Informatics in Medicine Unlocked 39 (2023) 101234

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: www.elsevier.com/locate/imu

A comparison of Generative Adversarial Networks for automated prostate cancer detection on T2-weighted MRI

Alexandros Patsanis [a,*], Mohammed R.S. Sunoqrot [a,b], Sverre Langørgen [b], Hao Wang [c], Kirsten M. Selnæs [b], Helena Bertilsson [d,e], Tone F. Bathen [a,b], Mattijs Elschot [a,b,**]

[a] Department of Circulation and Medical Imaging, Norwegian University of Science and Technology, Trondheim, Norway
[b] Department of Radiology and Nuclear Medicine, St. Olavs Hospital, Trondheim University Hospital, Trondheim, Norway
[c] Department of Computer Science - Big Data Lab, Norwegian University of Science and Technology, Gjøvik, Norway
[d] Department of Clinical and Molecular Medicine, St. Olavs Hospital, Trondheim University Hospital, Trondheim, Norway
[e] Department of Urology, St. Olavs Hospital, Trondheim University Hospital, Trondheim, Norway

ARTICLE INFO

Keywords:
Deep learning
Generative adversarial networks
Magnetic resonance imaging
Prostate cancer
CAD

ABSTRACT

Generative Adversarial Networks (GANs) have shown potential in medical imaging. In this study, several previously developed GANs were investigated for prostate cancer (PCa) detection on T2-weighted (T2W) magnetic resonance images (MRI).

T2W MRI from an in-house collected dataset (N=961) were used to train, validate, and test an automated computer-aided detection (CAD) pipeline. The open-access PROSTATEx training dataset (N=199) was used as an external test set. The CAD pipeline consisted of normalization, prostate segmentation, quality control, prostate gland cropping, and a GAN model. Six GANs (f-AnoGAN, HealthyGAN, StarGAN, StarGAN-v2, Fixed-Point-GAN and DeScarGAN) were evaluated for PCa detection on the patient-level using the area under the receiver operating characteristic curve (AUC). The best performing GAN (validation set) was trained with five different initializations and evaluated on the internal and external test sets to assess its robustness.

Fixed-Point-GAN performed best (validation, AUC 0.76) and was selected for further assessment. The highest performance on the internal and external test sets were an AUC of 0.73 (95% CI: 0.68-0.79) and 0.77 (95% CI: 0.70-0.83), respectively. The average AUCs ± standard deviation across all runs corresponded to 0.71 ± 0.01 and 0.71 ± 0.04, respectively.

Fixed-Point-GAN was identified as a promising GAN for the detection of PCa on T2W MRI. This model needs to be further investigated and trained on a larger dataset of multiparametric or biparametric MR images to assess its full potential as a support tool for radiologists.

1. Introduction

Prostate cancer (PCa) is the most commonly diagnosed cancer and the second highest cause of cancer-related death in men, worldwide [1]. Magnetic resonance imaging (MRI) prior to biopsy is increasingly used to guide biopsy sampling for suspected prostate cancer due to elevated blood prostate-specific antigen (PSA) levels and digital rectal examination (DRE) [2]. The multi-parametric MRI (mpMRI) protocol for PCa detection consists of T2-weighted imaging (T2W), diffusion-weighted imaging (DWI), and dynamic contrast-enhanced imaging (DCE). The use of mpMRI has improved diagnostic accuracy, and international guidelines have been established [3]. However, detection of PCa on mpMRI remains a difficult task, even for radiologists, and suffers from inter-reader variability [4,5].

Artificial intelligence could support radiological reading, aiming to increase accuracy and efficiency [6]. In PCa, a computer-aided detection (CAD) system would ideally present the radiologist with a patient-level probability of cancer while visualizing the suspected lesion(s) in the prostate gland. Several deep learning (DL) based methods that fulfill these criteria have already been proposed for detection of PCa on mpMRI, with relative success. For example, Saha et al. [7] introduced a multi-stage 3D supervised CNN for localization of clinically significant PCa (csPCa, ISUP ≥ 2) in bi-parametric MRI (bpMRI; T2W + DWI). Also other studies [8–10] have utilized supervised models to detect csPCA and shown promising results. These studies have in common that they rely on convolutional neural networks (CNN) with an

[*] Corresponding author.
[**] Corresponding author.
E-mail addresses: alexandros.patsanis@ntnu.no (A. Patsanis), mattijs.elschot@ntnu.no (M. Elschot).

https://doi.org/10.1016/j.imu.2023.101234
Received 23 November 2022; Received in revised form 22 March 2023; Accepted 27 March 2023
Available online 17 April 2023

encoder–decoder architecture (U-Net), which typically require supervised training on pixel-level annotations provided in the form of (manual) delineations of the lesions. However, these annotations are sparse due to the costly and time-consuming labeling process [11]. At the same time, inter-observer variability [4,5] can lead to an inaccurate, noisy, and annotator-dependent reference standard.

Given the strong dependence of AI performance on the size of the training data [8], unsupervised and weakly-supervised learning are interesting alternatives to supervised learning. In contrast to supervised learning, these approaches require no labels (unsupervised), or only image-level (weakly supervised, 2D) or patient-level (3D) labels for training, which are easier to obtain in practice than pixel-level annotations. In contrast to most CNNs in medical imaging, Generative Adversarial Networks (GANs) [12] can provide pixel-level output (i.e., images) while being trained in an unsupervised or weakly-supervised manner [13]. For a tumor detection task, GAN models can be trained to generate images from the negative class (i.e., no disease) by either learning the distribution from the negative class only (e.g f-AnoGAN [14]), or by learning to perform image-to-image translation to the negative class (e.g HealthyGAN [15], StarGAN [16], StarGAN-v2 [17], Fixed-Point GAN (FP-GAN) [18], DeScarGAN [19]). In both cases, anomalies such as PCa can then be highlighted during inference based on the difference between the input image and the generated negative output image.

The aim of this study is to assess the potential of several GAN models for the task of PCa detection on T2W MRI. To the best of our knowledge, such a comparison has not been reported in literature. Furthermore, GAN model performance for other medical problems is rarely reported on patient-level and/or using external test sets. Therefore, in this study, we compare several GAN models in a fully automated end-to-end pipeline for PCa detection and assess patient-level performance using internal and external test sets.

2. Methods

2.1. Dataset

Transverse T2W MR images from two datasets (N = 1160) were used in this study: an in-house collected dataset (N = 961) and the publicly available PROSTATEx Challenge [20] training dataset (N = 199). In both datasets, patients with a Gleason grade group (GGG) < 1 or those who did not undergo biopsy due to a clinical indication and Prostate Imaging Reporting and Data System (PI-RADS) score < 3 were defined as negative for PCa, whereas patients with GGG $\geq$ 1 were defined as positive for PCa. The task of any PCa (GGG $\geq$ 1) detection rather than csPCa (GGG $\geq$ 2) detection was chosen to compare the anomaly detection abilities of the models rather than their absolute performance on csPCa. Table 1 gives an overview of the patients' characteristics of the in-house collected dataset and Table S1 (Appendix A.) provides an overview of the GGG distributions of the internal and external test sets used in this work.

The in-house dataset consisted of diagnostic MR images from men enrolled in the standardized prostate cancer pathway at St. Olavs Hospital, Trondheim University Hospital, Trondheim, Norway, from January 2013 to December 2020, due to suspicion of PCa. The Regional Committee for Medical and Health Research Ethics (REC Mid Norway) approved the use of the dataset (identifier 2017/576) and granted permission for passive consent to be used. T2W imaging was performed on Magnetom Skyra (951 patients), Prisma (4 patients), or Biograph mMR (6 patients) 3T MRI systems (Siemens Healthineers, Erlangen, Germany) with a turbo spin-echo sequence (repetition time 3320–10 710 ms, echo time 93–108 ms, flip angle 120–160 degrees, matrix size 256 × 256–640 × 640, 16–36 slices with 3–3.5 mm slice thickness and 0.3 × 0.3–0.75 × 0.75 mm^2 in-plane resolution). 507 patients were classified as positive, and 454 patients were classified as negative for PCa. Lesion-level (and, consequently, image-level) annotations were available for 133 positive patients. These were based on delineations of PI-RADS $\geq$ 3 lesions performed by a radiology resident trained by a radiologist (S.L.) with more than a decade of experience. In this study, these 133 patients were selected for training due to the availability of image-level annotations. The patients were split into training (N = 266, 133 positive/133 negative), validation (N = 100, 50 positive/50 negative) and internal test (N = 595, 324 positive/271 negative) sets. The PROSTATEx challenge [20] training dataset (N = 199, 98 positive/101 negative) was used as an external test set. Details of the acquisition protocol are provided in [20]. Manual delineation of PI-RADS $\geq$ 3 lesions was performed by imaging experts (at Miller School of Medicine, Miami, FL, USA) based on targeted biopsy locations provided by the challenge organizers. 5 patients from the original dataset were excluded from analysis due to sub-optimal image or annotation quality.

Table 1
Overview of patients characteristics in the in-house collected dataset. Numbers are reported as mean (range) [number of patients with missing information]. PI-RADS refers to the PI-RADS score of the index lesion. The following abbreviations are listed in the table: Prostate-specific antigen (PSA), Gleason grade group (GGG), and transrectal ultrasound examination (TRUS).

	Negative	Positive
Patients (N = 961)	454	507
PSA (ng/mL2)	7 (0.4–99) [45]	8.5 (0.2–2843) [3]
Prostate volume	57 (15–312) [1]	39.5 (11.7–197) [17]
PSA density	0.13 (0.01–1.42) [42]	0.21 (0.02–66.7) [41]
Age	64 (21–80) [1]	68 (41–85)
Biopsy (Patients)	Negative (N = 208)/ not performed (N = 246)	Positive (N = 507)
Prostatectomy	–	N = 224
Systematic TRUS Biopsy (Patients)	N = 122 [6]	N = 76 [14]
Targeted biopsy (Patients)	N = 80	N = 417
PI-RADS (Index)		
1/2	370	78
3	41	47
4	22	121
5	21	261
GGG		
1	N/A	96 (19%)
2	N/A	179 (35%)
3	N/A	122 (24%)
4	N/A	41 (8%)
5	N/A	69 (14%)

2.2. End-to-end pipeline

Fig. 1 provides an illustration of the end-to-end pipeline to benchmark PCa detection using GANs. The pipeline consists of an image pre-processing module, a GAN module, and a post-processing module, each of which is explained in more detail below. The modules were implemented in Python (versions 2.7 & 3.7), except for normalization using AutoRef [21] and segmentation quality control (QC) [22], which were implemented using MATLAB (The Mathworks, Nattick, MA; version r2020b). All experiments were performed using Ubuntu 18.04.4L operating system with a single NVIDIA Tesla V100 PCIe 32 GB GPU.

2.2.1. Pre-processing module

As shown in Fig. 1, the pre-processing module performs T2W image intensity normalization, segmentation of the prostate gland, and sampling of the input images used for model training and inference. Intensity normalization was performed using AutoRef [21], a dual reference tissue (fat and muscles) normalization approach that also incorporates N4 bias field correction [24]. AutoRef produces pseudo-T2 images where prostate tissue is expected to have a value of approximately 80 ms. Image intensities larger than twice the expected pseudo-T2 value of fat (242 ms) were truncated to standardize the image intensity range (0–242). V-Net [25] and nnU-Net [26] models

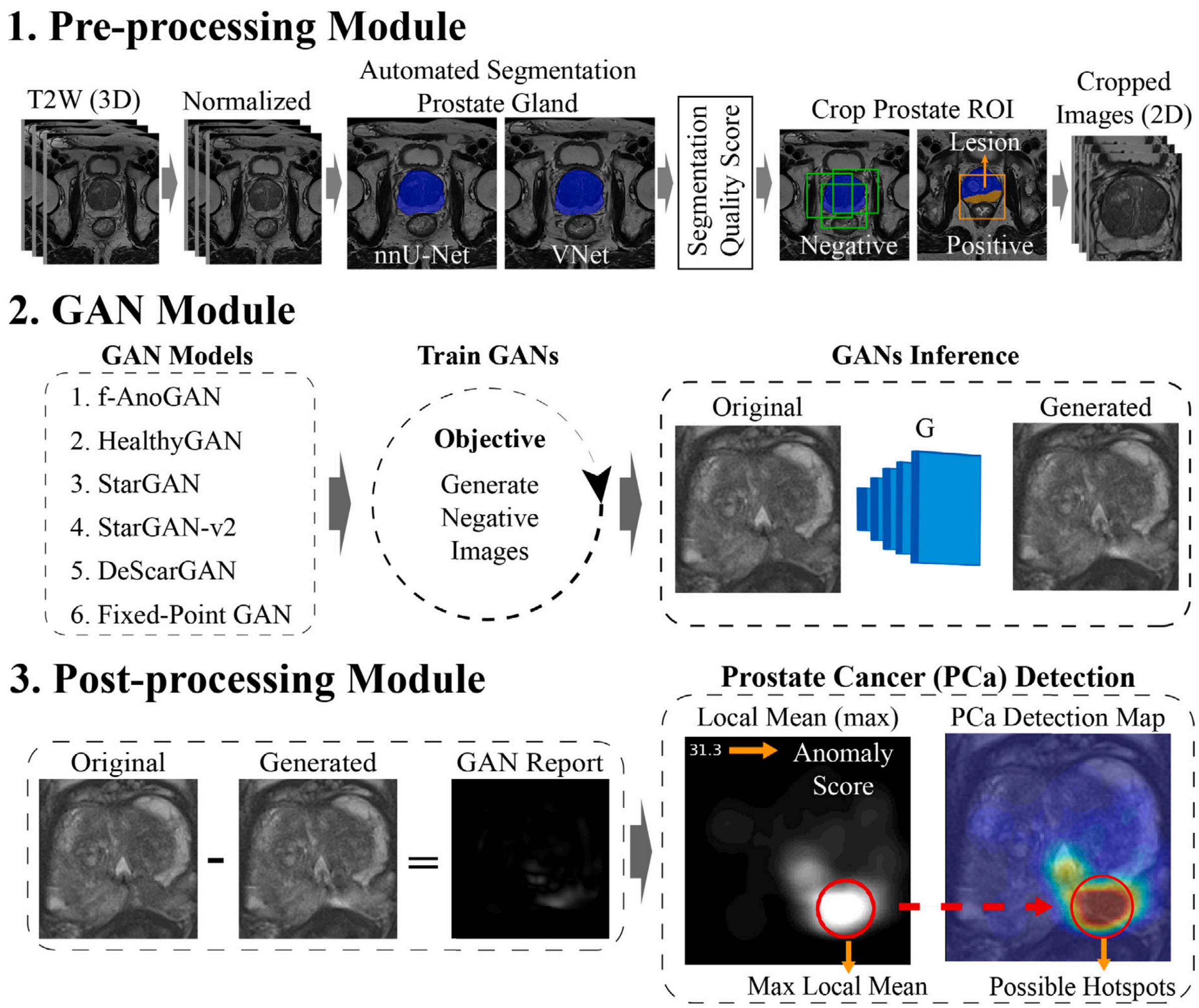

Fig. 1. shows the automated end-to-end pipeline for PCa detection. The pre-processing module (1) includes automated T2-weighted (T2W) normalization, prostate segmentation (3D), followed by a segmentation quality control step, and an automated sampling technique that crops image patches (2D). The cropped 2D images are used for training, validation and testing of the GANs. The GAN module (2) involves training and inference of the GANs on 2D MR images. The real cropped images are sampled from the original MRI images and the generated images are synthesized by the generator (G). During inference, the trained GAN models are used to generate negative images from input images (2D) of unknown health status. In the post-processing module (3), GAN reports (2D) are created as the difference between the input images and the generated images. The 2D GAN reports are then accumulated and smoothed with a local mean filter to produce a 3D cancer localization map. A patient-level anomaly score, defined as the maximum value in the 3D cancer localization map, represents the likelihood of detected PCa.

Table 2

Number of cropped images for each of the generated datasets for 4 different pixel spacing values (0.2 × 0.2, 0.3 × 0.3, 0.4 × 0.4, and 0.5 × 0.5 mm^2) using CROPro [23]. All GAN models were trained, validated, and tested with 266, 100, and 595 patients, respectively.

Pixel spacing (mm)	Train (N = 266, 133/133)		Validation (N = 100, 50/50)		Test (N = 595, 271/324)	
	Negative	Positive	Negative	Positive	Negative	Positive
0.5 × 0.5	7690	3011	771	688	4838	4591
0.4 × 0.4	12 397	4944	1104	811	7710	5701
0.3 × 0.3	19 099	8803	2815	1670	20 386	12 755
0.2 × 0.2	52 491	21 085	11 377	7014	77 484	52 517

were trained on 89 patients (in-house dataset) from whom manual segmentation of the prostate gland was available and were then used to automatically segment the 3D prostate volume on the normalized T2W images (see Appendix A., Table S2 for segmentation hyperparameters). Subsequently, a pre-trained quality control (QC) system [22] was used to select the best-performing segmentation model for each case. The segmentation with the highest QC score was selected, while cases with QC scores lower than 80% on both segmentations (N = 13, negative: 8, positive: 5) were excluded from analysis. Finally, CROPro [23], an in-house developed cropping technique, was used to resample the normalized T2W images and prostate masks to a new pixel-spacing (0.5 × 0.5, 0.4 × 0.4, 0.3 × 0.3, or 0.2 × 0.2 mm^2) and sample 2D input images of size 128 × 128 for training, validation, and testing of the GAN models. The cropping settings were based on the optimal crop size found for CNNs and Vision Transformers (~10 cm^2) in previous work [23]. Here, we selected a range from ~8 cm^2 to ~20 cm^2 to investigate the optimal cropping size for GANs. Separate models were trained and evaluated for each combination of pixel-spacing and GAN type. For the training set, input images were sampled randomly from the prostate mask (negative patients) or from the segmented lesions (positive patients). For the validation and test sets, images were sampled from the complete prostate mask using a stride of 32 pixels. Table 2 gives an overview of the number of cropped images for the train, validation, and internal test sets.

Table 3
The table presents differences in loss functions among all selected GANs. Five out of six models use the WGAN-GP [27] architecture, with different combinations of losses for both Generators ($\mathcal{L}_G$) and Discriminators ($\mathcal{L}_D$). There are various types of losses, including an adversarial loss ($\mathcal{L}_{adv}$), a domain classification loss ($\mathcal{L}_{cls/domain}$), a cycle consistency loss ($\mathcal{L}_{cyc}$) also referred to as a reconstruction loss ($\mathcal{L}_{rec}$), a conditional identity loss ($\mathcal{L}_{id}$), a focus loss ($\mathcal{L}_f$), a style reconstruction loss ($\mathcal{L}_{sty}$), a diversity-sensitive loss ($\mathcal{L}_{ds}$), and a $\mathcal{R}1$ regularization loss ($\mathcal{L}_{R1_{reg}}$). Even when models use the same loss, there may be differences in implementation.

Model	WGAN-GP	Generator ($\mathcal{L}_G$)							Discriminator ($\mathcal{L}_D$)		
		$\mathcal{L}_{adv}$	$\mathcal{L}_{cls/domain}$	$\mathcal{L}_{cyc/rec}$	$\mathcal{L}_{id}$	$\mathcal{L}_f$	$\mathcal{L}_{sty}$	$\mathcal{L}_{ds}$	$\mathcal{L}_{adv}$	$\mathcal{L}_{cls/domain}$	$\mathcal{L}_{R1_{reg}}$
f-AnoGAN [14]	✓	✓							✓		
HealthyGAN [15]	✓	✓		✓	✓	✓			✓		
StarGAN [16]	✓	✓	✓	✓					✓	✓	
StarGAN-v2 [17]		✓		✓			✓	✓	✓	✓	✓
FP-GAN [18]	✓	✓	✓	✓	✓				✓	✓	
DeScarGAN [19]	✓	✓	✓	✓	✓				✓	✓	

2.2.2. GAN module

Six relevant 2D GANs were selected for investigation: f-AnoGAN [14], HealhtyGAN [15], StarGAN [16], StarGAN-v2 [17], FP-GAN [18] and DeScarGAN [19]. For each model, the code was publicly available. All input annotations for training were on the image level, i.e., each input image was labeled as negative or positive. All models were trained to generate 2D images representing a negative patient.

There are several differences between the evaluated models. f-AnoGAN and HealthyGAN are unsupervised networks, whereas StarGAN, StarGAN-v2, FP-GAN, and DeScarGAN are weakly-supervised GANs that use the image-level class as a condition during training. Although the latter models perform supervised image-level detection and only weakly-supervised lesion localization, they are referred to as "weakly-supervised" in this work to be consistent with terminology used in the original papers. All models are image-to-image translation models except f-AnoGAN. f-AnoGAN aims to generate an image using a fast mapping (encoder) from input images to encodings in GAN's latent space, whereas HealthyGAN, StarGAN, StarGAN-v2, FP-GAN, and DeScarGAN aim to translate an input image to a different domain. f-AnoGAN uses only (unlabeled) negative images during training. HealthyGAN uses two unlabeled datasets during training, one with only negative images and one with a mix of negative and positive images. StarGAN, StarGAN-v2, FP-GAN, and DeScarGAN use both negative and positive labeled images during training. Also the loss functions differ between models, as summarized in Table 3. Furthermore, the size of the input images varies: HealthyGAN, StarGAN, StarGAN-v2, and FP-GAN are able to work with different input sizes, whereas f-AnoGAN and DeScarGAN have fixed input sizes of 64 × 64 and 256 × 256, respectively. For f-AnoGAN and DeScarGAN, the input images were resampled to meet the requirements of the networks.

To allow for a fair comparison of model performance, the same pre-processed images were used as input for all models. The hyperparameters of the GAN models were kept as proposed in their original implementations (see Appendix A., Table S3), except for the iteration number, which was expanded to 600K iterations for HealthyGAN, StarGAN, and FP-GAN. The generated output images of each model were further processed to provide a patient-level anomaly score (see 2.2.3), which was used to predict cancer status.

2.2.3. Post-processing module

The 2D output images of the GAN models represent a negative ("healthy") version of the input image, and differences between the generated negative image and the input image are expected to be indicative of PCa. For each image, a difference map (GAN report) was therefore calculated as the absolute difference between the cropped input image and the generated output image. For each image slice, overlapping difference maps (due to stride sampling) were pixel-wise averaged to create a 3D difference map of the same dimensions as the original T2W image. Finally, to reduce the impact of noisy pixels, this difference map was smoothed by using a mean filter with radius of 5 mm in-plane, producing the final 3D cancer detection map. The patient-level anomaly score was then defined as the highest value in the 3D cancer detection map. All values outside the segmented prostate gland were discarded.

2.3. Statistical analysis

The performance of PCa detection was compared between GAN models in the validation set. The performance of the most promising model was further assessed with internal and external test sets. Five additional training runs with random initialization on the same dataset were performed to determine the training stability of the best-performing model.

Classification performance was assessed on the patient-level using the area under the receiver-operating characteristics curve (AUC). The DeLong method was used to compare AUCs between different models. The 95% confidence interval (CI) of the AUC was calculated using the bootstrap method. Performance across multiple runs was reported as mean ± standard deviation (SD). The Wilcoxon rank sum test was used to assess statistical differences between anomaly scores of positive and negative patients.

3. Results

All models except f-AnoGAN and StarGAN-v2 performed best when trained on input images with a pixel spacing of 0.4 × 0.4 mm (Appendix A., Table S4). The performance of the GAN models on the validation set with these cropping settings is shown in Table 4. FP-GAN performed best, with an AUC of 0.76 (95% CI: 0.65–0.84). This was significantly better than the performance of f-AnoGAN and StarGAN-v2, but not HealthyGAN, StarGAN, and DeScarGAN.

Fig. 2 shows the output of all models for a patient with a histopathologically confirmed GGG 3 tumor in the left side of the peripheral zone. It can be observed that FP-GAN and StarGAN are the only models that convincingly visualize the tumor on the PCa detection map.

The performance of FP-GAN was further evaluation on the test sets. The AUC was 0.72 on both the internal and external test sets using the initial model (Table 5). Stable performance with a standard deviation of 1%–4% was observed across the five randomly initialized models (Fig. 3, Table 5). This is further illustrated in Fig. 4, which shows that all model initializations successfully detected the malignant area in a patient with a histopathologically confirmed GGG 2 tumor. Fig. 5 shows two additional examples, where the model fails and succeeds to detect the lesion. Fig. 6 shows that the anomaly scores corresponding to all model initializations are consistently higher ($p < 0.001$) for positive than negative patients in the test sets.

4. Discussion

In this study, we benchmarked several GAN models in an end-to-end CAD pipeline for PCa detection on T2W MR images. The method is designed to generate a 3D visual report for detection of lesions, together with a patient-level anomaly score representing the overall likelihood of cancer. The main objective of this study was to compare the performance of different GAN models for detection of PCa. Six GANs were selected based on availability and previous success in medical imaging tasks: f-AnoGAN, HealthyGAN, StarGAN, StarGAN-v2,

Table 4
Patient-level area under the receiver-operating characteristic curve (AUC) in the validation set for each of the six models. The model with the highest AUC was selected for further analysis (highlighted in bold).

Model	Type	Pixel spacing (mm)	Best found iteration	Validation patient-level AUC (CI 95%)
StarGAN-v2	Weakly-supervised	0.4 × 0.4	200 000	0.49 (0.37–0.60)
f-AnoGAN	Unsupervised	0.4 × 0.4	180 000	0.54 (0.43–0.66)
DeScarGAN	Weakly-supervised	0.4 × 0.4	190 000	0.68 (0.56–0.77)
HealthyGAN	Unsupervised	0.4 × 0.4	580 000	0.69 (0.58–0.78)
StarGAN	Weakly-supervised	0.4 × 0.4	570 000	0.70 (0.60–0.80)
Fixed-Point-GAN	Weakly-supervised	0.4 × 0.4	340 000	**0.76 (0.65–0.84)**

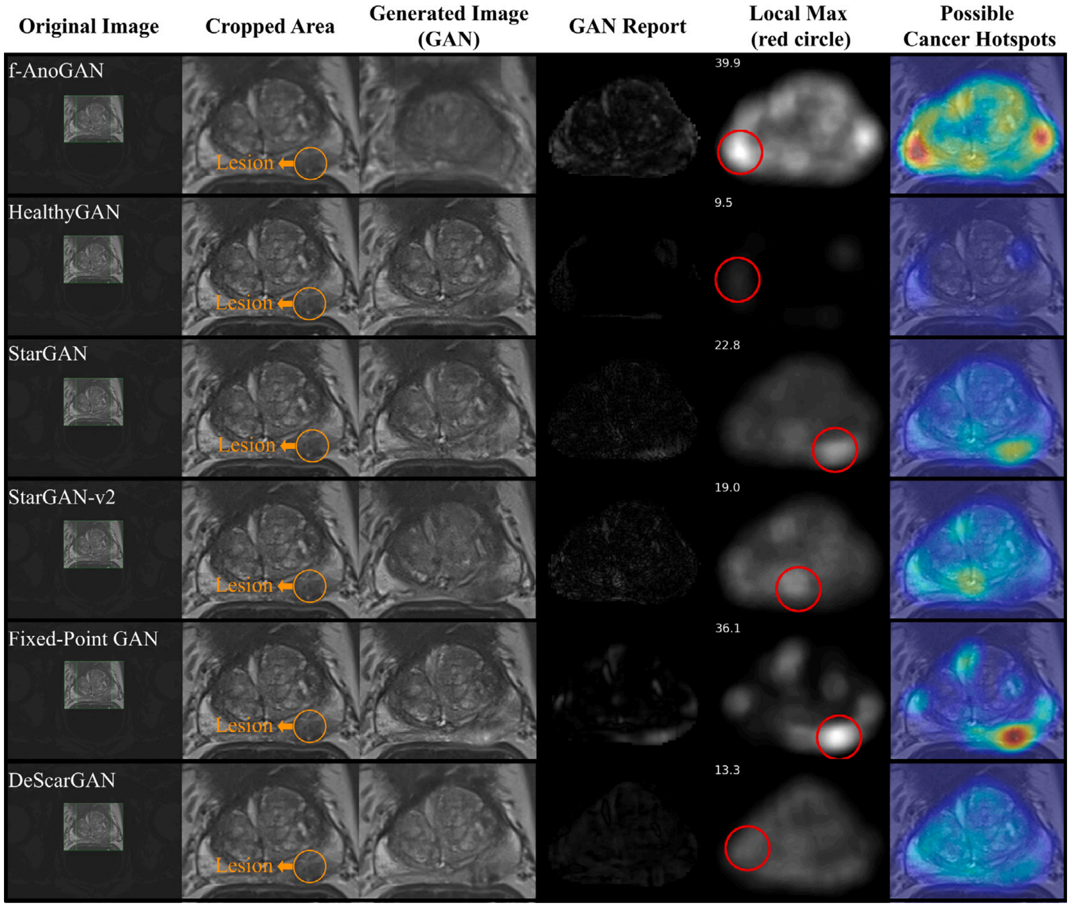

Fig. 2. shows a slice through the middle of the prostate of a positive patient (validation set) with a biopsy-confirmed GGG 3 lesion in the peripheral zone evaluated for all models. Columns represent the original image (green boxes represent the stride cropped images), the evaluated cropped area, the stitched generated images, the difference between the original and the generated image (GAN report), the filtered difference image where the red circle represents the maximum local mean, and the final PCa detection map. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

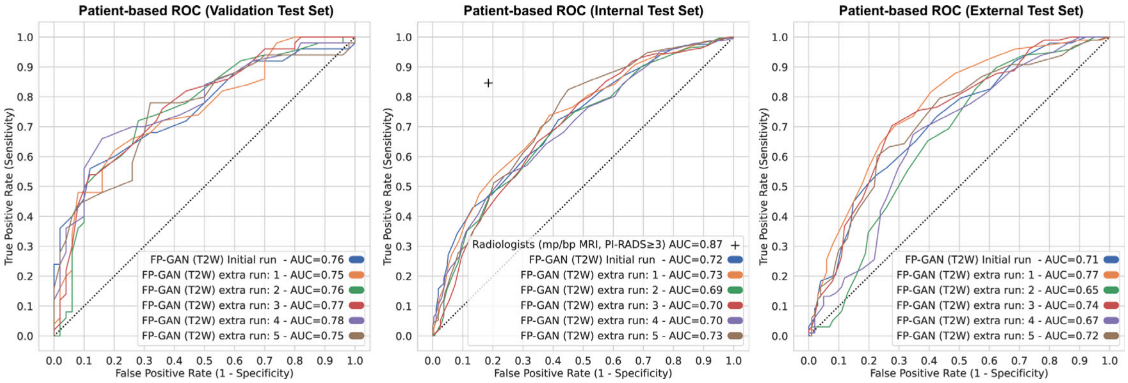

Fig. 3. The AUCs for validation, internal and external test sets are presented for the initial FP-GAN model and five additional models trained with random initialization. For the internal test set, the AUC for any PCa detection by a radiologists according to PI-RADS is presented for reference. Depending on the year of referral, patients may have been examined with mpMRI or bpMRI.

Table 5
AUCs of the best found model (FP-GAN) on the validation set trained for 5 more runs with the same split of the dataset to evaluate the stability of the model. The input images are 128 × 128 pixels and the pixel spacing is 0.4 × 0.4 mm^2.

Model runs	Best iteration	Validation AUC (CI 95%)	Internal AUC (CI 95%)	External AUC (CI 95%)
FP-GAN: 1	300 000	0.75 (0.64–0.84)	0.73 (0.69–0.77)	0.77 (0.70–0.83)
FPGAN: 2	280 000	0.76 (0.65–0.85)	0.69 (0.65–0.73)	0.65 (0.57–0.73)
FP-GAN: 3	550 000	0.77 (0.68–0.85)	0.70 (0.66–0.74)	0.74 (0.67–0.80)
FP-GAN: 4	140 000	0.78 (0.67–0.86)	0.70 (0.65–0.74)	0.67 (0.59–0.74)
FP-GAN: 5	160 000	0.75 (0.64–0.84)	0.73 (0.68–0.79)	0.72 (0.64–0.79)
FP-GAN: Initial	340 000	0.76 (0.65–0.84)	0.72 (0.67–0.75)	0.72 (0.64–0.78)
	Mean ± SD	0.76 ± 0.01	0.71 ± 0.01	0.71 ± 0.04

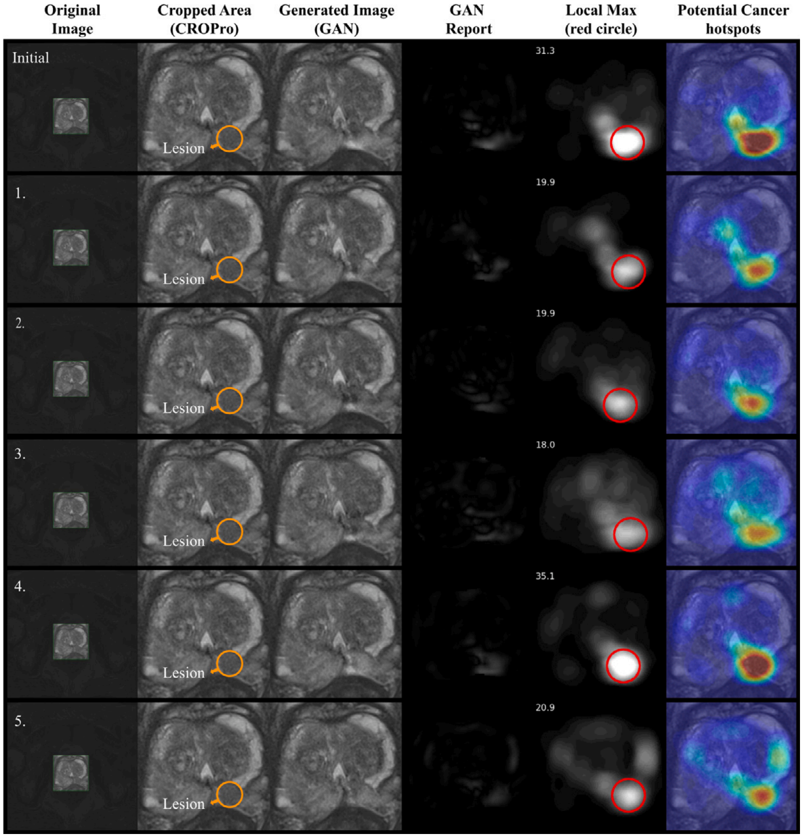

Fig. 4. shows a slice through the middle of the prostate from a patient (external test set) with a confirmed GGG 2 peripheral zone lesion. The rows represent the initial model and the models from 5 additional runs with random initialization. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

FP-GAN and DeScarGAN. Our results indicate that FP-GAN shows most potential for the patient-level detection of PCa on T2W MR images.

Previous studies have shown promising results for csPCA detection on bpMRI (e.g., [7,8]) and mpMRI (e.g., [9]) with supervised CNNs. Also the five best performing networks in the currently ongoing PI-CAI grand challenge [28] for detection of csPCa on bpMRI are based on supervised CNNs [29], with AUCs for patient-level detection of csPCa in the test set ranging from 0.871 to 0.889. However, these models require pixel-level labels, which are difficult to obtain in practice and may hinder future training on very large datasets. To overcome that limitation, Wang et al. [30] proposed a dual-path CNN, one for T2W and one for ADC, that generates cancer detection maps while only requiring image-level annotations, similar to the GANs benchmarked in this work. They reported an AUC of 0.979 ± 0.009 for detection of csPCa on bpMRI. However, the performance evaluation was limited to image-level analysis using manually selected slices where lesions were clearly visible, without external testing on an independent dataset.

GANs have been successfully used in prostate imaging tasks. They have been applied for prostate gland segmentation [31], fiducial marker detection [32], MRI intensity normalization [33], and hematoxylin and eosin staining [34]. Detection of csPCa using supervised fully connected networks (FCNs) with adversarial training is a relatively unexplored research area. Only one study has proposed such a method; Kohl et al. [35] used an adversarial network to discriminate between expert decisions and generated annotations from a U-Net-based [36] architecture for semantic segmentation of csPCa lesions. Of note, this approach requires pixel-level annotations from expert radiologists, whereas the GANs benchmarked in our study require only image-level annotations (weakly-supervised) or no annotations (unsupervised).

We investigated studies based on GANs that have been proposed for a variety of anatomical areas and diseases [13,15,18,37]. We identified nine unsupervised or weakly-supervised GAN models where implementation code was available. Three models were excluded from analysis; AnoGAN [38] was discarded because it is an earlier, much slower, version of f-AnoGan [14]. BigBiGAN was discarded based on the visual observation of the reconstructed images of the original paper [39], indicating it would not be capable to learn the heterogeneous distribution of T2W prostate MRI. VA-GAN [40] was discarded because it requires class label for each input image during inference.

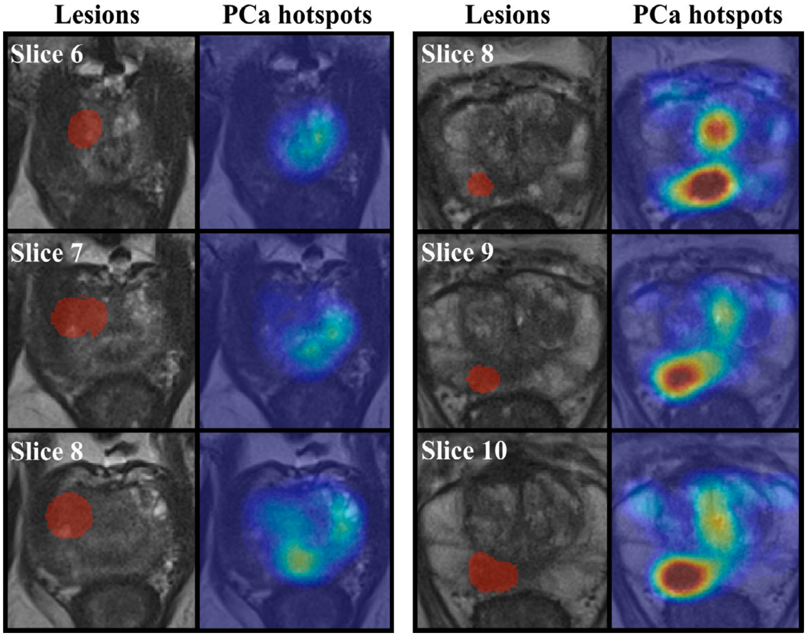

Fig. 5. presents results from the initial model with FP-GAN from three consecutive slices for two patients (external test set) with a confirmed GGG 2 lesion. The first column shows the T2W MRI image overlaid with the true positive lesions, while the second column shows the cancer hotspots generated from GANs. The case on the left represents a poor case where the model failed to detect the lesion, while the case on the right represents a case where the model successfully finds the area of the lesion.

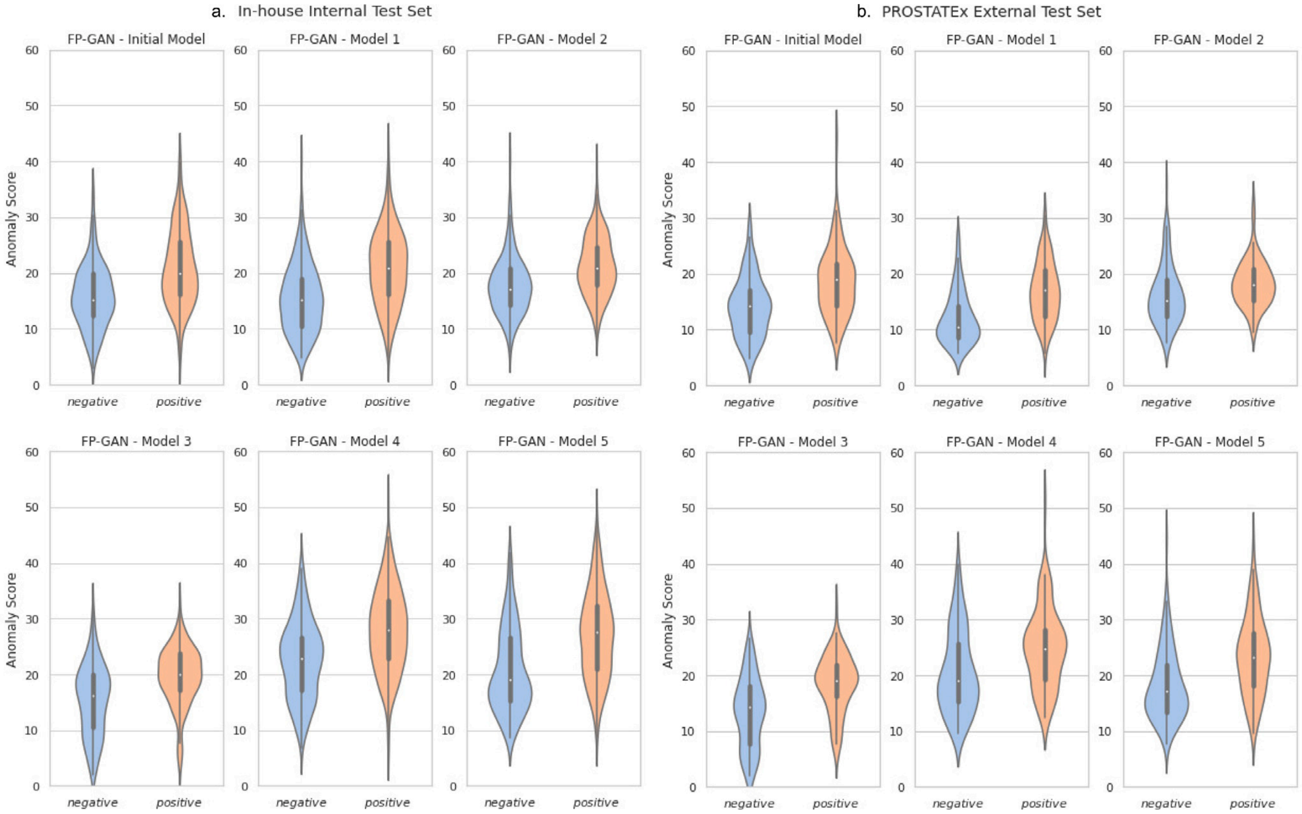

Fig. 6. shows violin plots of the patient-level anomaly scores in (a) the internal test set and (b) the external test set, for each of the FP-GAN model initializations. The mean anomaly score is significantly higher ($p < 0.001$) in positive than in negative patients in both internal and external datasets for all model initializations.

FP-GAN trained on 128 × 128 images with pixel spacing of 0.4 × 0.4 mm^2 performed best in the validation set. The relatively poor performance of f-AnoGan and HealthyGAN could be due to the difficulty of unsupervised approaches to learn the underlying distribution of the data space (see Appendix A., Figure S1), because prostate images vary in shape and size and negative ("healthy") prostate images often harbour benign lesions such as prostatic hyperplasia (BPH) and prostatitis. Another reason could be that unsupervised learning requires more training data than used in this study. In comparison to the other weakly-supervised methods, the success of FP-GAN may be attributed to the ability of its generator to produce an additive map rather than generating an entirely new image. FP-GAN is an extension of StarGAN and shows similar results to that model (Fig. 2). However, it can be observed that FP-GAN makes minimal "local" changes whereas more significant changes in the entire prostate region are seen for StarGAN. Furthermore, our results suggest that the poor performance of StarGAN-v2 may be due to the fact that the model aims to change the overall image style, resulting in widespread changes in the prostate (see Appendix A., Figure S3). Similar behavior was also observed with DeScarGAN (see Appendix A., Figure S4).

Due to the inherent difficulty of training GANs, which do not always converge [41], model performance was assessed by saving models and calculating the AUC for different iterations/epochs. One observation was that the optimal model (i.e, highest AUC) was not always found for the same iteration/epoch. This was evident when the same data set was used to train FP-GAN for 5 different runs, with the best iterations varying between 140 000 and 550 000. However, we used independent test data to investigate the generalizability of the model, which showed consistent performance across all runs in both the internal and external test set (Table 4). Furthermore, model stability was confirmed by visual presentation of the detection maps, which showed consistency throughout the slices (Fig. 5 and Appendix A., Figure S2) and per model initialization (Fig. 4).

One inherent limitation of the unsupervised and weakly-supervised GANs tested in this study is that the image features they learn to be important for detection do not necessarily need to originate from the tumor, but can be due to any systematic difference in the appearance of the negative and positive patients. Elevated PSA values and clinical symptoms can for example also be caused by benign BPH, which is not cancer but typically leads to an enlarged prostate. In our cohort, prostate volume was indeed significantly higher in negative patients than in positive patients. Although the visual presentation of detection maps indicates that the FP-GAN model mainly focuses on cancer lesions, this should be confirmed in more thorough analysis of its lesion detection capabilities, preferably based on mpMRI or bpMRI data.

In this study, we found that FP-GAN has potential for detecting PCa on T2W MR images. This is a solid basis for further method development, with the eventual goal of assessing the full potential of GANs as a support tool for radiologists to detect csPCa on bpMRI or mpMRI. However, the current work has several limitations. GGG from biopsy or radical prostatectomy (when available) was considered the gold standard, whereas in the absence of biopsy, only cases with PI-RADS 1 and 2 were used for training. Defining PCa is subject to limitations in biopsy sampling, which may miss the cancer [42], and inter/reader variability [4,5] of both radiologists and pathologists. In addition, we chose the task of detection of any PCa (GGG ≥ 1) for comparison between models, as this is arguably most consistent with anomaly detection. However, detection of csPCa (GGG ≥ 2) is regarded to be a more clinically useful task. Furthermore, we only reported patient-level classification performance and did not quantitatively evaluate lesion detection performance. We regard this subject of future research, as lesion detection based on PI-RADS also requires DWI and (to some extend) DCE sequences [3], whereas only T2-weighted scans were used in this study. Extending our pipeline to mpMRI or bpMRI will be necessary, as certain lesions may be missed when only T2W images are used. However, it remains to be seen whether GANs perform well on the more noisy DWI images. Finally, although pixel-level annotations are not required for training of the investigated GANs, we used the expert tumor segmentations when preparing the positive images in the training dataset. As a result, our training set was relatively small in terms of patient numbers, and results may be improved when the models are trained on a larger set of images.

5. Conclusion

We benchmarked six previously developed GAN models in an end-to-end pipeline for automated PCa detection on T2W MRI. The pipeline is designed to generate a 3D visual report for detection of lesions, together with a patient-level anomaly score representing the overall likelihood of cancer. FP-GAN was identified as the most promising GAN model for the given task, and was successfully tested on an internal and external dataset, showing generalizability. This work shows that GANs have potential for detection of PCa and serves as a solid basis for further development, including training and testing of FP-GAN for the task of detection and localization of csPCa on a larger set of bpMRI or mpMRI data.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Tone F. Bathen reports financial support was provided by Research Council of Norway. Mattijs Elschot reports financial support was provided by The Norwegian Cancer Society and Prostatakreftforeningen. Mattijs Elschot reports financial support was provided by The liaison Committee between the Central Norway Regional Health Authority and the Norwegian University of Science and Technology. Mohammed R. S. Sunoqrot reports financial support was provided by The liaison Committee between the Central Norway Regional Health Authority and the Norwegian University of Science and Technology. Tone F. Bathen, Mattijs Elschot, Mohammed R. S. Sunoqrot and Alexandros Patsanis has patent #GB2206589.0 pending to Licensee.

Acknowledgments

We wish to express our gratitude to SPIE, AAPM, NCI, and Radboud University Nijmegen (Nijmegen, The Netherlands), as well as the PROSTATEx Challenge organizers for providing their dataset. We would also like to thank Prof. Radka Stoyanova and her team at Miller School of Medicine (Miami, FL, USA) for providing the prostate delineations for the PROSTATEx dataset. We would also like to thank Elise Sandsmark for providing the prostate delineations for the in-house dataset. Finally, we would also like to thank Torill Eidhammer Sjøbakk for her technical assistance and compilation of the data for the in-house dataset.

Funding

This work was partly funded by the Research Council of Norway (295013), the Norwegian Cancer Society and Prostatakreftforeningen (215951), and the liaison Committee between the Central Norway Regional Health Authority and the Norwegian University of Science and Technology (983005100, 982992100, 90265300 and 81770928).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.imu.2023.101234.

References

[1] Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, Bray F. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: Cancer J Clin 2021;71(3):209–49. http://dx.doi.org/10.3322/caac.21660.

[2] Mottet N, van den Bergh RC, Briers E, Van den Broeck T, Cumberbatch MG, De Santis M, Fanti S, Fossati N, Gandaglia G, Gillessen S, Grivas N, Grummet J, Henry AM, van der Kwast TH, Lam TB, Lardas M, Liew M, Mason MD, Moris L, Oprea-Lager DE, van der Poel HG, Rouvière O, Schoots IG, Tilki D, Wiegel T, Willemse P-PM, Cornford P. EAU-EANM-ESTRO-ESUR-SIOG guidelines on prostate cancer—2020 update. part 1: Screening, diagnosis, and local treatment with curative intent. Eur Urol 2021;79(2):243–62. http://dx.doi.org/10.1016/j.eururo.2020.09.042.

[3] Weinreb JC, Barentsz JO, Choyke PL, Cornud F, Haider MA, Macura KJ, Margolis D, Schnall MD, Shtern F, Tempany CM, Thoeny HC, Verma S. PI-RADS prostate imaging – reporting and data system: 2015, version 2. Eur Urol 2016;69(1):16–40. http://dx.doi.org/10.1016/j.eururo.2015.08.052.

[4] Kohestani K, Wallström J, Dehlfors N, Sponga OM, Månsson M, Josefsson A, Carlsson S, Hellström M, Hugosson J. Performance and inter-observer variability of prostate MRI (PI-rads version 2) outside high-volume centres. Scand J Urol 2019;53(5):304–11. http://dx.doi.org/10.1080/21681805.2019.1675757.

[5] Westphalen AC, McCulloch CE, Anaokar JM, Arora S, Barashi NS, Barentsz JO, Bathala TK, Bittencourt LK, Booker MT, Braxton VG, Carroll PR, Casalino DD, Chang SD, Coakley FV, Dhatt R, Eberhardt SC, Foster BR, Froemming AT, Fütterer JJ, Ganeshan DM, Gertner MR, Mankowski Gettle L, Ghai S, Gupta RT, Hahn ME, Houshyar R, Kim C, Kim CK, Lall C, Margolis DJA, McRae SE, Oto A, Parsons RB, Patel NU, Pinto PA, Polascik TJ, Spilseth B, Starcevich JB, Tammisetti VS, Taneja SS, Turkbey B, Verma S, Ward JF, Warlick CA, Weinberger AR, Yu J, Zagoria RJ, Rosenkrantz AB. Variability of the positive predictive value of PI-RADS for prostate MRI across 26 centers: Experience of the society of abdominal radiology prostate cancer disease-focused panel. Radiology 2020;296(1):76–84. http://dx.doi.org/10.1148/radiol.2020190646.

[6] Mata LA, Retamero JA, Gupta RT, García Figueras R, Luna A. Artificial intelligence–assisted prostate cancer diagnosis: Radiologic-pathologic correlation. RadioGraphics 2021;41(6):1676–97. http://dx.doi.org/10.1148/rg.2021210020.

[7] 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. Med Image Anal 2021;73:102155. http://dx.doi.org/10.1016/j.media.2021.102155.

[8] Hosseinzadeh M, Saha A, Brand P, Slootweg I, de Rooij M, Huisman H. Deep learning–assisted prostate cancer detection on bi-parametric MRI: minimum training data size requirements and effect of prior knowledge. Eur Radiol 2022;32(4):2224–34. http://dx.doi.org/10.1007/s00330-021-08320-y.

[9] Arif M, Schoots IG, Castillo Tovar J, Bangma CH, Krestin GP, Roobol MJ, Niessen W, Veenland JF. Clinically significant prostate cancer detection and segmentation in low-risk patients using a convolutional neural network on multi-parametric MRI. Eur Radiol 2020;30(12):6582–92. http://dx.doi.org/10.1007/s00330-020-07008-z.

[10] Vente Cd, Vos P, Hosseinzadeh M, Pluim J, Veta M. Deep learning regression for prostate cancer detection and grading in bi-parametric MRI. IEEE Trans Biomed Eng 2021;68(2):374–83. http://dx.doi.org/10.1109/TBME.2020.2993528.

[11] 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. Proc IEEE 2021;109(5):820–38. http://dx.doi.org/10.1109/JPROC.2021.3054390.

[12] Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial networks. Commun ACM 2020;63(11):139–44.

[13] Xia X, Pan X, Li N, He X, Ma L, Zhang X, Ding N. GAN-based anomaly detection: A review. Neurocomputing 2022;493:497–535. http://dx.doi.org/10.1016/j.neucom.2021.12.093.

[14] Schlegl T, Seeböck P, Waldstein SM, Langs G, Schmidt-Erfurth U. F-anogan: Fast unsupervised anomaly detection with generative adversarial networks. Med Image Anal 2019;54:30–44. http://dx.doi.org/10.1016/j.media.2019.01.010.

[15] Rahman Siddiquee MM, Shah J, Wu T, Chong C, Schwedt T, Li B. Healthygan: Learning from unannotated medical images to detect anomalies associated with human disease. In: Simulation and synthesis in medical imaging: 7th International workshop, SASHIMI 2022, Held in conjunction with MICCAI 2022, Singapore, September 18, 2022, Proceedings. Springer; 2022, p. 43–54.

[16] Choi Y, Choi M, Kim M, Ha J, Kim S, Choo J. StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. In: 2018 IEEE/CVF Conference on computer vision and pattern recognition. 2018, p. 8789–97. http://dx.doi.org/10.1109/CVPR.2018.00916, ISSN: 2575-7075.

[17] Choi Y, Uh Y, Yoo J, Ha J-W. Stargan v2: Diverse image synthesis for multiple domains. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, 2020, p. 8185–94. http://dx.doi.org/10.1109/CVPR42600.2020.00821.

[18] Siddiquee MMR, Zhou Z, Tajbakhsh N, Feng R, Gotway M, Bengio Y, Liang J. Learning fixed points in generative adversarial networks: From image-to-image translation to disease detection and localization. In: 2019 IEEE/CVF International Conference on Computer Vision. ICCV, 2019, p. 191–200. http://dx.doi.org/10.1109/ICCV.2019.00028.

[19] Wolleb J, Sandk˘˝hler R, Cattin PC. Descargan: Disease-specific anomaly detection with weak supervision. In: Medical image computing and computer assisted intervention – MICCAI 2020. Springer International Publishing; 2020, p. 14–24. http://dx.doi.org/10.1007/978-3-030-59719-1_2.

[20] Armato SG, Huisman H, Drukker K, Hadjiiski L, Kirby JS, Petrick N, Redmond G, Giger ML, Cha K, Mamonov A, Kalpathy-Cramer J, Farahani K. PROSTATEx challenges for computerized classification of prostate lesions from multiparametric magnetic resonance images. J Med Imaging 2018;5(4):044501. http://dx.doi.org/10.1117/1.JMI.5.4.044501.

[21] Sunoqrot MR, Nketiah GA, Selnæ s KM, Bathen TF, Elschot M. Automated reference tissue normalization of T2-weighted MR images of the prostate using object recognition. Magn Reson Mater Phys, Biol Med 2021;34(2):309–21. http://dx.doi.org/10.1007/s10334-020-00871-3.

[22] Sunoqrot MRS, Selnæs KM, Sandsmark E, Nketiah GA, Zavala-Romero O, Stoyanova R, Bathen TF, Elschot M. A quality control system for automated prostate segmentation on T2-weighted MRI. Diagnostics 2020;10(9). http://dx.doi.org/10.3390/diagnostics10090714.

[23] Patsanis A, Sunoqrot MRS, Bathen TF, Elschot M. CROPro: a tool for automated cropping of prostate magnetic resonance images. J Med Imaging 2023;10(2):024004. http://dx.doi.org/10.1117/1.JMI.10.2.024004.

[24] Tustison NJ, Avants BB, Cook PA, Zheng Y, Egan A, Yushkevich PA, Gee JC. N4ITK: Improved N3 bias correction. IEEE Trans Med Imaging 2010;29(6):1310–20. http://dx.doi.org/10.1109/TMI.2010.2046908.

[25] Milletari F, Navab N, Ahmadi S-A. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth international conference on 3D vision (3DV). 2016, p. 565–71. http://dx.doi.org/10.1109/3DV.2016.79.

[26] Isensee F, Jaeger PF, Kohl SA, Petersen J, Maier-Hein KH. Nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods 2021;18(2):203–11. http://dx.doi.org/10.1038/s41592-020-01008-z.

[27] Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V, Courville AC. Improved training of wasserstein GANs. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R, editors. Advances in neural information processing systems. 30, Curran Associates, Inc.; 2017, URL https://proceedings.neurips.cc/paper/2017/file/892c3b1c6dccd52936e27cbd0ff683d6-Paper.pdf.

[28] Saha A, Twilt JJ, Bosma JS, van Ginneken B, Yakar D, Elschot M, Veltman J, Fütterer J, de Rooij M, Huisman H. Artificial intelligence and radiologists at prostate cancer detection in MRI: The PI-CAI challenge (study protocol). 2022, http://dx.doi.org/10.5281/zenodo.6667655.

[29] PI-CAI leaderboard. 2023, URL https://pi-cai.grand-challenge.org/evaluation/open-development-phase/leaderboard/, Accessed on 15 March 2023.

[30] Wang Z, Liu C, Cheng D, Wang L, Yang X, Cheng K-T. Automated detection of clinically significant prostate cancer in mp-MRI images based on an end-to-end deep neural network. IEEE Trans Med Imaging 2018;37(5):1127–39. http://dx.doi.org/10.1109/TMI.2017.2789181.

[31] Cem Birbiri U, Hamidinekoo A, Grall A, Malcolm P, Zwiggelaar R. Investigating the performance of generative adversarial networks for prostate tissue detection and segmentation. J. Imaging 2020;6(9). http://dx.doi.org/10.3390/jimaging6090083.

[32] Singhrao K, Fu J, Parikh NR, Mikaeilian AG, Ruan D, Kishan AU, Lewis JH. A generative adversarial network-based (GAN-based) architecture for automatic fiducial marker detection in prostate MRI-only radiotherapy simulation images. Med Phys 2020;47(12):6405–13. http://dx.doi.org/10.1002/mp.14498.

[33] DeSilvio T, Moroianu S, Bhattacharya I, Seetharaman A, Sonn G, Rusu M. Intensity normalization of prostate MRIs using conditional generative adversarial networks for cancer detection. In: Medical imaging 2021: Computer-aided diagnosis. 11597, SPIE; 2021, p. 121–6. http://dx.doi.org/10.1117/12.2582297.

[34] Nadarajan G, Doyle S. Conditional generative adversarial networks for HE to IF domain transfer: experiments with breast and prostate cancer. In: Medical imaging 2021: Digital pathology. SPIE; 2021, p. 144–54. http://dx.doi.org/10.1117/12.2581098.

[35] Kohl S, Bonekamp D, Schlemmer H-P, Yaqubi K, Hohenfellner M, Hadaschik B, Radtke J-P, Maier-Hein K. Adversarial networks for the detection of aggressive prostate cancer. 2017, arXiv preprint arXiv:1702.08014.

[36] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: International conference on medical image computing and computer-assisted intervention. Springer; 2015, p. 234–41. http://dx.doi.org/10.1007/978-3-319-24574-4.

[37] Park S, Lee KH, Ko B, Kim N. Unsupervised anomaly detection with generative adversarial networks in mammography. Sci Rep 2023;13(1):2925. http://dx.doi.org/10.1038/s41598-023-29521-z.

[38] Schlegl T, Seeböck P, Waldstein SM, Schmidt-Erfurth U, Langs G. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In: Niethammer M, Styner M, Aylward S, Zhu H, Oguz I, Yap P-T, Shen D, editors. Information processing in medical imaging. Cham: Springer International Publishing; 2017, p. 146–57. http://dx.doi.org/10.1007/978-3-319-59050-9_12.

[39] Donahue J, Simonyan K. Large scale adversarial representation learning. Adv Neural Inf Process Syst 2019;32.

[40] Baumgartner CF, Koch LM, Tezcan KC, Ang JX, Konukoglu E. Visual feature attribution using wasserstein gans. In: Proceedings of the IEEE Conference on computer vision and pattern recognition. 2018, p. 8309–19.

[41] Mescheder L, Geiger A, Nowozin S. Which training methods for GANs do actually converge? In: Dy J, Krause A, editors. Proceedings of the 35th International conference on machine learning. Proceedings of machine learning research, 80, PMLR; 2018, p. 3481–90, URL https://proceedings.mlr.press/v80/mescheder18a.html.

[42] Drost F-JH, Osses DF, Nieboer D, Steyerberg EW, Bangma CH, Roobol MJ, Schoots IG. Prostate MRI, with or without MRI-targeted biopsy, and systematic biopsy for detecting prostate cancer. Cochrane Database Syst Rev 2019;4:CD012663. http://dx.doi.org/10.1002/14651858.CD012663.pub2.