baumgartner, M., Roy, S., Maier-Hein, K., and Jaeger, P. F. (2024). **nnU-Net revisited: A call for rigorous validation in 3D medical image segmentation**. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 488–498. Springer.

Israël, B., van der Leest, M., Sedelaar, M., Padhani, A. R., Zamecnik, P., and Barentsz, J. O. (2020). **Multiparametric magnetic resonance imaging for the detection of clinically significant prostate cancer: what urologists need to know. part 2: interpretation**. *European urology*, 77(4):469–480.

Jeblick, K. (2024). **A whole-body PSMA-PET/CT dataset with manually annotated tumor lesions (PSMA-PET-CT-lesions)**. The Cancer Imaging Archive.

Johnsen, S. F., Taylor, Z. A., Clarkson, M. J., Hipwell, J., Modat, M., Eiben, B., Han, L., Hu, Y., Mertzanidou, T., Hawkes, D. J., and Ourselin, S. (2015). **NiftySim: A GPU-based nonlinear finite element package for simulation of soft tissue biomechanics**. *International journal of computer assisted radiology and surgery*, 10:1077–1095.

Kaji, T., Osanai, K., Takahashi, A., Kinoshita, A., Satoh, D., Nakata, T., and Tamaki, N. (2024). **Improvement of motion artifacts using dynamic whole-body 18F-FDG PET/CT imaging**. *Japanese Journal of Radiology*, 42(4):374–381.

Kasabwala, K., Patel, N., Cricco-Lizza, E., Shimpi, A. A., Weng, S., Buchmann, R. M., Motanagh, S., Wu, Y., Banerjee, S., Khani, F., Margolis, D. J., Robinson, B. D., and Hu, J. C. (2019). **The learning curve for magnetic resonance imaging/ultrasound fusion-guided prostate biopsy**. *European Urology Oncology*, 2(2):135–140.

Kasel-Seibert, M., Lehmann, T., Aschenbach, R., Guettler, F. V., Abubrig, M., Grimm, M.-O., Teichgraeber, U., and Franiel, T. (2016). **Assessment of PI-RADS v2 for the detection of prostate cancer**. *European journal of radiology*, 85(4):726–731.

Kerkmeijer, L. G. W., Groen, V. H., Pos, F. J., Haustermans, K., Monninkhof, E. M., Smeenk, R. J., Kunze-Busch, M., de Boer, J. C. J., van der Voort van Zijp, J., van Vulpen, M., Draulans, C., van den Bergh, L., Isebaert, S., and van der Heide, U. A. (2021). **Focal boost to the intraprostatic tumor in external beam radiotherapy for patients with localized prostate cancer: results from the FLAME randomized phase III trial**. *Journal of Clinical Oncology*, 39(7):787–796.

Khallaghi, S., Sánchez, C. A., Rasoulian, A., Nouranian, S., Romagnoli, C., Abdi, H., Chang, S. D., Black, P. C., Goldenberg, L., Morris, W. J., Spadinger, I., Fenster, A., Ward, A., Fels, S., and Abolmaesumi, P. (2015)a. **Statistical biomechanical surface registration: application to MR-TRUS fusion for prostate interventions**. *IEEE transactions on medical imaging*, 34(12):2535–2549.

Khallaghi, S., Sánchez, C. A., Rasoulian, A., Sun, Y., Imani, F., Khojaste, A., Goksel, O., Romagnoli, C., Abdi, H., Chang, S., Mousavi, P., Fenster, A., Ward, A., Fels, S., and Abolmaesumi, P. (2015)b. **Biomechanically constrained surface registration: Application to MR-TRUS fusion for prostate interventions**. *IEEE transactions on medical imaging*, 34(11):2404–2414.

Kirchhoff, Y., Rokuss, M. R., Roy, S., Kovacs, B., Ulrich, C., Wald, T., Zenk, M., Vollmuth, P., Kleesiek, J., Isensee, F., and Maier-Hein, K. (2024). **Skeleton recall loss for connectivity conserving and resource efficient segmentation of thin tubular structures**. In *European Conference on Computer Vision*, pages 218–234. Springer.

Kirimtat, A., Krejcar, O., and Selamat, A. (2020). **Brain MRI modality understanding: A guide for image processing and segmentation**. In *Bioinformatics and Biomedical Engineering: 8th International Work-Conference, IWBBIO 2020, Granada, Spain, May 6–8, 2020, Proceedings 8*, pages 705–715. Springer.

Klein, S., Staring, M., and Pluim, J. P. (2007). **Evaluation of optimization methods for nonrigid medical image registration using mutual information and b-splines**. *IEEE transactions on image processing*, 16(12):2879–2890.

Kohl, S., Bonekamp, D., Schlemmer, H.-P., Yaqubi, K., Hohenfellner, M., Hadaschik, B., Radtke, J.-P., and Maier-Hein, K. (2017). **Adversarial networks for the detection of aggressive prostate cancer**. In *Workshop on Machine Learning for Health (NIPS ML4H 2017)*, pages –.

Kono, K., Inoue, Y., Nakayama, K., Shakudo, M., Morino, M., Ohata, K., Wakasa, K., and Yamada, R. (2001). **The role of diffusion-weighted imaging in patients with brain tumors**. *American journal of neuroradiology*, 22(6):1081–1088.

Kovacs, B., Netzer, N., Baumgartner, M., Eith, C., Bounias, D., Meinzer, C., Jäger, P. F., Zhang, K. S., Floca, R., Schrader, A., Isensee, F., Gnirs, R., Görtz, M., Schütz, V., Stenzinger, A., Hohenfellner, M., Schlemmer, H.-P., Wolf, I., Bonekamp, D., and Maier-Hein, K. H. (2023)a. **Anatomy-informed data augmentation for enhanced prostate cancer detection**. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 531–540. Springer.

Kovacs, B., Netzer, N., Baumgartner, M., Schrader, A., Isensee, F., Weißer, C., Wolf, I., Görtz, M., Jaeger, P. F., Schütz, V., Floca, R., Gnirs, R., Stenzinger, A., Hohenfellner, M., Schlemmer, H.-P., Bonekamp, D., and Maier-Hein, K. H. (2023)b. **Addressing image misalignments in multi-parametric prostate MRI for enhanced computer-aided diagnosis of prostate cancer**. *Scientific Reports*, 13(1):19805.

Kovacs, B., Xiao, S., Rokuss, M., Ulrich, C., Isensee, F., and Maier-Hein, K. H. (2024). **Data-centric strategies for overcoming PET/CT heterogeneity: Insights from the autopet iii lesion segmentation challenge**. *arXiv preprint arXiv:2409.10120*. unpublished findings.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). **Imagenet classification with deep convolutional neural networks**. *Advances in neural information processing systems*, 25.

Kuhl, C. K., Bruhn, R., Krämer, N., Nebelung, S., Heidenreich, A., and Schrading, S. (2017). **Abbreviated biparametric prostate MR imaging in men with elevated prostate-specific antigen**. *Radiology*, 285(2):493–505.

Kuru, T. H., Wadhwa, K., Chang, R. T. M., Echeverria, L. M. C., Roethke, M., Polson, A., Rottenberg, G., Koo, B., Lawrence, E. M., Seidenader, J., Gnanapragasam, V., Axell, R., Roth, W., Warren, A., Doble, A., Muir, G., Popert, R., Schlemmer, H.-P., Hadaschik, B. A., and Kastner, C. (2013). **Definitions of terms, processes and a minimum dataset for transperineal prostate biopsies: a standardization approach of the Ginsburg study group for enhanced prostate diagnostics.** *BJU international*, 112(5).

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. (1989). **Backpropagation applied to handwritten zip code recognition**. *Neural computation*, 1(4):541–551.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). **Gradient-based learning applied to document recognition**. *Proceedings of the IEEE*, 86(11):2278–2324.

Lester, H. and Arridge, S. R. (1999). **A survey of hierarchical non-linear medical image registration**. *Pattern recognition*, 32(1):129–149.

Liang, Z., Hu, R., Yang, Y., An, N., Duo, X., Liu, Z., Shi, S., and Liu, X. (2020). **Is dynamic contrast enhancement still necessary in multiparametric magnetic resonance for diagnosis of prostate cancer: a systematic review and meta-analysis**. *Translational Andrology and Urology*, 9(2):553.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). **Microsoft coco: Common objects in context**. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. (2023). **Visual instruction tuning**. *Advances in neural information processing systems*, 36:34892–34916.

Loeb, S., Vellekoop, A., Ahmed, H. U., Catto, J., Emberton, M., Nam, R., Rosario, D. J., Scattoni, V., and Lotan, Y. (2013). **Systematic review of complications of prostate biopsy**. *European urology*, 64(6):876–892.

Lojanapiwat, B., Anutrakulchai, W., Chongruksut, W., and Udomphot, C. (2014). **Correlation and diagnostic performance of the prostate-specific antigen level with the diagnosis, aggressiveness, and bone metastasis of prostate cancer in clinical practice**. *Prostate international*, 2(3):133–139.

Lomas, D. J. and Ahmed, H. U. (2020). **All change in the prostate cancer diagnostic pathway**. *Nature reviews Clinical oncology*, 17(6):372–381.

Ma, J. (2021). **Cutting-edge 3D medical image segmentation methods in 2020: Are happy families all alike?** *arXiv preprint arXiv:2101.00232*. unpublished findings.

Ma, J., He, Y., Li, F., Han, L., You, C., and Wang, B. (2024)a. **Segment anything in medical images**. *Nature Communications*, 15(1):654.

Ma, J., Li, F., and Wang, B. (2024)b. **U-mamba: Enhancing long-range dependency for biomedical image segmentation**. *arXiv preprint arXiv:2401.04722*. unpublished findings.

Maier-Hein, L., Eisenmann, M., Reinke, A., Onogur, S., Stankovic, M., Scholz, P., Arbel, T., Bogunovic, H., Bradley, A. P., Carass, A., Feldmann, C., Frangi, A. F., Full, P. M., van Ginneken, B., Hanbury, A., Honauer, K., Kozubek, M., Landman, B. A., März, K., Maier, O., Maier-Hein, K., Menze, B. H., Müller, H., Neher, P. F., Niessen, W., Rajpoot, N., Sharp, G. C., Sirinukunwattana, K., Speidel, S., Stock, C., Stoyanov, D., Taha, A. A., van der

Sommen, F., Wang, C.-W., Weber, M.-A., Zheng, G., Jannin, P., and Kopp-Schneider, A. (2018). **Why rankings of biomedical image analysis competitions should be interpreted with care**. *Nature communications*, 9(1):5217.

Majkowska, A., Mittal, S., Steiner, D. F., Reicher, J. J., McKinney, S. M., Duggan, G. E., Eswaran, K., Cameron Chen, P.-H., Liu, Y., Kalidindi, S. R., Ding, A., Corrado, G. S., Tse, D., and Shetty, S. (2020). **Chest radiograph interpretation with deep learning models: assessment with radiologist-adjudicated reference standards and population-adjusted evaluation**. *Radiology*, 294(2):421–431.

Marenco, J., Orczyk, C., Collins, T., Moore, C., and Emberton, M. (2019). **Role of MRI in planning radical prostatectomy: what is the added value?** *World Journal of Urology*, 37: 1289–1292.

Markl, M. and Leupold, J. (2012). **Gradient echo imaging**. *Journal of Magnetic Resonance Imaging*, 35(6):1274–1289.

McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G. S., Darzi, A., Etemadi, M., Garcia-Vicente, F., Gilbert, F. J., Halling-Brown, M., Hassabis, D., Jansen, S., Karthikesalingam, A., Kelly, C. J., King, D., Ledsam, J. R., Melnick, D., Mostofi, H., Peng, L., Reicher, J. J., Romera-Paredes, B., Sidebottom, R., Suleyman, M., Tse, D., Young, K. C., De Fauw, J., and Shetty, S. (2020). **International evaluation of an AI system for breast cancer screening**. *Nature*, 577 (7788):89–94.

Meng, X., Rosenkrantz, A. B., Huang, R., Deng, F.-M., Wysock, J. S., Bjurlin, M. A., Huang, W. C., Lepor, H., and Taneja, S. S. (2018). **The institutional learning curve of magnetic resonance imaging-ultrasound fusion targeted prostate biopsy: temporal improvements in cancer detection in 4 years**. *The Journal of urology*, 200(5):1022–1029.

Miah, S., Eldred-Evans, D., Simmons, L. A., Shah, T. T., Kanthabalan, A., Arya, M., Winkler, M., McCartan, N., Freeman, A., Punwani, S., Moore, C. M., Emberton, M., and Ahmed, H. U. (2018). **Patient reported outcome measures for transperineal template prostate mapping biopsies in the PICTURE study**. *The Journal of Urology*, 200(6):1235–1240.

Mitchell, D. G., Bruix, J., Sherman, M., and Sirlin, C. B. (2015). **Li-rads (liver imaging reporting and data system): summary, discussion, and consensus of the li-rads management working group and future directions**. *Hepatology*, 61(3):1056–1065.

Monda, S. M., Vetter, J. M., Andriole, G. L., Fowler, K. J., Shetty, A. S., Weese, J. R., and Kim, E. H. (2018). **Cognitive versus software fusion for MRI-targeted biopsy: experience before and after implementation of fusion**. *Urology*, 119:115–120.

Moraal, B., Meier, D. S., Poppe, P. A., Geurts, J. J. G., Vrenken, H., Jonker, W. M. A., Knol, D. L., van Schijndel, R. A., Pouwels, P. J. W., Pohl, C., Bauer, L., Sandbrink, R., Guttmann, C. R. G., and Barkhof, F. (2009). **Subtraction MR images in a multiple sclerosis multicenter clinical trial setting**. *Radiology*, 250(2):506–514.

Moussa, A. S., Meshref, A., Schoenfield, L., Masoud, A., Abdel-Rahman, S., Li, J., Flazoura, S., Magi-Galluzzi, C., Fergany, A., Fareed, K., and Jones, J. S. (2010). **Importance of additional "extreme" anterior apical needle biopsies in the initial detection of prostate cancer**. *Urology*, 75(5):1034–1039.

Naz, Z., Khan, M. U. G., Saba, T., Rehman, A., Nobanee, H., and Bahaj, S. A. (2023). **An explainable AI-enabled framework for interpreting pulmonary diseases from chest radiographs**. *Cancers*, 15(1):314.

Netzer, N., Weißer, C., Schelb, P., Wang, X., Qin, X., Görtz, M., Schütz, V., Radtke, J. P., Hielscher, T., Schwab, C., Stenzinger, A., Kuder, T. A., Gnirs, R., Hohenfellner, M., Schlemmer, H.-P., Maier-Hein, K. H., and Bonekamp, D. (2021). **Fully automatic deep learning in bi-institutional prostate magnetic resonance imaging: Effects of cohort size and heterogeneity**. *Investigative radiology*, 56(12):799–808.

Netzer, N., Eith, C., Bethge, O., Hielscher, T., Schwab, C., Stenzinger, A., Gnirs, R., Schlemmer, H.-P., Maier-Hein, K. H., Schimmöller, L., and Bonekamp, D. (2023). **Application of a validated prostate MRI deep learning system to independent same-vendor multi-institutional data: demonstration of transferability**. *European Radiology*, 33(11): 7463–7476.

Nikolov, S., Blackwell, S., Zverovitch, A., Mendes, R., Livne, M., De Fauw, J., Patel, Y., Meyer, C., Askham, H., Romera-Paredes, B., Kelly, C., Karthikesalingam, A., Chu, C., Carnell, D., Boon, C., D'Souza, D., Moinuddin, S. A., Garie, B., McQuinlan, Y., Ireland, S., Hampton, K., Fuller, K., Montgomery, H., Rees, G., Suleyman, M., Back, T., Hughes, C. O., Ledsam, J. R., and Ronneberger, O. (2021). **Clinically applicable segmentation of head and neck anatomy for radiotherapy: Deep learning algorithm development and validation study**. *Journal of Medical Internet Research*, 23(7):e26151. doi: 10.2196/26151.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.-Y., Li, S.-W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., and Bojanowski, P. (2023). **Dinov2: Learning robust visual features without supervision**. *arXiv preprint arXiv:2304.07193*. unpublished findings.

Ouzzane, A., Puech, P., Lemaitre, L., Leroy, X., Nevoux, P., Betrouni, N., Haber, G.-P., and Villers, A. (2011). **Combined multiparametric MRI and targeted biopsies improve anterior prostate cancer detection, staging, and grading**. *Urology*, 78(6):1356–1362.

Panebianco, V., Giganti, F., Kitzing, Y. X., Cornud, F., Campa, R., De Rubeis, G., Ciardi, A., Catalano, C., and Villeirs, G. (2018). **An update of pitfalls in prostate mpMRI: a practical approach through the lens of PI-RADS v. 2 guidelines**. *Insights into imaging*, 9:87–101.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. (2019). **PyTorch: An imperative style, high-performance deep learning library**. *Advances in Neural Information Processing Systems*, 32.

Pati, S., Baid, U., Edwards, B., Sheller, M., Wang, S.-H., Reina, G. A., Foley, P., Gruzdev, A., Karkada, D., Davatzikos, C., Sako, C., Ghodasara, S., Bilello, M., Mohan, S., Vollmuth, P., Brugnara, G., Preetha, C. J., Sahm, F., Maier-Hein, K., Zenk, M., Bendszus, M., Wick, W., Calabrese, E., Rudie, J., Villanueva-Meyer, J., Cha, S., Ingalhalikar, M., Jadhav, M., Pandey, U., Saini, J., Garrett, J., Larson, M., Jeraj, R., Currie, S., Frood, R., Fatania, K., Huang, R. Y., Chang, K., Quintero, C. B., Capellades, J., Puig, J., Trenkler, J., Pichler, J., Necker, G., Haunschmidt, A., Meckel, S., Shukla, G., Liem, S., Alexander, G. S., Lombardo, J., Palmer, J. D., Flanders, A. E., Dicker, A. P., Sair, H. I., Jones, C. K., Venkataraman, A., Jiang, M., So, T. Y., Chen, C., Heng, P. A., Dou, Q., Kozubek, M., Lux, F., Michálek, J., Matula, P., Keřkovský, M., Kopřivová, T., Dostál, M., Vybíhal, V., Vogelbaum, M. A., Mitchell, J. R., Farinhas, J., Maldjian, J. A., Yogananda, C. G. B., Pinho, M. C., Reddy, D., Holcomb, J., Wagner, B. C., Ellingson, B. M., Cloughesy, T. F., Raymond, C., Oughourlian, T., Hagiwara, A., Wang, C., To, M.-S., Bhardwaj, S., Chong, C., Agzarian, M., Falcão, A. X., Martins, S. B., Teixeira, B. C. A., Sprenger, F., Menotti, D., Lucio, D. R., LaMontagne, P., Marcus, D., Wiestler, B., Kofler, F., Ezhov, I., Metz, M., Jain, R., Lee, M., Lui, Y. W., McKinley, R., Slotboom, J., Radojewski, P., Meier, R., Wiest, R., Murcia, D., Fu, E., Haas, R., Thompson, J., Ormond, D. R., Badve, C., Sloan, A. E., Vadmal, V., Waite, K., Colen, R. R., Pei, L., Ak, M., Srinivasan, A., Bapuraj, J. R., Rao, A., Wang, N., Yoshiaki, O., Moritani, T., Turk, S., Lee, J., Prabhudesai, S., Morón, F., Mandel, J., Kamnitsas, K., Glocker, B., Dixon, L. V. M., Williams, M., Zampakis, P., Panagiotopoulos, V., Tsiganos, P., Alexiou, S., Haliassos, I., Zacharaki, E. I., Moustakas, K., Kalogeropoulou, C., Kardamakis, D. M., Choi, Y. S., Lee, S.-K., Chang, J. H., Ahn, S. S., Luo, B., Poisson, L., Wen, N., Tiwari, P., Verma, R., Bareja, R., Yadav, I., Chen, J., Kumar, N., Smits, M., van der Voort, S. R., Alafandi, A., Incekara, F., Wijnenga, M.

M. J., Kapsas, G., Gahrmann, R., Schouten, J. W., Dubbink, H. J., Vincent, A. J. P. E., van den Bent, M. J., French, P. J., Klein, S., Yuan, Y., Sharma, S., Tseng, T.-C., Adabi, S., Niclou, S. P., Keunen, O., Hau, A.-C., Vallières, M., Fortin, D., Lepage, M., Landman, B., Ramadass, K., Xu, K., Chotai, S., Chambless, L. B., Mistry, A., Thompson, R. C., Gusev, Y., Bhuvaneshwar, K., Sayah, A., Bencheqroun, C., Belouali, A., Madhavan, S., Booth, T. C., Chelliah, A., Modat, M., Shuaib, H., Dragos, C., Abayazeed, A., Kolodziej, K., Hill, M., Abbassy, A., Gamal, S., Mekhaimar, M., Qayati, M., Reyes, M., Park, J. E., Yun, J., Kim, H. S., Mahajan, A., Muzi, M., Benson, S., Beets-Tan, R. G. H., Teuwen, J., Herrera-Trujillo, A., Trujillo, M., Escobar, W., Abello, A., Bernal, J., Gómez, J., Choi, J., Baek, S., Kim, Y., Ismael, H., Allen, B., Buatti, J. M., Kotrotsou, A., Li, H., Weiss, T., Weller, M., Bink, A., Pouymayou, B., Shaykh, H. F., Saltz, J., Prasanna, P., Shrestha, S., Mani, K. M., Payne, D., Kurc, T., Pelaez, E., Franco-Maldonado, H., Loayza, F., Quevedo, S., Guevara, P., Torche, E., Mendoza, C., Vera, F., Ríos, E., López, E., Velastin, S. A., Ogbole, G., Soneye, M., Oyekunle, D., Odafe-Oyibotha, O., Osobu, B., Shu'aibu, M., Dorcas, A., Dako, F., Simpson, A. L., Hamghalam, M., Peoples, J. J., Hu, R., Tran, A., Cutler, D., Moraes, F. Y., Boss, M. A., Gimpel, J., Veettil, D. K., Schmidt, K., Bialecki, B., Marella, S., Price, C., Cimino, L., Apgar, C., Shah, P., Menze, B., Barnholtz-Sloan, J. S., Martin, J., and Bakas, S. (2022). **Federated learning enables big data for rare cancer boundary detection**. *Nature communications*, 13(1):7346.

Payan, Y. and Ohayon, J., editors. (2017). *Biomechanics of living organs: hyperelastic constitutive laws for finite element modeling*. Academic Press, Oxford.

Pellicer-Valero, O. J., Marenco Jimenez, J. L., Gonzalez-Perez, V., Casanova Ramon-Borja, J. L., Martín García, I., Barrios Benito, M., Pelechano Gomez, P., Rubio-Briones, J., Rupérez, M. J., and Martín-Guerrero, J. D. (2022). **Deep learning for fully automatic detection, segmentation, and gleason grade estimation of prostate cancer in multiparametric magnetic resonance images**. *Scientific reports*, 12(1):2975.

Penzkofer, T. (2024). **Prostate-MRI reporting should be done with the aid of AI systems: Pros**. *European Radiology*, 34(12):7728–7730.

Perez, F., Vasconcelos, C., Avila, S., and Valle, E. (2018). **Data augmentation for skin lesion analysis**. In *OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis: First International Workshop, OR 2.0 2018, 5th International Workshop, CARE 2018, 7th International Workshop, CLIP 2018, Third International Workshop, ISIC 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16 and 20, 2018, Proceedings 5*, pages 303–311. Springer.

Pinto, P. A., Chung, P. H., Rastinehad, A. R., Baccala, A. A., Kruecker, J., Benjamin, C. J., Xu, S., Yan, P., Kadoury, S., Chua, C., Locklin, J. K., Turkbey, B., Shih, J. H., Gates, S. P.,

Buckner, C., Bratslavsky, G., Linehan, W. M., Glossop, N. D., Choyke, P. L., and Wood, B. J. (2011). **Magnetic resonance imaging/ultrasound fusion guided prostate biopsy improves cancer detection following transrectal ultrasound biopsy and correlates with multiparametric magnetic resonance imaging**. *The Journal of urology*, 186(4): 1281–1285.

Plodeck, V., Radosa, C. G., Hübner, H.-M., Baldus, C., Borkowetz, A., Thomas, C., Kühn, J.-P., Laniado, M., Hoffmann, R.-T., and Platzek, I. (2020). **Rectal gas-induced susceptibility artefacts on prostate diffusion-weighted MRI with epi read-out at 3.0 t: does a preparatory micro-enema improve image quality?** *Abdominal Radiology*, 45:4244–4251.

Puech, P., Rouvière, O., Renard-Penna, R., Villers, A., Devos, P., Colombel, M., Bitker, M.-O., Leroy, X., Mège-Lechevallier, F., Comperat, E., Ouzzane, A., and Lemaitre, L. (2013). **Prostate cancer diagnosis: multiparametric MR-targeted biopsy with cognitive and transrectal US–MR fusion guidance versus systematic biopsy—prospective multicenter study**. *Radiology*, 268(2):461–469.

Qasim, M., Puigjaner, D., Herrero, J., López, J. M., Olivé, C., Fortuny, G., and Garcia-Bennett, J. (2022). **Biomechanical modelling of the pelvic system: improving the accuracy of the location of neoplasms in MRI-TRUS fusion prostate biopsy**. *BMC cancer*, 22(1):338.

Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., Bakas, S., Galtier, M. N., Landman, B. A., Maier-Hein, K., Ourselin, S., Sheller, M., Summers, R. M., Trask, A., Xu, D., Baust, M., and Cardoso, M. J. (2020). **The future of digital health with federated learning**. *NPJ digital medicine*, 3(1):119.

Rohlfing, T. (2011). **Image similarity and tissue overlaps as surrogates for image registration accuracy: widely used but unreliable**. *IEEE transactions on medical imaging*, 31(2):153–163. doi: 10.1109/TMI.2011.2163944.

Rokuss, M., Kovacs, B., Kirchhoff, Y., Xiao, S., Ulrich, C., Maier-Hein, K. H., and Isensee, F. (2024). **From FDG to PSMA: A hitchhiker's guide to multitracer, multicenter lesion segmentation in PET/CT imaging**. *arXiv preprint arXiv:2409.09478*. unpublished findings.

Rokuss, M., Kirchhoff, Y., Akbal, S., Kovacs, B., Roy, S., Ulrich, C., Wald, T., Rotkopf, L. T., Schlemmer, H.-P., and Maier-Hein, K. (2025). **Lesionlocator: Zero-shot universal tumor segmentation and tracking in 3D whole-body imaging**. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 30872–30885.

Romaguera, L. V., Mezheritsky, T., Mansour, R., Carrier, J.-F., and Kadoury, S. (2021). **Probabilistic 4d predictive model from in-room surrogates using conditional generative networks for image-guided radiotherapy**. *Medical image analysis*, 74:102250.

Ronneberger, O., Fischer, P., and Brox, T. (2015). **U-net: Convolutional networks for biomedical image segmentation**. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer.

Rosenkrantz, A. B. (2016). *MRI of the Prostate: A Practical Approach*. Thieme, New York.

Roth, H. R., Chang, K., Singh, P., Neumark, N., Li, W., Gupta, V., Gupta, S., Qu, L., Ihsani, A., Bizzo, B. C., Wen, Y., Buch, V., Shah, M., Kitamura, F., Mendonça, M., Lavor, V., Harouni, A., Compas, C., Tetreault, J., Dogra, P., Cheng, Y., Erdal, S., White, R., Hashemian, B., Schultz, T., Zhang, M., McCarthy, A., Yun, B. M., Sharaf, E., Hoebel, K. V., Patel, J. B., Chen, B., Ko, S., Leibovitz, E., Pisano, E. D., Coombs, L., Xu, D., Dreyer, K. J., Dayan, I., Naidu, R. C., Flores, M., Rubin, D., and Kalpathy-Cramer, J. (2020). **Federated learning for breast density classification: A real-world implementation**. In *Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning: Second MICCAI Workshop, DART 2020, and First MICCAI Workshop, DCL 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4–8, 2020, Proceedings 2*, pages 181–191. Springer.

Rubod, C., Brieu, M., Cosson, M., Rivaux, G., Clay, J.-C., de Landsheere, L., and Gabriel, B. (2012). **Biomechanical properties of human pelvic organs**. *Urology*, 79(4):968–e17.

Rueckert, D., Sonoda, L. I., Hayes, C., Hill, D. L., Leach, M. O., and Hawkes, D. J. (1999). **Nonrigid registration using free-form deformations: application to breast MR images**. *IEEE transactions on medical imaging*, 18(8):712–721.

Sadowski, E. A., Thomassin-Naggara, I., Rockall, A., Maturen, K. E., Forstner, R., Jha, P., Nougaret, S., Siegelman, E. S., and Reinhold, C. (2022). **O-RADS MRI risk stratification system: guide for assessing adnexal lesions from the ACR O-RADS committee**. *Radiology*, 303(1):35–47.

Saha, A., Bosma, J., Linmans, J., Hosseinzadeh, M., and Huisman, H. (2021)a. **Anatomical and diagnostic bayesian segmentation in prostate MRI − should different clinical objectives mandate different loss functions?** *arXiv preprint arXiv:2110.12889*. unpublished findings.

Saha, A., Hosseinzadeh, M., and Huisman, H. (2021)b. **End-to-end prostate cancer detection in bpMRI via 3D CNNs: Effects of attention mechanisms, clinical priori**

and decoupled false positive reduction. *Medical Image Analysis*, 73:102155. ISSN 1361-8415. doi: 10.1016/j.media.2021.102155.

Saha, A., Bosma, J. S., Twilt, J. J., van Ginneken, B., Bjartell, A., Padhani, A. R., Bonekamp, D., Villeirs, G., Salomon, G., Giannarini, G., Kalpathy-Cramer, J., Barentsz, J., Maier-Hein, K. H., Rusu, M., Rouvière, O., van den Bergh, R., Panebianco, V., Kasivisvanathan, V., Obuchowski, N. A., Yakar, D., Elschot, M., Veltman, J., Fütterer, J. J., de Rooij, M., Huisman, H., Saha, A., Bosma, J. S., Twilt, J. J., van Ginneken, B., Noordman, C. R., Slootweg, I., Roest, C., Fransen, S. J., Sunoqrot, M. R., Bathen, T. F., Rouw, D., Immerzeel, J., Geerdink, J., van Run, C., Groeneveld, M., Meakin, J., Karagöz, A., Bône, A., Routier, A., Marcoux, A., Abi-Nader, C., Li, C. X., Feng, D., Alis, D., Karaarslan, E., Ahn, E., Nicolas, F., Sonn, G. A., Bhattacharya, I., Kim, J., Shi, J., Jahanandish, H., An, H., Kan, H., Oksuz, I., Qiao, L., Rohé, M.-M., Yergin, M., Khadra, M., Şeker, M. E., Kartal, M. S., Debs, N., Fan, R. E., Saunders, S., Soerensen, S. J., Moroianu, S., Vesal, S., Yuan, Y., Malakoti-Fard, A., Mačiūnien, A., Kawashima, A., de M.G. de Sousa Machadov, A. M., Moreira, A. S. L., Ponsiglione, A., Rappaport, A., Stanzione, A., Ciuvasovas, A., Turkbey, B., de Keyzer, B., Pedersen, B. G., Eijlers, B., Chen, C., Riccardo, C., Alis, D., Courrech Staal, E. F., Jäderling, F., Langkilde, F., Aringhieri, G., Brembilla, G., Son, H., Vanderlelij, H., Raat, H. P., Pikūnienė, I., Macova, I., Schoots, I., Caglic, I., Zawaideh, J. P., Wallström, J., Bittencourt, L. K., Khurram, M., Choi, M. H., Takahashi, N., Tan, N., Franco, P. N., Gutierrez, P. A., Thimansson, P. E., Hanus, P., Puech, P., Rau, P. R., de Visschere, P., Guillaume, R., Cuocolo, R., Falcão, R. O., van Stiphout, R. S., Girometti, R., Briediene, R., Grigienė, R., Gitau, S., Withey, S., Ghai, S., Penzkofer, T., Barrett, T., Tammisetti, V. S., Løgager, V. B., Černý, V., Venderink, W., Law, Y. M., Lee, Y. J., Bjartell, A., Padhani, A. R., Bonekamp, D., Villeirs, G., Salomon, G., Giannarini, G., Kalpathy-Cramer, J., Barentsz, J., Maier-Hein, K. H., Rusu, M., Obuchowski, N. A., Rouvière, O., van den Bergh, R., Panebianco, V., Kasivisvanathan, V., Yakar, D., Elschot, M., Veltman, J., Fütterer, J. J., de Rooij, M., and Huisman, H. (2024). **Artificial intelligence and radiologists in prostate cancer detection on MRI (PI-CAI): an international, paired, non-inferiority, confirmatory study**. *The Lancet Oncology*.

Sanyal, J., Banerjee, I., Hahn, L., and Rubin, D. (2020). **An automated two-step pipeline for aggressive prostate lesion detection from multi-parametric MR sequence**. *AMIA Summits on Translational Science Proceedings*, 2020:552.

Schelb, P., Kohl, S., Radtke, J. P., Wiesenfarth, M., Kickingereder, P., Bickelhaupt, S., Kuder, T. A., Stenzinger, A., Hohenfellner, M., Schlemmer, H.-P., Maier-Hein, K. H., and Bonekamp, D. (2019). **Classification of cancer at prostate MRI: deep learning versus clinical PI-RADS assessment**. *Radiology*, 293(3):607–617.

Schelb, P., Wang, X., Radtke, J. P., Wiesenfarth, M., Kickingereder, P., Stenzinger, A., Hohenfellner, M., Schlemmer, H.-P., Maier-Hein, K. H., and Bonekamp, D. (2021). **Simulated clinical deployment of fully automatic deep learning for clinical prostate MRI assessment**. *European radiology*, 31(1):302–313. doi: 10.1007/s00330-020-07086-z.

Schmid, J., Assassi, L., and Chênes, C. (2023). **A novel image augmentation based on statistical shape and intensity models: application to the segmentation of hip bones from CT images**. *European radiology experimental*, 7(1):39.

Schoots, I. G. and Padhani, A. R. (2020). **Risk-adapted biopsy decision based on prostate magnetic resonance imaging and prostate-specific antigen density for enhanced biopsy avoidance in first prostate cancer diagnostic evaluation**. *BJU international*, 127 (2):175.

Shannon, C. E. (1948). **A mathematical theory of communication**. *The Bell system technical journal*, 27(3):379–423.

Sheller, M. J., Edwards, B., Reina, G. A., Martin, J., Pati, S., Kotrotsou, A., Milchenko, M., Xu, W., Marcus, D., Colen, R. R., and Bakas, S. (2020). **Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data**. *Scientific reports*, 10(1):12598.

Shimofusa, R., Fujimoto, H., Akamata, H., Motoori, K., Yamamoto, S., Ueda, T., and Ito, H. (2005). **Diffusion-weighted imaging of prostate cancer**. *Journal of computer assisted tomography*, 29(2):149–153.

Shirk, J. D., Reiter, R., Wallen, E. M., Pak, R., Ahlering, T., Badani, K. K., and Porter, J. R. (2022). **Effect of 3-dimensional, virtual reality models for surgical planning of robotic prostatectomy on trifecta outcomes: a randomized clinical trial**. *Journal of Urology*, 208 (3):618–625.

Shish, L. and Zabell, J. (2024). **Digital rectal exam in prostate cancer screening and elevated PSA work-up—is there a role anymore?** *Current Urology Reports*, pages 1–7.

Shit, S., Paetzold, J. C., Sekuboyina, A., Ezhov, I., Unger, A., Zhylka, A., Pluim, J. P., Bauer, U., and Menze, B. H. (2021). **cldice-a novel topology-preserving loss function for tubular structure segmentation**. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16560–16569.

Shorten, C. and Khoshgoftaar, T. M. (2019). **A survey on image data augmentation for deep learning**. *Journal of Big Data*, 6(1):1–48.

Shukla, G., Alexander, G. S., Bakas, S., Nikam, R., Talekar, K., Palmer, J. D., and Shi, W. (2017). **Advanced magnetic resonance imaging in glioblastoma: a review**. *Chinese clinical oncology*, 6(4):40–40.

Siddiqui, M. M., Rais-Bahrami, S., Turkbey, B., George, A. K., Rothwax, J., Shakir, N., Okoro, C., Raskolnikov, D., Parnes, H. L., Linehan, W. M., Merino, M. J., Simon, R. M., Choyke, P. L., Wood, B. J., and Pinto, P. A. (2015). **Comparison of MR/ultrasound fusion–guided biopsy with ultrasound-guided biopsy for the diagnosis of prostate cancer**. *Jama*, 313(4):390–397.

Simard, P., Steinkraus, D., and Platt, J. (2003). **Best practices for convolutional neural networks applied to visual document analysis.** In *Icdar*, volume 3. Edinburgh.

Simonyan, K. and Zisserman, A. (2014). **Very deep convolutional networks for large-scale image recognition**. *arXiv preprint arXiv:1409.1556*. unpublished findings.

Studholme, C., Hill, D. L., and Hawkes, D. J. (1999). **An overlap invariant entropy measure of 3D medical image alignment**. *Pattern recognition*, 32(1):71–86.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). **Rethinking the inception architecture for computer vision**. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

Tang, Y., Yang, D., Li, W., Roth, H. R., Landman, B., Xu, D., Nath, V., and Hatamizadeh, A. (2022). **Self-supervised pre-training of swin transformers for 3D medical image analysis**. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20730–20740.

Tavakoli, A. A., Hielscher, T., Badura, P., Görtz, M., Kuder, T. A., Gnirs, R., Schwab, C., Hohenfellner, M., Schlemmer, H.-P., and Bonekamp, D. (2023). **Contribution of dynamic contrast-enhanced and diffusion MRI to pi-rads for detecting clinically significant prostate cancer**. *Radiology*, 306(1):186–199.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. (2023). **Llama: Open and efficient foundation language models**. *arXiv preprint arXiv:2302.13971*. unpublished findings.

Turkbey, B. and Choyke, P. L. (2018). **Future perspectives and challenges of prostate MR imaging**. *Radiologic Clinics*, 56(2):327–337.

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., Thoeny, H. C., Verma, S., Barentsz, J., and Weinreb, J. C. (2019). **Prostate imaging reporting and data system version 2.1: 2019 update of prostate imaging reporting and data system version 2**. *European urology*, 76(3):340–351.

Twilt, J. J., van Leeuwen, K. G., Huisman, H. J., Fütterer, J. J., and de Rooij, M. (2021). **Artificial intelligence based algorithms for prostate cancer classification and detection on magnetic resonance imaging: a narrative review**. *Diagnostics*, 11(6):959.

Twilt, J. J., Saha, A., Bosma, J. S., van Ginneken, B., Bjartell, A., Padhani, A. R., Bonekamp, D., Villeirs, G., Salomon, G., Giannarini, G., Kalpathy-Cramer, J., Barentsz, J., Maier-Hein, K. H., Rusu, M., Rouvière, O., van den Bergh, R., Panebianco, V., Kasivisvanathan, V., Obuchowski, N. A., Yakar, D., Elschot, M., Veltman, J., Fütterer, J. J., Huisman, H., de Rooij, M., Twilt, J. J., Saha, A., Bosma, J. S., van Ginneken, B., Noordman, C. R., Slootweg, I., Roest, C., Fransen, S. J., Sunoqrot, M. R., Bathen, T. F., Rouw, D., Geerdink, J., van Run, C., Groeneveld, M., Meakin, J., Immerzeel, J. J., Yakar, D., Elschot, M., Veltman, J., Fütterer, J. J., de Rooij, M., Huisman, H., Bjartell, A., Padhani, A. R., Bonekamp, D., Villeirs, G., Salomon, G., Giannarini, G., Kalpathy-Cramer, J., Barentsz, J., Maier-Hein, K. H., Rusu, M., Obuchowski, N. A., Rouviere, O., van den Bergh, R., Panebianco, V., Kasivisvanathan, V., Malakoti-Fard, A., Mačiūnien, A., Kawashima, A., Gaivão, A. M., Moreira, A. S., Ponsiglione, A., Rappaport, A., Stanzione, A., Ciuvasovas, A., Turkbey, B., Keyzer, B. D., Pedersen, B. G., Eijlers, B., Chen, C., Riccardo, C., Alis, D., Courrech Staal, E. F., Thimansson, E., Jäderling, F., Langkilde, F., Aringhieri, G., Brembilla, G., Son, H., van der Lelij, H., Raat, H. P., Pikūnienė, I., Macova, I., Schoots, I., Caglic, I., Zawaideh, J. P., Wallström, J., Bittencourt, L. K., Khurram, M., Choi, M. H., Takahashi, N., Tan, N., Franco, P. N., Gutierrez, P. A., Hanus, P., Puech, P., Rau, P. R., de Visschere, P., Guillaume, R., Cuocolo, R., Falcão, R. O., van Stiphout, R. S., Girometti, R., Briediene, R., Grigienė, R., Gitau, S., Withey, S., Ghai, S., Penzkofer, T., Barrett, T., Tammisetti, V. S., Løgager, V. B., Černý, V., Venderink, W., Law, Y. M., and Lee, Y. J. (2025). **Evaluating biparametric versus multiparametric magnetic resonance imaging for diagnosing clinically significant prostate cancer: An international, paired, noninferiority, confirmatory observer study**. *European urology*, 87(2):240–250.

Valerio, M., Donaldson, I., Emberton, M., Ehdaie, B., Hadaschik, B. A., Marks, L. S., Mozer, P., Rastinehad, A. R., and Ahmed, H. U. (2015). **Detection of clinically significant prostate cancer using magnetic resonance imaging–ultrasound fusion targeted biopsy: a systematic review**. *European urology*, 68(1):8–19.

Van der Leest, M., Israël, B., Cornel, E. B., Zámecnik, P., Schoots, I. G., van der Lelij,

H., Padhani, A. R., Rovers, M., van Oort, I., Sedelaar, M., Hulsbergen-van de Kaa, C., Hannink, G., Veltman, J., and Barentsz, J. (2019). **High diagnostic performance of short magnetic resonance imaging protocols for prostate cancer detection in biopsy-naïve men: the next step in magnetic resonance imaging accessibility**. *European urology*, 76 (5):574–581.

Van Ginneken, B., Armato III, S. G., de Hoop, B., van Amelsvoort-van de Vorst, S., Duindam, T., Niemeijer, M., Murphy, K., Schilham, A., Retico, A., Fantacci, M. E., Camarlinghi, N., Bagagli, F., Gori, I., Hara, T., Fujita, H., Gargano, G., Bellotti, R., Tangaro, S., Bolaños, L., Carlo, F. D., Cerello, P., Cristian Cheran, S., Lopez Torres, E., and Prokop, M. (2010). **Comparing and combining algorithms for computer-aided detection of pulmonary nodules in computed tomography scans: the ANODE09 study**. *Medical image analysis*, 14(6):707–722.

Van Leenders, G. J., van der Kwast, T. H., Grignon, D. J., Evans, A. J., Kristiansen, G., Kweldam, C. F., Litjens, G., McKenney, J. K., Melamed, J., Mottet, N., Paner, G. P., Samaratunga, H., Schoots, I. G., Simko, J. P., Tsuzuki, T., Varma, M., Warren, A. Y., Wheeler, T. M., Williamson, S. R., and Iczkowski, K. A. (2020). **The 2019 international society of urological pathology (ISUP) consensus conference on grading of prostatic carcinoma**. *The American journal of surgical pathology*, 44(8):e87–e99.

Vargas, H. A., Akin, O., Franiel, T., Goldman, D. A., Udo, K., Touijer, K. A., Reuter, V. E., and Hricak, H. (2012). **Normal central zone of the prostate and central zone involvement by prostate cancer: clinical and MR imaging implications**. *Radiology*, 262 (3):894–902.

Venderink, W., Bomers, J. G., Overduin, C. G., Padhani, A. R., de Lauw, G. R., Sedelaar, M. J., and Barentsz, J. O. (2020). **Multiparametric magnetic resonance imaging for the detection of clinically significant prostate cancer: what urologists need to know. part 3: targeted biopsy**. *European Urology*, 77(4):481–490.

Vickers, A. J., Cronin, A. M., Roobol, M. J., Hugosson, J., Jones, J. S., Kattan, M. W., Klein, E., Hamdy, F., Neal, D., Donovan, J., Parekh, D. J., Ankerst, D., Bartsch, G., Klocker, H., Horninger, W., Benchikh, A., Salama, G., Villers, A., Freedland, S. J., Moreira, D. M., Schröder, F. H., and Lilja, H. (2010). **The relationship between prostate-specific antigen and prostate cancer risk: the prostate biopsy collaborative group**. *Clinical Cancer Research*, 16(17):4374–4381.

Vilanova, J. C., Catalá, V., Algaba, F., and Laucirica, O., editors. (2018). *Atlas of multi-parametric prostate MRI: with PI-RADS approach and anatomic-MRI-pathological correlation*. Springer, Cham. doi: 10.1007/978-3-319-61786-2.

Viola, P. and Wells III, W. M. (1997). **Alignment by maximization of mutual information**. *International journal of computer vision*, 24(2):137–154.

Voyant, C., Biffi, K., Leschi, D., Briancon, J., and Lantieri, C. (2011). **Dosimetric uncertainties related to the elasticity of bladder and rectal walls: Adenocarcinoma of the prostate**. *Cancer/Radiothérapie*, 15(4):270–278.

Wasserthal, J., Breit, H.-C., Meyer, M. T., Pradella, M., Hinck, D., Sauter, A. W., Heye, T., Boll, D. T., Cyriac, J., Yang, S., Bach, M., and Segeroth, M. (2023). **TotalSegmentator: robust segmentation of 104 anatomic structures in CT images**. *Radiology: Artificial Intelligence*, 5(5):e230024.

Wegelin, O., van Melick, H. H., Hooft, L., Bosch, J. R., Reitsma, H. B., Barentsz, J. O., and Somford, D. M. (2017). **Comparing three different techniques for magnetic resonance imaging-targeted prostate biopsies: a systematic review of in-bore versus magnetic resonance imaging-transrectal ultrasound fusion versus cognitive registration. is there a preferred technique?** *European urology*, 71(4):517–531.

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., Thoeny, H. C., and Verma, S. (2016). **PI-RADS prostate imaging–reporting and data system: 2015, version 2**. *European urology*, 69(1):16–40.

Wekking, D., Porcu, M., De Silva, P., Saba, L., Scartozzi, M., and Solinas, C. (2023). **Breast MRI: clinical indications, recommendations, and future applications in breast cancer diagnosis**. *Current oncology reports*, 25(4):257–267.

Westphalen, A. C., McCulloch, C. E., Anaokar, J. M., Arora, S., Barashi, N. S., Barentsz, J. O., Bathala, T. K., Bittencourt, L. K., Booker, M. T., Braxton, V. G., Carroll, P. R., Casalino, D. D., Chang, S. D., Coakley, F. V., Dhatt, R., Eberhardt, S. C., Foster, B. R., Froemming, A. T., Fütterer, J. J., Ganeshan, D. M., Gertner, M. R., Mankowski Gettle, L., Ghai, S., Gupta, R. T., Hahn, M. E., Houshyar, R., Kim, C., Kim, C. K., Lall, C., Margolis, D. J. A., McRae, S. E., Oto, A., Parsons, R. B., Patel, N. U., Pinto, P. A., Polascik, T. J., Spilseth, B., Starcevich, J. B., Tammisetti, V. S., Taneja, S. S., Turkbey, B., Verma, S., Ward, J. F., Warlick, C. A., Weinberger, A. R., Yu, J., Zagoria, R. J., and Rosenkrantz, A. B. (2020). **Variability of the positive predictive value of PI-RADS for prostate MRI across 26 centers: experience of the society of abdominal radiology prostate cancer disease-focused panel**. *Radiology*, 296(1):76–84.

Winkel, D. J., Tong, A., Lou, B., Kamen, A., Comaniciu, D., Disselhorst, J. A., Rodríguez-Ruiz, A., Huisman, H., Szolar, D., Shabunin, I., Choi, M. H., Xing, P., Penzkofer, T.,

Grimm, R., von Busch, H., and Boll, D. T. (2021). **A novel deep learning based computer-aided diagnosis system improves the accuracy and efficiency of radiologists in reading biparametric magnetic resonance images of the prostate: results of a multireader, multicase study**. *Investigative radiology*, 56(10):605–613.

Woernle, A., Englman, C., Dickinson, L., Kirkham, A., Punwani, S., Haider, A., Freeman, A., Kasivisivanathan, V., Emberton, M., Hines, J., Moore, C. M., Allen, C., and Giganti, F. (2024). **Picture perfect: the status of image quality in prostate MRI**. *Journal of Magnetic Resonance Imaging*, 59(6):1930–1952.

Wolf, I., Vetter, M., Wegner, I., Böttger, T., Nolden, M., Schöbinger, M., Hastenteufel, M., Kunert, T., and Meinzer, H.-P. (2005). **The medical imaging interaction toolkit**. *Medical image analysis*, 9(6):594–604.

Woodhams, R., Ramadan, S., Stanwell, P., Sakamoto, S., Hata, H., Ozaki, M., Kan, S., and Inoue, Y. (2011). **Diffusion-weighted imaging of the breast: principles and clinical applications**. *Radiographics*, 31(4):1059–1084.

Woodrum, D. A., Gorny, K. R., Greenwood, B., and Mynderse, L. A. (2016). **MRI-guided prostate biopsy of native and recurrent prostate cancer**. In *Seminars in Interventional Radiology*, volume 33, pages 196–205. Thieme Medical Publishers.

Wysock, J. S., Rosenkrantz, A. B., Huang, W. C., Stifelman, M. D., Lepor, H., Deng, F.-M., Melamed, J., and Taneja, S. S. (2014). **A prospective, blinded comparison of magnetic resonance (MR) imaging–ultrasound fusion and visual estimation in the performance of MR-targeted prostate biopsy: the profus trial**. *European urology*, 66(2):343–351.

Xie, Y., Zhang, J., Shen, C., and Xia, Y. (2021). **Cotr: Efficiently bridging cnn and transformer for 3D medical image segmentation**. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24*, pages 171–180. Springer.

Yang, Q., Li, N., Zhao, Z., Fan, X., Chang, E. I.-C., and Xu, Y. (2020). **MRI cross-modality image-to-image translation**. *Scientific reports*, 10(1):3753.

Yang, X., Liu, C., Wang, Z., Yang, J., Le Min, H., Wang, L., and Cheng, K.-T. T. (2017). **Co-trained convolutional neural networks for automated detection of prostate cancer in multi-parametric MRI**. *Medical image analysis*, 42:212–227.

Ying, Y., He, W., Xiong, Q., Wang, Z., Wang, M., Chen, Q., Hua, M., Zeng, S., and Xu, C. (2023). **Value of digital rectal examination in patients with suspected prostate cancer: a prospective cohort analysis study**. *Translational Andrology and Urology*, 12(11):1666.

Youssofzadeh, V., McGuinness, B., Maguire, L. P., and Wong-Lin, K. (2017). **Multi-kernel learning with dartel improves combined MRI-PET classification of alzheimer's disease in AIBL data: group and individual analyses**. *Frontiers in human neuroscience*, 11:380.

Zawaideh, J. P., Sala, E., Shaida, N., Koo, B., Warren, A. Y., Carmisciano, L., Saeb-Parsy, K., Gnanapragasam, V. J., Kastner, C., and Barrett, T. (2020). **Diagnostic accuracy of biparametric versus multiparametric prostate MRI: assessment of contrast benefit in clinical practice**. *European radiology*, 30:4039–4049.

Zhai, X., Kolesnikov, A., Houlsby, N., and Beyer, L. (2022). **Scaling vision transformers**. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12104–12113.

Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., Wong, C., Tupini, A., Wang, Y., Mazzola, M., Shukla, S., Liden, L., Gao, J., Crabtree, A., Piening, B., Bifulco, C., Lungren, M. P., Naumann, T., Wang, S., and Poon, H. (2023). **Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs**. *arXiv preprint arXiv:2303.00915*. unpublished findings.

Zhao, B., Schwartz, L. H., Kris, M. G., and Riely, G. J. (2015). **Coffee-break lung CT collection with scan images reconstructed at multiple imaging parameters (version 3) [RIDER lung CT]**. *(No Title)*. doi: 10.7937/k9/tcia.2015.u1x8a5nr.

Zhao, T., Gu, Y., Yang, J., Usuyama, N., Lee, H. H., Kiblawi, S., Naumann, T., Gao, J., Crabtree, A., Abel, J., Moung-Wen, C., Piening, B., Bifulco, C., Wei, M., Poon, H., and Wang, S. (2025). **A foundation model for joint segmentation, detection and recognition of biomedical objects across nine modalities**. *Nature methods*, 22(1):166–176.

Zhou, H.-Y., Guo, J., Zhang, Y., Yu, L., Wang, L., and Yu, Y. (2021). **nnformer: Interleaved transformer for volumetric segmentation**. *arXiv preprint arXiv:2109.03201*. unpublished findings.

Zhou, S. K., Rueckert, D., and Fichtinger, G., editors. (2020). *Handbook of medical image computing and computer assisted intervention*. Academic Press, London and San Diego and Cambridge. doi: 10.1016/C2017-0-04608-6.

Own Contributions

This chapter gives an overview of my contributions in distinction to team efforts.

Own share in data acquisition and data analysis

This interdisciplinary thesis in the field of medical informatics was written within the Division of Medical Image Computing at the German Cancer Research Center (DKFZ) Heidelberg, headed by Prof. Dr. Klaus Hermann Maier-Hein, served as the primary supervisor. The research presented in this thesis was carried out in clinical collaboration with the Division of Radiology at DKFZ Heidelberg, headed by Prof. Dr. Heinz-Peter Schlemmer, particularly with the Prostate Cancer Working Group, led by Prof. Dr. David Bonekamp, who provided the data through co-supervision. Furthermore, I was in close collaboration with members of both divisions throughout the time of my doctoral work. This thesis is primarily based on two research projects:

1. **Soft Tissue Deformations of the Prostate**, which has been primarily presented in the following first-author journal publication:
   B. Kovacs, N. Netzer, M. Baumgartner, A. Schrader, F. Isensee, C. Weißer, I. Wolf, M. Görtz, P.F. Jaeger, V. Schütz, R. Floca, R. Gnirs, A. Stenzinger, M. Hohenfellner, H.P. Schlemmer, D. Bonekamp., K.H. Maier-Hein. **Nature Scientific Reports**. *Addressing image misalignments in multi-parametric prostate MRI for enhanced computer-aided diagnosis of prostate cancer*. `https://doi.org/10.1038/s41598-023-46747-z`

2. **Multi-Modal Misalignments in Prostate MRI**, which has been primarily presented in the following first-author conference publication:
   B. Kovacs, N. Netzer, M. Baumgartner, C. Eith, D. Bounias, C. Meinzer, P.F. Jaeger, K.S. Zhang, R. Floca, A. Schreder, F. Isensee, R. Gnirs, M. Görtz, V. Schütz, A. Stenzinger, M. Hohenfellner, H.P. Schlemmer, I. Wolf, D. Bonekamp, K.H. Maier-Hein. **In-**

ternational Conference on Medical Image Computing and Computer-Assisted Intervention - MICCAI 2023. *Anatomy-informed data augmentation for enhanced prostate cancer detection*. https://doi.org/10.1007/978-3-031-43990-2_50

Own share in Data Acquisition. The clinical foundation for the artificial intelligence (AI) experiments in both primary research projects, including magnetic resonance imaging (MRI) acquisition, clinical assessment, and data annotation was established by multiple collaborators from the Division of Radiology at DKFZ Heidelberg (Carolin Eith, Regula Gnirs, Clara Meinzer, Nils Netzer, Adrian Schrader, Cedric Weißer, Dr. Kevin S. Zhang, Prof. Dr. David Bonekamp, Prof. Dr. Heinz-Peter Schlemmer), Department of Urology at the University of Heidelberg Medical Center (Priv. Doz. Dr. Magdalena Görtz, Dr. Victoria Schütz, Prof. Dr. Markus Hohenfellner), and Institute of Pathology at the University of Heidelberg Medical Center (Prof. Dr. Albrecht Stenzinger).

Although I was not clinically qualified to participate directly in the acquisition of patient data, I was responsible for the systematic curation and preprocessing of the prostate MRI datasets used in both primary research projects to be able to answer my research questions. This process involved:

- Selecting the appropriate bi-parametric MRI (bpMRI) images, including the T2-weighted (T2w) scans, diffusion-weighted imaging (DWI) scans with both the lowest and highest b-values, and the apparent diffusion coefficient (ADC) maps.
- Curating the clinical organ segmentations for the prostate, bladder, and rectum.
- Filtering annotations of clinically significant PCa (csPCa) lesions on both T2w and ADC maps, using the systematic and targeted biopsy data for the SELGT annotations.

Furthermore, I also conducted comprehensive sanity checks on the lesion ground truth annotations across individual studies to ensure the highest possible dataset quality for my experiments:

1. **Soft Tissue Deformations of the Prostate:** I identified inaccuracies in the lesion annotations, including missing slices, and overlapping or adjacent instance labels labeled as a multiple lesion entity. While such errors may have had minimal impact on training for the task of semantic segmentation, they would have substantially affected the lesion-level evaluations.
2. **Multi-Modal Misalignments in Prostate MRI:** I detected inconsistencies between lesion annotations on T2w images and ADC maps, such as mismatched lesion identifiers, missing slices, and incorrect or overlapping labels. These inconsistencies

would have affected the assessment of lesion-level overlap across modalities used for evaluating image registration techniques.

All identified issues were corrected in collaboration with members from the Division of Radiology at DKFZ Heidelberg (Clara Meinzer, Nils Netzer, Cedric Weißer, Dr. Kevin S. Zhang, Prof. Dr. David Bonekamp).

Own share in Data Analysis.

1. **Soft Tissue Deformations of the Prostate:** The proposed method for simulating soft tissue deformations, named the anatomy-informed transformation was developed entirely by me. I also designed, conducted, and evaluated the Turing test to assess the realism of the synthetic images, where Carolin Eith, Clara Meinzer, and Dr. Kevin S. Zhang participated as expert raters. The integration of this method into the nnU-Net pipeline, the design and conduction of the AI experiments, the patient- and lesion-level evaluation were also performed entirely by me. Additionally, I adapted the proposed method to simulate lesion shape variations in a subsequent conference publication mentioned in the discussion, in which I contributed as a co-author:
   M. Rokuss, Y. Kirchhoff, S. Akbal, **B. Kovacs**, S. Roy, C. Ulrich, T. Wald, L.T. Rotkopf, H.P. Schlemmer, K.H. Maier-Hein. **IEEE/CVF Conference on Computer Vision and Pattern Recognition - CVPR 2025**. *LesionLocator: Zero-Shot Universal Tumor Segmentation and Tracking in 3D Whole-Body Imaging*. (Accepted, in press).

2. **Multi-Modal Misalignments in Prostate MRI:** The proposed misalignment-data augmentation was implemented by myself. Two registration algorithms were employed for different experimental settings. The B-spline rgistration used for a previous publication on this data cohort was provided by Nils Netzer and Prof. Dr. David Bonekamp. I adapted this algorithm for the GT-matching rigid registration. The integration of the misalignment augmentation and registration strategies into the nnU-Net pipeline, the design and conduction of the AI experiments, and the analysis of results were performed entirely by me. Furthermore, I adapted the proposed method for PET/CT images in the unpublised studies mentioned in the Discussion.

Throughout both research projects, I regularly received methodological advisory input from Michael Baumgartner, Dimitrios Bounias, Dr. Paul F. Jäger, Dr. Fabian Isensee, Dr. Ralf Floca, Prof. Dr. Ivo Wolf, and Prof. Dr. Klaus H. Maier-Hein. I received clinical insights from Carolin Eith, Regula Gnirs, Clara Meinzer, Nils Netzer, Adrian Schrader, Cedric Weißer, Dr. Kevin S. Zhang, and Prof. Dr. David Bonekamp.

Own Publications

In this section, all publications that I was a part of and contributed to during my Ph.D. work are listed. It is subdivided into *First Authorships* and *Co-Authorships*.

First Authorships - Peer Reviewed International Journal Publications

B. Kovacs, N. Netzer, M. Baumgartner, A. Schrader, F. Isensee, C. Weißer, I. Wolf, M. Görtz, P.F. Jaeger, V. Schütz, R. Floca, R. Gnirs, A. Stenzinger, M. Hohenfellner, H.P. Schlemmer, D. Bonekamp., K.H. Maier-Hein. **Nature Scientific Reports**. *Addressing image misalignments in multi-parametric prostate MRI for enhanced computer-aided diagnosis of prostate cancer*. https://doi.org/10.1038/s41598-023-46747-z.

B. Kovacs, F.A. Kraft, Z. Szabo, Y. Nazirizadeh, M. Gerken, R. Horvath. **Nature Scientific Reports**. *Near cut-off wavelength operation of resonant waveguide grating biosensors*. https://doi.org/10.1038/s41598-021-92327-4.

First Authorships - Peer Reviewed International Conference Publications

B. Kovacs, N. Netzer, M. Baumgartner, C. Eith, D. Bounias, C. Meinzer, P.F. Jaeger, K.S. Zhang, R. Floca, A. Schreder, F. Isensee, R. Gnirs, M. Görtz, V. Schütz, A. Stenzinger, M. Hohenfellner, H.P. Schlemmer, I. Wolf, D. Bonekamp, K.H. Maier-Hein. **International Conference on Medical Image Computing and Computer-Assisted Intervention - MICCAI 2023**. *Anatomy-informed data augmentation for enhanced prostate cancer detection*. https://doi.org/10.1007/978-3-031-43990-2_50.

Co-Authorships

D. Bounias, T. Führes, L. Brock, J. Graber, L.A. Kapsner, A. Liebert, H. Schreiter, J. Eberle, D. Hadler, D. Skwierawska, R. Floca, P. Neher, **B. Kovacs**, E. Wenkel, S. Ohlmeyer, M. Uder, K.H. Maier-Hein **Nature Communications**. *AI-Based Screening for Thoracic Aortic Aneurysms in Routine Breast MRI*.
https://doi.org/10.1038/s41467-025-59694-2.

M. Rokuss, Y. Kirchhoff, S. Akbal, **B. Kovacs**, S. Roy, C. Ulrich, T. Wald, L.T. Rotkopf, H.P. Schlemmer, K.H. Maier-Hein. **IEEE/CVF Conference on Computer Vision and Pattern Recognition - CVPR 2025**. *LesionLocator: Zero-Shot Universal Tumor Segmentation and Tracking in 3D Whole-Body Imaging*. (Accepted, in press).

T. Wald, B. Hamm, J.C. Holzschuh, R. El Shafie, A. Kudak, **B. Kovacs**, I. Pflüger, B. von Nettelbladt, C. Ulrich, M.A. Baumgartner, P. Vollmuth, J. Debus, K.H. Maier-

Hein, T. Welzel. **European Radiology Experimental**. *Enhancing deep learning methods for brain metastasis detection through cross-technique annotations on SPACE MRI*. https://doi.org/10.1186/s41747-025-00554-5.

P. A. Glemser, M. Freitag, **B. Kovacs**, N. Netzer, A. Dimitrakopoulou-Strauss, U. Haberkorn, K. Maier-Hein, C. Schwab, S. Duensing, B. Beuthien-Baumann, H.-P. Schlemmer, D. Bonekamp, F. Giesel, C. Sachpekidis **EJNMMI Reports**. *Enhancing the diagnostic capacity of [18F]PSMA-1007 PET/MRI in primary prostate cancer staging with artificial intelligence and semi-quantitative DCE: an exploratory study*. https://doi.org/10.1186/s41824-024-00225-5.

M.R. Rokuss, Y. Kirchhoff, S. Roy, S., **B. Kovacs**, C. Ulrich, T. Wald, M. Zenk, S. Denner, F. Isensee, P. Vollmuth, J. Kleesiek, K.H. Maier-Hein. **International Conference on Medical Image Computing and Computer-Assisted Intervention - MICCAI 2024**. *Longitudinal segmentation of MS lesions via temporal Difference Weighting*. https://doi.org/10.1007/978-3-031-84525-3_6.

Y. Kirchhoff, M.R. Rokuss, S. Roy, **B. Kovacs**, C. Ulrich, T. Wald, M. Zenk, P. Vollmuth, J. Kleesiek, F. Isensee, K.H. Maier-Hein. **European Conference on Computer Vision - ECCV 2024**. *Skeleton recall loss for connectivity conserving and resource efficient segmentation of thin tubular structures*.
https://doi.org/10.1007/978-3-031-72980-5_13.

Appendix A

Dataset Properties

A.1 Demographic Tables for In-House Datasets

Table A.1: Demographic and clinical characteristics of the in-house cohort #1. Table adapted from our previously published work Kovacs et al. (2023b), reused under the Creative Commons Attribution 4.0 License.

Characteristic	Training set (80 %)	Test set (20 %)
No. exams (421)	335	86
– without csPCa	222 (66.3 %)	54 (62.8 %)
– with csPCa	113 (33.7 %)	32 (37.2 %)
Median age (years)	64	64
Mean PSA (ng/mL)	8.79	10.05
Exams w/o MRI-detected lesion	13 (3.9 %)	4 (4.7 %)
MRI-detected index lesions		
– PI-RADS 2	43 (12.8 %)	10 (11.6 %)
– PI-RADS 3	83 (24.8 %)	19 (22.1 %)
– PI-RADS 4	129 (38.5 %)	34 (39.5 %)
– PI-RADS 5	67 (20.0 %)	19 (22.1 %)
No. csPCa/patient		
– 1 lesion	77	21
– 2 lesions	33	8
– 3 lesions	3	3
csPCa location		
– Peripheral zone	102	34
– Transition zone	42	12
– Multi zone	8	0
ISUP Grade		
– no PCa	148 (44.2 %)	36 (41.9 %)
– Gleason 3+3	74 (22.1 %)	18 (20.9 %)
– ISUP 2	63 (18.8 %)	22 (25.6 %)
– ISUP 3	21 (6.3 %)	3 (3.5 %)
– ISUP 4	12 (3.6 %)	2 (2.3 %)
– ISUP 5	17 (5.1 %)	5 (5.8 %)

Table A.2: Demographic and clinical characteristics of the in-house cohort #2. The cohort is used for our study Kovacs et al. (2023a) and values have been already partially published.

Characteristic	Training set (80 %)	Test set (20 %)
No. exams (774)	619	155
– without csPCa	394 (63.7 %)	99 (63.9 %)
– with csPCa	225 (36.3 %)	56 (36.1 %)
Median age (years)	64	65
Mean PSA (ng/mL)	8.67	9.90
Exams w/o MRI-detected lesion	13 (2.1 %)	7 (4.5 %)
MRI-detected index lesions		
– PI-RADS 2	64 (10.3 %)	16 (10.3 %)
– PI-RADS 3	163 (26.3 %)	43 (27.8 %)
– PI-RADS 4	265 (42.8 %)	48 (31.0 %)
– PI-RADS 5	114 (18.4 %)	41 (26.5 %)
No. csPCa/patient		
– 1 lesion	143	40
– 2 lesions	65	12
– 3 lesions	16	4
– 4 lesions	1	0
csPCa location		
– Transition zone	80	18
– Peripheral zone	234	56
– Multi zone	11	2
ISUP Grade		
– no PCa	265 (42.8 %)	66 (42.6 %)
– Gleason 3+3	129 (20.8 %)	33 (21.3 %)
– ISUP 2	130 (21.0 %)	31 (20.0 %)
– ISUP 3	38 (6.1 %)	9 (3.5 %)
– ISUP 4	26 (4.2 %)	5 (5.8 %)
– ISUP 5	31 (5.0 %)	11 (7.1 %)

A.2 MRI Protocol

Table A.3: Detailed sequence parameters for biparametric MRI of the in-house cohort #2 utilized for training of the deep learning system. Ranges represent the 5th and 95th percentile.

Scanner type	Siemens Prisma		Siemens Biograph mMR		Siemens Aera	
Field strength	3 T		3 T		1.5 T	
Acquisition plane	transversal		transversal		transversal	
Sequence name	T2w	DWI	T2w	DWI	T2w	DWI
b-values	-	0-1500, 2000	-	0-1500	-	0-1500
TE (ms)	145	48	143-146	80-86	123	68
TR (ms)	8080-9690	3300-4300	8000-9524	7640-9650	5610-6120	5300
In-plane resolution (mm)	0.3	2	0.3	2.2-3	0.6	2.6
Slice thicknes (mm)	3.0	3.0	3.3	3.0	3.5	3.0
FoV x (mm)	200	208	167-199	204-210	220-229	290
FoV y (mm)	200	280	199	280-300	220-229	290

Appendix B

Dokumentation zur Verwendung KI-basierter elektronischer Hilfsmittel

Alle wissenschaftlichen Ideen, Konzepte, Interpretationen und Schlussfolgerungen wurden unabhängig und ohne Einfluss KI-basierter Hilfsmittel entwickelt. Alle Methoden und Ergebnisse basieren auf der unabhängigen Arbeit des Autors unter Verwendung konventioneller wissenschaftlicher Methoden. Das KI-basierte Hilfsmittel wurde ausschließlich zur sprachlichen Überarbeitung durch Verbesserung der Grammatik von bereits selbst verfasstem Text verwendet.

B.1 Ziele der Verwendung KI-basierter Hilfsmittel

- Grammatische Empfehlungen selbst verfassten Textes
- Grammatische Prüfung eigener Übersetzung der englischen Zusammenfassung auf Deutsch

B.2 Verwendungsweise der KI-basierten Hilfsmittel

Der Arbeitsablauf bestand aus folgenden Schritten:

1. Manuelles Einfügen ausschließlich selbst verfasster Absätze in das KI Modell mit der Bitte um grammatikalische Vorschläge
2. Kritische Überprüfung der KI-Vorschläge durch den Autor
3. Umsetzung geeigneter Vorschläge, ohne wissenschaftlichen Inhalt oder Aussage zu verändern. Kein Textabschnitt wurde direkt oder vollständig von einer KI übernommen.

B.3 Übersicht der KI-Verwendung nach Kapiteln

KI-basierte Hilfsmittel wurden in folgenden Teilen der Dissertation eingesetzt:

- Kapitel 1-5: ChatGPT von OpenAI wurde verwendet, um grammatikalische Empfehlungen zu geben. Diese wurden kritisch geprüft und gegebenenfalls umgesetzt, wobei der wissenschaftliche Inhalt immer erhalten blieb und keine KI-generierten Informationen übernommen wurden.
- Kapitel 6: Die deutsche Übersetzung der englischen Zusammenfassung wurde durch ChatGPT von OpenAI grammatikalisch geprüft.

Acknowledgement

Not because convention dictates starting with the supervisor, but because he also deserves it, I express my gratitude to Prof. Dr. Klaus H. Maier-Hein. I am incredibly thankful that he welcomed me into his team and stood behind me with all his support in every situation. His supervision, reflecting his deep expertise and honest feedback, greatly shaped my development as a researcher. I will never forget the night of the MICCAI submission deadline, when he stayed with us at the institute over midnight, bombing us with last-minute suggestions for improvement. That day was just one example of how he could bring out the best from us. Beyond his scientific guidance, I deeply respect him as a person, for his thoughtful advice, and for his professionalism with calm presence even in high-pressure situations. He has shaped not only my scientific thinking but also my personal growth. For that, I will always think of him not just as a supervisor but as a mentor. Alongside Klaus, I would also like to thank Prof. Dr. Lena Maier-Hein for her kind support from time to time. Though not officially involved in my PhD, her encouraging words made me feel that she was also quietly following and supporting my journey.

I am particularly grateful to my internal Thesis Advisory Committee (iTAC) members: Michael Baumgartner, Dimitrios Bounias, and Dr. Ralf Floca. I truly appreciated their professional guidance and kind support throughout my PhD, investing tremendous time and energy into my development. What I particularly valued was their honesty and openness to discuss any topic at any time, which meant a lot to me. Although Dr. Paul Jäger had to leave my iTAC early on the beginning due to the establishment of his research group, and Dr. Fabian Isensee was never officially part of my iTAC, I was still able to rely on their support, for which I was always grateful. I was particularly happy to have Prof. Dr. Ivo Wolf in my TAC. Even as an external member, it was natural for him to respond insightfully, and in detail, based on his deep expertise as soon as possible whenever I needed it. His support helped me a lot during my PhD. The professionalism

and personalities of all of them left a strong impression on me that I am glad to carry forward on my future path.

I chose to do a PhD in the field of medical informatics, coming from an electrical engineering background, because I think that the intersection of engineering and medicine is one of the most impressive and demanding areas of research. That is why I appreciated the opportunity to directly engage with members of the Division of Radiology, Nuclear Medicine, and Intelligent Systems and Robotics in Urology, including Clara Meinzer, Carolin Eith, Regula Gnirs, Hanna Leisz, Julius Holzsuh, Nils Netzer, Adrian Schrader, Cedric Weißer, Dr. Kevin S. Zhang, Dr. Philip Glemser, Prof. Dr. Christos Sachpekidis, and Dr. Caelán Max Haney. They were always open to my questions, kindly and patiently introducing me to their clinical workflows and sharing insights into their field. These conversations were always a pleasure to listen to. A particular highlight for me was the opportunity to conduct a Turing test, something I had always wanted to try, which was made possible through the participation of Clara Meinzer, Carolin Eith, and Dr. Kevin S. Zhang. I would also like to thank Prof. Dr. David Bonekamp for providing the high-quality, well-maintained prostate MRI dataset reflecting his expertise, which was essential to this research.

I greatly appreciated the supportive working environment of SYMIC, where it was a pleasure to work. I was also happy to be part of the iTACs of Dimitrios Bounias, Benjamin Hamm, Yannik Kirchoff, and Max Rokuss. I learned a lot through their projects, and they were always kind and approachable whenever we spoke. I am grateful to them for involving me. I also thank our scientific coordinator team and the secretary office, including Michaela Gelz, Nina Kraft, Theresa Klocke, Stefanie Strzysch, Dr. Kathrin Brunk, Dr. Nina Decker, and Dr. Daniel Walther, for their continuous support. By helping with administrative tasks, always kindly and efficiently, they saved me valuable time and allowed me to focus more on my research. It was always a pleasure to pass by their offices and have a quick chat.

I will not remember my PhD only for its scientific achievements. Indeed, it was a time when I was surrounded by extraordinary people with a strong team spirit, many of whom became truly good friends. Canadian adventures with Michael and Alex S. (don't forget to stay in shape, guys, the Black Tusk is waiting for us next time!); the snowy Hawk Eye Peak attack with Tim, Constantin, Tassilo, and Karol (who was able to make it in Puma Sneaker!); the NAMIC Project Week with Deepa, Odile, Philipp, Max F., Marco, and Klaus; the lángos and grill events in our "garden", and our 25-meter move through it, with Clara, Lisa, Michael, Sebastian, Alex S., Dimitris, Shuhan, Nina, Michaela, Lars, Tabea, Karol, Carolin, Nils, Catana, Max Z., David Z., Constantin, Stefan, and Finja; the stand-up paddling with Silvia, Filipa, Jenny, Lukas, Max F., and Max Z.; the fun we had, with highlights like our Christmas videos, with my office mates Alex E., Shuhan, Silvia,

Max Z., Stefan, and Ole, made even the busiest days enjoyable; and so many more. Thank you all! Don't wait to knock if you are passing my door, you are always welcome!

No less importantly, during my PhD, Dora and I became a family, and we also had the joy of welcoming someone new into the world. This made the entire period even more special and meaningful. I would like to thank our families and friends for their support during this time, which often helped us manage the demands we faced. Above all, Dora was the steady point I could always rely on. In intense periods, such as paper deadlines or conferences, she gave me the space and strength to dedicate myself fully, helping make these projects successful. For that, I am deeply grateful.

Eidesstattliche Versicherung

1. Bei der eingereichten Dissertation zu dem Thema ***Data-Centric Artificial Intelligence for Enhanced Prostate Cancer Diagnosis on Magnetic Resonance Images*** handelt es sich um meine eigenständig erbrachte Leistung.
2. Ich habe nur die angegebenen Quellen und Hilfsmittel benutzt und mich keiner unzulässigen Hilfe Dritter bedient. Insbesondere habe ich wörtlich oder sinngemäß aus anderen Werken übernommene Inhalte als solche kenntlich gemacht.
3. Die Arbeit oder Teile davon habe ich bislang nicht an einer Hochschule des In- oder Auslands als Bestandteil einer Prüfungs- oder Qualifikationsleistung vorgelegt.
4. Die Richtigkeit der vorstehenden Erklärungen bestätige ich.
5. Die Bedeutung der eidesstattlichen Versicherung und die strafrechtlichen Folgen einer unrichtigen oder unvollständigen eidesstattlichen Versicherung sind mir bekannt. Ich versichere an Eides statt, dass ich nach bestem Wissen die reine Wahrheit erklärt und nichts verschwiegen habe.

Heidelberg, 15.06.2025 _____

Bálint Kovács

Angabe zu verwendeten KI-basierter Elektronischer Hilfsmittel

Zur Dokumentation der verwendeten Hilfsmittel ist der schriftlichen Ausarbeitung ein besonderer Anhang hinzugefügt, der eine Liste und Beschreibung aller verwendeter KI-basierter Hilfsmittel enthält. Der besondere Anhang zur Dokumentation der verwendeten Hilfsmittel erfüllt folgende Kriterien:

1. Auflistung der Ziele, für die die KI-basierten Hilfsmittel in der vorliegenden Arbeit eingesetzt wurden.

2. Dokumentation der Verwendungsweise der KI-basierten Hilfsmittel.

3. Nennung der Kapitel und Abschnitte der vorliegenden Arbeit, in denen die KI-basierten Hilfsmittel eingesetzt wurden, um Inhalte zu erzeugen.

Der Gebrauch dieser Hilfsmittel inklusive Art, Ziel und Umfang des Gebrauchs wurde mit meinem offiziellen Betreuer **Prof. Dr. Klaus H. Maier-Hein** abgesprochen.

Mir ist bewusst, dass insbesondere der Versuch einer nicht dokumentierten Nutzung KI- basierter Hilfsmittel als Täuschungsversuch zu werten ist:

Gem. § 16 Abs. 2 der Promotionsordnung „Dr. med. / dent.“: „Ergibt sich vor Aushändigung der Promotionsurkunde, dass der Kandidat/die Kandidatin bei einer Promotionsleistung getäuscht hat, so können einzelne oder alle Promotionsleistungen für ungültig erklärt werden. In schweren Fällen kann die Zulassung zum Promotionsverfahren zurückgenommen werden.“

Und § 16 Abs. 2 der Promotionsordnung „Dr. sc. hum.“: „Ergibt sich vor Aushändigung der Promotionsurkunde, dass der Doktorand / Doktorandin bei einer Promotionsleistung getäuscht hat, so kann der Promotionsausschuss diese Promotionsleistung oder

alle bisher erbrachten Promotionsleistungen für ungültig erklären. In besonders schweren Fällen kann der Promotionsausschuss die Annahme als Doktorand / Doktorandin endgültig widerrufen."

Heidelberg, 15.06.2025 _____

Bálint Kovács