Deep learning for an improved diagnostic pathway of prostate cancer in a small multi-parametric magnetic resonance data regime

by

Alvaro Fernandez-Quilez

Thesis submitted in fulfilment of
the requirements for the degree of
PHILOSOPHIAE DOCTOR
(PhD)

University
of Stavanger

Faculty of Health and Medicine
Department of Quality and Health Technology
2022

University of Stavanger
NO-4036 Stavanger
NORWAY
www.uis.no

ISBN: 978-82-8439-090-1
ISSN: 651
PhD: Thesis UiS No. 1819-1387

Preface

This thesis is submitted as partial fulfilment of the requirements for the degree of *Philosophiae Doctor* at the University of Stavanger, Norway. The author of this thesis has been employed as a research fellow at the Department of Quality and Health Technology, University of Stavanger, in collaboration with the Department of Radiology, Stavanger University Hospital.

The thesis includes a collection of six peer-reviewed and published (or accepted) research articles. The research articles have been reformatted for alignment with the format of the thesis and are included as part of the appendices of the thesis (hence, as chapters of the work).

Alvaro Fernandez-Quilez, March 2022

Abstract

Prostate Cancer (PCa) is the second most commonly diagnosed cancer among men, with an estimated incidence of 1.3 million new cases worldwide in 2018. The current diagnostic pathway of PCa relies on prostate-specific antigen (PSA) levels in serum. Nevertheless, PSA testing comes at the cost of under-detection of malignant lesions and a substantial over-diagnosis of indolent ones, leading to unnecessary invasive testing such biopsies and treatment in indolent PCa lesions.

Magnetic Resonance Imaging (MRI) is a non-invasive technique that has emerged as a valuable tool for PCa detection, staging, early screening, treatment planning and intervention. However, analysis of MRI relies on expertise, can be time-consuming, requires specialized training and in its absence suffers from inter and intra-reader variability and sub-optimal interpretations.

Deep Learning (DL) techniques have the ability to recognize complex patterns in imaging data and are able to automatize certain assessments or tasks while offering a lesser degree of subjectiveness, providing a tool that can help clinicians in their daily tasks. In spite of it, DL success has traditionally relied on the availability of large amounts of labelled data, which are rarely available in the medical field and are costly and hard to obtain due to privacy regulations of patients' data and required specialized training, among others.

This work investigates DL algorithms specially tailored to work in a limited data regime with the final objective of improving the current prostate cancer diagnostic pathway by improving the performance of DL algorithms for PCa MRI applications in a limited data regime scenario.

In particular, this thesis starts by exploring *Generative Adversarial Networks (GAN)* to generate synthetic samples and their effect on tasks such as prostate capsule segmentation and PCa lesion significance classification (triage). Following, we explore the use of *Auto-encoders*

(AEs) to exploit the data imbalance that is usually present in medical imaging datasets. Specifically, we propose a framework based on AEs to detect the presence of prostate lesions (tumours) by uniquely learning from *control* (healthy) data in an outlier detection-like fashion. This thesis also explores more recent DL paradigms that have shown promising results in natural images: *generative and contrastive self-supervised learning (SSL)*. In both cases, we propose specific prostate MRI image manipulations for a PCa lesion classification downstream task and show the improvements offered by the techniques when compared with other initialization methods such as ImageNet pre-training. Finally, we explore *data fusion* techniques in order to leverage different data sources in the form of MRI sequences (*orthogonal views*) acquired by default during patient examinations and that are commonly ignored in DL systems. We show improvements in a PCa lesion significance classification when compared to a single input system (axial view).

Acknowledgments

The time has finally come where I find myself writing the culmination of what it has been my work, passion and arguably, the epicentre of my life this last 3 years. This experience has allowed me to grow both from a professional and from a personal point of view in ways that I would have never even imagined. The balance during my PhD is definitely positive and during these years, I have met incredible people that have made this journey an unforgettable one (*in a positive way*).

I would like to start by extending my gratitude to my supervisor, *Ketil Oppedal* and co-supervisors, *Trygve Eftestøl* and *Thor Ole Gulsrud*. I would have never had this opportunity if you had never relied on me to carry out this work, in the first place. Thanks for the support, long conversations about my progress and trust. Special mention to *Svein Reidar Kjosavik,* whom in spite of not being (officially) part of my supervision team has also been a really important figure for my (personal) development and the realization of my work. Additionally, I am also eternally grateful for the trust and support during the application to continue working on the project in the near future as a postdoctoral researcher. In that regard, I would also like to show my gratitude to the University of Stavanger (UiS) and Stavanger University Hospital (SUS) for the resources and support offered during the PhD and for relying on my work in the form of funding to continue working on the project after the PhD (*HelseVest funding*). In that regard, I would also like to mention *Knut Sommerseth* and *Henriette Thune*, for all the help and long list of e-mails replied during these 3 years.

To all the people involved in SESAM at SUS, as I have also felt like I was officially part of their research group in spite of just being a collaborator. Special mention to *Dag Aarsland*, *Marthe Therese Gjestsen*, *Jon Arild*, *Helen Guthormsen*, *Anne Katrine*, *Martine Kajander, Khadija Khalifa and Solveig Hammonds.* Eternally grateful for the support shown during these years as well as those good times at the sport tournaments! To all the collaborators of the project, thank you. In particular, I would like to extend my deepest gratitude to *Tone F. Bathen* and specially, to *Mattijs Elschot* (NTNU) for your participation in the 50% and 90% seminars and all the useful inputs and contributions. I am looking forward to further collaborations in the future.

To my friends (back in Spain), thanks for the support even when I decided to move far away and start this crazy adventure in Norway almost four years ago. To the amazing friends I have made along this journey: I have no words to express my gratitude for the unconditional support during the good and not so-good times. I feel really grateful I have met you all and I know you will always have my back no matter what. I would have never made it here without your support: *Nicolás, Miguel, Edgar and Maria Camila*. To *Benji and Boni*, this work would have never happened without your barks, nose bites and Dalsnuten hikes.

Last but not least, to *my family*. If something kept me going at times where I felt lost was the possibility of visiting you or those long talks on the phone. I would like to express my deepest love and appreciation to my mother, father and sister as this would have never happened without their unconditional and inexhaustible support. I feel incredible lucky to have a family such as mine and I can proudly say that no matter what, I have always felt supported and loved by them. To the rest of my family: I have always felt the warmth and support from Spain and in times of need, you have always been there. Thanks for everything.

Elena, wherever you are, this is also for you. I still find it hard to accept that from a physical point of view, you are not with us anymore. I wish more fundamental research is carried out in areas like cancer, such that no one else needs to go through certain experiences. You will always be the lighthouse that brightens the night, no matter the amount of darkness.

"Y mientras escribo esto me doy cuenta de que la magnitud de lo que escriba o lo que diga nunca estará a la altura de lo que fuiste y del vacío que nos dejas. No queda más remedio que aprender a caminar sin ti, con nuestra tristeza bebiendo lluvia"

.

List of publications

The main part of the dissertation is made up of the following scientific papers, published in international conferences and journals. The current status of the paper is highlighted in *red* (*Accepted or published*).

Contribution A

Fernandez-Quilez, A., Larsen, S. V., Goodwin, M., Gulsrud, T. O., Kjosavik, S. R., & Oppedal, K. (2021, April). Improving prostate whole gland segmentation in t2-weighted MRI with synthetically generated data. In 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI) (pp. 1915-1919). IEEE. (*Published*)

Contribution B

Fernandez-Quilez, A., Parvez, O., Eftestøl, T., Kjosavik, S.R. & Oppedal, K. (2022). Improving prostate cancer triage with GAN-based synthetically generated prostate ADC MRI. In Medical Imaging 2021: Computer-aided Diagnosis. International society for Optics and Photonics. (*Accepted)*

Contribution C

Fernandez-Quilez, A., Ullah, H., Eftestøl, T., Kjosavik, S.R. & Oppedal, K. (2022). One class to detect them all: Detection and classification of prostate tumors in bi-parametric MRI based on autoencoders. In Medical Imaging 2021: Computer-aided Diagnosis. International society for Optics and Photonics. (*Accepted)*

Contribution D

Fernandez-Quilez, A., Eftestøl, T., Kjosavik, S.R. & Oppedal, K. (2022). Learning to triage by learning to reconstruct: A generative self-supervised approach for prostate cancer based on axial T2w MRI. In Medical Imaging 2021: Computer-aided Diagnosis. International society for Optics and Photonics. (*Accepted*)

Contribution E

Fernandez-Quilez, A., Eftestøl, T., Goodwin, M., Kjosavik, S.R. & Oppedal, K. (2022). Contrasting axial T2w MRI for prostate cancer triage: A self-supervised approach. In 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI) IEEE. (*Accepted*)

Contribution F

Fernandez-Quilez, A., Eftestøl, T., Goodwin, M., Kjosavik, S.R. & Oppedal, K. (2022). Multi-planar T2w MRI for an improved prostate cancer lesion classification. In 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI) IEEE. (*Accepted*)

Glossary

PCa	Prostate Cancer
DRE	Digital rectum examination
PSA	Prostate specific antigen
TRUS	Trans-rectal ultrasound
MRI	Magnetic resonance imaging
mp-MRI	Multi-parametric MRI
T2w	T2-weighted
DW	Diffusion-weighted
DCE	Dynamic contrast enhanced
DL	Deep Learning
TL	Transfer Learning
cS	Clinically significant
ncS	Non-clinically significant
TZ	Transition zone
PZ	Peripheral zone
CZ	Central zone
GS	Gleason score
ROI:	Region of interest
TNM	Tumors, nodes and metastases
CT	Computed tomography
AI	Artificial intelligence
NLP	Natural language processing
CV	Computer Vision
ML	Machine Learning
FFNN	Feed-forward neural network
MLP	Multi-layer perceptron
CNN	Convolutional Neural Network
ResNet	Residual Neural Network
MHA	Multi-head attention
LN	Layer normalization

VAE	Variational auto-encoder
VQ-VAE	Vector-quantized variational auto-encoder
GAN	Generative Adversarial Network
SGD	Stochastic Gradient Descent
TL	Transfer learning
SSL	Self-supervised learning
DA	Data augmentation
NCE	Noise contrastive estimation

Contents

PREFACE III

ABSTRACT V

ACKNOWLEDGMENTS VII

LIST OF PUBLICATIONS X

GLOSSARY XII

1 PROSTATE CANCER 16

1.1 DIAGNOSTIC PATHWAY 16
1.2 MAGNETIC RESONANCE IMAGING (MRI) IN PROSTATE CANCER 19
1.2.1 Multi-parametric MRI (mp-MRI) 20
1.2.2 PI-RADS score 21
1.2.3 Impact of mp-MRI in PCa 24
1.2.4 Radiological workflow in PCa 27

2 ARTIFICIAL INTELLIGENCE IN RADIOLOGY 29

2.1 ARTIFICIAL INTELLIGENCE IN MEDICAL IMAGING 29
2.1.1 Deep learning for medical imaging 30
2.2 APPLICATIONS OF DL IN IMAGING DATA IN PCA 39
2.3 OBSTACLES AND LIMITATIONS OF DL IN MEDICAL IMAGING AND PCA 42

3 DEEP LEARNING IN DATA-LIMITED SCENARIOS 44

3.1 DATA AUGMENTATION 44
3.1.1 Basic image manipulations 45
3.1.2 Synthetic data augmentation 47
3.2 REPRESENTATION LEARNING 48
3.3 EFFICIENT USE OF DIFFERENT DATA SOURCES AND DATA FUSION 50
3.4 TRANSFER LEARNING 51
3.5 SELF-SUPERVISED LEARNING 53
3.5.1 Contrastive SSL 56
3.5.2 Generative SSL 58

4 MATERIALS 60

4.1 DATA AND ETHICS 60
4.2 SOFTWARE 62

5 SUMMARY OF CONTRIBUTIONS 64

5.1 **CONTRIBUTION A:** *IMPROVING PROSTATE WHOLE GLAND SEGMENTATION IN T2-WEIGHTED MRI WITH SYNTHETICALLY GENERATED DATA.* 64

5.2 **CONTRIBUTION B:** *IMPROVING PROSTATE CANCER TRIAGE WITH GAN-BASED SYNTHETICALLY GENERATED PROSTATE ADC MRI.* 67

5.3 **CONTRIBUTION C:** *ONE CLASS TO RULE THEM ALL: DETECTION AND CLASSIFICATION OF PROSTATE TUMOURS PRESENCE IN BI-PARAMETRIC MRI BASED ON AUTO-ENCODERS.* 71

5.4 **CONTRIBUTION D:** *LEARNING TO TRIAGE BY LEARNING TO RECONSTRUCT: A GENERATIVE SELF-SUPERVISED LEARNING APPROACH FOR PROSTATE CANCER BASED ON AXIAL T2W MRI.* 74

5.5 **CONTRIBUTION E:** *CONTRASTING AXIAL T2W MRI FOR PROSTATE CANCER TRIAGE: A SELF-SUPERVISED LEARNING APPROACH.* 77

5.6 **CONTRIBUTION F:** *MULTI-PLANAR T2W MRI FOR AN IMPROVED PROSTATE CANCER LESION CLASSIFICATION.* 79

6 DISCUSSION, FUTURE WORK, WORK IN PROGRESS AND CONCLUSIONS 82

6.1 DISCUSSION 82

6.2 CONCLUSIONS...92

6.3 WORK IN PROGRESS AND FUTURE WORK 93

BIBLIOGRAPHY 94

APPENDICES: ARTICLES 122

CONTRIBUTION A 123

CONTRIBUTION B 136

CONTRIBUTION C 148

CONTRIBUTION D 164

CONTRIBUTION E 176

CONTRIBUTION F 189

1 Prostate Cancer

Prostate Cancer (PCa) is the second most commonly diagnosed cancer [1] with an estimated incidence of 1.3 million new cases among men in 2018 [2], and the fifth leading cause of death worldwide [3]. Furthermore, global trends have shown that PCa incidence is expected to increase during the next years due to aging of the population [4].

Treatments such as chemotherapy and immunotherapy cannot cure PCa once it has spread beyond the prostate gland [5]. Hereby, an early diagnosis and detection of PCa is crucial in order to be able to treat tumours when they are still confined to the prostate [6]. In spite of the urgency and the relevance of an accurate early diagnosis PCa screening in its current form remains as a controversial topic [5,7,8], with no clear benefits from it.

1.1 Diagnostic pathway

Traditionally, the diagnosis of PCa has been based on *digital rectum examination* (DRE). During DRE, the doctor inserts a gloved finger into the rectum and feels the prostate looking for hard, lumpy or abnormal areas (*Figure 1*). DRE relies on the experience of the health personnel performing the procedure and is heavily limited in terms of tumour detection, as some prostate areas are not reachable by the procedure [9]. Moreover, the invasive nature of

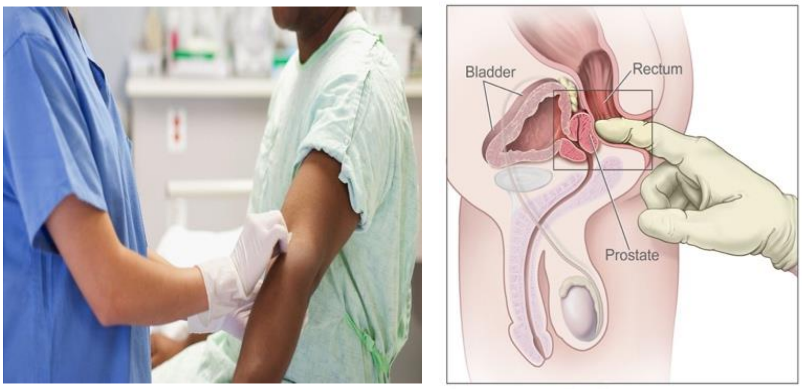


Figure 1. Blood sample extraction for PSA test (left) and digital rectum examination process (right). Image obtained from www.cancer.gov.

DRE results in an uncomfortable situation for the patient, which can lead to side effects such as bleeding [10].

Ever since the measurement of *prostate-specific antigen* (PSA) levels in serum (PSA testing) was approved as a screening test in the early 1990's, it became the main tool for PCa diagnosis and management [11]. Prostate specific antigen made the detection of tumours possible before they become palpable on DRE. Nevertheless, the benefits of PSA as a main test to distinguish between indolent PCa tumours (ncS) and clinically significant tumours (cS) (or in other words, those with potential to become malignant or already malignant [12]) is unclear. Randomized trials have not shown any clear association between a decrease in mortality and PSA as a screening test [13,14,15]. Moreover, PSA testing comes at the cost of substantial under detection of cS tumours and overdiagnosis [16] of indolent tumours, which leads to overtreatment and further unnecessary screening practices [17, 18].

Biopsies. Patients that are under the suspicion of suffering from PCa are referred to a *biopsy*, which is usually the last stage of the current screening practices. Biopsies are commonly used to confirm the stage of the patient and

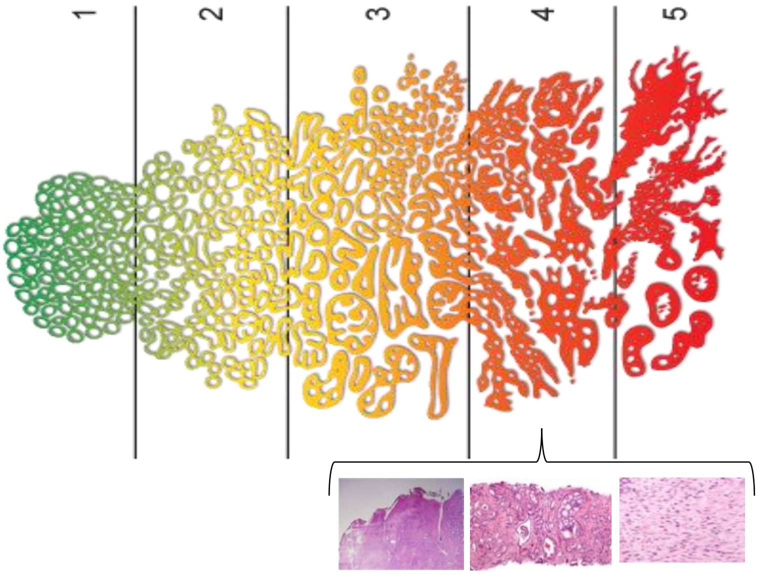


Figure 2. Gleason scoring system for biopsy samples. Source: *www.kreftlex.no*

assess the aggressiveness of the tumour. They aim to collect tissue samples from the prostate by inserting needles through the rectum of the patient [19], to then grade the samples based on the Gleason Score (GS) [19, 20]. In essence, the GS is a scoring system in which a score from 3 to 5 is assigned to the biopsy sample depending on how the cancer cells are arranged in the prostate (*Figure 2*). In order to obtain the final score, the two most prominent areas of tumour growth are determined and scored and then added together (i.e., 4+3 or 3+4). Based on the total sum of the scores, a grade group is assigned in relation to the patient risk (*Table 1*).

Risk Group	***Grade group***	***Gleason score***
Low	Grade group 1	Score <= 6
Intermediate (favourable)	Grade group 2	Score = 7 (3+4)
Intermediate (Unfavourable)	Grade group 3	Score = 7 (4+3)
High	Grade group 4	Score = 8
Very high	Grade group 5	Score = 9 or 10

Table 1. Gleason score risk stratification.

Two types of biopsy procedures can be distinguished: *trans-rectal ultrasound* (TRUS) guided biopsy and *transperine*al biopsy, being the main difference between them the passage used to insert the needle [21]. In the first case, if no external guidance in the form of imaging is used around 10-12 samples are extracted from different areas of the prostate. Some reports have highlighted that even after the repeated sampling, some non-indolent tumours might remain undetected [22, 23, 24, 25]. In addition to it, biopsies can cause pain or discomfort to the patient as well as adverse effects such as infections or rectal bleeding [26].

All in all, the current diagnostic pathway and screening practices of PCa calls for different alternatives able to overcome the current difficulties and to better detect and characterize the potential non-indolent tumours, while reducing the overdiagnosis that populates current practices [27, 28]. In that regard, thanks to recent advances in image acquisition and interpretation, magnetic resonance imaging (MRI) has emerged as a valuable tool for PCa detection, staging, treatment planning and intervention [29, 30]. In particular, multi-parametric MRI (mp-MRI) is already being adopted in clinical routine for PCa management, with positive results [31, 32] (*Figure 3*).

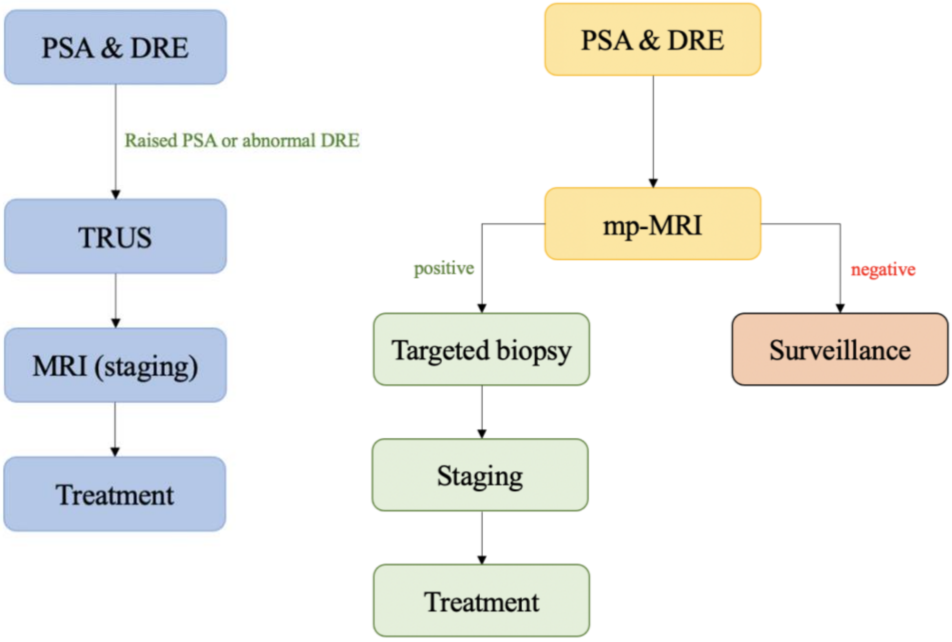


Figure 3. Current PCa diagnostic pathway (lef side) and diagnostic pathway incorporating mp-MRI as a triage test (right side).

1.2 Magnetic Resonance Imaging (MRI) in prostate cancer

Multi-parametric MRI (mp-MRI) is a *non-invasive technique* that can be defined as the combination of several MRI modalities: *T2-weighted* (T2w), *diffusion weighted* (DW), *dynamic contrast enhanced* (DCE) and *spectroscopy* (MRS), if desired [27, 30] (*Figure 4*). In particular, T2w images are acquired preferably in three perpendicular planes: axial, coronal and sagittal, obtaining three different sequences for the modality. Furthermore, *Apparent Diffusion Coefficient Maps* (ADC) are usually automatically computed by a software station when acquiring DW and broadly speaking, provide an average measure of the diffusion of a particular voxel in the image. Although many centres use the aforementioned combination of modalities, there is no standard combination yet, and on-going research is pointing in the direction that bi-parametric MRI (bp-MRI) consisting of T2w and DW sequences (or ADC) is able to match the performance of a mp-MRI approach for the diagnostic and detection of PCa [33].

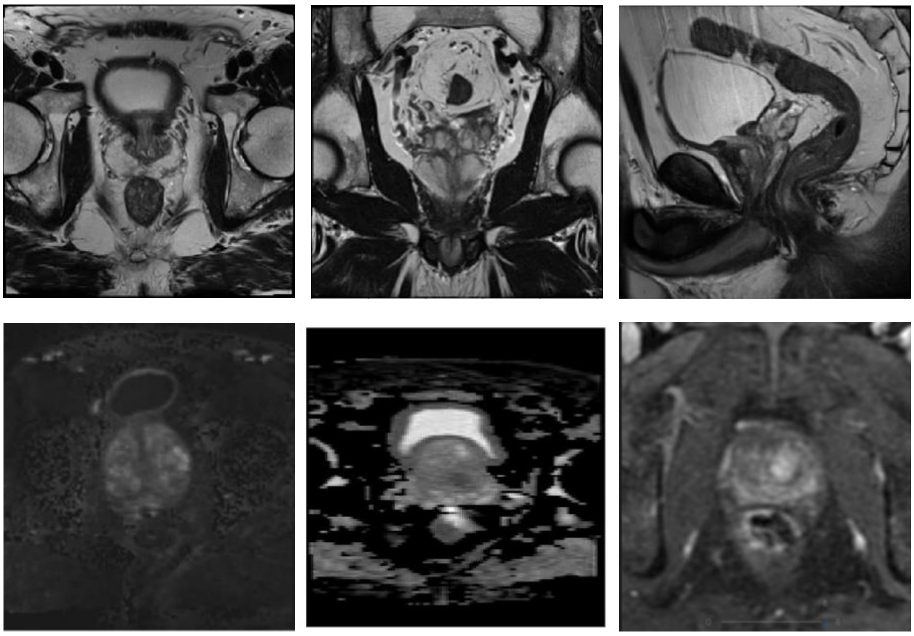


Figure 4. MRI sequences that commonly conform mp-MRI of the prostate. *Top row*: T2w and different acquisition views (from left to right: axial, coronal and sagittal), *bottom row*: DW (b value of 1400) and ADC and DCE.

1.2.1 Multi-parametric MRI (mp-MRI)

T2-weighted (T2w). *T2-weighted MRI* (T2w) shows anatomic-morphologic features of the prostate and morphologic-pathologic structures. Its acquisition in three perpendicular planes (axial, sagittal and coronal) shows the anatomic prostate zonal anatomy and the relation of the prostate to its surrounding structures. T2w allows to differentiate between the high-signal zone of the prostate (peripheral), the mixed-signal zone (transition) and the low-signal zone (central) (*Figure 5*). Furthermore, it allows to anatomically localize lesions and assess their shape, form and size, thanks to its *high inter-plane resolution* [34] (*Figure 4*, top row).

Diffusion-weighted (DWI). *Diffusion-weighted* MRI is the most important functional imaging technique because of its correspondence to histopathological findings which provides an improved evaluation of tissue characteristics and can be a useful tool for detection and staging of PCa in clinical practice [35]. In essence, DW MRI shows the velocity (diffusion) of intracellular water which is restricted for dense cellular tissue – which shows as a low signal (black) on the derived ADC map, whilst low cell density is represented as a high signal (white) on the ADC map [34, 36]. Diffusion-weighted MRI (*Figure 4, bottom row*) can be obtained with different *b values* which measure the degree of diffusion weighting applied, being $b \in [0, 1000]\ s/mm^2$ the recommended values for prostate mp-MRI [37].

Dynamic contrast enhanced (DCE). Dynamic contrast enhanced MRI of the prostate show tissue enhancement (vascularization) after injection of an MR contrast agent. DCE-MRI brings out its potential for the detection of local recurrences (i.e., after radiotherapy or after radical prostatectomy). In the case of patients that have not undergone any treatment, DCE-MRI (*Figure 4, bottom row*) helps to identify potential prostatitis and is of value in findings that might be controversial in the peripheral zone (PZ) [34].

Spectroscopy. MR-spectroscopy is sometimes included in prostate mp-MRI protocols depending upon 3rd party rules (such as hospitals or regions). Its value comes out when assessing the malignancy risk of a region of interest (ROI). Its utilization is usually reserved to research purposes as the process to obtain the image and analyse it is rather complex and time-intensive [27].

1.2.2 PI-RADS score

With the introduction of mp-MRI in the PCa diagnostic pathway and management, a standardized methodology and terminology to translate findings in mp-MRI into clinical practice in an unequivocal way was required. The

Prostate Imaging-Reporting and Data System (PI-RADS) was developed in 2013 in an effort to standardize mp-MRI evaluation of prostate MRI [38] and later updated in 2015 (PI-RADS v2) and 2019 (PI-RADS v2.1) [39]. The latest update introduced a sectoral map (*Figure 5*) for the prostate, redefined the scoring system aiming overcome conceptual confusion and differences present in the first PI-RAD scoring system and finally, relegated DCE to a minor classification role secondary to T2w and DW sequences [27, 40].

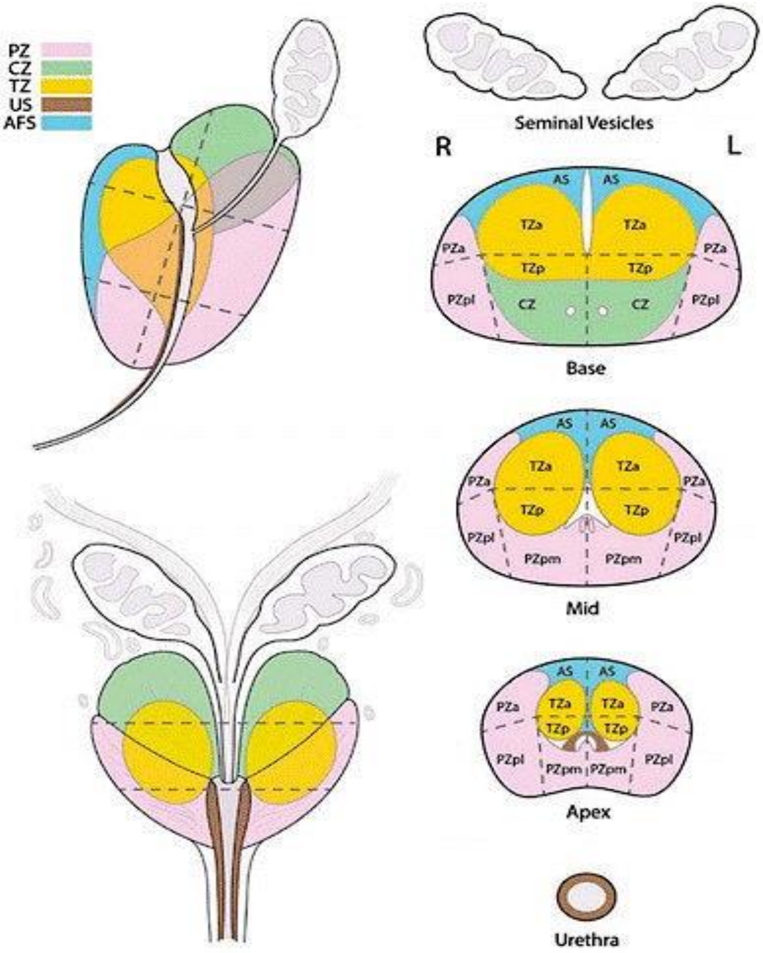


Figure 5. Sectoral map of the prostate according to PI-RADS v2. PZ: peripheral zone; CZ: central zone; TZ: Transition zone; US: urethral stroma; AFS: anterior fascial stroma. From American College of Radiology. MR Prostate Imaging Reporting and Data System version 2.0. http://www.acr.org/Quality-Safety/Resources/PIRADS/

PI-RADS v2 scoring system can be summarized as follows (*Figure 6*): The images of different sequences are obtained and a score ranging from 1-5 is given, depending on several criteria such as homogeneity and encapsulation of the detected lesion. The sequence of reference (and their contribution to the overall evaluation) depends on whether the lesion is located in the transition or peripheral zone of the prostate (*Figure 5*). Finally, the case is assigned one of the 5 assessment categories (*Table 2*) [39].

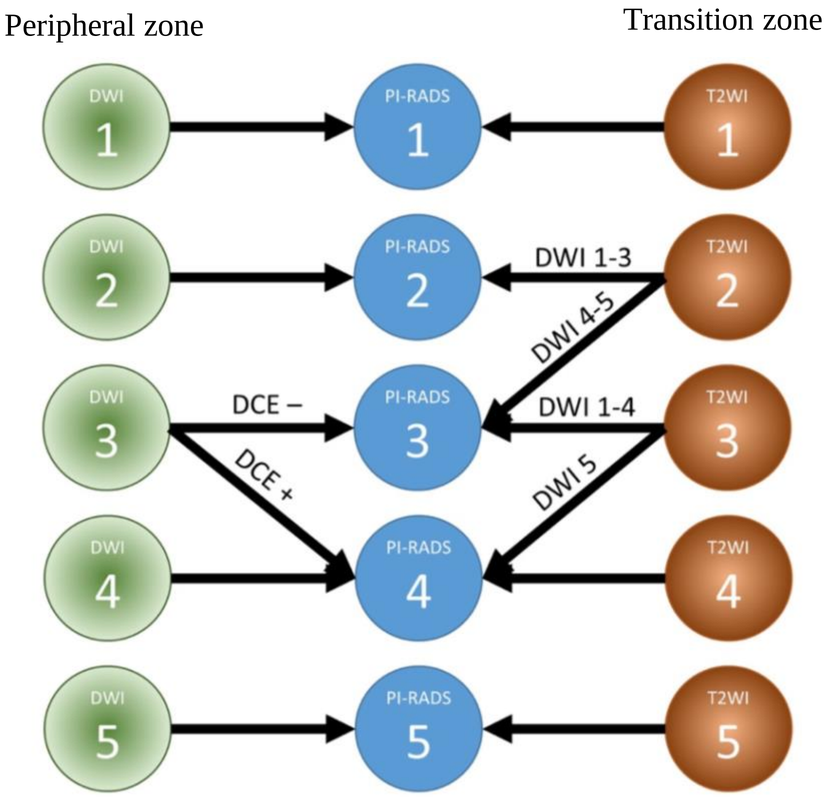


Figure 6. PI-RADS v2 flowchart to assign a category and grade the case. Case courtesy of Dr Francis Deng, Radiopaedia.org, rID: 70893

PI-RADS score	***T2w, DW and DCE score***	***Definition***
1	3-4	Most probably benign
2	5-6	Probably benign
3	7-9	Indeterminate
4	10-12	Probably malignant
5	13-15	Most probably malignant

Table 2. Assessment categories of PI-RADS scoring system.

1.2.3 Impact of mp-MRI in PCa

A major goal for PCa is *more accurate disease characterization* through the synthesis of anatomic (T2w), functional (DW) and molecular imaging information [41]. Arguably, such a characterization would improve the current diagnostic pathway by providing a tool that allows for a better patient management strategies and stratification in those who require an *active surveillance strategy* ("watchful waiting") or those who require immediate action in the form of further testing and treatment. Such is the interest in mp-MRI and its potential to improve the current practices, that its integration in the current diagnostic pathway is already gaining ground in different areas with positive outcomes for both the diagnosis and management of PCa [42]:

Triage test for men at risk. There is uncertainty and controversy surrounding PSA testing as a screening test and the attribution of grade D by the U.S. Preventive Services Task Force against PSA screening [13, 42] (moderate or high uncertainty that the service has no net benefit or that PSA is not fit for the purpose [43]). Introducing imaging techniques in the diagnostic pathway as a support for PSA testing for those men with elevated levels of serum PSA (and thus, at risk of suffering PCa) and before TRUS guided biopsies could address the problem of overdiagnosis of PSA since mp-MRI has been found to have reduced sensitivity for low GS grade tumours, and might systematically overlook ncS lesions. In addition, evidence is starting to accumulate reporting a high negative predictive value when it comes to ruling out cS lesions [44, 45] and similar approaches have already been successful when treating other solid organ cancers [46].

Disease characterization. Measures such as the prostate gland volume *in vivo* are commonly required in the management of prostate disorders, both benign and malignant. Knowledge of total prostatic volume is necessary in the calculation of PSA density (PSAD), a key indicator that elevated PSA is due to malignancy [47]. Fields such as *radiomics* [48] are rapidly evolving. Radiomics involves the extraction of quantitative features from images, that could, potentially, characterize the disease under consideration whilst enabling a more advanced understanding of it. Specifically, a high-quality delineation of the target area or region of interest (ROI) -such as the tumour or prostate gland- is the premise to ensure that the subsequent feature extraction is performed with acceptable quality. Imaging techniques such as T2w can delineate the normal prostate zonal anatomy, clearly showing the transition and peripheral zones [34] and hence playing a crucial role in the characterization of the disease.

Lesion localization and focal therapy. Standard biopsy techniques suffer from inadequacy of sampling. Such is the inadequacy that approximately one-third of patients undergoing active surveillance see an upgrade of the disease when undergoing TRUS guided biopsies [49]. Mp-MRI provides an alternative that can be used to detect, localize and characterize tumours as well as to track their progression and the pathological changes of the patient associated to it over time. There is already evidence that mp-MRI can act as an accurate monitoring tool for PCa progression in men undergoing an active surveillance program [50]. Additionally, an accurate detection and cancer localization might also help overcoming secondary effects in certain PCa practices such as radical prostatectomy [51] and to improve treatments such as focal therapy [52].

Initial staging and active surveillance. Tumours, nodes and metastases (TNM) is the reference standard for staging PCa [41], that has as a primary goal to define the anatomic extent of the tumours and to distinguish patients with organ-confined, locally invasive or metastatic disease. Staging contains 4 main subcategories (T1-T4) which are mainly based on a combination of findings obtained via palpability and after assessment of resected glandular tissue. Detecting extracapsular extension and locating the intraprostatic extent of the disease are important issues in the management of the disease and in the staging phase of it. Due to the difficulties in providing an accurate TNM staging, older men and men with significant health problems were traditionally diagnosed with stage A1 PCa and considered for an active surveillance program. Thanks to improved tumour localization and lymph node staging, a more optimal and tailored TNM assessment can be achieved along with an improved active surveillance program [53]. In that regard, contemporary active surveillance programs include low-risk patients with low tumour volumes, determined through imaging techniques.

Guided biopsies. To increase biopsy sensitivity and reduce the number of core biopsies required to detect cS PCa lesions, several technologies have been explored along with ultrasound [54]. Nevertheless, the ability of the explored techniques to discriminate benign from malignant tissue is low [55], and thus its application in guiding biopsies is compromised. Conventional MRI provides higher spatial and contrast resolution than ultrasound or computed tomography (CT), showing potential to be a suitable option to be used to guide prostate biopsies [56].

PCa management (recurrence of the disease). Once a patient has undergone radical prostatectomy or radiation therapy, a rise in PSA is commonly an indication of cancer recurrence. When a rise in PSA levels is observed in patients after radical prostatectomy or radiation therapy the next step is usually to determine whether cancer recurs locally or in distant organs. An accurate localization and determination of the extent of cancer is critical in selecting an appropriate treatment (local salvage therapy or systematic therapy). Hence, the primary role of MRI imaging in this kind of settings is to help distinguish local recurrence from distant metastatic diseases [43].

"Triage" test for men with confirmed lesions and test for negative first biopsies. One of the main reasons of why TRUS biopsies usually fail to sample the right location is because of their "blind" nature to the cancer location within the prostate. Specifically, cancers in the anterior prostate, apex and midline are either under-sampled (or never sampled) resulting in cS cancers going undetected [57]. Imaging can be used to assess the risk status of men with a previous negative biopsy and perform a follow-up biopsy that can be targeted to visible MRI lesions. Evidence has shown that when this strategy was adopted, 2/3 of men with 2 or more previous negative TRUS biopsies were diagnosed with cancer [58]. In addition, providing a timely treatment, further testing or active surveillance program is of crucial relevance for patient management. Imaging provides a way to perform targeted biopsies and determine the significance of the lesion under consideration, allowing for tailored feedback to the patient and improving the quality of life and outcomes of the disease by “triaging” the lesions based on their significance (GS) [59].

Therapy response & drug development. The role of imaging is not limited to delineating and localizing organs and structures but to detect at an early-stage changes occurring in tissues, enabling a tailored patient management including changes in real time and facilitating drug development. Specifically, data has already shown that DWI is able to show in a quantitative way the response of PCa bone metastases to treatment [60, 61].

1.2.4 Radiological workflow in PCa

Ever since the guidelines for PCa diagnostic and management were updated by the European Association of Urology (EAU) and the American College of Radiology (ACR), prostate MRI has been advised to be taken before a biopsy, instead of being relegated to a secondary role after undergoing a biopsy. But *what exactly does an MRI exam in PCa entail for the specialist in charge of carrying it out?* A patient exam takes in average, from 20 to 45 minutes during which the sequences conforming the mp-MRI (Section 1.1.2) are acquired [62].

Usually, after the image acquisition, a radiologist evaluates the obtained scans by performing anatomical measurements of the prostate (dimensions and volume) followed by calculations of PSAD [47]. Report and acquisition of the anatomical measurements used to include a manual delineation process of anatomic structures such as the gland of the prostate [63, 64], which was subject to *high inter-reader variability* and was a *time-intensive task*. Current approaches include semi-automatic tools that aid with the delineation of the ROIs. In spite of it, human interaction is still expected to some degree to provide points of interest (starting points to begin the delineation), guiding points or review the final results obtained by the tool (*Figure 7*).

Following, the radiologist assesses the MRI scans using a hanging protocol which usually includes T2w, DWI and ADC maps (Section 1.1.2). During the assessment, different zones of the prostate (Figure 3) are taken into consideration and DWI and ADC sequences lead if the peripheral zone is being assessed whilst T2w is the leading one if the transition one is under analysis and used as an additional input in case of doubt when analysing the peripheral zone. After assessment, all the derived information is put together to determine the PI-RADS score (Section 1.1.2) and the radiologist is in charge of creating a report to communicate the findings to the urologist and, if relevant, to be discussed with the rest of the team [65]. Assessment of the sequences and subsequent PI-RADS score assignment has been shown to be a time-intensive task which is subject to *high inter-reader variability* and the *amount of experience* of the radiologist, which can have a negative impact and consequences for the patient [66].

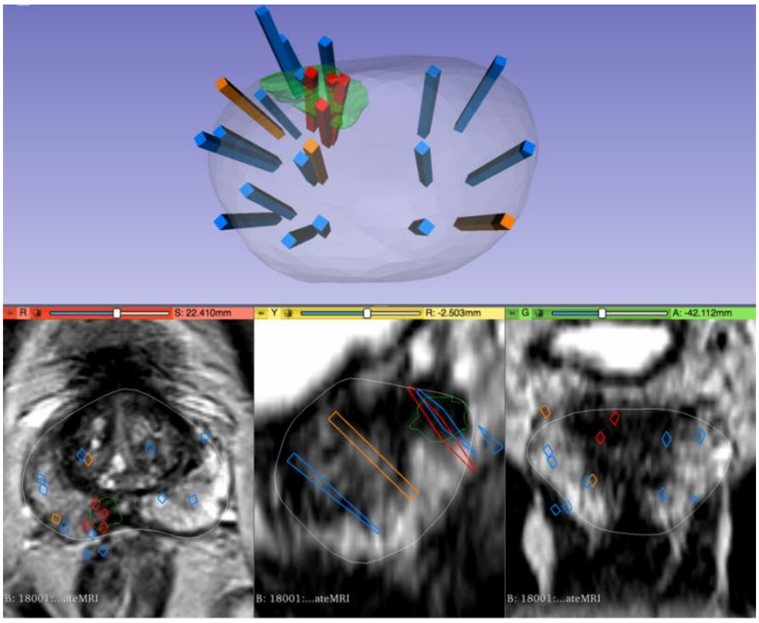


Figure 7. Example of a semi-automatic tool to delineate anatomical ROIs and to guide biopsies in prostate MRIs. Image obtained and reprinted from https://wiki.cancerimagingarchive.net/.

2 Artificial intelligence in radiology

The convergence of complex data (such as imaging) with artificial intelligence (AI) is leading to major advances in applications that range from self-driving vehicles to natural language processing (NLP) and computer vision (CV). The ability to better represent and interpret such complex data has allowed machines to automatize tasks that have, traditionally, been carried out by humans [67]. AI is becoming a major constituent of many applications within healthcare, including drug discovery, medical diagnostics and imaging, risk management, wearables, virtual assistants, virtual reality and patient monitoring, among others [68, 69]. Medical fields such as radiology, which rely on imaging data, are already seeing benefits from the implementation of AI methods [70, 71, 72].

Within radiology, physicians require specialized training to assess and analyse medical images and report findings to detect, characterize and monitor diseases. Such an assessment is often based on experience (along with many years of specialized training and education) and can be, at times, subjective. On the other hand, AI algorithms have the ability to recognize complex patterns in imaging data and are able to automatize certain assessments or tasks while offering a lesser degree of subjectiveness (subject to the ground truth it was trained on), as opposed to human-based assessment. Furthermore, with the proper deployment and when the right actions are in place, AI can also benefit the reproducibility of the results when integrated into the clinical workflow as a tool to assist physicians [67].

2.1 Artificial intelligence in medical imaging

One of the main driving factors behind the growth of AI in the medical imaging domain has been the search for greater efficacy and efficiency in clinical care. The disproportionate rate at which radiological data keeps growing coupled with an increasing lack of availability in specialized readers [73, 74], has forced health-care providers to dramatically increase radiologists' workload [75]. Such is the increase that in some cases, a radiologist must interpret one image every 3-4 seconds in an 8-hour workday to meet work demands [76]. As the workload demands increases it is inevitably that errors in the assessment arise, especially in a field like radiology where visual perception and decision making under uncertainty are particularly relevant [77].

Integrating AI within the imaging workflow of radiologists can increase efficiency, reduce errors and achieve the proposed objectives while reducing the manual input of the radiologist thanks to pre-screened images and identified features [78]. Furthermore, AI could aid with the increasing workload in the field due to the shortage of specialists. Hereby, a substantial effort is being made and policies are being put forward to facilitate the transition to a scenario in which AI helps and supports radiologists to carry out their duties.

2.1.1 Deep learning for medical imaging

We can mainly differentiate between two types of AI approaches that are widely used nowadays for radiology (*Figure 8*): *traditional machine learning (ML)* and *deep learning (DL)*, respectively. The first one aims to extract handcrafted features that are defined from a mathematical point of view (such as image texture) and can be quantified in an automatic or semi-automatic way by computer software [79] and that is usually followed by a feature selection step and a ML-based algorithm [80]. Although the extracted features are perceived to be discriminative for the tasks under consideration, they commonly rely on expert definition and hence, are subject to the limitations of their knowledge. Hereby, those features might not necessarily represent the most optimal feature quantification approach for the task at hand. Furthermore, features are usually “static” -specific to their imaging modality- and unable to adapt nor have the same success and impact with other imaging modalities with different signal-to-noise characteristics [67].

DL methods can automatically learn feature representations while suppressing the need for a human-expert intervention. Thanks to their data-driven approach, more general and informative features can be extracted. Additionally, DL gets rid of manual steps such as the definition of a ROI, which requires manual delineation by experts of the diseased tissues [81]. Given the right amount of data, DL is also often robust to undesired variations such as the inter-reader variability present among experts. Algorithms based on DL have seen an unprecedented success in different healthcare applications with a continuously-growing amount of software and products available for healthcare and, in particular, radiology, getting approval by the U.S Food & drugs administration (FDA) [82]. One could say that in some ways, DL is able to follow a similar process compared to the one radiologist’s follow, as opposed to traditional ML. That is, DL can identify parameters and features on the fly and assess their relevance on the basis of other factors to arrive at a clinical decision. When comparing DL models (deep models) with their ML-based counterparts, several studies have reported substantial improvements with DL methods [83, 84, 85]. Additionally, DL also has the benefit to have a faster development time as it only depends on curated data rather than domain expertise to extract useful features. ML methods have also reached a plateau in performance over the last years and generally speaking, they usually do not meet the minimum requirements for clinical utility and routine, resulting in only a few of the proposed systems to be translated into the clinic [86].

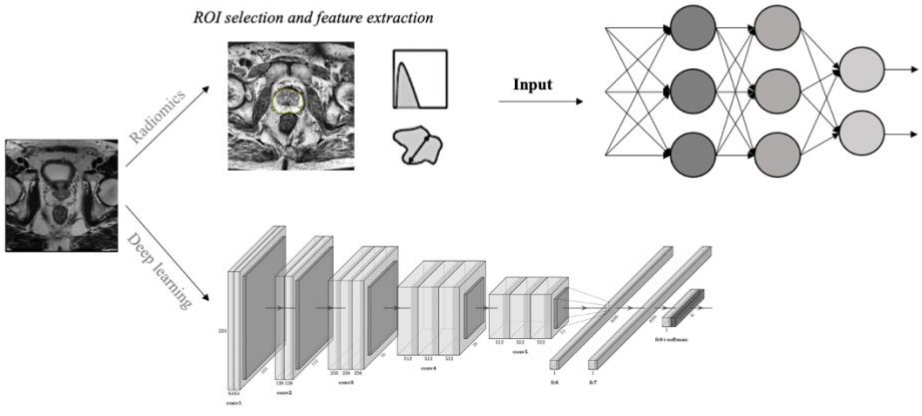


Figure 8. Radiomics vs DL approach for prostate MRI.

Deep models

The ***perceptron*** is the earliest trainable feed-forward neural network (FFNN) [87] with a single-layer architecture composed by an input layer and an output one, inspired by the structural elegance of the neural system. More complex architectures such as ***multi-layer perceptron*** (MLP) include a stack of layers composed by inputs, hidden layers and output layers. It is important to emphasize that in MLP the units (neurons) of neighbouring layers are fully

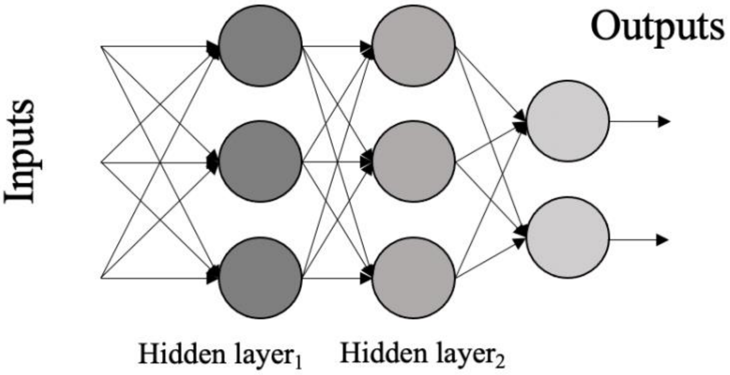


Figure 9. Multi-layer perceptron with two hidden layers.

connected to one another, but there are no connections among units in the same layer (*Figure 9*). In essence, each neuron performs three tasks: multiply each input with the respective weights, sum the resulting values of the previous step and apply an (non-linear) activation function to the result of the sum [88]. As it turns out, non-linear activation functions give us the power to represent arbitrary functions under certain technical conditions, even for a shallow MLP (i.e., with one single hidden layer). Hereby, they are regarded as universal approximators [89, 90]. Assuming a MLP with two hidden layers, we could represent the operations in a *vectorized form* as follows (*Equation 1*):

$$\begin{aligned} \mathbf{h}^{(1)} &= \phi^{(1)}\left(\sum_j \mathbf{W}^{(1)}\mathbf{x} + \mathbf{b}^{(1)}\right) \\ \mathbf{h}^{(2)} &= \phi^{(2)}\left(\sum_j \mathbf{W}^{(2)}\mathbf{h}^{(1)} + \mathbf{b}^{(2)}\right) \\ \mathbf{y} &= \phi^{(3)}\left(\sum_j \mathbf{W}^{(3)}\mathbf{h}^{(2)} + \mathbf{b}^{(3)}\right) \end{aligned} \qquad \textbf{(1)}$$

where $\mathbf{h}^{(l)}$ represents the activations of all the units of the layer *l*, each layer's weights are represented by a weight matrix $\mathbf{W}^{(l)}$, the bias vector of each layer is represented by $\mathbf{b}^{(l)}$ and the activation function -assuming the most generic case, that is, different ones- is represented by $\phi^{(l)}$. Fully-connected MLP are not optimal for the type of data highlighted in this thesis, that is, images. Since the resolution of an MRI can be of hundreds of pixels for each direction (for instance, 320x320), the use of MLP becomes impractical as the number of connections becomes extremely large and thus, the computational power required becomes exceedingly large too.

Convolutional Neural Networks (CNNs). Convolutional neural networks (CNNs) [91] (Figure 8, bottom architecture) are conceived to better utilize spatial information from neighboring pixels -or voxels, if used in 3D- by taking the full picture as an input, as opposed to traditional ML methods where vectorized features are used. Such a feat is accomplished by using *convolutional layers* which encourage *weight sharing, local receptive fields* and *spatial sub-sampling*. Thanks to those characteristics, CNNs have the benefit of being *invariant to affine transformations* of images, allowing them to recognize patterns that are shifted or tilted within images.

A typical CNN is composed by other layers besides convolutional ones, containing the classic CNN structure the following elements: multiple convolutional layers, non-linear activation functions and pooling layers [91]. In particular, VGG16 is a widely used CNN model based entirely on the previously defined layer structure [92]. Another remarkable CNN architecture is Residual Networks (ResNet) [93]. They follow a structure similar to VGG16 with the addition of residual connections -otherwise called skip connections- and batch normalization [94], which enabled to train in an efficient way deeper network architectures without falling into previous training pitfalls. CNNs have seen success in a variety of medical imaging tasks, such as classification [95, 96] and detection [97, 98].

Transformers. Transformers [99] are a sequence-to-sequence prediction architecture that has exhibited an outstanding performance in tasks such as natural language processing (NLP) [100, 101]. In particular, Transformers were designed to overcome the limitations in modeling explicit long-range relations of CNNs -due to their limited receptive field of convolution layers- and capture relations between arbitrary positions in the *input sequence* [99]. By using an *entire sequence,* in the form of *image patches* and relying on *self-attention* [99] the architecture is able to completely dispose of convolutions and model long-range dependencies in the image (or text).

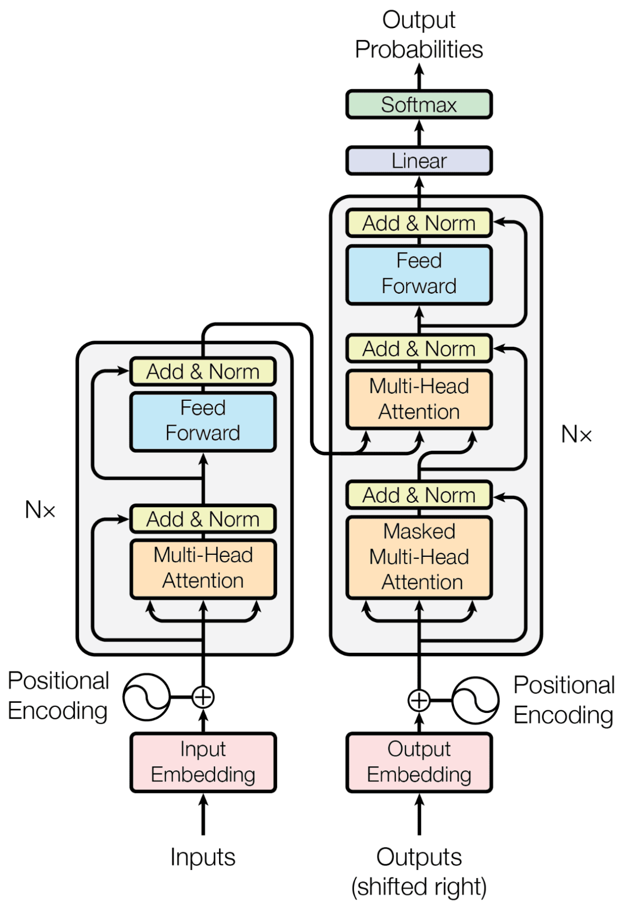


Figure 10. Transformer architecture. Figure reprinted with permission from [99].

The key elements in a Transformer architecture are the image representation as a sequence of patches, learnable positional embeddings, multi-head attention (MHA) mechanism and layer normalization (LN) [102] (*Figure 10*). In the first case, the concept of patch refers to p local areas of pixels (*Figure 11*) $\mathrm{x}_1, \mathrm{x}_2 \ldots \mathrm{x}_p \in \mathbb{R}^{H\mathrm{x}W\mathrm{x}C}$, where H is the height of the patch, W is the width, C represents the number of channels and $p = \left\lfloor \frac{H}{h} \mathrm{x} \frac{W}{w} \right\rfloor$, with h and w representing the height and width of the original image that are commonly obtained without overlap. The learnable positional embeddings aim to capture the order relationships between the low-dimensional p patches (spatial information).

MHA is arguably the core of the Transformer architecture. In essence, runs the inputs through the self-attention mechanism several times in which each time the *Key* ($\mathbf{K} \in \mathbb{R}^{mxd}$), *Query* ($\mathbf{Q} \in \mathbb{R}^{nxd}$) and *Value* ($\mathbf{V} \in \mathbb{R}^{mxf}$) matrices are mapped into different lower dimensional spaces and the attention is computed (commonly) with a scaled dot-product attention (*Equation 2*):

$$\text{attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{QK}^{\text{t}}}{\sqrt{d}}\right)\mathbf{V} \qquad \textbf{(2)}$$

Finally, LN [102] computes the mean μ_n and variance σ_n^2 across channels and spatial dimensions (*Equation 3*). LN yields to better performance than other normalization approaches such as batch normalization (BN) [94], thanks to the control achieved in the gradient computation [103].

$$\text{LN}(x_n, \mu_n, \sigma_n^2) = \gamma\left(\frac{x_n - \mu_n}{\sqrt{\sigma_n^2 + \epsilon}}\right) + \beta \qquad \textbf{(3)}$$

where γ and β are learnable parameters. Remarkable examples of specific Transformer architectures are ViT [104] and 3D ViT [105], which have seen success in a variety of medical imaging-based tasks such as classification and segmentation [106, 107].

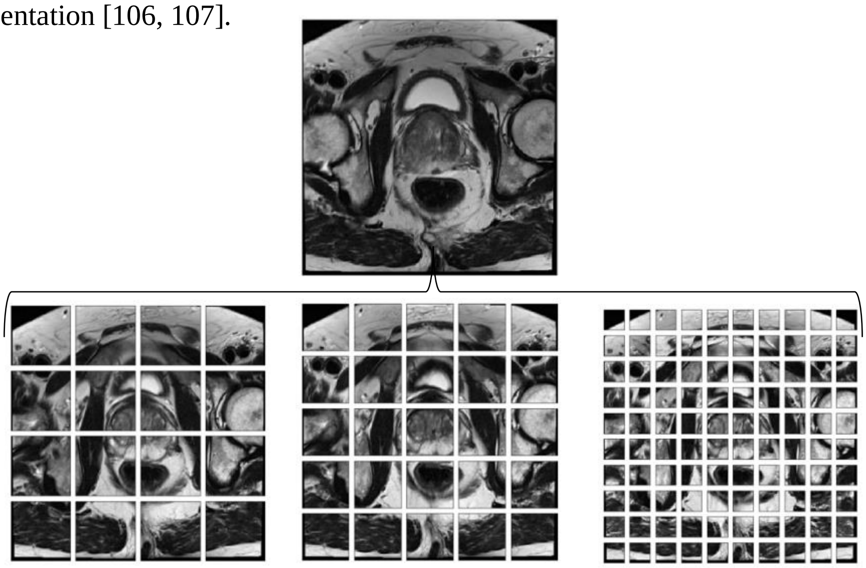


Figure 11. Patch extraction process. From left to right: 80x80, 64x64 and 32x32.

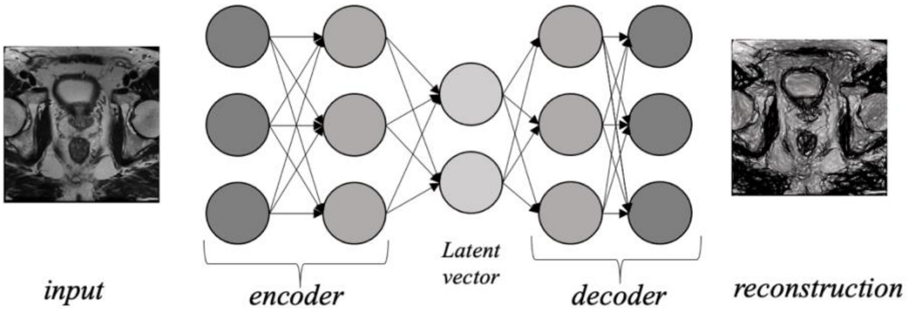


Figure 12. Example of an auto-encoder architecture.

Auto-encoders (AEs). Auto-encoders (AEs) (*Figure 12*) are a type of neural network where the output layer has the same dimensionality as the input layer. An auto-encoder looks to replicate the data from the input to the output in an *unsupervised manner* (defined in the next sections) and is therefore, sometimes, it is referred as a replicator [108]. One of the most important characteristics of AEs is their ability to learn or discover highly non-linear and complex patterns, such as relations between the input values assuming that there is some sort of structure in the data. In particular, AEs look to *project to a lower-dimension space* $\mathbb{R}^d$ the original input $\mathrm{x}_i \in \mathbb{R}^D$ where $d << D$ and obtain a reconstruction of the original input from the compressed version of the input. Such dimensionality reduction process is stored in the *bottleneck* component of AEs, which serves as a "bridge" between the two other main components of AEs: *encoder* and *decoder*. The first component, the *encoder*, can take the form of a deep neural network or its variants (i.e., FFNN, CNN or Transformer) and aims to *compress the input into a latent space representation*, obtaining a lower-dimension representation. On the other hand, the *decoder*, commonly "mirrors" (i.e., if the encoder was a CNN the decoder will follow the same structure) the structure of the encoder and is responsible for reconstructing the input back to the original dimensions from the reduced representation obtained by the encoder [109].

Some of the most iconic types of AEs are *Convolutional Auto-Encoders* (cAE) [110] and *Variational Auto-Encoders (*VAE) [111], which have been the predecessors of more advanced models such as *Quantized-Variational Auto-Encoder* (VQ-VAE) [112]. cAE are based on encoder-decoder structures that exploit convolutional layers, such that a more optimal encoding is learnt for images. VAEs are conceived to learn the probability distribution $p(x_i)$ of the input image x_i instead of learning the function $f(\cdot)$ that maps the input and the output. Thanks to that approach, VAEs are able to *generate* new images after successfully approximating $p(x_i)$. Some successful applications of AEs for medical imaging (based on their performance when compared with other methods) include detection and classification tasks [113, 114].

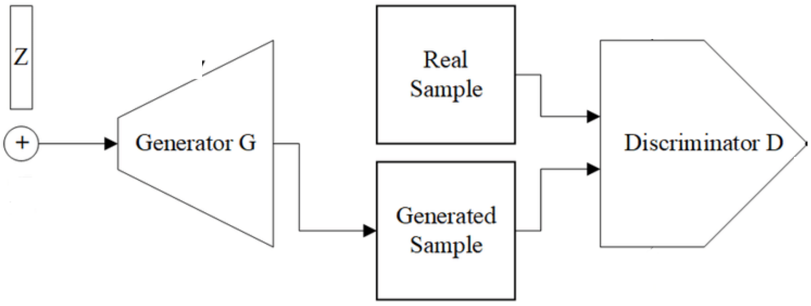


Figure 13. Example of a GAN architecture.

Generative Adversarial Networks (GAN). Generative Adversarial Networks (GAN) are party of the so-called data generation methods. GAN were conceived to be a *generative* method able to obtain higher quality synthetic samples with more diversity, compared to other generative methods such as VAE. The GAN architecture (*Figure 13*) is composed by two key elements: the *generator* (*G*) and the discriminator (*D*) [115]. The generator takes the form of a deep neural network (usually a CNN) and in its simplest form, the generator takes as a random vector **z**, that will help to obtain a non-deterministic output. Specifically, the output of the generator will be a synthetic sample (image) of a specific distribution $p(x_i)$, where x_i is the input image [115]. The discriminator is a classifier that instead of trying to classify an image in the correct class, focuses on learning the distribution of the class. In essence, it aims to quantify how representative the class is to the real class distribution. Both elements are trained in a dynamic scheme in which the generator tries to produce fake examples that are close to the real distribution $x_{fake} \sim p(x_i)$ such that the discriminator is "fooled" into thinking that the sample is *real* and not *generated* in a 2-player minmax game fashion [115].

Ever since their first appearance, GANs have seen how their design and applications became increasingly complex and some remarkable improvements were introduced. For instance, conditional GAN (cGAN) [116] lead to architectures such as pix2pix [116] and cycleGAN [117], which allow to learn a mapping f between an input image and an output image. Such applications are particularly useful in the medical domain, where images can come from different machines with different characteristics and thus, a *domain adaptation* in which images are translated to the training domain might prove useful [118, 119]. The generation of synthetic samples through GANs has also caught the interest of the medical imaging community, given the lack of annotated data and the difficulties to obtain them in the medical domain [120, 121, 122].

Broad categories of DL and training of deep models

Generally speaking, and from a classic perspective, DL models in medical imaging can be classified in two main different categories depending on the learning paradigm. Specifically, ***supervised learning*** is the most popular learning paradigm. Supervised learning is characterized by the availability of *labeled data* during the training process. In essence, given a training set $\mathcal{D} = \{ (x_1, y_1) \dots (x_N, y_N)\}$ where (x_i, y_i) represent the pair of input images and their corresponding labels (ground truth), the supervised learning paradigm aims to learn a function to map the input to the outputs $f\colon x \rightarrow y$ [123]. The second learning paradigm is ***unsupervised learning***, where the DL model learns to find hidden structure and relationships in the data by using a training set $\mathcal{D} = \{ x_1, .., x_n\}$ without labels [124]. Unsupervised learning has proven particularly useful in tasks such as dimensionality reduction and representation learning [125].

All in all, both learning paradigms require the definition and combination of an *internal* evaluation function (also called *objective function* or *scoring function*), *external evaluation function* (evaluation metric) and an *optimization technique* to search in the classifiers space the highest-scoring one in terms of the external evaluation score [126]. In particular, the internal evaluation function takes the form of a differentiable function that we commonly aim to minimize (i.e., cross-entropy [127]). The external evaluation function allows us to judge the performance of the model in the task under consideration (i.e., classification accuracy [128]). The most common choices for the optimizers are *Adam* [129] (and its variants [130]) and *Stochastic Gradient Descent* (SGD) [131].

2.2 Applications of DL in imaging data in PCa

Reading radiographic images, such as MRI, comes down to recognizing complex patterns which computers can be trained to do efficiently, reproducibly and fast. DL offers an alternative to standard human-based and analysis for a variety of PCa imaging-based applications. The applications can be categorized into *low-level processing methods* and *high-level image analysis*. In the first case, the applications deal with the classification of pixels in basic image tasks such as segmentation and registration. On the other hand, high-level applications provide information such as PCa detection, diagnosis, characterization and grading.

Segmentation. Prostate segmentation and accurate identification of the deformable prostate capsule (*Figure 14*) is important for a variety of applications such as radiation treatment planning, volume measurements, fusion-targeted biopsies or monitoring of the prostate disease over time [132]. Some examples of automatic segmentation of the prostate based on DL include the works of Li *et al.* [133] and Aldoj *et al.* [134] which make use of T2w MRI to segment in an automatic way the prostate in 2D (that is, making use of T2w slices). Other works have tried to exploit the inherent 3D nature of MRI in an isotropic way, with success such as in Meyer *et al.* [135]. Other remarkable examples include the work of Sanders *et al.* which evaluate in a prospective way the ability of DL algorithms to segment the prostate and organs at risk for radiation therapy assessment [136]. Some works have benefited from open

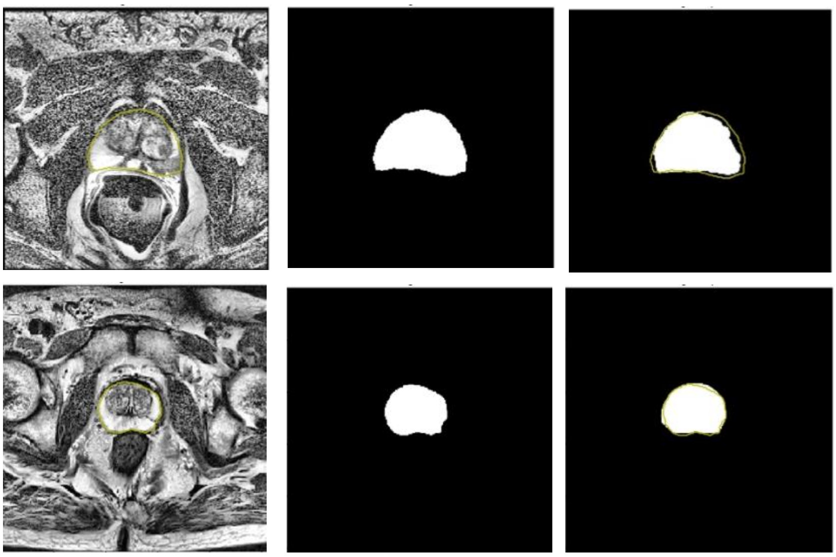


Figure 14. Segmentation of prostate capsule in T2w MRI. First column: results in the T2w slice. Second column: Ground truth Third column: Segmentation overlayed on the ground truth.

access databases (*Table 3*), which have played a key role in facilitating the research in prostate segmentation. For example, NCI-ISBI 2013 [137], PROMISE12 [138] and I2CVB [139] have allowed researchers to further develop prostate segmentation algorithms [140, 141].

Registration. Registration plays a crucial role for applications such as fusion biopsy with MRI and TRUS-targeted biopsy [132]. For example, registration of T2w and 3D TRUS volumes of the prostate through CNNs was proposed by Hu *et al*. [142]. Additionally, in the work of Haskins *et al.* [143] a DL approach to learn in an automatic way a similarity metric for MRI-TRUS registration such that an automatic registration and assessment of such a registration can be performed afterwards.

Diagnosis and prognosis. DL algorithms offer the possibility of automatizing the diagnostic and prognostic of PCa patients based on MRI. They enable the creation of diagnostic probability maps and to extract prognostic features within the pixels that might correlate with histological grading and clinical outcomes.

Dataset	***Field strength (T)***	***Manufactor***	***Number of cases***[1]
NCI-ISBI	1.5 and 3	Siemens and Philips	60
I2CVB	3	Siemens	20
PROMISE12	1.5 and 3	Siemens and GE	37

Table 3. Gleason score risk stratification.

As in segmentation tasks, Open access data has greatly facilitated the creation and research of tools for the automatic diagnosis and prognosis of PCa patients. For example, the Prostate MRI Gleason Grade Group challenge (ProstateX) [138] provided the research community with > 300 000 MRI slices from 347 patients who had MRI-guided biopsies. The “challenge” derived from that open access data obtained results that were similar to the ones obtained by human readers (radiologists), leading to the conclusion that DL methods were suitable to screen scans as a ‘first reader’ or act as an independent second reader in place of a review by a second radiologist [138].

Works such as the from Wang *et al.* and Ishioka *et al.* [144, 145] present a 2D lesion classification approach (slice level of the T2w MRI sequence), in which the different lesions -tumours- present in the radiographic images are classified depending on their GS in cS (GS ≥ 7) or ncS (GS < 7) using a CNN inspired by the VGG16 architecture. Other approaches such as the one presented in Le *et al.* [146] exploit bi-parametric MRI (ADC and T2w) and 2D patches extracted from them based on ROI around the tumour to, again, classify lesions based on their severity defined using the GS. Other approaches have tried to tackle the problem from a 3D perspective making full use of the volumetric information of MRI, such as the one presented in Mehrtash *et al.* and Saha *et al*. [147, 148] in which lesions are classified and detected, in the case of the work of Saha *et al.*

Treatment intervention. Some of the previous approaches can be adapted and extended to plan for PCa treatment and interventions. For example, brachytherapy and external beam radiation therapy can benefit from automatic

[1] https://liuquande.github.io/SAML/

detection of PCa lesions followed by an automatic registration from mp-MRI to CT such that the cancer regions can be used to generate targeted treatments plans [149]. Other honourable mentions of examples of applications in treatment intervention include the prediction of seeds required for low-dose radiation brachytherapy given a prostate volume, as presented in the work of Boussion *et al* [150]. An overview of DL applications in PCa treatment can be found in the work of Almeida *et al* [151].

2.3 Obstacles and limitations of DL in medical imaging and PCa

As highlighted in the previous section, DL has great potential to shape the future of radiology practices and alter the dynamics of it. The AI field is evolving at a really fast pace, with a huge support from industry in the form of heavy investments. Nevertheless, the success of DL is still hindered by several factors, including practices to ensure *fairness* and take into consideration *ethics of data* in the developed algorithms, the so-called *black-box* problem, and arguably the biggest problem of all – the *lack of large amounts of annotated data* and the difficulties associated to obtaining it.

Data bias happens to some degree in any collected data [152], and it can be defined as the differences in performance of the algorithm when encountering subpopulations of different characteristics (e.g., ethnical, economical or technical). In particular, *sampling bias* is quite common and prevalent in radiology, resulting in data with certain characteristics that is only available during the training of the algorithm but does not accurately reflect the characteristics of the data used for the evaluation or even during the deployment of the system [153, 154]. A really common example of selection bias is when data coming from single institution is used to develop and train the DL-algorithm resulting in an under-performing algorithm in the presence of other institutions populations' [155]. In spite of the relevance of an external evaluation protocol using data from other institutions, only 6% of the recent medical DL-papers included validation on an independent external data set [155]. *Data shift* is a subset of selection bias and among the biggest threats to the generalization of DL-systems. Data shift commonly happens because the data used to train the DL-system does not accurately reflect the characteristics of the data that will be used in the future. Whilst for a radiologist is common to assess and take into account technical differences in the acquisition of the data such as the scanner brand, the DL systems are not equipped, in general, to detect those differences if they have not been explicitly trained to take them int account the training phase of the model [152]. Radiology tools based on DL pose the risk to automatize and make biases invisible that are otherwise well-known if rigorous analysis of data used to train the system is not in place.

Questions remain regarding whether we can blindly trust a DL algorithm diagnosis. Historically, DL-systems have lacked mechanisms that allowed to understand why they reach certain decisions or make specific choices. The so-called *"black box"* nature of DL systems can be especially problematic in radiology where a trained radiologist should, under normal conditions, provide an explanation of the train of thoughts behind a certain decision. In a similar way, mechanisms that allow, to some degree, have some traceability and explainability of the DL-systems decisions are required [156].

One of the most prevalent problems and the biggest burden to develop *supervised* DL algorithms are the **difficulties associated to collecting annotated data**. In particular, the first steps to proceed with data collection in radiology usually involve local institutional review board (IRB) approval along with ensuring that all ethical and legal procedures are in place such as patient consent and data protection practices [157]. Such a process usually has the result of long delays in data retrieval and meeting unexpected difficulties during the data collection, along with problems when sharing data with other institutions that could help with *data biases and data shift*. Practically speaking, it is almost impossible to label (annotate) and store all the available data in the radiology domain. Moreover, the specialized knowledge that is usually required to obtain the annotations that are commonly used as ground truth for the supervised learning approach makes it even harder to obtain such annotations due to the economic costs associated to obtaining them [158] and the lack of time of the specialists to dedicate themselves to such a burdensome task. All in all, annotating data is a nuance for specialists and a bottleneck for building DL models that could, potentially, be more intelligent and general without requiring massive amounts of data and by making use of the vast amounts of unlabelled data available in the medical domain or already available open access data sets [159].

3 Deep learning in data-limited scenarios

In this chapter, we focus on the central topic of this thesis and lay the foundations of it: overcoming the lack of annotated data and the available techniques to accomplish it. As discussed in Section 2.3, the performance of *supervised* DL-based algorithms has a large dependency on the availability of large-scale annotated data [160, 161, 162, 163]. Given the difficulties associated to collecting large data-sets with enough diversity and high-quality images from multiple institutions to ensure the generalization of the model for clinical use and the limited radiologist availability, tedious annotation processes coupled with the complexities associated with data de-identification processes, alternative techniques able to deal with limited annotated training data are required.

We focus on different strategies able to boost the performance of DL methods in data-limited scenarios. In particular, we discuss standard techniques such as data augmentation (e.g., rotate an image) or transfer learning (TL), which are the base for the techniques such as generative self-supervised learning (SSL). In that regard, we also discuss more advanced techniques such as synthetic data augmentation by leveraging GANs, representation learning through AEs, and SSL techniques, while differentiating between contrastive and generative approaches (*Table 4*).

3.1 Data augmentation

In a *supervised learning* paradigm, we are interested in mapping an image to some output (label) $f: \mathrm{x} \to \mathrm{y}$ through a DL model. Naturally, the number of samples is close to being proportional to the number of parameters of the model to get a good performance. In addition, the number of parameters needed is usually proportional to the complexity of the task the DL model has to perform. Moreover, to build a useful DL model we are interested in having a decreasing validation error along with the training error, ensuring the generalization ability of our model. With small data sets, we risk *overfitting* the model. That is, having a good performance with the data used to develop the model but not having a good generalization ability.

Technique	***Main components***	***Explanation***
Synthetic augmentation	GAN	Generate new samples
Representation learning	AEs	Learn how normal samples look like (distribution).
Contrastive SSL	Encoder, pre-text and downstream task	Design a pre-text task to learn features via an AE that will be used in a down-stream task.
Generative SSL	AE, pre-text and downstream task	Learn the concept of similarity through an encoder.

Table 4. Summary of techniques for data-limited scenarios.

Data augmentation (DA) is a technique that approaches the overfitting problem from the root, the training set. DA assumes that more information can be extracted from the original dataset through certain image manipulations or synthetic generation of data, such that these manipulations or synthetically generated data artificially inflate the training dataset size while preserving the label y associated to the original data. In particular, data augmentation based on basic manipulations works on the premise that CNNs are invariant to translation, viewpoint, size or illumination (or a combination of all of them). In a real-world scenario, we might have a dataset of images taken in a limited set of conditions but the targeted application may exist in a variety of conditions which can be accounted for by training our neural network with additional transformed data or synthetic samples.

3.1.1 Basic image manipulations

The first studies on the effectiveness of DA focused on simple image manipulations such as *geometric transformations* [164] (*Figure 15*). Both *horizontal* and *vertical flipping* are common geometric transformations that have proven useful on datasets such as CIFAR-10, ImageNet and medical image-based applications [165, 166]. *Cropping* can be used as a pre-processing

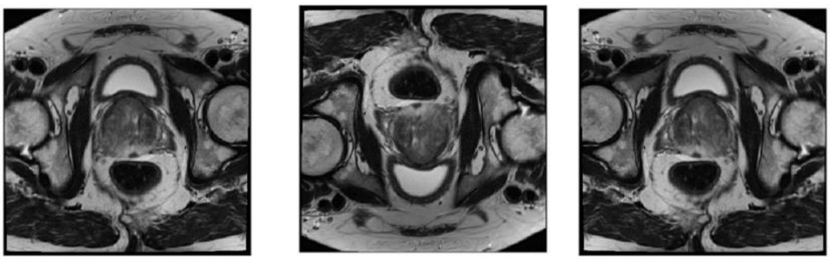


Figure 15. Examples of basic image manipulations. From left to right: original MRI, vertical flipping and horizontal flipping.

step in image-based analysis pipelines in which the data has mixed height and width dimensions by cropping a central patch (or around the ROI) for each image. In addition, *random cropping* can be used to provide a similar effect to augmentations such as *translation*. Both *translation* and ra*ndom cropping* can help to avoid positional bias in the data. Whilst *translation* preserve spatial dimensions, *random cropping* will reduce the size of the input. With *rotation* the image is rotated a certain degree. *Random erasing* [167] is inspired by other techniques such as *dropout regularization*. *Random erasing* forces the model to learn more descriptive features about an image by preventing it from overfitting to a certain visual feature in the image. The technique works by selecting a random patch of an image and masking it with either 0s, 255s or mean pixel intensity values.

The concept of *augmentation safety* is especially relevant for medical imaging. *Safety* refers to the likelihood of preserving the label of the original image *post-manipulation*. For instance, some manipulations such as *blurring* might lead to distorted characteristics of the original MRI which might completely change the diagnosis (label) or distort properties such as boundaries of certain structures (prostate) associated to that particular picture, leading to a sub-optimal learning and performance of the DL algorithm. Hereby, it is of particular relevance to evaluate whether the transformation that is being applied is suitable for the task and imaging data under consideration [164].

DA techniques based on basic image manipulations have been extensively studied for prostate MRI such as DWI. In the work of Ruqian *et al.* [168], where they are studied in the context of PCa detection and classification and evaluated based on AUC. In the study presented by Zia *et al.*, the authors present an evaluation of four AEs and different DA techniques in the context of semantic segmentation of the prostate in T2w [169]. In Cipollari *et al.*, the authors experiment with different DA techniques to boost the performance of a CNN-based system to determine the quality of MRI images that are subsequently used for other tasks, such that there is no degradation in that final task [170].

3.1.2 Synthetic data augmentation

Generative modelling has arisen as an exciting and effective alternative for DA. Generally speaking, generative modelling is the practice of creating artificial instances (synthetic samples) from a dataset such that they retain similar (and fundamental) characteristics to the original set. Synthetic generation of samples (*Figure 16*) offers the benefit of higher variability and less correlated data when compared to techniques such as basic image manipulation, providing more information to the algorithm in the training phase [171]. Synthetic samples are commonly created using GANs [88].

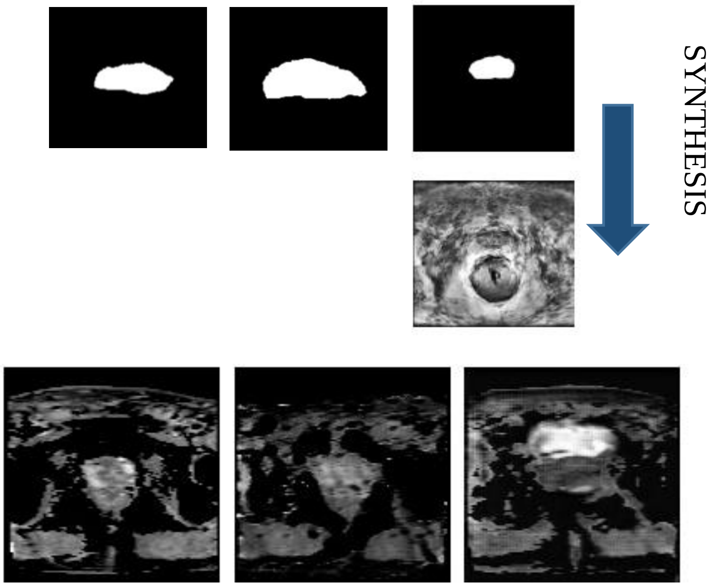


Figure 16. Synthetic generation of samples via GAN. Top row: generation via synthesis. Bottom row: Generation via GAN.

In radiology, GANs have been used to synthesize medical images like chest radiographs [172], CT scans with lung nodules [173], images from brain MRI sequences [174] and prostate MRI scans [120] with improved performance in different applications. For instance, in [173] the DL algorithm achieved a better sensitivity and specificity through the addition of synthetic samples. Another interesting application of synthetic samples is its usage as an oversampling technique to solve problems with class imbalance [175]. In spite of its potential, GANs also have some downsides; they usually require large amounts of data to be synthesize realistic samples [176] and it is computationally expensive to obtain high-resolution samples. In particular, more studies and use cases are required to evaluate the model performance when abnormalities are found in the training data and how it compares with the training with authentic images [163]. Moreover, GANs, and DA techniques in general, can be an amplifier of biases and shifts presents in the dataset and a thorough evaluation of the original characteristics should be required before increasing the sample size based on that data. A comprehensive review on DA with GANs and its shortcomings can be found in [176, 177].

GAN-based generation has been studied and applied in diverse PCa applications. In the work of Yu *et al.* [178] the authors make use of a capsule network-based GAN to generate prostate MRI which are later used for the classification of images in PCa or no PCa. In Xiaodan *et al.* the authors generate prostate MRI with different GS grades associated to them in order to avoid data biases during the training of the network to classify images in their respective GS grade groups due to the unbalanced nature of the classes [179]. Zhiwei *et al.* propose a novel GAN model to synthesize high-quality ADC images of cS PCa to fight, again, against potential data biases due to data imbalance and the difficulties associated to obtain cS data [180].

3.2 Representation learning and the concept of learning the "normal"

Generally speaking, in an annotated dataset the cases that present abnormalities (otherwise called positive) are generally scarce. As presented in *Section 2.1.1*, AEs are able to extract useful representations from the input data by reconstructing it from a compressed representation of it. By exploiting such an ability of AEs and abundant negative (controls, healthy or normal cases), the AE is able to learn the "concept of normal" and distinguish in an *unsupervised* *way* those cases that greatly differ from that normality (*Figure 17*). In particular, AEs can be trained exclusively with normal cases such that representative features are learnt only from them and subsequently, they can be used to distinguish abnormal from normal findings on the basis of the deviations of the features from the learnt ones that correspond to the normal class [181]. This process can also be understood as *outlier detection*, which has been extensively researched in other areas [182].

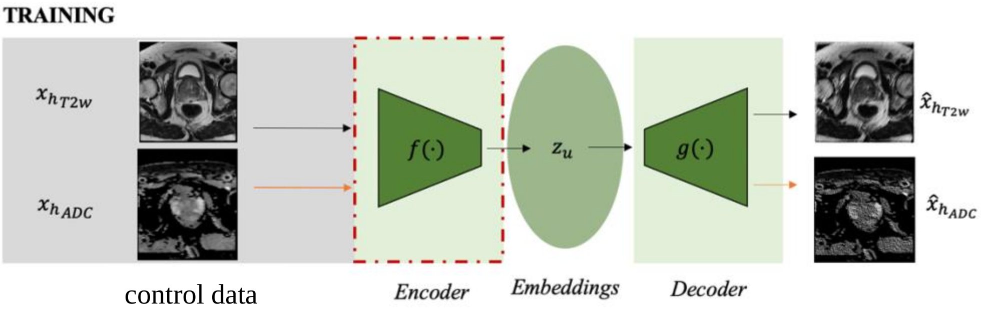


Figure 17. Learning the concept of normality through AEs.

In the work of Chen *et al*. [183] adversarial AEs are used with an extra regularization term to constrain the learnt representations in the embedding space. After being trained on healthy data exclusively, the authors show the usefulness of the trained model to detect out-of-distribution data and detect lesions in brain MRI with a performance that is on-pair with its supervised counterparts. On the other hand, in the work of Wong *et al*. [184] a CNN is trained using CT scans of a normal heart anatomy. The feature maps obtained through the CNN provide information to detect deviation from normal anatomy which can be further used to improve the detection performance in the presence of a limited number of positive samples.

Some authors have applied the concept of normality or anomaly detection for PCa. In the work of Jingya *et al*. the authors develop an end-to-end unsupervised framework to estimate which samples might degrade the performance of a previously developed detection PCa model by detecting which samples are out-of-distribution (OOD) [185] by means of an AE- based reconstruction [185] process. Other approaches have focused on PET and CT images to extract features via AEs such that anomalies (tumours) can be detected in the feature space by means of density estimation [186].

3.3 Efficient use of different data sources and data fusion

Information about the same phenomena can be acquired from different types of medical data. For example, the most common prostate MRI protocols include different MRI sequences and orthogonal views (for T2w) (*Section 1.2.1*). In the same way a human-based analysis usually requires or benefits from using data from different modalities, it is rare that a single modality or view provides complete knowledge or information of the phenomenon of interest [187]. Data fusion is particularly interesting in scenarios with a limited amount of data, as making use of all of the available sources in an efficient way virtually increases the amount of available data for the task under consideration whilst potentially improving the performance of the methodology under consideration.

Generally speaking, in a DL framework fusion can be defined at different levels being *early fusion and late fusion* levels the most common ones [188] (*Figure 18*). Specifically, early fusion integrates the multi-modal or multi-view information from the original space of low-level features [189, 190], merging the data at the input of the network. However, this approach has some limitations as discussed in [191], as it is difficult to discover highly non-linear relationships between low-level features more so when the modalities have significantly different statistical properties. On the other hand, *late fusion* approaches can be implemented by means of independent CNNs (acting as an encoder) for each data source, and fusing the outputs of the different networks in higher-level layers, allowing the discovery of highly non-linear relationships between the data. Fusion methodologies commonly require other complementary operations in the form of pre-processing. For instance, registration procedures might be required when misalignment is present in the imaging data [192]. In spite of the efforts, data fusion remains as a challenging area in which further developments are required to make use of the available information in an optimal way [187].

Fusion strategies have been employed for several PCa applications. For example, in the work of Meyer *et al.* [193] independent 3D CNNs are trained and fused at a late stage or ensembled to produce results that do not depend exclusively on axial plane but rather that exploit the information of the different views with the objective of improving prostate segmentation results. In the works Le *et al*. [194] the authors present an early fusion approach for different modalities of prostate MRI with the objective of improving PCa lesion classification 3D.

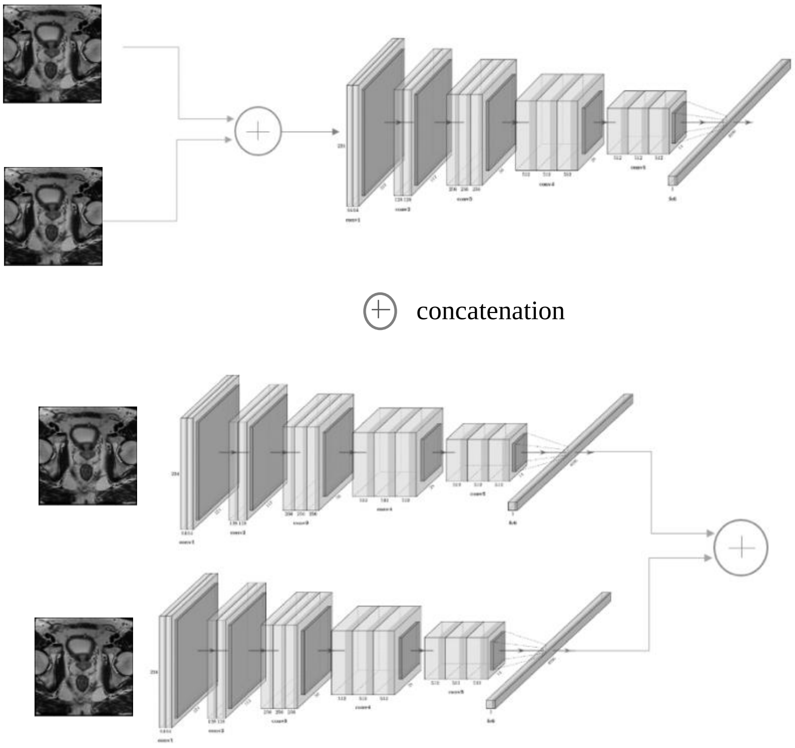


Figure 18. Early fusion (top row) and late fusion (bottom row) by concatenation.

3.4 Transfer learning

Transfer learning (TL) remains as one of the most widely used techniques to overcome the limitations in the availability of annotated data in the medical image domain [161]. In TL, a DL model is trained on a large-scale and annotated dataset (*pre-training)*, with the underlying assumption that the low-level learned features are generic enough such that they can be applied to the target domain with a limited amount of data [195, 196]. Different strategies can be adopted in TL: fixing the early layers of the DL architecture, re-training the high layers (*shallow tuning*) or fine-tuning the whole architecture (*deep tuning*) (*Figure 19*). In particular, fine-tuning is one of the most common TL approaches which allows the learned features that were obtained in the *pre-training* stage to adapt to the target domain. One of the critical factors in TL is the *similarity* between the *source domain* (the one from which the features were obtained in the pre-training stage) and the *target domain*. As shown in the work of Raghu *et al.*, using an *out-of-domain* TL approach (that is, with large differences between the source and target domain) leads to sub-optimal results [197].

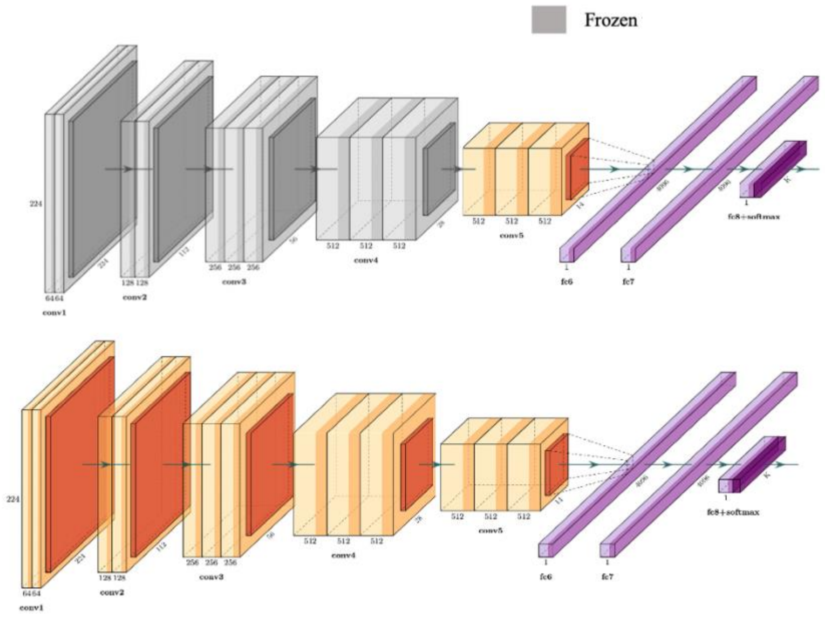


Figure 19. Shallow tuning (top) and fine-tuning (bottom) in transfer learning.

In spite of the sub-pair results obtained with out-of-domain TL, TL techniques have been widely adopted in 2D DL applications for radiologic images. For example, one of the first applications was a fine-tuned CNN that processed different views of mammographic images [198], which compared the performance with training from scratch. In the work of Aljundi *et al.* [199], a TL approach is applied to 2D MRI-based PCa screening with success, as shown by the gains in performance when compared to other approaches. More examples of successful applications of TL in radiological images can be found in the work of Shin *et al.* [171].

3D TL requires a special mention, as TL is easily applicable to 2D settings because of the abundance of pre-trained models but 3D models are not as established and hence, there is a limited amount of pre-trained models. In spite of it, several approaches have been proposed such as in Chen *et al.* [200], where they curated a large-scale 3D dataset and used it to train 3D CNNs that were later used as pre-trained models for lung segmentation and pulmonary nodule classification. Other works tried to accommodate 3D data into 2D to make use of existing pre-trained models, such as in Han *et al.* [201].

Transfer learning has been used extensively in 2D applications for PCa. For example, in the work of Yixuan *et al.* [202] the authors propose a three-branch architecture for mp-MRI images that make use of a pre-trained model to extract features that are concatenated in a late-fusion approach to classify PCa lesions. In Islam *et al.* the authors use TL to train two models based on VGG-like architectures to detect and identify lesions in PCa [203]. Finally, in the work of Hoar *et al.* TL (along with other techniques assumed to boost the performance of the model such as test-time augmentation and standard augmentation) is used in combination with mp-MRI data to boost the performance of the developed DL model for lesion segmentation [204].

3.5 Self-supervised learning

Self-supervised learning (SSL) made its first appearance in robotics, where labels are assigned to the training data by making use of the relations between the different inputs [205]. Broadly speaking, SSL methods gained popularity and raised as an alternative to mitigate the time-consuming and expensive data annotations process required to obtain large amounts of annotated data and keep improving the state-of-the-art results in computer vision tasks [206]. SSL falls in the category of *unsupervised learning* and in essence, SSL methods aim to learn visual features from large-scale unlabelled data by obtaining a supervisory signal from the unlabelled data itself and hence, usually, being an *in-domain* initialization method (assuming the unlabelled data is from the same domain as the one that will be used later on to evaluate the quality of the representations obtained by means of the SSL method) [206].

Some important concepts in SSL methods are the *pre-text task* and the *downstream task*. The *pre-text task* are pre-designed tasks which are designed for a certain deep network to solve with the objective of learning visual features by learning objective functions linked to the pre-text task. *Pseudo-labels* can be defined as labels that are automatically generated for a pre-defined pre-text task without involving any human annotation [206]. *Downstream tasks* are applications that are used to evaluate the quality of the features learned by means of the *pre-text* task, its associated objective function and *pseudo-labels*. Once the SSL training is finished, the learned features are transferred to the downstream task as pre-trained models to improve performance and overcome overfitting in the presence of small amounts of data. The general pipeline of SSL is shown in (*Figure 20*).

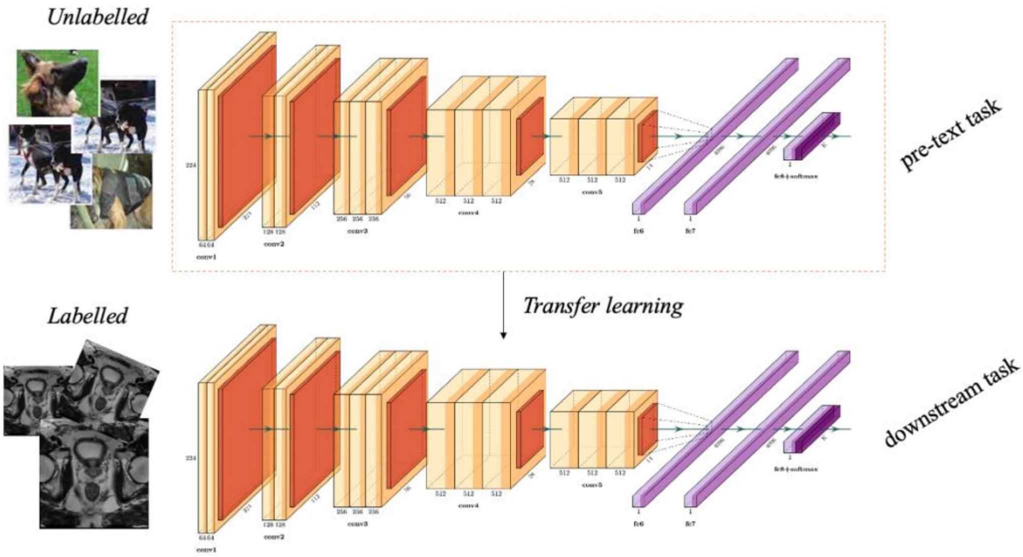


Figure 20. General framework of SSL methods.

SSL methods can be divided in *contrastive* and *generative*, depending on the nature of the pre-text task. As explained with more details in the next sections, the difference between the two categories lies in model architectures, pre-text task and objectives.

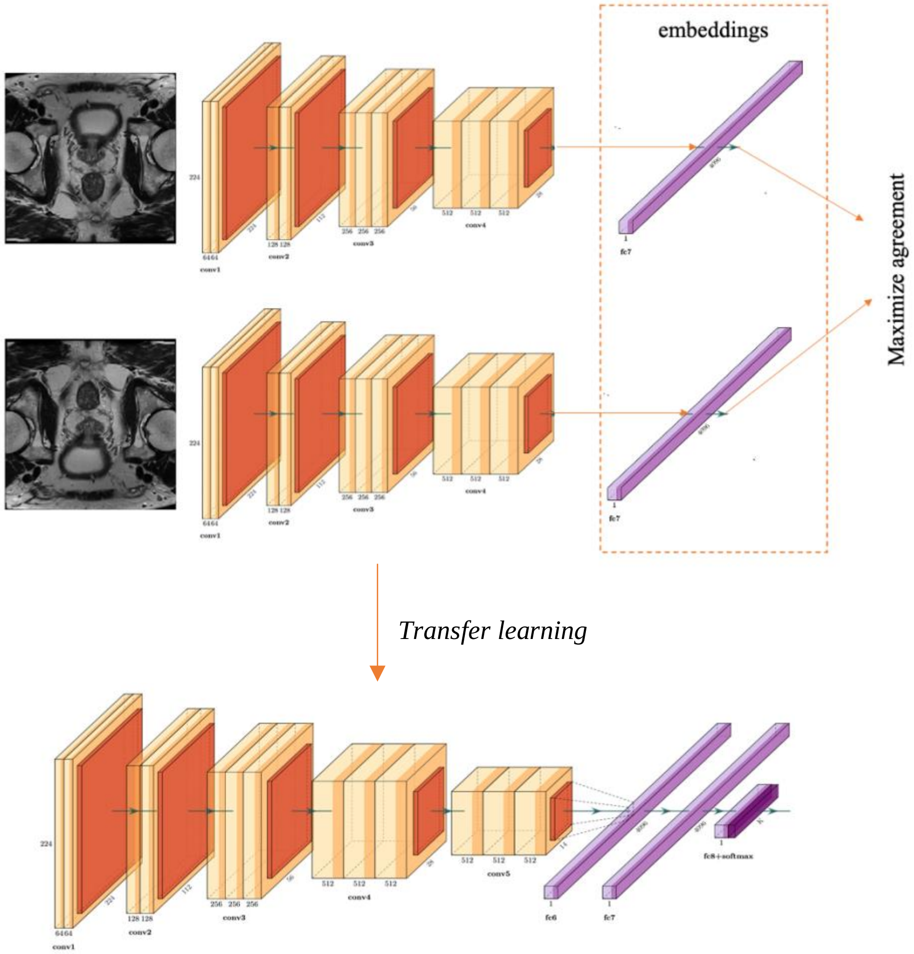


Figure 21. Contrastive SSL methodology.

3.5.1 Contrastive SSL

Contrastive SSL (*Figure 21*) aims to train an encoder to encode an input *x* into an explicit vector *z* to measure the similarity based on some metric of choice (e.g., mutual information maximization). Commonly, contrastive SSL makes use of an encoder that explicitly models *z* while making use of a contrastive similarity metric, being Noise Contrastive Estimation (NCE) [207] and InfoNCE [208] two of the most common similarity metrics. The similarity measure and the encoder architecture might vary from task to task but the overall idea and framework remain the same for all contrastive SSL approaches. Overall, *contrastive SSL* frameworks can be divided between two types: *context-instance* and *instance-instance* [205].

Context-instance contrastive SSL aims to model the relationship between the local feature of a sample and its global context representation. For instance, images contain rich spatial relations between parts of it (the head of a human is on top of the neck, for instance). Some *context-instance* SSL models focused on recognizing relative positions between parts of it as a pretext task [205], such as predicting patches position [209] or to recover the positions of shuffled patches like a jigsaw puzzle [210, 211].

The alternative to *context-instance* SSL is *instance-instance*, where the direct relationships between different samples. In *instance-level* contrastive SSL the focus is on the main instance (for example, in an image classified as a dog we want to focus on the dog and not on the context such as grass), with the hypothesis that what matters the most for the downstream task is the instance itself rather than the context. One of the first ways to study instance-instance based methods is through clustering [212], being SwAV the most recent and successful implementations [213], which incorporates multi-view augmentation and aims to assign views of the same images to the same prototype (clusters). Besides cluster-based approaches, CMC [214] proposed to adopt multiple different views of an image. MoCo [215] draws inspiration from CMC and the idea is further developed via *momentum contrast,* which increases the number of negative samples during the SSL method training. Nevertheless, MoCo adopts a fairly easy strategy: a pair of positive representations come from the same sample without any transformation or augmentation which makes the positive pairs (similar ones) easy to distinguish. In SimCLR [216], the authors introduce data augmentation to increase positive pairs whilst also introducing a learnable non-linear transformation between the representation and the contrastive loss [216].

Some remarkable examples of contrastive SSL techniques applied to medical imaging include the works of Tao *et al.* [217] in which a volume-wise transformation for context permutation is proposed and shown to offer improvements in tasks such as pancreas segmentation. In the work of Sowrirajan [218] *et al*. an existing SSL framework such as MoCo is used and shown to have a positive effect for downstream tasks in chest X-ray, obtaining better results than its counterparts in the presence of a limited amount of data.

Azizi *et al* [219] propose a multi-instance contrastive learning (MICLe) partially based on SimCLR, in which multiple images of the same pathology of the same patient are leveraged during the SSL stage. With their approach, they show improvements in dermatology and chest X-ray classification tasks and outperforming supervised baselines pre-trained on ImageNet. In the work of Li *et al*. a contrastive SSL approach is proposed to detect lesions in mammograms

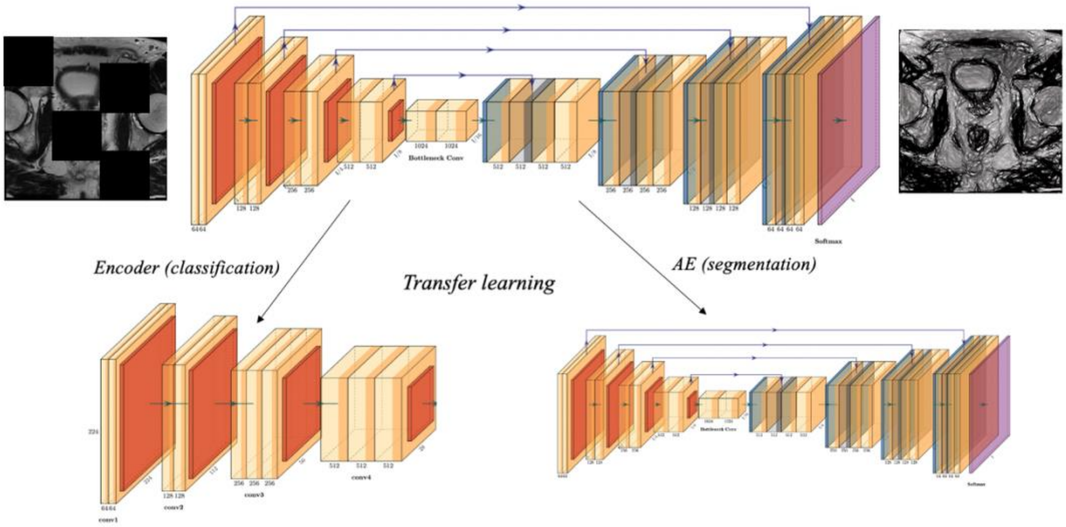


Figure 22. Generative SSL with a restoration pre-text task (occluded image).

and to learn invariant features to various vendor-styles [220]. When it comes to MRI and PCa, contrastive SSL remains as an unexplored field.

3.5.2 Generative SSL

In generative SSL (*Figure 22*), an encoder is trained to encode an input *x* into an explicit vector *z* and a decoder to reconstruct *x* from *z*. In contrast with contrastive SSL, generative SSL commonly makes use of a reconstruction loss [205]. The most common model used in generative SSL is the AE model, where the goal is to reconstruct (part of) inputs from corrupted versions of them (*Figure 22*). Some examples of generative SSL try to recover a partial input in which instead of asking the model to recover a whole input they provide models with a partial input and ask them to recover the rest of the parts. Some remarkable examples of applications that follow this principle are colorization [221, 222], inpainting [223] and super-resolution [224]. In some cases, the proposed methods leverage a discriminative loss function as the objective by making use of the idea of *adversarial learning*, which tries to reconstruct the original data distribution rather than the samples by minimizing the distributional divergence.

Some remarkable examples of generative SSL in medical imaging are the works of Zhou *et al.* [225] in which different types of transformations are applied to 2D and 3D images and learning is accomplished by reconstruction. In the work of Taleb *et al.* different tasks such as relative patch location are presented and shown to be useful for 3D downstream tasks [226]. In Chen *et al.* [227] context restoration is applied for 2D medical imaging tasks, in which the position of patches is swapped and they aim to recover the information that was originally in that position. In a similar fashion, in another work of Taleb *et al.* [228] multi-modal medical images are mixed in a patch-based way and the original content is then reconstructed from the mixed image, which serves as a pre-text task.

The examples of generative SSL applied to PCa are rather limited, but some of the most remarkable ones include the works of Bolus *et al.* [229] where a context restoration pre-text task is applied with “outer cuts” in the MRI and the process is shown to improve the ability to detect PCa with two different architectures. In the work of Qian *et al.* [230] a generative SSL approach is also applied where the input is distorted via injection of gaussian noise and then denoised, with the objective of learning features that might be useful for the tested downstream tasks such as segmentation. Finally, the authors show improvements in the final results thanks to the SSL procedure.

4 Materials

4.1 Data and Ethics

The data used in this work was obtained from different sources depending on the task under consideration. All the sources have something in common: they are publicly available databases. *Table 5* presents a summary of the data used for the project. In particular, we use the Prostate MR Image Segmentation challenge (PROMISE12) dataset [139]. The objective of the challenge and the dataset was to standardize the evaluation of the WG segmentation of the prostate and to objectively compare the performance of different algorithms. The dataset includes data from four different centres to account for the differences in pixel/voxel intensities due to different acquisition protocols: Haukeland University Hospital in Norway, Beth Israel Deaconess Medical Centre (BIDMC) in the US, University College London (UCL) in the United Kingdom and the Radboud University Nijmegen Medical Centre (RUNMC) in the Netherlands. Each of the centres provided 25 axial (transverse) T2-weighted MRI images, resulting in a total of 100 MRI sequences. Details of the acquisition protocols for the different centres can be found in *Table 6*. The acquisition plane was axial because of the number of anatomical details contained in it [231]. From those 100 MRI sequences, 50 are included as a training set, 30 as a test set and 20 for the live challenge. Each centre provided a reference segmentation performed on a slice-by-slice basis of the prostate WG provided by an experienced reader. The contouring required for the segmentation was performed in either 3DSlicer (www.slicer.org) or MeVisLab (www.mevislab.de). Following, the segmentations were double checked by a second expert that had no part in the initial segmentations.

Dataset	*Subjects*	*Field strength*	*Task*
ProstateX [232]	204	3T	Lesion classification
Promise12 [139]	50	1.5 and 3T	Prostate segmentation

Table 5. Summary of datasets used in the work.

Centre	*Field strength*	*Coil*	*Manufacturer*
Haukeland	1.5T	Yes	Siemens
BIDMC	3T	Yes	GE
UCL	1.5 and 3T	No	Siemens
RUNMC	3T	No	Siemens

Table 6. Acquisition protocols for each center involved in PROMISE12.

We have also used data from the ProstateX dataset [232], an open-access dataset that was the result of a collaborative effort sponsored by the SPIE, the AAPM and the NCI. In particular, we focus on the data from the first challenge "SPIE-AAPM-NCI Prostate MR Classification Challenge". The dataset consisted of a prostate mp-MRI cohort acquired from a single centre (Radboud University Medical Centre) in which each mp-MRI scan was read or supervised by an expert radiologist (20 years' experience) who indicated point-based suspicious findings and assigned a PI-RADS score. Findings with a PI-RADS score $\geq$ 3 were referred to a biopsy. Biopsy specimens were graded subsequently by a pathologist with over 20 years of experience, and these results were used as ground truth for the challenge [233]. Each mp-MRI scan included multiple orthogonal T2-weighted, dynamic contrast-enhanced (DCE) and diffusion-weighted imaging (DWI). Location and a reference thumbnail image were provided for each lesion, and each lesion had a known pathology-defined GS group which defined the ground truth for the challenge.

In particular, for the first challenge, those patients with a GS $\geq$ 7 were considered to have a clinically significant lesion while those with GS $<$ 7 were considered to have a non-clinically significant lesion. The challenge contained mp-MRI scans of 300 prostate lesions corresponding to 204 patients for the training set and 208 lesions corresponding to 140 patients for the test set, along with spatial location coordinates, anatomic zone location and known clinical significance of each lesion. In spite of not being part of the original dataset, segmentations of the prostate can also be found for the first challenge. Specifically, 66 cases selected at random were segmented and high-resolution segmentations were obtained by considering three scan directions: axial, sagittal and coronal [234]. The gland was manually delineated by a medical student, followed by a review and corrections of an expert urologist. Additionally, both lesion masks and prostate segmentations of the axial direction can be found on https://github.com/rcuocolo/PROSTATEx_masks. The segmentation data is the result of a lesion-by-lesion quality check conducted at the Department of Advanced Biomedical Sciences of the University of Naples "Federico II" for both T2-weighted and ADC images by two radiology residents and two experienced board-certified radiologists [235].

Finally, a special mention to datasets that are being used in our work in progress: PROSTATE-MRI [236], PROSTATE-DIAGNOSIS [237] and TCGA-PRAD [238]. In all cases, the nature of the studies is retrospective and consists of multi-view T2w MRI with different acquisition protocols. In the first case, the number of patients included in the study is 10, in the second case is 16 and in the last one is 10. All the data includes a significance level of the lesion based on GS from biopsy results analysed by an expert in pathology and urologist. Additionally, ground truth segmentations of the prostate WG are obtained for the multi-view data based on available software [234] and manual revision of the results. In all the cases, since the nature of the data was open access ethical approval was not required.

4.2 Software

3D Slicer. 3DSlicer is a widely used software package that provides automated and accurate analysis (including registration and interactive segmentation) and visualization (including volume rendering) of medical images and for research in image guided therapy [239]. Additionally, 3DSlicer is extensible as it has powerful plug-in capabilities for adding algorithms and applications.

To obtain visualizations of the prostate MRI (volume rendering too) and run quality checks of some of our developed applications such as the ones focusing on segmentation, we used the latest version available of 3DSlicer. Additionally, 3DSlicer is also been used to obtain in-house (Stavanger University Hospital) segmentations and will be the main tool to obtain the data used for future studies.

Python. The majority of the research carried out in this thesis has been based on code developed in Python. In particular, the development has been mainly supported on the basis of different libraries and frameworks: Numpy [240], Matplotlib [241], Tensorflow (Keras) [242], SimpleITK [243], classification models [244] and segmentation models [245], being the last two GitHub repositories, which host a variety of models ready to be deployed and tested in classification and segmentation applications that use a Tensorflow/Keras environment.

5 Summary of contributions

In this chapter, we present a summary of the seven papers (either as a proceeding contribution or journal article) included in the thesis. *Contributions A, B, C, D and E* conform the first part of the thesis, in which we focus on how to tackle data scarcity for prostate MRI. In particular, contributions A and B focus on GAN-generated data (*Section 3.1.2*), whilst contribution C focus on anomaly detection with AEs (*Section 3.2*) and contribution D and E focus on SSL applications for PCa (*Section 3.4*). *Contributions F and G* are part of the second part of the work, in which we explore how to efficiently make use of different data sources (multi-planar) acquired during MRI acquisition.

5.1 **Contribution A**: *Improving prostate whole gland segmentation in T2-weighted MRI with synthetically generated data*.

The main objective of ***contribution A*** [120] is to tackle the lack of segmentation annotations for the prostate capsule (data scarcity) and to provide an alternative to classic augmentation techniques (basic image manipulations). In particular, we argue that standard augmentation techniques produce highly correlated samples, limiting the variability of the data used to train the algorithm and the amount of information that the data can offer in the training process. We propose a framework based on GAN architectures that is able to generate prostate capsule (WG) masks and then synthesize T2-weighted MRI from them, obtaining paired synthetic samples in a semi-automatic way.

The main application of the study is to segment the prostate capsule and show that we are able to improve the quality of the segmentation by incorporating synthetically generated data obtained with our proposed framework, as compared a baseline trained with standard augmentation techniques (image manipulation) and without any extra image manipulations during the training phase. We apply our proposed framework to a collection of T2-weighted MRI corresponding to 50 patients, that are part of the PROMISE12 challenge [139], a multi-institutional and multi-vendor dataset (more details in Section 4.1).

We follow the steps depicted in *Figure 23* to implement the framework and evaluate it. As we work in 2D (slice level) and the organizers of the challenge

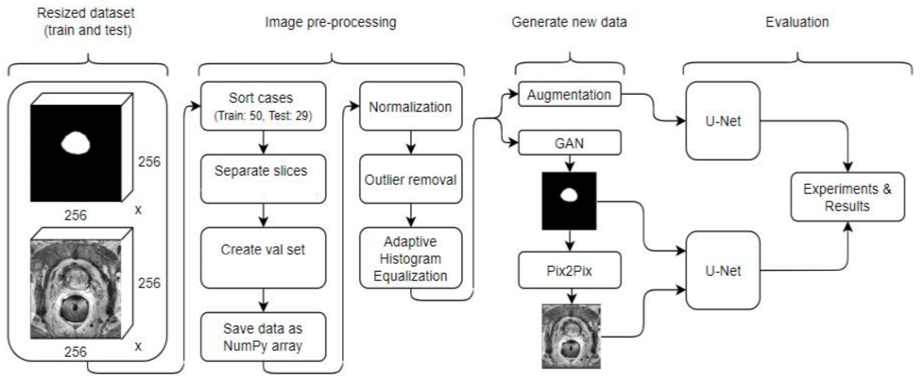


Figure 23. Technical approach to the project.

kept the test set hidden for evaluation purposes, we divide the training set in a 60%/20%/20% in a 2D-slice fashion (train, validation and test) and by patients, such that there is no patient data leakage between the different splits. We perform some pre-processing to the data such that we achieve a harmonization of the different scans coming from different centres and vendors. In particular, we resample the patients' MRI to a resolution of 256x256 (lowest resolution present in the data set) by linear interpolation. Following, the intensity of the MRI is normalized to an interval of [0, 1] and outlier removal is applied by forcing the pixel intensity values of the MRI images between the 1st and 99th percentiles. Finally, a contrast limited adaptative histogram equalization (CLAHE) is applied to improve local contrast and enhance the edge definition of the MRI [246].

We choose a 2D U-net architecture as our base segmentation architecture based on its popularity and previous results in the segmentation of different organs [134, 247]. As a first step to develop the framework to generate paired masks and T2-weighted MRI, we adopt the DCGAN architecture to generate synthetic WG prostate masks [248]. We observed disparities in the quality of the generated synthetic masks and hence, we applied a manual selection criterion of the generated masks based on the visual appearance of the image. For example, we observed that some synthetic masks contained disconnected objects (*Figure 24*), which is unrealistic. Following, the previously generated and selected masks are translated into a paired T2-weighted image, such that we end up with a T2-weighted slice and its corresponding mask. We base the second part of our framework in the pix2pix architecture [117]. Details about the training of the different architectures can be found in [120].

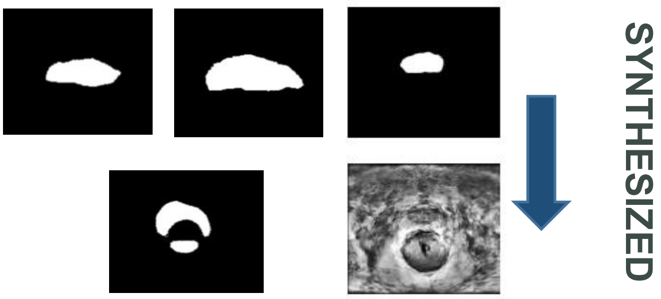


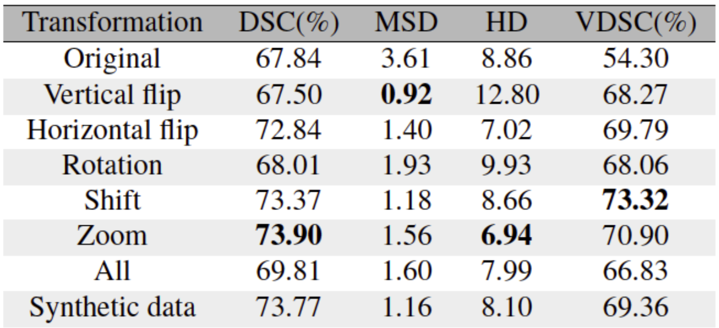
Figure 24. Mask generation and synthesis of new samples.

Finally, we evaluate our proposed framework by evaluating the quality of the segmentation results obtained by means of the U-net architecture, in the presence of synthetically generated samples, standard augmentation techniques and a plain U-net. Specifically, we evaluate the quality of the segmentation results based on dice score coefficient (DSC), mean volumetric (VDSC), mean surface distance (MSD) and mean Hausdorff distance (HD). The standard augmentation techniques included rotation (± 10 degrees), shifting (10 %), flipping and zooming ([1, 1.2 pixels] range). We generate 10000 synthetic T2-weighted MRI images and their corresponding masks for evaluation purposes, which corresponds to approximately 8 times the amount of original data.

Transformation	DSC (%)	MSD	HD	VDSC(%)
Original	67.84	3.61	8.86	54.30
Vertical flip	66.93	18.52	19.68	48.72
Horizontal flip	69.98	15.41	13.47	50.86
Rotation	73.07	3.59	8.53	59.78
Shift	71.33	12.24	9.44	56.16
Zoom	70.69	7.31	**7.74**	55.21
All	67.30	10.14	12.36	51.23
Synthetic data	**73.77**	**1.16**	8.10	**69.36**

Table 7. Results of simple augmentation and addition of synthetic samples.

Our results show a considerable improvement in terms of all the metrics used to evaluate the quality of the segmentation when synthetic data is added to the training process. In particular, metrics such as HD present an improvement of over 8% when comparing the addition of synthetic data and the vanilla process. Furthermore, using synthetic samples also surpasses by a considerable margin the results obtained with standard augmentation (*Table 7*). Using a combination of synthetic data and standard augmentation techniques yields to an even larger improvement when compared to using both techniques independently (*Table 8*).



Transformation	DSC(%)	MSD	HD	VDSC(%)
Original	67.84	3.61	8.86	54.30
Vertical flip	67.50	**0.92**	12.80	68.27
Horizontal flip	72.84	1.40	7.02	69.79
Rotation	68.01	1.93	9.93	68.06
Shift	73.37	1.18	8.66	**73.32**
Zoom	**73.90**	1.56	**6.94**	70.90
All	69.81	1.60	7.99	66.83
Synthetic data	73.77	1.16	8.10	69.36

Table 8. Results for synthetic samples combined with standard augmentation.

5.2 **Contribution B**: *Improving prostate cancer triage with GAN-based synthetically generated prostate ADC MRI*.

Contribution B [249] aims to tackle the data scarcity and lack of annotations in the context of lesion classification (clinical significance) and with ADC maps. In a similar fashion to ***Contribution A*** [120], we aim to provide an alternative to standard augmentation techniques (data manipulation) for PCa classification. We argue, again, that standard augmentation techniques might limit the amount of information provided to the final task under consideration and that even depending on the type of chosen basic manipulation the final results might not see any benefit from it but rather the opposite, the augmentations might distort the image in a way that the label (ground truth) might be distorted as well (*Section 3.1.1*, augmentation safety). To solve it, we propose a synthetic data generation pipeline based on conditional GAN (cGAN) [115] and

DCGAN and explore the effects of both of them on the chosen PCa application.

The main focus of the study is PCa triage, defined as the classification of the clinical significance (GS ≥ 7 or GS < 7) of PCa lesions such that patients can be sorted based on their treatment or further intervention (testing) needs. In particular, we aim to show that by using synthetically generated data during the training of the classification architectures we are able to improve the performance of them. Furthermore, we aim to compare cGAN and DCGAN generation with standard augmentation techniques and no extra manipulations, based on the classification performance results. To accomplish it, we make use of the ProstateX dataset [218] (Section 4.1) consisting of 204 patients diagnosed with PCa and 330 lesions. Among those lesions, 76 lesions are cS (GS ≥ 7) and 254 are ncS (GS < 7).

The technical approach to the project is depicted in *Figure 25*. We work on a 2D level (slice level). We start by applying some standard pre-processing to the data: re-sampling to an image size of 128x128 and normalization of the MRI intensities to a range of [0, 1].

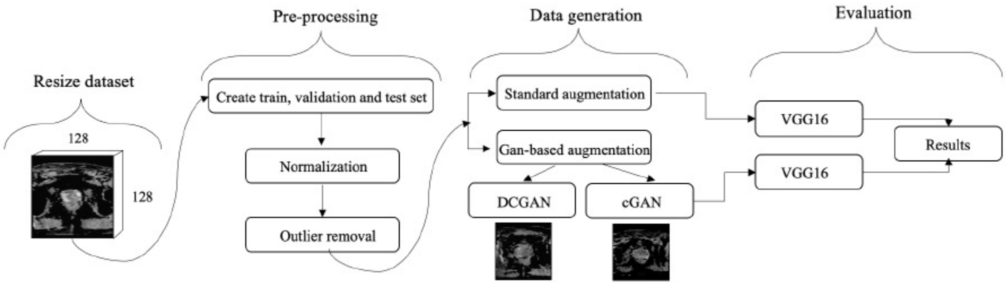


Figure 25. Technical approach for contribution B.

In order to evaluate the *classification architectures*, we split the original dataset in 70%/10%/20%, corresponding to training, validation and test set, respectively. Specifically, the splitting is done in a slice fashion (2D) and by patients, such that the resulting splits have no data leakage. When it comes to data generation, we make use of the full dataset, as the evaluation of the generation is considered to be implicit in the final classification task.

We choose a VGG16 [92] architecture as the classification architecture, based on previous results. As for the GAN architectures, we follow standard implementations described on original sources with small modifications based on experimentation with the architectures [115]. As a first step, we train both DCGAN and cGAN architectures to generate cS and ncS prostate ADC slices (2D) with the training protocol defined in Fernandez-Quilez *et al.* [235]. More specifically, two DCGAN are trained to generate samples for each class whilst one cGAN able to generate both types of samples (conditioned by the user) is trained. Examples of the generated samples with both architectures are shown in *Figure 26.*

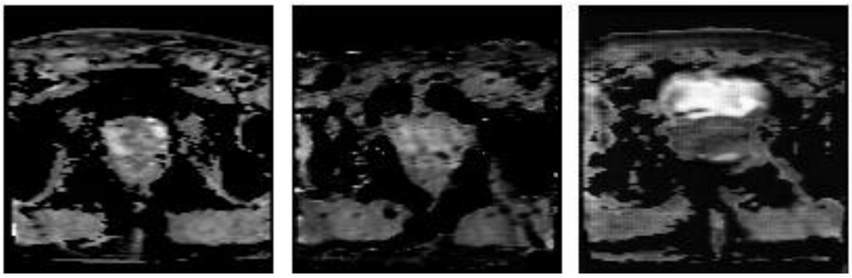
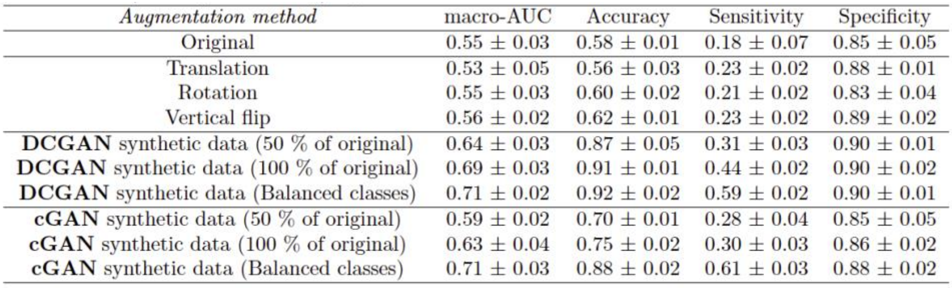


Figure 26. Examples of generate ADC samples with cGAN and DCGAN.

We evaluate the quality of the synthetic samples based on the final classification results obtained with the VGG16 architecture. We compare the performance of the DCGAN samples with cGAN ones, standard augmentation and a plain VGG16 with no extra data manipulation. Specifically, we evaluate the effect of rotation (±25 degrees), translation

(±0.4 pixels) and vertical flipping. Following, we evaluate the effect of using synthetic samples along with data manipulation during the training

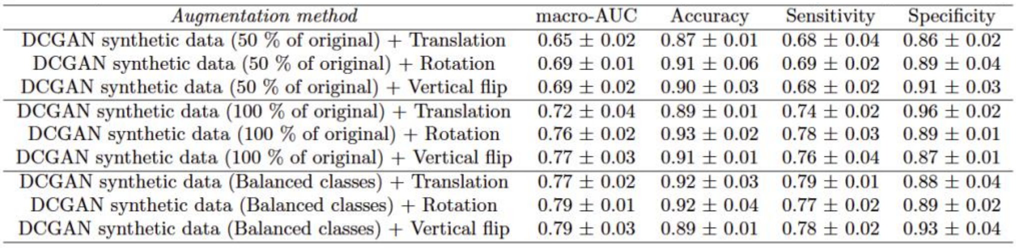


Augmentation method	macro-AUC	Accuracy	Sensitivity	Specificity
Original	0.55 ± 0.03	0.58 ± 0.01	0.18 ± 0.07	0.85 ± 0.05
Translation	0.53 ± 0.05	0.56 ± 0.03	0.23 ± 0.02	0.88 ± 0.01
Rotation	0.55 ± 0.03	0.60 ± 0.02	0.21 ± 0.02	0.83 ± 0.04
Vertical flip	0.56 ± 0.02	0.62 ± 0.01	0.23 ± 0.02	0.89 ± 0.02
DCGAN synthetic data (50 % of original)	0.64 ± 0.03	0.87 ± 0.05	0.31 ± 0.03	0.90 ± 0.01
DCGAN synthetic data (100 % of original)	0.69 ± 0.03	0.91 ± 0.01	0.44 ± 0.02	0.90 ± 0.02
DCGAN synthetic data (Balanced classes)	0.71 ± 0.02	0.92 ± 0.02	0.59 ± 0.02	0.90 ± 0.01
cGAN synthetic data (50 % of original)	0.59 ± 0.02	0.70 ± 0.01	0.28 ± 0.04	0.85 ± 0.05
cGAN synthetic data (100 % of original)	0.63 ± 0.04	0.75 ± 0.02	0.30 ± 0.03	0.86 ± 0.02
cGAN synthetic data (Balanced classes)	0.71 ± 0.03	0.88 ± 0.02	0.61 ± 0.03	0.88 ± 0.02

Table 9. Results of simple augmentation and addition of synthetic samples.

of the architecture. We test different amounts of generated data for every experiment that incorporates synthetic data: 50% of the original amount of data, 100% of the original amount of data and balanced classes (adding as many samples as needed to balance the number of samples of each class during training). The final PCa lesion classification results are evaluated base on AUC, accuracy, sensitivity and specificity and averaged over 5 independent runs with different partitions for training, validation and test sets.

As shown in *Table 9*, both cGAN and DCGAN synthetic samples have a positive effect on the final classification results. Larger improvements are obtained when more synthetic data is used. In particular, we can observe how DCGAN obtains slightly better results than cGAN at the expense of requiring two architectures: one for each class. We observe an even larger improvement when combining synthetic samples with basic data manipulation, as depicted in *Table 10*. The result highlights the add-on nature of synthetic samples, showing that can be added on top of other methodologies to improve the final results.



Augmentation method	macro-AUC	Accuracy	Sensitivity	Specificity
DCGAN synthetic data (50 % of original) + Translation	0.65 ± 0.02	0.87 ± 0.01	0.68 ± 0.04	0.86 ± 0.02
DCGAN synthetic data (50 % of original) + Rotation	0.69 ± 0.01	0.91 ± 0.06	0.69 ± 0.02	0.89 ± 0.04
DCGAN synthetic data (50 % of original) + Vertical flip	0.69 ± 0.02	0.90 ± 0.03	0.68 ± 0.02	0.91 ± 0.03
DCGAN synthetic data (100 % of original) + Translation	0.72 ± 0.04	0.89 ± 0.01	0.74 ± 0.02	0.96 ± 0.02
DCGAN synthetic data (100 % of original) + Rotation	0.76 ± 0.02	0.93 ± 0.02	0.78 ± 0.03	0.89 ± 0.01
DCGAN synthetic data (100 % of original) + Vertical flip	0.77 ± 0.03	0.91 ± 0.01	0.76 ± 0.04	0.87 ± 0.01
DCGAN synthetic data (Balanced classes) + Translation	0.77 ± 0.02	0.92 ± 0.03	0.79 ± 0.01	0.88 ± 0.04
DCGAN synthetic data (Balanced classes) + Rotation	0.79 ± 0.01	0.92 ± 0.04	0.77 ± 0.02	0.89 ± 0.02
DCGAN synthetic data (Balanced classes) + Vertical flip	0.79 ± 0.03	0.89 ± 0.01	0.78 ± 0.02	0.93 ± 0.04

Table 10. Results of simple augmentation along with synthetic samples.

5.3 **Contribution C**: *One class to rule them all: Detection and classification of prostate tumours presence in bi-parametric MRI based on auto-encoders.*

In ***contribution C*** [250] we tackle a similar problem as in ***contribution B*** [249]: prostate cancer lesion classification but from a different perspective (different approach): outlier detection with AEs. In addition, we also ***detect*** tumours besides classifying them. In particular, in this contribution we approach the data scarcity issue along with the imbalanced nature that is usually present in medical data (*Section 2.3*). In order to accomplish it, we

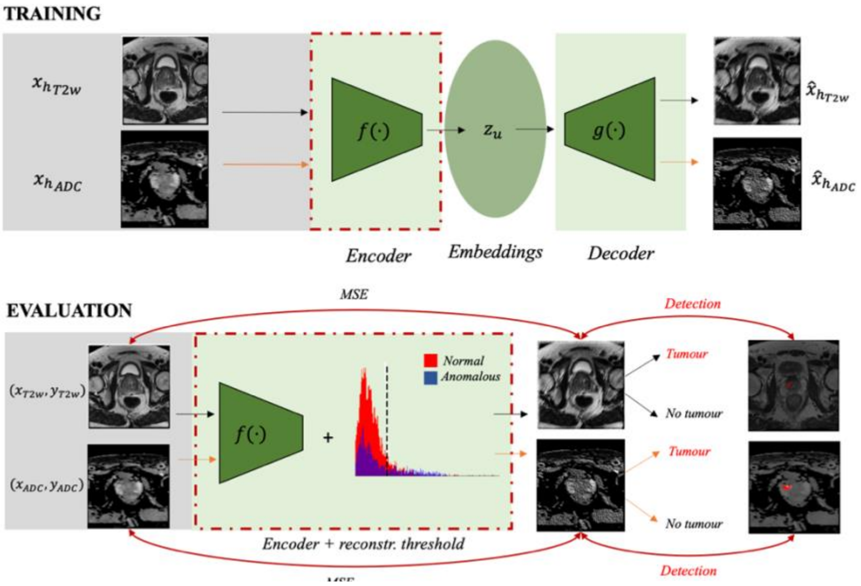


Figure 27. Learning the concept of normality with AEs.

exploit AEs to learn in an unsupervised way and based on the concept of "normal" to detect lesions and classify whether a prostate MRI slice contains one or not. We argue that specialists such as a radiologist are able to discern between normal (controls) and unhealthy cases after seeing a handful of normal cases (*Figure 27*) and even when no extensive training is present. Hereby, we aim to mimic that behaviour by exploiting the unbalanced nature of the data where a larger number of normal cases are available when compared to unhealthy ones.

We make use of the ProstateX [232] (*Section 4.1*) dataset with the same number of patients and lesions described in ***contribution B***. In addition, we also make use of the masks provided in [221] for the detection task. The proposed framework for ***contribution C*** makes use of two types of AEs: convolutional auto-encoder (cAE) and VAE to exploit the prevalence of slices without a lesion in the dataset. The general steps of the framework are depicted in *Figure 27*.

In this contribution, we work at the slice level (2D) with T2w and ADC prostate MRI. We start by applying some standard pre-processing to the data: re-sampling to 384x384 for T2w and 128x128 for ADC and normalization of the MRI intensities to a range of [0, 1]. In order to evaluate the detection and classification performance of the framework, we split the dataset in 70%/20%/10% by patients for training, validation and testing, respectively. We make use of all the "normal" (healthy or otherwise controls) cases available among the patients to train the AEs.

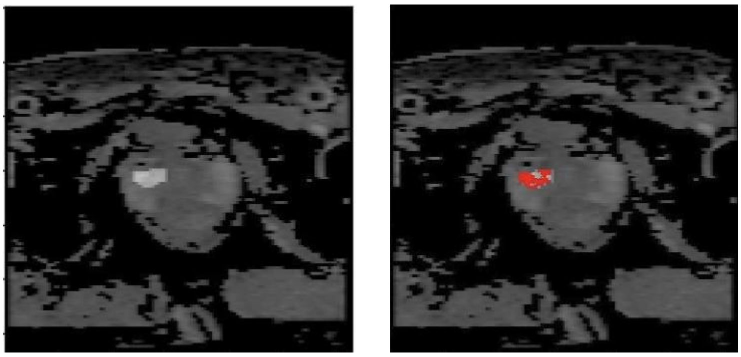
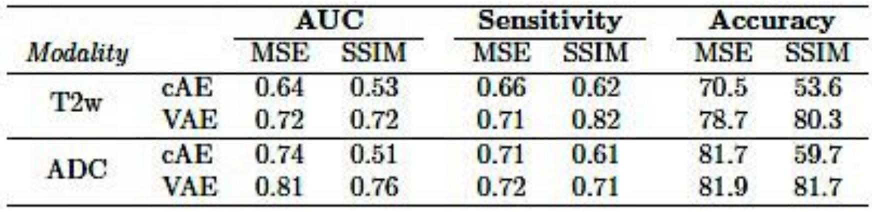


Figure 28. Lesion detection in prostate ADC MRI after thresholding.

As depicted in *Figure 27*, our framework consists of an AE (either cAE or VAE, we experiment with both) trained with “normal” cases such that the distribution of the normal slices is learnt. Once the distribution is learnt, we employ the previously trained AE architecture to obtain reconstructions of both healthy (no lesion present) and unhealthy (lesion present) prostate MRI slices. As the AE has only been trained on normal cases, we expect the reconstruction error of the unhealthy (non-normal) cases to be larger in the areas that are less similar to the healthy cases. That is, areas in which the lesion is located. Hence, we first conduct a classification of the slices based on the reconstruction error and those deemed as “unhealthy” are further analysed to find the areas in which the error is larger than another threshold (where the tumour is located). We compute the thresholds based on an interquartile range rule [251], which is commonly used in anomaly (outlier) detection works. Figure M shows an example of a lesion detected after applying a threshold found by means of IQR.

We evaluate our proposed framework in a quantitative way for the classification ability of it and in a qualitative way for the detection ability. In particular, we make use of AUC, sensitivity and accuracy as our metrics for the classification ability of the system. The same type of evaluation is carried out for both ADC and T2w, in addition to the investigation of the effect of cAE and VAE in the final results as well as the use of mean squared error (MSE) or structural similarity index (SSIM) as reconstruction metrics. As depicted in *Table 11* and as shown in [251], our framework achieves a good balance between false positives and false negatives and a higher AUC for T2w but competitive results for both MRI sequences when compared to other fully supervised approaches, as argued in [250]. Finally, our qualitative results show that the quality of the detection is reasonably good for the cases that are classified as unhealthy in the first step of the framework. An example of a correctly classified and further detected lesion is shown in *Figure 28*.



Modality		AUC		Sensitivity		Accuracy	
		MSE	SSIM	MSE	SSIM	MSE	SSIM
T2w	cAE	0.64	0.53	0.66	0.62	70.5	53.6
	VAE	0.72	0.72	0.71	0.82	78.7	80.3
ADC	cAE	0.74	0.51	0.71	0.61	81.7	59.7
	VAE	0.81	0.76	0.72	0.71	81.9	81.7

Table 11. Results of the proposed AE-based framework to detect and classify prostate lesions.

5.4 **Contribution D**: *Learning to triage by learning to reconstruct: A generative self-supervised learning approach for prostate cancer based on axial T2w MRI.*

Contribution D [252] is the first of the two contributions focusing on SSL to offer alternatives to the lack of annotated data. In particular, the main objective of the work is to stratify PCa lesions between ncS and cS ($GS < 7$ or $GS \geq 7$) in the presence of small amounts of labelled MRI data. To accomplish it, we propose a framework based on a generative SSL methodology in which the pre-text task is a reconstruction task (Section 3.4.2) from different image distortions applied at the patch level. As previously mentioned, the downstream task is a *binary classification task* aiming to discern between cS and ncS lesions.

To develop our generative SSL framework, we make use of the ProstateX [232] (*Section 4.1*) dataset with the same number of patients and lesions described in ***contribution B*** and ***contribution C***. Nevertheless, in this contribution we start by using *all the available data* without any annotation for the SSL approach, aiming to exploit the commonly available non-annotated data in the medical domain and, in particular, in PCa MRI. The general steps of the proposed generative SSL framework are depicted in *Figure 29*.

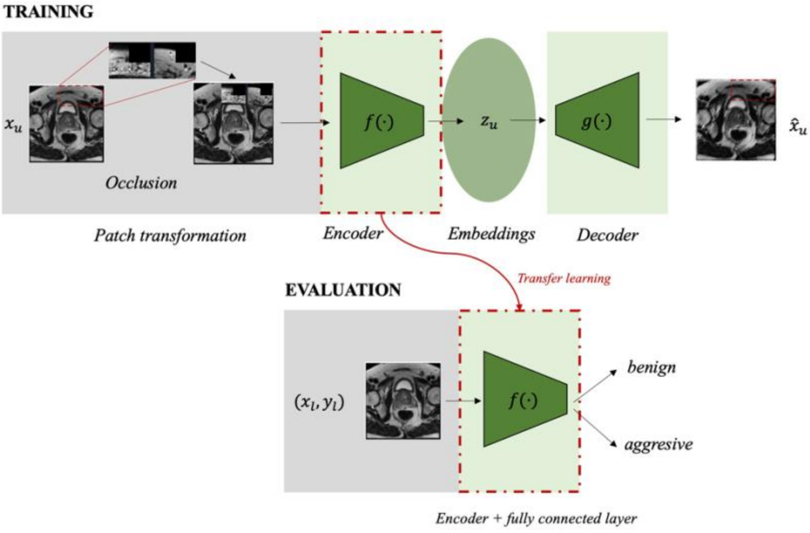
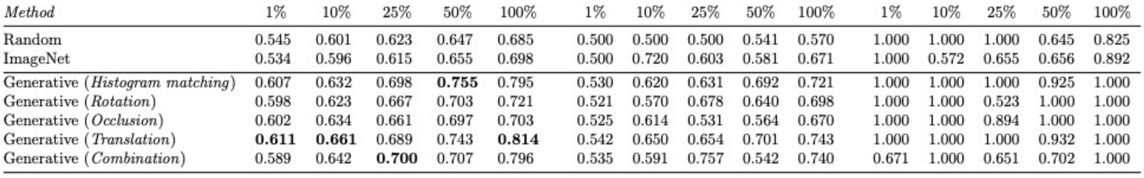


Figure 29. Proposed generative SSL framework.

As shown in *Figure 29*, we work at the slice level (2D) with T2w prostate MRI. As in previous contributions, we start by applying some pre-processing in the form of re-sampling to a common coordinate system with the desired dimensions (320x320) and normalization of the MRI intensities to a range of [0, 1]. In order to perform the evaluation of our proposed framework (downstream task), we split the original dataset following a 60%/20%/20% for training, validation and testing, respectively. The splitting is done by patients, such that no data leakage is present in our splits.

Our proposed framework consists of an AE which projects a *distorted* version of a T2w slice into a low-dimensional space through the encoder and tries to recover the original T2w MRI slice from that low-dimensional representation. Specifically, we obtain the distorted versions by sampling a transformation function from the following available ones: *patch histogram matching*, *patch rotation*, *patch occlusion* and *patch translation*. It is worth noting all the transformations are applied at the patch level (blocks of 64x64). All the transformations are applied to $N = 2$ randomly selected patches with a probability of $p = 0.5$. The different transformations are chosen on the basis of previous works [200] and thought to be useful for the final downstream task. That is, transformations that might preserve to some degree the label associated to the data in the supervised evaluation. Once the framework has been trained, we transfer the weights to perform the evaluation in the downstream task.



Method	1%	10%	25%	50%	100%	1%	10%	25%	50%	100%	1%	10%	25%	50%	100%
Random	0.545	0.601	0.623	0.647	0.685	0.500	0.500	0.500	0.541	0.570	1.000	1.000	1.000	0.645	0.825
ImageNet	0.534	0.596	0.615	0.655	0.698	0.500	0.720	0.603	0.581	0.671	1.000	0.572	0.655	0.656	0.892
Generative (*Histogram matching*)	0.607	0.632	0.698	**0.755**	0.795	0.530	0.620	0.631	0.692	0.721	1.000	1.000	1.000	0.925	1.000
Generative (*Rotation*)	0.598	0.623	0.667	0.703	0.721	0.521	0.570	0.678	0.640	0.698	1.000	1.000	0.523	1.000	1.000
Generative (*Occlusion*)	0.602	0.634	0.661	0.697	0.703	0.525	0.614	0.531	0.564	0.670	1.000	1.000	0.894	1.000	1.000
Generative (*Translation*)	**0.611**	**0.661**	0.689	0.743	**0.814**	0.542	0.650	0.654	0.701	0.743	1.000	1.000	1.000	0.932	1.000
Generative (*Combination*)	0.589	0.642	**0.700**	0.707	0.796	0.535	0.591	0.757	0.542	0.740	0.671	1.000	0.651	0.702	1.000

Table 12. Results of the proposed SSL generative framework in a linear evaluation setting.

We evaluate our framework in a *linear evaluation setting*. That is, we keep the transferred weights frozen such that we can obtain a proxy of the quality of the obtained representations during the training of the SSL framework. Furthermore, we evaluate the results with different % of labelled data to assess the robustness of the presented methodology in the presence of heavily scarce datasets. As depicted in *Table 12*, our proposed methodology achieves better results than both random initialization and ImageNet initialization for different fractions of labelled data and in a linear setting. Specifically, our method outperforms both initializations in the presence of extremely scarce datasets (1%) and when the original number of small amounts of labelled data is present (100%), showing the quality of the representations learnt by the generative SSL method. Both translation and histogram matching are the best performing patch manipulations overall, depicting the relevance of the patch position and pixel intensities for the lesion classification task.

5.5 **Contribution E**: *Contrasting axial T2w MRI for prostate cancer triage: A self-supervised learning approach.*

In ***contribution E*** [253] we approach the data scarcity problem with another SSL-based framework. In this case, we move on from "manually defined" pre-text tasks (***contribution D*** [252]) and apply a contrastive approach. That is, we aim to learn the concept of *similarity* among samples. Specifically, the application of this contribution is PCa lesion classification at the slice level (2D) between cS and ncS ($GS < 7$ or $GS \geq 7$), defined as the downstream task of the contribution.

We develop the contrastive SSL framework depicted in *Figure 30* by making use of T2w ProstateX [232] (*Section 4.1*) dataset with the same number of patients and lesions described in ***contribution B*** and ***contribution C***. Nevertheless, in this contribution we start by using *all the available data* without any annotation for the SSL approach, aiming to exploit the commonly available non-annotated data in the medical domain and, in particular, in PCa MRI.

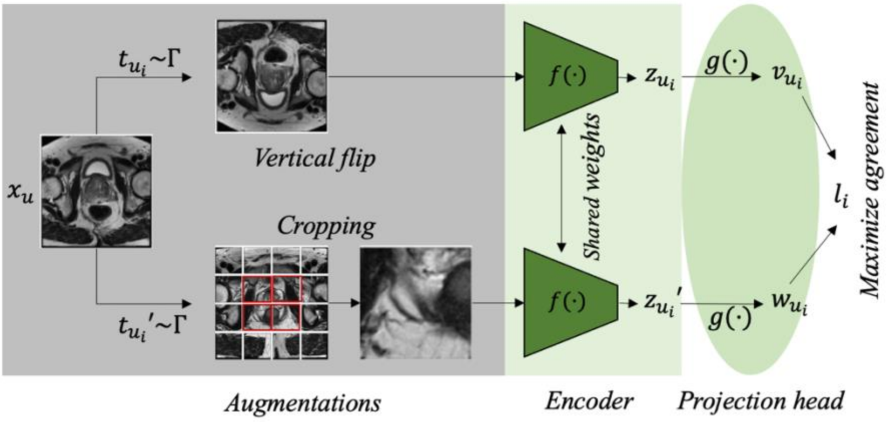
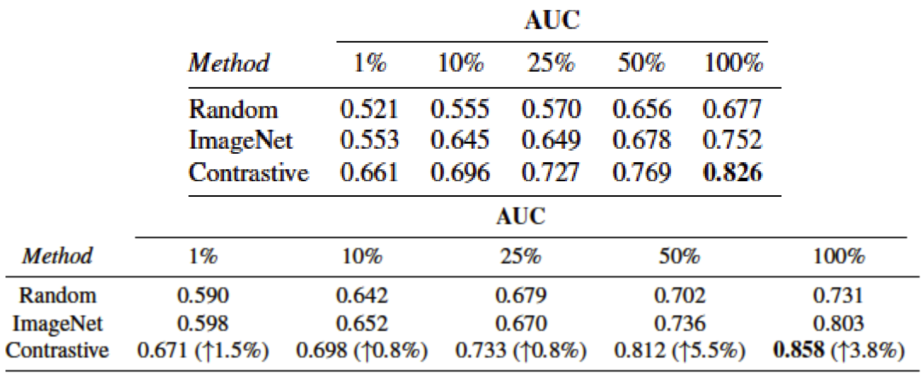


Figure 30. Proposed contrastive SSL framework.

The general steps of the SSL contrastive framework are depicted in *Figure 30*. In a similar fashion to ***contribution D*** [252], we apply some pre-processing in the form of re-sampling (resolution of 0.5 x 0.5 x 0.5 mm), normalization of MRI intensities to a range of [0,1] and outlier removal. We work, again, at the slice level (2D) with T2w MRI sequences of the prostate. To evaluate our proposed contrastive SSL framework in terms of the performance in the downstream task we split the original labelled dataset by patients following a 60%/20%/20% for training, validation and testing, respectively.

The contrastive SSL framework consists of an encoder (VGG16) that maximizes the agreement between different augmented views of the same data example using a contrastive loss [216] in the embedding space. To obtain the augmented views, we choose a family of image manipulations which we deem suitable to generate views without losing the interpretability of the diagnosis in the MRI slice under consideration. In particular, we make use of rotation (50 degrees), translation (range of 0.32 pixels), vertical flipping and cropping. We also experiment with different embedding dimensions (*Figure 30*) and different image resolutions.

We evaluate our framework in a *linear evaluation setting* and in a *fine-tuning setting*. In the first case, the pre-trained weights obtained from the contrastive approach are kept frozen and a randomly initialized linear head is trained for the task under consideration. This particular evaluation protocol is intended to give an idea of the quality of the learned features and their re-usability. On the other hand, in the fine-tuning scenario the whole encoder is unfrozen and the entire model is fine-tuned end-to-end. Both evaluation protocols are carried out with different fractions of labelled data in order to test the robustness of the approach when dealing with a limited amount of data, as a proxy for the real world.



Method	AUC 1%	10%	25%	50%	100%
Random	0.521	0.555	0.570	0.656	0.677
ImageNet	0.553	0.645	0.649	0.678	0.752
Contrastive	0.661	0.696	0.727	0.769	**0.826**

Method	AUC 1%	10%	25%	50%	100%
Random	0.590	0.642	0.679	0.702	0.731
ImageNet	0.598	0.652	0.670	0.736	0.803
Contrastive	0.671 (↑1.5%)	0.698 (↑0.8%)	0.733 (↑0.8%)	0.812 (↑5.5%)	**0.858** (↑3.8%)

Table 13. Results of the proposed SSL contrastive framework in a linear evaluation setting and fine-tuning one.

As depicted in *Table 13*, our contrastive SSL approach outperforms both ImageNet and random initialization for the different fractions of labelled data. Specifically, we can observe that our contrastive SSL approach outperforms the other initialization methods in the presence of really small amounts of data, supporting the hypothesis that the representations obtained with the SSL approach benefit the training of the downstream task in small data regimes. Regarding end-to-end fine-tuning, the proposed methodology outperforms by a large margin both ImageNet and random initializations and obtains better results than the linear setting for all the configurations.

5.6 **Contribution F**: *Multi-planar T2w MRI for an improved prostate cancer lesion classification.*

In ***contribution F*** [254], we tackle the data scarcity issue by making use of all the available data sources that are commonly acquired by default during MRI examination of the patients in PCa. Specifically, we argue that most of the works in PCa focus exclusively on the axial view of T2w whilst both coronal and sagittal are also available. Hence, we aim to exploit those views along with the axial one to improve the task under consideration. In this case, we focus on lesion classification and discriminating between cS and ncS cases, in a similar fashion to previous contributions.

We propose two different methodologies to make use in an efficient way of the different orthogonal views: simple multi-stream fusion (siMS) and inter-connected multi-stream (icMS), as depicted in *Figure 31*. Specifically, in the first case we propose three independent encoders to process the different views and a late fusion approach (by concatenation of the feature maps) once they have been processed. In the second case [254], dense connections are added to the chosen architectures along with connections between the convolutional blocks of each stream such that the fusion of information does not only happen at an early or late stage but rather in the whole architecture, allowing the network to have more freedom to learn more complex and abstract combinations between the sequences.

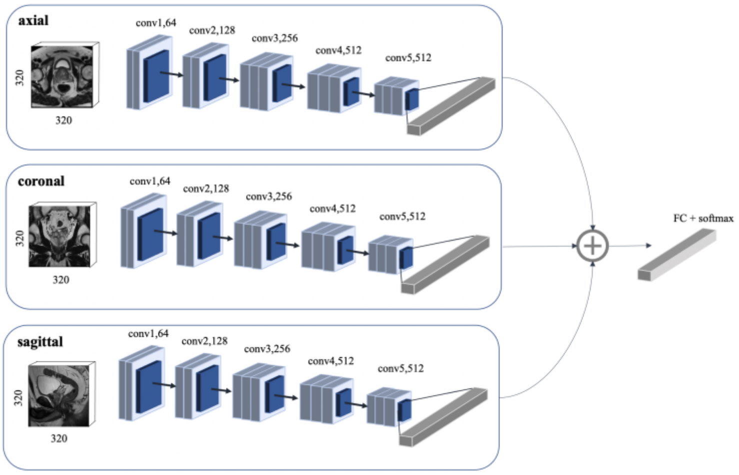
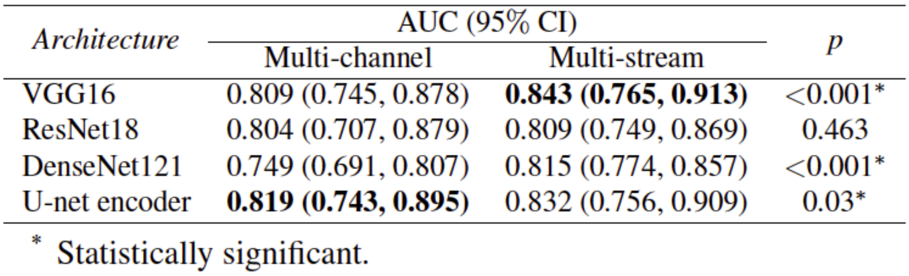


Figure 31. Proposed simple late fusion multi-stream approach.

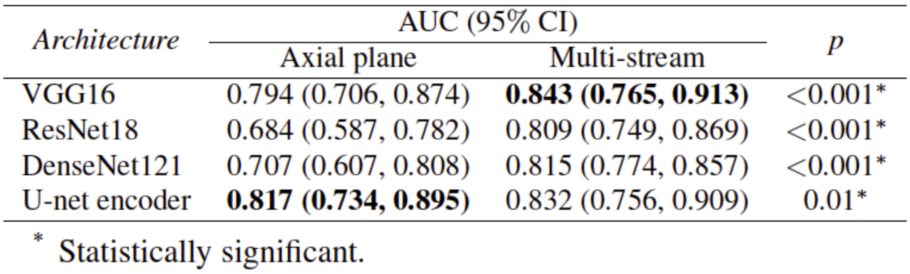
We develop the fusion framework depicted in *Figure 31* by making use of all the views available for T2w prostate MRI of the ProstateX dataset [232] (*Section 4.1*). The number of patients and lesions is the same as the one described in previous contributions. We apply some pre-processing in the form of data normalization, outlier removal and cropping along with re-sampling. Following, the original dataset is split by patients following a 60%/20%/20% for training, validation and testing.

We evaluate our results based on 95% confidence intervals (CI) obtained with $n = 100$ bootstrap replicates and quantify the different found between the results with Wilcoxon signed-rank test. Specifically, we compare the performance of our siMS approach against an axial-only approach and a multi-channel one (accommodating the different views in a 3-channel way) for different convolutional architectures. Our results showcase that by making use of all the available directions and siMS approach we are able to improve the performance of PCa lesion classification for diverse architectures (*Table 14*). In addition, we find out that for the best architectures (*Table 14*) fusing the features at different levels outperforms a simple late fusion approach showing the potential for multi-view data.



Architecture	AUC (95% CI)		*p*
	Multi-channel	Multi-stream	
VGG16	0.809 (0.745, 0.878)	**0.843 (0.765, 0.913)**	<0.001*
ResNet18	0.804 (0.707, 0.879)	0.809 (0.749, 0.869)	0.463
DenseNet121	0.749 (0.691, 0.807)	0.815 (0.774, 0.857)	<0.001*
U-net encoder	**0.819 (0.743, 0.895)**	0.832 (0.756, 0.909)	0.03*

* Statistically significant.



Architecture	AUC (95% CI)		*p*
	Axial plane	Multi-stream	
VGG16	0.794 (0.706, 0.874)	**0.843 (0.765, 0.913)**	<0.001*
ResNet18	0.684 (0.587, 0.782)	0.809 (0.749, 0.869)	<0.001*
DenseNet121	0.707 (0.607, 0.808)	0.815 (0.774, 0.857)	<0.001*
U-net encoder	**0.817 (0.734, 0.895)**	0.832 (0.756, 0.909)	0.01*

* Statistically significant.

Table 14. Results of the multi-stream approach (siMS) compared against axial only and multi-channel input.

6 Discussion, future work, work in progress and conclusions

6.1 Discussion

The common underlying factor of all the contributions presented in this work is the data scarcity issue, tackled from different angles. In spite of the differences between the different methodologies presented in this work, my belief is that it is relevant to put all the work in a general context and add a general discussion evaluating the different contributions and their impact from the common factor perspective: data scarcity in DL and PCa. Hence, in this section, I will present a specific discussion for each of the contributions and then build on that to present a general discussion.

Synthetic data augmentation*:* In ***contribution A*** [120] and ***contribution B*** [249], we focus on synthetic data augmentation by making use of GANs. In both cases, our results show a positive effect of the synthetic augmentation at training time by obtaining better results in terms of segmentation quality and classification ability, respectively. We develop both works with open access data (PROMISE12 and ProstateX, respectively – Section 4.1) which were made available with the main objective of providing a forum and a common ground to compare results for different PCa applications in the form of a challenge (competition). In particular, both challenges have a publicly accessible leader board which showcases the scores obtained from all the submitted solutions to the challenges[2]. When evaluating our results based on the public scores that can be found in such resources, we can observe that both leading scores (91.90 DSC and 0.95 AUC) are significantly larger than the best results presented in our contributions (73.90 DSC [120] and 0.79 AUC [249], respectively). Nevertheless, it is important to keep in mind when comparing the results that in our contributions the test set provided by the competition -and the one used by the competitors to obtain the results of the challenge board- which remains hidden until a formal evaluation of the method is performed by the organizers of the challenge has not been used but rather, we used an internal test set

[2] https://prostatex.grand-challenge.org/evaluation/challenge/leaderboard/
https://promise12.grand-challenge.org/evaluation/challenge/leaderboard/

extracted by dividing the training set provided by the competition. Hence, even if all the appropriate tools were applied to avoid big deviations of the results in the test set, small variations should be expected if the hidden test was used. Moreover, as the test set is not exactly the same, the results of the other participants could also differ from the ones presented in the board.

The results presented in our work can also be put into a more general context by comparing them with different approaches found in the literature and that might, potentially, not have submitted their solution to the challenge or used internal datasets to develop and test the methodologies. In that regard, Aldoj *et al.* [255] present a solution based on U-net with dense connections which obtains a final DSC of 92.1 ± 0.8 (four-fold cross-validation results) in an in-house dataset. In the work of Tian *et al.* [256] the authors fine-tune a fully CNN and reach a final DSC of 85.3 ± 3.2, by means of a cross-validation procedure. Other works such as the Wang *et al.* [257] present an architecture able to process 3D data along with some modifications such as strided convolutions. The final results show improvements over the introduced baselines on an internal dataset, reaching a DSC of 86.12 ± 0.4 an of 88.02 ± 0.5 for PROMISE12. In the work of Zhu *et al* [258] the authors introduce deep supervision in a standard U-net architecture, reaching a mean DSC of 88.5 and showing improvements over a standard U-net architecture.

Lesion classification approaches have also been proposed in the literature: Ishioka *et al.* [259] make use of a ResNet-like architectures to classify PCa lesions in 2D, reaching an AUC of 0.777 and 0.793 in two different external data-sets (available for in-house development). In Wang *et al.* [260] the authors use a VGG-like architecture and obtain an AUC of 0.84 in the lesion classification task, with significant differences ($p < 0.001$) when compared to other non-DL-based methods.

We can clearly observe how the different approaches presented in the literature obtain better results than our proposed approach. Nevertheless, once again, stablishing fair comparisons is hard based on the fact that most of the works use internal (in-house) data to develop and validate their methodology, making the comparison unfair as there is no common quantification (based on the same data) on how the methods perform under the same conditions. Nevertheless, the proposed methodology is intended to have a generic nature. That is, we move away from specific architecture modifications [255, 256] and instead, we propose an improvement that arguably, has an “add-on” capability which makes it suitable for a wide range of architectures in both ***contribution A*** [120] and ***contribution B*** [249]. Specifically, our proposed frameworks could be integrated in the previously presented works and potentially boost the final results thanks to the increase in data availability.

In terms of alternative approaches based on GAN generation for PCa, few alternatives have been proposed. Yu et al. [261] propose a GAN-based architecture to generate diverse prostate MRI images (T2w, DCE and ADC). The proposed GAN architecture introduces a modified discriminator which aims to obtain more equivariant features. The authors evaluate the quality of the generated images based on a divergence metric as well as in terms of a classification task. In particular, the authors extract features from the generated images and use them as an input for a ML classifier, showing improvements in the final results thanks to the generated images. In the work of Hu et al. [262] the authors combine a cGAN and DCGAN to generate DWI prostate MRI of different GS, in order to mitigate data bias in prostate applications. The authors provide a qualitative analysis of the results, showing from a visual perspective the generated images have a reasonable quality and have characteristics that are comparable to the original images used to train the architecture. Finally, in Wang et al. [263] the authors propose a GAN-based framework to generate cS and ncS prostate ADC. In particular, a novel layer is introduced in the framework to boost the quality of the final results. The authors quantify the quality of the generated pictures by means of classification accuracy in the discrimination ability of a CNN between cS and ncS prostate ADC MRI.

Our contributions differentiate from the presented ones in different ways: ***contribution A*** not only generates data but also *synthesizes* data from the generated masks, focusing on a segmentation task rather than classification (like most of the previously presented works). On the other hand, ***contribution B*** focuses on a classification task like the other works. However, our evaluation protocols are DL-based instead of radiomics-based [261]. Furthermore, our generated images are obtained with a higher resolution than the work presented in [263], hence avoiding up-sampling procedures and blurriness or lack of image quality derived from it. When compared to [262], we present a detailed evaluation based on classification results and not only visual quality of the generated images.

Our work presents several limitations. The first limitation comes from the nature of the data used to develop and evaluate both contributions: retrospective nature. Prospective studies should be carried out to verify the validity of the results and certify the usefulness of the approaches in real-world scenarios. Furthermore, in both cases, only one cohort has been used to quantify the results. External validation could enhance the study by providing proof of the robustness of the approach when tested with cohorts with different characteristics (i.e., acquisition parameters). In the same line, cross-validation or bootstrapping would also increase the quality of the study as of right now results are not averaged and based on one single split and no information about their variability is provided. Regarding the choice of metrics and quantification of the results, both methodologies present an indirect way of quantifying the quality of the generated synthetic samples via the final task (segmentation and classification) but a more direct and quantitively way of evaluating the results would most probably benefit the final results and the insights obtained from them (in order to improve the generation process).

Learning the concept of normality**:* In ***contribution C [250] we focus on an application of AEs to learn to "detect outliers" (or the concept of normality) by exploiting the imbalance present in medical datasets, where control data (healthy) is commonly prevalent when compared to the data presenting a specific pathology (unhealthy). In a similar fashion to previous contributions (***A and B***) we make use of publicly available data (ProstateX) to develop the AE-based framework[1]. As such, we can stablish comparisons by making use of the challenge scoreboard. Nevertheless, in this case, the main objective is to detect and classify *the presence* of tumours in prostate MRI slices (2D) and as such, the comparisons with the submissions to the challenge do not apply, as they focus on a different task. Importantly, in the event that such a challenge provided results for a similar task to the one presented in the work, we would need to be careful when stablishing the comparisons as in a similar way to the previously presented discussions we do not make use of the hidden test set and hence, our results in an external set could deviate from the ones presented in the work.

Looking at our results from a more general perspective, we can find several works in the literature focusing on the detection of prostate tumours and PCa classification from a supervised perspective. In particular, in the work of

Mehralivand *et al*. [264] a segmentation model is presented to obtain lesion segmentations with a DSC of 30.7 for the validation set compared to a plain U-net that obtains a DSC of 28.7. Xu *et al*. [265] presents a study using a ResNet to classify the presence of tumours in a fully supervised way and by making use of a patch-based approach to train the network, obtaining an AUC of 0.97 at the patch level. Aphinives *et al*. [266] present a study based on T2w prostate MRI in which they aim to detect prostate lesions in a fully supervised fashion and obtain a mean average precision of 13.1% and prediction rate of 31.58%. Finally, in the work of Saha *et al*. [267] the authors explore the effect of attention mechanisms, clinical priori and decoupled false positive reduction in PCa lesion detection, reaching a 0.882 ± 0.030 AUC in patient-based diagnosis (in 3D).

Based on the results that we can find in the literature; our results are slightly worse than the worse AUC presented in the small literature review (0.81 vs 0.882). In terms of lesion detection ability, as our work evaluates the task in a qualitative way instead of quantitative one it is hard to stablish a fair comparison between our results and the ones presented in the previous discussion (based on DSC or mean average precision). Similarly, to the previously presented discussion, it is hard to discuss and evaluate in a fair way our presented work in terms of the other contributions as most of them introduce in-house datasets for the development and validation of the algorithms. Nevertheless, in spite of the differences found in the final performance in terms of the metrics used to quantify them, an important factor to take into consideration is that our algorithm is trained *exclusively* on healthy (control) data and reaches the presented results in an *unsupervised fashion* (for the unseen class during training). Moreover, most of the previously discussed works require specialized architectures for both classification and detection, whilst our proposed framework is capable of obtaining both results with a single unspecialized AE-based architecture. The joint effect of both characteristics (single class training and single unspecialized architecture) is a large reduction in computational requirements, which can greatly benefit future real-world deployments of the application under consideration.

In terms of alternatives in the literature, there are no alternatives following a similar procedure to the one presented in our work for PCa. The closer alternatives to the one presented in this study exploit AEs to extract features (as a feature encoder) such as the one presented in the work of Abraham *et al.* [268]. Hence, our approach can be considered quite novel in the area of PCa and shows potential to be further developed.

Our work presents several limitations, being the retrospective nature of the data used to develop the framework one of them. In a similar fashion to previous contributions, prospective studies would be required to verify the validity and usefulness of the presented approach in a real-world scenario. Additionally, only one cohort was used to develop and evaluate the developed framework and further evaluation with external datasets should be carried out to account for potential data drifts in the form of different acquisition protocols or population characteristics. A more robust evaluation process would strengthen the validation of the study by quantifying the variability of the results and the uncertainty around them. In terms of specific details of the work, a quantitative way of evaluating the detection of the PCa lesions could enhance the work by providing a way to compare our contribution with (potential) other contributions that make use of the same data and have the same objective as our work. Another potential limitation of our work is the threshold implementation, as the threshold is not chosen on the basis of the final objective (i.e. detection or classification) but rather only on a general rule (IQR) which might not be optimal of the specific objectives of the work.

Self-supervised learning*:** Both ***contribution D [252] and ***contribution E*** [253] focus on SSL approaches to tackle the data scarcity issue in PCa applications. Specifically, ***contribution D*** focus on generative SSL whilst ***contribution E*** presents an approach based on contrastive SSL. In both cases, we aim to move away from traditional TL methods based on out-of-domain data (ImageNet) and instead, provide an in-domain initialization method based on the available unannotated data which might be robust for different applications and in the presence of small amounts of labelled data. In particular, we evaluate both applications in a 2D PCa lesion classification setting (Section 5). In both cases, we make use of the same publicly available dataset as in previous contributions: ProstateX (Section 4.1). As mentioned in previous contributions, the dataset has a public leaderboard for the task of lesion classification in cS and ncS[1]. The best score in the public leader board is of 0.95 (AUC), whilst we obtain an AUC score of 0.858 for ***contribution E*** [253] in a fine-tuning scenario and of 0.814 for ***contribution D*** [252] in the best transformation scenario and without fine-tuning. The best scores are, as seen in the case of ***contribution B*** [249] significantly larger than the ones obtained in our proposed SSL frameworks. Nevertheless, as mentioned in previous discussions, the comparison needs to be taken with caution as we do not make use of the hidden testing set but rather only the fully available training set for evaluation purposes. Hence, the results obtained with the hidden test set could differ from the ones presented in the work whilst the ones presented by the different algorithms proposed for the challenge could also differ when evaluated with the same data as the one used in our work.

When looking for contributions on PCa SSL applications to compare our work with, we find the work of Bolous *et al.* [269] which falls in the category of generative SSL. In particular, the authors present an SSL framework using context restoration as a pre-text task and evaluate the quality of the representations on a lesion detection task and showing improvements over a plain U-net trained from scratch in the presence of different % of labelled data, reaching a DSC of 0.57 and an AUC of 0.85. In the work of Qian *et al.* [270] the authors present a generative SSL approach based on reconstruction from a distorted version of the prostate MRI (by injecting noise) as previous step to the chosen downstream task: prostate segmentation. The results show an improvement by making use of the SSL approach in the detection of PCa lesion in T2w MRI, reaching a true positive ratio of 0.9182.

Similarly to the contributions present in the literature, our work is also able to improve the final downstream task results. Nevertheless, the downstream task nature is different as in our contributions we aim to stratify prostate lesions depending on their clinical significance and the SSL contributions in PCa have as an ultimate objective the detection of prostate tumours. Hence, a direct comparison in terms of final results cannot be established. However, the approaches can be compared in terms of methodology. Our SSL generative framework differentiates from the other approaches by introducing a variety of image manipulations other than noise injection or corruption and therefore, having the ability to learn more general features. Specifically, our proposed generative SSL framework applies the image manipulations at the patch level under the assumption that operating at a “lower level” might help to learn more discriminative features for the downstream classification task as lesions present in the prostate are relatively small. In terms of the contrastive approach, we present an SSL framework which does not depend on a "manually" designed pre-text task and hence, has a more generic nature than the contributions present in the literature and our generative SSL framework. In fact, when comparing our two SSL contributions that would be one of the key factors that differentiates them: the nature and design of the pre-text task, making the contrastive contribution more generic than the generative one where the pre-text task is, in fact, "supervised" to be more efficient for the downstream task under consideration.

In terms of limitations, both contributions present limitations due to the retrospective nature of the data used to develop the frameworks. Furthermore, both works lack external validation of the results to account for potential data shifts due to acquisition protocols or population characteristics. In addition, a more robust internal validation based on statistical testing and quantification of uncertainty around the results (see ***Contribution F***) would strengthen the final results. In particular, for contribution D, we also miss a fine-tuning setting in which results are obtained when the weights are unfrozen. More experimentation with a different number of manipulated patches or other basic manipulations would also benefit the final work by providing more results in the form of a sensitivity analysis and the effect of the amount of distortion on the final task. A more extensive comparison with other SSL approaches (see MoCo [218]) would also help to understand the effect of the chosen SSL approach and strengthen the hypothesis of the effectiveness of the methods when compared to its counterparts (for both contrastive and generative SSL methods).

Efficient use of different data sources & data fusion: In ***contribution F*** [254], we tackle the data scarcity issue from the point of view of an efficient use of the available data. In particular, we propose to make use of the different orthogonal views (axial, sagittal and coronal) which are acquired by default during the patients' visit instead of making use of only the axial view, which is the one used by default. We test our approach in a lesion classification setting in 2D, as in previous contributions (***contribution B, D, and E***). In a similar fashion, we develop the work by making use of the ProstateX dataset. When comparing our results to the public leader board, we can see a considerable difference in terms of the final AUC: 0.854 vs 0.95. Nevertheless, as in previous works, we make use of the training set exclusively (after proper division in training, validation and testing as explained in Section 5). Hence, a direct and fair comparison cannot be established due to validation differences in the dataset chosen for it. When compared to other lesion classification approaches present in the literature (discussion of ***contribution B***) our fusion approach obtains competitive results, obtaining better results than the ones presented in [259, 260].

Different fusion approaches can be found in the literature for prostate MRI. For instance, in the work of Lozoya *et al.* [272] and in the work of Meyer *et al.* [135] independent 2D CNNs are trained and fused at a late stage or ensembled to produce results that do not depend exclusively on axial plane but rather that exploit the information of the different views with the objective of improving prostate segmentation results. In the works of Yuan *et al.* and Le *et al.* [273, 274] the authors present early fusion approaches for different modalities of prostate MRI with the objective of improving PCa lesion classification results in 2D and 3D, respectively. In particular, in the work of Yuan *et al.* a final AUC of 0.89 is obtained whilst in the work of Le *et al.* a final AUC of 0.91 is obtained.

When comparing our contribution with the ones found in the literature, our results are significantly lower than results obtained with the same dataset such as in Le *et al.* [274]. Nevertheless, our procedure to obtain the fusion approach is way simpler than the one presented in other works which incorporate extra elements in the training of the network and in the pre-processing steps, considerably increasing the required computational power. Specifically, we also propose an inter-connected approach which allows to share features at different levels of the network instead of just making use of the features at an early or late level, like the approaches presented in the literature. Our inter-connected approach is generic enough such that it could be incorporated as an add-on to other fusion approaches (such as in Le *et al.)* potentially providing an extra improvement to the proposed solutions.

In terms of limitations, the work suffers from the retrospective nature of the data used to develop and evaluate the model and prospective evaluations would be required to validate the presented results. External validation should also be performed to evaluate potential data shifts and biases due to different acquisition protocols or population characteristics. Additionally, our method requires some pre-processing steps such as non-rigid registration which limits the generic and simplistic nature of the presented methodology. Hence, other fusion approaches which do not require any particular registration could be explored to further improve the final results.

We have presented a specific discussion and review of each of the contributions included in the work. From a more general perspective, we tackle the data scarcity issue from different angles. In spite of focusing on the same final objective (boosting the results in the presence of small amounts of data) the approaches are significantly different. **Contribution B, D, E** and **F** evaluate the efficiency of the approach from the point of view of lesion classification. Comparing the final results, we can observe how both contribution E and F obtain the highest AUC (0.858 and 0.854) in the presence of all the available testing data. Nevertheless, as **contribution F** does not evaluate the final results in terms of different fractions of data it is hard to determine the effectiveness of the method when an extreme data scarcity is present in the application under consideration. When evaluating the approaches based on the methodology, GAN-based methodologies offer an interesting alternative to increase the amount of data by generating new samples but at the cost of significant computational resources and instability in training time (in terms of reaching an optimal point in the loss landscape) along with struggles to generate high-resolution images, which is quite an important factor in the medical domain depending on the task under consideration [274].

Given such limitations, one could argue that GANs are not ready for prime time in the radiological domain but offer a promising (and interesting) alternative in data-limited scenarios. AE approaches are an interesting alternative to exploit what could look as a weakness at first sight: data imbalance. Nevertheless, based on the results obtained by means of other approaches (SSL and data fusion) and the “manual design” of the threshold-based methodology one can also argue that AE-based outlier detection is behind in terms of potential in data-limited scenarios when compared to approaches such as SSL. Furthermore, the principles applied in the AE scenario (reconstruction task and then detection and classification) form, in a way, the basis of the generative SSL approach presented in this thesis with the exception that they omit the “manual design” step and replace it with a TL approach.

The two most promising research directions based on the results of this thesis and previous arguments are, therefore, SSL approaches and efficient use of data by means of fusion. In particular, generative SSL approaches offer an interesting alternative by allowing to manually design a pre-text task which can, in a way, be conceived to be optimal for the final downstream task. In other words, the pre-text task can be "supervised" in terms of the evaluation task under consideration. Nevertheless, such a feature can be a double-edged sword as designing an optimal pre-text task might not be straightforward. The alternative are contrastive SSL approaches that have a "fixed" pre-text task and aim to learn the notion of similarity. Interestingly, in spite of having less degrees of freedom when it comes to the choice and design of a pre-text task, the efficiency of the approach is not compromised, as shown in our results and other results presented in the literature (Section 3.5). Nevertheless, further testing is required as contrastive SSL methods span a wide range and other methods might be more suitable for the context of medical images but might remain untested, as the field rapidly progresses. Finally, an efficient use of the available data is also a promising research direction as fusing different data sources (and even when the characteristics are similar) is challenging. However, as shown in our contribution [240] by moving away from simplistic fusion approaches such as early and late we can significantly improve the final results in the task under consideration. Furthermore, the fusion of information can be jointly used with SSL schemes making both approaches complementary and potentially improving the results obtained by making use of them in an independent way.

6.2 Conclusions

The common core of this thesis is the application of DL in data-limited scenarios in PCa. We present a variety of techniques to deal with the lack of annotations in the context of PCa and for different PCa applications: lesion classification, lesion detection and prostate segmentation. We show the effectiveness of the different approaches for data-limited situations in different applications and even in the presence of an extreme data scarcity (***contributions D and E***). Specifically, we provide some evidence that highlights the usefulness of alternatives that provide in-domain initializations for the medical imaging domain in PCa and that the findings and effects found and observed in natural images applications can also be translated to medical images, opening an interesting research path and laying the grounds for future research with more complex and tailored methods for an efficient training and deployment of DL-based systems in PCa based on MRI, moving away from the need of extremely large annotated datasets.

6.3 Work in progress and future work

Some of the future directions (and already work in progress) of the work based on this thesis are:

- Researching multi-modal approaches (efficient data usage) in the context of SSL methods.
- Research more complex fusion methodologies which, for instance, do not require previous registration steps.
- Investigate simpler in-domain initialization methods based on patches, based on the effectivity of them when training architectures such as Transformers [275].
- Explore the effect of making use of the inherent 3D nature of MRI data in our already developed applications or new ones.
- Improve evaluation protocols of the presented works by testing them with external data and in-house data. Include patient-level decisions, as they are the natural "dimension" in which radiologists work.
- Translate our research into clinical practice and carry out prospective studies to validate our results in a clinical scenario.

Finally, it is worth mentioning ***this work will continue during a postdoctoral position funded by HelseVest***, as I obtained funding to continue working at the intersection of PCa and DL. Future work will include part of the future plans highlighted in this section along with the task of translating and evaluating the impact of DL applications in clinical (radiology) practice. Furthermore, we will have an in-house dataset (Stavanger university hospital), which will be used along with the ones that have already been used to develop this work and that has already been collected and will be ready to use in the Postdoc position.

Bibliography

[1] Siegel, R.L., Miller, K.D. and Jemal, A., 2020. Cancer statistics, 2020. Ca-a Cancer Journal for Clinicians, 70(1), pp.7-30.

[2] Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians. 2018 Nov;68(6):394-424.

[3] Rawla P. Epidemiology of prostate cancer. World journal of oncology. 2019 Apr;10(2):63.

[4] Quon H, Loblaw A, Nam R. Dramatic increase in prostate cancer cases by 2021. BJU international. 2011 Dec;108(11):1734-8.

[5] Naitoh J, Zeiner RL, Dekernion JB. Diagnosis and treatment of prostate cancer. American family physician. 1998 Apr 1;57(7):1531.

[6] Wolf AM, Wender RC, Etzioni RB, Thompson IM, D'Amico AV, Volk RJ, Brooks DD, Dash C, Guessous I, Andrews K, DeSantis C. American Cancer Society guideline for the early detection of prostate cancer: update 2010. CA: a cancer journal for clinicians. 2010 Mar;60(2):70-98.

[7] Cuzick J, Thorat MA, Andriole G, Brawley OW, Brown PH, Culig Z, Eeles RA, Ford LG, Hamdy FC, Holmberg L, Ilic D. Prevention and early detection of prostate cancer. The lancet oncology. 2014 Oct 1;15(11):e484-92.

[8] Fernandez-Quilez A, Germán-Borda M, Leonardo-Carreño G, Castellanos-Perilla N, Soennesyn H, Oppedal K, Reidar-Kjosavik S. Prostate cancer screening and socioeconomic disparities in mexican older adults. salud pública de méxico. 2020 Feb 28;62(2, Mar-Abr):121-2.

[9] Tricoli JV, Schoenfeldt M, Conley BA. Detection of prostate cancer and predicting progression: current and future diagnostic markers. Clinical cancer research. 2004 Jun 15;10(12):3943-53.

[10] Romero FR, Romero AW, Brenny Filho T, Bark NM, Yamazaki DS, de Oliveira Júnior FC. Patients' perceptions of pain and discomfort during digital rectal exam for prostate cancer screening. Archivos espanoles de urologia. 2008;61(7):850-4.

[11] Catalona WJ, Richie JP, Ahmann FR, Hudson ML, Scardino PT, Flanigan RC, Dekernion JB, Ratliff TL, Kavoussi LR, Dalkin BL, Waters WB. Comparison of digital rectal examination and serum prostate specific antigen in the early detection of prostate

cancer: results of a multicenter clinical trial of 6,630 men. The Journal of urology. 1994 May;151(5):1283-90.

[12] Patel A. Benign vs Malignant Tumors. JAMA oncology. 2020 Sep 1;6(9):1488-.

[13] Schröder FH, Hugosson J, Roobol MJ, Tammela TL, Ciatto S, Nelen V, Kwiatkowski M, Lujan M, Lilja H, Zappa M, Denis LJ. Screening and prostate-cancer mortality in a randomized European study. New England journal of medicine. 2009 Mar 26;360(13):1320-8.

[14] Andriole GL, Crawford ED, Grubb III RL, Buys SS, Chia D, Church TR, Fouad MN, Gelmann EP, Kvale PA, Reding DJ, Weissfeld JL. Mortality results from a randomized prostate-cancer screening trial. New England Journal of Medicine. 2009 Mar 26;360(13):1310-9.

[15] O'Sullivan J. Controversies in PSA screening. BMJ Evidence-Based Medicine. 2017 Dec 1;22(6):198-.

[16] Barry MJ. Screening for prostate cancer--the controversy that refuses to die. New England Journal of Medicine. 2009 Mar 26;360(13):1351.

[17] Loeb S, Bjurlin MA, Nicholson J, Tammela TL, Penson DF, Carter HB, Carroll P, Etzioni R. Overdiagnosis and overtreatment of prostate cancer. European urology. 2014 Jun 1;65(6):1046-55.

[18] Moyer VA. Screening for prostate cancer: US Preventive Services Task Force recommendation statement. Annals of internal medicine. 2012 Jul 17;157(2):120-34.

[19] Gleason DF. Histologic grading of prostate cancer: a perspective. Human pathology. 1992 Mar 1;23(3):273-9.

[20] Gordetsky J, Epstein J. Grading of prostatic adenocarcinoma: current state and prognostic implications. Diagnostic pathology. 2016 Dec;11(1):1-8.

[21] Hodge KK, McNeal JE, Stamey TA. Ultrasound guided transrectal core biopsies of the palpably abnormal prostate. The Journal of urology. 1989 Jul 1;142(1):66-70.

[22] Eichler K, Hempel S, Wilby J, Myers L, Bachmann LM, Kleijnen J. Diagnostic value of systematic biopsy methods in the investigation of prostate cancer: a systematic review. The Journal of urology. 2006 May;175(5):1605-12.

[23] Manseck A, Fröhner M, Oehlschläger S, Hakenberg O, Friedrich K, Theissig F, Wirth MP. Is systematic sextant biopsy suitable for the detection of clinically significant prostate cancer?. Urologia Internationalis. 2000;65(2):80-3.

[24] Fink KG, Hutarew G, Lumper W, Jungwirth A, Dietze O, Schmeller NT. Prostate cancer detection with two sets of ten-core compared with two sets of sextant biopsies. Urology. 2001 Nov 1;58(5):735-9.

[25] Taira AV, Merrick GS, Galbreath RW, Andreini H, Taubenslag W, Curtis R, Butler WM, Adamovich E, Wallner KE. Performance of transperineal template-guided mapping biopsy in detecting prostate cancer in the initial and repeat biopsy setting. Prostate cancer and prostatic diseases. 2010 Mar;13(1):71-7.

[26] Loeb S, Vellekoop A, Ahmed HU, Catto J, Emberton M, Nam R, Rosario DJ, Scattoni V, Lotan Y. Systematic review of complications of prostate biopsy. European urology. 2013 Dec 1;64(6):876-92.

[27] Demirel HC, Davis JW. Multiparametric magnetic resonance imaging: Overview of the technique, clinical applications in prostate biopsy and future directions. Turkish journal of urology. 2018 Mar;44(2):93.

[28] Dirix P, Van Bruwaene S, Vandeursen H, Deckers F. Magnetic resonance imaging sequences for prostate cancer triage: two is a couple, three is a crowd? Translational andrology and urology. 2019 Dec;8(Suppl 5):S476.

[29] Shukla-Dave A, Hricak H. Role of MRI in prostate cancer detection. NMR in Biomedicine. 2014 Jan;27(1):16-24.

[30] Stabile A, Giganti F, Rosenkrantz AB, Taneja SS, Villeirs G, Gill IS, Allen C, Emberton M, Moore CM, Kasivisvanathan V. Multiparametric MRI for prostate cancer diagnosis: current status and future directions. Nature Reviews Urology. 2020 Jan;17(1):41-61.

[31] Bergdahl AG, Wilderäng U, Aus G, Carlsson S, Damber JE, Frånlund M, Geterud K, Khatami A, Socratous A, Stranne J, Hellström M. Role of magnetic resonance imaging in prostate cancer screening: a pilot study within the Göteborg randomised screening trial. European urology. 2016 Oct 1;70(4):566-73.

[32] Eldred-Evans D, Burak P, Connor MJ, Day E, Evans M, Fiorentino F, Gammon M, Hosking-Jervis F, Klimowska-Nassar N, McGuire W, Padhani AR. Population-Based prostate cancer screening with magnetic resonance imaging or ultrasonography: the IP1-PROSTAGRAM study. JAMA oncology. 2021 Mar 1;7(3):395-402.

[33] Xu L, Zhang G, Shi B, Liu Y, Zou T, Yan W, Xiao Y, Xue H, Feng F, Lei J, Jin Z. Comparison of biparametric and multiparametric MRI in the diagnosis of prostate cancer. Cancer Imaging. 2019 Dec;19(1):1-8.

[34] Israël B, van der Leest M, Sedelaar M, Padhani AR, Zámecnik P, Barentsz JO. Multiparametric magnetic resonance imaging for the detection of clinically significant prostate cancer: what urologists need to know. Part 2: interpretation. European urology. 2020 Apr 1;77(4):469-80.

[35] Ren J, Huan Y, Wang H, Zhao H, Ge Y, Chang Y, Liu Y. Diffusion-weighted imaging in normal prostate and differential diagnosis of prostate diseases. Abdominal imaging. 2008 Nov;33(6):724-8.

[36] Langer DL, van der Kwast TH, Evans AJ, Plotkin A, Trachtenberg J, Wilson BC, Haider MA. Prostate tissue composition and MR measurements: investigating the relationships between ADC, T2, K trans, ve, and corresponding histologic features. Radiology. 2010 May;255(2):485-94

[37] Manenti G, Nezzo M, Chegai F, Vasili E, Bonanno E, Simonetti G. DWI of prostate cancer: optimal-value in clinical practice. Prostate cancer. 2014 Oct;2014.

[38] Weinreb JC, Barentsz JO, Choyke PL, Cornud F, Haider MA, Macura KJ, Margolis D, Schnall MD, Shtern F, Tempany CM, Thoeny HC. PI-RADS prostate imaging–reporting and data system: 2015, version 2. European urology. 2016 Jan 1;69(1):16-40.

[39] Turkbey B, Rosenkrantz AB, Haider MA, Padhani AR, Villeirs G, Macura KJ, Tempany CM, Choyke PL, Cornud F, Margolis DJ, Thoeny HC. Prostate imaging reporting and data system version 2.1: 2019 update of prostate imaging reporting and data system version 2. European urology. 2019 Sep 1;76(3):340-51.

[40] De Visschere P, Lumen N, Ost P, Decaestecker K, Pattyn E, Villeirs G. Dynamic contrast-enhanced imaging has limited added value over T2-weighted imaging and diffusion-weighted imaging when using PI-RADSv2 for diagnosis of clinically significant prostate cancer in patients with elevated PSA. Clinical radiology. 2017 Jan 1;72(1):23-32.

[41] Kelloff GJ, Choyke P, Coffey DS. Challenges in clinical prostate cancer: role of imaging. American journal of roentgenology. 2009 Jun;192(6):1455-70.

[42] Dickinson L, Ahmed HU, Allen C, Barentsz JO, Carey B, Futterer JJ, Heijmink SW, Hoskin P, Kirkham AP, Padhani AR, Persad R. Clinical applications of multiparametric MRI within the prostate cancer diagnostic pathway. Urologic oncology. 2013 Apr;31(3):281.

[43] Chou R, Croswell JM, Dana T, Bougatsos C, Blazina I, Fu R, Gleitsmann K, Koenig HC, Lam C, Maltz A, Rugge JB. Screening for prostate cancer: a review of the evidence for the US Preventive Services Task Force. Annals of internal medicine. 2011 Dec 6;155(11):762-71.

[44] Villers A, Puech P, Mouton D, Leroy X, Ballereau C, Lemaitre L. Dynamic contrast enhanced, pelvic phased array magnetic resonance imaging of localized prostate cancer for predicting tumor volume: correlation with radical prostatectomy findings. The Journal of urology. 2006 Dec;176(6):2432-7.

[45] Villeirs GM, De Meerleer GO, De Visschere PJ, Fonteyne VH, Verbaeys AC, Oosterlinck W. Combined magnetic resonance imaging and spectroscopy in the assessment of high-grade prostate carcinoma in patients with elevated PSA: a single-institution experience of 356 patients. European journal of radiology. 2011 Feb 1;77(2):340-5.

[46] Burnside ES, Sickles EA, Bassett LW, Rubin DL, Lee CH, Ikeda DM, Mendelson EB, Wilcox PA, Butler PF, D'Orsi CJ. The ACR BI-RADS® experience: learning from history. Journal of the American College of Radiology. 2009 Dec 1;6(12):851-60.

[47] Magheli A, Rais-Bahrami S, Trock BJ, Humphreys EB, Partin AW, Han M, Gonzalgo ML. Prostate specific antigen versus prostate specific antigen density as a prognosticator of pathological characteristics and biochemical recurrence following radical prostatectomy. The Journal of urology. 2008 May;179(5):1780-4.

[48] Gillies RJ, Anderson AR, Gatenby RA, Morse DL. The biology underlying molecular imaging in oncology: from genome to anatome and back again. Clinical radiology. 2010 Jul 1;65(7):517-21.

[49] Porten SP, Whitson JM, Cowan JE, Cooperberg MR, Shinohara K, Perez N, Greene KL, Meng MV, Carroll PR. Changes in prostate cancer grade on serial biopsy in men undergoing active surveillance. Journal of clinical oncology: official journal of the American Society of Clinical Oncology. 2011 Jul 1;29(20):2795-800.

[50] Morgan VA, Riches SF, Thomas K, Vanas N, Parker C, Giles S, Desouza NM. Diffusion-weighted magnetic resonance imaging for monitoring prostate cancer progression in patients managed by active surveillance. The British journal of radiology. 2011 Jan;84(997):31-7.

[51] Labanaris AP, Zugor V, Takriti S, Smiszek R, Engelhard K, Nützel R, Kühn R. The role of conventional and functional endorectal magnetic resonance imaging in the decision of whether to preserve or resect the neurovascular bundles during radical retropubic prostatectomy. Scandinavian journal of urology and nephrology. 2009 Jan 1;43(1):25-31.

[52] Raz O, Haider MA, Davidson SR, Lindner U, Hlasny E, Weersink R, Gertner MR, Kucharcyzk W, McCluskey SA, Trachtenberg J. Real-time magnetic resonance imaging–guided focal laser therapy in patients with low-risk prostate cancer. European urology. 2010 Jul 1;58(1):173-7.

[53] Hricak H, Choyke PL, Eberhardt SC, Leibel SA, Scardino PT. Imaging prostate cancer: a multidisciplinary perspective. Radiology. 2007 Apr;243(1):28-53.

[54] Linden RA, Halpern EJ. Advances in transrectal ultrasound imaging of the prostate. InSeminars in Ultrasound, CT and MRI 2007 Aug 1 (Vol. 28, No. 4, pp. 249-257). WB Saunders.

[55] Halpern EJ, Verkh L, Forsberg F, Gomella LG, Mattrey RF, Goldberg BB. Initial experience with contrast-enhanced sonography of the prostate. American journal of roentgenology. 2000 Jun;174(6):1575-80.

[56] Hricak H, White S, Vigneron D, Kurhanewicz J, Kosco A, Levin D, Weiss J, Narayan P, Carroll PR. Carcinoma of the prostate gland: MR imaging with pelvic phased-array coils versus integrated endorectal--pelvic phased-array coils. Radiology. 1994 Dec;193(3):703-9.

[57] Presti Jr JC. Repeat prostate biopsy—when, where, and how. InUrologic Oncology: Seminars and Original Investigations 2009 May 1 (Vol. 27, No. 3, pp. 312-314). Elsevier.

[58] Hambrock T, Somford DM, Hoeks C, Bouwense SA, Huisman H, Yakar D, van Oort IM, Witjes JA, Fütterer JJ, Barentsz JO. Magnetic resonance imaging guided prostate biopsy in men with repeat negative biopsies and increased prostate specific antigen. The Journal of urology. 2010 Feb;183(2):520-8.

[59] Fehr D, Veeraraghavan H, Wibmer A, Gondo T, Matsumoto K, Vargas HA, Sala E, Hricak H, Deasy JO. Automatic classification of prostate cancer Gleason scores from multiparametric magnetic resonance images. Proceedings of the National Academy of Sciences. 2015 Nov 17;112(46):E6265-73.

[60] Lee KC, Sud S, Meyer CR, Moffat BA, Chenevert TL, Rehemtulla A, Pienta KJ, Ross BD. An imaging biomarker of early treatment response in prostate cancer that has metastasized to the bone. Cancer research. 2007 Apr 15;67(8):3524-8.

[61] Hanlon AL, Moore DF, Hanks GE. Modeling postradiation prostate specific antigen level kinetics: predictors of rising postnadir slope suggest cure in men who remain biochemically free of prostate carcinoma. Cancer. 1998 Jul 1;83(1):130-4.

[62] Lee DH, Nam JK, Lee SS, Han JY, Lee JW, Chung MK, Park SW. Comparison of multiparametric and biparametric MRI in first round cognitive targeted prostate biopsy in patients with PSA levels under 10 ng/mL. Yonsei Medical Journal. 2017 Sep 1;58(5):994-9.

[63] Becker AS, Chaitanya K, Schawkat K, Muehlematter UJ, Hötker AM, Konukoglu E, Donati OF. Variability of manual segmentation of the prostate in axial T2-weighted MRI: A multi-reader study. European journal of radiology. 2019 Dec 1;121:108716.

[64] Liechti MR, Muehlematter UJ, Schneider AF, Eberli D, Rupp NJ, Hötker AM, Donati OF, Becker AS. Manual prostate cancer segmentation in MRI: interreader agreement and volumetric correlation with transperineal template core needle biopsy. European Radiology. 2020 Sep;30(9):4806-15.

[65] Steiger P, Thoeny HC. Prostate MRI based on PI-RADS version 2: how we review and report. Cancer Imaging. 2016 Dec;16(1):1-9.

[66] Rosenkrantz AB, Ginocchio LA, Cornfeld D, Froemming AT, Gupta RT, Turkbey B, Westphalen AC, Babb JS, Margolis DJ. Interobserver reproducibility of the PI-RADS version 2 lexicon: a multicenter study of six experienced prostate radiologists. Radiology. 2016 Sep;280(3):793-804.

[67] Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJ. Artificial intelligence in radiology. Nature Reviews Cancer. 2018 Aug;18(8):500-10.

[68] Briganti G, Le Moine O. Artificial intelligence in medicine: today and tomorrow. Frontiers in medicine. 2020 Feb 5;7:27.

[69] van Leeuwen KG, Schalekamp S, Rutten MJ, van Ginneken B, de Rooij M. Artificial intelligence in radiology: 100 commercially available products and their scientific evidence. European radiology. 2021 Jun;31(6):3797-804.

[70] Irvin, J. et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In Proc. of the AAAI Conference on Artificial Intelligence Vol. 33, 590–597 (2019).

[71] Esteva A, Chou K, Yeung S, Naik N, Madani A, Mottaghi A, Liu Y, Topol E, Dean J, Socher R. Deep learning-enabled medical computer vision. NPJ digital medicine. 2021 Jan 8;4(1):1-9.

[72] Ding, J., Li, A., Hu, Z. & Wang, L. in Medical Image Computing and Computer Assisted Intervention—MICCAI 2017 559–567 (Springer International Publishing, 2017)

[73] Rimmer A. Radiologist shortage leaves patient care at risk, warns royal college. BMJ: British Medical Journal (Online). 2017 Oct 11;359.

[74] Bhargavan M, Sunshine JH, Schepps B. Too few radiologists?. American Journal of Roentgenology. 2002 May 1;178(5):1075-82.

[75] Boland GW, Guimaraes AS, Mueller PR. The radiologist's conundrum: benefits and costs of increasing CT capacity and utilization. European radiology. 2009 Jan;19(1):9-11.

[76] McDonald RJ, Schwartz KM, Eckel LJ, Diehn FE, Hunt CH, Bartholmai BJ, Erickson BJ, Kallmes DF. The effects of changes in utilization and technological advancements of cross-sectional imaging on radiologist workload. Academic radiology. 2015 Sep 1;22(9):1191-8.

[77] Fitzgerald R. Error in radiology. Clinical radiology. 2001 Dec 1;56(12):938-46.

[78] Wismüller A, Stockmaster L. A prospective randomized clinical trial for measuring radiology study reporting time on Artificial Intelligence-based detection of intracranial hemorrhage in emergent care head CT. InMedical Imaging 2020: Biomedical Applications in Molecular, Structural, and Functional Imaging 2020 Feb 28 (Vol. 11317, p. 113170M). International Society for Optics and Photonics.

[79] Castellino RA. Computer aided detection (CAD): an overview. Cancer Imaging. 2005;5(1):17.

[80] Sun Q, Lin X, Zhao Y, Li L, Yan K, Liang D, Sun D, Li ZC. Deep learning vs. radiomics for predicting axillary lymph node metastasis of breast cancer using

ultrasound images: don't forget the peritumoral region. Frontiers in oncology. 2020 Jan 31;10:53.

[81] Shen D, Wu G, Suk HI. Deep learning in medical image analysis. Annual review of biomedical engineering. 2017 Jun 21;19:221-48.

[82] Benjamens S, Dhunnoo P, Meskó B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. NPJ digital medicine. 2020 Sep 11;3(1):1-8.

[83] Papadimitroulas P, Brocki L, Chung NC, Marchadour W, Vermet F, Gaubert L, Eleftheriadis V, Plachouris D, Visvikis D, Kagadis GC, Hatt M. Artificial intelligence: Deep learning in oncological radiomics and challenges of interpretability and data harmonization. Physica Medica. 2021 Mar 1;83:108-21.

[84] Paul R, Hawkins SH, Balagurunathan Y, Schabath M, Gillies RJ, Hall LO, Goldgof DB. Deep feature transfer learning in combination with traditional features predicts survival among patients with lung adenocarcinoma. Tomography. 2016 Dec;2(4):388-95.

[85] Cheng JZ, Ni D, Chou YH, Qin J, Tiu CM, Chang YC, Huang CS, Shen D, Chen CM. Computer-aided diagnosis with deep learning architecture: applications to breast lesions in US images and pulmonary nodules in CT scans. Scientific reports. 2016 Apr 15;6(1):1-3.

[86] Van Ginneken B, Schaefer-Prokop CM, Prokop M. Computer-aided diagnosis: how to move from the laboratory to the clinic. Radiology. 2011 Dec;261(3):719-32.

[87] Rosenblatt F. The perceptron: a probabilistic model for information storage and organization in the brain. Psychological review. 1958 Nov;65(6):386.

[88] Goodfellow I, Bengio Y, Courville A. Deep learning. MIT press; 2016 Nov 10.

[89] Kim T, Adalı T. Approximation by fully complex multilayer perceptrons. Neural computation. 2003 Jul 1;15(7):1641-66.

[90] Pinkus A. Approximation theory of the MLP model in neural networks. Acta numerica. 1999 Jan;8:143-95.

[91] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems. 2012;25.

[92] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556. 2014 Sep

[93] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. InProceedings of the IEEE conference on computer vision and pattern recognition 2016 (pp. 770-778).

[94] Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. InInternational conference on machine learning 2015 Jun 1 (pp. 448-456). PMLR.

[95] Yadav SS, Jadhav SM. Deep convolutional neural network based medical image classification for disease diagnosis. Journal of Big Data. 2019 Dec;6(1):1-8.

[96] Le MH, Chen J, Wang L, Wang Z, Liu W, Cheng KT, Yang X. Automated diagnosis of prostate cancer in multi-parametric MRI based on multimodal convolutional neural networks. Physics in Medicine & Biology. 2017 Jul 24;62(16):6497.

[97] Cao Z, Duan L, Yang G, Yue T, Chen Q. An experimental study on breast lesion detection and classification from ultrasound images using deep learning architectures. BMC medical imaging. 2019 Dec;19(1):1-9.

[98] Lundervold AS, Lundervold A. An overview of deep learning in medical imaging focusing on MRI. Zeitschrift für Medizinische Physik. 2019 May 1;29(2):102-27.

[99] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. Advances in neural information processing systems. 2017;30.

[100] Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. 2018 Oct 11.

[101] Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692. 2019 Jul 26.

[102] Xiong R, Yang Y, He D, Zheng K, Zheng S, Xing C, Zhang H, Lan Y, Wang L, Liu T. On layer normalization in the transformer architecture. InInternational Conference on Machine Learning 2020 Nov 21 (pp. 10524-10533). PMLR.

[103] Xu J, Sun X, Zhang Z, Zhao G, Lin J. Understanding and improving layer normalization. Advances in Neural Information Processing Systems. 2019;32.

[104] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929. 2020 Oct 22.

[105] Arnab A, Dehghani M, Heigold G, Sun C, Lučić M, Schmid C. Vivit: A video vision transformer. InProceedings of the IEEE/CVF International Conference on Computer Vision 2021 (pp. 6836-6846).

[106] Wu Y, Liao K, Chen J, Chen DZ, Wang J, Gao H, Wu J. D-Former: A U-shaped Dilated Transformer for 3D Medical Image Segmentation. arXiv preprint arXiv:2201.00462. 2022 Jan 3.

[107] Dai Y, Gao Y, Liu F. Transmed: Transformers advance multi-modal medical image classification. Diagnostics. 2021 Aug;11(8):1384.

[108] Bourlard H, Kamp Y. Auto-association by multilayer perceptrons and singular value decomposition. Biological cybernetics. 1988 Sep;59(4):291-4.

[109] Rumelhart DE, Hinton GE, Williams RJ. Learning internal representations by error propagation. California Univ San Diego La Jolla Inst for Cognitive Science; 1985 Sep 1.

[110] Masci J, Meier U, Cireşan D, Schmidhuber J. Stacked convolutional auto-encoders for hierarchical feature extraction. InInternational conference on artificial neural networks 2011 Jun 14 (pp. 52-59). Springer, Berlin, Heidelberg.

[111] Pu Y, Gan Z, Henao R, Yuan X, Li C, Stevens A, Carin L. Variational autoencoder for deep learning of images, labels and captions. Advances in neural information processing systems. 2016;29.

[112] Van Den Oord A, Vinyals O. Neural discrete representation learning. Advances in neural information processing systems. 2017;30.

[113] Shvetsova N, Bakker B, Fedulova I, Schulz H, Dylov DV. Anomaly detection in medical imaging with deep perceptual autoencoders. IEEE Access. 2021 Aug 24;9:118571-83.

[114] Baur C, Wiestler B, Albarqouni S, Navab N. Fusing unsupervised and supervised deep learning for white matter lesion segmentation. InInternational Conference on Medical Imaging with Deep Learning 2019 May 24 (pp. 63-72). PMLR.

[115] Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial nets. Advances in neural information processing systems. 2014;27.

[116] Mirza M, Osindero S. Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784. 2014 Nov 6.

[117] Zhu JY, Park T, Isola P, Efros AA. Unpaired image-to-image translation using cycle-consistent adversarial networks. InProceedings of the IEEE international conference on computer vision 2017 (pp. 2223-2232).

[118] Palladino JA, Slezak DF, Ferrante E. Unsupervised domain adaptation via CycleGAN for white matter hyperintensity segmentation in multicenter MR images. In16th International Symposium on Medical Information Processing and Analysis 2020 Nov 3 (Vol. 11583, p. 1158302). International Society for Optics and Photonics.

[119] Chen C, Dou Q, Chen H, Qin J, Heng PA. Synergistic image and feature adaptation: Towards cross-modality domain adaptation for medical image segmentation. InProceedings of the AAAI conference on artificial intelligence 2019 Jul 17 (Vol. 33, No. 01, pp. 865-872).

[120] Fernandez-Quilez A, Larsen SV, Goodwin M, Gulsrud TO, Kjosavik SR, Oppedal K. Improving prostate whole gland segmentation in t2-weighted mri with synthetically generated data. In2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI) 2021 Apr 13 (pp. 1915-1919). IEEE.

[121] Han C, Hayashi H, Rundo L, Araki R, Shimoda W, Muramatsu S, Furukawa Y, Mauri G, Nakayama H. GAN-based synthetic brain MR image generation. In2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018) 2018 Apr 4 (pp. 734-738). IEEE.

[122] Skandarani Y, Jodoin PM, Lalande A. Gans for medical image synthesis: An empirical study. arXiv preprint arXiv:2105.05318. 2021 May 11.

[123] Singh A, Thakur N, Sharma A. A review of supervised machine learning algorithms. In2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom) 2016 Mar 16 (pp. 1310-1315). Ieee.

[124] Dayan P, Sahani M, Deback G. Unsupervised learning. The MIT encyclopedia of the cognitive sciences. 1999 Oct:857-9.

[125] Segovia F, Górriz JM, Ramírez J, Martínez-Murcia FJ, García-Pérez M. Using deep neural networks along with dimensionality reduction techniques to assist the diagnosis of neurodegenerative disorders. Logic Journal of the IGPL. 2018 Nov 27;26(6):618-28.

[126] Domingos P. A few useful things to know about machine learning. Communications of the ACM. 2012 Oct 1;55(10):78-87.

[127] Zhang Z, Sabuncu M. Generalized cross entropy loss for training deep neural networks with noisy labels. Advances in neural information processing systems. 2018;31.

[128] Chen RC, Dewi C, Huang SW, Caraka RE. Selecting critical features for data classification based on machine learning methods. Journal of Big Data. 2020 Dec;7(1):1-26.

[129] Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980. 2014 Dec 22.

[130] Loshchilov I, Hutter F. Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101. 2017 Nov 14.

[131] Ruder S. An overview of gradient descent optimization algorithms. arXiv preprint arXiv:1609.04747. 2016 Sep 15.

[132] Goldenberg SL, Nir G, Salcudean SE. A new era: artificial intelligence and machine learning in prostate cancer. Nature Reviews Urology. 2019 Jul;16(7):391-403.

[133] Li S, Chen Y, Yang S, Luo W. Cascade dense-unet for prostate segmentation in mr images. InInternational Conference on Intelligent Computing 2019 Aug 3 (pp. 481-490). Springer, Cham.

[134] Aldoj N, Biavati F, Michallek F, Stober S, Dewey M. Automatic prostate and prostate zones segmentation of magnetic resonance images using DenseNet-like U-net. Scientific reports. 2020 Aug 31;10(1):1-7.

[135] Meyer A, Mehrtash A, Rak M, Schindele D, Schostak M, Tempany C, Kapur T, Abolmaesumi P, Fedorov A, Hansen C. Automatic high-resolution segmentation of the

prostate from multi-planar MRI. In 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018) 2018 Apr 4 (pp. 177-181). IEEE.

[136] Sanders JW, Kudchadker RJ, Tang C, Mok H, Venkatesan AM, Thames HD, Frank SJ. Prospective Evaluation of Prostate and Organs at Risk Segmentation Software for MRI-based Prostate Radiation Therapy. Radiology: Artificial Intelligence. 2022 Jan 26:e210151.

[137] Bloch, N., Madabhushi, A., Huisman, H., Freymann, J., et al.: NCI-ISBI 2013 Challenge: Automated Segmentation of Prostate Structures. 2014

[138] Lemaître G, Martí R, Freixenet J, Vilanova JC, Walker PM, Meriaudeau F. Computer-aided detection and diagnosis for prostate cancer based on mono and multi-parametric MRI: a review. Computers in biology and medicine. 2015 May 1;60:8-31.

[139] Litjens G, Toth R, van de Ven W, Hoeks C, Kerkstra S, van Ginneken B, Vincent G, Guillard G, Birbeck N, Zhang J, Strand R. Evaluation of prostate segmentation algorithms for MRI: the PROMISE12 challenge. Medical image analysis. 2014 Feb 1;18(2):359-73.

[140] Liu Q, Dou Q, Heng PA. Shape-aware meta-learning for generalizing prostate MRI segmentation to unseen domains. InInternational Conference on Medical Image Computing and Computer-Assisted Intervention 2020 Oct 4 (pp. 475-485). Springer, Cham.

[141] Liu Q, Dou Q, Yu L, Heng PA. MS-Net: multi-site network for improving prostate segmentation with heterogeneous MRI data. IEEE transactions on medical imaging. 2020 Feb 17;39(9):2713-24.

[142] Hu Y, Modat M, Gibson E, Li W, Ghavami N, Bonmati E, Wang G, Bandula S, Moore CM, Emberton M, Ourselin S. Weakly-supervised convolutional neural networks for multimodal image registration. Medical image analysis. 2018 Oct 1;49:1-3.

[143] Haskins G, Kruecker J, Kruger U, Xu S, Pinto PA, Wood BJ, Yan P. Learning deep similarity metric for 3D MR–TRUS image registration. International journal of computer assisted radiology and surgery. 2019 Mar;14(3):417-25.

[144] Wang X, Yang W, Weinreb J, Han J, Li Q, Kong X, Yan Y, Ke Z, Luo B, Liu T, Wang L. Searching for prostate cancer by fully automated magnetic resonance imaging

classification: deep learning versus non-deep learning. Scientific reports. 2017 Nov 13;7(1):1-8.

[145] Ishioka J, Matsuoka Y, Uehara S, Yasuda Y, Kijima T, Yoshida S, Yokoyama M, Saito K, Kihara K, Numao N, Kimura T. Computer-aided diagnosis of prostate cancer on magnetic resonance imaging using a convolutional neural network algorithm. BJU international. 2018 Sep;122(3):411-7.

[146] Le MH, Chen J, Wang L, Wang Z, Liu W, Cheng KT, Yang X. Automated diagnosis of prostate cancer in multi-parametric MRI based on multimodal convolutional neural networks. Physics in Medicine & Biology. 2017 Jul 24;62(16):6497.

[147] Mehrtash A, Sedghi A, Ghafoorian M, Taghipour M, Tempany CM, Wells III WM, Kapur T, Mousavi P, Abolmaesumi P, Fedorov A. Classification of clinical significance of MRI prostate findings using 3D convolutional neural networks. InMedical Imaging 2017: Computer-Aided Diagnosis 2017 Mar 3 (Vol. 10134, p. 101342A). International Society for Optics and Photonics.

[148] Saha A, Hosseinzadeh M, Huisman H. End-to-end prostate cancer detection in bpMRI via 3D CNNs: Effects of attention mechanisms, clinical priori and decoupled false positive reduction. Medical image analysis. 2021 Oct 1;73:102155.

[149] Shiradkar R, Podder TK, Algohary A, Viswanath S, Ellis RJ, Madabhushi A. Radiomics based targeted radiotherapy planning (Rad-TRaP): a computational framework for prostate cancer treatment planning with MRI. Radiation oncology. 2016 Dec;11(1):1-4.

[150] Boussion N, Valeri A, Malhaire JP, Visvikis D. Predicting the number of seeds in ldr prostate brachytherapy using machine learning and 320 patients. inradiotherapy and oncology 2018 apr 1 (vol. 127, pp. s477-s478). elsevier house, brookvale plaza, east park shannon, co, clare, 00000, ireland: elsevier ireland ltd.

[151] Almeida G, Tavares JM. Deep learning in radiation oncology treatment planning for prostate cancer: a systematic review. Journal of medical systems. 2020 Oct;44(10):1-5.

[152] Geis JR, Brady AP, Wu CC, Spencer J, Ranschaert E, Jaremko JL, Langer SG, Kitts AB, Birch J, Shields WF, van den Hoven van Genderen R. Ethics of artificial

intelligence in radiology: summary of the joint European and North American multi-society statement. Canadian Association of Radiologists Journal. 2019 Nov;70(4):329-34.

[153] Group SI, Community FR. Artificial intelligence and medical imaging 2018: French Radiology Community white paper. Diagnostic and interventional imaging. 2018 Nov 1;99(11):727-42.

[154] Geis JR, Brady AP, Wu CC, Spencer J, Ranschaert E, Jaremko JL, Langer SG, Kitts AB, Birch J, Shields WF, van den Hoven van Genderen R. Ethics of artificial intelligence in radiology: summary of the joint European and North American multisociety statement. Canadian Association of Radiologists Journal. 2019 Nov;70(4):329-34.

[155] Kim DW, Jang HY, Kim KW et al (2019) Design characteristics of studies reporting the performance of artificial intelligence algorithms for diagnostic analysis of medical images: results from recently published papers. Korean J Radiol 20:405–410

[156] Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. The Lancet Digital Health. 2021 Nov 1;3(11):e745-50.

[157] Altman DG. Statistics and ethics in medical research: collecting and screening data. BMJ 1980; 281:1399–1401

[158] Kshetri N. Data Labeling for the Artificial Intelligence Industry: Economic Impacts in Developing Countries. IT Professional. 2021 Mar 31;23(2):96-9.

[159] Ghesu FC, Georgescu B, Mansoor A, Yoo Y, Neumann D, Patel P, Vishwanath RS, Balter JM, Cao Y, Grbic S, Comaniciu D. Self-supervised Learning from 100 Million Medical Images. arXiv preprint arXiv:2201.01283. 2022 Jan 4.

[160] Zhou SK, Greenspan H, Davatzikos C, Duncan JS, Van Ginneken B, Madabhushi A, Prince JL, Rueckert D, Summers RM. A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. Proceedings of the IEEE. 2021 Feb 26.

[161] Tajbakhsh N, Shin JY, Gurudu SR, Hurst RT, Kendall CB, Gotway MB, Liang J. Convolutional neural networks for medical image analysis: Full training or fine tuning?. IEEE transactions on medical imaging. 2016 Mar 7;35(5):1299-312.

[162] Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, Cui C, Corrado G, Thrun S, Dean J. A guide to deep learning in healthcare. Nature medicine. 2019 Jan;25(1):24-9.

[163] Candemir S, Nguyen XV, Folio LR, Prevedello LM. Training Strategies for Radiology Deep Learning Models in Data-limited Scenarios. Radiology: Artificial Intelligence. 2021 Oct 6;3(6):e210014.

[164] Shorten C, Khoshgoftaar TM. A survey on image data augmentation for deep learning. Journal of big data. 2019 Dec;6(1):1-48.

[165] Elgendi M, Nasir MU, Tang Q, Smith D, Grenier JP, Batte C, Spieler B, Leslie WD, Menon C, Fletcher RR, Howard N. The effectiveness of image augmentation in deep learning networks for detecting COVID-19: A geometric transformation perspective. Frontiers in Medicine. 2021;8.

[166] Sánchez-Peralta LF, Picón A, Sánchez-Margallo FM, Pagador JB. Unravelling the effect of data augmentation transformations in polyp segmentation. International journal of computer assisted radiology and surgery. 2020 Dec;15(12):1975-88.

[167] Zhong Z, Zheng L, Kang G, Li S, Yang Y. Random erasing data augmentation. arXiv. arXiv preprint arXiv:1708.04896. 2017.

[168] Hao R, Namdar K, Liu L, Haider MA, Khalvati F. A comprehensive study of data augmentation strategies for prostate cancer detection in diffusion-weighted MRI using convolutional neural networks. Journal of Digital Imaging. 2021 Aug;34(4):862-76.

[169] Khan Z, Yahya N, Alsaih K, Ali SS, Meriaudeau F. Evaluation of deep neural networks for semantic segmentation of prostate in T2W MRI. Sensors. 2020 Jan;20(11):3183.

[170] Cipollari S, Guarrasi V, Pecoraro M, Bicchetti M, Messina E, Farina L, Paci P, Catalano C, Panebianco V. Convolutional neural networks for automated classification of prostate multiparametric magnetic resonance imaging based on image quality. Journal of Magnetic Resonance Imaging. 2022 Feb;55(2):480-90.

[171] Shin HC, Roth HR, Gao M, Lu L, Xu Z, Nogues I, Yao J, Mollura D, Summers RM. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE transactions on medical imaging. 2016 Feb 11;35(5):1285-98.

[172] Kora Venu S, Ravula S. Evaluation of deep convolutional generative adversarial networks for data augmentation of chest x-ray images. Future Internet. 2021 Jan;13(1):8.

[173] Chuquicusma MJ, Hussein S, Burt J, Bagci U. How to fool radiologists with generative adversarial networks? A visual turing test for lung cancer diagnosis. In2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018) 2018 Apr 4 (pp. 240-244). IEEE.

[174] Kazuhiro K, Werner RA, Toriumi F, Javadi MS, Pomper MG, Solnes LB, Verde F, Higuchi T, Rowe SP. Generative adversarial networks for the creation of realistic artificial brain magnetic resonance images. Tomography. 2018 Dec;4(4):159-63.

[175] Lim SK, Loo Y, Tran NT, Cheung NM, Roig G, Elovici Y. Doping: Generative data augmentation for unsupervised anomaly detection with gan. In 2018 IEEE International Conference on Data Mining (ICDM) 2018 Nov 17 (pp. 1122-1127). IEEE.

[176] Salimans T, Goodfellow I, Zaremba W, Cheung V, Radford A, Chen X. Improved techniques for training GANs. In: Lee D, Sugiyama M, Luxburg U, Guyon I, Garnett R, eds.Advances in neural information processing systems 29 (NIPS 2016).Red Hook, NY:Curran Associates,2016;2234–2242.

[177] Yi X, Walia E, Babyn P. Generative adversarial network in medical imaging: A review. Medical image analysis. 2019 Dec 1;58:101552.

[178] Yu H, Zhang X. Synthesis of prostate MR images for classification using capsule network-based GAN model. Sensors. 2020 Jan;20(20):5736.

[179] Hu X, Chung AG, Fieguth P, Khalvati F, Haider MA, Wong A. Prostategan: Mitigating data bias via prostate diffusion imaging synthesis with generative adversarial networks. arXiv preprint arXiv:1811.05817. 2018 Nov 14.

[180] Wang Z, Lin Y, Liao C, Cheng KT, Yang X. StitchAD-GAN for Synthesizing Apparent Diffusion Coefficient Images of Clinically Significant Prostate Cancer. In BMVC 2018 (p. 240).

[181] Cheplygina V, de Bruijne M, Pluim JP. Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. Medical image analysis. 2019 May 1;54:280-96.

[182] Chen J, Sathe S, Aggarwal C, Turaga D. Outlier detection with autoencoder ensembles. InProceedings of the 2017 SIAM international conference on data mining 2017 Jun 30 (pp. 90-98). Society for Industrial and Applied Mathematics.

[183] Chen X, Konukoglu E. Unsupervised detection of lesions in brain MRI using constrained adversarial auto-encoders. arXiv preprint arXiv:1806.04972. 2018 Jun 13.

[184] Wong KC, Karargyris A, Syeda-Mahmood T, Moradi M. Building disease detection algorithms with very small numbers of positive samples. InInternational Conference on Medical Image Computing and Computer-Assisted Intervention 2017 Sep 10 (pp. 471-479). Springer, Cham.

[185] Liu J, Lou B, Diallo M, Meng T, von Busch H, Grimm R, Tian Y, Comaniciu D, Kamen A, Winkel D, Tong A. Detecting Out-of-Distribution via an Unsupervised Uncertainty Estimation for Prostate Cancer Diagnosis.

[186] Rubinstein E, Salhov M, Nidam-Leshem M, White V, Golan S, Baniel J, Bernstine H, Groshar D, Averbuch A. Unsupervised tumor detection in Dynamic PET/CT imaging of the prostate. Medical image analysis. 2019 Jul 1;55:27-40.

[187] Lahat D, Adali T, Jutten C. Multimodal data fusion: an overview of methods, challenges, and prospects. Proceedings of the IEEE. 2015 Aug 20;103(9):1449-77.

[188] Dolz J, Gopinath K, Yuan J, Lombaert H, Desrosiers C, Ayed IB. HyperDense-Net: a hyper-densely connected CNN for multi-modal image segmentation. IEEE transactions on medical imaging. 2018 Oct 30;38(5):1116-26.

[189] Kamnitsas K, Ledig C, Newcombe VF, Simpson JP, Kane AD, Menon DK, Rueckert D, Glocker B. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. Medical image analysis. 2017 Feb 1;36:61-78.

[190] Zhang W, Li R, Deng H, Wang L, Lin W, Ji S, Shen D. Deep convolutional neural networks for multi-modality isointense infant brain image segmentation. NeuroImage. 2015 Mar 1;108:214-24.

[191] Srivastava N, Salakhutdinov RR. Multimodal learning with deep boltzmann machines. Advances in neural information processing systems. 2012;25.

[192] Pellicer-Valero OJ, Marenco Jiménez JL, Gonzalez-Perez V, Casanova Ramón-Borja JL, Martín García I, Barrios Benito M, Pelechano Gómez P, Rubio-Briones J, Rupérez MJ, Martín-Guerrero JD. Deep Learning for fully automatic detection, segmentation, and Gleason Grade estimation of prostate cancer in multiparametric Magnetic Resonance Images. Scientific reports. 2022 Feb 22;12(1):1-3

[193] Meyer A, Chlebus G, Rak M, Schindele D, Schostak M, van Ginneken B, Schenk A, Meine H, Hahn HK, Schreiber A, Hansen C. Anisotropic 3D multi-stream CNN for accurate prostate segmentation from multi-planar MRI. Computer Methods and Programs in Biomedicine. 2021 Mar 1;200:105821.

[194] Le MH, Chen J, Wang L, Wang Z, Liu W, Cheng KT, Yang X. Automated diagnosis of prostate cancer in multi-parametric MRI based on multimodal convolutional neural networks. Physics in Medicine & Biology. 2017 Jul 24;62(16):6497.

[195] Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg AC. ImageNet large scale visual recognition challenge (2014). arXiv preprint arXiv:1409.0575. 2014;2(3).

[196] Yosinski J, Clune J, Bengio Y, Lipson H. How transferable are features in deep neural networks?. Advances in neural information processing systems. 2014;27.

[197] Raghu M, Zhang C, Kleinberg J, Bengio S. Transfusion: Understanding transfer learning for medical imaging. Advances in neural information processing systems. 2019;32.

[198] Carneiro G, Nascimento J, Bradley AP. Unregistered multiview mammogram analysis with pre-trained deep learning models. InInternational Conference on Medical Image Computing and Computer-Assisted Intervention 2015 Oct 5 (pp. 652-660). Springer, Cham.

[199] Aljundi R, Lehaire J, Prost-Boucle F, Rouvière O, Lartizien C. Transfer learning for prostate cancer mapping based on multicentric MR imaging databases. InMedical learning meets medical imaging 2015 Jul 11 (pp. 74-82). Springer, Cham.

[200] Chen S, Ma K, Zheng Y. Med3d: Transfer learning for 3d medical image analysis. arXiv preprint arXiv:1904.00625. 2019 Apr 1.

[201] Han X. Automatic liver lesion segmentation using a deep convolutional neural network method. arXiv preprint arXiv:1704.07239. 2017 Apr 24.

[202] Yuan Y, Qin W, Buyyounouski M, Ibragimov B, Hancock S, Han B, Xing L. Prostate cancer classification with multiparametric MRI transfer learning model. Medical physics. 2019 Feb;46(2):756-65.

[203] Abdelmaksoud IR, Shalaby A, Mahmoud A, Elmogy M, Aboelfetouh A, El-Ghar A, El-Melegy M, Alghamdi NS, El-Baz A. Precise Identification of Prostate Cancer from DWI Using Transfer Learning. Sensors. 2021 Jan;21(11):3664.

[204] Hoar D, Lee PQ, Guida A, Patterson S, Bowen CV, Merrimen J, Wang C, Rendon R, Beyea SD, Clarke SE. Combined Transfer Learning and Test-Time Augmentation Improves Convolutional Neural Network-Based Semantic Segmentation of Prostate Cancer from Multi-Parametric MR Images. Computer Methods and Programs in Biomedicine. 2021 Oct 1;210:106375.

[205] Liu X, Zhang F, Hou Z, Mian L, Wang Z, Zhang J, Tang J. Self-supervised learning: Generative or contrastive. IEEE Transactions on Knowledge and Data Engineering. 2021 Jun 22.

[206] Jing L, Tian Y. Self-supervised visual feature learning with deep neural networks: A survey. IEEE transactions on pattern analysis and machine intelligence. 2020 May 4;43(11):4037-58.

[207] Gutmann M, Hyvärinen A. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In Proceedings of the thirteenth international conference on artificial intelligence and statistics 2010 Mar 31 (pp. 297-304). JMLR Workshop and Conference Proceedings.

[208] Oord AV, Li Y, Vinyals O. Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748. 2018 Jul 10.

[209] Doersch C, Gupta A, Efros AA. Unsupervised visual representation learning by context prediction. InProceedings of the IEEE international conference on computer vision 2015 (pp. 1422-1430).

[210] Kim D, Cho D, Yoo D, Kweon IS. Learning image representations by completing damaged jigsaw puzzles. In2018 IEEE Winter Conference on Applications of Computer Vision (WACV) 2018 Mar 12 (pp. 793-802). IEEE.

[211] Noroozi M, Favaro P. Unsupervised learning of visual representations by solving jigsaw puzzles. InEuropean conference on computer vision 2016 Oct 8 (pp. 69-84). Springer, Cham.

[212] Caron M, Bojanowski P, Joulin A, Douze M. Deep clustering for unsupervised learning of visual features. In Proceedings of the European conference on computer vision (ECCV) 2018 (pp. 132-149)

[213] Caron M, Misra I, Mairal J, Goyal P, Bojanowski P, Joulin A. Unsupervised learning of visual features by contrasting cluster assignments. Advances in Neural Information Processing Systems. 2020;33:9912-24.

[214] Tian Y, Krishnan D, Isola P. Contrastive multiview coding. In European conference on computer vision 2020 Aug 23 (pp. 776-794). Springer, Cham.

[215] X. Chen, H. Fan, R. Girshick, and K. He. Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297, 2020

[216] Chen T, Kornblith S, Norouzi M, Hinton G. A simple framework for contrastive learning of visual representations. InInternational conference on machine learning 2020 Nov 21 (pp. 1597-1607). PMLR.

[217] Tao X, Li Y, Zhou W, Ma K, Zheng Y. Revisiting Rubik’s cube: self-supervised learning with volume-wise transformation for 3D medical image segmentation. InInternational Conference on Medical Image Computing and Computer-Assisted Intervention 2020 Oct 4 (pp. 238-248). Springer, Cham.

[218] Sowrirajan H, Yang J, Ng AY, Rajpurkar P. Moco pretraining improves representation and transferability of chest x-ray models. InMedical Imaging with Deep Learning 2021 Aug 25 (pp. 728-744). PMLR.

[219] Azizi S, Mustafa B, Ryan F, Beaver Z, Freyberg J, Deaton J, Loh A, Karthikesalingam A, Kornblith S, Chen T, Natarajan V. Big self-supervised models advance medical image classification. InProceedings of the IEEE/CVF International Conference on Computer Vision 2021 (pp. 3478-3488).

[220] Li Z, Cui Z, Wang S, Qi Y, Ouyang X, Chen Q, Yang Y, Xue Z, Shen D, Cheng JZ. Domain Generalization for Mammography Detection via Multi-style and Multi-view Contrastive Learning. InInternational Conference on Medical Image Computing and Computer-Assisted Intervention 2021 Sep 27 (pp. 98-108). Springer, Cham.

[221] Larsson G, Maire M, Shakhnarovich G. Learning representations for automatic colorization. InEuropean conference on computer vision 2016 Oct 8 (pp. 577-593). Springer, Cham.

[222] Larsson G, Maire M, Shakhnarovich G. Colorization as a proxy task for visual understanding. InProceedings of the IEEE conference on computer vision and pattern recognition 2017 (pp. 6874-6883).

[223] Iizuka S, Simo-Serra E, Ishikawa H. Globally and locally consistent image completion. ACM Transactions on Graphics (ToG). 2017 Jul 20;36(4):1-4.

[224] Ledig C, Theis L, Huszar F, Caballero J, Cunningham A, Acosta A, Aitken A, Tejani A, Totz J, Wang Z, Shi W. Photo-realistic single image super-resolution using a generative adversarial network. arXiv 2016. arXiv preprint arXiv:1609.04802. 2016.

[225] Zhou Z, Sodha V, Pang J, Gotway MB, Liang J. Models genesis. Medical image analysis. 2021 Jan 1;67:101840.

[226] Taleb A, Loetzsch W, Danz N, Severin J, Gaertner T, Bergner B, Lippert C. 3d self-supervised methods for medical imaging. Advances in Neural Information Processing Systems. 2020;33:18158-72.

[227] Chen L, Bentley P, Mori K, Misawa K, Fujiwara M, Rueckert D. Self-supervised learning for medical image analysis using image context restoration. Medical image analysis. 2019 Dec 1;58:101539.

[228] Taleb A, Lippert C, Klein T, Nabi M. Multimodal self-supervised learning for medical image analysis. InInternational Conference on Information Processing in Medical Imaging 2021 Jun 28 (pp. 661-673). Springer, Cham.

[229] Bolous A, Seetharaman A, Bhattacharya I, Fan RE, Soerensen SJ, Chen L, Ghanouni P, Sonn GA, Rusu M. Clinically significant prostate cancer detection on MRI with self-supervised learning using image context restoration. InMedical Imaging 2021: Computer-Aided Diagnosis 2021 Feb 15 (Vol. 11597, p. 115971M). International Society for Optics and Photonics.

[230] Qian Y, Zhang Z, Wang B. ProCDet: A New Method for Prostate Cancer Detection Based on MR Images. IEEE Access. 2021 Sep 22;9:143495-505.

[231] Barentsz JO, Richenberg J, Clements R, Choyke P, Verma S, Villeirs G, Rouviere O, Logager V, Fütterer JJ. ESUR prostate MR guidelines 2012. European radiology. 2012 Apr;22(4):746-57.

[232] Litjens G, Debats O, Barentsz J, Karssemeijer N, Huisman H. Computer-aided detection of prostate cancer in MRI. IEEE transactions on medical imaging. 2014 Jan 30;33(5):1083-92.

[233] Armato SG, Huisman H, Drukker K, Hadjiiski L, Kirby JS, Petrick N, Redmond G, Giger ML, Cha K, Mamonov A, Kalpathy-Cramer J. PROSTATEx Challenges for computerized classification of prostate lesions from multiparametric magnetic resonance images. Journal of Medical Imaging. 2018 Nov;5(4):044501.

[234] Meyer A, Chlebus G, Rak M, Schindele D, Schostak M, van Ginneken B, Schenk A, Meine H, Hahn HK, Schreiber A, Hansen C. Anisotropic 3D multi-stream CNN for accurate prostate segmentation from multi-planar MRI. Computer Methods and Programs in Biomedicine. 2021 Mar 1;200:105821.

[235] Cuocolo R, Stanzione A, Castaldo A, De Lucia DR, Imbriaco M. Quality control and whole-gland, zonal and lesion annotations for the PROSTATEx challenge public dataset. European Journal of Radiology. 2021 May 1;138:109647.

[236] Choyke, P. Turkbey, B., Pinto, P., Merino M, Wood, B. Data from prostate-mri, The Cancer Imaging Archive 9 (2016).

[237] Bloch, B. N., Jain, A., Jaffe, C. C. Data from prostatediagnosis. the cancer imaging archive, The Cancer Imaging Archive 9 (2015).

[238] Zuley, M.L., Jarosz, R., Drake, B.F., Rancilio, D. Klim, A., Rieger-Christ, K., Lemmerman, J. Radiology data from the cancer genome atlas prostate adenocarcinoma [tcga-prad] collection, Cancer Imaging Arch 9 (2016).

[239] Fedorov A, Beichel R, Kalpathy-Cramer J, Finet J, Fillion-Robin JC, Pujol S, Bauer C, Jennings D, Fennessy F, Sonka M, Buatti J. 3D Slicer as an image computing platform for the Quantitative Imaging Network. Magnetic resonance imaging. 2012 Nov 1;30(9):1323-41.

[240] Harris CR, Millman KJ, Van Der Walt SJ, Gommers R, Virtanen P, Cournapeau D, Wieser E, Taylor J, Berg S, Smith NJ, Kern R. Array programming with NumPy. Nature. 2020 Sep;585(7825):357-62.

[241] Hunter JD. Matplotlib: A 2D graphics environment. Computing in science & engineering. 2007 May 1;9(03):90-5.

[242] Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, Corrado GS, Davis A, Dean J, Devin M, Ghemawat S. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:1603.04467. 2016 Mar 14.

[243] Yaniv Z, Lowekamp BC, Johnson HJ, Beare R. SimpleITK image-analysis notebooks: a collaborative environment for education and reproducible research. Journal of digital imaging. 2018 Jun;31(3):290-303.

[244] Yakubovskiy P. Classification models Keras. Github, Github repository. https://github.com/qubvel/classification_models.

[245] Yakubovskiy P. Segmentation models Keras. Github, Github repository. https://github.com/qubvel/segmentation_models.

[246] Van der Walt S, Schönberger JL, Nunez-Iglesias J, Boulogne F, Warner JD, Yager N, Gouillart E, Yu T. scikit-image: image processing in Python. PeerJ. 2014 Jun 19;2:e453.

[247] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. InInternational Conference on Medical image computing and computer-assisted intervention 2015 Oct 5 (pp. 234-241). Springer, Cham.

[248] Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434. 2015 Nov 19.

[249] Fernandez-Quilez, A., Parvez, O., Eftestøl, T., Kjosavik, S.R. & Oppedal, K. (2021). Improving prostate cancer triage with GAN-based synthetically generated

prostate ADC MRI. In Medical Imaging 2022: Computer-aided Diagnosis. International society for Optics and Photonics.

[250] Fernandez-Quilez, A., Ullah, H., Eftestøl, T., Kjosavik, S.R. & Oppedal, K. (2021). One class to rule them all: Detection and classification of prostate tumors presence in bi-parametric MRI based on auto-encoders. In Medical Imaging 2022: Computer-aided Diagnosis. International society for Optics and Photonics.

[251] Vinutha, H., Poornima, B., and Sagar, B., "Detection of outliers using interquartile range technique from intrusion dataset," in [Information and Decision Sciences], 511–518, Springer (2018).

[252] Fernandez-Quilez, A., Eftestøl, T., Kjosavik, S.R. & Oppedal, K. (2021). Learning to triage by learning to reconstruct: A generative self-supervised learning approach for prostate cancer based on axial T2w MRI. In Medical Imaging 2022: Computer-aided Diagnosis. International society for Optics and Photonics.

[253] Fernandez-Quilez, A., Eftestøl, T., Goodwin, M., Kjosavik, S.R. & Oppedal, K. (2022). Contrasting axial T2w MRI for prostate cancer triage: A self-supervised approach. In 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI) IEEE.

[254] Fernandez-Quilez, A., Eftestøl, T., Goodwin, M., Kjosavik, S.R. & Oppedal, K. (2022). Multi-planar T2w MRI for an improved prostate cancer lesion classification. In 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI) IEEE.

[255] Aldoj N, Biavati F, Michallek F, Stober S, Dewey M. Automatic prostate and prostate zones segmentation of magnetic resonance images using DenseNet-like U-net. Scientific reports. 2020 Aug 31;10(1):1-7.

[256] Tian Z, Liu L, Fei B. Deep convolutional neural network for prostate MR segmentation. International journal of computer assisted radiology and surgery. 2018 Nov;13(11):1687.

[257] Wang B, Lei Y, Tian S, Wang T, Liu Y, Patel P, Jani AB, Mao H, Curran WJ, Liu T, Yang X. Deeply supervised 3D fully convolutional networks with group dilated convolution for automatic MRI prostate segmentation. Medical physics. 2019 Apr;46(4):1707-18.

[258] Zhu Q, Du B, Turkbey B, Choyke PL, Yan P. Deeply-supervised CNN for prostate segmentation. In2017 international joint conference on neural networks (IJCNN) 2017 May 14 (pp. 178-184). IEEE.

[259] Ishioka J, Matsuoka Y, Uehara S, Yasuda Y, Kijima T, Yoshida S, Yokoyama M, Saito K, Kihara K, Numao N, Kimura T. Computer-aided diagnosis of prostate cancer on magnetic resonance imaging using a convolutional neural network algorithm. BJU international. 2018 Sep;122(3):411-7.

[260] Wang X, Yang W, Weinreb J, Han J, Li Q, Kong X, Yan Y, Ke Z, Luo B, Liu T, Wang L. Searching for prostate cancer by fully automated magnetic resonance imaging classification: deep learning versus non-deep learning. Scientific reports. 2017 Nov 13;7(1):1-8.

[261] Yu H, Zhang X. Synthesis of prostate MR images for classification using capsule network-based GAN model. Sensors. 2020 Jan;20(20):5736.

[262] Hu X, Chung AG, Fieguth P, Khalvati F, Haider MA, Wong A. Prostategan: Mitigating data bias via prostate diffusion imaging synthesis with generative adversarial networks. arXiv preprint arXiv:1811.05817. 2018 Nov 14.

[263] Wang Z, Lin Y, Liao C, Cheng KT, Yang X. StitchAD-GAN for Synthesizing Apparent Diffusion Coefficient Images of Clinically Significant Prostate Cancer. InBMVC 2018 (p. 240).

[264] Mehralivand S, Yang D, Harmon SA, Xu D, Xu Z, Roth H, Masoudi S, Kesani D, Lay N, Merino MJ, Wood BJ. Deep learning-based artificial intelligence for prostate cancer detection at biparametric MRI. Abdominal Radiology. 2022 Jan 31:1-0.

[265] Xu H, Baxter JS, Akin O, Cantor-Rivera D. Prostate cancer detection using residual networks. International journal of computer assisted radiology and surgery. 2019 Oct;14(10):1647-50.

[266] Aphinives C, Aphinives P. Artificial intelligence development for detecting prostate cancer in MRI. Egyptian Journal of Radiology and Nuclear Medicine. 2021 Dec;52(1):1-5.

[267] Saha A, Hosseinzadeh M, Huisman H. End-to-end prostate cancer detection in bpMRI via 3D CNNs: Effects of attention mechanisms, clinical priori and decoupled false positive reduction. Medical image analysis. 2021 Oct 1;73:102155.

[268] Abraham B, Nair MS. Computer-aided diagnosis of clinically significant prostate cancer from MRI images using sparse autoencoder and random forest classifier. Biocybernetics and Biomedical Engineering. 2018 Jan 1;38(3):733-44.

[269] Bolous A, Seetharaman A, Bhattacharya I, Fan RE, Soerensen SJ, Chen L, Ghanouni P, Sonn GA, Rusu M. Clinically significant prostate cancer detection on MRI with self-supervised learning using image context restoration. InMedical Imaging 2021: Computer-Aided Diagnosis 2021 Feb 15 (Vol. 11597, p. 115971M). International Society for Optics and Photonics.

[270] Qian Y, Zhang Z, Wang B. ProCDet: A New Method for Prostate Cancer Detection Based on MR Images. IEEE Access. 2021 Sep 22;9:143495-505.

[271] Meyer A, Chlebus G, Rak M, Schindele D, Schostak M, van Ginneken B, Schenk A, Meine H, Hahn HK, Schreiber A, Hansen C. Anisotropic 3D multi-stream CNN for accurate prostate segmentation from multi-planar MRI. Computer Methods and Programs in Biomedicine. 2021 Mar 1;200:105821.

[272] Lozoya RC, Iannessi A, Brag J, Patriti S, Oubel E. Assessing the relevance of multi-planar MRI acquisitions for prostate segmentation using deep learning techniques. InMedical Imaging 2018: Imaging Informatics for Healthcare, Research, and Applications 2018 Mar 6 (Vol. 10579, p. 105791A). International Society for Optics and Photonics.

[273] Yuan Y, Qin W, Buyyounouski M, Ibragimov B, Hancock S, Han B, Xing L. Prostate cancer classification with multiparametric MRI transfer learning model. Medical physics. 2019 Feb;46(2):756-65.

[274] Sabottke CF, Spieler BM. The effect of image resolution on deep learning in radiography. Radiology: Artificial Intelligence. 2020 Jan 22;2(1):e190015.

[275] Fernandez-Quilez, A., Eftestøl, T., Goodwin, M., Kjosavik, S. R., and Oppedal, K., “Self-transfer learning via patches: A prostate cancer triage approach based on bi-parametric mri,” arXiv preprint arXiv:2107.10806 (2021).

Appendices: Articles

Contribution A

IMPROVING PROSTATE WHOLE GLAND SEGMENTATION IN T2-WEIGHTED MRI WITH SYNTHETICALLY GENERATED DATA

Alvaro Fernandez-Quilez[1,2,†] *Steinar Valle Larsen*[2,3,†]
Morten Goodwin [4] Thor Ole Gulsrud [1] Svein Reidar Kjosavik [5] Ketil Oppedal[2,3,6]

[1]Department of Quality and Health Technology, University of Stavanger, Norway.
[2]Stavanger Medical Imaging Laboratory (SMIL), Stavanger University Hospital, Norway.
[3] Department of Electrical Engineering and Computer Science, University of Stavanger, Norway.
[4] Department of ICT, University of Agder, Grimstad, Norway.
[5] General Practice and Care Coordination Research Group, Stavanger University Hospital, Norway.
[6] Centre for Age-Related Medicine, Stavanger University Hospital, Norway.
[†]Shared first authorship.

ABSTRACT

Whole gland (WG) segmentation of the prostate plays a crucial role in detection, staging and treatment planning of prostate cancer (PCa). Despite promise shown by deep learning (DL) methods, they rely on the availability of a considerable amount of annotated data. Augmentation techniques such as translation and rotation of images present an alternative to increase data availability. Nevertheless, the amount of information provided by the transformed data is limited due to the correlation between the generated data and the original. Based on the recent success of generative adversarial networks (GAN) in producing synthetic images for other domains as well as in the medical domain, we presenta pipeline to generate WG segmentation masks and synthesize T2-weighted MRI of the prostate based on a publicly available multi-center dataset. Following, we use the generated data as a form of data augmentation. Results show an improvement in the quality of the WG segmentation when compared to standard augmentation techniques.

Index Terms**—** MRI, prostate, segmentation, convolutional neural networks, generative adversarial networks

1. INTRODUCTION

Prostate cancer (PCa) is the second most common diagnosedcancer [1], with an estimated incidence of 1.3 million new cases among men worldwide in 2018 [2, 3].

Thanks to recent advances in image acquisition and interpretation, MRI has proven to be a valuable tool for PCa detection, staging, treatment planning and intervention [4]. Segmentation of the prostate from MRI plays a crucial role. For instance, radiotherapy planning, MRI-transrectal ultra- sound fusion guided biopsy or radiation dose planning in brachytherapy are highly dependent on an accurate delineation of the prostate in imaging data [5]. Current practices include manual contouring by a specialist in a slice-by-slice basis, which is a time and labor-intensive task as well as susceptible to intra-observer and interobserver variability.

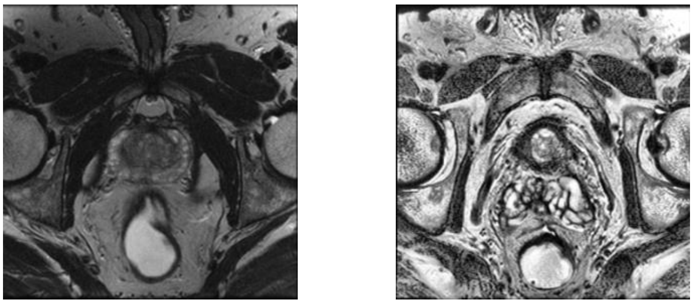


Fig. 1: Example of an original slice (left side) and preprocessed slice (right side).

In recent years, convolutional neural networks (CNNs) have shown promise in segmentation, where U-net was a major breakthrough [6]. Nevertheless, without further refinement, the training of CNN-based methods require a considerable amount of data [7]. Moreover, imbalanced data or data with low variability might lead to sub-optimal results [8]. Augmentation techniques (e.g. translation and rotation) haveproven useful but they also produce highly correlated data, limiting the amount of information provided to the algorithmin the training phase.

Generative Adversarial Networks (GAN) [9] have gained a considerable amount of attention in the DL community. Several variations of these generative models have been developed, such as the deep convolutional GAN (DCGAN) [10] or pix2pix [11] which are able to generate realistic images after learning the distribution of the original dataset and have been used to generate T1-weighted brain MRI [12]. The pix2pix architecture has been used to translate brain masksto images [13]. We propose a pipeline to generate paired prostate masks and T2-weighted MRI of the prostate for data augmentation. Our contributions in this work are:

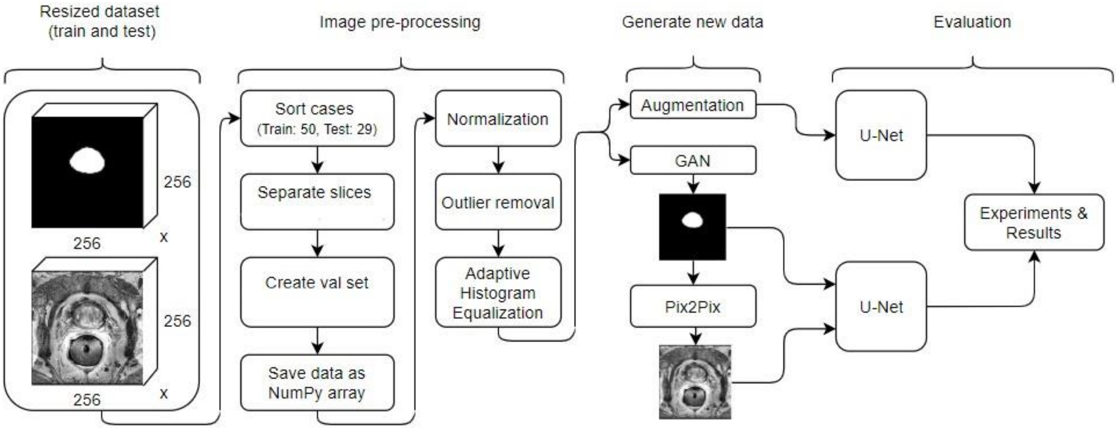


Fig. 2: Technical approach to the project.

1. We propose a DCGAN-based architecture to generate whole gland prostate masks from T2-weighted MRI.

2. We propose a pix2pix-based architecture to translate the synthetic WG prostate MRI masks into T2-weighted prostate MRI and to obtain paired training samples.

3. We provide a comprehensive comparison of different data augmentation techniques and their effect on the WG segmentation of the prostate as well as the effect of adding synthetic data to those standard augmentation techniques.

2. METHODS

In this work, we propose a semi-automatic pipeline able to generate synthetic pairs of T2-weighted prostate MRI and their respective WG mask. Figure 2 presents an overview of the steps followed in the work. First, we provide a description of the dataset, the pre-processing steps and the architectures training process.

2.1. Dataset

We use the PROMISE12 data set [14], containing T2-weighted axial MRI of 50 patients for training and 30 for testing. Ground truths of the WG of the prostate annotated by the experts are only available in the training set whilst the testing ones can only be accessed when submitting the results[1].

2.1.1. Pre-processing and data splitting

We perform four different steps. First, MRI are re-sampled by linear interpolation to 256x256, which is the lowest resolution present in the data set. Following, the intensity is then normalized to an interval of [0,1]. Outlier removal is performed by forcing the pixel intensity values of the im-age between the 1st and 99th percentiles. Finally, a contrast limited adaptative histogram equalization (CLAHE) is applied to improve local contrast and enhance the edge definition [15]. Figure 1 shows an example of the result obtainedafter applying the preprocessing steps. In order to evaluate the methods, we split the original dataset by patients following a 60%/20%/20% split for the training, validation and testing set, respectively. Eventually, an independent assessment is done using the PROMISE12 test set.

2.2. Segmentation architecture

Our segmentation architecture is based on the original U-netarchitecture [6].

2.2.1. Training of the network

Dice coefficient (DSC) [16] is used to evaluate the model performance during training. The architecture is trained for 200 epochs with a batch size of 32 on a 16gb NVIDIA Tesla P100, based on the validation set. Adam optimizer [17] is used with a learning rate scheduler. The learning rate started with a value of 1e-3 and reduced by a factor of 10 if the loss did not decrease during 10 epochs.

2.3. Generation of synthetic WG masks

We adopt the DCGAN architecture [10] to generate syntheticWG prostate masks. The DCGAN consists of two main com-ponents, a generator (*G*) and a discriminator (*D*), where bothG and *D* are CNNs.

[1] For up-to-date information refer to https://promise12.grand- challenge.org/

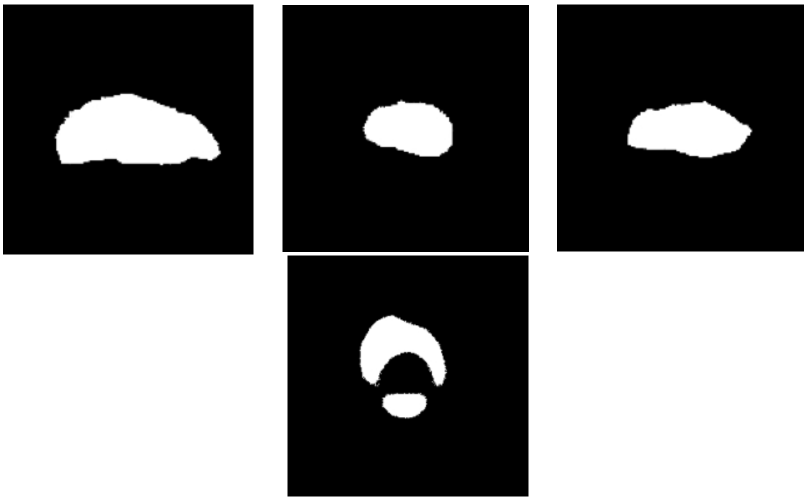


Fig. 3: Realistically-looking masks generated by DCGAN (top row) and a mask deemed as unrealistic (bottom).

The generator *G* gets inputs samples z from a distribution which has a normal prior imposed to it $p_z \sim N(\mu, \sigma^2)$. The role of *G* is to map such samples to the original data space while inducing a specific distribution p_{data} and thus synthesizing samples that follow such distribution $\hat{x}$ = G(z)). On the other hand, the role of D is to dis- criminate between real data samples x and generated ones $\hat{x}$.

2.3.1. Training of the network

The generator network takes a vector *z* of 100 random numbers drawn from a normal distribution $p_z \sim N(\mu, \sigma^2)$ as in- puts and outputs 256x256 WG segmentation masks. The generator architecture follows the one proposed in [10]. Nevertheless, after an extensive testing, we added two transposed convolutional layers. In addition, the original paper makes use of a rectified linear unit (ReLU) activation for all the generator layers except for the output whereas this work makes use of the LeakyReLU [18]. The discriminator implementation follows, again, the original one [10] with the addition of two extra convolutional layers. The architecture is trained for 1500 epochs on a 16gb NVIDIA Tesla P100, based on the validation set. Each epoch took approximately 100 seconds to finish. The batch size used for the

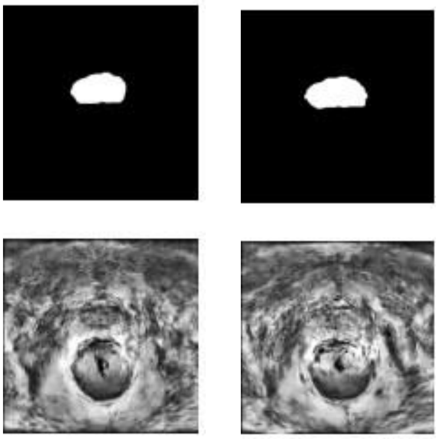


Fig. 4: Synthesized T2 weighted MRI (bottom) from GAN- generated WG masks (top row).

training was 32. Adam optimizer is used with default parameters and a learning rate of 0.0002.

2.3.2. Mask selection

The selection criteria for the synthetic masks is based on the visual appearance of the image and done in a manual way. For instance, some of the synthetic masks might contain disconnected prostate glands. Figure 3 shows an example of realisticWG prostate masks as well as an unrealistic one.

2.4. Mask-to-image translation: T2-weighted MRI

We base our architecture on [11]. In the particular case of translation, the generator (G) aims to map a source domain image $x_s \sim p_{xs}$ into the corresponding target image $x_t \sim p_{xt}$ via the mapping function $G(x_s, x_t)$. In this case, the discriminator tries to discriminate between the source image and its corresponding ground truth by classifying them as real while classifying the input and the transformation as fake.

2.4.1. Training of the network

The input of the architecture is a 256x256 segmentation mask,while the output was a synthesized T2-weighted MRI from the input mask. Figure 4 shows two examples of synthesized T2-weighted prostate MRI. We train the network for 200 epochs with a batch size of 1 on an Nvidia Tesla P100 16gb, based on the validation set. Each epoch took approximately 300 seconds to finish. The optimizer is Adam with a learning rate of 0.0002.

3. RESULTS

We evaluate the segmentation results following the metrics used in the PROMISE12 challenge as well as additional ones: DSC, mean volumetric DSC (VDSC), mean surface distance (MSD) and mean hausdorff distance (HD). The quality of the synthetic data was manually evaluated in a visual way in an intermediate step. The segmentation metrics are both representative of slice-based metrics as well as volume-based. More details on the metrics can be found in [14, 19]. Standard augmentation included rotation (±10 degrees), shifting (10%) total height and width), flipping and zooming ([1, 1.2] range). All results are based on U-net and the effect of the segmentation architecture was considered to be out of the scope of thepaper. Generated data results are based on 10000 synthetic T2-weighted images and their corresponding masks, which isapproximately 8 times the amount of original data. Quantitative results on the usage of synthetic data as well as standard augmentation techniques can be found in table 1 while table 2depicts the effect of the combination of the standard augmentation techniques and synthetically generated data.

A mean DSC of 73.77% was obtained for the WG ofthe prostate when adding synthetically generated data, both masks and the T2-weighted synthesized MRI. On the other hand, a DSC of 67.84% was obtained with a vanilla U-net without any augmentation. When comparing HD, the synthetically augmented dataset also improved the vanilla U-netresults (8.86 mm) by more than 8%. In addition, the MSD has also shown a considerable improvement when using a synthetically augmented dataset.

Transformation	DSC (%)	MSD	HD	VDSC(%)
Original	67.84	3.61	8.86	54.30
Vertical flip	66.93	18.52	19.68	48.72
Horizontal flip	69.98	15.41	13.47	50.86
Rotation	73.07	3.59	8.53	59.78
Shift	71.33	12.24	9.44	56.16
Zoom	70.69	7.31	**7.74**	55.21
All	67.30	10.14	12.36	51.23
Synthetic data	**73.77**	**1.16**	8.10	**69.36**

Table 1: Standard augmentation techniques and syntheticdata effect on WG segmentation results.

Finally, the volumetric DSCalso increased by more than 15% when using synthetic data. When comparing standard augmentation with the synthetic one, the latest surpasses by a considerable margin all the othertechniques when it comes to MSD and VDSC as well as bya small margin the DSC. Amongst the standard augmentationtechniques rotation obtained the best results.

Using synthetic data in combination with standard augmentation techniques yield to a larger improvement over the previous results. The combination resulted in an improvement of the metrics for all the standard augmentation techniques with the exception of rotation. In particular, zoom showed the best results amongst all of them with improvements with respect to the standard augmentation as well as the baseline.

Further experiments were performed to explore the effect of the synthetic sample size on the DSC value. Different % of synthetic data with respect to the size of the original sample were tested and we found out that the increase in the results was consistent with the amount of synthetic data used to augment the original set. In particular, we observed that every time the amount of synthetic data was doubled theDSC increased around 2%, reaching the peak with the largest amount of data tested (10000 cases).

Transformation	DSC(%)	MSD	HD	VDSC(%)
Original	67.84	3.61	8.86	54.30
Vertical flip	67.50	**0.92**	12.80	68.27
Horizontal flip	72.84	1.40	7.02	69.79
Rotation	68.01	1.93	9.93	68.06
Shift	73.37	1.18	8.66	**73.32**
Zoom	**73.90**	1.56	**6.94**	70.90
All	69.81	1.60	7.99	66.83
Synthetic data	73.77	1.16	8.10	69.36

Table 2: Combination of standard augmentation techniqueswith synthetic data

4. CONCLUSIONS

The objective of this work is to provide a pipeline able to pro- vide an alternative to the standard augmentation techniques by making use of GAN-based architectures. We propose a GAN-based framework to generate prostate WG segmentation masks and synthesize T2-weighted MRI from them. We evaluate our method against the standard augmentation techniques while providing a comprehensive comparison of the effect of them in the prostate WG segmentation in T2 weighted MRI. Results have shown that our method was ableto obtain better results when used as a standalone technique than the standard augmentation techniques whilst also im- proving the results of the standard augmentation techniques when used in combination with them.

To the best of our knowledge, this is the first approach that makes use of generated synthetic T2 weighted prostate MRI and their generated paired WG masks to improve WG segmentation. Whilst our method showed promise improving WG prostate segmentation results, the framework would benefit from inclusion of an automatic way to assess the quality of the generated images and select those deemed as more realistic. Furthermore, refinement of the synthesis architecture(pix2pix) could be explored in order to have more realistic details and less blurriness around the gland boundary.

5. COMPLIANCE WITH ETHICAL STANDARDS

This research study was conducted retrospectively using data made available in open access by the PROMISE12 challengeorganizers. Ethical approval was not required.

7. REFERENCES

[1] Rebecca L. Siegel, Kimberly D. Miller, and Ahmedin Jemal, “Cancer statistics, 2020,” CA: A Cancer Journal for Clinicians, vol. 70, no. 1, pp. 7–30, 2020.

[2] Freddie Bray, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L. Siegel, Lindsey A. Torre, and Ahmedin Jemal, “Global cancer statistics 2018,” CA: A Can- cer Journal for Clinicians, vol. 68, no. 6, pp. 394–424, 2018.

[3] Alvaro Fernandez-Quilez, Miguel Germa´n-Borda, Gabriel Leonardo, Nicola´s Castellanos, Hogne Soen- nesyn, Ketil Oppedal, and Svein Reidar-Kjosavik, “Prostate cancer screening and socioeconomic disparities in mexican older adults,” salud pu´blica de me´xico, vol. 62, no. 2, Mar-Abr, pp. 121–122, 2020.

[4] Amita Shukla-Dave and Hedvig Hricak, “Role of mri in prostate cancer detection,” NMR in Biomedicine, vol. 27, no. 1, pp. 16–24, 2014.

[5] Maria A Schmidt and Geoffrey S Payne, “Radiotherapy planning using mri,” Physics in Medicine & Biology, vol. 60, no. 22, pp. R323, 2015.

[6] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, “U-net: Convolutional networks for biomedical image segmentation,” in International Conference on Med- ical image computing and computer-assisted interven- tion. Springer, 2015, pp. 234–241.

[7] Pedro Domingos, “A few useful things to know about machine learning,” Communications of the ACM, vol. 55, no. 10, pp. 78–87, 2012.

[8] Hoo-Chang Shin, Holger R Roth, Mingchen Gao, Le Lu, Ziyue Xu, Isabella Nogues, Jianhua Yao, Daniel Mollura, and Ronald M Summers, “Deep convolutional neural networks for computer-aided detection: Cnn ar- chitectures, dataset characteristics and transfer learn- ing,” IEEE transactions on medical imaging, vol. 35, no. 5, pp. 1285–1298, 2016.

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversar- ial nets," in Advances in neural information processing systems, 2014, pp. 2672–2680.

[10]Alec Radford, Luke Metz, and Soumith Chintala, "Un- supervised representation learning with deep convolu- tional generative adversarial networks," arXiv preprint arXiv:1511.06434, 2015.

[11]Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros, "Image-to-image translation with conditional ad- versarial networks," in Proceedings of the IEEE confer- ence on computer vision and pattern recognition, 2017, pp. 1125–1134.

[12]Koshino Kazuhiro, Rudolf A Werner, Fujio Toriumi, Mehrbod S Javadi, Martin G Pomper, Lilja B Solnes, Franco Verde, Takahiro Higuchi, and Steven P Rowe, "Generative adversarial networks for the creation of realistic artificial brain magnetic resonance images," To- mography, vol. 4, no. 4, pp. 159, 2018.

[13]Hoo-Chang Shin, Neil A Tenenholtz, Jameson K Rogers, Christopher G Schwarz, Matthew L Senjem, Jeffrey L Gunter, Katherine P Andriole, and Mark Michalski, "Medical image synthesis for data augmen- tation and anonymization using generative adversarial networks," in International workshop on simulation and synthesis in medical imaging. Springer, 2018, pp. 1–11.

[14]Geert Litjens, Robert Toth, Wendy van de Ven, Caro- line Hoeks, Sjoerd Kerkstra, Bram van Ginneken, Gra- ham Vincent, Gwenael Guillard, Neil Birbeck, Jindang Zhang, et al., "Evaluation of prostate segmentation al- gorithms for mri: the promise12 challenge," Medical image analysis, vol. 18, no. 2, pp. 359–373, 2014.

[15]Stefan Van der Walt, Johannes L Scho¨nberger, Juan Nunez-Iglesias, Franc¸ois Boulogne, Joshua D Warner, Neil Yager, Emmanuelle Gouillart, and Tony Yu, "scikit-image: image processing in python," PeerJ, vol. 2, pp. e453, 2014.

[16]Anthony D Yao, Derrick L Cheng, Ian Pan, and Felipe Kitamura, "Deep learning in neuroradiology: A system- atic review of current algorithms and approaches for the new wave of imaging technology," Radiology: Artificial Intelligence, vol. 2, no. 2, pp. e190026, 2020.

[17]Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic

optimization," arXiv preprint arXiv:1412.6980, 2014.

[18]Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li, "Empirical evaluation of rectified activations in convo- lutional network," arXiv preprint arXiv:1505.00853, 2015.

[19]Guido Gerig, Matthieu Jomier, and Miranda Chakos, "Valmet: A new validation tool for assessing and im- proving 3d object segmentation," in International con- ference on medical image computing and computer-assisted intervention. Springer, 2001, pp. 516–523.

Contribution B

Improving prostate cancer triage with GAN-based synthetically generated prostate ADC MRI

Alvaro Fernandez-Quilez[a, b, e, *], Omer Parvez[b,*], Trygve Eftestøl[c], Svein Reidar Kjosavik[d], and Ketil Oppedal[b, e]

[a]Department of Quality and Health Technology, University of Stavanger, Stavanger, Norway.
[b]Stavanger Medical Imaging Laboratory (SMIL), Department of Radiology, Stavanger University Hospital, Stavanger, Norway.
[c]Department of Electrical Engineering and Computer Science, University of Stavanger, Stavanger, Norway.
[d]General Practice and Care Coordination Research Group, Stavanger University Hospital, Stavanger, Norway.
[e]Centre for Age-Related Medicine, Stavanger University Hospital, Stavanger, Norway.

[*]Shared first authorship.

ABSTRACT

Tumor classification in clinically significant (cS, Gleason score 7) or non-clinically significant (ncS, Gleason score < 7) plays a crucial role in patient management of prostate cancer (PCa), allowing to triage those patients that might benefit from an active surveillance approach from those that require an immediate action in the form of further testing or treatment. In spite of it, the current diagnostic pathway of PCa is substantially hampered by over-diagnosis of ncS lesions and under-detection of cS ones. Magnetic Resonance Imaging (MRI) has proven to be helpful in the stratification of tumors, but it relies on specialized training and experience. Despite the promise shown by deep learning (DL) methods, they are data-hungry approaches and rely on the availability of large amounts of annotated data. Standard augmentation techniques such as image translation have become the by default option to increase variability and data availability. However, the correlation between transformed data and original one limits the amount of information provided by them. Generative Adversarial Networks (GAN) present an alternative to classic augmentation techniques by creating synthetic samples. In this paper, we explore a conditional GAN (cGAN) architecture and a deep convolutional one (DCGAN) to generate synthetic apparent diffusion coefficient (ADC) prostate MRI. Following, we compare classic augmentation techniques with our GAN-based approach in a prostate cancer triage (classification of tumors) setting. We show that by adding synthetic ADC prostate MRI we are able to improve the final classification AUC of cS vs ncS tumors when compared to classic augmentation.

Keywords: Prostate Cancer, MRI, GAN, Classification, ADC

1. INTRODUCTION

Prostate Cancer (PCa) is the third most commonly diagnosed cancer[1] among men, with an estimated incidence of 1.3 million new cases worldwide in 2018.[2, 3] Current diagnostic and management methods of PCa rely on prostate-specific antigen (PSA) levels in serum. However, PSA testing comes at the cost of substantial over- diagnosis of clinically significant (cS) lesions, which leads to unnecessary testing such as biopsies and treatment of indolent PCa lesions.[4, 5]

Thanks to recent advances in medical image acquisition, magnetic resonance imaging (MRI) has been pro- posed as an alternative to classic diagnostic methods and found to be particularly useful for triage applications.[6,7] That is, to stratify PCa patients depending on the aggressiveness of the lesions to either provide further testing (or treatment) or to propose an active surveillance program. However, MRI analysis requires specialized training, suffers from inter-reader variability and depends on the reader experience. Moreover, in the absence of specialized training, the analysis of the MRI might be sub-optimal and time-expensive.[8, 9]

Deep learning (DL) techniques have shown promise in tasks such as classification and detection, being successful in several medical applications.[10] Nevertheless, traditional DL-based applications rely on large amounts of annotated data which are rarely available on the medical domain and are expensive and hard to obtain. Moreover, imbalanced data or with low variability might lead to sub-optimal results[11] highlighting even more the necessity for large amounts of annotated and high-quality data. Classic augmentation techniques such as translation and rotation have become the by default approach when dealing with a limited amount of data. However, such techniques produce highly correlated data, limiting the amount of information provided to the algorithm in the training phase.[11]

Generative Adversarial Networks (GAN)[12] have emerged as an alternative to standard augmentation techniques. In particular, conditional GAN (cGAN)[13] allows to generate synthetic data samples of a particular class, which results quite convenient for classification tasks. In addition, deep convolutional GAN (DCGAN)[14] has also seen success in generating synthetic medical images.[11] In this work, we exploit such an ability to generate synthetic samples of a particular class (cS or ncS) to improve the classification ability a DL architecture by increasing the amount and variability of ADC prostate MRI data.

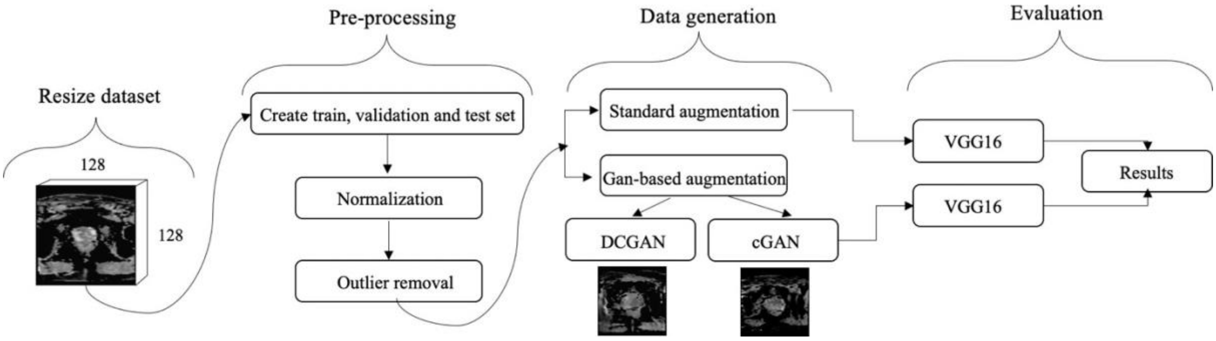


Figure 1: Technical approach to the project.

2. METHODS

2.1 Task definition

Our work is developed in two stages. In the first stage, we synthesize prostate ADC MRI slices with DCGAN or cGAN. In order to perform such a task, we assume an input (x_u, y_u) where yu are the associated labels to the ADC prostate MRI inputs denoted as xu. In the case of DCGAN, we train one model for each class. On the other hand, for cGAN, we train a single model able to generate a specific class given the right label. Following, we obtain a defined number of synthetic samples and their respective labels, denoted as (x_s, y_s). After visual inspection, we make use of the most realistic looking pictures x_s, y_s as an input for a VGG16[15], with the objective of discriminating between cS and ncS prostate ADC MRI - a 2D classification task. We test a different number of generated images with respect to the original ones and their effect on the final results, quantified by the area under the curve (AUC), sensitivity, specificity and accuracy averaged over 3 independent runs.

2.2 GAN augmentation for prostate cancer triage

Our methodology is based on[11, 16]. Figure 1 shows the technical approach to the project. Labelled ADC prostate MRI x_u are used to generate $x_s \sim p_{x_u}$ in which p_{x_u} is the distribution of the original data, by means of DCGAN and cGAN architectures. In particular, DCGAN aims to learn p_{x_u} where p_{x_u} is the distribution of the data for a specific class, namely cS or ncS. On the other hand, cGAN aims to learn $p_{x_u|y_u}$that is, the conditional probability of the original data. Hereby, in the first case two architectures are trained to be able to generate samples from both classes whilst on the other case a single architecture is enough, since it is able to generate samples based on the selected class. Our loss function follows the standard two-player minimax game.[13]

Once the DCGAN and cGAN architectures have been trained, we use the generated samples to train the classification model (VGG16). Examples of generated samples are shown in Figure 2. Following, we compare both architectures in terms of the quality of the generated pictures based on the final task: classification of prostate MRI tumors in cS and ncS, based on AUC, sensitivity, specificity and accuracy.

2.3 Implementation and experiments

2.3.1 Data pre-processing and splitting

We make use of the ProstateX dataset,[17] which is open source*. The cohort included in the study consisted of 204 patients diagnosed with PCa and 330 lesions. Among those lesions, 76 lesions were cS and 254 were ncS. The nature of the study is retrospective and includes different MRI modalities from which ADC is used in this work. Standard pre-processing is applied to the data, including re-sampling to a common coordinate system with the desired dimensions (128x128) and normalization of the MRI intensities to a range of [0, 1]. All the images are provided with results of patients' biopsies, which are used as the reference standard for this work. The significance level of the lesions is based on the Gleason Score (GS) obtained from the biopsies. That is, if GS $\geq$7 the slice (2D) is classified as cS and if GS < 7 the slice (2D) is classified as ncS. The data-set is split following a 70%/10%/20% for training, validation and testing, respectively. The splitting is done by patients, avoiding cross-contamination in the form of data leakage.

*https://wiki.cancerimagingarchive.net/display/Public/SPIE-AAPM-NCI+PROSTATEx+Challenges

2.3.2 Architectures and training

We evaluate the performance of two different GAN-based architectures: DCGAN and cGAN. Both architectures follow a standard implementation, as explained in[13] and,[14] respectively. We experiment with different learning rates and number of iterations and evaluate the quality of the pictures in a qualitative way to determine the best training scheme for both architectures. After experimenting with different configurations, we obtain the final synthetic ADC MRI training for 5000 iterations the DCGAN architecture and for 6000 epochs the cGAN one. Both architectures use a learning rate of 0.0002 and an Adam optimizer. All the training is carried out on a NVIDIA Tesla V100 GPU.

As for the classification stage of the work, we make use of a VGG16 architecture,[15] based on previous results.[18] The architecture is trained for 1000 epochs with an early stopping mechanism of 20 epochs, which halts the training if no improvement is seen in the AUC for the previously defined number of epochs. We train the network with Adam optimizer and a learning rate of 0.001. The training and inference, are, again, carried out on a NVIDIA Tesla V100 GPU.

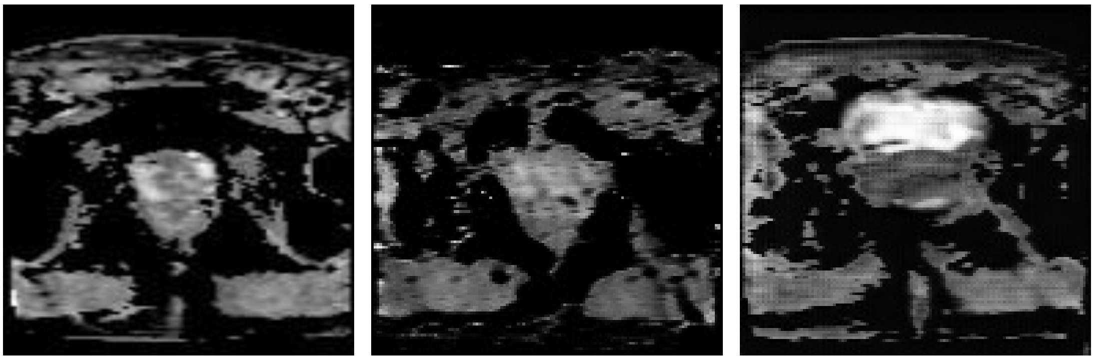


Figure 2: Examples of a real ADC image (left), generated ADC MRI with cGAN (middle) and DCGAN (right).

2.3.3 Evaluation of the results

Since the main objective of the work is to improve the lesion classification results by means of synthesized ADC maps, we evaluate the quality of the generated pictures by the different GAN architectures by means of the classification results obtained with them. In particular, we first train a plain VGG16 and use those results as the baseline for our work. Following, we experiment with different classic augmentation techniques, namely: rotation (25 degrees), translation (0.4 pixels) and vertical flipping, based on.[18] The same configuration in terms of hyperparameters is used for all the comparisons. In the second evaluation stage we experiment with different amounts of generated data by the GAN architectures. That is, we increase the data available in the training set by a certain % (50 and 100). In the case of a 50% increase we would generate half of the amount of cS and ncS data present in the original training set and add the generated samples to it whilst for 100% we would generate the same amount of cS and ncS samples present in the original training set and add them to it. Lastly, we evaluate a "balanced configuration" in which we generate and add the necessary amount of cS and ncS samples such that the presence of both classes is balanced in the training set. Finally, we present comparisons in terms of the results obtained with GAN-based augmentation (both DCGAN and cGAN), classic augmentation and a combination of both. The evaluation of the classic augmentation and GAN-based augmentation is carried out with the GAN architecture that obtained the best results as a standalone technique. All in all, the results are presented in terms of the following parameters:

- **Baseline**: VGG16 without augmentation is considered to be the baseline of this work. Further experiments use the same configuration for the architecture and the same architecture choice.
- **Classic augmentation**: We experiment with translation (0.4), rotation (25 degrees) and vertical flipping.
- **GAN architectures**: We experiment with cGAN and DCGAN. We evaluate the quality of the generated images based on the results obtained in the final 2D lesion classification task.

- **% of generated data**: We test different amounts of generated data with respect to the original one x_u (i.e., 100% implies generating the same amount of data as the original amount present in the original training set and 50% would generate half of the amount present in the original training set), including a balanced configuration in which we generate the necessary amount of synthetic samples for each class such that we end up with the same amount of samples in both classes at training time.

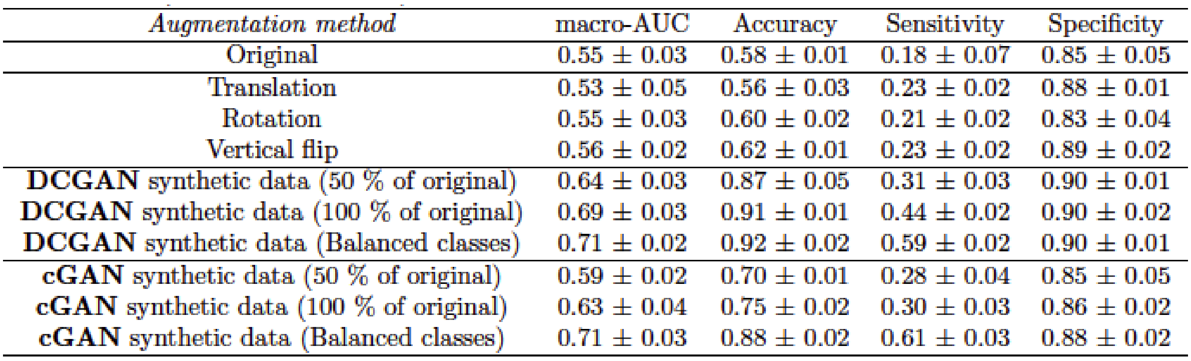
3. RESULTS

In this section, we provide the results obtained with the proposed GAN-augmentation technique. As shown in Table 1, both the cGAN and DCGAN-generated synthethic samples have a positive effect on the final macro-AUC when using VGG16 as the classication architecture. We see an increase in the averaged macro-AUC for both DCGAN and cGAN. In particular, the increase becomes notably larger when more synthetic data is used (100%). We can also see the classifer struggles to classify cS lesions based on the sensitivity results but there is a considerable improvement when balancing the synthesis of classes. Generally speaking, DCGAN obtains slightly higher results in terms of macro-AUC when compared to cGAN ones. Nevertheless, the results com at the expense of the need of two specific DCGAN architectures: one for each lesion class.

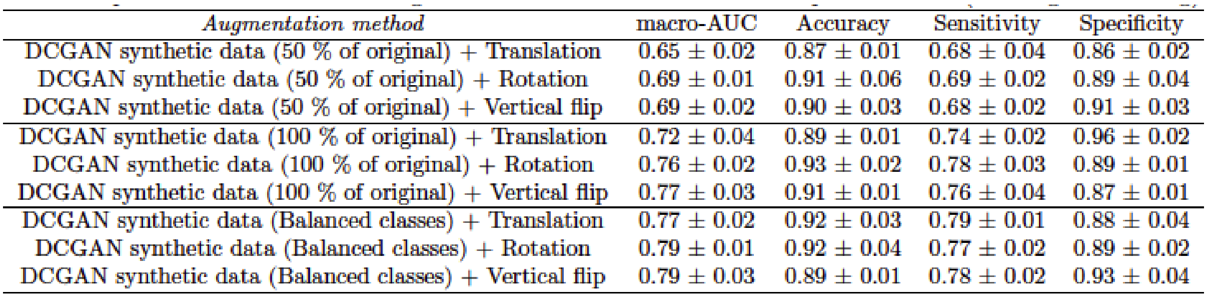
As Table 2 shows, combining both classic augmentation and synthetic samples further increases the final AUC, showing that both techniques can be used in a complementary way. The results also highlight the add-on nature of synthetic data, which can be added on top of more complex architectures and other methodologies. In particular, we observe how the combination of DCGAN and rotation outperforms all the other approaches by a considerable margin, whilst keeping a good balance in the classification of both classes.

When comparing the final results presented in Table 2 with other approaches, we observe that our method, in spite of the simplicity of the architecture used to carry out all the testing reaches competitive results. Other works obtain an AUC of 0.78 with the same architecture and fine-tuning,18 or AUC of 0.80 by combining different modalities and a 2D and 3D approach,19 whilst the leading results in the ProstateX challenge reach AUC's of 0.95 but make use of all the available data (that is, not only ADC). The results presented in this work show the usefulness of synthetic data as a complementary approach to classic augmentation techniques and can be used as an extension of existing works, thus not being limited to the specific architecture used in this work.

Table 1: Classic augmentation techniques and synthetic data effect on prostate triage results, based on 3independent runs (training and testing) reported in terms of mean ± standard deviation.



Augmentation method	macro-AUC	Accuracy	Sensitivity	Specificity
Original	0.55 ± 0.03	0.58 ± 0.01	0.18 ± 0.07	0.85 ± 0.05
Translation	0.53 ± 0.05	0.56 ± 0.03	0.23 ± 0.02	0.88 ± 0.01
Rotation	0.55 ± 0.03	0.60 ± 0.02	0.21 ± 0.02	0.83 ± 0.04
Vertical flip	0.56 ± 0.02	0.62 ± 0.01	0.23 ± 0.02	0.89 ± 0.02
DCGAN synthetic data (50 % of original)	0.64 ± 0.03	0.87 ± 0.05	0.31 ± 0.03	0.90 ± 0.01
DCGAN synthetic data (100 % of original)	0.69 ± 0.03	0.91 ± 0.01	0.44 ± 0.02	0.90 ± 0.02
DCGAN synthetic data (Balanced classes)	0.71 ± 0.02	0.92 ± 0.02	0.59 ± 0.02	0.90 ± 0.01
cGAN synthetic data (50 % of original)	0.59 ± 0.02	0.70 ± 0.01	0.28 ± 0.04	0.85 ± 0.05
cGAN synthetic data (100 % of original)	0.63 ± 0.04	0.75 ± 0.02	0.30 ± 0.03	0.86 ± 0.02
cGAN synthetic data (Balanced classes)	0.71 ± 0.03	0.88 ± 0.02	0.61 ± 0.03	0.88 ± 0.02



Augmentation method	macro-AUC	Accuracy	Sensitivity	Specificity
DCGAN synthetic data (50 % of original) + Translation	0.65 ± 0.02	0.87 ± 0.01	0.68 ± 0.04	0.86 ± 0.02
DCGAN synthetic data (50 % of original) + Rotation	0.69 ± 0.01	0.91 ± 0.06	0.69 ± 0.02	0.89 ± 0.04
DCGAN synthetic data (50 % of original) + Vertical flip	0.69 ± 0.02	0.90 ± 0.03	0.68 ± 0.02	0.91 ± 0.03
DCGAN synthetic data (100 % of original) + Translation	0.72 ± 0.04	0.89 ± 0.01	0.74 ± 0.02	0.96 ± 0.02
DCGAN synthetic data (100 % of original) + Rotation	0.76 ± 0.02	0.93 ± 0.02	0.78 ± 0.03	0.89 ± 0.01
DCGAN synthetic data (100 % of original) + Vertical flip	0.77 ± 0.03	0.91 ± 0.01	0.76 ± 0.04	0.87 ± 0.01
DCGAN synthetic data (Balanced classes) + Translation	0.77 ± 0.02	0.92 ± 0.03	0.79 ± 0.01	0.88 ± 0.04
DCGAN synthetic data (Balanced classes) + Rotation	0.79 ± 0.01	0.92 ± 0.04	0.77 ± 0.02	0.89 ± 0.02
DCGAN synthetic data (Balanced classes) + Vertical flip	0.79 ± 0.03	0.89 ± 0.01	0.78 ± 0.02	0.93 ± 0.04

Table 2: Results for DCGAN synthetically generated data in combination with classic augmentation results.Results are presented in terms of average and standard deviation for 3 independent runs (training and testing).

4. CONCLUSION

We presented a GAN-based augmentation approach as an alternative to classic augmentation techniques. Our results show that our proposed approach outperforms classic augmentation in a 2D prostate cancer lesion classification and when using VGG16 as the classification architecture choice. In particular, we observe that DCGAN- generated data produces better results in the final classification task at the cost of requiring two class-specific architectures to generate samples for each class. Moreover, we show that classic augmentation and GAN-based augmentation can be used in a complementary way to improve the results obtained with classic augmentation techniques. Our results show that GAN-generated samples have the potential to help with data scarcity in medical applications and that can be used on top of a given architecture and task to improve the final results by increasing the amount of available data to train the model.

5. ABOUT THE WORK

The work has not been submitted nor is planned to be submitted anywhere else.

REFERENCES

[1] T˘ataru, O. S., Vartolomei, M. D., Rassweiler, J. J., Virgil, O., Lucarelli, G., Porpiglia, F., Amparore, D., Manfredi, M., Carrieri, G., Falagario, U., et al., "Artificial intelligence and machine learning in prostate cancer patient management—current trends and future perspectives," Diagnostics 11(2), 354 (2021).

[2] Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., and Jemal, A., "Global cancer statistics 2018," CA: A Cancer Journal for Clinicians 68(6), 394–424 (2018).

[3] Fernandez-Quilez, A., Germ´an-Borda, M., Leonardo, G., Castellanos, N., Soennesyn, H., Oppedal, K., and Reidar-Kjosavik, S., "Prostate cancer screening and socioeconomic disparities in mexican older adults," salud pu´blica de m´exico 62(2, Mar-Abr), 121–122 (2020).

[4] Barry, M. J., "Screening for prostate cancer — the controversy that refuses to die," *New England Journal of Medicine* **360**, 1351–1354 (Mar 2009).

[5] Hodge, K. K., McNeal, J. E., and Stamey, T. A., "Ultrasound guided transrectal core biopsies of the palpablyabnormal prostate," *Journal of Urology* **142**, 66–70 (Jul 1989).

[6] Dirix, P., Bruwaene, S. V., Vandeursen, H., and Deckers, F., "Magnetic resonance imaging sequences for prostate cancer triage: two is a couple,

three is a crowd?," *Translational Andrology and Urology* **8**, S476–S479 (Dec 2019).

[7] Lomas, D. J. and Ahmed, H. U., "All change in the prostate cancer diagnostic pathway," *Nature Reviews Clinical Oncology* **17**, 372–381 (Feb 2020).

[8] Turkbey, B., Rosenkrantz, A. B., Haider, M. A., Padhani, A. R., Villeirs, G., Macura, K. J., Tempany, C. M., Choyke, P. L., Cornud, F., Margolis, D. J., and et al., "Prostate imaging reporting and data system version 2.1: 2019 update of prostate imaging reporting and data system version 2," *European Urology* **76**, 340–351 (Sep 2019).

[9] Gaziev, G., Wadhwa, K., Barrett, T., Koo, B. C., Gallagher, F. A., Serrao, E., Frey, J., Seidenader, J., Carmona, L., Warren, A., et al., "Defining the learning curve for multiparametric magnetic resonance imag-ing (mri) of the prostate using mri-transrectal ultrasonography (trus) fusion-guided transperineal prostate biopsies as a validation tool," *BJU international* **117**(1), 80–86 (2016).

[10] Bohr, A. and Memarzadeh, K., "The rise of artificial intelligence in healthcare applications," in [*Artificial Intelligence in healthcare*], 25–60, Elsevier (2020).

[11] Fernandez-Quilez, A., Larsen, S. V., Goodwin, M., Gulsrud, T. O., Kjosavik, S. R., and Oppedal, K., "Improving prostate whole gland segmentation in t2-weighted mri with synthetically generated data," in [*2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*], 1915–1919, IEEE (2021).

[12] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio,Y., "Generative adversarial nets," *Advances in neural information processing systems* **27** (2014).

[13] Mirza, M. and Osindero, S., "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784* (2014).

[14] Radford, A., Metz, L., and Chintala, S., "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434* (2015).

[15] Simonyan, K. and Zisserman, A., "Very deep convolutional networks

for large-scale image recognition," *arXiv preprint arXiv:1409.1556* (2014).

[16] Frid-Adar, M., Diamant, I., Klang, E., Amitai, M., Goldberger, J., and Greenspan, H., "Gan-based syn- thetic medical image augmentation for increased cnn performance in liver lesion classification," *Neurocomputing* **321**, 321–331 (2018).

[17] Litjens, G., Debats, O., Barentsz, J., Karssemeijer, N., and Huisman, H., "Computer-aided detection of prostate cancer in mri," *IEEE transactions on medical imaging* **33**(5), 1083–1092 (2014).

[18] Fernandez-Quilez, A., Eftestøl, T., Goodwin, M., Kjosavik, S. R., and Oppedal, K., "Self-transfer learning via patches: A prostate cancer triage approach based on bi-parametric mri," *arXiv preprint arXiv:2107.10806* (2021).

[19] Mehrtash, A., Sedghi, A., Ghafoorian, M., Taghipour, M., Tempany, C. M., Wells III, W. M., Kapur, T., Mousavi, P., Abolmaesumi, P., and Fedorov, A., "Classification of clinical significance of mri prostate findings using 3d convolutional neural networks," in [Medical Imaging 2017: Computer-Aided Diagnosis], 10134, 101342A, International Society for Optics and Photonics (2017).

Contribution C

One class to rule them all: Detection and classification of prostate tumors presence in bi-parametric MRI based on auto-encoders

Alvaro Fernandez-Quilez[a, b, e, *,] Habib Ullah[b, c, *], Trygve Eftestøl[c], Svein Reidar Kjosavik[d], and Ketil Oppedal[b, e]

[a]Department of Quality and Health Technology, University of Stavanger, Stavanger, Norway.
[b]Stavanger Medical Imaging Laboratory (SMIL), Department of Radiology, Stavanger University Hospital, Stavanger, Norway.
[c]Department of Electrical Engineering and Computer Science, University of Stavanger, Stavanger, Norway.
[d]General Practice and Care Coordination Research Group, Stavanger University Hospital, Stavanger, Norway.
[e]Centre for Age-Related Medicine, Stavanger University Hospital, Stavanger, Norway.

[*]Shared first authorship.

ABSTRACT

Prostate Cancer (PCa) is the fifth leading cause of death and the second most common cancer diagnosed among men worldwide. Current diagnostic practices suffer from a substantial overdiagnosis of indolent tumors. Deep Learning (DL) holds promise in automatizing prostate MRI analysis and enabling computer-assisted systems able to improve current practices. Nevertheless, large amounts of annotated data are commonly required for DL systems success. On the other hand, an experienced clinician is typically able to discern between a normal (no lesion) and an abnormal (contains PCa lesions) case after seeing a few normal cases, ultimately reducing the amount of data required to detect abnormal cases. This work exploits such an ability by making use of normal cases at training time and learning their distribution through auto-encoder-based architectures. We propose to use a threshold approach based on interquartile range to discriminate between normal and abnormal cases at evaluation time, quantified through the area under the curve (AUC). Furthermore, we show the ability of our method to detect lesions in those cases deemed as abnormal in an unsupervised way in T2w and apparent diffusion coefficient maps (ADC) MRI modalities.

Keywords: Auto-encoders, Prostate, Unsupervised, Detection, MRI

1. INTRODUCTION

Prostate Cancer (PCa) is the second most commonly diagnosed cancer1 and one of the leading causes of death worldwide.[2,3] Current diagnostic and management methods of PCa rely on prostate-specific antigen (PSA) levels in serum. However, PSA use remains as a controversial topic due to unclear benefits of it as a screening technique and a substantial under-diagnosis of aggressive tumors as well as over-diagnosis of indolent tumors, leading to unnecessary biopsies and treatment of the indolent lesions.[4, 5]

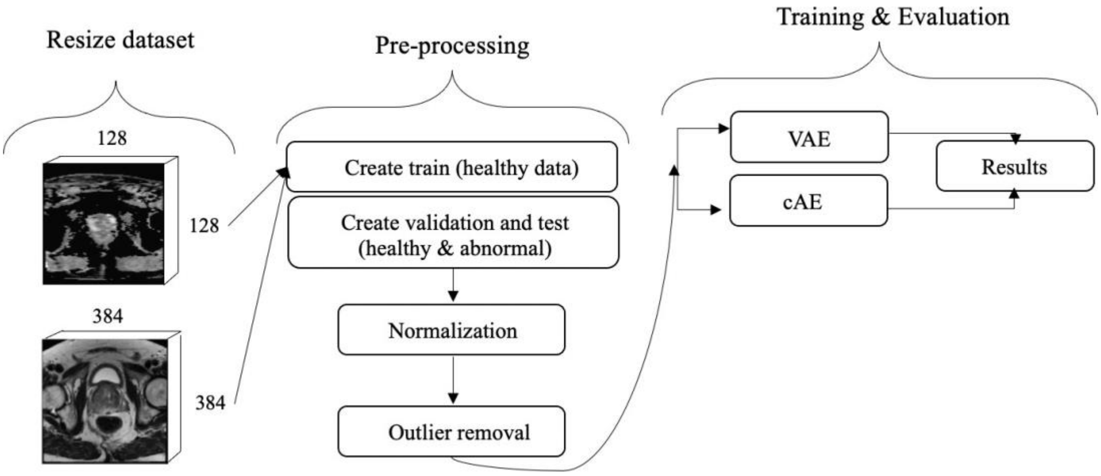


Figure 1: Technical approach to the project.

Magnetic resonance imaging (MRI) id s non-invasive technique that has shown potential to improve the current PCa diagnostic and management pathway.[6, 7] In addition, MRI has proven to be a highly valuable tool for lesion detection and treatment planning.[8] Nevertheless, analysis of MRI suffers from inter-reader variability and sub-optimal interpretation in the absence of specialized training. Furthermore, it can be time-expensive.[9] Deep learning (DL) techniques have emerged as an alternative able to automatize MRI analysis, showing its potential in other studies.[10, 11] However, traditional DL-based applications rely on large amounts of annotated data which are rarely available in the medical domain. On the other hand, experts are able discern abnormal (contains PCa lesion) cases from normal (no lesion) ones as well as to detect lesions after seeing a handful of control cases, just by mere comparison and even when no extensive specialized training is present.[12] Furthermore, the same expert is usually able to carry out several tasks such as detection and classification of MRI slices. Motivated by it, we focus on an unsupervised detection of tumors and classification of slices in abnormal or normal by learning the prior distribution of normal prostate MRI slices.

Convolutional auto-encoders (cAE)[13] and models related to it, such as variational auto-encoder (VAE)[14] have been successful in tasks such as outlier detection and high-dimensional data compression,[15] which are closely related to the task presented in this work. Auto-encoder models are of particular interest in this work due to their ability of approximating the likelihood of a given data point with respect to the learnt distribution from the data points they were trained on.

In an effort to palliate the lack of annotated data and mimic radiologists' behavior, we investigate VAE and cAE architectures to detect prostate lesions and stratify between abnormal and normal cases in ADC and T2w MRI modalities, while exclusively making use of normal data at training time. Following, we make use of the learnt distributions and show in our experimental evaluation the potential of our approach for lesion presence classification and detection in both ADC and T2w prostate MRI slices.

2. METHODS

2.1 Task definition

The unsupervised detection of lesions and classification between abnormal and normal MRI slices is performed in two stages using the same architecture for both tasks. We assume an input X_h = (x_{h_1} , x_{h_2} ..., x_{h_N}), where x_{h_i} , i = 1...*N* are normal (healthy) T2w or ADC prostate MRI slices (2D) with a fixed resolution. Our ultimate goal is to learn the distribution $p(X_h)$ through an auto-encoder architecture. The rationale behind the process is that we hypothesize that models will not be able to reconstruct abnormal images accurately, due to the fact that they have only been trained with control ones, hereby learning the non-anomalous data distribution as a prior $p(X_h)$.

Once the prior distribution has been learnt, we use a mix of 2D prostate MRI slices as an input for the already trained model. The data consists of slices that contain tumors, denoted as X_a = (x_{a_1}, x_{a_2} ..., x_{a_N}) and normal images, which are only used for validation purposes. Such a process allows us to obtain an estimate of the mean squared error (MSE) or structural similarity index (SSIM) distribution of X_a and of X_h, which is further use to determine a classification threshold t_{MSE} or t_{SSIM} to distinguish between abnormal and normal prostate MRI slices. Following, lesion regions are detected through pixel-wise difference between the MRI slices deemed as abnormal and the reconstruction obtained after using them as an input for the already trained auto-encoder Finally, some post-processing is applied to the detected lesions in the form of threshold application, such that MSE <= ϵ is set to 0. The threshold is obtained, again, using the MSE distribution of the difference between the original MRI slice and their reconstruction.

2.2 Unsupervised detection and lesion presence classification

Our methodology is based on.[12] The technical approach to the project can be found in Figure 1. We perform the anomaly classification and lesion detection in two stages with the same architecture. In the first stage, normal ADC or T2w MRI slices of the prostate (2D) are used to learn the distribution $p(\mathrm{X}_\mathrm{h})$ through an auto-encoder, either VAE or cAE. In order to achieve such objective, we experiment with MSE (Equation 1) and SSIM (Equation 2)[16] as our choice for the metrics used to quantify the quality of the reconstruction:

$$MSE = \sum_{k=1}^{N} \left(x_{h_k} - \hat{x}_{h_k}\right)^2 \qquad (1)$$

$$SSIM = \sum_{k=1}^{N} l\left(x_{h_k} - \hat{x}_{h_k}\right) c\left(x_{h_k} - \hat{x}_{h_k}\right) s\left(x_{h_k} - \hat{x}_{h_k}\right) \qquad (2)$$

Where *c* represents the contrast, *l* the luminance, *s* the structure of the images, $\hat{x}_{h_k}$represents the *k*th reconstructed image and *N* is the number of images in the batch under consideration.

In the second stage, we make use of the already trained auto-encoder[12] and the validation set composed by a mix of normal and abnormal cases with the objective of obtaining an estimate of the optimal threshold t_MSE or t_SSIM, based on the distribution of MSE or SSIM (Figure 2) for each MRI modality individually (ADC or T2w). In essence, such a threshold is ultimately used to distinguish prostate MRI slices that are normal from those that are abnormal so that we are able to further analyse the abnormal ones. The threshold value is obtained by means of interquartile range (IQR),[17] which does not assume a specific underlying distribution. All the images with an associated MSE > t_MSE or SSIM > t_SSIM are deemed as abnormal (Figure 2). Finally, those cases deemed as abnormal are further analysed and a pixel-wise intensity difference is computed between the reconstructed and the original images, which were used as an input for the trained auto-encoder (Figure 3). Again, a threshold is computed by means of IQR and those regions where the error is larger than the obtained threshold are considered to be abnormal regions and therefore, regions that might potentially contain a lesion.

2.3 Implementation and experiments

2.3.1 Data pre-processing and splitting

We make use of the ProstateX dataset,[18] which is open source*. The cohort included in the study consisted of 204 patients diagnosed with PCa and 330 lesions. Among those lesions, 76 lesions were aggressive and 254 were indolent. The nature of the study is retrospective and includes different MRI modalities from which axial T2w and ADC are used in this work. Standard pre-processing is applied to the data, including re-sampling to a common coordinate system with the desired dimensions (384x384 for T2w and 128x128 for ADC, respectively) and normalization of the MRI intensities to a range of [0, 1].[19] All the images are provided with results of the patients' biopsies. We use publicly available lesion masks[20] as the ground truth for the lesion detection and anomaly classification (slices are labelled as anomalous if they contain a lesion, otherwise they are considered normal). We split the original data-set by patients, avoiding cross-contamination in the form of data leakage. The data-set is split following a 70%/10%/20% for training, validation and testing, respectively. It is worth mentioning that as stated in previous sections, only normal slices (without a lesion) are used during the training phase whilst both slices with lesions and without them are used for validation and testing phases.

*https://wiki.cancerimagingarchive.net/display/Public/SPIE-AAPM-NCI+PROSTATEx+Challenges

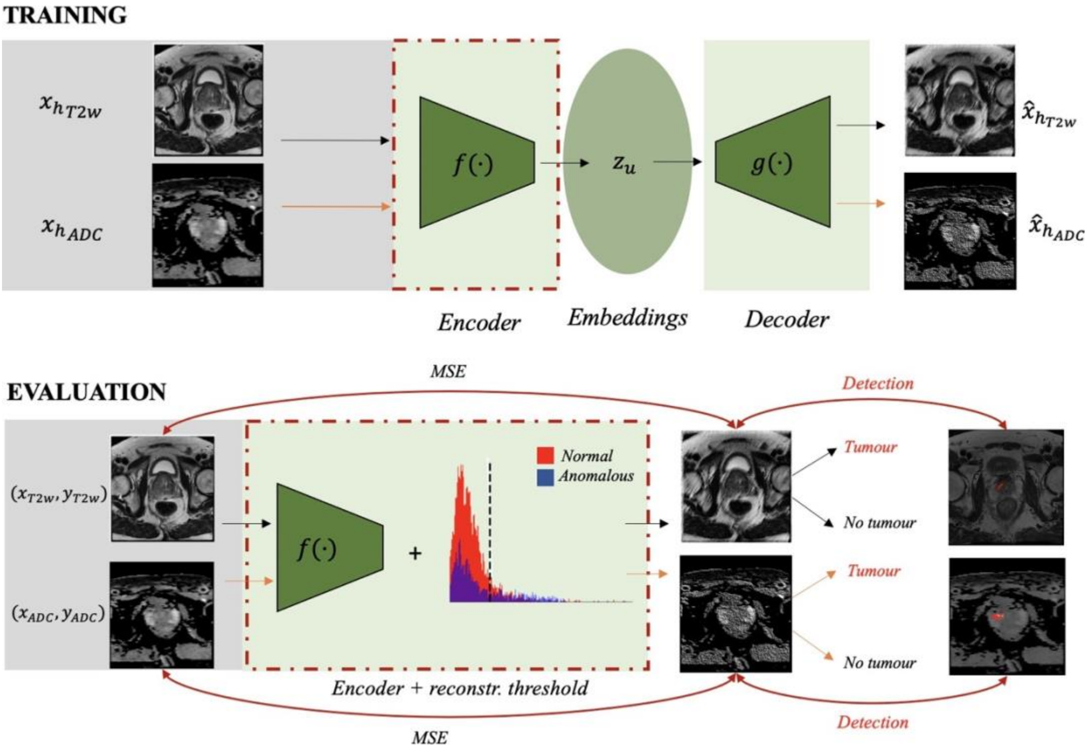


Figure 2: Training and evaluation of the proposed methodology.

2.3.2 Architectures and training

We experiment with two auto-encoder-based architectures: cAE and VAE. Generally speaking, both architectures follow an encoder-decoder structure. In particular, cAE obtains a low-dimensional representation (embedding) of the input and tries to reconstruct the original image from its low-dimensional projection (Figure 2). Our cAE architecture follows a U-net like structure,21 including 2D convolution filters followed by a rectified linear unit (ReLU), pooling operations and dense layers. The architecture is trained using a binary-cross entropy loss function for both ADC and T2w, with a learning rate of 1e−4, 1000 epochs and batch size of 32.

Similarly to cAE, our VAE architecture follows a U-net-like architecture composed by 2D convolution filters, ReLU, pooling operations and dense layers. Nevertheless, VAE includes a latent inference enabled by stochastic sampling in the latent space. The model is trained by minimizing a Kullback-

Leibler (KL) divergence with a learning rate of 1e−4, 1000 epochs and batch size of 32. All the architectures are trained using an NVIDIA v100 GPU.

2.3.3 Evaluation of the results

In order to evaluate our proposed methodology, we provide a comparison between SSIM and MAE as the choice for the image reconstruction quality and loss function for the proposed auto-encoder architectures. In addition to it, several configurations for the encoder-decoder structure of cAE and VAE are tested with different latent dimensions and convolutional layers, from which we report the results obtained with the best of them for every MRI modality. Final results are reported in terms of area under the curve (AUC), sensitivity and accuracy percentage to quantify the discrimination ability of our method and in terms of MSE and qualitative results to show the detection ability of it. All in all, the results are reported in terms of the following parameters:

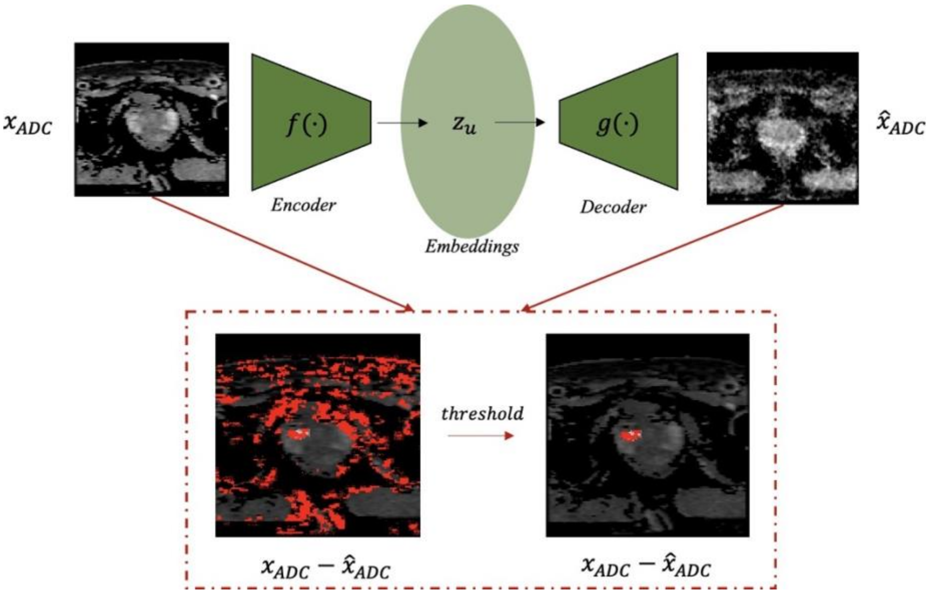


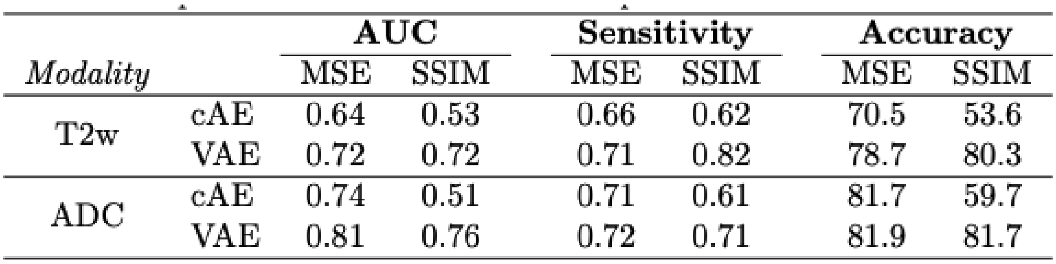
Figure 3: Example of threshold application with an MSE distribution and VAE architecture to obtain an ADC lesion

• **Architectures**: cAE and VAE are explored for both classification and detection in T2w and ADC modalities. Different encoder-decoder structures are also tested for both cAE and VAE, including but not limited to a different number of convolutional layers and dimensions of the embedding space.
• **Image reconstruction**: We experiment with MSE and SSIM as the choice for the metrics to quantify the image reconstruction quality and as a loss function and report the results for both of them.

Following, we provide some details on the effect of the chosen thresholds to classify MRI slices between abnormal (containing a PCa lesion) and normal (no lesion). It is worth insisting in the fact that the choice of the threshold is done using the validation set, which consists of a mix of normal and abnormal cases. Figure 4 shows some examples of threshold calculations. At testing time, we make use again of a mix of normal and abnormal cases whilst during training only healthy cases are used.

3. RESULTS

We start by evaluating our results in terms of the discrimination ability of both cAE and VAE with different reconstruction metrics (SSIM and MSE) in terms of AUC, accuracy and sensitivity.



		AUC		**Sensitivity**		**Accuracy**	
Modality		MSE	SSIM	MSE	SSIM	MSE	SSIM
T2w	cAE	0.64	0.53	0.66	0.62	70.5	53.6
	VAE	0.72	0.72	0.71	0.82	78.7	80.3
ADC	cAE	0.74	0.51	0.71	0.61	81.7	59.7
	VAE	0.81	0.76	0.72	0.71	81.9	81.7

Table 1: Comparison between different unsupervised classification methods.

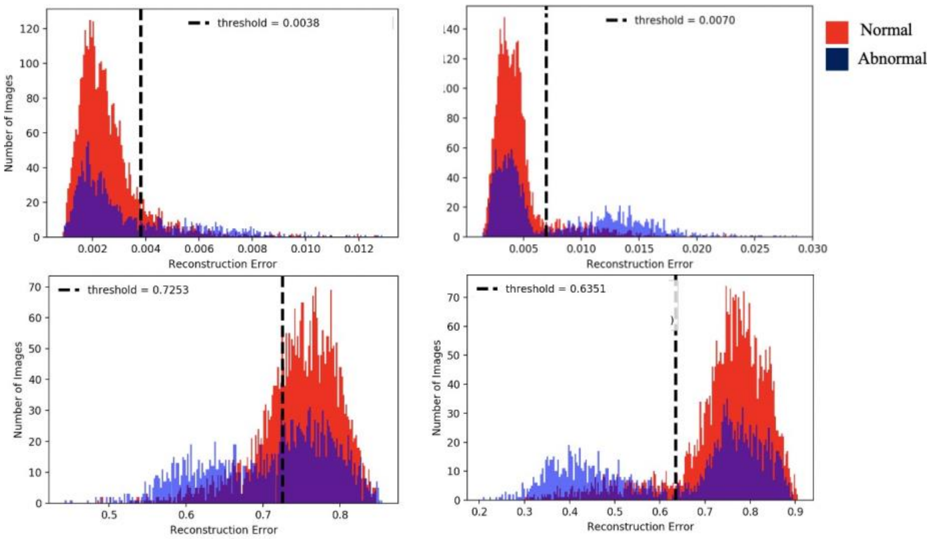


Figure 4: Threshold calculation using interquartile range for MSE error distribution (top row) and T2w (first column) and SSIM error distribution (bottom row) and ADC (second column).

As it can be observed in Table 1, MSE obtains a better performance overall. Specifically, the best results are achieved with VAE: an AUC of 0.81 for ADC and an AUC of 0.72 for T2w.

Our method achieves a good balance between false positives and false negatives, as depicted in Figure 5. Regarding the effect of the threshold in the final results, we include a comparison of the different results obtained with a variation in increments (or decrements) of 0.0005 and0.002 of the optimal value in Figure 6 for MSE results. As depicted in the figure, both precision and recall are significantly affected by the threshold choice highlighting the importance of it in the design of our proposed solution, especially for PCa classification given the relevance of the false positives.

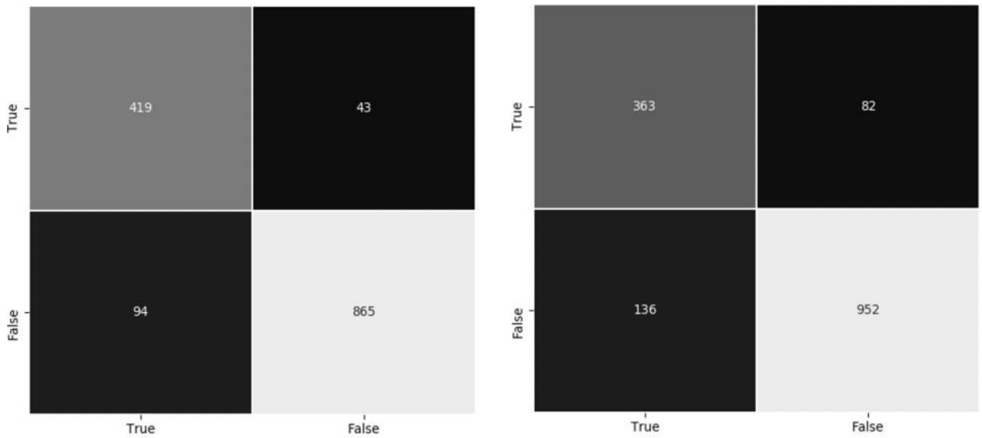


Figure 5: Confusion matrix for the optimal classification threshold. Left side depicts the results for ADC and right side for T2w

Finally, we present qualitative results of the lesion detection obtained by the proposed system for both ADC and T2w, when the MRI slice under consideration is classified as anomalous. Figure 7 shows 3 examples (bottom row) of T2w lesion detection with VAE and MSE for cases that were correctly classified as anomalous whilst the top row shows 3 examples of ADC lesion detection for VAE and MSE. From a qualitative point of view, the detection quality is reasonably good after applying a threshold based, again, on MSE distribution and IQR. In our quantitative results we observe that the MSE distribution around the lesion area is notably higher than in the other areas of the MRI slices, being the averaged MSE of 0.687 for the lesion areas in T2w and of 0.167 for the rest of the slice whilst for ADC is of 0.727 for the lesion areas and of 0.172 for the rest of the slice.

When compared to other approaches, our method presents the advantage of requiring only normal cases during training, which are easier to obtain in the medical domain. Moreover, we make use of the same architecture for both classification and detection, moving away from two-stage systems that require several specialized architectures to achieve the same objective. In particular, our method achieves competitive results when compared to fully supervised and with all classes available such as *FocalNet* [22] (0.81 AUC) or 2-stage 2D U-net (0.86 AUC),[23] but with the previously mentioned advantages.

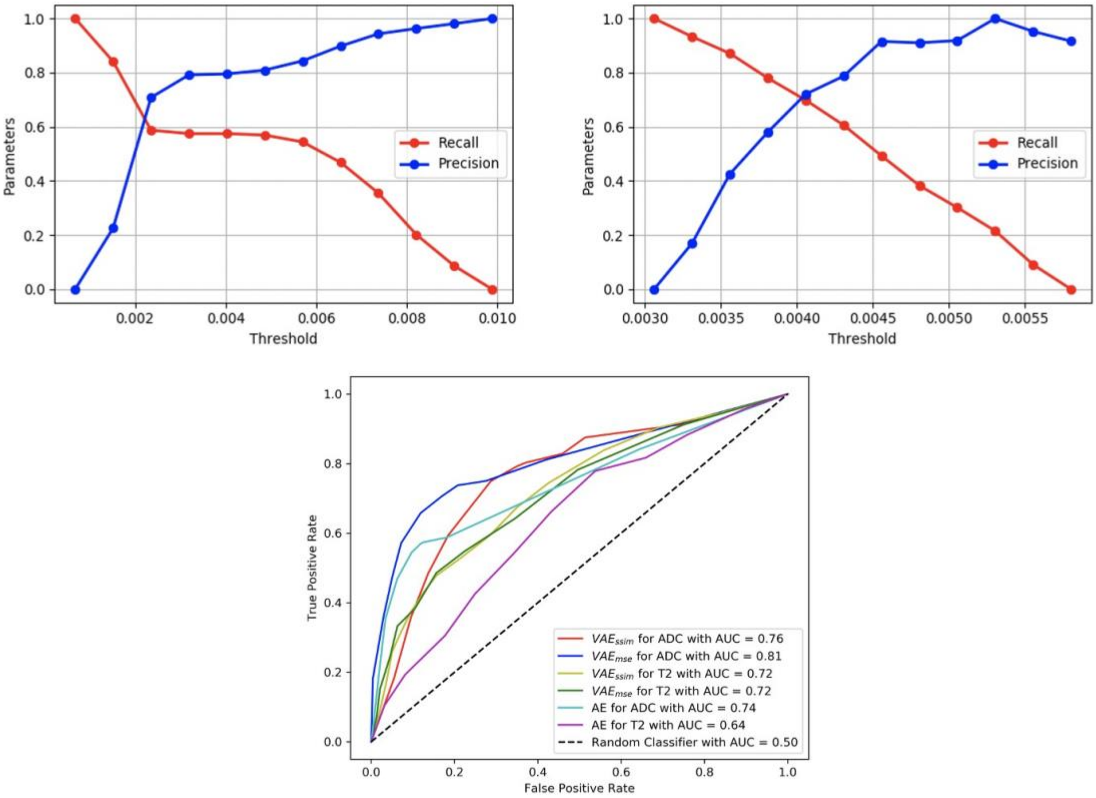


Figure 6: Effect of classification threshold on precision and recall for ADC (left) and T2w (right), with a VAE architecture and MSE metric (top row). Bottom row depicts the AUC for the best cAE configuration and MSE along with VAE results for both modalities.

4. CONCLUSIONS AND REMARKS

We presented an unsupervised prostate lesion detection approach based on auto-encoders. Our VAE-based approach outperformed the cAE one by a large margin in both prostate MRI modalities: ADC and T2w. Our

results show that our approach has the potential to reduce the amount of data needed to perform such tasks by making use of healthy data exclusively during training time, with competitive results when compared to their fully supervised counterpart. Our results suggest that

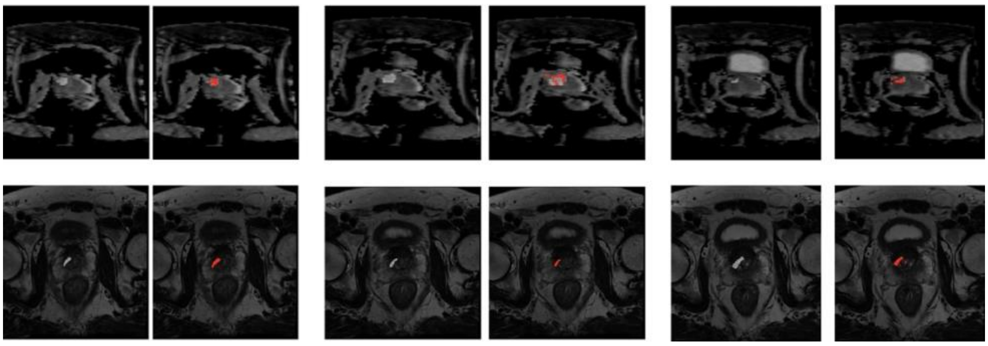


Figure 7: Lesion detection results after threshold application for slices that were correctly classified as abnormal. Top row depicts ADC and bottom row T2w, respectively.

the methodology could be applied to a broader spectrum of MRI PCa applications, such as quality control of segmentation results in an unsupervised way or segmentation of the prostate gland. Our work presents several limitations, being the retrospective nature of the study one of them. Experimentation with more complex auto-encoder architectures and a quantitative way to evaluate the quality of the lesion detection is reserved for future work. Additionally, a better assessment of the impact of the wrong slice classification and subsequent lesion detection will be carried out in future works. Moreover, a combination of both ADC and T2w modalities could be of interest, as other works have shown better results in different applications when using a combination of the MRI modalities instead of using them independently.

5. ABOUT THE WORK

The work has not been submitted nor is planned to be submitted anywhere else.

REFERENCES

[1] Siegel, R. L., Miller, K. D., and Jemal, A.,"Cancer statistics, 2020," CA: A Cancer Journal for Clinicians 70(1), 7–30 (2020).

[2] Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., and Jemal, A., "Global cancer statistics 2018," CA: A Cancer Journal for Clinicians 68(6), 394–424 (2018).

[3] Fernandez-Quilez, A., German-Borda, M., Leonardo, G., Castellanos, N., Soennesyn, H., Oppedal, K., and Reidar-Kjosavik, S., “Prostate cancer screening and socioeconomic disparities in mexican older adults,” salud publica de m´exico 62(2, Mar-Abr), 121–122 (2020).

[4] Barry, M. J., “Screening for prostate cancer — the controversy that refuses to die,” New England Journal of Medicine 360, 1351–1354 (Mar 2009).

[5] Hodge, K. K., McNeal, J. E., and Stamey, T. A., “Ultrasound guided transrectal core biopsies of the palpably abnormal prostate,” Journal of Urology 142, 66–70 (Jul 1989).

[6] Dirix, P., Bruwaene, S. V., Vandeursen, H., and Deckers, F., “Magnetic resonance imaging sequences for prostate cancer triage: two is a couple, three is a crowd?,” Translational Andrology and Urology 8, S476–S479 (Dec 2019).

[7] Lomas, D. J. and Ahmed, H. U., “All change in the prostate cancer diagnostic pathway,” Nature Reviews Clinical Oncology 17, 372–381 (Feb 2020).

[8] Shukla-Dave, A. and Hricak, H., “Role of mri in prostate cancer detection,” NMR in Biomedicine 27(1), 16–24 (2014).

[9] Turkbey, B., Rosenkrantz, A. B., Haider, M. A., Padhani, A. R., Villeirs, G., Macura, K. J., Tempany,

C. M., Choyke, P. L., Cornud, F., Margolis, D. J., and et al., “Prostate imaging reporting and data system version 2.1: 2019 update of prostate imaging reporting and data system version 2,” European Urology 76, 340–351 (Sep 2019).

[10] Fernandez-Quilez, A., Eftestøl, T., Goodwin, M., Kjosavik, S. R., and Oppedal, K., “Self-transfer learning via patches: A prostate cancer triage approach based on bi-parametric mri,” arXiv preprint arXiv:2107.10806 (2021).

[11] Saha, A., Hosseinzadeh, M., and Huisman, H., “End-to-end prostate cancer detection in bpmri via 3d cnns: Effect of attention mechanisms, clinical priori and decoupled false positive reduction,” arXiv preprint arXiv:2101.03244 (2021).

[12] Chen, X. and Konukoglu, E., “Unsupervised detection of lesions in brain mri using constrained adversarial auto-encoders,” arXiv preprint arXiv:1806.04972 (2018).

[13] Masci, J., Meier, U., Ciresan, D., and Schmidhuber, J., “Stacked convolutional auto-encoders for hierarchical feature extraction,” in

[International conference on artificial neural networks], 52–59, Springer (2011).

[14] Kingma, D. P. and Welling, M., “Auto-encoding variational bayes,” arXiv preprint arXiv:1312.6114 (2013).

[15] Kiran, B. R., Thomas, D. M., and Parakkal, R., “An overview of deep learning-based methods for unsuper- vised and semi-supervised anomaly detection in videos,” Journal of Imaging 4(2), 36 (2018).

[16] Renieblas, G. P., Nogu´es, A. T., Gonz´alez, A. M., Le´on, N. G., and Del Castillo, E. G., “Structural similarity index family for image quality assessment in radiological images,” Journal of medical imaging 4(3), 035501 (2017).

[17] Vinutha, H., Poornima, B., and Sagar, B., “Detection of outliers using interquartile range technique from intrusion dataset,” in [Information and Decision Sciences], 511–518, Springer (2018).

[18] Litjens, G., Debats, O., Barentsz, J., Karssemeijer, N., and Huisman, H., “Computer-aided detection of prostate cancer in mri,” IEEE transactions on medical imaging 33(5), 1083–1092 (2014).

[19] Fernandez-Quilez, A., Larsen, S. V., Goodwin, M., Gulsrud, T. O., Kjosavik, S. R., and Oppedal, K., “Improving prostate whole gland segmentation in t2-weighted mri with synthetically generated data,” in [2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)], 1915–1919, IEEE (2021).

[20] Cuocolo, R., Stanzione, A., Castaldo, A., De Lucia, D. R., and Imbriaco, M., “Quality control and whole- gland, zonal and lesion annotations for the prostatex challenge public dataset,” European Journal of Radiology 138, 109647 (2021).

[21] Ronneberger, O., Fischer, P., and Brox, T., “U-net: Convolutional networks for biomedical image seg- mentation,” in [International Conference on Medical image computing and computer-assisted intervention], 234–241, Springer (2015).

[22] Cao, R., Bajgiran, A. M., Mirak, S. A., Shakeri, S., Zhong, X., Enzmann, D., Raman, S., and Sung, K., “Joint prostate cancer detection and gleason score prediction in mp-mri via focalnet,” IEEE transactions on medical imaging 38(11), 2496–2506 (2019).

[23] Sanyal, J., Banerjee, I., Hahn, L., and Rubin, D., “An automated two-step pipeline for aggressive prostate lesion detection from multi-parametric mr

sequence," AMIA Summits on Translational Science Proceedings 2020, 552 (2020).

Contribution D

Learning to triage by learning to reconstruct: A generative self-supervised learning approach for prostate cancer based on axial T2w MRI

Alvaro Fernandez-Quilez[a, b, e], Trygve Eftestøl[c], Svein Reidar Kjosavik[d], and Ketil Oppedal[b, e]

[a]Department of Quality and Health Technology, University of Stavanger, Stavanger, Norway.
[b]Stavanger Medical Imaging Laboratory (SMIL), Department of Radiology, Stavanger
University Hospital, Stavanger, Norway.
[c]Department of Electrical Engineering and Computer Science, University of Stavanger, Stavanger, Norway.
[d]General Practice and Care Coordination Research Group, Stavanger University Hospital, Stavanger, Norway.
[e]Centre for Age-Related Medicine, Stavanger University Hospital, Stavanger, Norway.

ABSTRACT

Prostate cancer (PCa) is the second most commonly diagnosed cancer worldwide among men. In spite of it, its current diagnostic pathway is substantially hampered by over-diagnosis of indolent lesions and under-detection of aggressive ones. Imaging techniques like magnetic resonance imaging (MRI) have proven to add additional value to the current diagnostic practices, but they rely on specialized training and can be time-intensive. Deep learning (DL) has arisen as an alternative to automatize tasks such as MRI analysis. Nevertheless, its success relies on large amounts of annotated data which are rarely available in the medical domain. Existing work tackling data scarcity commonly relies on ImageNet pre-training, which is sub-optimal due to the existing gap between the training and the task domain. We propose a generative self-supervised learning (SSL) approach to alleviate such issues. We show that by making use of an auto-encoder architecture and by applying different patch-level transformations such as pixel intensity or occlusion transformations to T2w MRI slices and then trying to recover the original T2w slice we are able to learn robust medical visual representations that are domain-specific. Furthermore, we show the usefulness of our approach by making use of the representations as an initialization method for PCa lesion classification downstream task. Following, we show how our method outperforms ImageNet initialization and how the performance gap increases as the amount of the available labelled data decreases. Furthermore, we provide a detailed sensitivity analysis of the different pixel manipulation transformations and their effect on the downstream task performance.

Keywords: Self-supervised, Prostate cancer, MRI, Classification

1. INTRODUCTION

Prostate Cancer (PCa) is the second most commonly diagnosed cancer,[1] with an estimated incidence of 1.3 million new cases among men worldwide in 2018.[2, 3] The current diagnostic pathway of PCa relies on prostate- specific antigen (PSA) levels in serum. Nevertheless, PSA testing comes at the cost of substantial over-diagnosis of indolent PCa

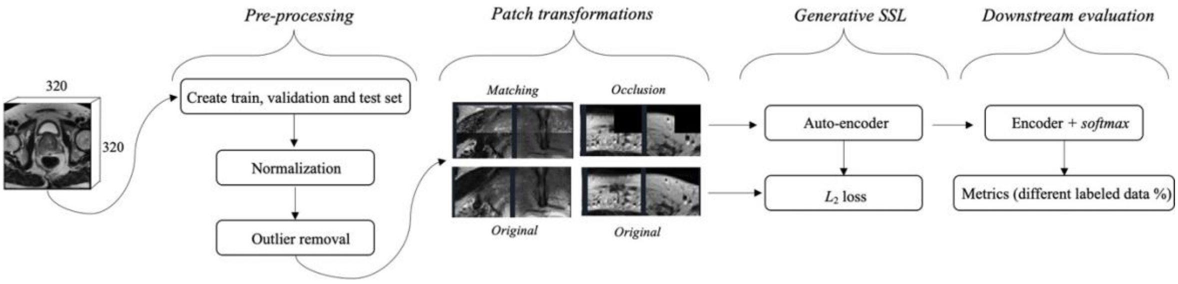


Figure 1: Proposed approach including pre-processing, the self-supervised approach, and evaluation methodology for prostate cancer lesion classification.

lesions and under-detection of aggressive ones, leading to unnecessary biopsies and treatment of indolent PCa lesions.[4, 5]

Magnetic resonance imaging (MRI) has arisen as a suitable option to be used in the current PCa diagnostic pathway and has been proposed as an alternative to current PCa diagnostic pathway approaches or as a support tool for them.[6, 7] However, analysis of MRI requires specialized training and, in its absence, it suffers from inter-reader variability and sub-optimal interpretation.[8, 9] Deep learning (DL) techniques have shown potential in clinical applications.[10] However, traditional DL-based applications rely on large amounts of annotated data which are rarely available in the medical domain.

With the objective to tackle the lack of annotated data, most works use a transfer learning approach from ImageNet weights.[11] Such an approach has been shown to be sub-optimal, due to the existing gap between ImageNet domain and the targeted medical domain.[12] Self-supervised learning (SSL) is a subset of DL techniques aiming to exploit unlabelled data in order to obtain a more efficient initialization method. Existing works have shown promising results when applying generative-based SSL approaches in chest CT and X-ray.[13, 14] In particular, generative SSL aims to offer an autodidactic framework that does not require labelled data as well as robust learning by making use of different self-supervised tasks.

In light the results obtained by generative SSL approaches in other domains and with other types of data, we propose a generative SSL approach and test its robustness and performance in the presence of small amounts of data for PCa lesion classification, the downstream task. Furthermore, we propose different prostate MRI transformations aimed to exploit the downstream task: pixel intensity manipulation. In addition, we test a variety of other transformations and compare them to our proposed one.

2. METHODS

2.1 Task definition

We start by giving a formal description of our SSL generative methodology. We assume an input X_u = $(x_{u_1}, x_{u_2}, \ldots, x_{u_N})$, where x_{u_i} , i = 1...N is the ith unlabelled T2w prostate MRI slice with a fixed resolution. Our goal is to learn a parametrized encoder function fu that maps the input to a low-dimensional embedding zu and a parametrized decoder function g_u that recovers the original input from the low-dimensional embedding z_u. We are interested in transferring the encoder function f_u into downstream tasks. In this particular case, a binary classification of PCa lesions. For evaluation, we assume a labelled input {x_l, y_l} where y_l are the associated labels to the T2w prostate MRI input slice xl. Figure 2 depicts our proposed SSL generative approach.

2.1.1 Context encoding from T2w axial MRI

Our methodology is based on.[13] The technical approach to the project is depicted in Figure 1. We project every T2w prostate MRI slice into a low dimensional space through the encoder function fu, which results into an embedding $z_u \in \mathbb{R}^d$. For each input sample we obtain a distorted version of it $\hat{x}_u$ from X_u with a sampled transformation function $t_u \sim \Gamma$, where Γ is a family of image transformations functions composed by the following transformations: Patch histogram matching, patch rotation, patch occlusion and patch translation. The transformations are applied at the patch level, that is, to small blocks $x_u \in \mathbb{R}^{64\text{x}64}$of the T2w slice under consideration, based on the results and procedures described in[15] (Figure 2). The transformations are applied with a probability p = 0.5 for every image patch under consideration. In addition, if several transformations are applied at the same time, the order in which they are applied is randomized. Once z_u has been obtained, the decoder function g_u tries to recover the original T2w MRI slice, resulting in $\hat{x}_u = g_u$ (z_u).

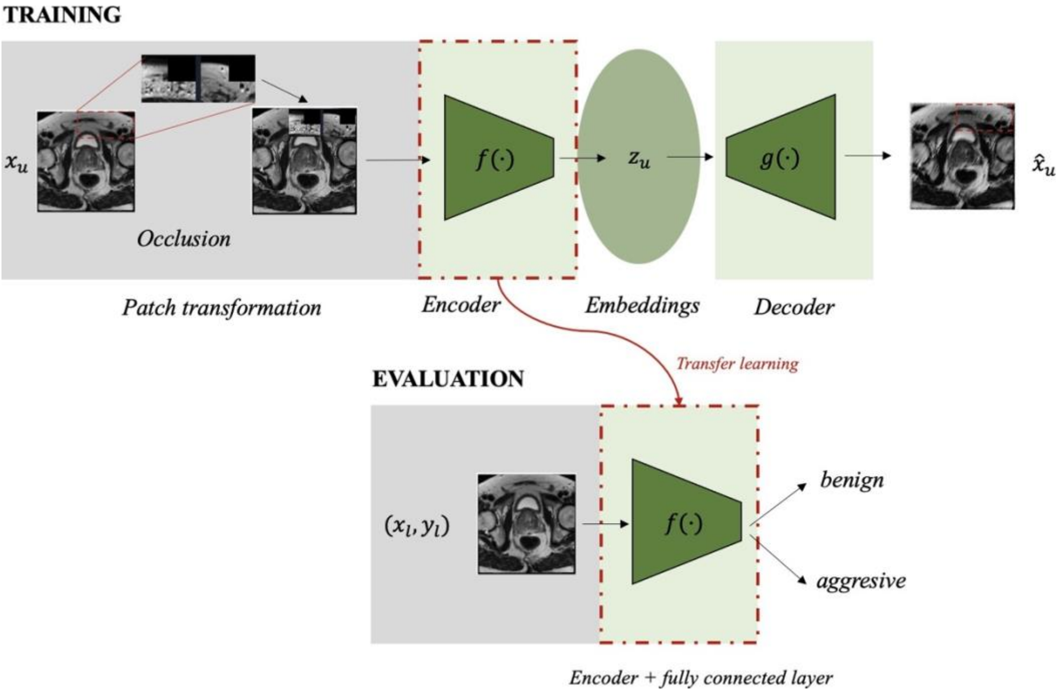


Figure 2: Training and evaluation of our proposed generative self-supervised approach.

At training time, image views are obtained on the fly. The training loss of our methodology is a mean squared error loss:

$$l_u = \sum_{k=1}^{N} \left(x_{u_k} - \hat{x}_{u_k}\right)^2 \qquad (1)$$

where *k* indicates the sample under consideration for the calculation of the loss.

2.2 Implementation and experiments

We introduce the dataset used to carry out the SSL experiments and the downstream task. Following, we introduce the baseline methods used for comparison purposes with our proposed SSL methodology. Finally, we provide an explanation of the different experiments and settings used in our work.

2.2.1 Data pre-processing and splitting

We make use of the ProstateX dataset,[16] which is open source*. The cohort included in the study consisted of 204 patients diagnosed with PCa and 330 lesions. Among those lesions, 76 lesions were clinically significant/aggressive (cS) and 254 were non-clinically significant/benign (ncS). The nature of the study is retrospective and includes different MRI modalities, from which axial T2w is used in this work. Standard pre-processing is applied to the data, including re-sampling to a common coordinate system with the desired dimensions (320x320) and normalization of the MRI intensities to a range of [0, 1]. All the images are provided with results of their biopsies, which is used as the reference standard to determine whether a lesion is cS or ncS (cS if Gleason score is ≥ 7, ncS otherwise). We split the original dataset following a 60%/20%/20% for training, validation and testing, respectively. The splitting is done by patients, such that no data contamination in the form of leakage is present in our splits. respectively. The splitting is done by patients, such that no data contamination in the form of leakage is present in our splits.

2.2.2 Evaluation of the downstream task

We evaluate our pre-trained encoder in one specific downstream task: 2D PCa lesion classification. That is, we classify lesions present in T2w axial slices. We make use of a VGG16 architecture as the encoder of our proposed approach, based on previous results.[15]

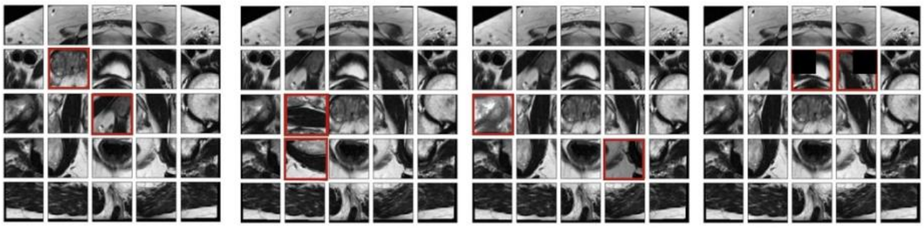


Figure 3: Image transformations from left to right: Patch translation, rotation, histogram matching and occlusion.

We use a linear evaluation protocol to evaluate our results, which consists in keeping the weights of the encoder frozen while training a randomly initialized linear head added on top of the VGG16 encoder. The linear evaluation setting serves the purpose of being a proxy to evaluate the re-usability and quality of the learnt features by the SSL approach. We follow other works on representation learning to choose such an evaluation protocol.[14]

The linear evaluation protocol is carried out with different fractions of labelled data, such that we test the robustness of our approach in a limited data regime. Specifically, we use 1%, 10%, 25%, 50% and 100% of the available T2w axial slice test labels in the evaluation of our approach. Two types of baselines in terms of initialization method are trained following the % reduction of labels approach: ImageNet pre-training and random initialization. In the first case, we initialize the network with weights obtained from training on the standard ImageNet ILSVRC-2012 task.[17] In the second case, we train from scratch a randomly initialized network. The models and training are configured in the following way: batch size of 64, learning rate of 1e−5 and 500 epochs with early stopping if the validation loss does not improve over 40 epochs. All the training and evaluation is carried out on an NVIDIA v100 GPU.

2.2.3 On image views and patch transformations

In order to capture different T2w axial MRI characteristics during the pre-training phase of our SSL approach, we leverage different transformations of the data at the patch level (Figure 2). Specifically, the transformations aim to capture robust image representations by restoring the original image from the family of transformations Γ. In particular, we propose patch translation, patch histogram matching, patch pixel occlusion and patch rotation (Figure 3). All the transformations are applied to $N = 2$ randomly selected patches and with a probability of occurrence of $p = 0.5$.

- **Translation**: We shuffle $N = 2$ randomly selected patches, resulting in a change of position of the selected patches, which encourages the model to learn textures of objects.
- **Rotation**: We rotate $N = 2$ randomly selected patches between 0 and 90 degrees. We expect to encourage the model to learn the spatial layout of objects in medical images by restoring the rotated patches.

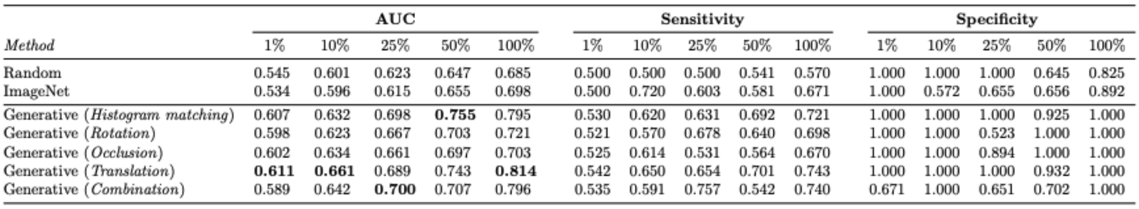
• **Histogram matching**: We apply histogram matching to N = 2 randomly selected patches, resulting in a change in the distribution of the pixel intensities of the selected patches. With this transformation we expect the model to learn appearance of anatomic structures that appear in the image

• **Occlusion**: We occlude a randomly selected 16x16 region of N = 2 64x64 randomly selected patches. The choice of occluding a sub-region of the patch is to be able to capture neighbouring details such that the auto-encoder architecture is able to learn the context of the patch under consideration. With the occlusion, we expect the model to learn local context of the image.

3. RESULTS

We start by evaluating our results in terms of the quality of the representations obtained by our SSL approach. We make use of a linear evaluation protocol (Section 2.2.2) and the area under the curve (AUC), specificity and sensitivity for different fractions of labels. We compare our initialization method against ImageNet and random initialization. Following, we provide a discussion on current lesion classification methods and our SSL approach.

Table 1: Comparison between different initialization methods and different percentage of labelled data in a linear classification setting.



	AUC					Sensitivity					Specificity				
Method	1%	10%	25%	50%	100%	1%	10%	25%	50%	100%	1%	10%	25%	50%	100%
Random	0.545	0.601	0.623	0.647	0.685	0.500	0.500	0.500	0.541	0.570	1.000	1.000	1.000	0.645	0.825
ImageNet	0.534	0.596	0.615	0.655	0.698	0.500	0.720	0.603	0.581	0.671	1.000	0.572	0.655	0.656	0.892
Generative (*Histogram matching*)	0.607	0.632	0.698	**0.755**	0.795	0.530	0.620	0.631	0.692	0.721	1.000	1.000	1.000	0.925	1.000
Generative (*Rotation*)	0.598	0.623	0.667	0.703	0.721	0.521	0.570	0.678	0.640	0.698	1.000	1.000	0.523	1.000	1.000
Generative (*Occlusion*)	0.602	0.634	0.661	0.697	0.703	0.525	0.614	0.531	0.564	0.670	1.000	1.000	0.894	1.000	1.000
Generative (*Translation*)	**0.611**	**0.661**	0.689	0.743	**0.814**	0.542	0.650	0.654	0.701	0.743	1.000	1.000	1.000	0.932	1.000
Generative (*Combination*)	0.589	0.642	**0.700**	0.707	0.796	0.535	0.591	0.757	0.542	0.740	0.671	1.000	0.651	0.702	1.000

As table 1 depicts, our SSL approach yields to better results in terms of AUC than ImageNet and random initialization. In particular, in reduced data regimes (lower than 50% of the original number of labelled T2w axial slices), the SSL method AUC is larger by a considerable margin when compared with the other initialization methods. Furthermore, we can also observe a better performance overall in the presence of the original amount of testing data (fraction of 100%), showing the quality of the representations learnt by our SSL methodology. We can also observe that the different transformations yield to different result in terms of the different fractions of labelled data used to evaluate the methodology. As table 1 shows, translation and histogram matching are the best performing ones overall. Since the downstream task is lesion classification, one could argue that the relative position of the patch (translation) is an important factor when determining the severity of the lesion, as most severe lesions are located in specific zones of the prostate.[15] The results obtained by histogram matching depict the relevance of the pixel intensity distribution and the anatomic and appearance structures of the prostate.

When focusing on the 100% fraction of labelled axial T2w slices, we observe the best performance is an AUC of 0.814 (Table 1). When compared to other approaches tackling the same problem, our SSL approach reaches a similar level of AUC as of other methods such as *FocalNet*[10] (0.81 AUC), 2-stage 2D U-net[18] (0.86 AUC), a 2.5 HED architecture that reaches an AUC of 0.8019 or a multi-vendor architecture that reaches an AUC of 0.93,[20] among others. In spite of not reaching the same AUC as the best performing approaches, our method offers the advantage of requiring less data and being applied to a simple architecture (VGG16), whereas methods that evaluate on the same downstream task use bigger datasets or more complex architectures to reach similar or higher AUC scores. Furthermore, our method can be used as an add-on method in more complex architectures, potentially improving and obtaining higher AUC scores than the ones presented in this work. Finally, the leading methods in the AUC score board of the ProstateX challenge[†] reach AUC scores of 0.95, which are considerably larger to the one presented in this study. Nevertheless, based on the available details of such methodologies, the proposed architectures have higher complexity than the ones used in this study and our method could be used as an add-on to such methods, potentially improving the final results.

4. CONCLUSION AND REMARKS

In this work, we presented an SSL generative approach for PCa lesion classification. Our approach has consistently outperformed common

[†] https://prostatex.grand-challenge.org/evaluation/challenge/leaderboard/

initialization methods used in the medical image domain such as ImageNet or random one, achieving good results even in settings with a highly reduced amount of labelled data. In particular, we observed our proposed methodology is able to obtain high quality and transferable representations by means of a linear evaluation protocol. Our results suggest the possibility of a broad application of SSL methods for prostate MRI applications in the presence of a limited amount of labelled data and the re-usability of our methodology as an add-on for models with higher complexity than the ones presented in this work. Patient-level classification and more experiments in terms of the proposed transformations will be carried out in the future, along with multi- parametric MRI data, with potential for further improvement and understanding of the effect of the different pre-text tasks.

5. ABOUT THE WORK

This work has not been submitted nor is planned to be submitted anywhere else.

REFERENCES

[1] Siegel, R. L., Miller, K. D., and Jemal, A.,"Cancer statistics, 2020," CA: A Cancer Journal for Clini-cians 70(1), 7–30 (2020).

[2] Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., and Jemal, A., "Global cancer statistics 2018," CA: A Cancer Journal for Clinicians 68(6), 394–424 (2018).

[3] Fernandez-Quilez, A., German-Borda, M., Leonardo, G., Castellanos, N., Soennesyn, H., Oppedal, K., and Reidar-Kjosavik, S., "Prostate cancer screening and socioeconomic disparities in mexican older adults," salud pu´blica de m´exico 62(2, Mar-Abr), 121–122 (2020).

[4] Barry, M. J., "Screening for prostate cancer — the controversy that refuses to die," New England Journal of Medicine 360, 1351–1354 (Mar 2009).

[5] Hodge, K. K., McNeal, J. E., and Stamey, T. A., "Ultrasound guided transrectal core biopsies of the palpably abnormal prostate," Journal of Urology 142, 66–70 (Jul 1989).

[6] Dirix, P., Bruwaene, S. V., Vandeursen, H., and Deckers, F., "Magnetic resonance imaging sequences for prostate cancer triage: two is a couple, three is a crowd?" Translational Andrology and Urology 8, S476–S479 (Dec 2019).

[7] Lomas, D. J. and Ahmed, H. U., “All change in the prostate cancer diagnostic pathway,” Nature Reviews Clinical Oncology 17, 372–381 (Feb 2020).

[8] Turkbey, B., Rosenkrantz, A. B., Haider, M. A., Padhani, A. R., Villeirs, G., Macura, K. J., Tempany,

C. M., Choyke, P. L., Cornud, F., Margolis, D. J., and et al., “Prostate imaging reporting and data system version 2.1: 2019 update of prostate imaging reporting and data system version 2,” European Urology 76, 340–351 (Sep 2019).

[9] Gaziev, G., Wadhwa, K., Barrett, T., Koo, B. C., Gallagher, F. A., Serrao, E., Frey, J., Seidenader, J., Carmona, L., Warren, A., et al., “Defining the learning curve for multiparametric magnetic resonance imag- ing (mri) of the prostate using mri-transrectal ultrasonography (trus) fusion-guided transperineal prostate biopsies as a validation tool,” BJU international 117(1), 80–86 (2016).

[10] Cao, R., Bajgiran, A. M., Mirak, S. A., Shakeri, S., Zhong, X., Enzmann, D., Raman, S., and Sung, K., “Joint prostate cancer detection and gleason score prediction in mp-mri via focalnet,” IEEE transactions on medical imaging 38(11), 2496–2506 (2019).

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

[12] Raghu, M., Zhang, C., Kleinberg, J., and Bengio, S., “Transfusion: Understanding transfer learning for medical imaging,” arXiv preprint arXiv:1902.07208 (2019).

[13] Zhou, Z., Sodha, V., Siddiquee, M. M. R., Feng, R., Tajbakhsh, N., Gotway, M. B., and Liang, J., “Models genesis: Generic autodidactic models for 3d medical image analysis,” in [International Conference on Medical Image Computing and Computer-Assisted Intervention], 384–393, Springer (2019).

[14] Haghighi, F., Taher, M. R. H., Zhou, Z., Gotway, M. B., and Liang, J., “Transferable visual words: Ex- ploiting the semantics of anatomical patterns for self-supervised learning,” IEEE transactions on medical imaging (2021).

[15] Fernandez-Quilez, A., Eftestøl, T., Goodwin, M., Kjosavik, S. R., and Oppedal, K., “Self-transfer learning via patches: A prostate cancer triage approach based on bi-parametric mri,” arXiv preprint arXiv:2107.10806 (2021).

[16] Litjens, G., Debats, O., Barentsz, J., Karssemeijer, N., and Huisman, H., “Computer-aided detection of prostate cancer in mri,” IEEE transactions on medical imaging 33(5), 1083–1092 (2014).

[17] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., B

ernstein, M., et al., “Imagenet large scale visual recognition challenge,” International journal of computer vision 115(3), 211–252 (2015).

[18] Sanyal, J., Banerjee, I., Hahn, L., and Rubin, D., “An automated two-step pipeline for aggressive prostate lesion detection from multi-parametric mr sequence,” AMIA Summits on Translational Science Proceed- ings 2020, 552 (2020).

[19] Seetharaman, A., Bhattacharya, I., Chen, L. C., Kunder, C. A., Shao, W., Soerensen, S. J., Wang, J. B., Teslovich, N. C., Fan, R. E., Ghanouni, P., et al., “Automated detection of aggressive and indolent prostate cancer on magnetic resonance imaging,” Medical Physics (2021).

[20] Sumathipala, Y., Lay, N. S., Turkbey, B., Smith, C., Choyke, P. L., and Summers, R. M., “Prostate cancer detection from multi-institution multiparametric mris using deep convolutional neural networks,” Journal of Medical Imaging 5(4), 044507 (2018).

Contribution E

CONTRASTING AXIAL T2W MRI FOR PROSTATE CANCER TRIAGE: A SELF-SUPERVISED LEARNING APPROACH

Alvaro Fernandez-Quilez[1,2]
Trygve Eftestøl[3] Svein Reidar Kjosavik[4] Morten Goodwin[5]
Ketil Oppedal[2,3]

[1]Department of Quality and Health Technology, University of Stavanger, Norway.
[2]Stavanger Medical Imaging Laboratory, Dept. of Radiology, Stavanger University Hospital, Norway.
[3]Department of Electrical Engineering and Computer Science, University of Stavanger, Norway.
[4]General Practice and Care Coordination Research Group, Stavanger University Hospital, Norway.
[5]Centre for Artificial Intelligence Research (CAIR), Department of ICT, University of Agder, Norway

ABSTRACT

Current diagnostic practices for prostate cancer (PCa) suffer from over-diagnosis of indolent lesions and under-detection of aggressive ones. Deep learning (DL) techniques have shown potential in automatizing tasks and helping clinicians. Nevertheless, their success depends on the availability of large amounts of labelled data, which are rarely available in the medical field. Hence, transfer learning using ImageNet has become the de facto approach but it has been shown to be sub- optimal for medical images. Contrastive learning is a form of self-supervised learning (SSL) that leverages unlabelled data to produce pre-trained models and has shown promising results on natural images. However, its application to MRI interpretation has been rather limited. In this work, we pro- pose a contrastive approach (SimCLR) to produce models with better initializations for 2D PCa lesion classification. Our results show that linear and end-to-end fine-tuned models trained on our SSL pre-trained representations outperform ImageNet and random initialization.

Index Terms— Contrastive learning, Self-supervised learning, SimCLR,
MRI, Prostate

1. INTRODUCTION

Prostate Cancer (PCa) is the second most commonly diagnosed cancer [1], with an estimated incidence of 1.3 million new cases among men worldwide in 2018 [2, 3]. Current diagnostic and management methods of PCa rely on prostate-specific antigen (PSA) levels in serum. However, PSA testing comes at the cost of under-detection of aggressive lesions and over-diagnosis of indolent ones, leading to unnecessary biopsies and treatment of the indolent ones [4].

Magnetic resonance imaging (MRI) has arisen as an alternative to the current diagnostic pathway tests thanks to recent advances in image acquisition and technology innovation. Nevertheless, MRI analysis requires expertise and specialized training, and in its absence, it suffers from inter- reader variability and sub-optimal interpretation [5]. Deep learning (DL) techniques have emerged as an alternative able to

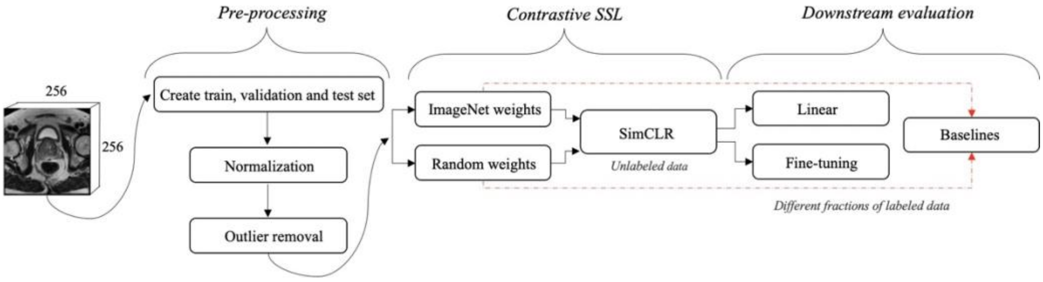


Fig. 1. Proposed approach including pre-processing, self-supervised learning (SimCLR) and evaluation for prostate cancer triage.

automatize and democratize applications. Its potential has already been proven with some successful applications in the medical domain [6]. However, traditional DL-based applications rely on large amounts of annotated data which are rarely available in the medical domain [7].

In order to palliate the lack of annotated data, existing works use a transfer learning approach using ImageNet weights [8]. Such an approach has been shown to be sub- optimal, due to the existing domain gap between ImageNet and the targeted medical domain (out-of-domain initialization) [9]. Self-supervised learning (SSL) is a subset of DL techniques which exploit unlabelled data in order to obtain a more efficient initialization method (in-domain). In particular, SimCLR has been shown to obtain promising results when applied to natural images, even in a limited data regime [10, 11].

SimCLR uses a variety of data augmentations during its training; however, the nature of some of those augmentations such as blurring might perturb characteristics of the image that are disease-specific. In the light of it, we apply a similar method with data augmentations specially tailored to our down-stream application: prostate cancer triage. We show how our method outperforms other initialization strategies such as ImageNet or random initialization. Furthermore, we show how our method works well under small data regimes and evaluate its performance in terms of different parameters such as image resolution and embedding dimensions.

2. METHODS

We start by giving a formal description of our SSL contrastive methodology. Figure 1 presents an overview of the steps followed in the work. In terms of contrastive SSL, we assume an input $X_u = (x_{u_1}, x_{u_2}, \dots, x_{u_N})$, where x_u are unlabelled T2w prostate MRI slices (2D), with a pre-defined resolution. Our goal is to learn a parametrized encoder function fu that maps the input to a low-dimensional embedding z_u. After learning such a function, we are interested in transferring the learned function fu into PCa downstream tasks. In this particular case, a binary classification between indolent and non- indolent lesions (triage) task. For the evaluation of the task, we assume an input (x_l, y_l) where y_l are the associated labels to the T2w prostate MRI input slices x_l.

2.1. Dataset: Prostate T2w axial MRI

We make use of the ProstateX dataset [12], which is open source[1]. The cohort included in the study consisted of 204 patients diagnosed with PCa and 330 lesions. Among those lesions, 76 lesions are clinically significant (cS) and 254 are non-clinically significant (ncS). The nature of the study is retrospective and includes different MRI modalities from which axial T2w is used in this work. All images are provided with biopsy results, which are used to obtain the ground truth for the downstream task. The clinical significance label of the MRI slices is based on the Gleason score, where the lesion is considered cS if the score is 7 or higher and ncS otherwise.

2.1.1. Pre-processing and data splitting

Standard pre-processing is applied to the data, including re- sampling to a common coordinate system by linear interpolation with a resolution of 0.5x0.5x0.5mm, normalization of the MRI intensities to a range of [0, 1] and outlier removal by forcing the intensity values of the image between the 1st and the 99th percentiles. In order to evaluate the downstream method, we split the original labelled dataset by patients following a 60%/20%/20% for training, validation and testing, respectively.

[1] https://wiki.cancerimagingarchive.net/display/
Public/SPIE-AAPM-NCI+PROSTATEx+Challenges

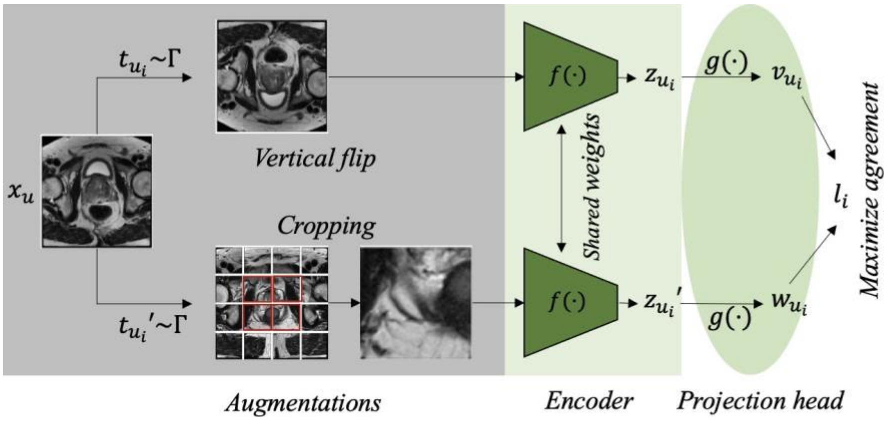


Fig. 2. Contrastive learning using different augmentations of the same slice. Example with cropping and vertical flipping.

2.2. SimCLR

We adapt the SimCLR-pretraining procedure to prostate T2w axial MRI. SimCLR maximizes the agreement between different augmented views of the same data example using a contrastive loss (Equation 1) in the embedding space [10]. Figure 2 shows a schematic of how data augmentations are used to learn embeddings in an unsupervised fashion. Given a mini-batch of images, each image xu of the mini-batch is aug mented twice using a sampled augmentation $t_u \sim \Gamma$, where Γ is a family of image augmentations. Following, both views of x_u are projected into a low-dimensional space through the encoder function f_u, which results into an embedding $z_u \in \mathbb{R}^d$for each of the two image views. Once z_u has been obtained, a non-linear projection g_u is applied through a projection head (multi-layer perceptron head)[2], $v_u = g_u(z_u)$, $v_u \in \mathbb{R}^D$, for both image views, which are used for the contrastive loss (Equation 1).

With a mini-batch of N encoded examples, the contrastive loss for the ith a pair of embeddings can be defined as follows:

$$MSE = \sum_{k=1}^{N} \left(x_{h_k} - \hat{x}_{h_k}\right)^2 \qquad (1)$$

where we use (v_{u_i}, w_{u_i}) to denote the *i*th embedding pair and <· , ·> is used to denote the similarity measure used between the embeddings. In particular, a dot product similarity is assumed.

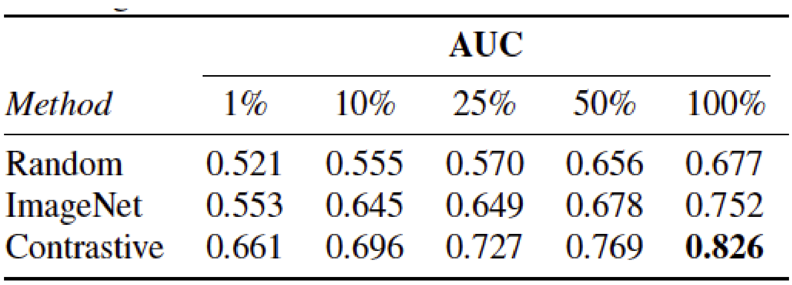
2.2.1. Pre-training for 2D Prostate Cancer Triage

SimCLR pre-training is performed on the entire ProstateX unlabelled dataset. We apply SimCLR on an encoder initialized with both ImageNet weights and random ones, with no extra cost due to the large availability of ImageNet-pretrained models. Our choice to use SimCLR is driven by its previous success in both natural images [10] and medical images [13]. We choose a family of augmentations Γ which we deem suitable to generate views without losing the interpretability of the diagnosis in the MRI slice under consideration. In particular, we make use of random rotation (50 degrees), translation (range of 0.32 pixels), vertical flipping and cropping, based on previous works [14]. The proposed cropping mechanism targets areas of the prostate in which lesions are commonly located by centre cropping the image.

The overall training steps are depicted in Figure 2. We kept the hyperparameters used in the original work, with the exception of the batch size which is set to 512 [13] and a swish activation function in the projection head g(·)[3]. Additionally, we experiment with different embedding dimensions D = {128, 256, 512}. Furthermore, we test different image resolutions and evaluate their effect on the final performance of the linear evaluation. Checkpoints from the SimCLR pre- training of the top performing epochs are obtained and used for the downstream task evaluation. A VGG16 architecture is used as an encoder, based on previous works [14].

2.2.2. Linear and Fine-Tuning Evaluation

We evaluate SimCLR pre-training following previous work on unsupervised visual representation learning [15]. The Sim- CLR pre-trained encoder is evaluated under two different set- tings for the downstream task: linear classification and fine- tuning. In the first case, the pre-trained weights obtained from SimCLR pre-training are frozen and a randomly initialized linear head is trained for the task under consideration. Linear evaluation is intended to give an idea about the quality of the learned features and their re-usability [16]. On the other hand, in the fine-tuning scenario the whole encoder is unfrozen and the entire model is fine-tuned end-to-end.



	AUC				
Method	1%	10%	25%	50%	100%
Random	0.521	0.555	0.570	0.656	0.677
ImageNet	0.553	0.645	0.649	0.678	0.752
Contrastive	0.661	0.696	0.727	0.769	**0.826**

Table 1: Comparison between different initialization methods and different percentage of labelled data, in a linear classification setting.

Both evaluation protocols are carried out with different fractions of labelled training data (x_l, y_l) in order to test the robustness of the approach when dealing with a limited amount of data and as a proxy for the real world, where only a small amount of labelled data is available. In particular, 1%, 10%, 25%, 50% and 100% fraction of the labels are used in the evaluation protocols. Two types of baselines are trained fol lowing the reduced fraction protocol: ImageNet pre-training and random initialization. In the first case, we use an ImageNet pre-trained network which was not subject to SimCLR pre-training. In the second case, we train from scratch a randomly initialized network.

We follow the same configuration for all the fine-tuned models: batch size of 64, learning rate of 1e−5 and 500 epochs with early stopping, which halted the training if the validation loss did not improve over 40 epochs. All the training and fine-tuning is carried out on an NVIDIA v100 GPU.

2.2.3. On Image Size and Embedding Dimension

SimCLR requires large batches to obtain good results [10]. Given that medical image sizes are commonly larger than natural ones, we explore the effect of reducing the image size on the SimCLR pre-training phase and on the down- stream one.

[2,3] More details of the implementation on https://github.com/ alvfq/simclr-2D-prostate-triage

The objective of such a test is to verify whether reducing the size of the images used to pre-train SimCLR can be beneficial in terms of final performance in the down- stream task and computational power requirements of Sim- CLR. Moreover, since the optimal embedding dimension D is related to the shape of the inputs, we also experiment with D = {128, 256, 512} for different image sizes.

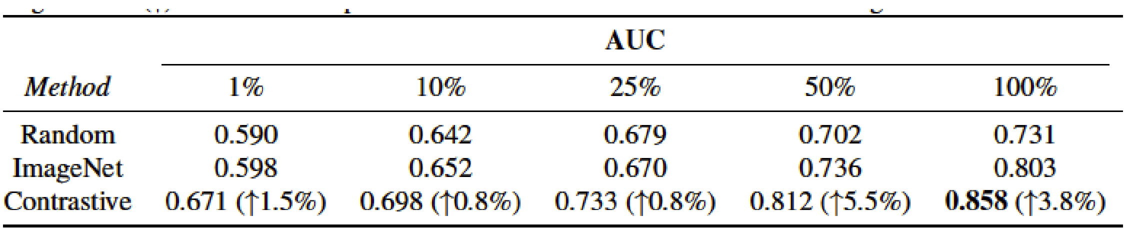
3. RESULTS

We start by evaluating our results based on the quality of the representations obtained by SimCLR and comparing them against ImageNet pre-training and a random initialization of the weights. In order to evaluate the representations, we used a linear evaluation [16], where the base model is frozen (VGG16) and a linear classifier is trained on top of it. Following, the test performance is used as a proxy to evaluate

the quality of the representations (Section 2.2.2). The results are quantified in terms of area under the curve (AUC) [17], at different label fractions (Table 1). Following, we investigate whether SimCLR pre-training is able to obtain a higher performance when fine-tuned end-to-end. Similarly to the first evaluation, we obtain the results for different label fractions (Table 2). Based on Table 1 we can observe how when trained on small label fractions, the SimCLR pre-training approach shows a significantly larger AUC than their counterparts; random initialization and ImageNet ones. In particular, we can see how when trained in a limited data regime (1% label fraction) the SimCLR-based model achieves a 0.671 AUC whereas a random initialization obtains a 0.590 and the ImageNet one a 0.598. Moreover, we can see how the drop in the AUC is not as drastic when moving from a 10% label fraction to a 1% as compared to ImageNet and random initialization, supporting the hypothesis that SimCLR pre-training benefits the quality of the obtained representations in small data regimes.

Regarding end-to-end fine-tuning, as shown in Table 2 we found that SimCLR pre-training and fine-tuned end-to-end outperforms by a large margin ImageNet pre-training when working under really small label fractions or large ones (1% and 100%). In particular, for a label fraction of 1% the AUC of SimCLR end-to-end is 0.671 whilst ImageNet one is 0.589 and for a label fraction of 100% SimCLR obtains an AUC of 0.858 whilst ImageNet obtains an AUC of 0.803. The results show

Table 2. Comparison between different initialization methods and different percentage of labelled data, in a fine-tuning classification setting. Arrows (↑) indicates the improvement in % over the linear classification setting in Table 1.



	AUC				
Method	1%	10%	25%	50%	100%
Random	0.590	0.642	0.679	0.702	0.731
ImageNet	0.598	0.652	0.670	0.736	0.803
Contrastive	0.671 (↑1.5%)	0.698 (↑0.8%)	0.733 (↑0.8%)	0.812 (↑5.5%)	**0.858** (↑3.8%)

that SimCLR pre-training with an end-to-end fine- tuning yields a boost in performance for end-to-end training, even when the fraction of labels is small, which is consistent with previous findings [10, 11].

Finally, as explained in Section 2.2.3 we investigate the relationship between different image sizes and embedding dimension. Furthermore, we explore different initialization methods for the SimCLR framework. As depicted in Figure 3, a larger image size yields a better performance when paired with a larger embedding size for a VGG16 encoder. In particular, we found that a good compromise between image size, embedding dimension and training speed is accomplished at an image resolution of 256x256 with D = 256, which obtains the best AUC (Figure 3). Furthermore, our experiments revealed that ImageNet initialization obtained better results in terms of AUC and training speed when compared to random initialization in the SimCLR framework, being the differences of more than 30% difference in the final AUC in the linear and fine-tuning evaluation (Figure 1, contrastive SSL).

4. CONCLUSIONS

We found that SimCLR is able to obtain high quality representations for PCa triage based on T2w axial MRI. In particular, the initialization provided by SimCLR outperforms ImageNet and random ones in small data regimes, showing their quality even in situations where data is scarce. To the best of our knowledge, this work is the first to highlight the benefit of SimCLR across label fractions for PCa triage. Limitations of the work include the retrospective nature of the data and the lack of more experimentation with other SSL methods.

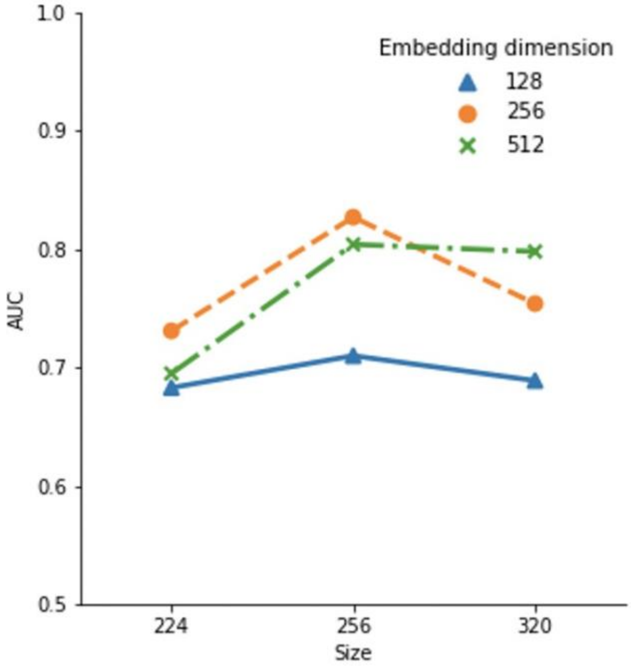


Fig. 3. Results with different image sizes and embedding dimensions.

The results suggest the possibility for a broad application of SSL approaches beyond natural images.

5. COMPLIANCE WITH ETHICAL STANDARDS

This research study was conducted retrospectively using open access data. Ethical approval was not required.

6. ACKNOWLEDGMENTS

This work has been funded by the University of Stavanger. The authors have no relevant financial or non-financial interests to disclose.

7. REFERENCES

[1] Rebecca L. Siegel, Kimberly D. Miller, and Ahmedin Jemal, “Cancer statistics, 2020,” CA: A Cancer Journal for Clinicians, vol. 70, no. 1, pp. 7–30, 2020.

[2] Freddie Bray, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L. Siegel, Lindsey A. Torre, and Ahmedin Je- mal, “Global cancer statistics 2018,” CA: A Can- cer Journal for Clinicians, vol. 68, no. 6, pp. 394–424, 2018.

[3] Alvaro Fernandez-Quilez, Miguel German-Borda, Gabriel Leonardo, Nicola´s Castellanos, Hogne Soen- nesyn, Ketil Oppedal, and Svein Reidar-Kjosavik, “Prostate cancer screening and socioeconomic disparities in mexican older adults,” salud publica de me´xico, vol. 62, no. 2, Mar-Abr, pp. 121–122, 2020.

[4] Kathryn K. Hodge, John E. McNeal, and Thomas A. Stamey, “Ultrasound guided transrectal core biopsies of the palpably abnormal prostate,” Journal of Urology, vol. 142, no. 1, pp. 66–70, Jul 1989.

[5] Andrew B Rosenkrantz, Luke A Ginocchio, Daniel Cornfeld, Adam T Froemming, Rajan T Gupta, Baris Turkbey, Antonio C Westphalen, James S Babb, and Daniel J Margolis, “Interobserver reproducibility of the pi-rads version 2 lexicon: a multicenter study of six ex- perienced prostate radiologists,” Radiology, vol. 280, no. 3, pp. 793–804, 2016.

[6] Ruiming Cao, Amirhossein Mohammadian Bajgiran, Sohrab Afshari Mirak, Sepideh Shakeri, Xinran Zhong, Dieter Enzmann, Steven Raman, and Kyunghyun Sung, “Joint prostate cancer detection and gleason score prediction in mp-mri via focalnet,” IEEE transactions on medical imaging, vol. 38, no. 11, pp. 2496–2506, 2019.

[7] Alvaro Fernandez-Quilez, Steinar Valle Larsen, Morten Goodwin, Thor Ole Gulsrud, Svein Reidar Kjosavik, and Ketil Oppedal, “Improving prostate whole gland segmentation in t2-weighted mri with synthetically gen-

erated data," in 2021 IEEE 18th International Sympo- sium on Biomedical Imaging (ISBI), 2021, pp. 1915– 1919.

[8] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun, "Dermatologist-level classification of skin can- cer with deep neural networks," nature, vol. 542, no. 7639, pp. 115–118, 2017.

[9] Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio, "Transfusion: Understanding trans- fer learning for medical imaging," arXiv preprint arXiv:1902.07208, 2019.

[10] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton, "A simple framework for contrastive learning of visual representations," in International con- ference on machine learning. PMLR, 2020, pp. 1597– 1607.

[11] Ting Chen, Simon Kornblith, Kevin Swersky, Mo- hammad Norouzi, and Geoffrey Hinton, "Big self- supervised models are strong semi-supervised learners," arXiv preprint arXiv:2006.10029, 2020.

[12] Geert Litjens, Oscar Debats, Jelle Barentsz, Nico Karssemeijer, and Henkjan Huisman, "Computer-aided detection of prostate cancer in mri," IEEE transac- tions on medical imaging, vol. 33, no. 5, pp. 1083–1092, 2014.

[13] Shekoofeh Azizi, Basil Mustafa, Fiona Ryan, Zachary Beaver, Jan Freyberg, Jonathan Deaton, Aaron Loh, Alan Karthikesalingam, Simon Kornblith, Ting Chen, et al., "Big self-supervised models advance medical image classification," arXiv preprint arXiv:2101.05224, 2021.

[14] Minh Hung Le, Jingyu Chen, Liang Wang, Zhiwei Wang, Wenyu Liu, Kwang-Ting Tim Cheng, and Xin Yang, "Automated diagnosis of prostate cancer in multi- parametric mri based on multimodal convolutional neu- ral networks," Physics in Medicine & Biology, vol. 62, no. 16, pp. 6497, 2017.

[15] Hari Sowrirajan, Jingbo Yang, Andrew Y Ng, and Pranav Rajpurkar, "Moco-cxr: Moco pretraining im- proves representation and transferability of chest x-ray models," arXiv preprint arXiv:2010.05352, 2020.

[16] Simon Kornblith, Jonathon Shlens, and Quoc V Le, "Do better imagenet models transfer better?," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 2661–2671.

[17] Christopher M Florkowski, "Sensitivity, specificity, receiver-operating characteristic (roc) curves and like- lihood ratios: communicating the

performance of diagnostic tests," The Clinical Biochemist Reviews, vol. 29, no. Suppl 1, pp. S83, 2008.

Contribution F

MULTI-PLANAR T2W MRI FOR AN IMPROVED PROSTATE CANCER LESION CLASSIFICATION

Alvaro Fernandez-Quilez[1,2]
Trygve Eftestøl[3] Svein Reidar Kjosavik[4] Morten Goodwin[5]
Ketil Oppedal[2,3]

[1]Department of Quality and Health Technology, University of Stavanger, Norway.
[2]Stavanger Medical Imaging Laboratory, Dept. of Radiology, Stavanger University Hospital, Norway.
[3]Department of Electrical Engineering and Computer Science, University of Stavanger, Norway.
[4]General Practice and Care Coordination Research Group, Stavanger University Hospital, Norway.
[5]Centre for Artificial Intelligence Research (CAIR), Department of ICT, University of Agder, Norway

ABSTRACT

Prostate cancer (PCa) is the fifth leading cause of death worldwide. In spite of the urgency for a timely and accurate diag- nostic, the current PCa diagnostic pathway suffers from over- diagnosis of indolent lesions and under-diagnosis of highly invasive ones. The advent of deep learning (DL) techniques has enabled automatic and accurate computer-assisted systems that rival human performance. However, current approaches for PCa diagnostic are heavily reliant on T2w axial MRI, which suffer from low out-of-plane resolution. Sagittal and coronal MRI scans are usually acquired by default along with the axial one but are generally ignored by DL classification algorithms. We propose a multi-stream approach to accommodate sagittal, coronal and axial planes and improve the performance of PCa lesion classification. We evaluate our method on a publicly available dataset and demonstrate that it provides better results when compared with a single-plane approach over a range of different DL architectures.

Index Terms— MRI, lesion classification, Multi-planar, Multi-stream, Prostate

1. INTRODUCTION

Prostate Cancer (PCa) is one of the most prevalent cancers in men and the fifth leading cause of death worldwide [1, 2]. Traditionally, the diagnostic of PCa was based on digital rectum examination (DRE) but ever since its approval as a screening test, prostate-specific antigen (PSA) levels in serum became the main tool for PCa diagnostic and man- agement [3]. However, its use remains as a controversial topic due to unclear benefits of it as a screening technique [4], over-diagnostic of indolent (non-clinically significant, ncS) lesions and under-detection of potentially lethal (clinically significant, cS) PCa lesions [5]. Magnetic resonance imaging (MRI) is a non-invasive technique that has shown potential to compensate for current main diagnostic tests shortcomings [6, 7]. In particular, T2-weighted MRI (T2w) shows anatomic-morphological features of the prostate and morphological-pathological

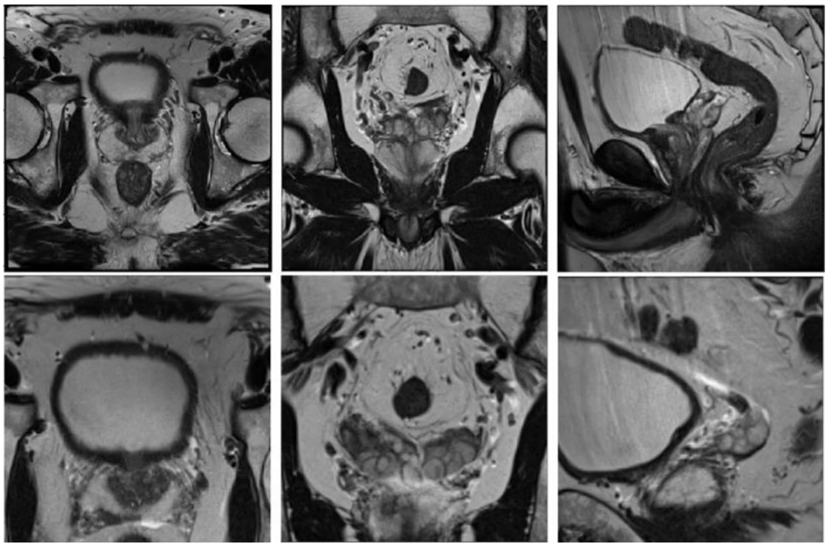


Fig. 1: Original axial, coronal and sagittal slice (top row) and cropped images after pre-processing (bottom row).

structures and it is usually acquired in three planes: sagittal, coronal and axial [6]. In spite of its potential, MRI analysis requires specialized training and in the event of its absence, it suffers from sub-optimal interpretation and inter-reader variability [8].

Deep learning (DL) models are becoming increasingly popular for PCa applications thanks to their ability to automatize time-intensive applications while requiring minimal human intervention [9]. In particular, models have been proposed for prostate segmentation [10], detection and lesion classification [11, 12, 13]. Nevertheless, the proposed methods purely rely on axial T2w MRI scans of the prostate. However, sagittal and coronal MRI scans are commonly avail- able, as multiple directions are acquired by default for better clinical interpretation and commonly taken into account by radiologists [14].

In this work, we propose a multi-stream approach to ex- ploit axial, coronal and sagittal prostate MRI scans such that better features are obtained for 2D PCa lesion classification. In addition to it, we propose to further improve the informa- tion sharing between the orthogonal scans by adding connections

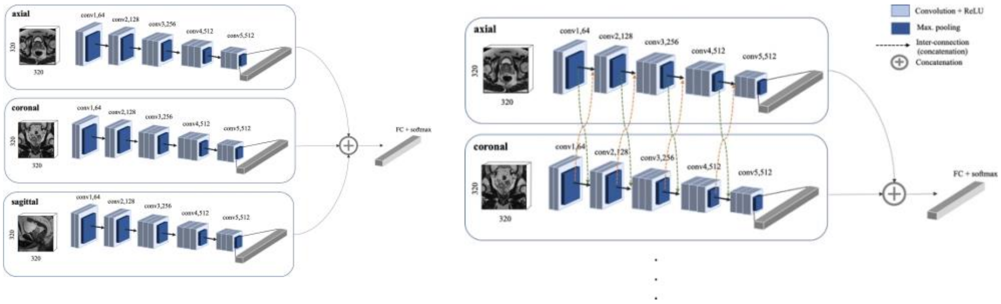


Fig. 2: Exemplification of our proposed multi-stream approaches based on VGG16.

between the different input streams [15]. We demonstrate the effectiveness of our approaches over a range of different DL architectures and that by incorporating information from the different scan directions we are able to substantially improve the performance of the different architectures for 2D PCa lesion classification when compared to training exclu- sively on T2w axial scans or a multi-channel approach. In particular, our contributions are:

1. We propose a simple 2D multi-stream approach (siMS) to process orthogonal T2w scan directions simultaneously.
2. We propose a 2D inter-connected multi-stream approach (icMS) to improve feature sharing and re-use between the different orthogonal scan direction streams.
3. We demonstrate the effectiveness of our approaches over a range of different DL architectures, showing the generic nature of it.
4. We provide a comprehensive comparison between single stream (for sagittal, coronal and axial), multi-stream (siMS and icMS) and multi-channel approaches for a range of architectures.

2. METHODS

We propose a 2D multi-stream approach to accommodate the different T2w orthogonal scan directions to obtain better feature representations for 2D classification. Figure 2 presents our proposed multi-stream approaches. We start by providing a description of the dataset, the pre-processing steps and the architectures training process.

2.1.Dataset: Multi-planar prostate T2w MRI

We use the ProstateX dataset [16], which is open source[1]. The nature of the study is retrospective and includes different MRI modalities from which axial, sagittal and coronal T2w are used in this work. The cohort included in the study consisted of 204 patients diagnosed with PCa and 330 lesions. Among those lesions, 76 lesions are clinically significant (cS) and 254 are non-clinically significant (ncS). All images are provided with biopsy results, serving as the reference standard. The significance level of the lesions is based on the Gleason score where the lesion is considered cS if the Gleason score is 7 or higher and ncS otherwise. In the case that a slice contains several lesions, we obtain the label from the dominant lesion (higher score) in terms of Gleason score.

2.1.1. Pre-processing and data splitting

Standard pre-processing is applied to the data, including resampling to a common coordinate system by linear interpolation with a resolution of 0.5x0.5x0.5 mm, which is close to the 3mm slice thickness of the best single in-plane scans. Fol- lowing, the slices are cropped to the area corresponding to the intersection between the different scan directions, as Figure 1 exemplifies. By cropping following the intersection of the different directions, we have a flexible approach which can be used in different data-sets in contrast with hard-cropping approaches. We apply normalization of the MRI intensities to a range of [0, 1] and outlier removal by forcing the intensity values between the 1st and the 99th percentiles as part of the pre-processing. We split the original dataset by patients following a 60%/20%/20% for training, validation and testing, respectively.

2.1.2. Statistical analysis

We compare the performance of the architectures trained with and without our proposed siMS and icMS along with multi- channel approaches using the area under the receiver oper- ating characteristic curve (AUC) [17]. In order to evaluate whether the differences between

[1] https://wiki.cancerimagingarchive.net/display/Public/SPIE-AAPM-NCI+PROSTATEx+Challenges

the performance of the models on the test set were significant, we used a non-parametric bootstrap of $n = 100$ bootstrap replicates from the test set to estimate the variability of the model performance. Following this, we obtain the 95% bootstrap confidence intervals (95% CI) and assess the significance at the $p = 0.05$ level by means of Wilcoxon signed-rank [18], which does not assume an underlying Gaussian distribution of the the bootstrap results.

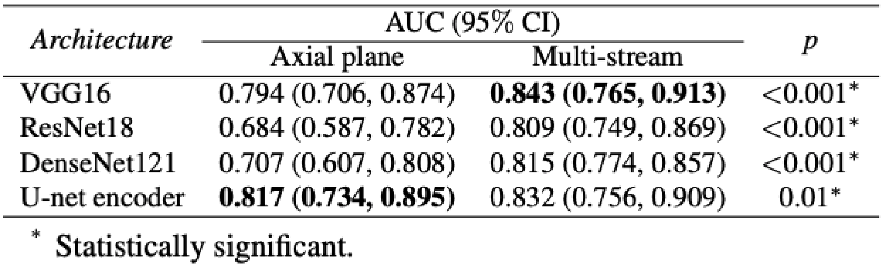
2.2. Classification architectures

In this section, we provide a description of the trained ar- chitectures, considered to be the baseline of the work when trained with a single stream (input) which processes cropped 2D prostate slices. Additionally, we consider a multi-channel approach in which the three T2w directions are stacked and jointly processed in a 3 channel single stream-way.

2.2.1. Baselines

We choose VGG16, ResNet18, DensNet121 and the encoder part of the U-net architecture to evaluate the consistency of our findings across model architectures. Our choice of ar- chitectures is motivated by previous works [11, 13, 19]. We provide more details of the implementations on the source code, available on GitHub[2]. In the following sections we provide a description of the implementation of our proposed multi-stream architectures: siMS and icMS.

Training: We experiment with learning rates of 10−3, 10−4 and 10−5, and investigate their effect on the final classifi- cation performance. We train all the architectures for 200 epochs and saved the weights for the top performing epochs for subsequent model evaluation. We apply L2 regularizationwith a value of 1e−5 to avoid over-fitting of the architectures. We use online augmentation based on previous works [19]: rotation (50 degrees), translation (pixel range of 0.32) and vertical flipping. We run all the experiments on a Tesla V100 with 30GB of RAM.



Architecture	AUC (95% CI)		*p*
	Axial plane	Multi-stream	
VGG16	0.794 (0.706, 0.874)	**0.843 (0.765, 0.913)**	<0.001*
ResNet18	0.684 (0.587, 0.782)	0.809 (0.749, 0.869)	<0.001*
DenseNet121	0.707 (0.607, 0.808)	0.815 (0.774, 0.857)	<0.001*
U-net encoder	**0.817 (0.734, 0.895)**	0.832 (0.756, 0.909)	0.01*

* **Statistically significant.**

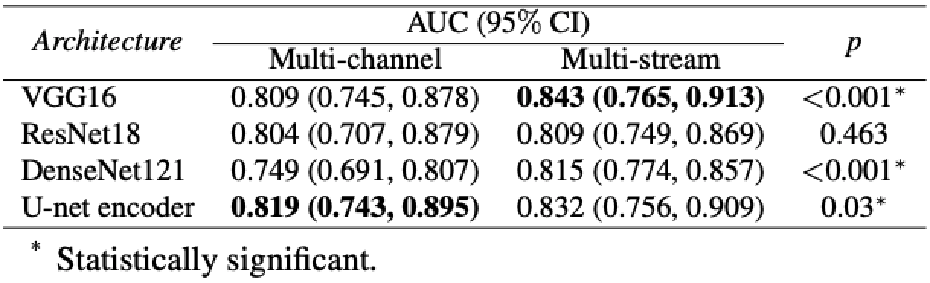
Table 1: 2D PCa lesion classification results for different ar-chitectures and single plane (axial) or simple multi-stream (siMS). Results are shown in terms of mean (95% CI).

2.2.2. Simple multi-stream classification

We propose siMS architecture to accommodate the differ- ent scan directions and process them individually to extract direction-specific features in each stream. Figure 2a illustrates a triple-planar model which processes axial, coronal and sagittal scans. Once the scans have been processed by the chosen architecture, we perform a late fusion approach in which the feature maps extracted by each individual stream are concatenated. Following, we obtain the final classification of the PCa lesion. The training protocol follows the one described in Section 2.2.1.

2.2.3. Inter-connected multi-stream classification

We propose to improve the information flow and obtain more powerful representations by adding inter-connections between the different streams instead of processing them independently followed by a concatenation of the feature maps (Section 2.2.2). We extend the work of [15] to the best performing architecture (Table 1) in a multi-stream configuration among the ones tested in this work and to accommodate or- thogonal scans with different directions. In particular, in icMS, we include inter-connections in which we concate- nate the feature maps obtained by each individual stream in each block of the architecture under consideration (Figure 2 shows an example with two streams and VGG16) as opposed to only the concatenation of the final feature maps obtained by each stream as in siMS. Finally, icMS also incorporates a late fusion of the final feature maps in the form of concatenation.



Architecture	AUC (95% CI)		*p*
	Multi-channel	Multi-stream	
VGG16	0.809 (0.745, 0.878)	**0.843 (0.765, 0.913)**	<0.001*
ResNet18	0.804 (0.707, 0.879)	0.809 (0.749, 0.869)	0.463
DenseNet121	0.749 (0.691, 0.807)	0.815 (0.774, 0.857)	<0.001*
U-net encoder	**0.819 (0.743, 0.895)**	0.832 (0.756, 0.909)	0.03*

* Statistically significant.

Table 2: 2D PCa lesion classification results for different ar-chitectures and simple multi-stream (siMS) or multi-channel. Results are shown in terms of mean (95% CI).

3. RESULTS

We evaluated our proposed approach in terms of AUC and 95% CI, obtained with *n* = 100 bootstrap replicates. Figure 3 portrays the results for the different architectures and different T2w directions, when evaluated independently. In particular, Figure 3 depicts how the coronal and sagittal slices reached significantly lower average AUC when compared to the axial direction for VGG16, the best performing architecture for the siMS approach (0.794 axial vs 0.765 coronal, $p < 0.001$ and 0.794 axial vs 0.761 sagittal, $p < 0.001$). The same trend was observed for ResNet18, DenseNet121 and U-net encoder (Figure 3), confirming the hypothesis that axial slices outper- form coronal and sagittal ones when used as a single input.

Table 1 presents the quantitative results of our proposed multi-stream approach and axial T2w MRI for a range of ar- chitectures. For the sake of simplicity, we compare the siMS approach with axial direction only as it is the one that is pre- dominant in the literature. The siMS approach outperforms the axial-only based one for all the architectures tested in this work. In particular, VGG16 outperforms the rest of the ar- chitectures in a multi-planar setting, achieving an averaged AUC of 0.843 (vs 0.794 for the axial plane, $p < 0.001$) with the siMS approach. As table 2 shows, the siMS approach also outperformed the multi-channel one, showing that an independent processing of the orthogonal planes provides richer features leading to improved lesion classification re- sults. All the computed differences were found to be significant (p < 0.05) except for ResNet18 (Table 2).

In our second experiment, we compared the siMS approach (Figure 2) with icMS for VGG16, the best performing architecture based on AUC in the siMS set- ting (Table 1 and Table 2) and thus the preferred architecture for further comparisons. We found that the icMS approach achieved an averaged AUC of 0.854 (0.838, 0.870) vs 0.843 (0.765, 0.913) of the siMS one. The difference between them was found to be statistically significant (p < 0.001). Additionally, we can observe how the icMS approach seems to be more robust based on the variability around the 95% CI when compared to the siMS one.

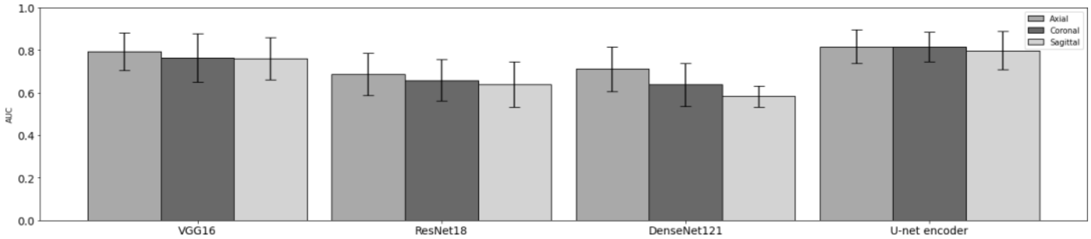


Fig. 3: Results for the different plane directions and different architectures. Bars represent the 95% CI.

4. CONCLUSIONS

We found that by making use of axial, coronal and sagittal MRI directions with multi-stream approaches (siMS and icMS) we are able to significantly improve the 2D lesion classification results in PCa when compared to axial and multi channel approaches for a range of architectures. To the best of our knowledge, the only works that make use of a multi- stream approach for prostate cancer have segmentation as a final objective. Moreover, icMS approaches have yet to be explored for PCa classification. Our work highlights the benefits of using a multi-stream approach for lesion classification without requiring any additional effort in terms of image ac- quisition, as the different orthogonal scan directions are usually acquired by default. Limitations of the work include the retrospective nature of the data and the criteria of the archi tecture choice, based solely on AUC instead of a combina- tion of metrics such as AUC and required computational resources. Future works will explore other types of fusion and architectures (transformers) and their effect on the final performance along with their computational requirements. Finally, future developments will extend the work to a 3D setting, were richer features can be extracted from the whole volume of the patients.

5. COMPLIANCE WITH ETHICAL STANDARDS

This research study was conducted retrospectively using open access data made available by the ProstateX challenge organizers. Ethical approval was not required.

6. ACKNOWLEDGMENTS

This work has been funded by the University of Stavanger. The authors have no relevant financial or non-financial interests to disclose.

7. REFERENCES

[1] Prashanth Rawla, “Epidemiology of prostate cancer,” World journal of oncology, vol. 10, no. 2, pp. 63, 2019.

[2] Alvaro Fernandez-Quilez, Miguel German-Borda, Gabriel Leonardo, Nicola´s Castellanos, Hogne Soen- nesyn, Ketil Oppedal, and Svein Reidar-Kjosavik, “Prostate cancer screening and socioeconomic dispari- ties in mexican older adults,” salud pu´blica de me´xico, vol. 62, no. 2, Mar-Abr, pp. 121–122, 2020.

[3] William J Catalona, Jerome P Richie, Frederick R Ah- mann, M’Liss A Hudson, Peter T Scardino, Robert C Flanigan, Jean B Dekernion, Timothy L Ratliff, Louis R Kavoussi, Bruce L Dalkin, et al., “Comparison of dig- ital rectal examination and serum prostate specific anti- gen in the early detection of prostate cancer: results of a multicenter clinical trial of 6,630 men,” The Journal of urology, vol. 151, no. 5, pp. 1283–1290, 1994.

[4] Jack Cuzick, Mangesh A Thorat, Gerald Andriole, Otis W Brawley, Powel H Brown, Zoran Culig, Ros- alind A Eeles, Leslie G Ford, Freddie C Hamdy, Lars Holmberg, et al., “Prevention and early detection of prostate cancer,” The lancet oncology, vol. 15, no. 11, pp. e484–e492, 2014.

[5] Kathryn K. Hodge, John E. McNeal, and Thomas A. Stamey, “Ultrasound guided transrectal core biopsies of the palpably abnormal prostate,” Journal of Urology, vol. 142, no. 1, pp. 66–70, Jul 1989.

[6] Rianne RM Engels, Bas Israe¨l, Anwar R Padhani, and Jelle O Barentsz, “Multiparametric magnetic reso- nance imaging for the detection of clinically significant prostate cancer: what urologists need to know. part 1: acquisition,” European urology, vol. 77, no. 4, pp. 457– 468, 2020.

[7] Linda M Johnson, Baris Turkbey, William D Figg, and Peter L Choyke, “Multiparametric mri in prostate cancer management,” Nature reviews Clinical oncology, vol. 11, no. 6, pp. 346–353, 2014.

[8] Andrew B Rosenkrantz, Luke A Ginocchio, Daniel Cornfeld, Adam T Froemming, Rajan T Gupta, Baris Turkbey, Antonio C Westphalen, James S Babb, and Daniel J Margolis, “Interobserver reproducibility of the pi-rads version 2 lexicon: a multicenter study of six ex- perienced prostate radiologists,” Radiology, vol. 280, no. 3, pp. 793–804, 2016.

[9] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun, “Dermatologist-level classification of skin can- cer with deep neural networks,” nature, vol. 542, no. 7639, pp. 115–118, 2017.

[10] Alvaro Fernandez-Quilez, Steinar Valle Larsen, Morten Goodwin, Thor Ole Gulsrud, Svein Reidar Kjosavik, and Ketil Oppedal, “Improving prostate whole gland segmentation in t2-weighted mri with synthetically generated data,” in 2021 IEEE 18th International Sympo- sium on Biomedical Imaging (ISBI), 2021, pp. 1915– 1919.

[11] Anindo Saha, Matin Hosseinzadeh, and Henkjan Huis- man, “End-to-end prostate cancer detection in bpmri via 3d cnns: Effect of attention mechanisms, clinical priori and decoupled false positive reduction,” arXiv preprint arXiv:2101.03244, 2021.

[12] Sunghwan Yoo, Isha Gujrathi, Masoom A Haider, and Farzad Khalvati, “Prostate cancer detection using deep convolutional neural networks,” Scientific reports, vol. 9, no. 1, pp. 1–10, 2019.

[13] Nader Aldoj, Steffen Lukas, Marc Dewey, and Tobias Penzkofer, “Semi-automatic classification of prostate cancer on multi-parametric mr imaging using a multi- channel 3d convolutional neural network,” European radiology, vol. 30, no. 2, pp. 1243–1253, 2020.

[14] Baris Turkbey, Andrew B. Rosenkrantz, Masoom A. Haider, Anwar R. Padhani, Geert Villeirs, Katarzyna J. Macura, Clare M. Tempany, Peter L. Choyke, Francois Cornud, Daniel J. Margolis, and et al., “Prostate imag- ing reporting and data system version 2.1: 2019 update of prostate imaging reporting and data system version 2,” European Urology, vol. 76, no. 3, pp. 340–351, Sep 2019.

[15] Jose Dolz, Karthik Gopinath, Jing Yuan, Herve Lom- baert, Christian Desrosiers, and Ismail Ben Ayed, “Hyperdense-net: a hyper-densely connected cnn for multi-modal image segmentation,” IEEE transactions on medical imaging, vol. 38, no. 5, pp. 1116–1126, 2018.

[16] Geert Litjens, Oscar Debats, Jelle Barentsz, Nico Karssemeijer, and Henkjan Huisman, “Computer-aided detection of prostate cancer in mri,” IEEE transac- tions on medical imaging, vol. 33, no. 5, pp. 1083–1092, 2014.

[17] Christopher M Florkowski, “Sensitivity, specificity, receiver-operating characteristic (roc) curves and like- lihood ratios: communicating the performance of diag- nostic tests,” The Clinical Biochemist Reviews, vol. 29, no. Suppl 1, pp. S83, 2008.

[18] RF Woolson, “Wilcoxon signed-rank test,” Wiley ency- clopedia of clinical trials, pp. 1–3, 2007.

[19] Minh Hung Le, Jingyu Chen, Liang Wang, Zhiwei Wang, Wenyu Liu, Kwang-Ting Tim Cheng, and Xin Yang, “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, 2017.