WARSAW UNIVERSITY OF TECHNOLOGY

DISCIPLINE OF SCIENCE INFORMATION AND COMMUNICATION TECHNOLOGY/
FIELD OF SCIENCE ENGINEERING AND TECHNOLOGY

Ph.D. Thesis

Piotr Sobecki, M.Sc.

The use of domain knowledge in methods of computer-aided prostate cancer diagnosis

Supervisor

prof. dr hab. inż. Artur Przelaskowski

WARSAW 2022

The use of domain knowledge in methods of computer-aided prostate cancer diagnosis

Abstract

Medical imaging plays a key role in the noninvasive diagnosis of prostate cancer, which is the second most common cancer among men worldwide. Current diagnostic standards define the terminology used in the description of radiological findings and methods for assessing the clinical significance of prostate lesions by evaluating significant imaging sets of features. The results of radiology examinations are subjective in their interpretations of imaging features and in the narrative form of their reports; these frequently contribute to diminished diagnostic value.

This thesis considers solutions that are designed to structure and normalise the text contained in such reports. It also proposes conceptual and computational extensions of the representation of diagnostic description. Selected semantic forms were integrated with computational models based on analyses of features relevant to diagnosis and therapy. Domain knowledge gathered by means of content- and quality-standardised diagnostic protocols were used to construct solutions that infer interpretations of objectified diagnostic images to increase the effectiveness of clinical decisions. The bulk of this thesis concerns the development and verification of the effective use of objectified domain knowledge in computer-aided prostate cancer diagnosis solutions. This form of support in report generation procedures serves to compile representative and structured datasets; their analysis contributes to the development of more reliable and credible computational models.

The concept of computer-assisted structural reporting supplemented with calculated and explained interpretations of data included in reports enables both the accuracy of diagnosticians' conclusions to be supported and effectiveness of decisions made by clinicians to be improved. Our experiments have confirmed the validity of these conclusions.

Keywords: domain knowledge, prostate cancer, structured reporting

Contents

Glossary of Common Terms

ACR	American College of Radiology
ADC	apparent diffusion coefficient
AI	artificial intelligence
AS	anterior fibromuscular stroma (of the prostate)
AUC	area under the receiver operating characteristics curve
BI-RADS	breast imaging reporting and data system
BPH	benign prostate hyperplasia
BPMN	business process modelling notation
bpMRI	biparametric magnetic resonance imaging
CAD	computer-aided diagnosis
CAR/DS	computer-assisted reporting and decision support
CDE	common data element
CMA	common model architecture
CNN	convolutional neural network
csPCa	clinically significant prostate cancer
CZ	central zone (of the prostate)
DCE	dynamic contrast enhancement

DMN	decision model and notation
DRE	digital rectal examination
DWI	diffusion-weighted imaging
EAU	European Association of Urology
mpMRI	multiparametric magnetic resonance imaging
MRI	magnetic resonance imaging
PA	percent concordance
PCa	prostate cancer
PI-RADS	prostate imaging reporting and data system
PSA	prostate-specific antigen
PZ	peripheral zone (of the prostate)
RADS	reporting and data system
RIS	radiology information system
T2W	T2-weighted imaging
TZ	transition zone (of the prostate)
US	ultrasonography
UX	usability
XAI	explainable artificial intelligence

Chapter 1

Introduction

The prostate gland (prostate) is a reproductive organ that is responsible for the production of the alkaline liquid that carries sperm; this liquid accounts for 30% of ejaculate [1]. The gland is located below the bladder, on the anterior side of the rectum, and measures approximately 25 x 25 x 30 mm. The weight of the organ is estimated to range between seven and sixteen grams in a typical adult man. The size of the prostate increases during two phases: first during puberty, when it reaches its standard size, and second above the age of sixty, when it becomes enlarged and may cause benign prostatic hyperplasia (BPH).

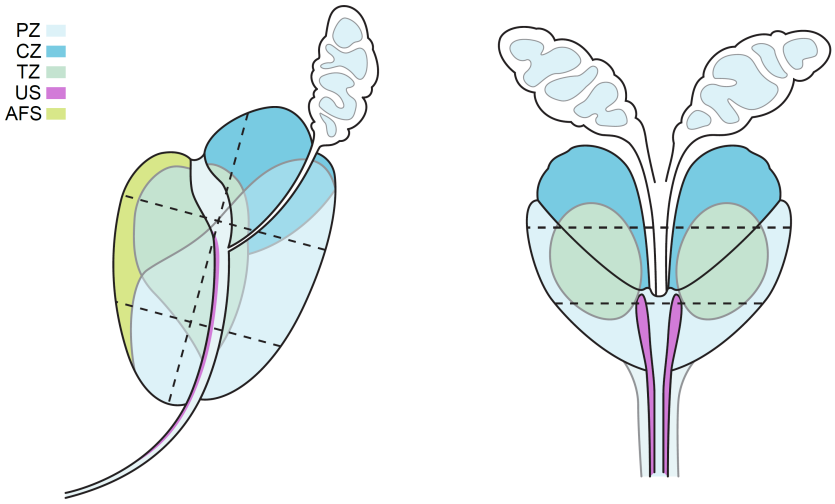


Figure 1.0.1: Prostate zones. Source [2]

The prostate gland is located around the urethra. Anatomically, the prostate is divided into four zones [Figure 1.0.1], according to McNeal's classification [3]:

- the peripheral zone (PZ), which contains 70% of the glandular tissue
- the central zone (CZ), which contains 25% of the glandular tissue

- the transition zone (TZ), which contains 5% of the glandular tissue
- the anterior fibromuscular stroma (AS), which contains no glandular tissue

The part close to the bladder is called the base and the part close to the urethral sphincters is called the apex.

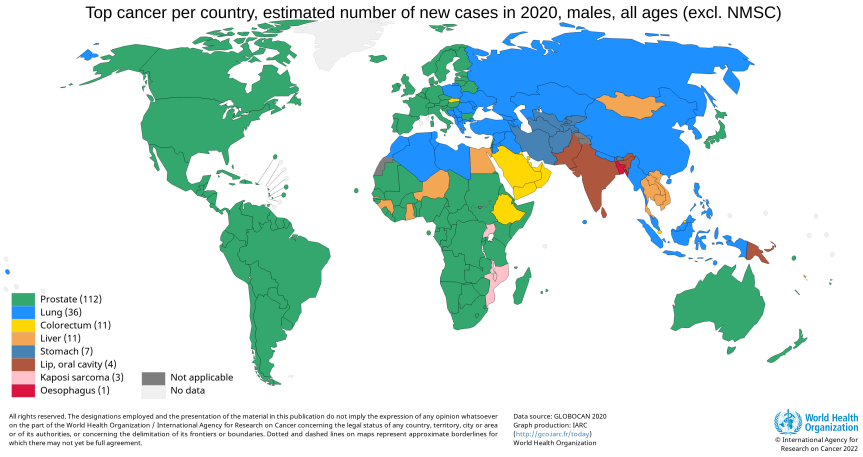


Figure 1.0.2: The most common malignancies in men by country. Source: World Health Organization, GLOBOCAN 2020 [4].

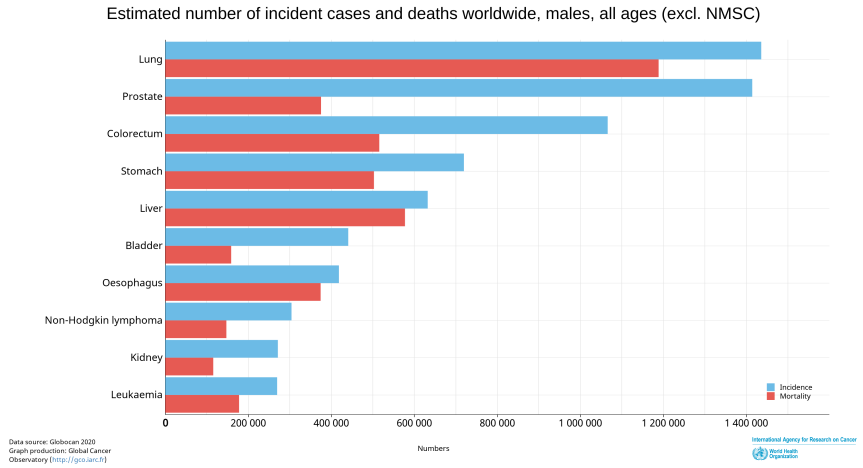


Figure 1.0.3: The most common malignancies diagnosed in men in terms of incidence and mortality. Source: World Health Organization, GLOBOCAN 2020 [4].

Statistically, one in seven males will suffer from prostate cancer (PCa) during their lifetime. It is estimated that prostate cancer contributed to 3.8% of all deaths in 2020; in the same year, more than 1.4 million men were diagnosed with PCa and over 375,000 died from the disease [Figure 1.0.3] [4]. Prostate cancer is the fourth most commonly diagnosed malignancy worldwide. It is the second most commonly diagnosed malignancy in men globally, and the first in Europe, Australia, North, Central and South

America, and parts of Africa [Figure 1.0.2]. Poland reported over 17,000 new cases and over 5,500 deaths from PCa in 2019 [5].

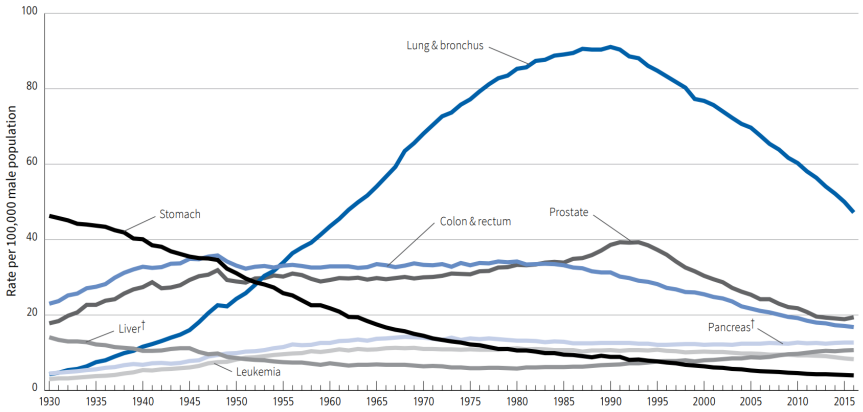


Figure 1.0.4: Cancer mortality rates from 1930 to 2016, based on the data from the National Center for Health Statistics in the United States. Source: American Cancer Society [6]

PCa is caused by cell proliferation of glandular tissue. It is diagnosed in very few individuals under the age of 50 (<1% of all patients) [7]. The average age of PCa patients is 66 years and 86% of all diagnoses are made in patients over the age of 60 [6]. The aetiology of PCa has also been linked to racial ethnicity [8], family history [9], and genetic factors [9]. Moreover, evidence suggests that height [9], [10], obesity [8], and smoking [11] increase the risk of developing clinically significant forms of PCa, while physical activity decreases it [12]. PCa mortality rates have decreased in recent years due to the development of techniques that enable early detection and treatment [Figure 1.0.4] [13].

1.1. Domain knowledge of prostate cancer management

Risk factors alone are insufficient in the diagnosis of PCa; screening tests are also required to ensure early detection and treatment. Currently, prostate-specific antigen (PSA) testing [14] and digital rectal examinations (DRE) [15] are used to detect cancer early. The widespread global introduction of PSA testing has influenced the patterns seen in case epidemiology; this was particularly evident in the 1990s [Figure 1.0.4], when the introduction of screening resulted in a dramatic increase in PCa detection [14], [16].

The classical oncological diagnostic management scheme involves referral to urology clinics. Patients with suspected PCa then qualify for prostate biopsies. A urologist performs a rectal examination, interprets the PSA test result, and refers the patient for multiparametric magnetic resonance imaging (mpMRI) of the prostate, in accordance with the guidelines of the European Association of Urology (EAU) [17].

Decisions on management are based chiefly on the evaluation of serum PSA, DRE, risk group for recurrence, life expectancy, comorbidity, performance status, and symptoms of dysuria [17]. The steps following a diagnosis are established cooperatively by clinicians and patients. Imaging diagnostics—particularly mpMRI—play an increasingly important role in the successful diagnosis of PCa, qualification for prostate biopsies, treatment, and conservative management.

Diagnostic and treatment procedures are invasive, and cause a significant number of side effects and complications. Active treatment options include radical prostatectomy, radiotherapy, hormonal therapy, and investigational therapies (e.g. cryotherapy, high-intensity focused ultrasound, and focal therapy) [17]. Such therapies may lead to urosepsis, urinary incontinence, erectile dysfunction, radiation reactions, psychological disorders, and complications [18]. Patients are presented with different treatment options depending on the stage of their illness. High-risk PCa patients have an increased risk of PSA failure, metastatic progression, and death [17]. The EAU suggests that treatment of patients with limited life expectancy (lower than ten years) may be deferred to avoid loss of quality of life.

Advancements in knowledge and the growing number of variables that have resulted from new technologies impact patient management. Failure to account for all aspects may result in inaccurate assessment and poor therapeutic decisions. The largest issue

facing patients with low-risk diseases is overtreatment. Overdiagnosis of clinically insignificant lesions leads to unnecessary biopsies, a high percentage of referrals to active treatment, and inadequate patient management. A need exists for new diagnostic methods that allow the effective selection of men for active and deferred therapies.

The detection of clinically significant cancer of the prostate gland (csPCa) is a complex process that must be managed carefully. PSA levels and DRE methods of assessment are characterised by their low predictive value and are ineffective in the selection of proper methods of clinical intervention [19]–[21]. Several phenomena mimic PCa by increasing PSA levels in patients' blood or by causing palpable nodules that are subsequently diagnosed as DRE abnormalities. Balancing the PSA threshold is a choice between sensitivity and specificity, which, in practice, results in a large number of unnecessary painful and invasive biopsies, diagnoses of clinically irrelevant cases, and no translation into a decrease in mortality [22]. Although such diagnosis methods are limited, their use is recommended in European guidelines [23] due to their low cost, availability, and universality.

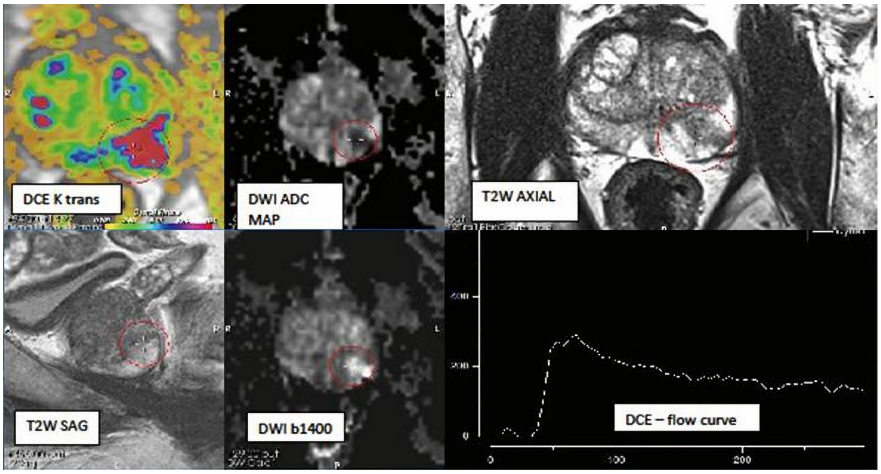


Figure 1.1.1: mpMRI modalities in the visualisation of the prostate gland and lesion assessment. Source: [24]

Multivariate models that employ various other noninvasive biomarkers can help in more individualised risk assessment [25]. Diagnostic imaging, particularly mpMRI, has gained considerable recognition in recent years as a tool that can further improve the detection of csPCa. Such methods also play an increasingly important role in qualification for prostate biopsy, treatment, or conservative management. mpMRI considers the following combination of imaging methods [Figure 1.1.1] [24]:

- **T2-weighted imaging (T2W)** highlights the differences in T2 tissue relaxation time by sequence weighting. While normal PZ tissue is characterised by high signal intensity, the TZ is characterised by heterogeneous nodularity and the possible coexistence of BPH features. T2W is the most useful sequence in an analysis of anatomical features—for example, BPH or lesion extraprostatic extension.
- **Diffusion-weighted imaging (DWI)** involves using magnetic gradients (known as 'b' values) for quantification of the Brownian motion of free water protons. Normal glandular prostate tissues do not constrict water diffusion; thus, showing low signal intensity on high b-value images. Analysis of DWI images in PCa assessment is complemented by visualisation of the apparent diffusion coefficient (ADC) maps, constructed using multiple conventional diffusion images with different amounts of weighting, which present an assessment of water diffusion. Contrary to DWI images, signal intensity on ADC maps correlates inversely with lesion malignancy.
- **Perfusion-weighted imaging (DCE: Dynamic Contrast Enhancement)** is used in the assessment of tumour vascularity through T1-weighted scanning sequences that are performed before, during, and after the administration of contrast agents. PCa, due to neoangiogenesis and more permeable vessels than are found in normal tissue, shows rapid enhancement and early washout; this corresponds to high signal intensity on the DCE images.

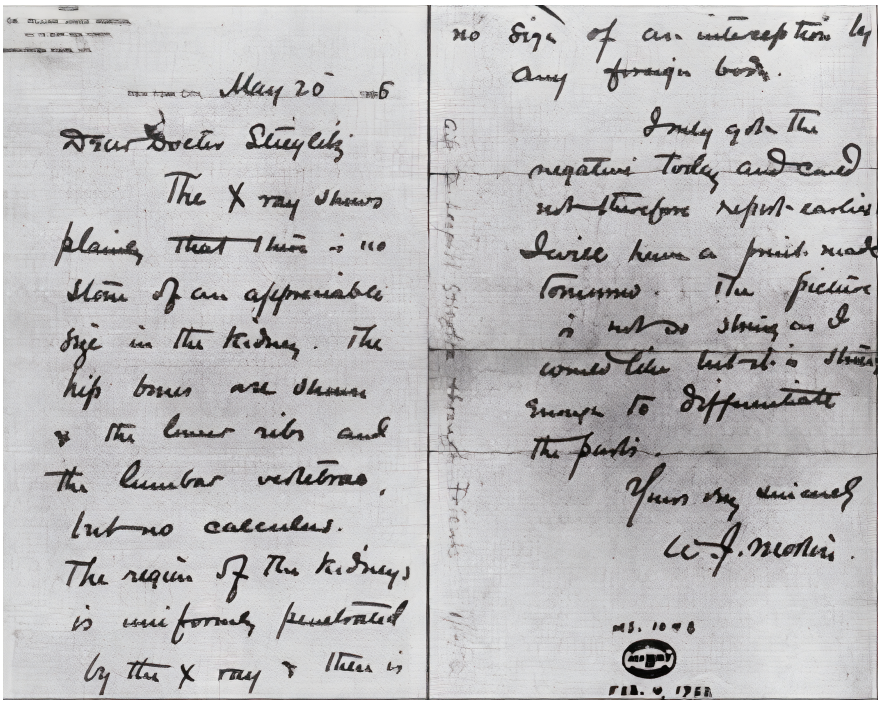
Due to the heterogeneous character of prostate gland tissue, mpMRI modalities vary in their diagnostic value. This depends of the location of lesions[1]. The variable amounts of glandular and stromal tissue in the TZ (often resulting from BPH), creates difficulty in the identification of cancer from T2W images. Lesions located in the PZ can be confused on T2W with abnormalities resulting from prostatitis, haemorrhage, glandular atrophy, BPH, biopsy related scars, or therapy. For such lesions, DWI demonstrates higher diagnostic value because it enhances the regions that display restricted diffusion. Due to the technique employed and the resulting high resolution of T2W, it is frequently used in assessing and differentiating anatomic features, establishing sectoral locations, and measuring both the prostate gland and the lesions

[1] The role of the particular mpMRI modalities in PCa assessment is described in detail as part of the PI-RADS 2019 v2.1 guidelines [2].

assessed. Early enhancement displayed on DCE can be indicative of csPCa, but is not definitive, as similar results can be observed in the case of BPH nodules. Moreover, the absence of enhancement does not exclude the possibility of PCa. On account of the variable kinetics of PCa enhancement, its diagnostic value remains debatable. The specialisation of mpMRI modalities complements PCa assessment; drawing diagnostic conclusions requires a fusion of information on lesions' characteristics.

The use of mpMRI in localisation, risk assessment, and lesion classification is becoming more widespread as more radiologists become experienced in findings analysis and interpretation. Due to its higher predictive performance in comparison to PSA and DRE testing [26]–[30], the role of mpMRI as a screening tool is under consideration. Using mpMRI as a method of noninvasive diagnosis reduces the number of patients referred for biopsies without increasing the number of clinically significant cases missed [31], [32].

1.1.1. Radiological diagnostic standards


May 20 96

Dear Doctor Stuyckz

The X ray shows plainly that there is no stone of an appreciable size in the kidney. The hip bones are shown & the lower ribs and the lumbar vertebrae, but no calculus. The region of the kidneys is uniformly penetrated by the X ray & there is no sign of an interception by any foreign body.

I only got the negative today and cannot therefore report earlier. I will have a print made tomorrow. The picture is not as strong as I would like but it is strong enough to differentiate the parts.

Yours very sincerely

W. J. Morton.

MS. 1078

FEB. 9, 1953

Figure 1.1.2: A radiology report was written in 1896 by James Morton that presents an assessment of X-ray imaging of the abdominal area. It is widely considered to be the first known radiology report [33].

The effectiveness of radiologists' work depends strongly on their experience levels and interpretative skills in the identification and assessment of pathologies on medical

images. Since the dawn of medical imaging, radiologists and other clinicians have communicated primarily through reports. Figure 1.1.2, which refers to the findings of an abdominal X-ray, and assessments of the kidneys and hip bone, constitutes one of the earliest examples of narrative medical reporting. Reports must be presented in a manner that facilitates their clarity and legibility for referring clinicians, as well as providing clear answers to key clinical questions. Choosing the correct diagnosis and prognosis for cancer patients depends not only on image data, but also heavily on the content of radiological reports. The high competence of experts is insufficient for optimal patient management if the findings of radiologists' work are expressed unclearly.

The need for standardisation of medical reporting has been apparent for as long as radiology has been practised. It was expressed in 1904 by Preston Hickey, who postulated the need for nomenclatural standardisation of radiographic descriptions [34] as a result of observations he had made on the individual styles used in medical reports. Hickey's research indicated that the low quality of reports, which frequently omitted crucial information, diminished their clinical value and applicability in diagnosis [35].

The problems resulting from the ambiguities and style variety in medical reporting continue to impede diagnostic processes [36]. The need for standardisation of medical reporting and the resulting solutions is best illustrated in the history of breast imaging reporting. With increased utilisation of mammography [37] in the 1980s, a host of problems arose from inconsistency in reporting and data acquisitions standards, vague descriptions, and ambiguous recommendations [38].

In response, the American College of Radiology (ACR) charged a committee of medical experts with developing guidelines on the reporting of breast imaging to introduce methods for the precise communication of findings [37], [39]. This work resulted in the first management guidelines: the breast imaging–reporting and data system (BI-RADS). The standard included semantic lexicons, and recommendations regarding the way that mammographic imaging should be conducted, the structure of medical reports, and assessment categories [37]. This was mandated by six standard final assessment codes [40]. Figure 1.1.3 presents two examples of reports prepared according to the BI-RADS standard. Since the establishment of BI-RADS in 1993, it

has become a formally accepted and successful standard component of patient management pathways.

```
Patient's Name :        Patient's Surname:
Patient's Birth date:
Referent doctor:

Bilateral screening mammography.
March 1rst, 1998.

Clinical history : history of family breast cancer screening.
There are bilateral, disseminated fibro-glandular opacities.
The mammograms are compared with the
preceding ones (Wichita Clinic), dated May 6, 1995.

SYNTHESIS
Incomplete examination. Mass circumscribed to center-left part.
An echographic examination is recommended.
The patient must get an appointment for it.

Further investigation is needed.

BI-RADS CATEGORIE 0
```

(a) Structured BI-RADS report, 1998 [41]

```
Patient History:
Patient is postmenopausal. No known family history of cancer.
Last mammogram was performed 1 year ago.

Reason for exam: screening.

Bilateral digital CC and MLO view(s) were taken.

Technologist: XXXX, RT (R)(M)

Prior study comparison: 2008, 2007

The breast tissue is heterogeneously dense.
This may lower the sensitivity of mammography.
NO masses, calcifications or other abnormalities are seen.
No significant interval changes when compared to prior studies.

ASSESMENT: Negative (BI-RADS Category 1)

Recommendation: Routine screening mammogram in 1 year.
```

(b) Structured BI-RADS report, 2009 [42]

Figure 1.1.3: Two sample radiology reports of breast imaging prepared according to the BI-RADS standard. Despite the records having been written over ten years apart by radiologists of different backgrounds who worked at different centres and in different countries, they follow a remarkably similar structure, are based on common terminology, and refer to the concrete assessment codes.

The breadth of applications of radiology imaging techniques has led to a need for the standardisation of other diagnostic protocols. As a result of the advanced work of expert communities, various RADS standards have been proposed for the assessment of specific types of cancer and noncancer pathologies. The standards provide a common terminology for describing radiological findings, grading structure, and classification for reporting and data acquisition. Guidelines are updated periodically by multidisciplinary committees of medical experts, based on the most recent advances in methods of noninvasive diagnosis. Notable examples of RADS diagnostic standards that are currently used include [43]:

- Liver imaging reporting and data system (LI-RADS), which is based on assessment of computer tomography, MRI, ultrasonography, and contrast-enhanced ultrasonography
- Lung imaging reporting and data system (Lung-RADS), which is based on assessment of CT
- CT colonography reporting and data system (C-RADS), which is based on assessment of CT colonography
- Coronary Artery Disease reporting and data system (CAD-RADS), which is based on assessment of CT angiography

- Neck imaging reporting and data system (NI-RADS), which is based on assessment of positron emission tomography , computer tomography, and MRI
- Ovarian-Adnexal reporting and data system (O-RADS), which is based on assessment of ultrasonography
- Thyroid imaging reporting and data system (TI-RADS), which is based on assessment of ultrasonography

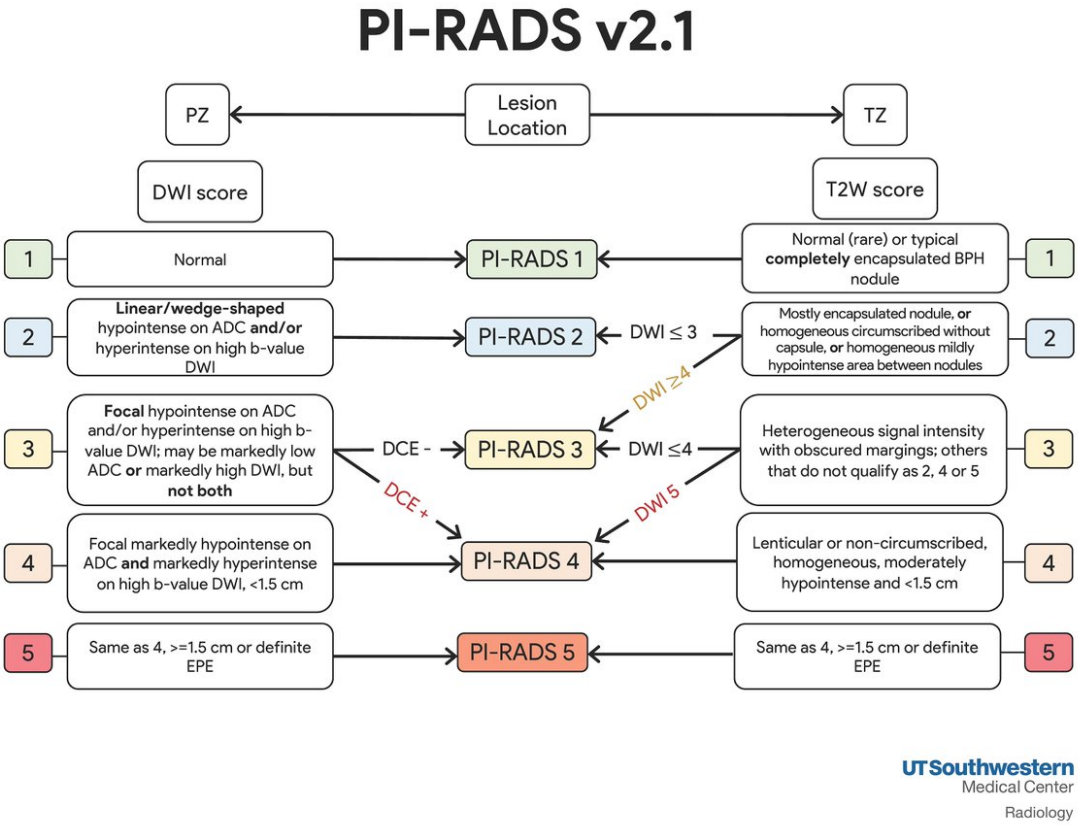


Figure 1.1.4: A PI-RADS v2.1 flowchart prepared by the Department of Radiology at the University of Texas Southwestern Medical Center. Lesion imaging features assessed separately on T2W, DWI and DCE images correspond to the assessment codes that are further fused into the final PI-RADS category based on lesion location.

Diagnostic schemes for PCa are defined within the prostate imaging reporting and data system (PI-RADS) guidelines introduced in 2011 [44]. The system was initially based on independently assessed sequences of mpMRI on a five-point scale. They lacked instructions on how to rate the likelihood of clinically significant cancer for specific lesions [45].

PI-RADS version 2 was announced in 2015 and introduced the concept of a dom-

inant sequence, as well as lesion assessment rules [2]. Since then, the standard has evolved to describe how to report imaging findings and assess the likelihood of lesions' clinical significance on a five-point Likert scale [2], [46]. The assessment rules were developed based on the correlation between certain features of lesions observed on the T2W, DWI-ADC, and DCE sequences and their likelihood of neoplasticity. The most recent update of PI-RADS was published in 2019 as version 2.1 (presented graphically in Figure 1.1.4) and addressed the standard's tendency to score lesions inconclusively by assigning the third category [47].

1.1.2. Importance of PI-RADS in patient care

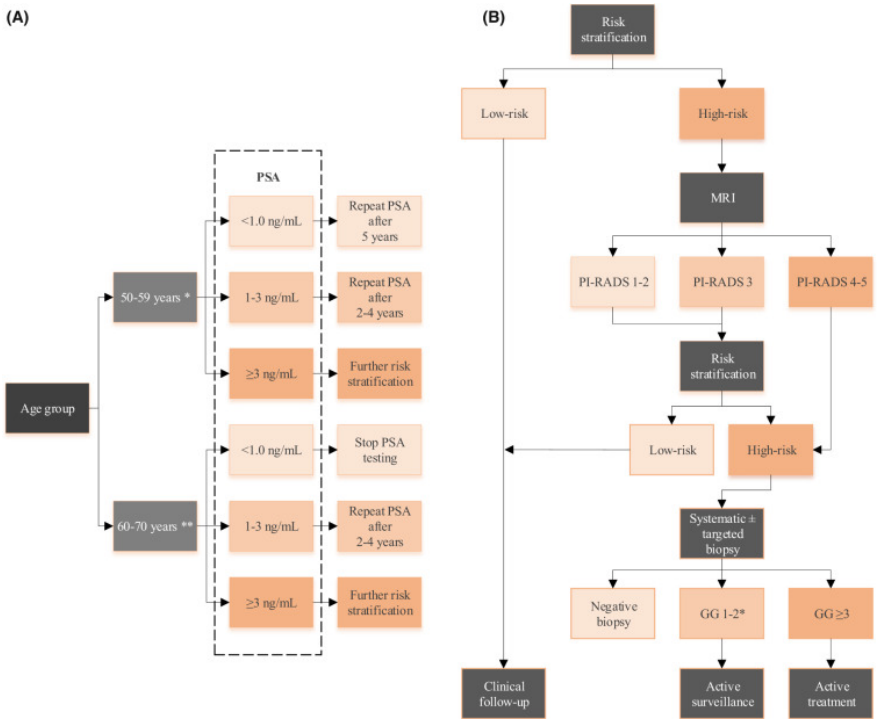


Figure 1.1.5: Part of the EAU PCa early detection pathway. Source: [48]

According to the EAU, the decision to perform a biopsy should be based on mpMRI evaluation using the PI-RADS standard [17]. In the case of a positive MRI (that indicates the presence of a PI-RADS lesion with assigned category 3), urologists decide on subsequent steps in the patient management process with consideration for the patient's clinical picture—including their symptoms and life expectancy [Figure 1.1.5]. A cognitive fusion biopsy involves a targeted biopsy of the focus lesion and a random biopsy of the remainder of the gland. The suspicious lesion is plotted on a diagram of the prostate gland as part of the radiology report, or is viewed using a display adjacent

to the ultrasound machine. Samples are then obtained from sections of the marked regions of the prostate.

Histopathological samples are then rated on the Gleason scale based on their visual assessment to allow distinction of low-, medium-, and high-risk cancers. The clinical significance of PCa is determined by the pathomorphological evaluation of samples, based on analysis of microscopic tissue images [49]. This distinction is crucial due to the nature of neoplastic lesions that occur in the prostate, whose growth may be limited to the prostate gland, and those that may lead to metastasis (most commonly to the bone), which directly affects mortality [50]–[52]. Lowered risk means that there is a lower chance of a cancer progressing and spreading. The Gleason score is a key factor in determining the proper treatment option based on the prognosis of progression.

A correlation can be observed between the PI-RADS scores and the Gleason scores of assessed lesions [53]. This indicates that mpMRI evaluation may help to distinguish low- and high-risk PCa, which is integral in combating overdiagnosis. According to the EAU guidelines, a decision to perform a prostate biopsy is made based on an mpMRI assessment using PI-RADS guidelines that displays high sensitivity but low specificity in the assessment of clinically significant lesions [54]. Focusing on improving mpMRI assessment protocols by increasing their specificity is crucial in limiting the number of unnecessary biopsies, which often result in patient discomfort and avoidable costs.

As a direct result of the developments in PI-RADS, which was designed to improve and standardise the performance and reporting of mpMRI lesions, the standard is now widely known and practised. Since its introduction, it has proved to hold great predictive potential in the assessment of lesion progression and severity [55].

Nevertheless, using the PI-RADS standard entails some limitations. Due to the amount of information generated by multimodal imaging and the relative complexity of the reporting procedure and PI-RADS rules, the standard's use in clinical practice is associated with difficulties and requires high competence. The recent update of PI-RADS addressed the issue of inter-observer variability [56], which was particularly noticeable in the evaluations of less experienced radiologists [57].

1.2. Standardisation of diagnostics processes

Healthcare clinicians and patients alike access radiology reports. This means that reports must be written in a manner that is easy to understand and contains key clinical information [58]. PI-RADS does not specify requirements for the structure of medical reports; the ACR diagnostic protocols outline the scope of information needed in medical reports, but do not specify how the information should be structured and presented. Narrative reports are characterised by considerable subjectivity and high variability in terms of form, language, length, and style.

The likelihood of human error is high due to observer dependency, subjectivity, and a lack of standards in the structure of reports. The use of ambiguous language and assumptions can give the appearance of uncertainty of the contributing radiologist. Problems that occur during the reporting process can negatively affect the patient management processes. The quality of radiological reporting in various clinical applications might be improved by the introduction of tailored structured reporting models.

Structured reporting, which is based on the organisation of text into structured documents that contain dedicated sections, improves the quality of radiological reports. It brings numerous benefits for both radiologists and clinicians by improving the communication of findings, accuracy of descriptions, readability, form, and accessibility. Automation and the adoption of workflow in report generation additionally improve radiologists' work ergonomics. PI-RADS assessments can be improved through the use of structured reporting systems that provide interactive templates for examination descriptions and reproducible final report schemes. Nevertheless, to create effective standardised communication methods and promote interoperability, improvements in the terminology of defined PI-RADS lexicon can be implemented. The terms used in PI-RADS are not unified nor standardised in a way that is shared and reaches beyond the lexicon defined as part of the guidelines.

As medical records, imaging, and reporting are being digitised, a need has arisen for a unified language with which they can be retrieved and compared. Specialists use common, controlled terminology; there is no single unified source of truth, however, concerning the conclusive meaning of those terms. Most of the medical reports in clinical databases are stored in an unstructured, narrative form that allows only rudimentary indexing based on patient admission data. This increases the difficulty

of querying the records, as well as hampering patient management, diagnosis and the performance of research based on clinical data. Standardised terms are necessary to remove the ambiguity introduced both by specialists and by computer systems to medical records.

The standardisation of medical reporting and introduction of a common language to communicate radiological findings could be performed on the basis of domain-specific ontologies. The Radiological Society of North America (RSNA) presents the following definition of an ontology: 'An ontology consists of a standardised set of concepts or terms and the relationships between those concepts'[2]. Those concepts can be expressed in the form of lexicons: catalogues of entries and their definitions that describe particular domains. Radiological lexicons are used to express common terminology used in diagnostic practice. They are prepared and expanded using the clinical experience of radiologists and the results of research on diagnosis protocols. Dedicated communities continuously update lexicons using current knowledge in related fields.

Integrating radiological lexicons with structured reporting systems can solve the current reporting solutions' poor data accessibility. Using structured reporting systems based on standardised terms enables continuous curation of high-quality datasets during clinical practice. Terms can be used for annotating, indexing, and retrieving medical image data. As well as reducing the number of ambiguities that result from radiological reports, tools that employ standardised terminology with well-defined value domains have the potential to improve the communication of findings between clinicians.

RADS focuses heavily on the diagnosis and management of patients through assessment involving analysis of key disease features. Guidelines form as combinations of these features' values indicate the qualities of clinically significant lesions grouped into rules that are assigned to the RADS categories. These protocols act as a representation of diagnostic domain knowledge. Structured reporting systems can integrate the RADS guidelines using the forms of standardised domain knowledge representation—particularly radiological lexicons. These could be further improved by clear definitions, and differentiation of concepts (assessed features) and their value domains.

[2] https://www.rsna.org/practice-tools/data-tools-and-standards/radlex-radiology-lexicon

The exchange of radiology information can be improved using well defined common units of information. This has been achieved by defining radiology terms within RadElement common data elements (CDEs). This formulation enables the integration of accurate definitions of observations in diagnoses [59]. CDEs are key terms; units of information used to describe and standardise application areas. Their definition represents a question that acts as a key, and a set of available answers that are mapped associated values. Using such defined terms allows precise expression of diagnostic observations. For example, the following features can be defined as CDEs with dedicated value domains:

- CDE 'signal characteristics' with three permissible values: 'hyperintense', 'hypointense', and 'isointense'
- CDE 'image quality' with three permissible values: 'adequate', 'suboptimal', and 'non-diagnostic'

Representation of the RADS diagnostics guidelines as formal descriptions entails many potential benefits, including a reduction in the ambiguities concerning rule interpretation and improvement in how updates to the assessment standards are introduced. Moreover, basing the assessment on standardised terminology and domain values enables research on the most suitable definitions of the terms used and their effect on diagnoses characteristics.

Diagnostic guidelines and best practices can be standardised and presented in the form of clinical pathways, a method of describing clinical processes in a standardised way. Clinical pathways benefit from their readability, documentation, communication, potential for optimisation, cost analysis, and quality assurance. Multiple processes that follow the guidelines of PCa management have been modelled as clinical pathways—at least in form of visual representation. For example, the National Institute for Health and Care Excellence (NICE) presents algorithms for diagnosis and staging as clinical pathways [Figure 1.2.1], and the EAU presented pathways of PCa early detection in visual form in its recent whitepaper [Figure 1.1.5]. Clinical pathways are important in the communication of guidelines in a visually readable and easily digestible manner.

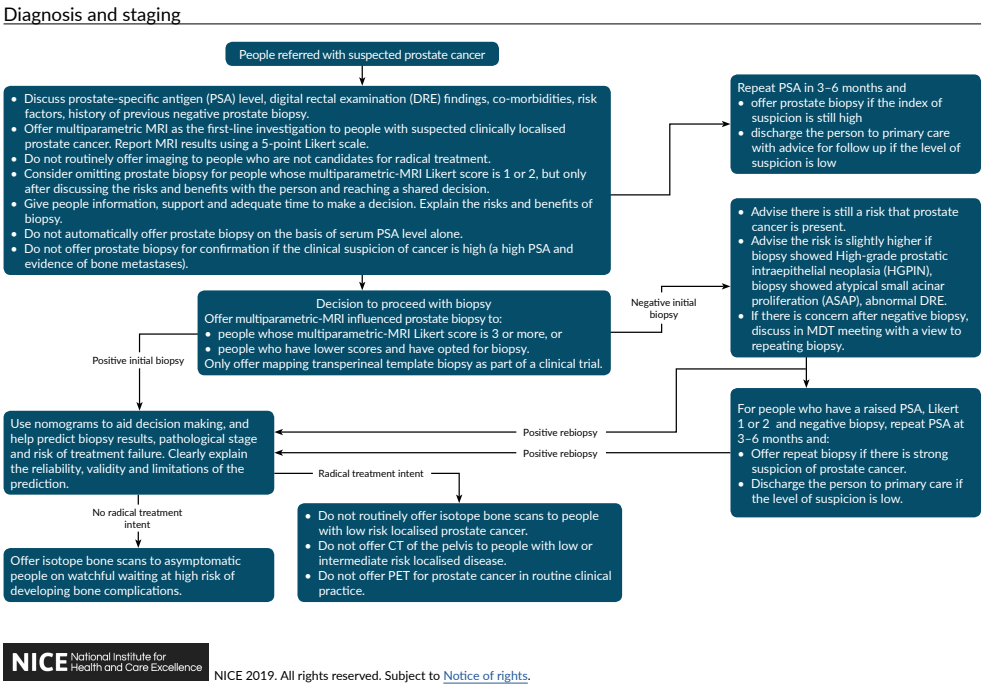


Figure 1.2.1: A simplified flowchart visual representation of PCa diagnosis and staging. NICE presents clinical pathways as hierarchical diagrams on the NICE Pathway platform[3], which enables ease of navigation through guidelines of clinical subprocesses. Source: NICE Pathways [60]

The widespread introduction of clinical practice standards is useful to knowledge engineers in the formalisation of knowledge bases. Diagnostic standards, such as those presented in RADS guidelines can be presented as deterministic decision tables that allow constant updates to be implemented and maintenance to ease. This form of domain knowledge representation introduces a foundation for discussion and agreement on the scope of decision making. Decision tables are commonly used in business to represent rules in both computer- and human-readable forms. The method has also been applied to model clinical practice guidelines, in which the tables serve as a unifying representation of knowledge [61].

[3] the service will be discontinued in spring 2022, as it has become obsolete.

1.3. Integration of domain knowledge in computer-aided diagnosis

Imaging diagnostics has developed in decades alongside processes of digitalisation in medicine. Technology and imaging systems have evolved to the point at which digital images have replaced radiographic films. As a result, the work of radiologists has changed radically, as their workplaces have gained digital diagnostic stations and new working tools. New challenges have also arisen concerning the enormity of data to be analysed — which is often complex and multimodal — and the need to synthesise large amounts of information. In the context of staff shortages and the prevalence of burnout among radiologists [62], modern technologies that improve work ergonomics and enhance radiologists' cognitive processes have adopted a new, crucial meaning.

The zoned anatomy of the prostate gland and its heterogeneous structure causes the interpretation of MRI scans based on the evaluation of multiple sequences to be a multifaceted and tedious task. Formal descriptions can be used by workflow and decision-support systems then integrated into structured reporting systems. This means not only that radiologists benefit from automatic predictions based on inputted information, but also that high-quality data annotations are collected for the assessed images during day-to-day work. Applying this form of assistance during structured reporting has the potential to create a feedback loop in which updates to the diagnostic standards could be introduced as modifications of defined rules based not only on experts' knowledge, but also as a result of analysis of the data that is collected.

Introducing such assistance requires representation of the PI-RADS guidelines as standardised decision tables that are based on standardised terms. Guidelines would have to be expressed using a similar form to that of CDEs: with clearly defined attributes and value domains. Using a defined domain of features assessed during an mpMRI evaluation, guideline rules could be expressed by composing the predefined values into sets that reflect the PI-RADS categories. This process is tedious and requires the engagement of both IT specialists and experienced radiologists to decompose the narrative guidelines into underlying terms and intermediate variables.

Using computerised methods of image analysis in PCa diagnosis has the potential to address the current issues concerning the high subjectivity of image interpretation.

Interest in artificial intelligence (AI) applications in medicine is already high and is growing rapidly [63]. AI is the study of intelligent agents: systems that perform actions based on the perceived environment [64]. The aim of researching and developing AI methods is to provide tools that can match or surpass human performance in particular domains. The construction of such agents can be achieved through several means. Machine learning is a subfield of AI that focuses on using data to improve the probability of achieving defined goals. It utilises algorithms that can generalise from examples, learn from datasets, and adjust certain actions according to the expected input data and output results [65]. For many years, machine learning has enabled the processing and analysis of radiological data.

Machine learning algorithms are utilised in computer-aided diagnosis (CAD) tools: a class of systems that assists specialists in interpreting medical images through imaging registration and the segmentation, detection, and classification of abnormalities [66]. The aim of CAD systems is to provide aid that improves the quality of diagnostic task outcomes by extending human capabilities. Tools only aid radiologist with hints - suggestions resulting from the quantified characteristics of the case, which the expert must complement with the broader clinical context. Machine learning methods methods applied in CAD could be used to introduce objective measures for analysed features, potentially improving interobserver agreement and the robustness of diagnostic standards. The main aspect is to make better use of the capacity of the diagnostician through supportive tools.

Recently, solutions based on deep learning, a subfield of machine learning, have played a major role in the development of CAD. Such techniques are capable of learning high-level feature representations based on analysis of raw data using artificial neural networks: methods loosely inspired by biological neural structures. The development of deep learning has accelerated as a result of the availability of large datasets, and the creation of new algorithms and network architectures. These development opportunities are paired, however, with threats and challenges [63]. In contrast to the current direction of AI development, which is measured by its ability to replicate human performance, research into computational techniques used in CAD focuses on solutions that improve diagnostic processes through meaningful integration that support, enhance and extend capabilities of diagnosticians. Deep learning methods should communicate

diagnostic decisions in a way that is understandable to radiologists, allowing them to benefit fully from AI support.

Introducing CAD methods to assist in the process of medical reporting could lead to a higher quality of diagnosis and affect rates of referral to active treatment. Structured reporting solutions can be integrated with methods of automated image analysis. Machine learning algorithms combined with computer-assisted structured reporting are capable of data extraction and annotation, supporting the process with additional knowledge in the report-writing phase. This approach reaches beyond organisation of report structures based on electronic report template, which merely specify certain image categories and descriptions. The use of such systems introduces protocols that enforce a particular order in how imaging evaluation is conducted and provides radiologists with crucial information that enhances their cognition before final diagnostic decisions are made.

In the use of AI in decision support and knowledge discovery, it is vital that the appropriate CDEs are identified and used to express the variables that influence diagnostic decisions. These should be defined using radiological lexicons, which establish a common vocabulary and provide explicit representation of radiological data. Integration of computational assistance in report generation process helps to eliminate the subjectivity in medical imaging evaluation by providing objective measures to the features being assessed by the diagnostician. In this way, AI can become part of diagnostic guidelines to further improve the specificity of diagnostic procedures. These methods should be developed as convenient tools that act as reporting assistants to enhance radiologists' cognition, improve their workflow and allay tedious diagnostic tasks.

The radiologists assess the features of medical images and base their decisions on guidelines that describe the features that characterise clinically significant lesions. No consensus has emerged on how to optimally design, develop and integrate computational methods that are tailored to application in noninvasive cancer diagnostics based on MRI assessment. The design of those methods would benefit from domain knowledge that derives from clinical practice to enable full integration with diagnostic tools. Moreover, the results of automated medical image analysis should be communicated in a human-readable way to reassure clinicians about the usefulness and reliability of

the technologies used. Domain knowledge must be considered during the design of computer-aided diagnosis methods to provide solutions that can detect and present the inconsequences, contradictions, and shortcomings in radiology evaluations during imaging assessments.

Diagnostic processes could be aided by the integration of computational methods with the structured reporting systems that incorporate the formalised assessment guidelines expressed as rule sets on the basis of the identified CDEs. Expression of automatically assessed features would rely on the standardised terms that comprise radiological lexicons and extend radiologists' cognition by providing objective measures through dedicated image descriptors. This thesis explores these subjects and aims to answer how the domain knowledge contained within diagnostic standards can be used to construct and improve the design of solutions that assist the diagnostic processes of radiological PCa assessment.

1.4. Theses

It is possible to use formalised domain knowledge to enhance the cognitive abilities of diagnostician (use of deep learning solutions to support interpretation processes of imaging examinations) and to improve accuracy of formulated decisions (support of diagnostic report formalisation procedures through imposed order, inspection, and verification of diagnostic protocols, as well as inference of decision-making suggestions).

Experimental verification of the effectiveness of the models and computational methods as well as interfaces and forms of interaction with knowledge resources indicates their usefulness in diagnosis support. The proposed way of integrating domain knowledge models into the real conditions of diagnostic processes enables significant improvements in the efficiency of diagnosticians' work.

The formalisation of domain knowledge in computerised assisted reporting contributes to improving diagnostic procedures. Integration of well-defined, standardised terminology in imaging feature assessment allows research to be conducted on the quality, consistency, and variability of diagnostic procedures utilising datasets curated during clinical use in medical reporting.

Chapter 2

Domain knowledge applied in computational models

Abstract The research described in this chapter aimed to establish whether it is possible to use PI-RADS v2.1 features as sources to identify an effective set of image descriptors. From the proposed set of feature descriptors, an optimal subset was selected as a result of an experimental feature engineering process to construct effective machine learning models that are capable of assessing the probability of prostate lesions' clinical significance. The chapter also presents an innovative method of computerised assessment with the use of deep learning methods of domain-knowledge-inspired architectures. It proposes an intervention in the adjustment of the architecture definition of multi-modal convolutional neural networks (CNNs) using routing. A custom fitness function is also proposed to support the training process and simulate the overall PI-RADS v2.1 algorithm for assessing prostate lesions depending on their zonal location. We found that models based on extracted features and models based on deep learning matched and outperformed inexperienced and experienced radiologists, respectively. Results indicate that introducing changes into the CNN architecture results in faster convergence than the classic multimodal approach does. The chapter concludes that domain knowledge of diagnostic standards can be used to facilitate feature engineering and improve the training processes of deep learning models.

2.1. Introduction

Due to an increase in the demand for diagnostic imaging specialists and rising patient numbers, waiting times and the cost of diagnostic results have increased. Interviews conducted with radiologists reveal that interpretation of mpMRI according to the PI-RADS assessment standards is a tedious task that requires approximately thirty minutes to prepare a single examination report. Moreover, as PCa assessment requires specialisation, the quality of diagnosis differs between experienced and inexperienced specialists; this is reflected in the results of multiple retrospective studies on PI-RADS diagnostic accuracy involving raters of varying experience levels [30]. The low interrater agreement can be explained partially by the ambiguity and subjectivity of features contributing to problems using diagnostic standards. These problems can be solved partially by methods of CAD, the use of which has the potential to shorten the time of diagnosis and simultaneously play the role of an additional diagnostician [63], [66]. The success of AI methods in medical diagnostics depends on reliable verification of the methods and tools developed, and the creation of solutions based on the explainable artificial intelligence (XAI) concept, meaning that the reasoning behind answers provided by computational models can be reviewed by their users.

In recent years, multiple machine-learning-based solutions that attempt to automate the estimation of prostate lesions' clinical significance have been proposed. This was summarised in a systematic review by Castillo et al in 2020 [67], in which twenty-seven of the 2,846 articles the authors queried were analysed. Thirteen studies eligible for meta-analysis were included in the review. Most of the papers were published in 2018 and 2019; however, the earliest study included dates back 2013. The median area under the receiver operating characteristics curve (AUC) achieved by the solutions was 0.79 (interquartile range 0.77–0.87). According to the meta-analysis, SVM, the linear mix model, and k-nearest neighbours algorithms demonstrated the highest performance; the authors noted, however, that most studies did not use external sets for validation—making the results incomparable and likely overestimated. Moreover, studies indicated that the features selected for classification tasks are of more relevance to the results than the classifiers themselves. None of the studies tested reported improvements in PCa assessment that employed tools integrated within CAD systems in clinical workflows.

The quality and applicability of the computational solutions proposed in the literature is difficult to compare, as several factors may influence the performance of machine learning models in PCa assessment. First, the definition of ground truth differs between studies; some use prostatectomy results, while others base their evaluations on biopsy as the reference standard. Then, the task itself may be defined in terms of distinguishing clinically significant vs. nonsignificant prostate lesions based on Gleason score (or its estimation), or on International Society of Urological Pathology Grade Group classification. Castillo et al report [67] that most of the studies analysed include ADC and T2W sequences, but DCE was used in only around half of them. The largest dataset included in the studies contained data on 344 patients, while the smallest contained data on only 36 (interquartile range of 71-193).

Radiomics is a subfield of AI methods applied in radiology, which aims to improve the process of image analysis and interpretation through the extraction of features using image descriptors tailored to particular application domains [68]. Such methods use mathematical models to obtain and quantify imaging features and enable the discovery of new disease signatures. Features include first-order statistics as well as shape, size or texture estimation using dedicated descriptors [69]. The advantage of radiomics is that indications can be explained and justified in language that is understandable to radiologists. The linking of patterns in imaging and clinical data enables the steering of a disease's course and accurate prediction of the possible treatment responses. This facilitates the application of personalised medicine into patient management processes. Studies indicate optimistic results of machine learning methods based on radiomics features in PCa diagnosis [67]; however, their lack of reproducibility and validation are considered major challenges.

The vast array of possible descriptors (the image biomarker standardisation initiative [IBSI] defines 169 validated radiomics features [70]) and the variability of methodology and reported performance in the literature on machine learning methods applied in PCa diagnosis cause difficulty in concluding which descriptors are significant. Domain knowledge contained within the PCa assessment standards and diagnostic rules can be used to identify important imaging features assessed during mpMRI interpretation and to design optimal algorithms. Most of the crucial features used in the PI-RADS diagnostics guidelines correspond to signal intensity characteristics, such as

degree, homogeneity, and regularity. This constitutes a valuable insight that derives from established diagnostics practice; one that can facilitate the process of feature engineering and the selection of adequate descriptors.

Unlike solutions based on feature engineering, deep learning methods automatically extract and learn features from raw data. The development of artificial neural networks, a subset of machine learning techniques, has birthed deep learning methods that, empowered by GPU computing power, are capable of producing powerful solutions to modelled problems. Deep learning can be applied to the analysis of types of data that are otherwise too complex to be properly managed and analysed [66]. As input data is transformed through multiple processing layers, such models are optimised to accurately represent modelled problems by learning increasingly more sophisticated features [71]. This process conceals the reasoning processes of trained deep learning models and complicates the revelation of the characteristics of optimised feature descriptors. Contrary to the methods that utilise feature engineering techniques, how to apply deep learning solutions in decision support in a manner in which the reasoning behind the predictions can be understandable to radiologists remains a challenge. The same challenge hinders efforts to discover new important factors that could meaningfully enhance the diagnostic protocols.

In [72], the authors revise the major deep learning concepts pertinent to medical image analysis and summarise over 300 contributions to the field—including the application areas of image classification, object detection, segmentation, and registration. Concerning the current state of the art, the authors present a critical discussion of open challenges and directions for future research. Although the works they present are predominantly research papers, several AI-based software solutions for clinical radiology practice, with CE or FDA certification, are available. The wide range of limitations, weaknesses, and threats that exist for the practical application of AI in radiology remains valid. To learn how to perform specific clinical tasks, deep learning algorithms demand a large volume of training data. Separate problems concern the models, which, if not properly validated during training, may suffer from overfitting and loss of their generalisation capabilities. Moreover, reliable validation requires the participation of professionals to approve or refute the recommendations made by software or algorithms.

The use of AI in radiology has already become a reality [66]. Computational methods used in the analysis of examinations fulfil the role of a second observer, providing a method of reporting unbiasedly. Deep convolutional neural networks have now become the primary tool in computer vision. They comprise multiple layers, which use convolution filters to alter their input towards meaningful output by convolution operations that are used to identify the significant signal patterns. Each CNN is composed of a stack, which includes an input layer and multiple hidden layers that transform the input data into a network output. The hidden layers typically comprise convolutional, pooling, fully connected, and normalisation layers [65]. With the use of CNNs, radiomics is evolving from feature-engineered to non-feature-engineered methods, which involve using deep learning models as feature extractors [68].

Solutions based on deep learning have achieved promising results in applications in PCa diagnostics. Table 2.1.1 presents a brief review of different approaches to csPCa detection using CNNs. It can be observed that the use of CNNs in the diagnosis of PCa is strongly differentiated. Computational PCa assessment tasks can be formulated in two different ways: as classification problems [73]–[77] or as a semantic-segmentation problems [73], [78]–[80]. In the first case, a patch-based classification of suspected tissue samples is typically performed, which retrospectively exploits annotated image patches. The result of this classification is valid for the whole image patch (unified for all pixels). The second approach utilises a pixel-level classification, the goal of which is to assign a label to each pixel, indicating its association to a proper class (typically cancer tissue, normal organ, or background). The selection of basic architecture depends on the task formulation. VGG [73], [76], ImageNet [74], GoogLeNet [75]–[77], and ResNet [76] have all been used in the classification approach. The segmentation approach favours encoder-decoder architectures, and promotes U-Net [73], [80], ResNet [81], SegNet [79], and VGG16 [78] architectures. The selection of a network is usually motivated by the opportunity to use its most interesting features, or by its effective performance in other application areas.

Publication	Cohort	Dataset	Ground truth	Data	Input	Approach	Multi-modal	Base architecture	Reported results
[81]	335 patients (211 PCas and 124 NCs)	301 (189 PCas and 112 NCs) for training 34 (22 PCas and 12 NCs) for testing	MRI-targeted biopsy	1,5T MRI	T2W	semantic segmentation (pixel-level classification)	no	U-Net + ResNet	AUC 0.645 / AUC 0.636
[73]	PROSTATEx training set; 195 patients divided 159/17/19 for training/validation/testing	444 (215 PCas and 229 NCs)for training 48 (23 PCas and 25 NCs) for validation 55 (23 PCas and 32 NCs) for testing	biopsy qualified with an initial PI-RADS assessment	3T mpMRI	T2W, DWI, ADC	classification (patch based)	yes	VGG	AUC 0.944
[74]	172 patients (79 PCas and 93 NCs)	155 (71 PCas and 84 NCs) for training 17 (8 PCas and 9 NCs) for testing	biopsy	3T mpMRI	T2W	classification (patch based)	no	ImageNet (pre-trained)	AUC 0.84
[75]	160 patients (72 PCas and 88 NCs)	1679 (300 PCas and 1379 NCs)for training 120 (60 PCas and 60 NCs) for testing	TRUS-guided biopsy	3T mpMRI	T2W, ADC	augmented classification (cancer response maps)	yes	GoogLeNet	sensitivity of 0.92 at 1 FP/normal patient
[76]	364 patients (276 PCas and 88 NCs)	639 (324 PCas and 315 NCs)for training 274 (139 PCas and 135 NCs) for testing	12-core systematic TRUS-guided plus targeted prostate biopsy	3T mpMRI	T2W, ADC	classification (patch based)	yes	VGG, GoogLeNet, ResNet	sensitivity of 0.90 and specificity of 0.96
[78]	I2CVB prostate subset; 19 patients (17 PCas and 2 NCs)	2356 multi-channel slices (1413 slices for training / 236 slices for validation / 707 slices for testing)	n/a	3T mpMRI	T2W	semantic segmentation (pixel-level classification)	no	encoder-decoder based on modified VGG16	AUC 0.995 (0.894 accuracy and 0.928 recall)
[79]	PROSTATEx training set; 202 patients	824 multi-channel slices (5-fold cross-validation)	biopsy qualified with an initial PI-RADS assessment	3T mpMRI	T2W, DWI, ADC DCE	semantic segmentation (pixel-level classification)	yes	SegNet	AUC 0.834
[77]	364 patients (276 PCas and 88 NCs)	402 pathologically-validated lesions (detailed division n/a.)	12-core systematic TRUS-guided plus targeted prostate biopsy	3T mpMRI	T2W, ADC	classification (patch based)	yes	GoogLeNet (pretrained)	0.84 detection rate and 0.91 precision
[80]	312 patients (250 PCas and 62 NCs)	369 (104 PCas and 265 NCs)for training 88 (33 PCas and 55 NCs) for testing	targeted and extended systematic MRI–transrectal US fusion biopsy	3T mpMRI	T2W, DWI, ADC	semantic segmentation (pixel-level classification)	yes	U-Net	sensitivity of 0.88 and specificity of 0.50

Table 2.1.1: Review of different approaches to csPC detection using CNNs. Abbreviation used: PCas - prostate cancer sample; NCa - non-cancer sample; TRUS – transrectal ultrasound;

Comparing the effectiveness of the models is a challenging and ambiguous task. The possibility of multiple formulations of the PCa assessment task hinders the definition of consistent benchmarking criteria. In addition, the datasets are characterised by different sizes, differently defined 'ground truth', and varying ratios of training, validation, and test sets. The best results were mostly obtained on the smallest datasets. Except for [81], most of the authors used mpMRI data acquired on 3.0 T MRI machines. In addition to three or more mpMRI series [73], [79], [80], the reduced concept of biparametric magnetic resonance imaging (bpMRI) was often used, exploiting only the T2W and ADC series, and eliminating the use of the dynamic DCE series [75]–[77]. In some cases, only T2W images were used [74], [78]. Some studies have attempted to experimentally verify the reduction of mpMRI to bpMRI and its impact on the accuracy of the CNN model [79], [80].

One crucial aspect that occurs widely in medical imaging is the multimodality of the image data. In the case of PCa, multimodality is expressed in the multiparametric form of MRI scans. The problem of multimodal fusion in CNNs was analysed extensively in [82], whose authors proposed various strategies. Most solutions in PCa detection exploit the concept of input-level fusion or of decision-level fusion. In the input-level fusion strategy, multiparametric images are fused before being passed to the network. The most common form of input-level fusion is image registration, in which coregistered multiparametric image series constitute input for network training [73], [79].

The use of an input-level fusion strategy is usually simple and allows analysis of information from different modalities in all layers of a CNN. Decision-level fusion usually assumes the use of individual networks for each multiparametric series [75]–[77], [80]. Each network can learn unique and mutually complementary information from different multiparametric images. This allows the creation of modality-specific feature representations. The results from individual networks are integrated and fused at the classification stage to reach a final decision.

There is no established consensus on how domain knowledge can be best utilised in the design the optimal deep learning solutions; many possibilities exist, however. We propose methods that focus on adjusting the network architectures and facilitating the model training processes. First, the optimal base network architecture can be identified based on the qualities of imaging data and characteristics of diagnos-

tic evaluation—for example, by selecting the multimodal architectures for analysis of mpMRI. The integration of reasoning behind radiology diagnostics guidelines into the CNN models could be performed by limiting the degrees of freedom in network definitions guided by the specifics of applied domain knowledge. To further enhance the training processes, a custom loss function can be used to integrate the characteristics of diagnostic procedures. Using those interventions, it is possible to assist the training process in formulating the internal knowledge representation of classification problems.

We propose applying these concepts to integrate the domain knowledge of the PCa diagnostic guidelines into the CNN architecture. Integration of the assessment algorithms is achieved using the decision-level fusion in adapted multimodal architecture with routing. Network optimisation is enhanced using a complex loss function that includes the results of subnetwork predictions and integrates the overall PI-RADS assessment algorithm (loss depends on the location of the lesion being assessed). The potential benefits resulting from such interventions are investigated by comparing the performance of the resulting model with the baseline model definition.

2.2. Methods

International competitions based on open, public datasets have enabled comparisons of the quality between the predictions that computational models provide. Challenge organisers formulate the problems and provide datasets, which are used by competitors to develop and submit their optimal solutions for evaluation. Using credible reference datasets standardises experimental methodology and enables validation and benchmarking of, and comparison between machine learning algorithms to draw meaningful conclusions on the optimal approach to modelled problems. This makes machine learning and data science competitions an important method of discovery and the sites of initial verification of innovative techniques. To evaluate proposed methods, we conducted a series of studies based on a publicly available reference dataset published as part of a PCa classification challenge. This research methodology allowed us to compare model performance results with other solutions submitted by competitors.

Table 2.2.1: Lesions and their locations in the ProstateX dataset

Lesion location	Not significant	Significant	Total
Peripheral zone	155	36	191
Transition zone	73	9	82
Anterior Stroma	24	31	55
Seminal Vesicle	2	0	2
Total	**254**	**76**	**330**

A PCa classification challenge held in 2017 (ProstateX) provided a way of comparing methods of supporting PCa diagnosis [83]. Competitors' methods were evaluated on the task of differentiating between clinically significant and nonsignificant lesions. We used the ProstateX dataset to develop and validate computational solutions. We then selected thirty-two lesions from the training dataset to evaluate model performance compared with that of experienced and inexperienced radiology specialists using assessments collected during the retrospective study described in the *Methods* section of the third chapter. We tracked the performance of our methods using ROC curve analysis (calculating AUC). This approach allowed us to compare the diagnostic accuracy of our methods with results reported by other competing teams and obtained by human raters.

2.2.1. Defining the radiomics workflows

Our studies on feature extraction methods aimed to develop a comprehensive radiomics analysis pipeline for automated extraction of properties and features from individual lesions, and a radiomics framework adapted for discrimination of clinically significant and nonsignificant prostate lesions.

Analysis of the PI-RADS lexicon reveals descriptions of multiple characteristics of prostate lesions that correspond to statistical and texture features. Abnormalities are assessed by estimation of their focality, uniformity, evaluation of margins (e.g. well-defined, ill-defined, blurred), and signal characteristics (having a higher or lower signal intensity that corresponds to a brighter and darker appearance on MRI). Based on that insight, we selected statistical features (mean, skewness, and range of percentiles able to capture the signal hypo- and hyperintensity), and texture features derived from the grey level co-occurrence matrix (GLCM) [84] as a base for predictions. Research on mpMRI analysis methods indicates that these texture descriptors (e.g. Haralick features [85], [86]), combined with statistical features, can be used to differentiate the PCa of various Gleason gradings [87]–[90] and achieve high performance in classification tasks [67]. Restricting the feature descriptors considered allowed us to limit the degrees of freedom in the research.

First, we focused on developing a simple prototype method that bases its predictions on imaging features that are essential in mpMRI diagnosis. This work resulted in two models that were submitted as entries to the ProstateX 2017 Challenge. Following two prototype configurations were established for evaluation[91]:

Prototype #1 relied on 126 attributes corresponding mostly to statistical features:

- age of the patient obtained from image metadata
- provided prostate zone of the lesion location as a feature
- single voxel signal intensity
- statistical features of the signal intensity for each of the modalities on the whole-slide of provided lesion location and five, ten, and fifteen millimetre lesion margins: average, standard deviation, skewness, kurtosis, and percentiles (5th, 10th, 15th, 25th, 75th, 90th, and 95th)

Prototype #2 involved 72 additional texture features: estimated Haralick fea-

tures (contrast, dissimilarity, homogeneity, ASM, energy, and correlation) of whole slides, and ROI at five, ten, and fifteen millimetre lesion margins.

Lesion locations provided as part of the dataset were used to obtain square, lesion centred regions of interest (ROIs) and to calculate features on 2D image slices with various lesion margins (five, ten, and fifteen millimetres). Capping the margin at fifteen millimetres was dictated by the PI-RADS guidelines, as lesions larger than that size are assumed to be clinically significant. We employed histogram normalisation, rescaling the input values to 128 intensity levels.

The classifier and optimal configuration were established as a result of experimental performance evaluation involving an array of classification methods. We used three-fold stratified cross-validation on the ProstateX training dataset to estimate the performance of the models. Folding the set into three parts was inspired by the relative sizes of the test and training sets. Both prototypes relied on multilayer feed-forward artificial neural networks trained using stochastic gradient descent and maxout activation function. The layer configuration comprised two hidden layers with seventy-five and fifty neurons, respectively. The use of other classifiers and different configurations were also analysed; the artificial neural network achieved superior performance in the validation phase.

Feature extraction based on image patches proposed in prototype methods produces a vast number of features, which causes difficulty in determining the influence of particular descriptors on classification outcomes. The goals of the follow-up experiments performed on the feature selection were to determine an optimal subset of features and to further investigate the potential performance of PI-RADS-inspired imaging descriptors. This study aimed to perform an in-depth analysis on which modalities, features, and normalisation methods contribute to the best classification results, and which carry the best predictive power in the clinical significance of lesions. We expanded the set of imaging descriptors under consideration[92] and established the potential performance of the selected methods of image pre-processing and feature extraction through a series of feature selection experiments.

Following extension of feature set applied in prototype configurations was considered to describe the characteristics of each image sample:

- The co-occurrences computed in four directions were used to compute standard

GLCM texture features: contrast, dissimilarity, homogeneity, ASM, energy, and correlation. For these, we experimented with the degree of spatial relation between the pixels that configured different offsets: the closest (1), the closest and one separated by two pixels (1, 3), and the closest with one separated by two pixels and one separated by four pixels (1, 3, 5);

- Haralick features: angular second moment, contrast, correlation, the sum of squares—variance, inverse difference moment, sum average, sum variance, sum entropy, entropy, difference variance, difference entropy, and the information measure of correlations (1 and 2). The features then were reduced using statistical methods: average, standard deviation, skewness, and kurtosis.

Texture features from MRI images depend on imaging and pre-processing parameters [88]. We considered different approaches to image normalisation to test the effect of image pre-processing methods on model performance (no pre-processing, standardisation, and dividing the signal intensity by the sum of mean and doubled standard deviation [93]). The images were then rescaled to the scale of 255 intensity values. Additionally, we considered the 2D and 3D regions of the margins surrounding the lesions. We also expanded the array of margin sizes considered (ranging from 2.5 to 45 mm) and features obtained from the whole slices presenting the regions. We extracted separate features for all mpMRI modalities and three (sagittal, transverse, and coronal) views in the T2W modality. Image normalisation and pre-processing methods are described in detail in the following published papers, which present the results of our research on feature engineering methods in PCa assessment [91], [92].

Lesions from MRI were analysed at several levels of detail with the extraction of first-order features like intensity histogram distribution and extracted higher-order features that describe different aspects of tissue texture. Given the very large number of attributes, a feature selection step was included using a genetic algorithm [94], which allowed the search of complex feature space for variable interactions. Fitness function was defined in terms of the cross-validation results, whereby the most predictive features were used for the final classification and prediction of relevant pathological features.

To find an optimal, trimmed set of features that allowed us to draw conclusions on which parameters influenced model performance the most, we decided to test feature

set performance based on a simple k-nearest neighbours classifier with different neighbours (n= 1, 3, 5). The model performance (AUC) was defined as a mean three-fold cross-validation result. The best k-nearest neighbours classifier score and feature set configuration were stored for each generation. The following configuration was used for the genetic algorithm: generation quantity 1,000, mutation probability 0.1, mating probability 0.5, and tournament selection. Specimen per generation quantity was equal to the feature vector (for example, the generation counted 200 specimens if a subset of 200 features was considered).

Second-order statistics were obtained from GLCM using the open-source computer vision Python library, Mahotas [95], following the recommendations of IBSI [70].

2.2.2. Integration of domain knowledge into the CNN architectures

To investigate the benefits that insights from diagnostics standards like PI-RADS can bring into the solutions based on deep learning methods, we proposed a method of integrating domain knowledge into CNN architectures. Focusing our work on modifications of one popular architecture allowed us to reduce the number of variables considered in the studies. Literature analysis shows that VGG is a commonly applied architecture that demonstrates high performance in PCa classification tasks. It was also used in one of the top-scoring models in the ProstateX challenge. We opted to use it as a baseline for research on methods of a priori knowledge integration into the network structures and training processes.

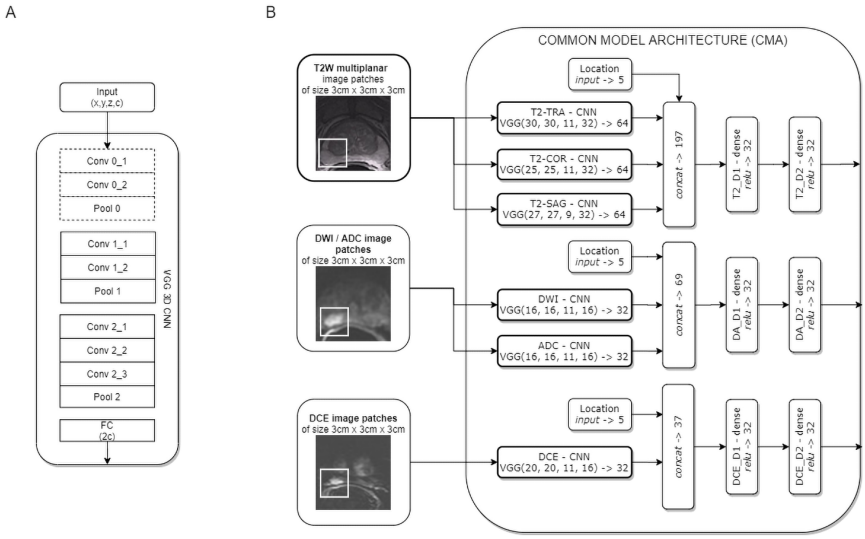


Figure 2.2.1: Common model architecture reflects multi-modal VGG (A) dedicated to T2W, DWI-ADC and DCE sequences. CMA part of the network architectures (B) was the same for the two proposed models that underwent experimentation. Each modality subnetwork has been parametrised [abbreviation 'VGG(x,y,z,c)' is explained in Table 2.2.2]. Source: [96]

We decided to integrate the following setups into the CNN architecture, which were inspired by the diagnostic standard guidelines on feature evaluation:

1. PI-RADS v2.1 deliberately states that the shape and margin features findings should be assessed in at least two planes on T2W MRI; therefore all T2 axes (traverse, coronal, and sagittal) were used for evaluation on separate subnetworks. The system analyses 3D image fragments for all modalities to allow recognition and estimation of high-level features, such as lesions' dimensions, shape, and invasiveness.

2. As in diagnostic standard recommendations, the system analyses mpMRI modalities separately; the input image modalities are not fused and treated in disjunction.
3. The PI-RADS assessment algorithm considers sets of imaging features to establish the probability of lesions' clinical significance. Separate subnetworks composed of convolutional/pooling layers are used for feature extraction designed for the input modalities analysed. Analogous architecture is parameterised to address the different resolutions of T2, DWI /ADC, and DCE images. Outputs of subnetworks that extract features from T2 views (traverse, coronal, and sagittal) as well as DWI and ADC modality pairs are combined before the dense layers. This replicates high-level feature estimation based on the low-level descriptors.
4. Radiological assessment algorithms differ for PZ and TZ lesions, and lesions' zonal locations are considered in the estimation of clinical significance during modality assessment. The zonal position of the lesions is integrated after the feature extraction stage to allow composition of the high-level feature representation, depending on lesions' locations.

Id	Operation	Filter	Strides	Width	Height	Depth	Channels
Conv 0_1	Convolution	3x3x1	1x1x1	2x	2y	z	*c*/2
Conv 0_2	Convolution	3x3x1	1x1x1	2x	2y	z	*c*/2
Pool 0	Max pooling	3x3x1	2x2x1	x	y	z	*c*/2
Conv 1_1	Convolution	3x3x1	1x1x1	x	y	z	c
Conv 1_2	Convolution	3x3x1	1x1x1	x	y	z	c
Pool 1	Max pooling	3x3x1	2x2x1	*x*/2	*y*/2	z	c
Conv 2_1	Convolution	3x3x3	1x1x1	*x*/2	*y*/2	z	2c
Conv 2_2	Convolution	3x3x3	1x1x1	*x*/2	*y*/2	z	2c
Conv 2_3	Convolution	3x3x3	1x1x1	*x*/2	*y*/2	z	2c
Pool 2	Max pooling	3x3x3	2x2x2	*x*/2	*y*/2	$\lfloor z/2 \rfloor$ 1	2c
FC	Average pooling	global	global	-	-	-	2c

Table 2.2.2: Details of parametrised VGG-inspired CMA architecture [Figure 2.2.1] abbreviated as 'VGG(x,y,z,c)', where: x and y = layer width and height, z = layer depth and c = number of channels. Conv 0_1, 0_2 and Pool 0 layers are used only in T2W modality to compensate modality higher resolution rich in textural features. Source: [96]

These assumptions allowed us to devise a baseline multimodal architecture, which, in turn, allowed us to expand and represent approaches to the classification task in respect to high-level features extracted from the input images. The output of each T2W,

DWI-ADC, and DCE subnetwork is a thirty-two-neuron dense layer, which enables interpretation of the models as high-level feature extractors. The whole model, comprising the subnetworks, was named common model architecture (CMA) and further integrated into the classification models [Figure 2.2.1].

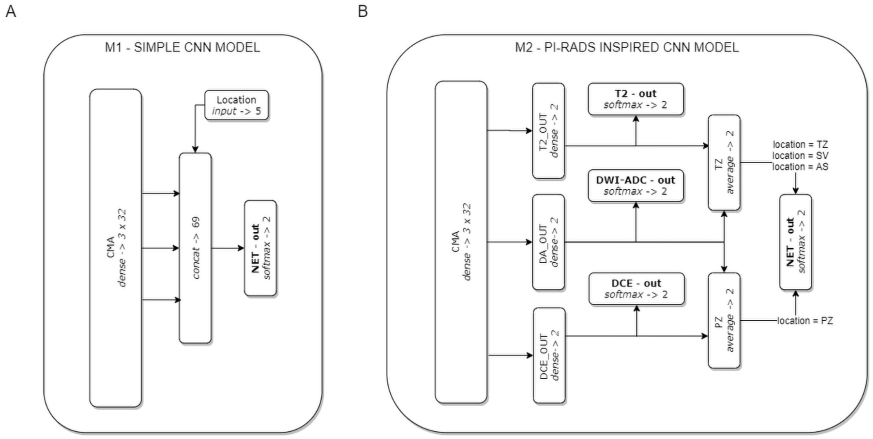


Figure 2.2.2: Diagrams representing architectures of two deep learning models: M1(A) and M2(B) used in the experiments. Source: [96]

We devised two models that can represent two different ways of interpreting the mpMRI images [Figure 2.2.2]. These are based on the same CMA model and include information on the zonal location of lesions.

In the first case, the simple CNN model (M1) represents a classical approach to diagnostics, which assumes that input image modalities contribute equally to the output prediction. This can be treated as analogous to the PI-RADS v1 standard, in which modalities were analysed separately, but the Likert scores were summed to represent an output evaluation. This is common among the models described in the literature, in which information from multiple modalities is fused before the output prediction [82].

To integrate the domain knowledge of diagnostic standards into the model architecture, we used the CMA output for submodalities to form separate predictions and produce the final prediction in regards to lesion location. Instead of one network output, as in the M1 model, the M2 model has four separate outputs that represent the predictions from the subnetworks (T2, DWI-ADC, and DCE-out), as well as the final prediction (NET-out). The final layer of the M2 network integrates routing between two top-level subnetworks that represent classification models for TZ and PZ lesions (as in PI-RADS v2.1 guidelines, in which two separate algorithms are proposed for TZ

and PZ lesions). The softmax activation function is used to estimate the probability of a lesion's clinical significance.

The PI-RADS standard does not deliberately define diagnostic algorithms for evaluation of lesions located in the AS and seminal vesicles; however, to balance data distribution, we decided to include those lesions in the analysis as 'TZ' lesions for the purpose of diagnostic evaluation; this was consulted with radiologists and reflected diagnostic practice. Depending on the lesion location, the network outputs the prediction from the PZ or the TZ (for TZ, AS, and seminal vesicle lesions) subnetworks.

Network	Location	M1	M2
NET	-	100	100
Subnetworks			
DCE	TZ	-	5
DCE	PZ	-	20
T2W	PZ	-	5
T2W	TZ	-	20
DWI-ADC	PZ	-	12.5
DWI-ADC	TZ	-	12.5

Table 2.2.3: Minor weights for the defined complex loss function - total loss is the sum of the weighted average of sub-losses given the lesion location. M1 loss includes only global, NET output. For M2 model, total loss includes auxilary losses resulting from intermediate subnetwork outputs [Figure 2.2.2]. Loss varies for PZ and TZ lesions reflecting the domain knowledge of PCa diagnostics - preference of modalities is dependent on a lesion location. Specific weight values have been set experimentally through analysis of learning curves. Source: [96]

A training process using the minor loss weights was introduced to enable the subnetworks to be trained simultaneously to the integrated model. To achieve that, a complex loss function—defined as a weighted average of subnetwork predictions and final prediction—was devised, in which the weights are preset in regards to the subnetwork type and lesion zonal location. A detailed definition of a proposed complex loss function is offered in a published article that describes our methods [96]. The idea originates from the assumption resulting from the domain knowledge of lesion assessment that T2/DCE and DWI-ADC subnetworks' specialisation should depend on lesion location. PI-RADS standard guidelines emphasise T2 and DWI-ADC modalities for TZ lesions, and the DWI-ADC and DCE modalities for PZ lesions. This is reflected in the weights of the complex loss functions presented in Table 2.2.3. Low

Parameter	Values
Batch size	4, 8, 16, **32**, 64
Training optimization algorithm	**mini-batch SGD**, RMSprop, Adam, Adagrad
Learn rate	0.001, 0.01, **0.05**, 0.1
Momentum	**0.9**
Network weight initialization	random normal, random uniform, **Xavier**
Neuron activation function	**leaky relu**, relu
Weight constraint	0, 0.01, **0.1**, 0.2
Dropout regularization	0, **0.125**, 0.25, 0.5, 0.75

Table 2.2.4: CNN Hyperparameters. Boldened values are considered to be optimal. Source: [96]

weight is also applied to the DCE predictions on TZ lesions and T2 predictions on PZ lesions so that the subnetworks are capable of making their predictions independently. This was implemented to reflect the diagnostic procedures, as lesions are assessed on all modalities during mpMRI evaluation.

We performed a series of hyperparameter tuning tasks while optimising the experimental method. Table 2.2.4 presents the results of the architecture and training process configuration. This proved to achieve the best and most consistent results. The definition of the complex weight function and routing in architecture made the model prone to stagnation and gradient loss, which was solved primarily using a Leaky reLU neuron activation function. Weighted cross-entropy was optimised using the stochastic gradient descent. We found that the use of more sophisticated optimisation algorithms often led to overfitting or contradicted the idea of proposed complex fitness function (for example by maintaining the learning rate for each network weight instead of using a single learning rate [97]).

The data was standardised (Z-score normalisation) and augmented ten-folds using random rotation and was cropped with a 1.5 centimetre lesion margin. As with the feature engineering pipelines, we selected that value as lesions' maximum dimensions reaching beyond 1.5 centimetres is a strong indicator of clinical significance, according to the diagnostic protocols. A smaller margin was not considered, as the three-centimetre ROI size accounted for the potential misalignment of the lesion pixel coordinates provided in the training dataset. Online augmentation was performed during training with random histogram shifting and stretching, Gaussian noise addition, and data flipping along three dimensions.

We trained the models twice for 500 epochs and stored the five-fold stratified cross-validation results for each run, resulting in ten evaluations. The training was stopped at the twenty-fifth, seventy-fifth, and one-hundredth epochs to compare the performance of the models and estimate the model convergence rate. This allowed us to estimate whether regulation of network degrees of freedom and the introduction of the complex fitness function significantly affected the training processes.

The CNN models were constructed and trained using the Tensorflow 1.12.0 [98] for Python framework and evaluated on a local Windows 10 machine with a i7-7700K IntelCore CPU, 32GB RAM, and an NVIDIA GeForce GTX 1080 Ti GPU graphics card. We selected Tensorflow as the framework for the CNN definitions as it enables far-reaching customisations in network and training process configurations using its low-level API.

2.2.3. Statistical analysis

AUC was used as a measure of performance for each of the models. In the study on the integration of domain knowledge into the deep learning architectures, we repeated the model training ten times to perform a statistical analysis of the differences between the performance of the M1 and M2 models. We deployed the Wilcoxon signed-rank test [99] to compare the results. The test set scores were provided by the ProstateX challenge organisers, who performed ROC curve analysis based on the binormal ROC model with AUC as a figure of merit to estimate the potential of methods in reducing the number of unnecessary biopsies [83].

The processes of data cleaning, restructuration, statistical analysis, and visualisation were performed in Python (v 3.6.9) using the Pandas (v1.3.5), Scipy (v1.4.1), and Seaborn (v0.11.2) packages. All scripts were written in the Google Collaboratory tool using dedicated notebooks.

2.3. Results

2.3.1. Performance of the defined radiomics pipelines

Training set cross-validation results indicate that the texture features included in the second prototype model contributed to a significant increase in its performance (AUC = 0.723 ± 0.009) vs. (AUC = 0.692 ± 0.048). These results were confirmed by evaluation on the test set performed as a result of the submission of models during the challenge. A more sophisticated prototype, which incorporates second-order texture descriptors, scored 0.73 AUC on the test set, while the simpler one, which uses first-order features, achieved 0.63 AUC.

Feature		Region	T2-TRA	T2-SAG	T2-COR	DCE	DWI-ADC
Intensity		Voxel	0,678****	0,687****	0,642****	0,668****	0,713****
		2D	0,819***	0,771**	0,812*	0,778*	0,812***
		3D	0,794**	0,804**	**0,836***	0,756**	0,803**
GLCM:	D=1*	2D	0,738**	0,742**	0,807***	0,731***	0,772***
	D=1,3*	2D	0,759***	0,752*	0,821**	0,739***	0,785***
	D=1,3,5*	2D	0,778**	0,749**	0,823*	0,745*	0,777*
Haralick:	Mean	2D	0,786**	0,759***	0,835*	0,743*	0,790*
		3D	0,755**	0,796***	0,833*	0,738**	0,777**
Haralick:	Skewness	2D	**0,840***	**0,840****	0,806**	**0,824***	0,792**
		3D	0,812***	0,807*	0,813*	0,807***	**0,812***
Haralick:	Kurtosis	2D	0,830***	0,786*	0,805*	0,793*	0,805***
		3D	0,785***	0,820***	0,819***	0,785***	0,800*

Normalization method used:
* none, ** method #2, *** method #3, **** no effect of applied normalization

Table 2.3.1: Maximum AUC for modalities depending on used features and normalization method. Source: [92]

Table 2.3.1 presents partial results of the feature set optimisation. First, the effect of feature extraction on model performance on single modalities was analysed. The performance of the models ranged between 0.64 and 0.84 AUC. Overall, the best performing models were obtained on T2W modalities that scored equally in the T2W sagittal and transverse views, followed by the coronal view. Interestingly, the performance of models with DWI-ADC image source was inferior in comparison with the those that based their predictions on features obtained from DCE.

Experiments showed that the models based on Haralick features were superior to others with the exception of the best model configuration for the T2W coronal view

that was optimised to base on statistical features (n = 12). In most of the cases, models of features calculated on images that were not normalised displayed the best performance. Although the qualities of T2W modality images were similar across the views, different Haralick features were selected as optimal for each. The simplest model in terms of the number of features (n = 5) included was optimised for DWI-ADC and included three Haralick features. Model based on DCE was optimised to include twelve features. Details on selected best configuration of features for modalities are presented in [92].

Based on the results of experiments performed on single modalities, we used the model configurations and reran the experiments on a combined feature set of features from subsets of modalities. Then we compared the mean obtained cross-validation results. Combining the features from T2W views enabled improvements in the performance of models that based their predictions on single views, achieving (AUC = 0.88). A combination of T2W and DWI-ADC features achieved superior performance (AUC = 0.90) to the T2W and DCE (AUC = 0.88) images. Overall, the model that produced the best results included features from all of the modalities (AUC = 0.92).

2.3.2. Effects of CNN architecture modification

In this section, we present a comparison of the two deep learning models, highlighting the effect of introducing PI-RADS v2.1 guidelines into the CNN architectures.

Model	Averaged CV results	Test set result
M1	0.831 ±0.019	0.82
M2	0.843 ±0.021	0.84

Table 2.3.2: Validation and test set results for models M1 and M2. The averaged CV results proved to be comparable with the test set results.

Table 2.3.2 presents the results of the deep learning models. The M2 model produced a superior test set result of 0.84 AUC. The Wilcoxon signed-rank test indicated that the differences in mean cross-validation results between the models were statistically significant ($Z = 7$, $p < .05$). Statistically significant differences in the cross-validation-scored model performance were confirmed by the test set results. The results obtained on the test dataset indicate that the analysis of model performance using mean cross-validation results proved to be the most reliable method of establish-

ing model performance; the results obtained on the test set fell within the margin of error of the mean cross-validation results (across epochs).

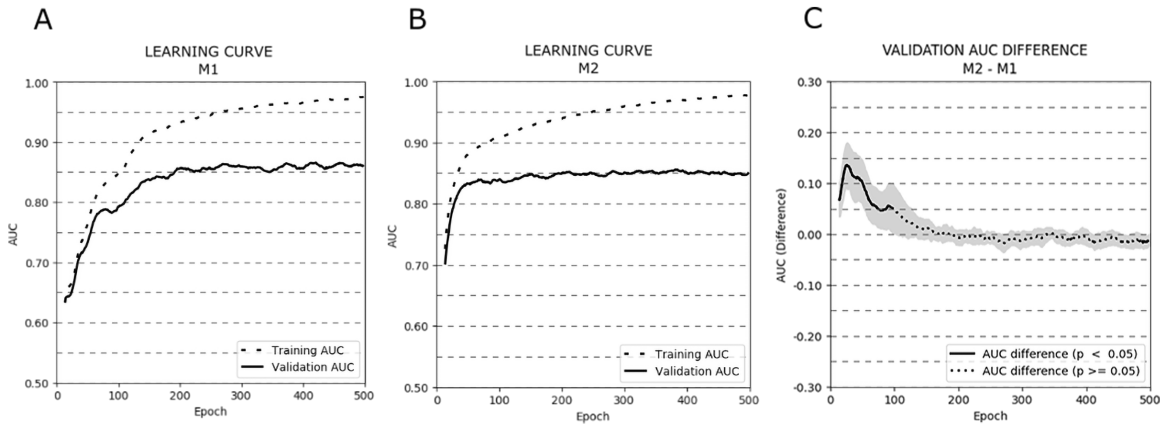


Figure 2.3.1: M1(A) and M2(B) learning curves expressed by mean cross-validation results of multiple model instances (n=10). Differences between obtained performance in regards to the epoch is presented on C. Regions of statistically significant differences are marked with bold line. Source: [96]

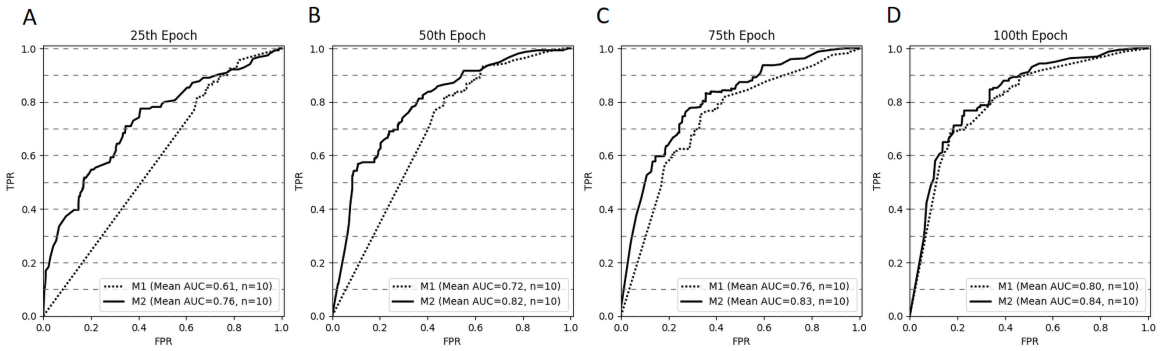
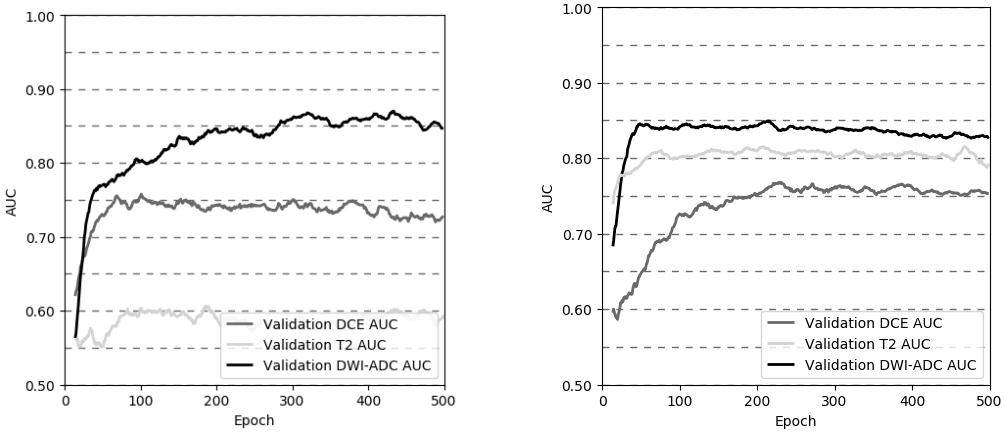


Figure 2.3.2: ROC curve analysis has been performed to investigate predictions of M1 and M2 models by early stopping the training at 25th (A), 50th (B), 75th (C) and 100th (D) epochs. Source: [96]

The effect of the proposed methods of using domain knowledge to optimise deep learning training processes was investigated using learning curves and estimations of model performance at several stages of training [Figure 2.3.2]. Analysis of the learning curves of models M1 and M2 reveals faster convergence in the M2 model. This is visible in the first 100 epochs of training, in which the differences between AUC are statistically significant ($p < .05$) [Figure 2.3.1]. No significant differences in model performance were observed after that. The models demonstrated a significant difference in the performance of 0.15 AUC after the first twenty-five epochs ($Z=0$, $p < .001$), decreasing to 0.1 AUC at the fiftieth epoch ($Z = 0$, $p < .001$), 0.07 AUC at

the seventy-fifth epoch (Z = 5, p < .05), and an insignificant difference of 0.04 AUC (Z = 17, p = .28) at the one-hundredth epoch.



(a) M2 learning curve, PZ lesions

(b) M2 learning curve, TZ lesions

Figure 2.3.3: M2 learning curves of modality subnetworks depending on lesion location.

Using modality subnetworks combined with the complex loss function allowed us to investigate the predictive performance of models based on single modality images. Figure 2.3.3 presents the learning curve of the model M2 PZ and TZ subnetworks. In the case of PZ lesions, performance of subnetworks closely follows the PI-RADS guidelines, in which DCE serves as a supporting modality to the DWI-ADC evaluation. However, in the case of TZ lesions, subnetwork performance indicates that T2 performance is lower than DWI-ADC, which is contrary to the PI-RADS guidelines (according to which, the DWI-ADC should be referenced in the case of inconclusiveness in T2 evaluation).

2.3.3. Comparison with radiology specialists

Using the results from the study that involved inexperienced and experienced radiology specialists evaluating part of the ProstateX dataset, we compared the humans' with the final models' performance.

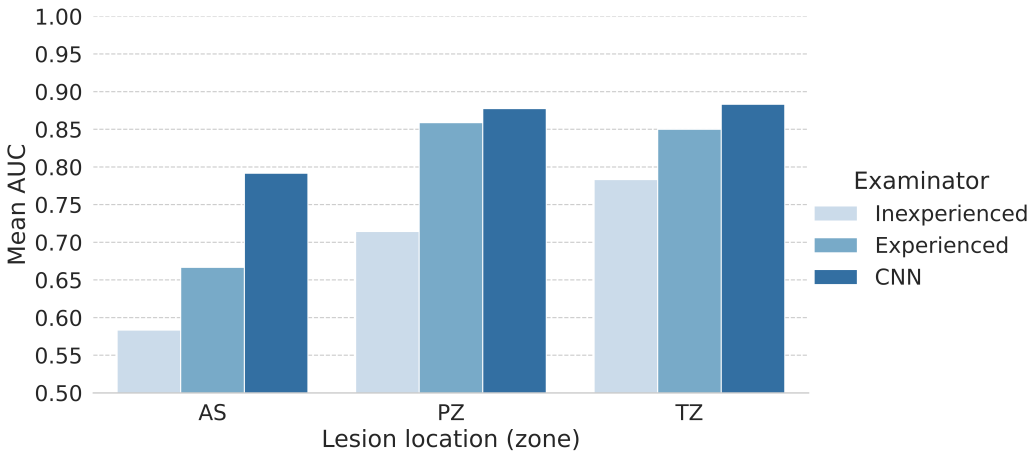


Figure 2.3.4: Mean AUC of Inexperienced, Experienced and CNN raters for AS, PZ and TZ lesions.

The results achieved by the CNN model (AUC = 0.836) demonstrate superior diagnostic accuracy in comparison with both experienced (AUC = 0.811) and inexperienced (AUC = 0.714) specialists in the evaluation of lesions' clinical significance using the PI-RADS v2.1 standard. This is particularly visible in the case of lesions located in the AS [Figure 2.3.4], where deep learning solution provides higher quality estimations of lesions' clinical significance in comparison both to experienced and inexperienced radiologists (AUC = 0.792 vs. AUC = 0.667 vs. AUC = 0.583). In the case of other lesion locations, the differences between the neural network and the experienced radiology specialists were less pronounced: PZ (AUC = 0.878 vs. AUC = 0.858 vs. AUC = 0.714) and TZ (AUC = 0.883 vs. AUC = 0.850 vs. AUC = 0.783).

2.4. Discussion

Experiments demonstrate that it is possible to use diagnostic standards as a guideline to design, construct and improve the computational methods of PCa assessment.

We used the domain knowledge to identify the significant intensity, and statistical and textural features considered during mpMRI assessments. The results of experiments on the engineering of the optimal feature set prove that it is possible to construct effective methods of diagnosis support that are based on first-order statistical and second-order texture features tailored to reflect the specifics of PCa characteristics on mpMRI. The performance predictions generated by the machine learning classifiers based on selected features can be compared to the diagnostic accuracy of inexperienced radiology specialists (based on comparison with a result prototype submitted to the ProstateX challenge that based its prediction on texture features). This indicates the potential of the method in guiding diagnostic decisions by providing significant indicators that correlate with features of clinically significant PCa.

Overall, the results indicate that most optimal models base their predictions on features that derive from all of the mpMRI modalities analysed. This aligns with the PI-RADS standard and diagnostic practice, during which lesion characteristics are inspected on all modalities. The integrated analysis of complementary imaging methods plays a major role in lesion identification and sizing, and estimation of assessed features. The models based on the T2 and DWI image modalities scored higher than the T2 and DCE-based ones on classification tasks. According to the diagnostic standards, DCE plays a minor role in determining PI-RADS assessment when T2W and DWI are of diagnostic quality [2]. The results obtained on feature selection based on the fusion of features from various modalities align with the (recently widely studied) use of biparametric MRI [100], [101].

A study of the effects of the feature selection that constructs an optimal machine learning solution allowed us to investigate the effect of particular features on model performance. The results of the initial studies suggested that similar classification models with texture features scored higher than those based on basic statistical analyses of imaging signal intensity. These results were replicated during the study on model optimisation: texture descriptors yet again proved beneficial in comparison with the sole statistical signal intensity analysis in PCa classification on mpMRI images. The

results confirm that the research reported in the literature suggesting that Haralick (energy, contrast, and entropy) features correlate with the indications of clinically significant PCa on mpMRI images [87], [88]. This can be explained by the significance of lesion and lesion margin imaging features that correspond to the homogeneity assessment in PCa assessment, according to the diagnostic standards. The proposed optimal set of descriptors that correspond to the features estimated by radiologists during mpMRI assessment can be integrated with CAD systems to aid diagnostic decisions by introducing objective measures and indicators of imaging features.

Nevertheless, our study involved limitations. The methods of feature selection and model optimisation that are based on the maximisation of stratified cross-validation results using genetic algorithms are prone to overfitting. This is visible when analysing the selected optimal configuration of features included in the best DCE-based model. The PI-RADS standard bases its predictions on this sequence mostly on signal intensity analysis and acknowledges its poor diagnostic accuracy [2]. DCE image modality is low resolution, and the benefit of in-depth texture analysis in its case is doubtful; however, the selected optimal configuration shows otherwise as it uses ten texture descriptors. The accuracy of feature selection for DCE is doubtful, as the best DWI-ADC model uses only five texture descriptors, despite the modality being richer in information due to the imaging method.

The AUC for the seventy-two methods of the groups that competed in the 2017 ProstateX PCa classification challenge ranged between 0.45 and 0.87 [83]. There was one winning group (AUC 0.87) and two groups that tied for second place (AUC 0.84); however, no statistically significant differences were observed in performance among the top-performing methods [83]. The top-scoring models submitted to the ProstateX challenge were based on CNNs [102], [103]. This can be explained by the complexity of prostate lesion assessment based on the evaluation of compound features like lesion shape and invasiveness [97]. This, along with the results achieved by the machine learning methods illustrate that more sophisticated solutions that are capable of assessing the high-level features are required to create effective computational methods of PCa assessment.

We demonstrated that the domain knowledge contained within the PI-RADS v2.1 diagnostic standard can be used to improve the deep learning solutions. This was

achieved by modifying the architecture and introducing a complex loss function. The deep learning model with architecture that reflects the PI-RADSv2.1 guidelines achieved a score close to the best models submitted during the ProstateX challenge. Although test results demonstrated superior performance of the model in comparison to the baseline, the differences between the models were minor (0.02 AUC). This was expected, as both models are similarly complex and carry the potential to discover patterns within the imaging data.

The experiments highlighted a mismatch between the PI-RADS v2.1 guidelines in establishing the base and supporting modalities in lesion evaluation and the performance results of modality subnetworks in case of TZ lesions. This can be explained by the key role that DWI-ADC plays in distinguishing clinically significant and insignificant lesions in cases of inconclusive evaluation [47].

The primary advantage of the method of a priori knowledge integration was illustrated by its faster convergence than the experimental method. Significant changes were introduced, which affected the training process and allowed the method to achieve the same results in a lower number of iterations. This is an important discovery, which demonstrates that domain knowledge of PCa diagnostics is correctly represented in the architecture. Its role can be observed in its increased robustness and stabilised learning. Accordingly, domain knowledge played a role in network regularisation. Faster convergence, and thus learning, can lead to lower numbers of iterations with potential benefits in hyperparameter tuning. Decreasing training time allows us to search more network configurations and maximises predictive accuracy.

Comparing the neural network model's performance with that of radiology specialists based on the assessment of a subset of suspicious lesions leads us to conclude that the deep learning model demonstrates diagnostic accuracy in predicting clinically significant PCa at least at the level of experienced radiologists. The largest differences were observed in the assessment quality of lesions located in the AS where the PI-RADS diagnostic guidelines do not provide a tailored algorithm for clinical significance evaluation. This exemplifies the potential of introducing methods of computer aid to diagnosis and expanding diagnostic protocols through various applications. The predictions of sophisticated models can be complementary to PI-RADS evaluation: estimated probabilities of lesions' clinical significance can be included as part

of radiology reports, expressed in the form of degree of confidence in cases in which assessment is inconclusive. Integrating the model that utilises subnetworks capable of forming predictions on mpMRI modalities separately aids diagnostic decisions on all partial PI-RADS assessments. Moreover, the method could be extended to estimate imaging descriptors using the dataset that involves the estimation of features assessed during the diagnostic process by the radiologists. Both constituting features and final PI-RADS levels could be predicted and assigned credibility scores based on the objectively measured features. That could lead to greater explainability and credibility in diagnostic processes.

Although the differences between the diagnostic accuracy of the deep learning models and those of inexperienced and experienced radiologists are statistically significant, this comparison has limitations; only thirty-two prostate lesions were assessed during the process. To settle that matter, the experiment should be repeated on a larger annotated multicenter mpMRI dataset. An additional evaluation should be performed that incorporates models integrated with the PCa CAD system to verify the methods' clinical applicability.

2.5. Conclusions

Computational methods designed for the evaluation of lesions on mpMRI images that are based on texture analysis have been repeatedly proved to be viable and optimal in the recognition of PCa. This has been demonstrated by multiple studies, including the most recent ones that employ publicly available datasets. The ProstateX challenge allowed comparisons to be made between multiple approaches to the application of AI in PCa diagnostics. Our research on solutions that base their predictions on statistical and textural features suggests that domain knowledge described in diagnostic standards offers crucial insight into the construction of machine learning models. RADS guidelines can be used as a source of inspiration in the selection of imaging descriptors to implement an optimal method of automated diagnosis. This concept reaches beyond prostate evaluation to diagnosis of other cancer pathologies on medical imaging that employs magnetic resonance.

RADS standards include more features in their guidelines that are high level and require more sophisticated solutions. Texture features alone cannot represent the complete array of rules in the PI-RADS standard, which include high-level descriptors, such as lesion dimension, shape, and invasiveness estimations. This is confirmed by the high efficiency of solutions based on deep learning, which can automatically discover and represent high-order patterns.

Recent advancements in the application of deep learning in PCa diagnosis present a variety of approaches that employ CNNs . Our study failed to demonstrate any clear benefits of domain knowledge encoding in terms of improving model performance. This was expected, as the number of possible states between models was similar—as was their capacity to learn, recognise patterns, and form problem representations. Instead, our research illustrates that encoding domain knowledge in neural architectures brings benefits during training in the forms of stable performance and faster convergence.

To conclude, if the method of integrating domain knowledge into neural network architectures can be generalised, the experiments should be repeated using other analogous domains of cancer diagnostics. Similar architecture modifications could be made to networks that simulate other RADS standards.

Chapter 3

Structured reporting with integrated formal descriptions

Abstract The research described in this chapter aims to investigate the effects of introducing assistance to diagnostic processes in the form of structured reporting based on the standardised lexicons. Domain knowledge of PCa diagnosis was integrated within the system as constructed decision tables that model the PI-RADS v2.1 guidelines. Based on the defined lesion properties, the tool automatically estimates the assessment category for the given modalities, and, using the assessments verified by radiologists, generates an overall category. The system's usability was examined during research involving both experienced and inexperienced radiology specialists on retrospective data and in clinical environment as part of the diagnostic procedure. Data curated during the interaction of users with a structured reporting system was used to examine the agreement between raters and their diagnostic accuracy. The experiments described in this section were performed on the web-based eRADS platform, which is specially designed for research on the standardisation of medical reports.

3.1. Introduction

The radiological reports of prostate mpMRIs play a significant role in patient care, as according to the EAU, the choice to perform a biopsy should be based on the radiological findings. The PI-RADS radiological report of a prostate mpMRI contains an assessment category PI-RADS score, which is used—along with factors that contribute to the patient's overall clinical picture—by the referring clinician to decide subsequent steps in patient management. Depending on the radiologist's experience and the terminology used to convey the relevant diagnostic information, the structure of the report—as well as its readability and utility—varies greatly and can directly contribute to impeding treatment outcomes.

The potential of structured reporting in radiology was investigated in a recent integrated review that analysed 223 publications on the subject to select thirty-two for in-depth evaluation [33]. Multiple aspects of structured reporting were considered and studied within the relevant literature: evaluation of terminology, accuracy, completeness or consistency (twelve papers), clarity, readability or quality (nine papers), efficiency, and effectiveness (six papers). Six papers attempted to establish whether structured reports helped in clinical decision making, and only three papers investigated whether such reports brought any benefits to patients' health. The conclusions of the papers included observations of increased accuracy, integrity, consistency, clarity, readability, and overall quality when structured reporting was used. Introducing the tools of structured reporting made reports more effective and brought improvements to patients' health.

Studies show that structured reporting has the potential to reduce the time necessary to prepare reports, decrease the incidence of errors, and improve the quality of diagnoses [33]. By improving the specificity of applied methods of PCa assessment, it supports clinical decisions by reducing the number of invasive diagnostic procedures and improving the share of patients referred for active surveillance instead of active treatment. A reduction in unnecessarily performed prostate biopsies decreases patients' discomfort and reduces the number of complications that result from such procedures [104]. Both reductions in the time needed to prepare reports and the potential to avoid active treatment have positive effects on the social and economic costs of PCa management.

An overview of structured reporting in radiology was presented in a paper by the European Society of Radiology [105], in which a three-level structure of reports was proposed based on the previous work by Weiss and Bolos [106]:

1. The structured format includes paragraphs, subheadings that provide sections designed for clinical information, examination protocol, radiological findings, and conclusions
2. Consistent logical organisation of reports based on ordering the information using an internal logical order
3. Use of dedicated terminology related to domain ontologies

The inclusion of integration with standardised lexicons in the definition of structured reporting enables the indexing of information resulting from imaging interpretations and the reusability of radiology reports. The ease of creating interactive forms has resulted in a range of solutions for building structured reports being available. One example is the RadReport application [107] under the supervision of the RSNA, which acts as a source of standardised report templates based on the best practices in reporting. Key clinical observations are captured on the resulting reports using the terminology, measurements, technical parameters, and annotations appropriate to the problem domain.

Many reporting solutions are simple and based on forms with inputs dedicated to reporting sections. Reports generated from those systems are based on the intermediate variables defined, but often are exported in text form to radiology information systems (RISs): a class of systems that support the management of medical imagery resulting from the clinical workflow. Formulating the diagnostic conclusions in narrative textual form does not enforce report completion nor compatibility of structure. It also does not allow means of aiding diagnosticians in the generating reports by providing suggestions and hints regarding possible mistakes. Documentation stored in text form is stripped from the important annotations resulting from variables that underline the diagnostic processes. The data specified during report generation that could be used to curate the datasets for research purposes is lost. Such data could be only prepared by manual report text decomposition, which is tedious and prone to error.

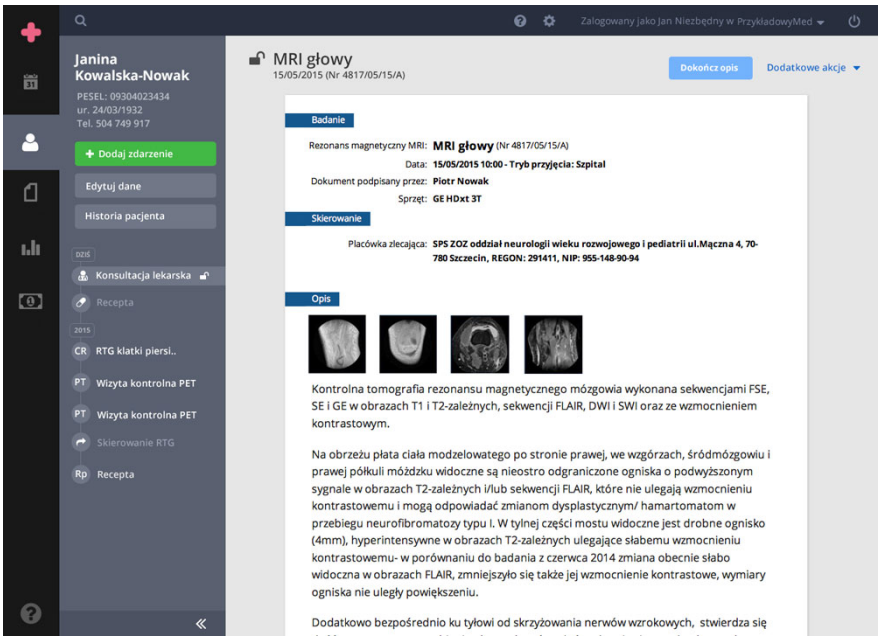


Figure 3.1.1: Report of head MRI examination presented in the Meddo+ RIS. The generated narrative report does not include the intermediate features assessed during the diagnostic process, as the reporting module supports report generation using predefined report templates. Source: Meddo24[1]

Moreover, not all variables assessed during the reporting stage are included in final reports. Unstructured reports in text form do not contain all information that is considered when making a diagnosis decisions [Figure 3.1.1]; the assessment of the intermediate variables that describe lesion characteristics is one prime example. In the case of PCa assessment, reports present only dimensions of lesions, their locations, and the estimated probability of their clinical significance expressed by a score on the five-point PI-RADS scale. They do not contain information on homogeneity, level, nor type of signal intensity that characterises lesions. As a result, information that could be used to expand clinicians' knowledge on the efficiency of diagnostic procedures is lost.

To maximise the benefit of structured reporting, the data that results from imaging interpretation must be integrated with full clinical information to enhance clinical decision-making and, therefore, the quality of diagnosis and patient management. Structured reporting should be integrated with methods that allow storage of research data to make imaging data more accessible, quantitative, uniformed, and structured [108]. Data must be stored in a way that allows effective querying; this can be achieved

[1] https://www.meddo24.pl/system-ris-funkcje

by indexing medical images using the significant variables assessed during medical reporting. This enables the introduction of tools of statistical analysis, knowledge discovery, data mining, and integration with AI and clinical-decision-support systems [109] to improve the quality of diagnostic decisions. Analysis of curated data can be used to evaluate and compare radiologists' performance, optimise the reporting process, introduce quality control, and provide departmental quality indicators [110]. The interoperability required to integrate the clinical data can be achieved by applying the common nomenclature of structured reports that goes beyond the domain of diagnostic standards for given pathologies (i.e. the PI-RADS lexicon) to the generalised concepts of radiological image assessment.

RADS provide standardised terminology for imaging results, which reduces the reported variability and improves the communication of findings [109]; each system, however, is based on disease-specific lexicons. This stands in the way of data interoperability by introducing ambiguities in definitions, and interpretations that result from a lack of standardisation across the widely recognised lexicons. For example, the interchangeable use of the terms 'invasiveness' and 'aggressiveness' in the description of lesions assessed during prostate mpMRI examinations is caused by ambiguity in the terms' definitions across problem domains, and strongly depends on a specialist's experience.

Structured reporting could result in improved diagnostic decision specificity while improving, or at least maintaining, the degree of sensitivity. This outcome can be accomplished in mpMRI assessments by improved adherence to PI-RADS guidelines by the standardised structured form of reporting, which is based on an established conceptual terminology and formalised rules. This can be achieved by the expression of RADS guidelines using the concept of CDEs, and by improving the process of data organisation and management in structured reporting.

The efficient exchange of information between reporting systems and structured databases is crucial for the reproducibility of reports and simplified data exchange, and can facilitate knowledge extraction and adaptation of clinical decision support. Many forms of standards, lexicons, and term sets currently utilised in radiology include:

- ACR Index, created in 2004 as an online system that allows specialists to access

the ACR index for radiological diagnosis, which was created for the indexing of image-based teaching files [111];

- SNOMED-CT, created in 1999, contains over 350,000 concepts organised into hierarchies, with unique meanings and formal logic definitions [112];
- LOINC (Logical Observation Identifiers Names and Codes), developed and updated since 1994, contains over 71,000 'observation terms', each record includes fields used for unique specifications [113];
- RadLex, proposed by the RSNA, is a specialised ontological radiology lexicon that contains imaging terms and their relationships with each other [114], adopts the best features of existing terminology systems and currently contains more than 34,000 terms [115];
- ICD (International Statistical Classification of Diseases and Related Health Problems) is a medical classification created by the World Health Organisation (WHO) (On 1 January, 2022 the last revision, ICD-10, which contains 14,400 positions, was replaced by ICD-11, which contains over 55,000 diseases [116]);
- DICOM (Digital Imaging and Communications in Medicine), which has functioned since 1993, was developed by the ACR and the (US) National Electrical Manufacturers Association to organise methods of medical data exchange and interpretation, and is used chiefly in medical imaging [117].

The RadLex lexicon—which was designed to provide radiologists and educators with an online index of educational materials—demonstrates the most promise for application in radiological structured reporting. Radiology experts have expanded the domain of defined terms to create a unified source of standardised terminology that can be referred to by researchers, clinicians, and developers. The RadLex lexicon aims to form a single source for medical imaging terminology that integrates other medical terminology systems, such as SONOMED-CT and ICD while addressing the missing areas through continuous extensions. Systems that use RadLex facilitate the analysis of radiological data and allow the uniform indexing of medical datasets [118]. The defined set of radiology terms can be used not only in knowledge discovery and decision-support systems, but also in education and research. Using the CDEs based on RadLex definitions in the standardisation of diagnostic domain knowledge

unlocks the opportunity to utilise formal descriptions that can be used by workflow and decision-support systems to ensure adherence to diagnostic guidelines.

Clinical practice guidelines have been already formalised for various cases using the Arden syntax, a language released in 1992 that is applied for representing and sharing medical knowledge and is used as an executable format by clinical decision-support systems. Rules are represented in the form of medical logic modules (MLM) that assist in medical decisions [Figure 3.1.2]. The program's code is designed in a format that is readable to humans and resembles natural language, and its structure makes it easier to be understood by non-programmers. Using Arden-based decision rules allows interpretations to be generated and communicated to clinicians. It is currently part of Health Level Seven International, a health data interoperability standard. Scripts written in Arden can be understood and validated by experts in a particular clinical field, but their development and verification requires some programming knowledge.

```
maintenance:
    mlmname:    UTI_SUTI;;
    arden:  version 2.5;;
    [...]
knowledge:
    [...]
    data:
        (Stay, Date) := argument;
        Temperature :=
            read {temp (Stay, Date)};
            /* Body temperature. */
            if Temperature >= 38
                then Fever := true;
            endif;
            Urgency :=
                read {urge_urinate (Stay, Date)};
            /* Urge to urinate? */
            Micturition :=
                read {mict (Stay, Date)};
            /* Increased frequency of urination? */
```

(a) Part of the variables specification section

```
then Urine_culture := true;
    endif;
    ;;
    evoke:
    ;;
    logic:
        UTI_SUTI := (Fever OR Urgency OR
                Micturition OR Dysuria OR
                Suprapubic_tenderness)
                AND Urine_culture;
        conclude true;
        ;;
    action:
        return UTI_SUTI;
end:
```

(b) Part of the logic underlying the decision support

Figure 3.1.2: An example of a medical logic module categories in Arden Syntax. The example shows a module that helps determine whether a patient has symptoms of a urinary tract infection (UTI). [119]

Management information systems have introduced clinical pathway solutions analogously to business processes in the form of business process modelling notation (BPMN) and Unified Modelling Language diagrams, flow charts, process chains, or dedicated languages. Some initiatives promote BPMN as a tool to model clinical pathways, as it is a widely recognised standard in industry, used for modelling business processes. Its widespread use makes it possible to allocate human resources to the highly demanding task of modelling clinical pathways. Moreover, using the standard brings additional benefits in the form of possible deployment on dedicated workflow engines in an ex-

ecutable form. BPMN is well suited to the modelling of clinical processes due to its widespread use and support. Some advocate BPMN as the dedicated modelling tool for clinical pathways in the academic community; for example, Braun et al recommend the extension model, BPMN4CP as a tool dedicated to clinical pathway modelling [120], [121].

Another form of modelling standard, decision model and notation (DMN) is a language used to define the precise specifications of business decisions and business rules. Tables are represented by sets of rules based on the attributes introduced and their values. Correct rules are found to be applied for input cases based on the hit policies (resolution strategies that match the input parameters towards finding the first matching rule).

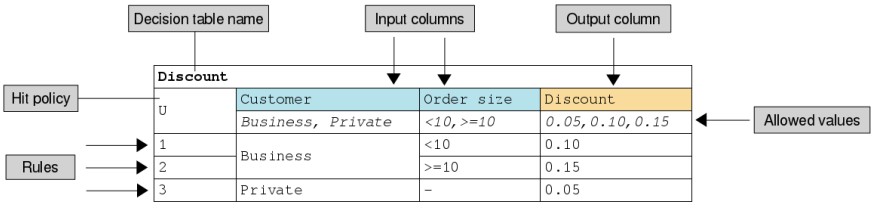


Figure 3.1.3: An example of a DMN decision table that selects discounts based on customer type and order size. Source: [122]

Decision tables benefit from their use of visual interpreters and editors. Many common software modelling tools, such as Enterprise Architect[2] and Camunda Modeller[3] support the DMN standard and allow editing rules by directly modifying entries in tables. This makes the process easy to manage by users that are not experts in computer science. DMN tables can be also deployed to decision engines, such as the Camunda Workflow engine[4], and be integrated into the BPMN processes.

It is possible to express the PI-RADS v2.1 guidelines using DMN notation based on the CDEs definitions. Such a method would benefit from possible integration within the clinical pathway that is defined using BPMN processes. Constructing the decision tables requires the decomposition of guideline rules into the sets of identified CDEs to which the diagnostic algorithms refer. Using the DMN standard as a method of modelling the domain knowledge contained within RADS standards removes the ambiguities of rule formulation and can improve how changes in PI-RADS versions

[2] https://sparxsystems.com/
[3] https://camunda.com/download/modeler/
[4] https://camunda.com

are communicated. Moreover, formalised guidelines presented as decision tables could be used to automatically generate predictions based on the predefined features according to current domain knowledge—thus aiding radiologists throughout the diagnostic process.

This methodology is compatible with computer-assisted reporting and decision support (CAR/DS), a novel class of systems that integrate the guidelines as clinical decision structured reporting tools [123]–[125]. CAR/DS is an open framework, initially proposed by the ACR, for integrating clinical decision support tools with radiology reporting systems. It implements clinical guidelines in the form of modules deployed in the radiology workflow system, embedded in the reporting environment [125]. An article published in February 2022 investigates the impact of CAR/DS application in abdominal CT assessment; its results suggest increased compliance with follow-up imaging and improved adherence to guidelines [124].

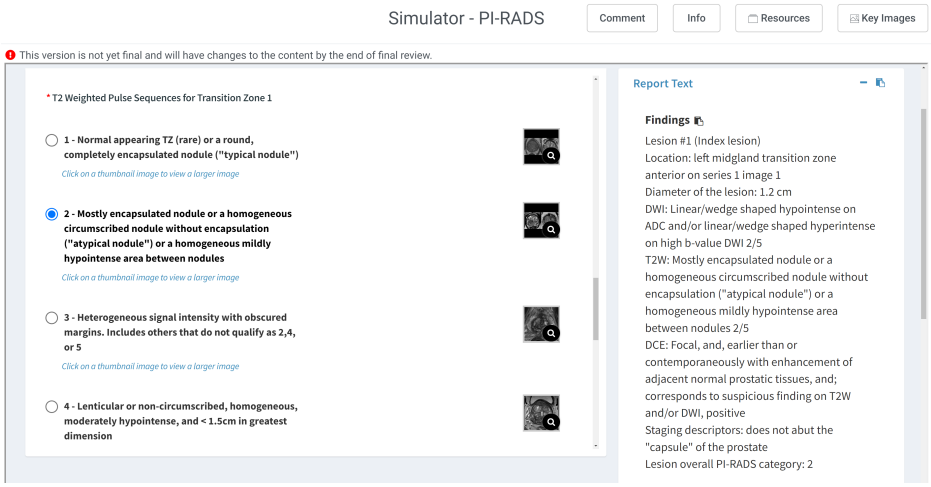


Figure 3.1.4: An T2W assessment section of the ACR Assist PI-RADS CAR/DS module[5]. Radiologist selects the rule matching the lesion picture on the T2W image. Overall PI-RADS category is automatically estimated based on the specified grades for modalities. The assessment with the CAR/DS tool concludes with a structured report text that can be copied to the local system.

ACR assist is an open web-based platform supporting CAR/DS models [125]. It integrates a tool to support reporting according to PI-RADS guidelines. Decision support within this module is limited to estimating the overall PI-RADS score based on manually selected appropriate T2W, DWI and DCE algorithm rules (assessment categories) matching the lesion features [Figure 3.1.4]. What is missing from this form

[5] The ACR Assist PI-RADS module is available at https://assist.acr.org/assistweb/PI-RADS

of assistance is the exploitation of the potential benefits of expanding the collected set of annotations resulting from the intermediate variables that make up the PI-RADS rules. The ACR assist PI-RADS module is currently under final verification.

The quality of the PI-RADS lexicon's formulation can be verified by conducting studies that involve the assessment of mpMRI features by experienced and inexperienced radiologists. Results from research reported in the literature on agreement among raters using PI-RADS demonstrate varied agreement among specialists. Most of the research analysed the assessment of PI-RADS categories, not the composite features that affect the diagnosis outcomes. The evaluation of agreement among experts who use the PI-RADS lexicon [55], [103] expands understanding of the reproducibility, uniformity, and quality of diagnoses [55]. The development of a computer-assisted structured reporting system integrated with formalised standard definitions allows research to be conducted on the qualities of mpMRI assessment based on continuously curated data during system use [108]. The indexing of medical imaging using the radiological lexicons resulting from application within structured reporting systems has already been proposed [108], [126]; however, the most common utilisation of the common terminology is report text standardisation. Knowledge resulting from the assessment of intermediate variables that constitute diagnostic evaluation in the form of CDEs provides insights on the quality of radiological evaluation that goes beyond domain-specific tasks.

In this chapter, we present our work on enhancing diagnostic procedures using a CAR/DS system of mpMRI PCa assessment integrated with formal descriptions based on the RadLex lexicon. We propose an innovative formalisation of the PI-RADS guidelines that utilises decision tables composed of rules that are based on CDE attributes. We demonstrate that it is possible to investigate the quality of radiological assessment using the datasets created during interaction with the tool in the course of clinical practice. To achieve this, we identified the sources of variability among experts, which continues to pose a key challenge in the development of PI-RADS. We present an analysis of interobserver agreement of CDE-standardised features that expand beyond the PI-RADS lexicon. Moreover, by performing repeated assessments of imaging by the same radiologists, we evaluated the intraobserver agreement on feature estimation, which allows investigation into the consistency of image interpretation.

3.2. Methods

Our goal was to develop an optimal proposition for CAR/DS in PCa diagnosis. The reporting procedure is based on the PI-RADS structured reporting scheme for mpMRI assessment. An article published in 2021 described a method of automatic PI-RADS assignment that employed formal methods generated on the basis of extracted radiomics features from DWI-ADC mpMRI from ninety-one patients [127]. This approach, however, does not guarantee adherence to the PI-RADS guidelines and is input-data dependent [88]. We propose a formalisation of the diagnostic guidelines in the form of defined rule sets expressed as decision tables modelled using DMN. This approach benefits from its ease of introducing iterative improvements and updates, and full transparency in regards to the inner logic of decision support. Rules can be easily investigated using the DMN visual interpreters by users without specialist technical expertise. Possible integration within workflow engines enables further utilisation of formalised guidelines as part of clinical processes.

3.2.1. Formalisation of PI-RADS diagnostic guidelines

An iterative analytic process was established in a series of consultations with an experienced radiologist (who had more than five years' experience using the PI-RADS standard in clinical practice) to identify the set of CDEs and develop the decision tables. Our work on reproducing the PI-RADS guidelines in the form of decision tables was divided into two stages:

1. A subset of RadLex and non-RadLex terms was defined that reflect the concepts in the PI-RADS guidelines to propose a set of CDEs
2. Based on the selected and defined CDEs, PI-RADS score decision tables were prepared that base their rule definition on CDE values and their outputs on lesion evaluation scores

The catalogue of the defined CDEs included elements that appear in the PI-RADS lexicon (relating to various categories, including abnormality, shape, margins, signal characteristics, etc.), as well as elements that exceed the standard and refer to clinically significant features or morphometric lesion features. The CDEs and decision rules were initially defined by the decomposition of the PI-RADS narrative guidelines into

Variable	Label	Related Radlex Terms	Possible Values
Input Variables			
lesion_dim_max	Lesion Max Dimension (mm)	Diameter [RID13432]	<5, >=5, >=15
lesion_location	Zone	Zone of prostate [RID38890]	PZ, TZ, NOT AVAILABLE
t2w_present_and_adequate	T2W present and adequate	Adequate [RID39308]	YES, NO
t2w_abnormality	T2W lesion present	Lesion [RID 38780]	YES, NO
t2w_invasive	T2W Invasive	Invasive [RID5680]	YES, NO
t2w_signal_intensity_type	T2W Signal Intensity Type	Signal characteristic [RID6049]	HYPOINTENSITIVITY, ISOINTENSITIVITY, HYPERINTENSITIVITY
t2w_signal_intensity	T2W Signal Intensity Scale	Signal characteristic [RID6049]	MILD, MODERATE, MARKEDLY
t2w_uniformity	T2W Lesion uniformity	Uniformity descriptor [RID43293]	HOMOGENEOUS, HETEROGENEOUS
t2w_focality	T2W Focality	Focal [RID5702]	YES, NO
t2w_shape	T2W Shape	Morphologic descriptor [RID5863]	LINEAR, WEDGE, LENTICULAR, WATER-DROP
t2w_shape_category	T2W Shape category	Morphologic descriptor [RID5863]	LINEAR, ROUND, IRREGULAR
t2w_margin	T2W Margin	Margin [RID5972]	PARTLY_ENCAPSULATED, ENCAPSULATED, WELL_DEFINED
t2w_margin_category	T2W Margin category	Margin [RID5972]	CIRCUMSCRIBED, NON_CIRCUMSCRIBED
Output Variables			
t2w_pirads	T2W PI-RADS Evaluation	PZ [RID50301], TZ [RID50307]	1,2,3,4,5, X
t2w_pirads_description	T2W PI-RADS Rule Descripton		<string>

Table 3.2.1: Input and output variables of decision table representing the PI-RADS T2W assessment algorithm

logical statements using a subset of identified RadLex terms. The resulting decision tables were used as a base for review by an experienced radiologist and introduced improvements through interviews. Overall, the decision tables were concluded to be final and complete after three iterations. Then, to validate the decision tables, an independent experienced radiologist who was not engaged in the development of the decision tables, verified the rules and declined to recommend the introduction of any changes; the tables were then assumed to be valid and complete.

Four decision tables were prepared, which reflected the PI-RADS algorithm. Separate decision tables calculate the partial scores for T2W [Table 3.2.3], DCE, and DWI-ADC modalities, as well as the final PI-RADS score based on the partial evaluations and lesion location[6]. Each rule was defined using a subset of CDEs, an output variable (PI-RADS score), and a description that acts as an explanation for a given decision. Decision tables were prepared using the DMN standard, which allows ease in communication and the introduction of minor updates.

Although definition of the decision table input variables and their possible values was based on the identified related RadLex terms [Table 3.2.1], it was also necessary to introduce additional variables that were not included in the lexicon. This was dictated by the distributed character of RadLex terms: they are frequently duplicated or inconsistent due to the inclusion of domain-specific lexicons. A prime example of this can be found when signal characteristic property is considered. The terms 'Hypointense' [RID35804], 'T2 hypointensity' [RID49501] and 'Markedly hypointense' [RID49500] are all contained in the RadLex lexicon; however, the corresponding terms,

[6] DCE, DWI-ADC and Overall PI-RADS decision tables are included in the *Appendix*

PI-RADS	Rule Description
X	[#1] PI-RADS evaluation not avaible for the selected zone
1	[#2] No lesions
5	[#3] Zone: PZ or TZ; Invasive; Max dim.>=5mm
5	[#4] Zone: PZ, Circumscribed, Focal, Moderate/Markedly, Hypointense, Homogenous, Non-invasive; Max dim.>=15mm
4	[#5] Zone: PZ, Circumscribed, Focal, Moderate/Markedly, Hypointense, Homogenous, Non-invasive; Max dim.[5, 15)mm
3	[#6] Zone: PZ, Heterogenous, Non-invasive; Max dim.>=5mm
3	[#7] Zone: PZ, Round, NonCircumscribed, Moderate/Markedly, Hypointense, Non-invasive; Max dim.>=5mm
2	[#8] Zone: PZ, Linear/Wedge shaped, NonCircumscribed, Hypointense, Non-invasive; Max dim.>=5mm
2	[#9] Zone: PZ, NonCircumscribed, Non-focal, Mild, Hypointense, Non-invasive; Max dim.>=5mm
5	[#10] Zone: TZ, Lenticula/Water-drop shaped, Moderate/Markedly, Hypointense, Homogenous, Non-invasive, Max dim.>=15mm
5	[#11] Zone: TZ: NonCircumscribed, Moderate/Markedly, Hypointense, Homogenous, Non-invasive, Max dim.>=15mm
4	[#12] Zone: TZ, Lenticula/Water-drop shaped, NonCircumscribed, Moderate/Markedly, Hypointense, Homogenous, Non-invasive, Max dim. [5, 15) mm
4	[#13] Zone: TZ: NonCircumscribed, Moderate/Markedly, Hypointense, Homogenous, Non-invasive, Max dim.[5, 15)mm
3	[#14] Zone: TZ, NonCircumscribed, Heterogenous, Non-invasive; Max dim.>=5mm
2	[#15] Zone: TZ, Partly Ecapsulated, Focal, Round, Non-invasive, Max dim.>=5mm
2	[#16] Zone: TZ, Well-defined, Focal , Round, Homogenous, Non-invasive, Max dim.>=5mm (Atypical nodule)
2	[#17] Zone: TZ, Non-focal, Mild, Hypointense, Non-invasive; Max dim.>=5mm
1	[#18] Zone: TZ, Encapsulated, Focal, Round, Non-invasive, Max dim.>=5mm (Typical nodule)
3	[#19] Others that do not qualify as PI-RADS 2, 4, or 5.
X	[#20] T2W image is unavailable
X	[#21] Unclassified case

Table 3.2.2: Decomposed rules of the PI-RADS T2W assessment algorithm

'DWI Hypointensity' or 'Moderately hypointense' are absent, as they do not occur in domain lexicons. For that reason, we opted to define a specific set of CDEs and to relate those to the RadLex terms.

The high number of rules derived from the PI-RADS guidelines is a result of the translation of descriptive characteristics into logical statements. This reduced ambiguities in the diagnostic standard rules' definitions. For example, the PI-RADS v2.1 T2W evaluation rule for TZ lesion states that PI-RADS category 2 is assigned for lesions that display the following characteristic: 'A mostly encapsulated nodule OR a homogeneous circumscribed nodule without encapsulation ("atypical nodule") OR a homogeneous mildly hypointense area between nodules'. This statement was translated into three rules in the decision tables [Tables 3.2.2 and 3.2.3]:

[#15] Lesion: TZ, Partly Encapsulated, Focal, Round, Noninvasive, Max dim.>=5mm

[#16] Lesion: TZ, Well-defined, Focal, Round, Homogenous, Noninvasive, Max dim.>=5mm (Atypical nodule)

[#17] Lesion: TZ, Non-focal, Mild, Hypointense, Noninvasive; Max dim.>=5mm

Based on radiologists' insights, two additional CDEs were incorporated to simplify the rule sets. According to the experts, assessment of particular shape and margin type features on mpMRI images is highly subjective and could potentially decrease diagnostic accuracy if implemented as part of a formalised model. Instead, categorisation was suggested: mapping the particular values into more general feature types to simplify the defined rules. As a result, eight shapes described as part of the PI-RADS lexicon (round, oval, lenticular, lobulated, tear-shaped, wedge-shaped, linear, and ir-

regular) were simplified to three shape types (linear, round, and irregular) and seven types of lesion margin (circumscribed, noncircumscribed, indistinct, obscured, spiculated, encapsulated, and erased charcoal sign) were simplified to two significant types (circumscribed and noncircumscribed).

Rule	Lesion Max Dim	Lesion location	Adequate	Abormality	Invasive	Signal Intensity Type	Signal Intensity Scale	Uniformity	Focality	Shape	Shape Cat.	Margin	Margin Cat.	PI-RADS
1		NOT_AVAILABLE	YES	YES										X
2		PZ,TZ	YES	NO										1
3	>=5	PZ,TZ	YES	YES	YES									5
4	>=15	PZ	YES	YES	NO	HYPOINTENSITIVITY	MODERATE,MARKEDLY	HOMOGENEOUS	YES				CIRCUMSCRIBED	5
5	[5..15)	PZ	YES	YES	NO	HYPOINTENSITIVITY	MODERATE,MARKEDLY	HOMOGENEOUS	YES				CIRCUMSCRIBED	4
6	>=5	PZ	YES	YES	NO			HETEROGENEOUS						3
7	>=5	PZ	YES	YES	NO	HYPOINTENSITIVITY	MODERATE,MARKEDLY		YES		ROUND		NON_CIRCUMSCRIBED	3
8	>=5	PZ	YES	YES	NO	HYPOINTENSITIVITY				LINEAR,WEDGE			NON_CIRCUMSCRIBED	2
9	>=5	PZ	YES	YES	NO	HYPOINTENSITIVITY	MILD		NO				NON_CIRCUMSCRIBED	2
10	>=15	TZ	YES	YES	NO	HYPOINTENSITIVITY	MODERATE,MARKEDLY	HOMOGENEOUS		LENTICULAR,WATER-DROP				5
11	>=15	TZ	YES	YES	NO	HYPOINTENSITIVITY	MODERATE,MARKEDLY	HOMOGENEOUS					NON_CIRCUMSCRIBED	5
12	[5..15)	TZ	YES	YES	NO	HYPOINTENSITIVITY	MODERATE,MARKEDLY	HOMOGENEOUS		LENTICULAR,WATER-DROP				4
13	[5..15)	TZ	YES	YES	NO	HYPOINTENSITIVITY	MODERATE,MARKEDLY	HOMOGENEOUS					NON_CIRCUMSCRIBED	4
14	>=5	TZ	YES	YES	NO			HETEROGENEOUS					NON_CIRCUMSCRIBED	3
15	>=5	TZ	YES	YES	NO				YES		ROUND	PARTLY_ENCAPSULATED		2
16	>=5	TZ	YES	YES	NO			HOMOGENEOUS	YES		ROUND	WELL_DEFINED		2
17	>=5	TZ	YES	YES	NO	HYPOINTENSITIVITY	MILD	HOMOGENEOUS	NO					2
18	>=5	TZ	YES	YES	NO				YES		ROUND	ENCAPSULATED		1
19	>=5	PZ,TZ	YES	YES										3
20			NO											X
21														X

Table 3.2.3: Decision table representing the PI-RADS T2W assessment algorithm

3.2.2. PI-RADS CAR/DS form

Based on the agreed CDEs, an interactive electronic form was developed and made available for collecting and annotating data on the dedicated platform. A special mpMRI evaluation form was constructed as a system module, built based on selected groups of lesion-relevant attributes for which radiologists mark specific values, such as 'shape: round' and 'margin: circumscribed', according to the properties of defined CDEs. An imposed order of assessment that uses a computer-assisted reporting tool assures adherence to diagnostic guidelines and guarantees complete evaluation, thus allowing verification of intermediate steps of diagnosis.

Defined decision tables were integrated with the structured reporting form. Using the marked lesion features, the system automatically suggests the value of PI-RADS scores using the implemented PI-RADS 2.1 decision rules. Such suggestions play the role of a 'second opinion' and allow radiologists to make final decisions in adherence with the diagnostic guidelines. This constitutes a form of diagnosis support, in which users receive feedback from the platform based on the domain knowledge of the diagnostic standard on which the system's logic is based. A diagnosis protocol and decision-making suggestions can be inspected and verified to ensure efficacy. Given the specified input variables, radiology specialists are informed of the assessment category and the specific rule that applies to the inputted case. The resulting automated assessment can then be manually corrected if a specialist disagrees with the suggestion.

Two experienced radiology consultants were engaged in the process of form design and development. First, all defined descriptions of decision table rules, CDEs, and related RadLex-lexicon terms were translated to Polish language, as their definitions and values had to match the language of the narrative report text the system generates. Based on the knowledge gathered, several form prototypes were iteratively developed and reviewed by diagnosticians, who identified flaws introduced during the form's development and recommended changes to improve the quality of user experience (UX). An additional two CDEs related to the categories of shape and margin type CDEs were included on the form, and their value was automatically set based on the estimation of base CDE.

Diagnosticians can benefit from integration with PI-RADS v2.1 decision tables, which automatically suggest PI-RADS scores for the modalities. The system auto-

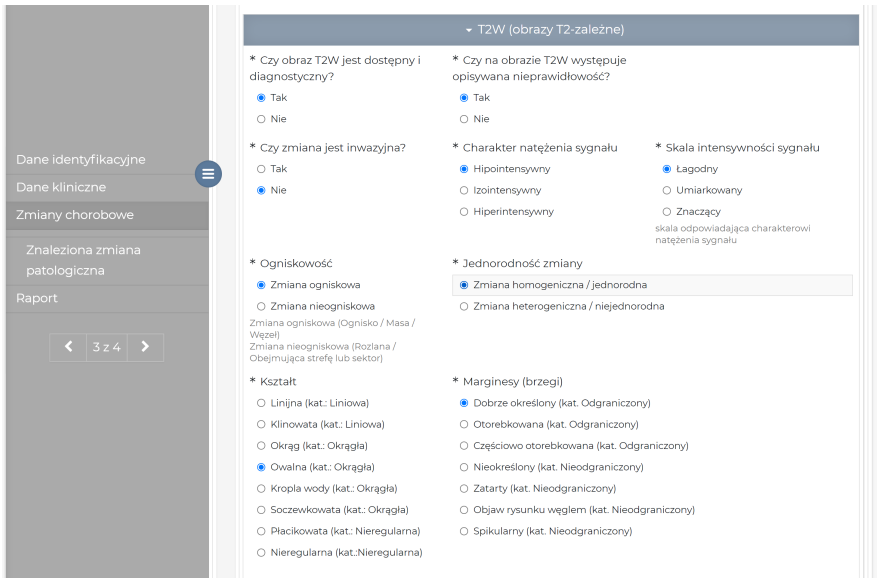


Figure 3.2.1: Part of PI-RADS CAR/DS form for assessment of lesion features in T2W images (corresponding to the T2W decision table input variables [Table 3.2.1]). The left sidebar allows for navigation between sections. The selected section, *Zmiany chorobowe* (lesions) allows for evaluation of detected lesions in mpMRI. Each lesion is evaluated separately.

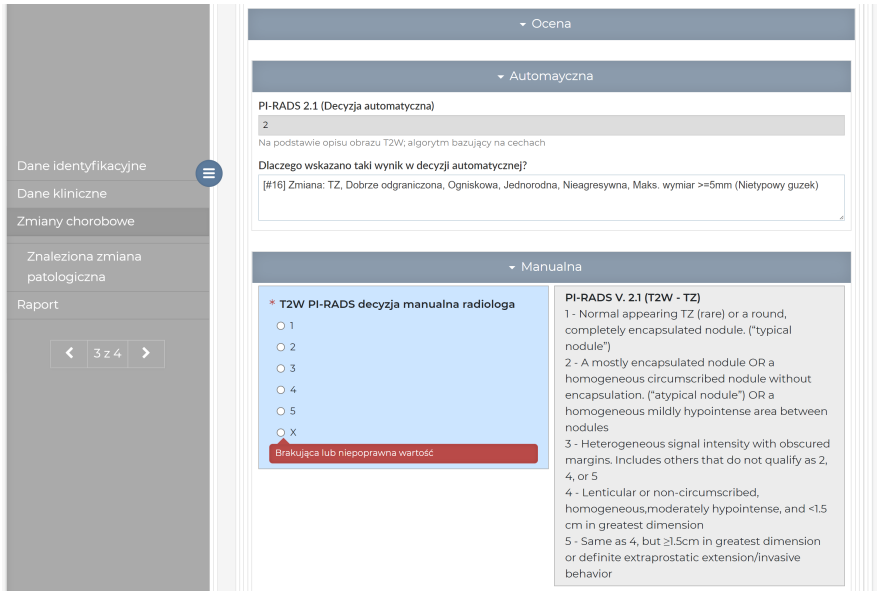


Figure 3.2.2: Radiologists are presented with semi-automatically determined PI-RADS v2.1 scores (here, using the T2W decision table [Table 3.2.3]) and explanations of which rules have been detected [Table 3.2.2].

matically generates a text-based, synoptic final report that follows the report formula proposed in the PI-RADS guidelines. The structured reporting form is divided into the following sections: patient identification and clinical data, identified pathological lesions, and narrative report. During the main assessment phase, radiologists can

access multiple lesions within the prostate gland. Each lesion evaluation involves specifying lesion sectoral location using the prostate sector map, lesion dimensions, assessment of lesion characteristics on T2W, DWI, ADC and DCE images, and finalising the assessment by deciding on the overall PI-RADS score. The completed structured reporting form is then used to generate the report narrative text [Figure 3.2.3].

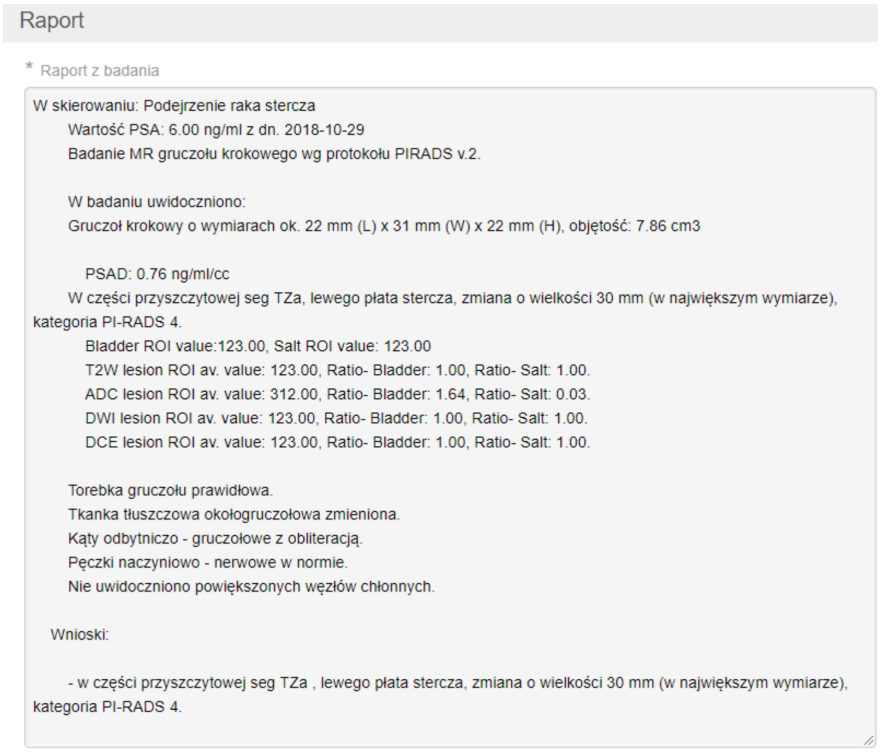


Figure 3.2.3: In the report section, an automatically generated text is presented to the user in a structured form. The report includes clinical information and a summary of the findings, in addition to prostate volume estimated using the dimensions provided, calculated PSA density, and parameters that were included to conduct research on signal intensity normalisation.

Using the PI-RADS CAR/DS form, clinicians can perform complete mpMRI assessments according to the current PCa diagnostic standards. Instead of writing reports, clinicians interact with the system through a web browser. The form comprises mostly radio buttons that are used to define CDEs' values that correspond to the visually assessed lesion properties. Based on the completed form, the generated text of the report can be manually copied to the local HIS/RIS; this means that the tool is ready for production use.

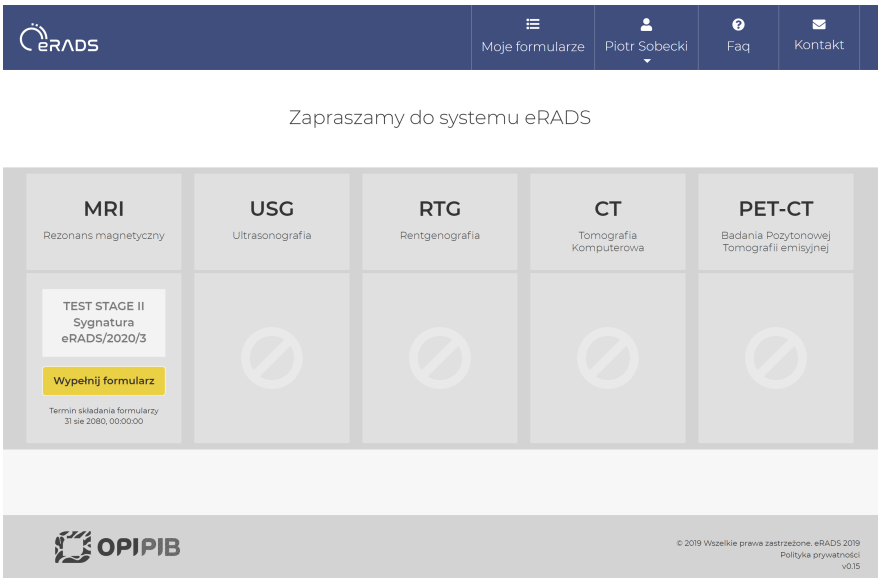


Figure 3.2.4: eRADS system homepage

3.2.3. CAR/DS research platform: eRADS

During our work on an optimal CAR/DS solution, we developed a research platform named eRADS [Figure 3.2.4], which was designed to enable the standardisation of radiological reporting and integration with decision-support solutions. It was used as a research tool for the elaboration, implementation, and evaluation of the structured reporting models. The design, analysis and development of structured reporting is tailored to specific clinical problems and includes algorithms and schemes for descriptions of pathological lesions. The platform's goal is to develop methods and practices for creating good and reliable structured reports in radiology, with particular emphasis on the reporting of cancer lesions backed with integrated methods of decision support.

The eRADS platform possesses modular architecture [Figure 3.2.5], which allows rapid implementation of newly-created structured reporting models. For image data management, eRADS integrates with the XNAT imaging informatics software platform with archiving and management functionalities. Integration with the OHIF web-based DICOM image viewer is provided for medical image viewing and analysis. eRADS integrates with the Camunda business process model engine and supports a decision table engine for models defined using DMN.

The eRADS platform and integrated systems are accessible via web browser and have been made available for production use. Radiologists can access the program using a local station and use the functionalities that enable structured report generation

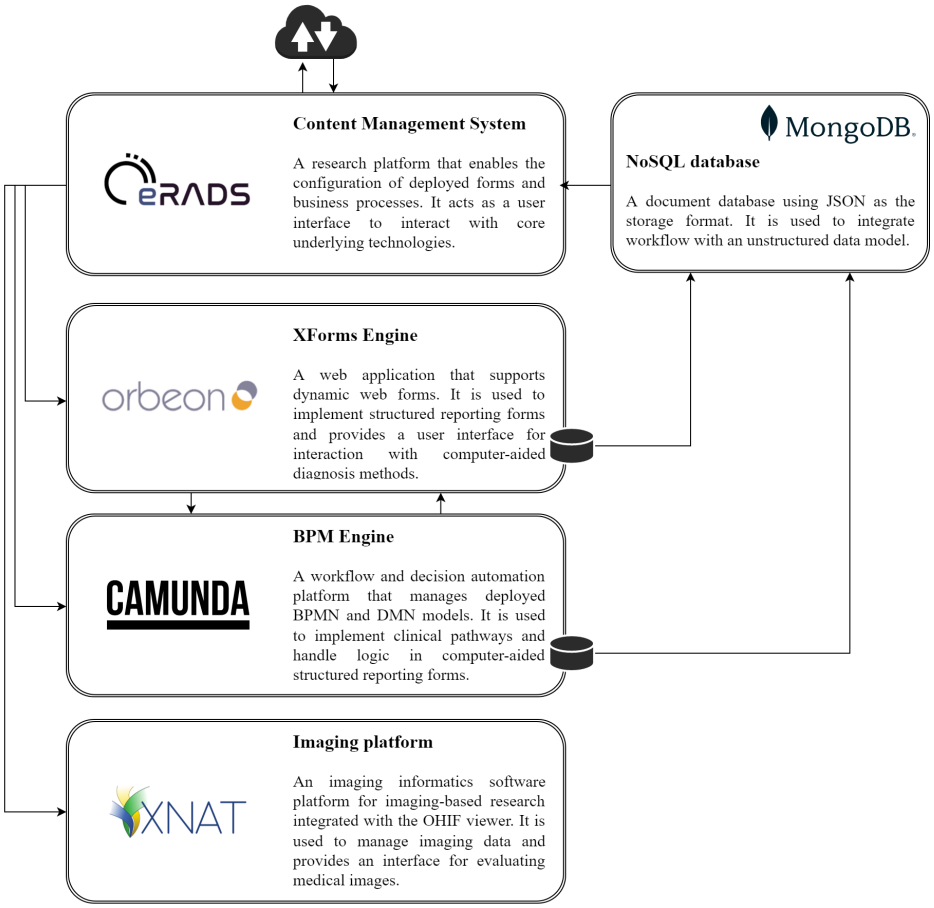


Figure 3.2.5: eRADS System architecture

based on the manually specified features. Structured reporting forms are developed using XForms technology allowing dynamic interactive forms to be constructed using XML definitions [128], [129]. Forms are deployed on the Orbeon platform, which is used as an XForms engine integrated into the eRADS system. Orbeon provides a form-builder user interface that allows developers to visually construct forms using drag-and-drop controls. The engine also supports calling web services, which enables integration with the models of decision support. This feature is used to integrate the benefits of decision tables deployed on the Camunda platform into the form logic. The possible extension and direct integration with hospital information system are possible via web-service calls from using XForms within the form definition; that, however, is dependent on software API and would require secure network access to local hospital infrastructure from the platform or localised deployment in a target diagnostic centre.

Using integrated Orbeon and Camunda applications on a custom content manage-

[6] https://www.orbeon.com/

ment system platform decreases the cost of introducing changes to the functionalities. Business logic can be extended or modified without engaging developers in the process. Forms (XForms), processes (BPMN), and decision tables (DMN) are defined with the support of editors using a graphic user interface. This enables modularity and prototyping. New forms and decision tables can be deployed onto the platform independently with limited risk of regression errors.

3.2.4. Experiments

To investigate the effects of our method of computer-aided diagnosis, we conducted experiments that involved radiology specialists interacting with the CAR/DS form deployed on the eRADS platform. Two studies were conducted. The retrospective study involved two evaluations of preselected subsets of lesions sourced from the ProstateX training dataset. The study was conducted by six radiologists accessing the platform remotely using home workstations, outside of the clinical environment. The imaging was accessible through the integrated, web-based OHIF viewer. The study aimed to investigate inter- and intrarater agreement among experienced and inexperienced specialists, as well as verifying the proposed interpretation of the PI-RADS guidelines as decision tables. The prospective study, which involved two radiology specialists occurred in a clinical environment during usual work practice. The imaging was assessed on a hospital's certified diagnostic workstations. The aim of studying the tool in a clinical setting was to verify the applicability of the assessment method beyond the research environment. Contrary to the retrospective study, no biopsy results were available for the lesions analysed, as the mpMRI assessment practised at the facility was performed without further feedback on patients outcomes.

Retrospective study

This study was performed on a specially prepared mpMRI dataset, and the results of targeted magnetic resonance biopsies. The data was balanced, and included sixteen clinically significant and sixteen clinically nonsignificant lesions. The nonsignificant lesions included abnormalities scored as PI-RADS <= 2 for which biopsy had not been performed and those for which histopathology assessment had assigned the first Gleason category. The experiment was performed on the preselected subset of the ProstateX challenge training dataset. The selected lesions were located in the AS

(seven, of which four were clinically significant), TZ (eleven, of which five were clinically significant), and PZ (fourteen, of which seven were clinically significant).

The experiment was conducted with the participation of experienced and inexperienced radiologists. These experts were not involved in the development process of the methodology. The study was conducted on a group of radiology specialists who used the RADS standards during the diagnostic practice:

- Three specialists with diagnostic experience of one to five years
- Three specialists with more than ten years' diagnostic experience and at least five years' experience using the PI-RADS standard (since the first version of the standard)

The study involved two sessions that required the full assessment of thirty-two selected lesions using the structured reporting forms deployed on the eRADS system. In the first phase, the radiologists assessed all cases by specifying the imaging features (the values of the identified CDEs) and assigned the manual PI-RADS categories for lesions. The second phase was conducted two weeks after the first to eliminate the memory effect. All cases were assessed using the computer-assisted structured report tool integrated with the decision tables. The form included text fields that acted as comments that could be filled by radiologists to express their reasons for disagreement with suggestions resulting from the formalised model's assessments. The time spent on interaction with the computer-assisted reporting form during each mpMRI examination assessment was measured automatically.

The interactive form enhanced the procedure of mpMRI assessment, supporting the data collection process and suggesting the PI-RADS scores based on the answers provided on the structured reporting form. The results of both study stages were compared to establish inter- and intrarater concordance in assessing the imaging features and PI-RADS categories, and in determining the resulting quality of diagnosis.

UX and ergonomics tests of the CAR/DS tool were conducted between the two study sessions. The goal was to verify the implementation of the support method—in particular, the verification of the decision tables and the mapping of the evaluation rules according to the PI-RADS guidelines. UX tests were conducted in a specially prepared environment, which was adapted for usability testing. This allowed us to consider the specific needs of radiologists and the clinical conditions of mpMRI exam-

ination evaluation. The interviews with the radiologists were conducted and recorded to gather their insights on the quality of interaction with the solution and to estimate its potential in clinical settings.

Prospective study

To validate the solutions in a clinical setting, we conducted a prospective study on the premises of a radiology diagnostic centre during a weekend shift of the specialists. The study was conducted on two radiology specialists of the same radiology department, who used the PI-RADS standard in daily diagnostic practice. The study included a specialist with one to five years' diagnostic experience and more than one year's experience using the PI-RADS standard, and a specialist with ten years' diagnostic experience and more than five years' experience using the PI-RADS standard. The structured report generated by the tool was copied manually into the hospital RIS system. Patient data stored in the eRADS database was anonymised by referring only to the generated imaging study identifiers to identify the cases.

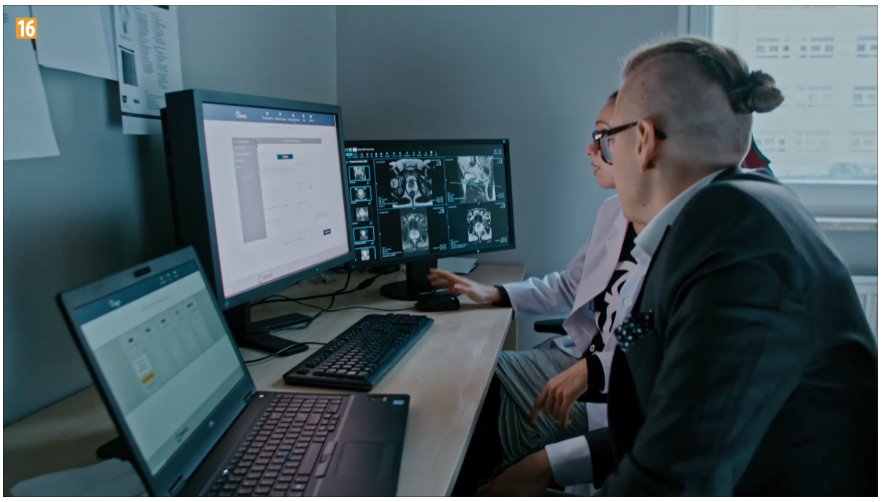


Figure 3.2.6: Setup of the diagnostics workstation during validation of the method in a clinical setting at the Department of Radiology, Centre of Postgraduate Medical Education in Warsaw, Poland. The radiologists were interviewed after the conducted study. Photo obtained from: [130]

The specialists were instructed not to contact each other during the study to discuss the cases they had assessed. The prostate imaging examinations of eighteen patients were assessed by both radiologists. The data was not specifically preselected and reflected the true nature of diagnostic work during clinical practice [Figure 3.2.6]. The data was acquired using three-tesla mpMRI; the data was complete and all images

analysed were diagnostic. Due to the type of diagnostic work performed at the institution and the prospective nature of the study, the data did not include the biopsy results of reported findings. For that reason, it was not possible to perform analysis of the diagnostic accuracy.

3.2.5. Statistical analysis

To compare the results of interrater agreement of the CDEs to similar research on interobserver agreement of the PI-RADS v2 lexicon published by Mussi et. al. [131] in 2020, we followed similar statistical analysis protocols. We present the percent concordance (PA) and first-order agreement coefficient (AC1) [132] obtained using Gwet's method for the CDEs. Additionally, the intrarater agreement was estimated using the same measures by comparing the assessments between the two study sessions performed on the retrospective data. Statistical significance levels were set at 5% and interpretation of agreement levels was defined as excellent for AC1 values (≥ 0.81), good (0.61–0.80), moderate (0.41–0.60), fair (0.21–0.40), and poor (≤ 20). We used a Wilcoxon signed-rank test to compare the means of assessed features. AUC, Recall and Precision were used as measures of the diagnostic methods' performance.

Data cleaning, restructuration, and visualisation were performed in Python (v3.7.12) using the Pandas (v1.3.5) and Plotly (v5.5.0) packages. Statistical analysis was performed using R (v4.1.2) and irrCAC (v1.0)[133] package. All scripts were written in the Google Collaboratory tool using the dedicated notebooks.

3.3. Results

In this section, we present the results of the retrospective study conducted on thirty-two prostate lesions drawn from the ProstateX training dataset. Those lesions were preselected for evaluation by six radiologists. The results are compared to that of the prospective study, which was conducted in a clinical environment and involved two specialists evaluating eighteen mpMRI studies that were not pre-selected, but resulted from the clinicians' regular work during their shift.

3.3.1. Quality and variability of PI-RADS v2.1 assessment

Interrater agreement of defined CDEs

		Session 1		Session 2		Overall	
		PA	AC1 (95% CI)	PA	AC1 (95% CI)	PA	AC1 (95% CI)
Modality	Feature						
	Lesion >= 1.5cm	72.3	0.45 (0.26; 0.63)	69.2	0.40 (0.21; 0.58)	70.7	0.42 (0.29; 0.55)
	Zone (selected)	76.9	0.66 (0.51; 0.81)	79.4	0.70 (0.56; 0.84)	78.1	0.68 (0.58; 0.78)
T2W	Abnormality	94.0	0.94 (0.87; 1.00)	99.0	0.99 (0.97; 1.00)	96.5	0.96 (0.93; 0.99)
	Focality	66.6	0.48 (0.30; 0.65)	74.4	0.65 (0.49; 0.80)	70.5	0.57 (0.45; 0.68)
	Homogeneity	63.9	0.38 (0.19; 0.56)	62.7	0.37 (0.19; 0.54)	63.3	0.37 (0.25; 0.50)
	Invasiveness	68.1	0.45 (0.23; 0.66)	72.7	0.57 (0.38; 0.77)	70.4	0.51 (0.37; 0.65)
	Margin	26.5	0.13 (0.06; 0.20)	28.2	0.18 (0.11; 0.24)	27.3	0.16 (0.12; 0.21)
	Margin cat.	73.0	0.56 (0.36; 0.76)	69.2	0.50 (0.29; 0.71)	71.1	0.53 (0.39; 0.67)
	Shape	27.4	0.18 (0.13; 0.23)	23.0	0.13 (0.07; 0.18)	25.2	0.15 (0.12; 0.19)
	Shape cat.	46.2	0.22 (0.10; 0.35)	50.6	0.31 (0.16; 0.45)	48.4	0.26 (0.17; 0.35)
	Signal int.	45.2	0.24 (0.15; 0.33)	54.4	0.38 (0.24; 0.52)	49.8	0.37 (0.30; 0.44)
	Signal int. type	95.3	0.95 (0.90; 1.00)	100.0		97.7	0.98 (0.95; 1.00)
DWI	Abnormality	85.0	0.81 (0.69; 0.93)	94.0	0.93 (0.87; 1.00)	89.5	0.87 (0.81; 0.94)
	Focality	73.0	0.61 (0.41; 0.80)	77.2	0.68 (0.50; 0.86)	75.1	0.64 (0.52; 0.77)
	Homogeneity	70.0	0.54 (0.33; 0.75)	77.0	0.69 (0.53; 0.85)	73.6	0.62 (0.49; 0.75)
	Invasiveness	65.1	0.42 (0.22; 0.62)	70.0	0.52 (0.32; 0.72)	67.6	0.47 (0.33; 0.61)
	Shape	25.2	0.16 (0.08; 0.23)	20.3	0.10 (0.04; 0.15)	22.7	0.12 (0.08; 0.17)
	Shape cat.	46.8	0.24 (0.10; 0.38)	51.2	0.31 (0.17; 0.46)	49.0	0.28 (0.18; 0.37)
	Signal int.	50.2	0.26 (0.13; 0.40)	55.5	0.34 (0.18; 0.50)	52.9	0.38 (0.29; 0.47)
	Signal int. type	89.0	0.88 (0.75; 1.00)	94.6	0.94 (0.87; 1.00)	91.9	0.91 (0.84; 0.98)
ADC	Abnormality	92.3	0.91 (0.84; 0.99)	94.4	0.94 (0.88; 1.00)	93.3	0.93 (0.88; 0.98)
	Focality	74.2	0.64 (0.46; 0.82)	76.4	0.67 (0.52; 0.83)	75.3	0.66 (0.54; 0.78)
	Homogeneity	65.3	0.49 (0.33; 0.65)	73.6	0.63 (0.46; 0.80)	69.5	0.56 (0.44; 0.68)
	Invasiveness	67.1	0.45 (0.25; 0.65)	71.2	0.54 (0.34; 0.75)	69.2	0.50 (0.36; 0.64)
	Shape	23.5	0.13 (0.08; 0.19)	21.1	0.11 (0.05; 0.16)	22.3	0.12 (0.08; 0.16)
	Shape cat.	46.5	0.23 (0.09; 0.37)	50.5	0.30 (0.17; 0.44)	48.5	0.27 (0.17; 0.36)
	Signal int.	47.4	0.24 (0.08; 0.41)	59.9	0.44 (0.27; 0.61)	53.6	0.39 (0.29; 0.48)
	Signal int. type	94.8	0.95 (0.87; 1.00)	100.0		97.4	0.97 (0.94; 1.00)
DCE	Abnormality	77.9	0.64 (0.47; 0.82)	75.0	0.65 (0.52; 0.78)	76.5	0.65 (0.54; 0.75)
	BPH Features	82.7	0.77 (0.57; 0.95)	85.7	0.81 (0.67; 0.95)	84.3	0.79 (0.68; 0.91)
	Enhancement	79.3	0.75 (0.57; 0.93)	75.7	0.60 (0.41; 0.79)	77.3	0.68 (0.55; 0.80)
	Focality	43.7	0.32 (0.17; 0.47)	45.6	0.35 (0.19; 0.51)	44.7	0.34 (0.23; 0.44)

Table 3.3.1: Inter-observer agreement of PI-RADS v2.1 CDEs (p<.01)

Based on the results obtained from the two stages of the retrospective study, the mean interrater percentage agreement and AC1 values with 95% confidence intervals are presented in Table 3.3.1 for estimated values of PI-RADS v2.1 CDEs. Overall, the table presents the mean of fifteen pairs of radiologists' evaluations that were compared to estimate their concordance.

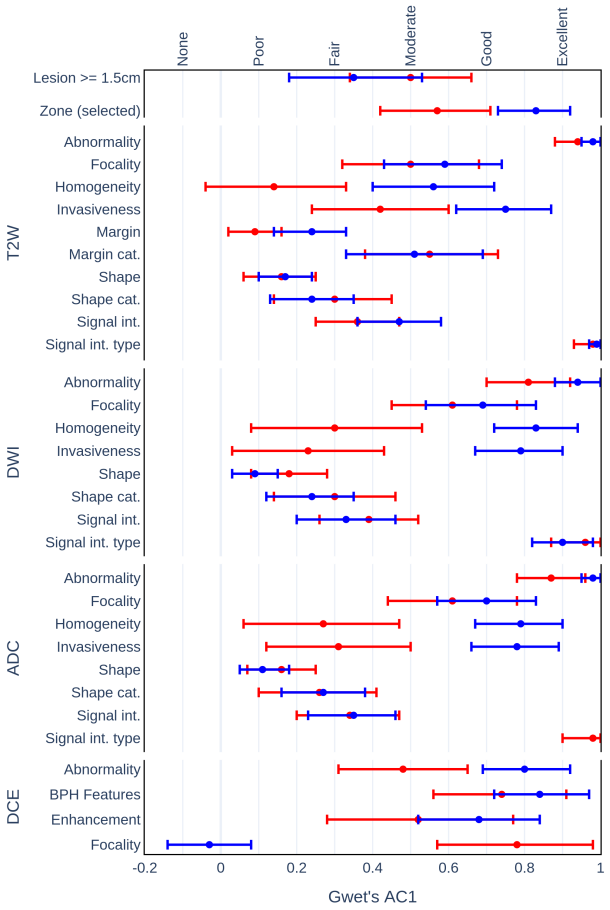


Figure 3.3.1: Mean interrater agreement among experienced (red) and inexperienced (blue) raters and 95% confidence intervals.

The highest agreement between the radiologists was observed for abnormality detection, assessment of signal intensity type, and BPH feature CDEs. The results were expected for the first two features, as radiologists were presented with reference images of abnormalities on all modalities as a guide. Signal intensity type connects strongly with the occurrence of lesions on those modalities; hypointensity indicated the abnormalities for the T2W and ADC images and hyperintensity for the DWI images. Given a study design that assesses the preselected potentially clinically significant lesions, those features demonstrated high agreement between raters; this, however, was not observed for DCE images, for which abnormalities were not observed in all

of the cases analysed. This resulted in decreased agreement of abnormality detection (PA = 76.5%) and enhancement indication (PA = 77.3%), which suggests that not all abnormalities are evident on all mpMRI sequences and that the evaluation of signal enhancement on DCE is subjective. The lowest agreement was observed for highly subjective features: shape, signal intensity level, and type of lesion margins. The simplification of lesion shape and margin features by grouping the values into types improved the concordance between raters.

Analysis of differences in interrater agreement among the experienced and inexperienced raters (within groups) reveals several significant differences in agreement values [Figure 3.3.1]. The results present mean comparisons of five pairs of assessments for inexperienced and inexperienced groups, each of which was represented by three experts. The largest difference between the groups was observed in their assessment of focality on DCE images: agreement among inexperienced raters was not statistically significant (AC1 = -0.03, p = .37) and was good for experienced raters (AC1 = 0.78, p < .001). However, this was the only case for which the experienced raters agreed more on the feature assessment. The opposite tendency was observed for:

- Zonal locations of lesions
  - Experienced AC1 = 0.57, p < .001 vs. Inexperienced AC1 = 0.83, p < .001
- Homogeneity on:
  - T2W (AC1 = 0.14, p = .13 vs. AC1 = 0.56, p < .001),
  - DWI (AC1 = 0.30, p < .01 vs. AC1 = 0.83, p < .001),
  - ADC (AC1 = 0.27, p < .05 vs. AC1 = 0.79, p < .001);
- Invasiveness on:
  - T2W (AC1 = 0.42, p < .001 vs. AC1 = 0.75, p < .001),
  - DWI (AC1 = 0.30, p < .01 vs. AC1 = 0.83, p < .001),
  - ADC (AC1 = 0.27, p < .05 vs. AC1 = 0.79, p < .001);
- Abnormality detection on:
  - DCE (AC1 = 0.48, p < .001 vs. AC1 = 0.80, p < .001).

Figure 3.3.2 presents the concordance analysis of each CDE evaluated on lesions located in the PZ, TZ, and AS in comparison to the overall results. The results reveal that evaluation of AS lesion features demonstrated lower agreement among raters in

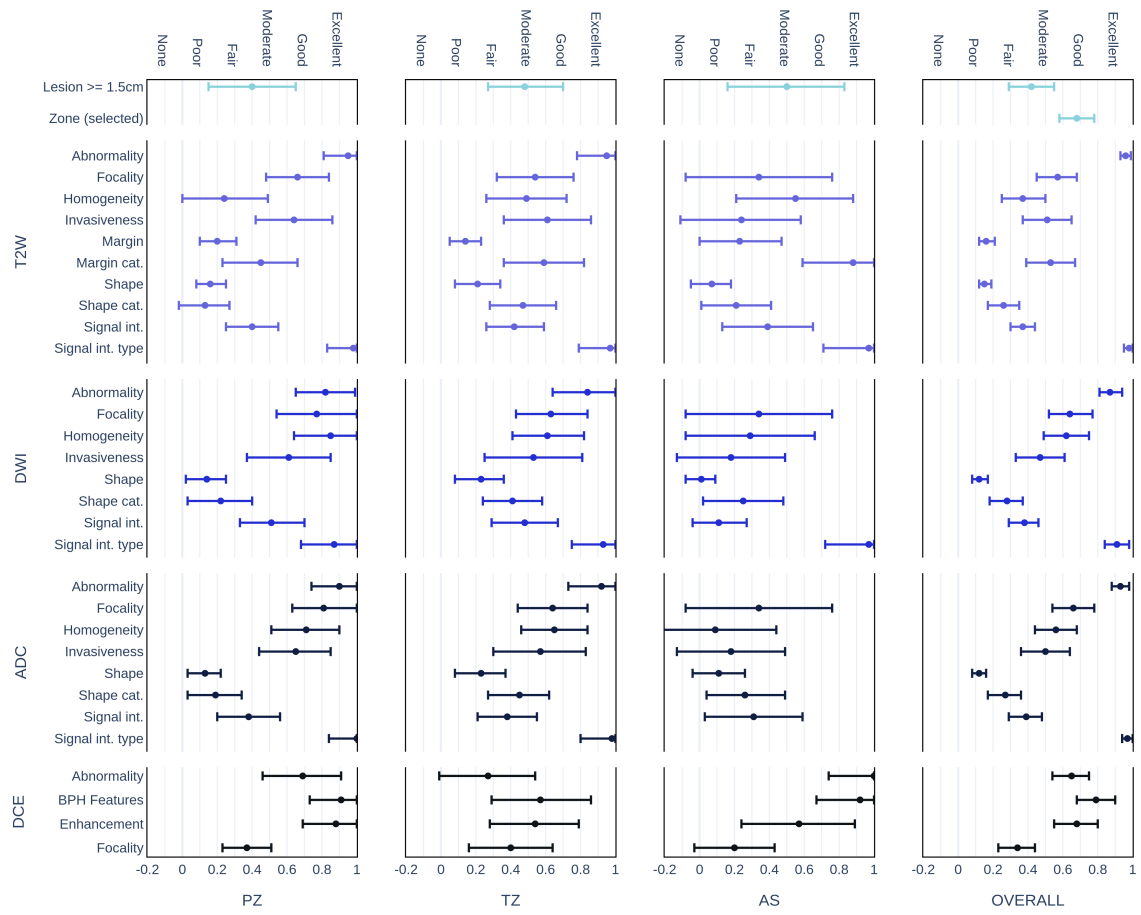


Figure 3.3.2: Mean interrater agreement (AC1) of composite PI-RADS CDEs in the PZ, TZ, and AS zones and overall results. The colours correspond to the modality sources of the features.

comparison to PZ and TZ lesions. The overall wider range of confidence intervals can be explained partially by the smaller number of lesions (seven) evaluated in that zone. Overall, it was observed that the shape and signal intensity features demonstrated the lowest agreement between raters. Analysis of interrater agreement dependent on lesion locations indicate that overall no statistically significant differences were demonstrated in agreement between PZ (mean PA = 69.7%), TZ (mean PA = 67.3%) and AS (mean PA = 65.5%) features. The agreement between radiologists displayed high variability. The highest deviations in agreement between the experts were observed for features of lesions located in the AS. This was particularly visible for highly subjective features based on the evaluation of signal intensity, focality and texture features (homogeneity).

Intrarater agreement of defined CDEs

Analysis of intrarater agreement indicated that most feature evaluations displayed moderate or good agreement between the study stages [Figure 3.3.3]. The lowest intrarater agreement was observed for the highly subjective low-level features (except homogeneity estimation). For example, the signal intensity evaluation, in which the repeated feature estimation on ADC images demonstrated no significant agreement in rater estimations between the study sessions. Overall, the inexperienced raters displayed higher consistency in their evaluations compared with the experienced radiologists, except in focality assessment on DCE images.

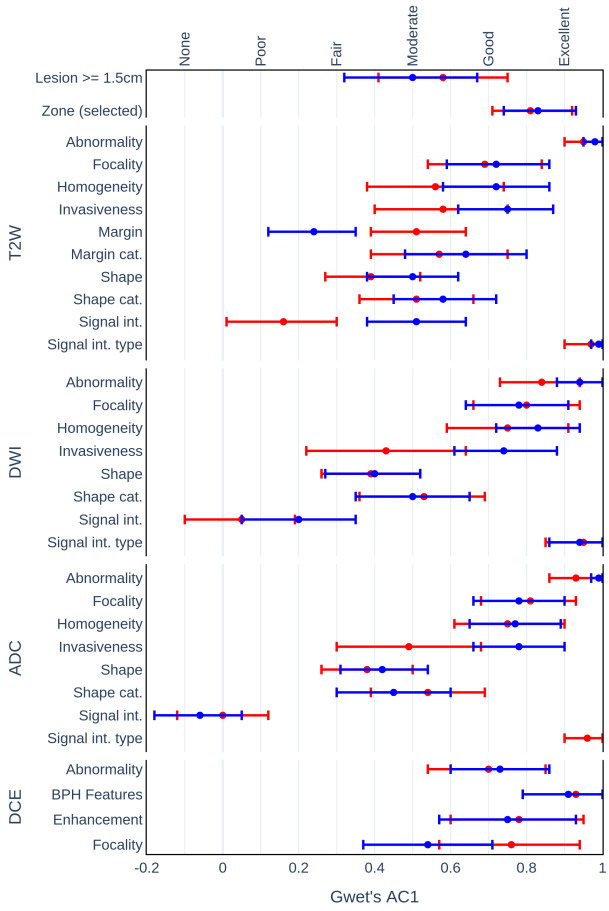


Figure 3.3.3: Mean intrarater agreement (AC1) among the experienced (red) and inexperienced (blue) raters and 95% confidence intervals.

Agreement of the assessment categories

Statistical analysis of the interrater agreement of the PI-RADS categories for the same evaluated lesions between stages [Table 3.3.2] was generally fair to moderate ($0.2 < AC1 < 0.6$). The highest percentage agreement was observed for DCE PI-RADS

Modality	Algorithm	Session 1 PA	Session 1 AC1 (95% CI)	Session 2 PA	Session 2 AC1 (95% CI)	Overall PA	Overall AC1 (95% CI)
OVERALL	PI-RADS Auto.			42.1	0.29 (0.19; 0.39)	42.1	0.29 (0.19; 0.39)
	PI-RADS Man.	47.1	0.35 (0.25; 0.46)	42.7	0.30 (0.19; 0.41)	44.9	0.33 (0.26; 0.40)
T2W	PI-RADS Auto.			41.1	0.28 (0.17; 0.39)	41.1	0.28 (0.17; 0.39)
	PI-RADS Man.	47.1	0.35 (0.26; 0.45)	43.8	0.31 (0.19; 0.43)	45.4	0.33 (0.26; 0.41)
DWI	PI-RADS Auto.			53.0	0.44 (0.33; 0.54)	53.0	0.44 (0.33; 0.54)
	PI-RADS Man.	50.0	0.39 (0.28; 0.50)	49.2	0.38 (0.29; 0.47)	49.6	0.39 (0.32; 0.46)
DCE	PI-RADS Auto.			69.6	0.41 (0.23; 0.59)	69.6	0.41 (0.23; 0.59)
	PI-RADS Man.	69.5	0.48 (0.28; 0.68)	69.8	0.42 (0.24; 0.60)	69.6	0.45 (0.31; 0.58)

Table 3.3.2: Inter-observer agreement of the PI-RADS v2.1 category assessments that were assigned manually (PI-RADS Man.) and determined by the decision tables (PI-RADS Auto.) (p<.001).

evaluation, but it is crucial to note that this evaluation type allows only three outcomes: positive, negative, and unavailable (X). The overall PI-RADS scores assigned by the experienced radiologists (mean = 4.58, standard deviation = 0.71) to the clinically significant lesions were higher (Z = 147, p < .001) than those assigned by the inexperienced radiologists (mean = 4.09; standard deviation = 1.05).

Modality	Algorithm	PA	AC1 (95% CI)	P-value
OVERALL	PI-RADS Man.	64.2	0.56 (0.48; 0.65)	<.001
T2W	PI-RADS Man.	66.3	0.59 (0.50; 0.67)	<.001
DWI	PI-RADS Man.	61.5	0.53 (0.44; 0.62)	<.001
DCE	PI-RADS Man.	76.0	0.55 (0.41; 0.69)	<.001

Table 3.3.3: Intra-observer agreement of manual PI-RADS v2.1 assessment

Table 3.3.3 presents the intraobserver agreement of manual PI-RADS v2.1 category assessments according to the T2W, DWI, DCE, and Overall algorithms between the study stages. All scoring methods demonstrate similar, moderate statistically significant (p < .001) interobserver agreement with respect to AC1 scores.

Agreement between the manually assessed PI-RADS scores and the categories estimated using the PI-RADS algorithm decision table rules [Table 3.3.4] display excellent agreement (PA > 88%, AC1 > 0.86). The agreement between the automatically assigned PI-RADS scores based on the manually specified composite features and manually assessed categories was higher than both the intra- and interobserver agreement for all PI-RADS algorithms.

Modality		PA	AC1 (95% CI)	P-value
OVERALL	PI-RADS	88.3	0.86 (0.80; 0.92)	<.001
T2W	PI-RADS	89.4	0.87 (0.81; 0.93)	<.001
DWI	PI-RADS	92.1	0.90 (0.84; 0.96)	<.001
DCE	PI-RADS	99.5	0.99 (0.97; 1.00)	<.001

Table 3.3.4: Intra-observer agreement of manual and automatic PI-RADS v2.1 assessment

Diagnostic accuracy based on category assessment

To investigate the quality of the radiologists' diagnoses, we used the manually assessed PI-RADS v2.1 categories as a measure of the probability of each lesion's clinical significance. Diagnostic accuracy was assessed by assuming the EAU guidelines of consideration for patient active treatment, in which PI-RADS >= 3 lesions are considered clinically significant and recommended to be histopathologically evaluated.

The AUC results suggest that despite the lower interrater agreement between the experienced radiologists in both estimated features and PI-RADS category assessment, their diagnoses demonstrated superior performance compared with that of the inexperienced radiologists[7]. This applied to lesions located in all zones [Figure 2.3.4]. The assessments of the experienced radiologists showed higher sensitivity (recall = 0.97 vs. 0.85) and precision (0.61 vs. 0.58). Overall the diagnostic decisions demonstrated excellent sensitivity (>0.81, mean = 0.91); the precision, however, was moderate (0.5 < precision < 0.67, mean = 0.58). Maximum observed specificity was 0.50 and lowest 0.06 (mean = 0.34). No statistically significant differences were observed between the results of the first and second stages in terms of assessment quality.

3.3.2. Method validation in a clinical setting

Below we present the results of the prospective study. To analyse feature agreement with respect to the lesions assessed in eighteen patients, the lesions were identified by their sectors by the experts. The majority of lesions were located in the PZ (thirty-eight assessments); lesions were observed three times in the TZ and once in the CZ and AS zones. This aligns with the general tendencies for PCa of incidence frequency

[7] results are presented in the second chapter in the *Comparison with radiology specialists* subsection of the *Results* section.

concerning lesions' zonal locations. Overall, the experts differed in the number of lesions they located. In nine cases, the same number of findings was assessed; in five cases, one of the experts failed to locate any lesions; in four cases, the number of findings differed significantly.

Modality	Algorithm	PA	AC1 (95% CI)	P-value
OVERALL	PI-RADS Auto.	56.2	0.48 (0.13; 0.83)	<.01
	PI-RADS Man.	50.0	0.41 (0.04; 0.78)	<.05
T2W	PI-RADS Auto.	43.8	0.33 (-0.00; 0.67)	.05
	PI-RADS Man.	43.8	0.33 (-0.00; 0.67)	.05
DWI	PI-RADS Auto.	42.9	0.30 (-0.26; 0.85)	.26
	PI-RADS Man.	50.0	0.40 (-0.12; 0.91)	.12
DCE	PI-RADS Auto.	68.8	0.50 (0.01; 0.98)	<.05
	PI-RADS Man.	81.2	0.72 (0.30; 1.00)	<.01

Table 3.3.5: Inter-observer agreement of PI-RADS category assessment during the prospective study

Table 3.3.5 presents the inter-observer agreement of the PI-RADS category assessments that were assigned manually and automatically using the formal model. In terms of AC1 values, agreement ranged from fair to good; however, only the DCE and final PI-RADS scores agreement between raters indicated statistical significance. In only three cases did the experts agree that the images did not contain any clinically significant lesions; in ten cases, they agreed on the lesions' significance. The experts disagreed in their assessments in five cases; in one case, one of them located PCa, while the other did not locate any clinically significant lesions.

Modality		PA	AC1 (95% CI)	P-value
OVERALL	PI-RADS	84.2	0.82 (0.67; 0.98)	<.001
T2W	PI-RADS	97.5	0.97 (0.88; 1.00)	<.001
DWI	PI-RADS	90.0	0.89 (0.63; 1.00)	<.001
DCE	PI-RADS	93.0	0.92 (0.82; 1.00)	<.001

Table 3.3.6: Intra-observer agreement of manual and automatic PI-RADS category assessment during the prospective study

The evaluation of the PI-RADS algorithm was performed by analysing the inter-rater agreement between the assigned assessment categories. Table 3.3.6 presents the

levels of agreement between the automatically determined assessment based on defined features and the manual assessment performed by the radiologists. As with the retrospective study, excellent agreement was observed (AC1 > 0.80, p < .001) between the results of the automatic assessment algorithms and the expert assessments. The highest agreement occurred in the assessment of PI-RADS categories based on the T2W images (PA = 97.5%) and the lowest (PA = 84.2%) in the determination of the final PI-RADS category for a given lesion.

Modality	Feature	PA	AC1 (95% CI)	P-value
	Lesion >= 1.5cm	75.0	0.50 (0.02; 0.98)	<.05
	Zone (selected)	100.0	1.00 (0.67; 1.00)	<.001
T2W	Focality	50.0	0.04 (-0.52; 0.61)	.88
	Homogeneity	50.0	0.08 (-0.48; 0.63)	.77
	Invasiveness	86.7	0.84 (0.48; 1.00)	<.001
	Margin cat.	78.6	0.74 (0.32; 1.00)	<.01
	Shape cat.	28.6	-0.01 (-0.40; 0.38)	.95
	Signal int.	46.7	0.20 (-0.21; 0.62)	.32
	Signal int. type	100.0		
DWI	Focality	60.0	0.28 (-0.29; 0.85)	.32
	Homogeneity	53.3	0.14 (-0.46; 0.74)	.64
	Invasiveness	80.0	0.77 (0.40; 1.00)	<.001
	Shape cat.	33.3	0.04 (-0.36; 0.44)	.84
	Signal int.	46.7	0.25 (-0.16; 0.66)	.22
	Signal int. type	93.3	0.92 (0.58; 1.00)	<.001
ADC	Focality	66.7	0.43 (-0.11; 0.96)	.11
	Homogeneity	53.3	0.09 (-0.47; 0.65)	.74
	Invasiveness	86.7	0.84 (0.48; 1.00)	<.001
	Shape cat.	33.3	0.04 (-0.35; 0.43)	.84
	Signal int.	60.0	0.46 (0.07; 0.85)	<.05
	Signal int. type	100.0		
DCE	Enhancement	84.6	0.83 (0.42; 1.00)	<.001

Table 3.3.7: Inter-observer agreement of PI-RADS CDEs during the prospective study

Analysis of interobserver agreement of the estimated values of PI-RADS CDEs demonstrates excellent agreement for the evaluation of the enhancement on DCE images, invasiveness, and signal intensity type [Table 3.3.7]. The lowest agreement was observed for the shape category and signal intensity assessment on the ADC, DWI, and T2W images. Due to the limited size of the dataset, only a handful of features that displayed high percentage agreement can be considered statistically significant

according to the analysis of AC1 values. The high-level features that are critical in the estimation of prostate lesions' clinical significance demonstrated highly statistically significant agreement: maximum lesion dimension (AC1 = 0.50, $p < .05$); estimation of invasiveness on T2W (AC1 = 0.84, $p < .001$), DWI (AC1 = 0.77, $p < .001$), ADC (AC1 = 0.84, $p < .001$).

3.3.3. Usability tests and conducted interviews

The median time required to assess a single lesion according to the measured interaction time with the tool during the retrospective study was nine minutes and fifteen seconds; this, however is incomparable with clinical practice, as, due to the design of the study, only single-lesion prostate mpMRIs were evaluated, which is not typical for PCa assessment.

Analysis of the comments entered in the text fields in the cases of disagreement between raters and the suggestions that resulted from the integrated decision tables allowed us to investigate the reasons behind some of the disagreements between the automatic and the manual PI-RADS assessments. We have interviewed radiologists involved in both studies and established following reasons for the disagreements:

- manual assessment of PI-RADS category in cases in which the dedicated algorithm was inapplicable (lesions located in the CZ in the prospective study)
- manual assessment of PI-RADS 5 category in cases in which a lesion's maximum dimension extended fifteen millimetres, but estimated features did not qualify the lesion to a higher PI-RADS category according to the official guidelines and defined formal model
- manual assessment of lower categories in cases in which a lesion's maximum dimension did not extend fifteen millimetre, but demonstrated invasive behaviour and qualified for a PI-RADS 5 category according to the official guidelines and defined formal model
- errors made by radiologists during completion of the forms (for example, marking a lesion as hypointense on DWI instead of hyperintense), which resulted in inapplicable rules
- deliberate disagreement with the PI-RADS guidelines and modification of suggested final assessment

- 'intuition' was also stated as reason of disagreement with automatic assessments

The conclusions we drew from the interviews during the retrospective and prospective studies have resulted in several improvements in the design of the proposed CAD/DS form. Changes concerning users' interaction with the tool involved improvements in form layout, limitations of unnecessary fields, improvements in navigation between form sections, and simplification of assessment (for example, hiding unrelated controls, introducing 'YES'/'NO' questions instead of CDE value selection, and automatic categorisation of features). We have also removed the redundant signal characteristic control that corresponds to the signal intensity type, as hypointensity is typical for abnormalities on T2W and ADC, while DWI lesions are hyperintense; this was confirmed by analysis of interrater variability demonstrating almost full agreement in the estimation of signal intensity type.

3.4. Discussion

Our research demonstrates that it is possible to curate annotated datasets through the standardised-structured reporting methods used during diagnostic practice. Interfaces prepared using diagnostic lexicons and the collective knowledge of experienced radiologists enables high-resolution assessment, and the capturing of the intermediate variables that compose final radiological evaluations. By analysing those characteristics, the quality of radiology assessments can be investigated more deeply, enabling the introduction of data-driven improvements to the diagnostic protocols and defined terminology in the radiology lexicons.

Radiologists differ in their assessment of lesions' qualities, number, and the probability of their clinical significance. Our research has shown that inexperienced radiologists tend to underestimate the PI-RADS assessment scores of clinically significant lesions compared with experienced radiologists. This has been also noticed in the work by Mussi et al [131], which indicates that moderately experienced raters were more likely than highly experienced ones to score lesions inconclusively as PI-RADS 3 category than indicating their clinical significance (PI-RADS 4 and 5). These findings suggest that studies on the consistency of PI-RADS evaluation are important, as potential differences in diagnosis contribute to lower recall rates and, thus, the possibility of failing to identify clinically significant lesions. Introducing dedicated computational indicators that estimate the confidence in cases of inconclusive assessment could improve diagnostic accuracy.

Results show that high-level features that require expert knowledge and subjective interpretation demonstrate decreased agreement between raters. During the interviews, the radiologists established that disagreement existed in their interpretations of the 'invasiveness' feature: for some, that feature indicates an extraprostatic extension behaviour; for others, that definition also incorporates lesions that extended to the surrounding zones/sectors. According to the PI-RADS standard, the latter interpretation is correct when considering the assessment rules. High concordance was observed for other high-level features, including part of the DCE algorithm and evaluation of the 'BPH features', which indicate that the gland presents features of benign prostate hyperplasia. Overall, the experienced radiologists demonstrated less agreement than the inexperienced ones did. This was particularly evident in their

evaluations of invasiveness and homogeneity at all stages and focality at the second stage. This contradicts other findings, in which the less experienced raters displayed inferior agreement when evaluating the MRI features [131].

Due to the subjective nature of mpMRI assessment, interrater agreement varies for particular features. No 'gold standard' can be defined by the estimations of a particular radiologist. To construct a reference dataset and assure high-quality annotations, a committee of experienced diagnosticians would have to be involved in rating a substantial dataset of prostate mpMRI. Such data could be then used to enhance the formal model using radiomics to provide objective measures and confidence levels for the features. Setting a gold standard with the help of an expert panel was beyond our organisational and financial capacity. However, in 2021 we received funding for a project in which such verification will be reliably carried out during a multi-centre study[8].

Analysis of interrater agreement performed on the results of the retrospective study reveals that although both the experienced and inexperienced raters differed in their assessments of the PI-RADS categories for the preselected lesions, their evaluations demonstrated high recall scores. Results indicate that the method shows low specificity, meaning that mpMRI diagnosis using the PI-RADS standard can lead to many unnecessary biopsies. Significant differences in the predictive value of the experienced and inexperienced radiologists' diagnoses can be explained by the low agreement between specialists in assessing the high-level features that indicate a lesion's clinical significance, such as invasiveness. Correctly evaluating those traits requires experience in PCa diagnosis.

The excellent agreement between the manual assessments and those resulting from the decision tables support the thesis that the terms of the radiological lexicons can be used to automate the mechanics of estimating the assessment of prostate lesions according to the PI-RADS guidelines. This was verified during the controlled study and in clinical settings. Overall, no cases have been identified in which improper rules, contradictory to the PI-RADS guidelines, were matched by the decision tables based on the features assessed. In all cases of differing manual and automatic assessments, in both the retrospective and prospective studies, the disagreements were caused by

[8] details of this project are presented in the *Future Work* section of the *General Discussion* chapter

differences in interpretation of the diagnostic standards for specific cases and deviation (deliberate or unintentional) from the official PI-RADS guidelines.

One limitation of the prospective study is that in many dimensions, the agreement of responses cannot be considered as statistically significant due to the low number of cases analysed and raters engaged. For assessment of the agreement of most of the individual features that comprise final diagnoses, the power of the AC1 test was insufficient for the results to be evaluated as meaningful. Moreover, the experts-engaged in the prospective study worked at the same department, which potentially introduced interpretation bias; this, however, was not the case in the controlled study, in which the radiologists did not have a history of cooperation. Confirming the effects observed in the clinical setting would require a study involving multiple radiologists on a larger number of cases.

Another limitation resulted from the design of the retrospective study, which involved two sessions of lesion assessment using the structured reporting tool. To assess the overall usefulness of the tool and not merely the added value of automated PI-RADS assessment using integrated decision support, an additional, initial session should occur that involves lesion assessment without any method of computer aid. Due to the limited resources that confined our research to only two sessions, we decided to conduct the first study stage using the structured reporting form that gathered the assessment of CDEs: intermediate variables of mpMRI assessment that used the PI-RADS standard. That allowed us to estimate interrater variability and consistency in assessment using the tool; as a result, however, we were unable to compare the diagnostic accuracy to that resulting from usual diagnostic practice.

We collected opinions of the diagnosticians participating in the test, who pointed to a number of usability advantages, including: verification of inference through suggestions for compliance with diagnostic guidelines, simplicity of report creation (minimising the use of the keyboard in favour of the mouse when completing the form), and clarity and uniformity of the resulting textual reports. Radiologists confirmed the potential of the tool in increasing the availability and reliability of diagnostic standards in clinical practice. The tool allowed radiologists to verify entered parameters in case of discrepancies between manual and suggested assessments. Both experienced and inexperienced professionals noted the potential of the tool in supporting compliance

with diagnostic standards for radiologists in training. Experienced radiologists pointed out that the greatest benefit would provide a solution that reduces the time needed to assess the examination (primarily the time needed to prepare the examination report after visual assessment of the imaging). Additionally, they pointed out the subjectivity of radiological assessment in estimating individual parameters and indicated the potential for introducing objective measures that could assist them in estimating imaging features.

3.5. Conclusions

Structured reporting remains a developing field of research. It enables improvements to the workflows in diagnostics based on medical imaging by reducing ambiguities in the communication of clinical findings to patients and specialists.

We have demonstrated that it is possible to develop structured reporting systems of radiological assessment in PCa diagnosis that integrate with formal descriptions. The domain knowledge contained within diagnostic standards can be integrated with the concept of standardised radiology through the decomposition of guidelines to rules based on well-defined terminology. The use of decision tables based on the features specified by radiologists allows the current standards to be integrated into clinicians' workflows. The diagnostic tool that resulted from the methodology has been verified in a clinical setting and during a controlled study.

Knowledge representation of diagnostic guidelines based on the decision tables applied in structured reporting systems enables constant curation of high-quality datasets. Collecting the data during medical assessments makes the data more accessible for drawing conclusions and further improving the diagnostics without retrospective curation and data annotation. Moreover, it facilitates the introduction of iterative improvements to clinical workflows and improves diagnostic standards based on the insights gathered. This constitutes a feedback loop in which applied domain knowledge enables the collection of datasets that allow further improvements in diagnostic standards.

Our research has demonstrated the need for further work to clarify the concepts and features considered in PI-RADS assessment.

Chapter 4

General Discussion

The contributions of our work must be considered in the context of patient management guidelines, in which noninvasive diagnostics play a major role in patient referral to active surveillance, watchful waiting, and active treatment. Expanding radiologists' cognition through enhanced domain knowledge solutions carries the potential to improve diagnostic decisions; this, in turn, improves patients' quality of life by avoiding unnecessary biopsies and improves active treatment outcomes by early pathology detection and correct identification.

We have demonstrated that domain knowledge can be efficiently applied to construct and improve the machine learning models of PCa diagnosis. Our research on radiomics shows that the feature domain can be designed based on the CDEs that derive from diagnostic standards. Those can be used to identify significant image descriptors that can be used in machine learning models. Radiomics can be used in tandem with mpMRI assessment based on the PI-RADS v2.1 standard and enhance radiologists' cognitive processes by providing concrete objective measures of the features analysed. The results of our work on this subject were presented at two international conferences in 2017 [91], [92]. Our work has contributed towards the machine learning solutions and radiomics used in PCa assessment[1][134]–[137].

Based on the domain knowledge and the set of defined CDEs, we can identify areas that are not covered by feature extraction and might constitute limitations to our method. We have demonstrated that automated solutions of PCa lesion assessment based on intensity and texture features are unable to match the diagnostic performance

[1] Resulting articles were published in conference proceedings and achieved a total of fifteen citations according to Google Scholar as of 10.03.2022.

of experienced radiology specialists. In PCa diagnostics, evaluation of lesions' shape and invasiveness are considered key features.

In addition to being employed in the identification of methods' weaknesses, domain knowledge can also be integrated into deep learning solutions. We have shown that the PI-RADS algorithm can be represented using subnetworks integrated with routing on multimodal CNNs, and tailored fitness functions that favour specific modalities, which are more efficient in the diagnosis of lesions, depending on their locations. Our results show that this intervention resulted in an improved convergence rate in the model—a significant benefit. The results of our work have been published in the PeerJ journal (IF=2.984) [96]. The contributions of our work have been recognised as key in improving the methods of CAD: our published article was among the top five most-viewed 'Radiology and Medical Imaging' and 'Urology' papers published in PeerJ Journal in 2021, reaching 1,057 views and over 250 downloads. This indicates that deep learning methods are gaining more attention in terms of their applications in the automatic diagnosis of PCa. This is indicative of the importance of our study's subject matter in the expansion of diagnostic protocols, in which deep learning can provide crucial new indicators in the assessment of lesions' clinical significance.

Computational methods show great promise in aiding PCa diagnosis. A review of AI methods applied in the field illustrates the dominance of CNN models in their diagnostic accuracy in lesions' clinical significance assessment. This can be explained by the method's ability to construct tailored image descriptors and extract high-level features. Using deep learning signatures alone, or those integrated with classical radiomics signatures can unlock a new method of lesion characterisation. Computer-assisted reporting systems can be enhanced with objective measures of confidence estimated that utilise computational methods as 'second observers', aiding radiologists in cases of doubt or inconclusiveness, and, therefore, supporting specialists' clinical decisions. Integration of tailored descriptors within diagnostic processes can improve consistency among radiologists in the evaluation of certain imaging features.

Multiple ideas and solutions exist that have the potential to improve the work of radiologists by making the domain knowledge better defined, accessible, and applicable during diagnostics workflows. These include structured reporting tools, radiology lexicons, CDEs, and assessment and reporting standards; the integration of these

solutions, however, is lacking. RADS assessment methodologies require continuous improvements that translate into clinical decisions. Diagnostic standards in many cases are not defined based on CDEs, which means that it is impossible to correlate them with other radiology lexicons and create interoperable databases. Moreover, this hinders research on the standards themselves and the features defined in diagnostic decision rules.

In this work, we have demonstrated that it is possible to express RADS rules in terms of CDEs—for example, the PI-RADS diagnostics standard. The data generated through the interaction of radiologists with structured reporting systems facilitates research by constructing annotated datasets that can be utilised to investigate assessment qualities and introduce improvements in diagnostic protocols. Based on the retrospective and prospective studies, which involved multiple inexperienced and experienced radiologists, we have proved that it is possible to formalise diagnostic standards using DMN decision tables. Such tables can easily be updated and integrated within CAR/DS systems to assist radiologists in diagnosis by ensuring assessments' validity and completeness through adherence to diagnostic guidelines.

Basing the computer-assisted structure reporting on CDEs is a form that, through the application, results in possible in-depth analysis of the problem domain by the curation of high-quality annotated datasets. We have demonstrated that data collected through interaction with our system during PCa assessment can be used to analyse the characteristics of features that comprise the PI-RADS guidelines. It is possible to identify the descriptors that characterise poor intra- and interrater agreement; these could potentially benefit from redefinition in radiological lexicons or from the integration of automatically quantified image features. The domain knowledge identified and applied to developing the solutions can be continuously extended by the data-driven conclusions. Overall, we have collaborated on this subject with twelve radiologists—six of whom are experts in the field with over ten years' experience. The interviews we conducted with specialists who had interacted with the CAR/DS system allowed us to improve the solution's usability and confirmed its promise in improving work ergonomics and, by further integration with computational methods, in providing an interface for CAD through structured reporting.

Structured reporting continues to advance radiology. In the future, radiology re-

ports will become more structured, standardised, and patient-specific. The data from structured reports can be analysed to support clinical decisions and enhance patient management through data mining and knowledge discovery. We have demonstrated how domain knowledge may be used to improve the computerised methods of PCa diagnosis and how the solutions can be integrated into diagnostic procedures. Moreover, the experience from radiologists' practice can be used to extend the domain of descriptors, or CDEs—potentially expanding the definitions of diagnostic guidelines.

4.1. Future work

The integration of AI algorithms with the CAR/DS systems can enhance radiologists' work by strengthening the processes of recognition and interpretation of changes with image analysis results that are based on explainable and reliable machine learning methods. This leads directly to the realisation of AI-enhanced radiology. Enhancement of clinical pathways that use computational methods integrated with computer-assisted reporting opens new prospects for personalised medicine by possible integration of clinical pictures resulting from risk factors—as well as clinical, radiomics, genetic, and histopathology data at the stage of medical imaging. Our work in this field contributes to this area of research, which remains open to new solutions—particularly in the case of the evolving RADS guidelines in multiple fields of noninvasive diagnostics.

Reporting processes can integrate formalised models and computational methods to improve diagnosis processes and enable the integration of vast amounts of data to personalise clinicians' approach to diagnostics. This approach can expand beyond radiology, extending its support to clinicians to improve the diagnostic and patient management processes, and, as a result, provide more personalised final clinical decisions. The eRADS system could be expanded into a clinical workflow tool that integrates pathways represented as BPMN diagrams with integrated DMN decision tables that reflect formalised domain knowledge. This advances the formalisation of diagnostic guidelines presented in this thesis into the much broader concept of complete clinical pathway standardisation.

AI systems are often implemented as 'black boxes' in which the steps taken to

reach output are uninterpretable by humans. This incurs ethical and legal implications that impede such systems' use in medical applications and leaves radiologists distrustful towards the technology [66]. The implementation of deep learning solutions that can provide readable interpretations of predictions remains a problematic area of active research. Researchers are attempting to prove that deep learning models are capable of making decisions that match or surpass human performance, and provide understandable justifications for their decisions [62]. The CAR/DS concept can be enhanced using machine learning methods, as modern structured reporting demands AI as an integrated element. Basing computational methods on domain knowledge and high-quality reference datasets enables the creation of explainable and reliable decision-support systems that can be back-integrated with structured reporting tools. Structured reporting can be further enhanced by the integration of feedback from models that are already trained. Such models could be used to automatically complete structured reporting forms and facilitate the process of reporting, which presents radiologists with pre-prepared assessments of imaging. This would decrease the time necessary for experts to produce narrative reports and would alter radiologists' role in the diagnostic process to that of reviewing experts who investigate the diagnosis of independent observers (AI models).

Low interrater agreement in the assessment of mpMRI features negatively affects the quality of the annotations that are assigned to medical examinations. The creation of high-quality reference datasets that could be used for research and the further development of the computational method is crucial to achieving advancements in AI-enhanced CAR/DS systems. This would require the conductance of a multicentre retrospective study involving multiple experienced radiologists who evaluate representative datasets of prostate imaging. A 'gold standard' could then be established by estimating the confidence levels for variables based on expert estimations. The assessments of multiple experienced radiologists are crucial in capturing the intuitions involved in estimating the values of the defined CDEs—for example, the meaning of moderate signal intensity on T2W images: establishing the requirements to classify signal intensity as moderate would require multiple ratings of varying imaging characteristics. This expands to other features— particularly those with high interrater variability.

We propose a particular methodology of integrating the AI models with CAR/DS systems. The reference datasets that utilise the data captured during reporting can be used to feed the AI models; these, in turn, can be used to predict the intermediate variables determined during preparation of reports. Automatic estimation of the attributes that compose final structured reports allows radiologists to investigate and understand the decisions made by the computational methods; this constitutes a way of introducing explainable AI into diagnostic workflows. These models can improve diagnostic accuracy, locate lesions, and act as a 'second opinion', which enables faster and more accurate identification and reporting of suspicious or positive cases. This form of support and decision-making assistance is an important aid in the context of human fatigue, distraction, and concentration problems that often result from overwork (which is common among radiologists due to staff shortages). Basing AI solutions on domain knowledge enables the development of explainable and reliable support systems.

We have received funding for further research on AI-enhanced CAR/DS for PCa assessment[2]. A major part of the project funding is designated for the development of a high-quality reference dataset comprising hundreds of annotated mpMRI images by multiple radiologists. The dataset will be used to develop computational models that are able to pre-fill report forms based on the estimated CDEs from imaging data. The project will result in a state-of-the-art e-learning platform for learning structural reporting of prostate mpMRI studies.

[2] The project, *AI-augmented radiology - detection, reporting and clinical decision making in prostate cancer diagnosis* has been funded by the National Centre for Research and Development with 7,347,082.50 PLN as part of the INFOSTRATEG I programme (INFOSTRATEG-I/0036/2021-00). The project's realisation is planned from 2022 to 2025 at the National Information Processing Institute in Warsaw, Poland. The imaging database and medical subtasks will be implemented in collaboration with the Lower Silesian Oncology Center in Wroclaw, Poland.

Chapter 5

Conclusions

This thesis verifies a research hypothesis concerning the diagnosis of clinically significant PCa. The standardisation of radiological reporting—alongside the elaboration of the PI-RADS reporting standards and the development of CAD methods for PCa detection on mpMRI—remains an active area of research worldwide.

We have demonstrated that it is possible to improve diagnostic procedures by formalising domain knowledge in the CAR/DS system. We have elaborated and technically validated the reporting model on retrospective data and validated its usability in clinically realistic settings during the prospective study. To create the standardised structured reporting method, we utilised common lexicons, reporting templates, and clinical decision-support algorithms. By integrating well-defined, standardised terminology into imaging assessment, the quality and reliability of diagnostic procedures could be investigated.

Our experiments have demonstrated the usefulness of the computational methods and models in supporting the diagnosis process. The domain knowledge can be used to effectively construct and improve tools that support PCa diagnosis. Our results prove that diagnostic guidelines can be integrated into CNN models to facilitate the convergence rate. The research methodology, which is based on the integration of domain knowledge into methods of structured reporting and computational methods, is universal and can easily be adapted to the needs of the radiological reporting and diagnostics of other types of cancer.

The results have further development potential and can contribute significantly to structured reporting methods and the application of machine learning in radiology. Multicentre reference datasets must be curated with a significant number of

CDE-based data annotations to implement AI-enhanced CAR/DS systems. Such methods have the potential to increase the availability of AI-based clinical guidance in radiological reporting, improve communication between radiologists and referring physicians, and lead to practical, reliable, and ethical applications of AI in medicine.

Bibliography

[1] C. Huggins, W. W. Scott, and J. H. Heinen, "Chemical composition of human semen and of the secretions of the prostate and seminal vesicles," *American Journal of Physiology-Legacy Content*, vol. 136, no. 3, pp. 467–473, 1942, Publisher: American Physiological Society.

[2] J. C. Weinreb, J. O. Barentsz, P. L. Choyke, *et al.*, "Pi-rads prostate imaging–reporting and data system: 2015, version 2," *European urology*, vol. 69, no. 1, pp. 16–40, 2016.

[3] J. E. McNeal, "The zonal anatomy of the prostate," *The prostate*, vol. 2, no. 1, pp. 35–49, 1981, Publisher: Wiley Online Library.

[4] H. Sung, J. Ferlay, R. L. Siegel, *et al.*, "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries," en, *CA: A Cancer Journal for Clinicians*, vol. 71, no. 3, pp. 209–249, 2021, ISSN: 1542-4863. DOI: 10.3322/caac.21660. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.3322/caac.21660 (visited on 02/19/2022).

[5] K. R. Nowotworów, "National Cancer Registry," 2018. [Online]. Available: http://onkologia.org.pl/raporty/.

[6] W. Street, "Cancer Facts & Figures 2019," *American Cancer Society: Atlanta, GA, USA*, 2019.

[7] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and

mortality worldwide for 36 cancers in 185 countries," *CA: a cancer journal for clinicians*, vol. 68, no. 6, pp. 394–424, 2018, Publisher: Wiley Online Library.

[8] C. L. Amling, R. H. Riffenburgh, L. Sun, *et al.*, "Pathologic variables and recurrence rates as related to obesity and race in men with prostate cancer undergoing radical prostatectomy," *Journal of Clinical Oncology*, vol. 22, no. 3, pp. 439–445, 2004, Publisher: American Society of Clinical Oncology.

[9] D. G. Bostwick, H. B. Burke, D. Djakiew, *et al.*, "Human prostate cancer risk factors," *Cancer: Interdisciplinary International Journal of the American Cancer Society*, vol. 101, no. S10, pp. 2371–2490, 2004, Publisher: Wiley Online Library.

[10] G. Markozannes, I. Tzoulaki, D. Karli, *et al.*, "Diet, body size, physical activity and risk of prostate cancer: An umbrella review of the evidence," *European Journal of Cancer*, vol. 69, pp. 61–69, 2016, Publisher: Elsevier.

[11] L. A. Plaskon, D. F. Penson, T. L. Vaughan, and J. L. Stanford, "Cigarette smoking and risk of prostate cancer in middle-aged men," *Cancer Epidemiology and Prevention Biomarkers*, vol. 12, no. 7, pp. 604–609, 2003, Publisher: AACR.

[12] B. D. Hayes, L. Brady, M. Pollak, and S. P. Finn, "Exercise and prostate cancer: Evidence and proposed mechanisms for disease modification," *Cancer Epidemiology and Prevention Biomarkers*, vol. 25, no. 9, pp. 1281–1288, 2016, Publisher: AACR.

[13] P. Rawla, "Epidemiology of prostate cancer," *World journal of oncology*, vol. 10, no. 2, p. 63, 2019, Publisher: Elmer Press.

[14] W. J. Catalona, D. S. Smith, T. L. Ratliff, and J. W. Basler, "Detection of organ-confined prostate cancer is increased through prostate-specific antigen—based screening," *Jama*, vol. 270, no. 8, pp. 948–954, 1993, Publisher: American Medical Association.

[15] W. H. Cooner, B. Mosley, C. L. Rutherford, *et al.*, "Prostate cancer detection in a clinical urological practice by ultrasonography, digital rectal examination and prostate specific antigen," *The Journal of urology*, vol. 143, no. 6, pp. 1146–1152, 1990, Publisher: Wolters Kluwer Philadelphia, PA.

[16] D. S. Smith, W. J. Catalona, and J. D. Herschman, "Longitudinal screening for prostate cancer with prostate-specific antigen," *Jama*, vol. 276, no. 16, pp. 1309–1315, 1996, Publisher: American Medical Association.

[17] P. C. N. Mottet *et al.*, *Eau - eanm - estro - esur - isup - siog guidelines on prostate cancer*, 2021.

[18] F. Koerber, R. Waidelich, B. Stollenwerk, and W. Rogowski, "The cost-utility of open prostatectomy compared with active surveillance in early localised prostate cancer," *BMC health services research*, vol. 14, no. 1, pp. 1–15, 2014, Publisher: BioMed Central.

[19] M. J. Roobol, H. A. van Vugt, S. Loeb, *et al.*, "Prediction of prostate cancer risk: The role of prostate volume and digital rectal examination in the ERSPC risk calculators," *European urology*, vol. 61, no. 3, pp. 577–583, 2012, Publisher: Elsevier.

[20] T. Kosaka, R. Mizuno, T. Shinojima, *et al.*, "The implications of prostate-specific antigen density to predict clinically significant prostate cancer in men 50 years," *American journal of clinical and experimental urology*, vol. 2, no. 4, p. 332, 2014, Publisher: e-Century Publishing Corporation.

[21] I. M. Thompson, D. P. Ankerst, C. Chi, *et al.*, "Operating characteristics of prostate-specific antigen in men with an initial PSA level of 3.0 ng/ml or lower," *Jama*, vol. 294, no. 1, pp. 66–70, 2005, Publisher: American Medical Association.

[22] H. B. Carter, "Prostate cancers in men with low PSA levels—must we find them?" *The New England journal of medicine*, vol. 350, no. 22, p. 2292, 2004, Publisher: NIH Public Access.

[23] N. Mottet, J. Bellmunt, M. Bolla, *et al.*, "EAU-ESTRO-SIOG Guidelines on Prostate Cancer. Part 1: Screening, Diagnosis, and Local Treatment with Curative Intent," *European Urology*, vol. 71, no. 4, 2017, ISSN: 18737560. DOI: 10.1016/j.eururo.2016.08.003.

[24] L. C. Thompson and M. R. Pokorny, "Multiparametric MRI in the diagnosis of prostate cancer-a generational change," *Australian family physician*, vol. 44, no. 8, pp. 597–602, 2015.

[25] F. M. Fennessy, A. Fedorov, M. G. Vangel, *et al.*, "Multiparametric MRI as a Biomarker of Response to Neoadjuvant Therapy for Localized Prostate Cancer–A Pilot Study," *Academic radiology*, vol. 27, no. 10, pp. 1432–1439, 2020, Publisher: Elsevier.

[26] H. U. Ahmed, A. E.-S. Bosaily, L. C. Brown, *et al.*, "Diagnostic accuracy of multi-parametric MRI and TRUS biopsy in prostate cancer (PROMIS): A paired validating confirmatory study," *The Lancet*, vol. 389, no. 10071, pp. 815–822, 2017, Publisher: Elsevier.

[27] K. Mistry and G. Cable, "Meta-analysis of prostate-specific antigen and digital rectal examination as screening tests for prostate carcinoma," *The Journal of the American Board of Family Practice*, vol. 16, no. 2, pp. 95–101, 2003, Publisher: Am Board Family Med.

[28] M. Adhyam and A. K. Gupta, "A review on the clinical utility of PSA in cancer prostate," *Indian journal of surgical oncology*, vol. 3, no. 2, pp. 120–129, 2012, Publisher: Springer.

[29] C. Tempany, P. Carrol, and M. Leapman, "The role of magnetic resonance imaging in prostate cancer," *UpToDate. Waltham (MA): UpToDate*, 2018.

[30] C. P. Smith and B. Türkbey, "PI-RADS v2: Current standing and future outlook," *Turkish journal of urology*, vol. 44, no. 3, p. 189, 2018, Publisher: Turkish Association of Urology.

[31] L. Boesen, N. Nørgaard, V. Løgager, *et al.*, "Assessment of the diagnostic accuracy of biparametric magnetic resonance imaging for prostate cancer in biopsy-naive men: The biparametric MRI for detection of prostate cancer (BIDOC) study," *JAMA network open*, vol. 1, no. 2, e180219–e180219, 2018, Publisher: American Medical Association.

[32] S. Mehralivand, J. H. Shih, S. Rais-Bahrami, *et al.*, "A magnetic resonance imaging–based prediction model for prostate biopsy risk stratification," *JAMA oncology*, vol. 4, no. 5, pp. 678–685, 2018, Publisher: American Medical Association.

[33] D. M. Rocha, L. M. Brasil, J. M. Lamas, G. V. Luz, and S. S. Bacelar, "Evidence of the benefits, advantages and potentialities of the structured

radiological report: An integrative review," *Artificial intelligence in medicine*, vol. 102, p. 101 770, 2020, Publisher: Elsevier.

[34] P. Hickey, "The interpretation of radiographs," *J Mich Med Soc*, vol. 3, pp. 496–9, 1904.

[35] ——, "Standardization of roentgen-ray reports," *AJR Am J Roentgenol*, vol. 9, no. 422, e6, 1922.

[36] A. Wallis and P. McCoubrie, "The radiology report—are we getting the message across?" *Clinical radiology*, vol. 66, no. 11, pp. 1015–1022, 2011, Publisher: Elsevier.

[37] E. S. Burnside, E. A. Sickles, L. W. Bassett, *et al.*, "The ACR BI-RADS® experience: Learning from history," *Journal of the American College of Radiology*, vol. 6, no. 12, pp. 851–860, 2009, Publisher: Elsevier.

[38] W. Scott, "Establishing mammographic criteria for recommending surgical biopsy," *Report of the Council on Scientific Affairs. Chicago, IL: American Medical Association*, 1989.

[39] C. D'orsi and D. Kopans, "Mammography interpretation: The BI-RADS method.," *American family physician*, vol. 55, no. 5, pp. 1548–1551, 1997, Publisher: American Academy of Family Physicians.

[40] F. M. Hall, "The radiology report of the future," *Radiology*, vol. 251, no. 2, pp. 313–316, 2009, Publisher: Radiological Society of North America.

[41] S. Demigha and C. Rolland, "Training-aided system in senology: Methodologies and techniques," in *Medical Imaging 2003: PACS and Integrated Medical Information Systems: Design and Evaluation*, vol. 5033, SPIE, 2003, pp. 339–349.

[42] H. Ojeda-Fournier and J. Q. Nguyen, "How to improve your breast cancer program: Standardized reporting using the new American College of Radiology Breast Imaging-Reporting and Data System," *Indian Journal of Radiology and Imaging*, vol. 19, no. 04, pp. 266–277, 2009, Publisher: Thieme Medical and Scientific Publishers Private Ltd.

[43] J. Y. An, K. M. Unsdorfer, and J. C. Weinreb, "BI-RADS, C-RADS, CAD-RADS, LI-RADS, Lung-RADS, NI-RADS, O-RADS, PI-RADS,

TI-RADS: Reporting and Data Systems," *RadioGraphics*, vol. 39, no. 5, pp. 1435–1436, 2019, Publisher: Radiological Society of North America.

[44] H. Shaish, W. Feltus, J. Steinman, E. Hecht, S. Wenske, and F. Ahmed, "Impact of a structured reporting template on adherence to Prostate Imaging Reporting and Data System version 2 and on the diagnostic performance of prostate MRI for clinically significant prostate cancer," *Journal of the American College of Radiology*, vol. 15, no. 5, pp. 749–754, 2018, Publisher: Elsevier.

[45] R. T. Gupta, K. A. Mehta, B. Turkbey, and S. Verma, "PI-RADS: Past, present, and future," *Journal of Magnetic Resonance Imaging*, vol. 52, no. 1, pp. 33–53, 2020, Publisher: Wiley Online Library.

[46] R. Likert, "A technique for the measurement of attitudes.," *Archives of psychology*, 1932.

[47] B. Turkbey, A. B. Rosenkrantz, M. A. Haider, *et al.*, "Prostate imaging reporting and data system version 2.1: 2019 update of prostate imaging reporting and data system version 2," *European urology*, 2019, Publisher: Elsevier.

[48] H. Van Poppel, R. Hogenhout, P. Albers, R. C. van den Bergh, J. O. Barentsz, and M. J. Roobol, "Early detection of prostate cancer in 2020 and beyond: Facts and recommendations for the European Union and the European Commission," *Screening*, vol. 73, p. 56, 2021.

[49] J. I. Epstein, "An update of the Gleason grading system," *The Journal of urology*, vol. 183, no. 2, pp. 433–440, 2010, Publisher: Wolters Kluwer Philadelphia, PA.

[50] G. L. Lu-Yao, P. C. Albertsen, D. F. Moore, *et al.*, "Outcomes of localized prostate cancer following conservative management," *Jama*, vol. 302, no. 11, pp. 1202–1209, 2009, Publisher: American Medical Association.

[51] G. Oster, L. Lamerato, A. G. Glass, *et al.*, "Natural history of skeletal-related events in patients with breast, lung, or prostate cancer and metastases to bone: A 15-year study in two large US health systems," *Supportive Care in Cancer*, vol. 21, no. 12, pp. 3279–3286, 2013, Publisher: Springer.

[52] L. Ye, H. G. Kynaston, and W. G. Jiang, "Bone metastasis in prostate cancer: Molecular and cellular mechanisms," *International journal of molecular medicine*, vol. 20, no. 1, pp. 103–111, 2007, Publisher: Spandidos Publications.

[53] H. Cash, A. Maxeiner, C. Stephan, *et al.*, "The detection of significant prostate cancer is correlated with the Prostate Imaging Reporting and Data System (PI-RADS) in MRI/transrectal ultrasound fusion biopsy," *World Journal of Urology*, vol. 34, no. 4, pp. 525–532, Apr. 2016, ISSN: 1433-8726. DOI: 10.1007/s00345-015-1671-8. [Online]. Available: https://doi.org/10.1007/s00345-015-1671-8.

[54] S. Polanec, T. H. Helbich, H. Bickel, *et al.*, "Head-to-head comparison of PI-RADS v2 and PI-RADS v1," *European Journal of Radiology*, vol. 85, no. 6, pp. 1125–1131, 2016, ISSN: 0720-048X. DOI: https://doi.org/10.1016/j.ejrad.2016.03.025.

[55] S. Y. Park, D. C. Jung, Y. T. Oh, *et al.*, "Prostate cancer: PI-RADS version 2 helps preoperatively predict clinically significant cancers," *Radiology*, vol. 280, no. 1, pp. 108–116, 2016, Publisher: Radiological Society of North America.

[56] R. Bhayana, A. O'Shea, M. A. Anderson, *et al.*, "PI-RADS versions 2 and 2.1: Interobserver agreement and diagnostic performance in peripheral and transition zone lesions among six radiologists," *American Journal of Roentgenology*, vol. 217, no. 1, pp. 141–151, 2021, Publisher: Am Roentgen Ray Soc.

[57] A. B. Rosenkrantz, L. A. Ginocchio, D. Cornfeld, *et al.*, "Interobserver reproducibility of the PI-RADS version 2 lexicon: A multicenter study of six experienced prostate radiologists," *Radiology*, vol. 280, no. 3, pp. 793–804, 2016, Publisher: Radiological Society of North America.

[58] C. I. Farmer, A. M. Bourne, D. O'Connor, J. G. Jarvik, and R. Buchbinder, "Enhancing clinician and patient understanding of radiology reports: A scoping review of international guidelines," *Insights into imaging*, vol. 11, pp. 1–10, 2020, Publisher: Springer.

[59] D. L. Rubin and C. E. Kahn Jr, "Common data elements in radiology," *Radiology*, vol. 283, no. 3, pp. 837–844, 2016, Publisher: Radiological Society of North America.

[60] N. G. U. T. UK, "Prostate cancer: Diagnosis and management," 2019, Publisher: National Institute for Health and Care Excellence (UK).

[61] R. N. Shiffman, "Representation of clinical practice guidelines in conventional and augmented decision tables," *Journal of the American Medical Informatics Association*, vol. 4, no. 5, pp. 382–393, 1997, Publisher: BMJ Group BMA House, Tavistock Square, London, WC1H 9JR.

[62] N. Zha, M. N. Patlas, and R. Duszak, "Radiologist burnout is not just isolated to the United States: Perspectives from Canada," *J Am Coll Radiol*, vol. 16, no. 1, pp. 121–123, 2019.

[63] J. H. Thrall, X. Li, Q. Li, *et al.*, "Artificial intelligence and machine learning in radiology: Opportunities, challenges, pitfalls, and criteria for success," *Journal of the American College of Radiology*, vol. 15, no. 3, pp. 504–508, 2018, Publisher: Elsevier.

[64] S. Russell and P. Norvig, "Artificial intelligence: A modern approach," 2002.

[65] G. Choy, O. Khalilzadeh, M. Michalski, *et al.*, "Current applications and future impact of machine learning in radiology," *Radiology*, vol. 288, no. 2, pp. 318–328, 2018, Publisher: Radiological Society of North America.

[66] T. M. Noguerol, F. Paulano-Godino, M. T. Martín-Valdivia, C. O. Menias, and A. Luna, "Strengths, weaknesses, opportunities, and threats analysis of artificial intelligence and machine learning applications in radiology," *Journal of the American College of Radiology*, vol. 16, no. 9, pp. 1239–1247, 2019, Publisher: Elsevier.

[67] J. M. Castillo T, M. Arif, W. J. Niessen, I. G. Schoots, J. F. Veenland, *et al.*, "Automated classification of significant prostate cancer on MRI: A systematic review on the performance of machine learning applications," *Cancers*, vol. 12, no. 6, p. 1606, 2020, Publisher: Multidisciplinary Digital Publishing Institute.

[68] M. Avanzo, L. Wei, J. Stancanello, *et al.*, "Machine and deep learning methods for radiomics," *Medical physics*, vol. 47, no. 5, e185–e202, 2020, Publisher: Wiley Online Library.

[69] G. Langs, S. Röhrich, J. Hofmanninger, *et al.*, "Machine learning: From radiomics to discovery and routine," *Der Radiologe*, vol. 58, no. 1, pp. 1–6, 2018, Publisher: Springer.

[70] A. Zwanenburg, M. Vallières, M. A. Abdalah, *et al.*, "The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping," *Radiology*, vol. 295, no. 2, pp. 328–338, May 2020, Publisher: Radiological Society of North America, ISSN: 0033-8419. DOI: 10.1148/radiol.2020191145. [Online]. Available: https://pubs.rsna.org/doi/full/10.1148/radiol.2020191145 (visited on 02/20/2022).

[71] G. Litjens, O. Debats, J. Barentsz, N. Karssemeijer, and H. Huisman, *SPIE-AAPM PROSTATEx Challenge Data*, 2017. DOI: 10.7937/K9TCIA.2017.MURS5CL. [Online]. Available: https://wiki.cancerimagingarchive.net/x/iIFpAQ.

[72] G. Litjens, T. Kooi, B. E. Bejnordi, *et al.*, "A survey on deep learning in medical image analysis," *Medical image analysis*, vol. 42, pp. 60–88, 2017, Publisher: Elsevier.

[73] Y. Song, Y.-D. Zhang, X. Yan, *et al.*, "Computer-aided diagnosis of prostate cancer using a deep convolutional neural network from multiparametric MRI," *Journal of Magnetic Resonance Imaging*, vol. 48, no. 6, pp. 1570–1577, 2018. DOI: 10.1002/jmri.26047. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/jmri.26047.

[74] X. Wang, W. Yang, J. Weinreb, *et al.*, "Searching for prostate cancer by fully automated magnetic resonance imaging classification: Deep learning versus non-deep learning," *Scientific Reports*, vol. 7, Dec. 2017. DOI: 10.1038/s41598-017-15720-y.

[75] X. Yang, C. Liu, Z. Wang, *et al.*, "Co-trained convolutional neural networks for automated detection of prostate cancer in multi-parametric MRI," *Medical Image Analysis*, vol. 42, pp. 212–227, 2017, ISSN: 1361-8415. DOI: https://doi.org/10.1016/j.media.2017.08.006. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1361841517301299.

[76] M. H. Le, J. Chen, L. Wang, *et al.*, "Automated diagnosis of prostate cancer in multi-parametric MRI based on multimodal convolutional neural networks," *Physics in Medicine & Biology*, vol. 62, no. 16, pp. 6497–6514, Jul.

2017, Publisher: IOP Publishing. DOI: 10.1088/1361-6560/aa7731. [Online]. Available: https://doi.org/10.1088%2F1361-6560%2Faa7731.

[77] X. Yang, Z. Wang, C. Liu, *et al.*, "Joint detection and diagnosis of prostate cancer in multi-parametric MRI based on multimodal convolutional neural networks," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2017, pp. 426–434.

[78] R. Alkadi, F. Taher, A. El-baz, and N. Werghi, "A Deep Learning-Based Approach for the Detection and Localization of Prostate Cancer in T2 Magnetic Resonance Images," *Journal of Digital Imaging*, vol. 32, no. 5, pp. 793–807, Oct. 2019, ISSN: 1618-727X. DOI: 10.1007/s10278-018-0160-1. [Online]. Available: https://doi.org/10.1007/s10278-018-0160-1.

[79] A. P. Kiraly, C. A. Nader, A. Tuysuzoglu, *et al.*, "Deep convolutional encoder-decoders for prostate cancer detection and classification," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2017, pp. 489–497.

[80] P. Schelb, S. Kohl, J. P. Radtke, *et al.*, "Classification of cancer at prostate MRI: Deep Learning versus Clinical PI-RADS Assessment," *Radiology*, vol. 293, no. 3, 2019, ISSN: 15271315. DOI: 10.1148/radiol.2019190938.

[81] J. Ishioka, Y. Matsuoka, S. Uehara, *et al.*, "Computer-aided diagnosis of prostate cancer on magnetic resonance imaging using a convolutional neural network algorithm," *BJU International*, vol. 122, no. 3, pp. 411–417, 2018. DOI: 10.1111/bju.14397. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/bju.14397.

[82] T. Zhou, S. Ruan, and S. Canu, "A review: Deep learning for medical image segmentation using multi-modality fusion," *Array*, p. 100 004, 2019, Publisher: Elsevier.

[83] S. G. Armato, H. Huisman, K. Drukker, *et al.*, "PROSTATEx Challenges for computerized classification of prostate lesions from multiparametric magnetic resonance images," *Journal of Medical Imaging*, vol. 5, no. 4, p. 044 501, 2018, Publisher: International Society for Optics and Photonics.

[84] A. Gebejes and R. Huertas, "Texture characterization based on grey-level co-occurrence matrix," *Databases*, vol. 9, no. 10, pp. 375–378, 2013.

[85] E. Miyamoto and T. Merryman, "Fast calculation of Haralick texture features," *Human computer interaction institute, Carnegie Mellon University, Pittsburgh, USA. Japanese restaurant office*, 2005.

[86] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural Features for Image Classification," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-3, no. 6, pp. 610–621, Nov. 1973, ISSN: 0018-9472. DOI: 10.1109/TSMC.1973.4309314.

[87] A. Wibmer, H. Hricak, T. Gondo, *et al.*, "Haralick texture analysis of prostate MRI: Utility for differentiating non-cancerous prostate from prostate cancer and differentiating prostate cancers with different Gleason scores," *European radiology*, vol. 25, no. 10, pp. 2840–2850, 2015, Publisher: Springer.

[88] P. Brynolfsson, D. Nilsson, T. Torheim, *et al.*, "Haralick texture features from apparent diffusion coefficient (ADC) MRI images depend on imaging and pre-processing parameters," en, *Scientific Reports*, vol. 7, no. 1, p. 4041, Jun. 2017, Number: 1 Publisher: Nature Publishing Group, ISSN: 2045-2322. DOI: 10.1038/s41598-017-04151-4. [Online]. Available: https://www.nature.com/articles/s41598-017-04151-4 (visited on 02/19/2022).

[89] A. S. Kierans, H. Rusinek, A. Lee, *et al.*, "Textural differences in apparent diffusion coefficient between low-and high-stage clear cell renal cell carcinoma," *American Journal of Roentgenology*, vol. 203, no. 6, W637–W644, 2014, Publisher: American Roentgen Ray Society.

[90] A. Vignati, S. Mazzetti, V. Giannini, *et al.*, "Texture features on T2-weighted magnetic resonance imaging: New potential biomarkers for prostate cancer aggressiveness," *Physics in Medicine & Biology*, vol. 60, no. 7, p. 2685, 2015, Publisher: IOP Publishing.

[91] P. Sobecki, D. Życka-Malesa, I. Mykhalevych, K. Sklinda, and A. Przelaskowski, "MRI imaging texture features in prostate lesions classification," in *EMBEC & NBC 2017*, Springer, 2017, pp. 827–830.

[92] P. Sobecki, D. Życka-Malesa, I. Mykhalevych, A. Gora, K. Sklinda, and A. Przelaskowski, "Feature extraction optimized for prostate lesion

classification," in *Proceedings of the 9th International Conference on Bioinformatics and Biomedical Technology*, ACM, 2017, pp. 22–27.

[93] J. T. Kwak, S. Xu, B. J. Wood, *et al.*, "Automated prostate cancer detection using T2-weighted and high-b-value diffusion-weighted magnetic resonance imaging," *Medical physics*, vol. 42, no. 5, pp. 2368–2378, 2015, Publisher: Wiley Online Library.

[94] O. H. Babatunde, L. Armstrong, J. Leng, and D. Diepeveen, "A genetic algorithm-based feature selection," 2014, Publisher: IJECCE.

[95] L. P. Coelho, "Mahotas: Open source software for scriptable computer vision," *arXiv preprint arXiv:1211.4907*, 2012.

[96] P. Sobecki, R. Jóźwiak, K. Sklinda, and A. Przelaskowski, "Effect of domain knowledge encoding in cnn model architecture—a prostate cancer study using mpmri images," *PeerJ*, vol. 9, e11006, 2021.

[97] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[98] M. Abadi, P. Barham, J. Chen, *et al.*, "Tensorflow: A system for large-scale machine learning," in *12th USENIX Symposium on Operating Systems Design and Implementation OSDI 16)*, 2016, pp. 265–283.

[99] R. F. Woolson, "Wilcoxon signed-rank test," *Wiley encyclopedia of clinical trials*, pp. 1–3, 2007, Publisher: Wiley Online Library.

[100] M. Scialpi, A. D'Andrea, E. Martorana, *et al.*, "Biparametric MRI of the prostate," *Turkish journal of urology*, vol. 43, no. 4, p. 401, 2017, Publisher: Turkish Association of Urology.

[101] L. Boesen, N. Nørgaard, V. Løgager, *et al.*, "Assessment of the diagnostic accuracy of biparametric magnetic resonance imaging for prostate cancer in biopsy-naive men: The biparametric mri for detection of prostate cancer (bidoc) study," *JAMA Network Open*, vol. 1, no. 2, e180219–e180219, 2018.

[102] A. Mehrtash, A. Sedghi, M. Ghafoorian, *et al.*, "Classification of clinical significance of MRI prostate findings using 3D convolutional neural networks," in *Medical Imaging 2017: Computer-Aided Diagnosis*, vol. 10134, International Society for Optics and Photonics, 2017, 101342A.

[103] S. Liu, H. Zheng, Y. Feng, and W. Li, "Prostate cancer diagnosis using deep learning with 3D multiparametric MRI," in *Medical Imaging 2017: Computer-Aided Diagnosis*, vol. 10134, International Society for Optics and Photonics, 2017, p. 1 013 428.

[104] S. Loeb, A. Vellekoop, H. U. Ahmed, *et al.*, "Systematic Review of Complications of Prostate Biopsy," en, *European Urology*, vol. 64, no. 6, pp. 876–892, Dec. 2013, ISSN: 0302-2838. DOI: 10.1016/j.eururo.2013.05.049. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0302283813005587 (visited on 02/18/2022).

[105] , "ESR paper on structured reporting in radiology," *Insights into imaging*, vol. 9, pp. 1–7, 2018, Publisher: Springer.

[106] D. L. Weiss and P. R. Bolos, "Reporting and dictation," in *Practical Imaging Informatics*, Springer, 2009, pp. 147–162.

[107] K. Juluru, M. E. Heilbrun, and M. D. Kohli, "Describing disease-specific reporting guidelines: A brief guide for radiologists," *RadioGraphics*, vol. 39, no. 5, pp. 1233–1235, 2019, PMID: 31498741. DOI: 10.1148/rg.2019190182. eprint: https://doi.org/10.1148/rg.2019190182. [Online]. Available: https://doi.org/10.1148/rg.2019190182.

[108] N. M. Safdar, E. Siegel, B. J. Erickson, and P. Nagy, "Enabling comparative effectiveness research with informatics: Show me the data!" *Academic radiology*, vol. 18, no. 9, pp. 1072–1076, 2011.

[109] E. R. Ranschaert, S. Morozov, and P. R. Algra, *Artificial Intelligence in Medical Imaging: Opportunities, Applications and Risks*. Springer, 2018.

[110] J. W. Nance Jr, C. Meenan, and P. G. Nagy, "The Future of the Radiology Information System," *American Journal of Roentgenology*, vol. 200, no. 5, pp. 1064–1070, May 2013, Publisher: American Roentgen Ray Society, ISSN: 0361-803X. DOI: 10.2214/AJR.12.10326. [Online]. Available: https://www.ajronline.org/doi/full/10.2214/AJR.12.10326 (visited on 02/20/2022).

[111] C. P. Langlotz, *RadLex: A new method for indexing online educational materials*, Issue: 6 Pages: 1595–1597 Publication Title: Radiographics Volume: 26, 2006.

[112] K. Donnelly *et al.*, "SNOMED-CT: The advanced terminology and coding system for eHealth," *Studies in health technology and informatics*, vol. 121, p. 279, 2006, Publisher: IOS Press; 1999.

[113] C. J. McDonald, S. M. Huff, J. G. Suico, *et al.*, "LOINC, a universal standard for identifying laboratory observations: A 5-year update," *Clinical chemistry*, vol. 49, no. 4, pp. 624–633, 2003, Publisher: Oxford University Press.

[114] F. Jungmann, G. Arnhold, B. Kämpgen, *et al.*, "A hybrid reporting platform for extended RadLex coding combining structured reporting templates and natural language processing," *Journal of digital imaging*, vol. 33, no. 4, pp. 1026–1033, 2020, Publisher: Springer.

[115] J. M. Bosmans, E. Neri, O. Ratib, and C. E. Kahn Jr, "Structured reporting: A fusion reactor hungry for fuel," *Insights into imaging*, vol. 6, no. 1, pp. 129–132, 2015, Publisher: SpringerOpen.

[116] J. Hirsch, G. Nicola, G. McGinty, *et al.*, "ICD-10: History and context," *American Journal of Neuroradiology*, vol. 37, no. 4, pp. 596–599, 2016, Publisher: Am Soc Neuroradiology.

[117] M. Mustra, K. Delac, and M. Grgic, "Overview of the DICOM standard," in *2008 50th International Symposium ELMAR*, vol. 1, IEEE, 2008, pp. 39–44.

[118] D. L. Rubin, "Creating and curating a terminology for radiology: Ontology modeling and analysis," *Journal of digital imaging*, vol. 21, no. 4, pp. 355–362, 2008, Publisher: Springer.

[119] T. Vetterlein, H. Mandl, and K.-P. Adlassnig, "Fuzzy Arden Syntax: A fuzzy programming language for medicine," en, *Artificial Intelligence in Medicine*, vol. 49, no. 1, pp. 1–10, May 2010, ISSN: 0933-3657. DOI: 10.1016/j.artmed.2010.01.003. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0933365710000047 (visited on 03/10/2022).

[120] R. Braun, H. Schlieter, M. Burwitz, and W. Esswein, "BPMN4CP: Design and implementation of a BPMN extension for clinical pathways," in *2014*

IEEE international conference on bioinformatics and biomedicine (BIBM), IEEE, 2014, pp. 9–16.

[121] ——, "BPMN4CP Revised–Extending BPMN for Multi-perspective Modeling of Clinical Pathways," in *2016 49th Hawaii International Conference on System Sciences (HICSS)*, IEEE, 2016, pp. 3249–3258.

[122] *Designing a decision service using DMN models*, Red Hat Customer Portal. [Online]. Available: https://access.redhat.com/documentation/en-us/red_hat_decision_manager/7.0/html/designing_a_decision_service_using_dmn_models/dmn-elements-ref.

[123] B. C. Bizzo, R. R. Almeida, and T. K. Alkasab, "Computer-Assisted Reporting and Decision Support in Standardized Radiology Reporting for Cancer Imaging," *JCO Clinical Cancer Informatics*, no. 5, pp. 426–434, Dec. 2021, Publisher: Wolters Kluwer. DOI: 10.1200/CCI.20.00129. [Online]. Available: https://ascopubs.org/doi/abs/10.1200/CCI.20.00129 (visited on 02/20/2022).

[124] R. R. Almeida, B. C. Bizzo, R. Singh, K. P. Andriole, and T. K. Alkasab, "Computer-assisted Reporting and Decision Support Increases Compliance with Follow-up Imaging and Hormonal Screening of Adrenal Incidentalomas," eng, *Academic Radiology*, vol. 29, no. 2, pp. 236–244, Feb. 2022, ISSN: 1878-4046. DOI: 10.1016/j.acra.2021.01.019.

[125] T. K. Alkasab, B. C. Bizzo, L. L. Berland, S. Nair, P. V. Pandharipande, and H. B. Harvey, "Creation of an Open Framework for Point-of-Care Computer-Assisted Reporting and Decision Support Tools for Radiologists," English, *Journal of the American College of Radiology*, vol. 14, no. 9, pp. 1184–1189, Sep. 2017, Publisher: Elsevier, ISSN: 1546-1440, 1558-349X. DOI: 10.1016/j.jacr.2017.04.031. [Online]. Available: https://www.jacr.org/article/S1546-1440(17)30549-5/fulltext (visited on 02/20/2022).

[126] A. K. Goel, W. S. Campbell, and R. Moldwin, "Structured data capture for oncology," *JCO Clinical Cancer Informatics*, vol. 5, pp. 194–201, 2021.

[127] L. Brunese, M. C. Brunese, M. Carbone, V. Ciccone, F. Mercaldo, and A. Santone, "Automatic PI-RADS assignment by means of formal methods,"

en, *La radiologia medica*, vol. 127, no. 1, pp. 83–89, Jan. 2022, ISSN: 1826-6983. DOI: 10.1007/s11547-021-01431-y. [Online]. Available: https://doi.org/10.1007/s11547-021-01431-y (visited on 02/20/2022).

[128] W. W. W. Consortium *et al.*, "Xforms 1.1," 2009.

[129] N. van den Bleeken, "Building rich web applications using xforms 2.0," *XML LONDON 2013*, 2013.

[130] *Documentary tv series czarna domena (eng. black domain), season 1 episode 8 sztuczna inteligencja (eng. artificial intelligence)*, Warsaw, 2019.

[131] T. C. Mussi, F. I. Yamauchi, C. F. Tridente, *et al.*, "Interobserver agreement of PI-RADS v. 2 lexicon among radiologists with different levels of experience," *Journal of Magnetic Resonance Imaging*, vol. 51, no. 2, pp. 593–602, 2020, Publisher: Wiley Online Library.

[132] K. L. Gwet, "Computing inter-rater reliability and its variance in the presence of high agreement," *British Journal of Mathematical and Statistical Psychology*, vol. 61, no. 1, pp. 29–48, 2008, Publisher: Wiley Online Library.

[133] K. L. Gwet, "irrCAC: Computing chance-corrected agreement coefficients (CAC)," *R Package version*, vol. 1, 2019.

[134] Y. Yuan, W. Qin, M. Buyyounouski, *et al.*, "Prostate cancer classification with multiparametric MRI transfer learning model," en, *Medical Physics*, vol. 46, no. 2, pp. 756–765, 2019, ISSN: 2473-4209. DOI: 10.1002/mp.13367. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/mp.13367 (visited on 02/13/2022).

[135] Y. Sun, H. M. Reynolds, B. Parameswaran, *et al.*, "Multiparametric MRI and radiomics in prostate cancer: A review," en, *Australasian Physical & Engineering Sciences in Medicine*, vol. 42, no. 1, pp. 3–25, Mar. 2019, ISSN: 1879-5447. DOI: 10.1007/s13246-019-00730-z. [Online]. Available: https://doi.org/10.1007/s13246-019-00730-z (visited on 02/13/2022).

[136] R. Cuocolo, M. B. Cipullo, A. Stanzione, *et al.*, "Machine learning for the identification of clinically significant prostate cancer on MRI: A meta-analysis," en, *European Radiology*, vol. 30, no. 12, pp. 6877–6887, Dec. 2020, ISSN: 1432-1084. DOI: 10.1007/s00330-020-07027-w. [Online].

Available: https://doi.org/10.1007/s00330-020-07027-w (visited on 02/13/2022).

[137] C. Jensen, J. Carl, L. Boesen, N. C. Langkilde, and L. R. Østergaard, "Assessment of prostate cancer prognostic Gleason grade group using zonal-specific features extracted from biparametric MRI using a KNN classifier," en, *Journal of Applied Clinical Medical Physics*, vol. 20, no. 2, pp. 146–153, 2019, ISSN: 1526-9914. DOI: 10.1002/acm2.12542. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/acm2.12542 (visited on 02/13/2022).

List of Figures

List of Tables

Appendix

A1. Decision tables definitions

A1.1. DWI PI-RADS variables and rules

Variable	Label	Related Radlex Terms	Possible Values
Input Variables			
lesion_dim_max	Max dim (mm)	Diameter [RID13432]	<5, >=5, >=15
lesion_location	Zone	Zone of prostate [RID38890]	PZ, TZ, NOT_AVAILABLE
adc_present_and_adequate	ADC present and adequate	Adequate [RID39308]	YES, NO
dwi_present_and_adequate	DWI present and adequate	Adequate [RID39308]	YES, NO
adc_abnormality	ADC lesion present	Lesion [RID38780]	YES, NO
adc_invasive	ADC Invasive	Invasive [RID5680]	YES, NO
adc_signal_intensity_type	ADC Signal Intensity Type	Signal characteristic [RID6049]	HYPOINTENSITIVITY, ISOINTENSITIVITY, HYPERINTENSITIVITY
adc_signal_intensity	ADC Signal Intensity Scale	Signal characteristic [RID6049]	MILD, MODERATE, MARKEDLY
adc_focality	ADC Focality	Focal [RID5702]	YES, NO
adc_shape	ADC Shape	Morphologic descriptor [RID5863]	LINEAR, WEDGE, LENTICULAR, WATER-DROP
adc_shape_category	ADC Shape category	Morphologic descriptor [RID5863]	LINEAR, ROUND, IRREGULAR
dwi_abnormality	DWI lesion present	Lesion [RID38780]	YES, NO
dwi_invasive	DWI Invasive	Invasive [RID5680]	YES, NO
dwi_signal_intensity_type	DWI Signal Intensity Type	Signal characteristic [RID6049]	HYPOINTENSITIVITY, ISOINTENSITIVITY, HYPERINTENSITIVITY
dwi_signal_intensity	DWI Signal Intensity Scale	Signal characteristic [RID6049]	MILD, MODERATE, MARKEDLY
dwi_focality	DWI Focality	Focal [RID5702]	YES, NO
dwi_shape	DWI Shape	Morphologic descriptor [RID5863]	LINEAR, WEDGE, LENTICULAR, WATER-DROP
dwi_shape_category	DWI Shape category	Morphologic descriptor [RID5863]	LINEAR, ROUND, IRREGULAR
Output Variables			
dwi_pirads	PI-RADS Evaluation	PI-RADS DWI Lesion Assessment Category [RID50313]	1, 2, 3, 4, 5, X
dwi_pirads_description	PI-RADS Rule Descripton	-	<string>

Input and output variables of decision table representing the PI-RADS DWI-ADC assessment algorithm

PI-RADS	Description
X	[#1] PI-RADS evaluation not avaible for the selected zone
1	[#2] No lesions
5	[#3] ADC: Invasive; DWI: Invasive; Max dim. >=5 mm
5	[#4] ADC: Invasive; DWI: Non-invasive; Max dim. >=5 mm
5	[#5] ADC: Non-invasive; DWI: Invasive; Max dim. >=5 mm
5	[#6] ADC: Focal, Markedly, Hypointense, Non-invasive; DWI: Focal, Markedly, Hyperintense, Non-invasive; Max dim. >=15 mm
4	[#7] ADC: Focal, Markedly, Hypointense, Non-invasive; DWI: Focal, Markedly, Hyperintense, Non-invasive; Max dim. [5, 15) mm
3	[#8] ADC: Focal, Mild/Moderate/Markedly, Hypointense, Non-invasive; DWI: Focal, Mild/Moderate, Hyperintense, Non-invasive; Max dim. >=5 mm
3	[#9] ADC: Focal, Mild/Moderate, Hypointense, Non-invasive; DWI: Focal, Mild/Moderate/Markedly, Hyperintense, Non-invasive; Max dim. >=5 mm
3	[#10] ADC: Focal, Mild/Moderate/Markedly, Hypointense, Non-invasive; Max dim. >=5 mm
3	[#11] DWI: Focal, Mild/Moderate/Markedly, Hypointense, Non-invasive; Max dim. >=5 mm
2	[#12] ADC: Non-focal, Linear/Wedge shaped, Hypointense, Non-invasive; DWI: Non-focal, Linear/Wedge shaped, Hyperintense, Non-invasive; Max dim. >=5 mm
2	[#13] ADC: Non-focal, Linear/Wedge shaped, Hypointense, Non-invasive; Max dim. >=5 mm
2	[#14] DWI: Non-focal, Linear/Wedge shaped, Hyperintense, Non-invasive; Max dim. >=5 mm
X	[#15] The lesion can not be evaluated (no PI-RADS algorithm).
X	[#16] PI-RADS evaluation not avaible: at least one of the ADC / DWI images is unavailable
X	[#17] PI-RADS evaluation not avaible: at least one of the ADC / DWI images is unavailable
X	[#18] Unclassified case

Decomposed rules of the PI-RADS DWI-ADC assessment algorithm

A1.2. DCE PI-RADS decision table

Variable	Description	Related Radlex Term	Possible Values
Input Variables			
lesion_location	Zone	Zone of prostate [RID38890]	PZ, TZ, NOT_AVAILABLE
dce_present_and_adequate	Is DCE present and adequate?	Adequate [RID39308]	YES, NO
dce_abnormality	Does an abnormality appear on the DCE image?	Lesion [RID38780]	YES, NO
dce_enhancement	Enhancement Pattern	Enhancement pattern [RID6058]	POSITIVE_DCE, NEGATIVE_DCE
dce_corresponds_to	Corresponds to finding	MR tissue contrast attribute (Mr procedure attribute) [RID10791]	T2, DWI, NOT_AVAILABLE
dce_bph_features	BPH features on T2	Benign prostatic hyperplasia [RID3784]	YES, NO
Output Variables	Output Variables		
dce_pirads	PI-RADS Evaluation	PI-RADS DCE Lesion Assessment Category [RID50319]	1, 2, 3, 4, 5, X
dce_pirads_description	PI-RADS Rule Descripton	-	<string>

Input and output variables of decision table representing the PI-RADS DCE assessment algorithm

Location	Adequate	Abnormality	Enhancement	Corresponds to	BPH Features	DCE PI-RADS	Description
PZ,TZ	YES	YES	POSITIVE_DCE	T2 DWI,DWI T2	NO	POSITIVE	[#1] Positive Enhancement, Corresponds to finding T2 and DWI, no BPH on T2
PZ,TZ	YES	YES	POSITIVE_DCE	T2	NO	POSITIVE	[#2] Positive Enhancement, Corresponds to finding T2, no BPH on T2
PZ,TZ	YES	YES	POSITIVE_DCE	DWI	NO	POSITIVE	[#3] Positive Enhancement, Corresponds to finding DWI, no BPH on T2
PZ,TZ	YES	YES				NEGATIVE	[#4] Negative Enhancement
PZ,TZ	YES	NO				NEGATIVE	[#5] Negative Enhancement
PZ,TZ	NO					X	[#6] DCE image is unavailable
NOT_AVAILABLE						X	[#7] PI-RADS evaluation not avaible for the selected zone
						X	[#8] Unclassified case

Decision table representing the PI-RADS DCE assessment algorithm

A1.3. OVERALL PI-RADS decision table

Variable	Label	Related Radlex Terms	Possible Values
Input Variables			
lesion_location	Zone	Zone of prostate [RID38890]	PZ, TZ, NOT_AVAILABLE
t2w_pirads_score	T2W score	PI-RADS T2W Lesion Assessment Category: PZ [RID50301], TZ [RID50307]	1, 2, 3, 4, 5, X
dwi_pirads_score	DWI / ADC score	PI-RADS DWI Lesion Assessment Category [RID50313]	1, 2, 3, 4, 5, X
dce_pirads_score	DCE score	PI-RADS DCE Lesion Assessment Category [RID50319]	POSITIVE, NEGATIVE, X
Output Variables			
pirads	PI-RADS Evaluation	PI-RADS Overall Assessment Category [RID50294]	1, 2, 3, 4, 5, X
pirads_description	PI-RADS Rule Description	-	<string>

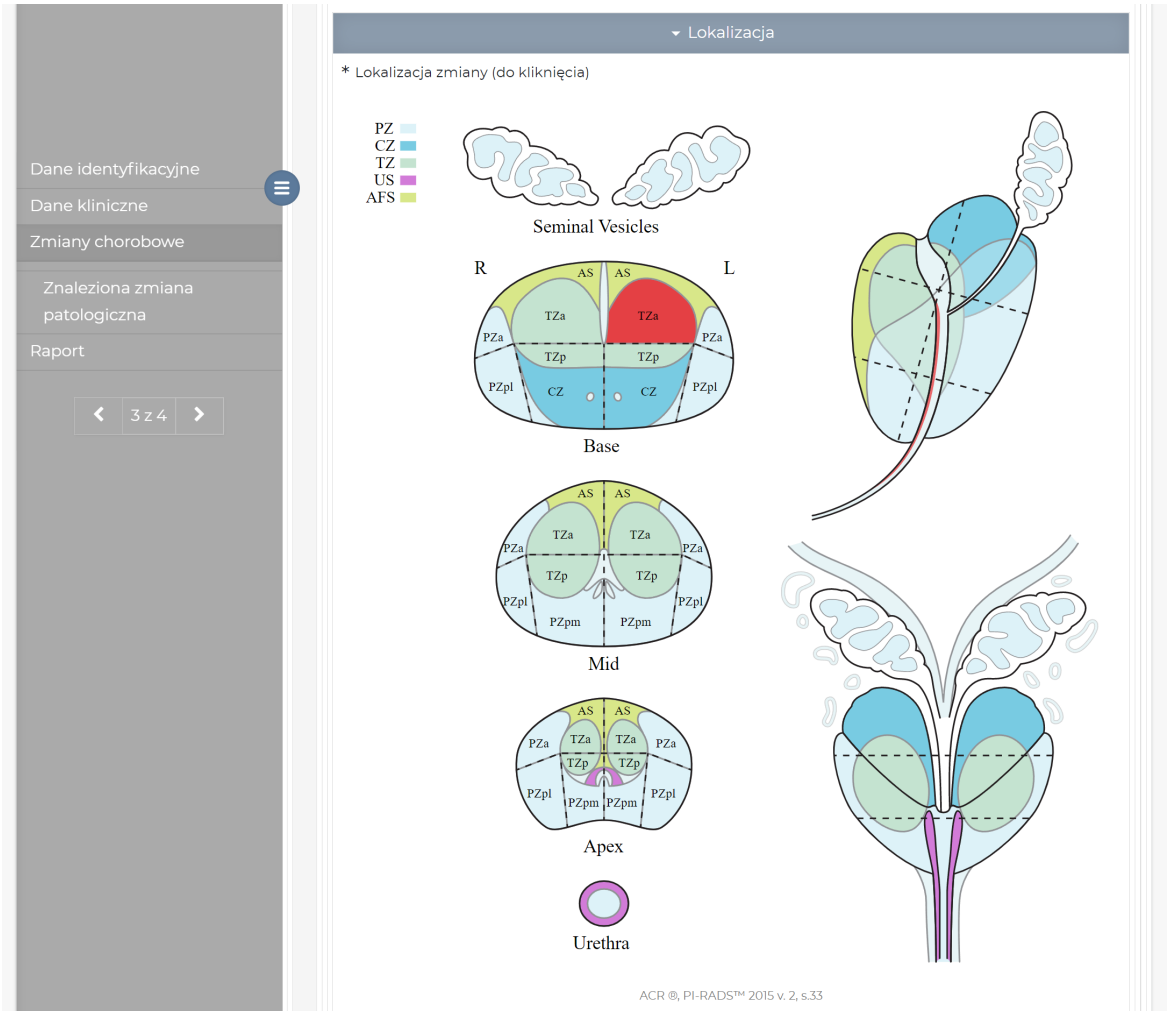
Input and output variables of decision table representing the PI-RADS Overall assessment algorithm

Location	T2W PI-RADS	DWI PI-RADS	DCE PI-RADS	PI-RADS	Description
PZ	1,2,3,4,5	1	POSITIVE,NEGATIVE	1	[#1] PZ; DWI/ADC = 1; T2W = Any; DCE = Any
PZ	1,2,3,4,5	2	POSITIVE,NEGATIVE	2	[#2] PZ; DWI/ADC = 2; T2W = Any; DCE = Any
PZ	1,2,3,4,5	3	NEGATIVE	3	[#3] PZ; DWI/ADC = 3; T2W = Any; DCE -
PZ	1,2,3,4,5	3	POSITIVE	4	[#4] PZ; DWI/ADC = 3; T2W = Any; DCE +
PZ	1,2,3,4,5	4	POSITIVE,NEGATIVE	4	[#5] PZ; DWI/ADC = 4; T2W = Any; DCE = Any
PZ	1,2,3,4,5	5	POSITIVE,NEGATIVE	5	[#6] PZ; DWI/ADC = 5; T2W = Any; DCE = Any
TZ	1	1,2,3,4,5	POSITIVE,NEGATIVE	1	[#7] TZ; DWI/ADC = Any; T2W = 1; DCE = Any
TZ	2	1,2,3	POSITIVE,NEGATIVE	2	[#8] TZ; DWI/ADC = 1-3; T2W = 2; DCE = Any
TZ	2	4,5	POSITIVE,NEGATIVE	3	[#9] TZ; DWI/ADC = 4-5; T2W = 2; DCE = Any
TZ	3	1,2,3,4	POSITIVE,NEGATIVE	3	[#10] TZ; DWI/ADC = 1-4; T2W = 3; DCE = Any
TZ	3	5	POSITIVE,NEGATIVE	4	[#11] TZ; DWI/ADC = 5; T2W = 3; DCE = Any
TZ	4	1,2,3,4,5	POSITIVE,NEGATIVE	4	[#12] TZ; DWI/ADC = Any; T2W= 4; DCE = Any
TZ	5	1,2,3,4,5	POSITIVE,NEGATIVE	5	[#13] TZ; DWI/ADC = Any; T2W= 5; DCE = Any
PZ,TZ	1	X	POSITIVE,NEGATIVE	1	[#14] PZ/TZ; DWI/ADC = X; T2W = 1; DCE = Any
PZ,TZ	2	X	POSITIVE,NEGATIVE	2	[#15] PZ/TZ; DWI/ADC = X; T2W = 2; DCE = Any
PZ,TZ	3	X	NEGATIVE	3	[#16] PZ/TZ; DWI/ADC = X; T2W = 3; DCE -
PZ,TZ	3	X	POSITIVE	4	[#17] PZ/TZ; DWI/ADC = X; T2W = 3; DCE +
PZ,TZ	4	X	POSITIVE,NEGATIVE	4	[#18] PZ/TZ; DWI/ADC = X; T2W = 4; DCE = Any
PZ,TZ	5	X	POSITIVE,NEGATIVE	5	[#19] PZ/TZ; DWI/ADC = X; T2W = 5; DCE = Any
PZ	1,2,3,4,5	1	X	1	[#20] PZ; DWI/ADC=1; T2W = Any; DCE = X
PZ	1,2,3,4,5	2	X	2	[#21] PZ; DWI/ADC=2; T2W = Any; DCE = X
PZ	1,2,3,4,5	3	X	3	[#22] PZ; DWI/ADC=3; T2W = Any; DCE = X
PZ	1,2,3,4,5	4	X	4	[#23] PZ; DWI/ADC=4; T2W = Any; DCE = X
PZ	1,2,3,4,5	5	X	5	[#24] PZ; DWI/ADC=5; T2W = Any; DCE = X
TZ	1	1,2,3,4,5	X	1	[#25] TZ; DWI/ADC = Any; T2W = 1; DCE = X
TZ	2	1,2,3	X	2	[#26] TZ; DWI/ADC = 1-3; T2W = 2; DCE = X
TZ	2	4,5	X	3	[#27] TZ; DWI/ADC = 4-5; T2W = 2; DCE = X
TZ	3	1,2,3,4	X	3	[#28] TZ; DWI/ADC = 1-4; T2W = 3; DCE = X
TZ	3	5	X	4	[#29] TZ; DWI/ADC = 5; T2W = 3; DCE = X
TZ	4	1,2,3,4,5	X	4	[#30] TZ; DWI/ADC = Any; T2W = 4; DCE = X
TZ	5	1,2,3,4,5	X	5	[#31] TZ; DWI/ADC = Any; T2W = 5; DCE = X
PZ,TZ	1,2,3,4,5	X	X	X	[#32] PZ/TZ; DWI/ADC = X; T2W = Any; DCE = X.
PZ,TZ	X	1,2,3,4,5,X	POSITIVE,NEGATIVE,X	X	[#33] PZ/TZ; T2W = X. No PI-RADS algorithm
NOT_AVAILABLE	1,2,3,4,5,X	1,2,3,4,5,X	POSITIVE,NEGATIVE,X	X	[#34] PI-RADS evaluation not avaible for the selected zone
				X	[#35] Unclassified case

Decision table representing the PI-RADS Overall assessment algorithm

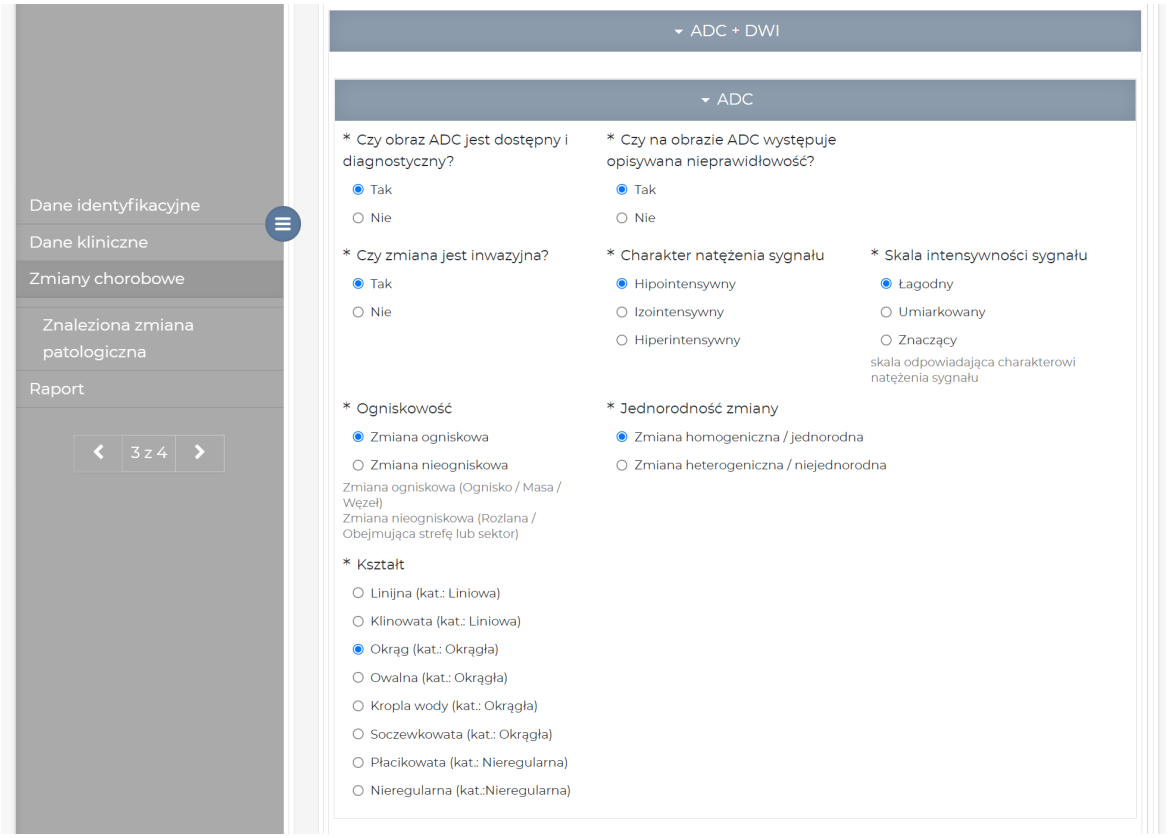
A2. CAR/DS form screenshots

A2.1. Sectoral location subsection



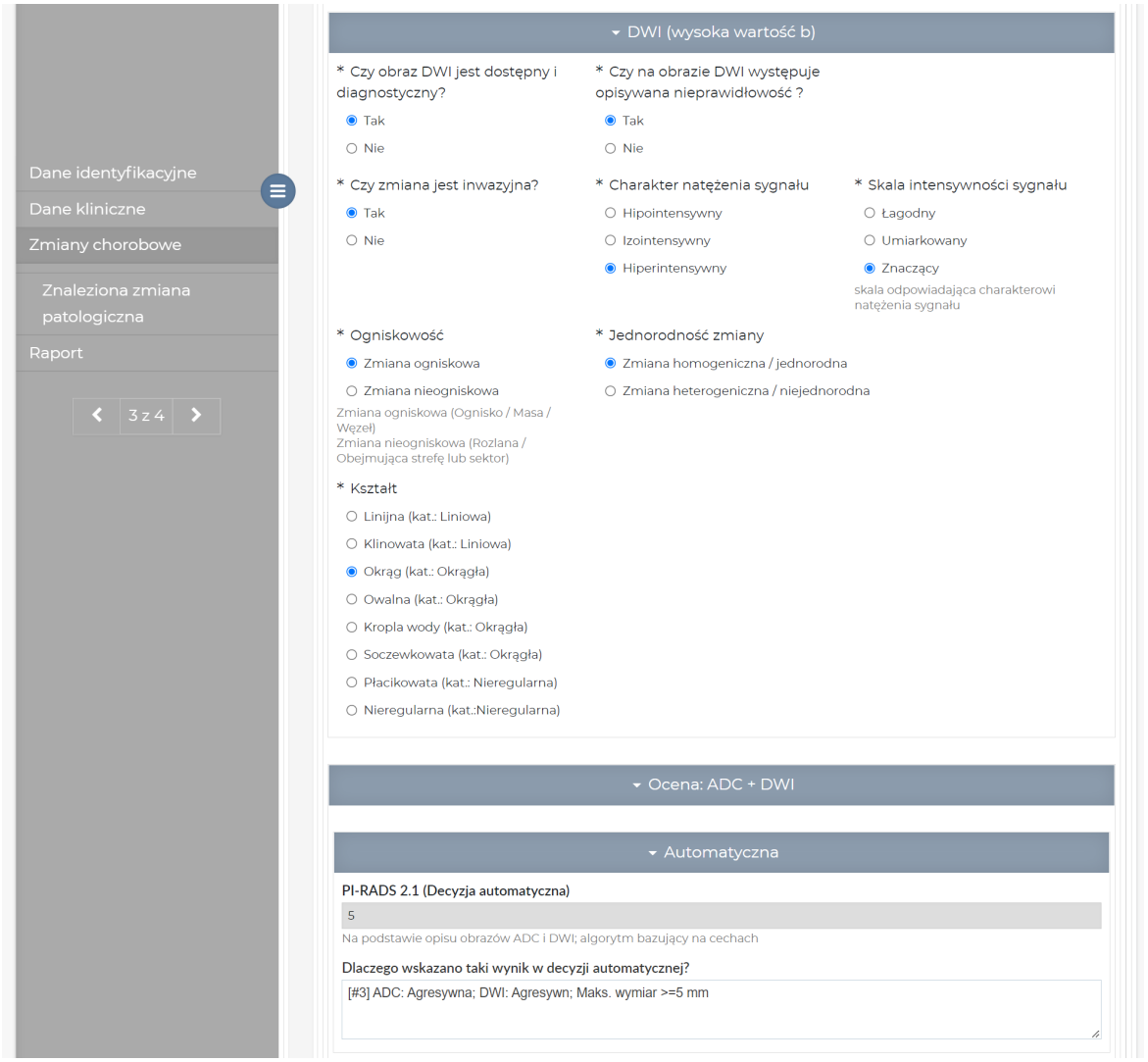
Part of PI-RADS CAR/DS form for specifying lesion location. Multiple sectors can be selected.

A2.2. ADC subsection



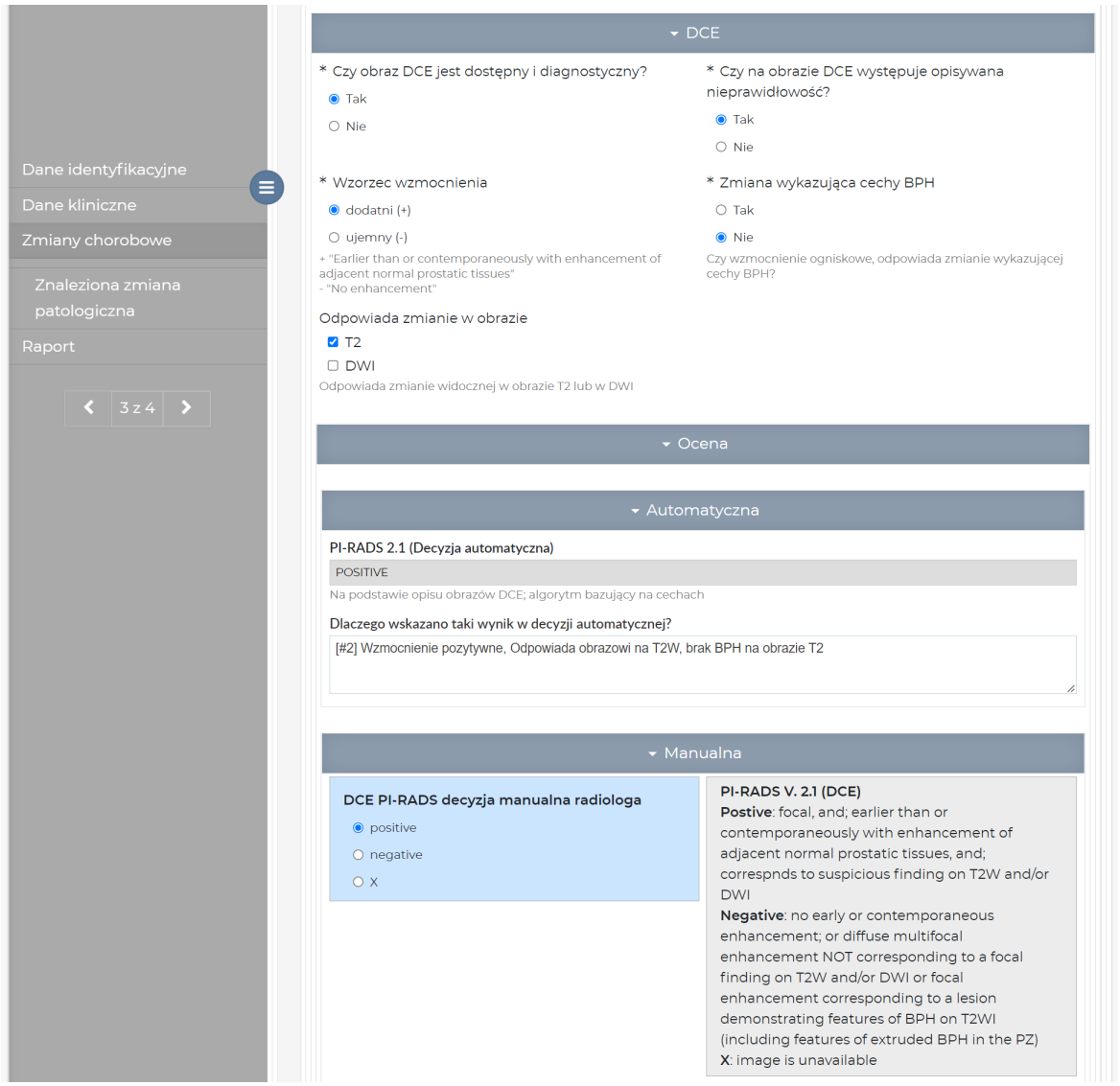
Part of PI-RADS CAR/DS form for assessment of lesion features in ADC images.

A2.3. DWI subsection



Part of PI-RADS CAR/DS form for assessment of lesion features in DWI images. User is presented with automatically estimated category based on the specified DWI and ADC features.

A2.4. DCE subsection



Part of PI-RADS CAR/DS form for assessment of lesion features in DCE images.

A2.5. Final assessment

Part of PI-RADS CAR/DS form for final assessment.