cancers

MDPI

Article

Classification of Clinically Significant Prostate Cancer on Multi-Parametric MRI: A Validation Study Comparing Deep Learning and Radiomics

Jose M. Castillo T. [1,†], Muhammad Arif [1,†], Martijn P. A. Starmans [1], Wiro J. Niessen [1,2], Chris H. Bangma [3], Ivo G. Schoots [1] and Jifke F. Veenland [1,4,*]

1 Department of Radiology and Nuclear Medicine, Erasmus MC, 3015 GD Rotterdam, The Netherlands; j.castillotovar@erasmusmc.nl (J.M.C.T.); a.muhammad@erasmusmc.nl (M.A.); m.starmans@erasmusmc.nl (M.P.A.S.); w.niessen@erasmusmc.nl (W.J.N.); i.schoots@erasmusmc.nl (I.G.S.)
2 Faculty of Applied Sciences, Delft University of Technology, Lorentzweg 1, 2628 CJ Delft, The Netherlands
3 Department of Urology, Erasmus MC, 3015 GD Rotterdam, The Netherlands; c.h.bangma@erasmusmc.nl
4 Department of Medical Informatics, Erasmus MC, 3015 GD Rotterdam, The Netherlands
* Correspondence: j.veenland@erasmusmc.nl
† These authors contributed equally to this study.

Simple Summary: Computer-aided diagnosis systems to improve significant prostate cancer (PCa) diagnoses are being reported in the literature. These methods are based on either deep learning or radiomics. However, there is a lack of scientific evidence comparing these methods on the same external validation sets. The aim of our study was to compare the performance of a deep-learning model with the performance of a radiomics model for significant-PCa diagnosis on various cohorts. We collected multiparametric magnetic resonance images and pathology data from four patient cohorts (644 patients in total). One of the cohorts was used to develop a deep-learning model and a radiomics model. Both models were tested on the three remaining cohorts. The comparison shows that whereas the performance of the deep-learning model was higher on the training cohort, the radiomics model outperformed the deep-learning model in all the testing cohorts, making it a more accurate tool with which to detect clinically significant prostate cancer.

Abstract: The computer-aided analysis of prostate multiparametric MRI (mpMRI) could improve significant-prostate-cancer (PCa) detection. Various deep-learning- and radiomics-based methods for significant-PCa segmentation or classification have been reported in the literature. To be able to assess the generalizability of the performance of these methods, using various external data sets is crucial. While both deep-learning and radiomics approaches have been compared based on the same data set of one center, the comparison of the performances of both approaches on various data sets from different centers and different scanners is lacking. The goal of this study was to compare the performance of a deep-learning model with the performance of a radiomics model for the significant-PCa diagnosis of the cohorts of various patients. We included the data from two consecutive patient cohorts from our own center (n = 371 patients), and two external sets of which one was a publicly available patient cohort (n = 195 patients) and the other contained data from patients from two hospitals (n = 79 patients). Using multiparametric MRI (mpMRI), the radiologist tumor delineations and pathology reports were collected for all patients. During training, one of our patient cohorts (n = 271 patients) was used for both the deep-learning- and radiomics-model development, and the three remaining cohorts (n = 374 patients) were kept as unseen test sets. The performances of the models were assessed in terms of their area under the receiver-operating-characteristic curve (AUC). Whereas the internal cross-validation showed a higher AUC for the deep-learning approach, the radiomics model obtained AUCs of 0.88, 0.91 and 0.65 on the independent test sets compared to AUCs of 0.70, 0.73 and 0.44 for the deep-learning model. Our radiomics model that was based on delineated regions resulted in a more accurate tool for significant-PCa classification in the three unseen test sets when compared to a fully automated deep-learning model.

Citation: Castillo T., J.M.; Arif, M.; Starmans, M.P.A.; Niessen, W.J.; Bangma, C.H.; Schoots, I.G.; Veenland, J.F. Classification of Clinically Significant Prostate Cancer on Multi-Parametric MRI: A Validation Study Comparing Deep Learning and Radiomics. *Cancers* **2022**, *14*, 12. https://doi.org/10.3390/cancers14010012

Academic Editors: Claudio Fiorino, Alan Hutson and Song Liu

Received: 2 November 2021
Accepted: 3 December 2021
Published: 21 December 2021

Keywords: prostate carcinoma; clinically significant; radiomics; machine learning; deep learning; comparison; mpMRI; classification; model; prediction; Gleason score

1. Introduction

Prostate cancer (PCa) diagnosis using prostate-specific antigen (PSA) and transperineal ultrasound-guided biopsies combined with multiparametric magnetic resonance imaging (mpMRI) is recommended by the European guidelines [1] and common practice. In the literature, it was shown that the multiparametric MRI (mpMRI)-targeted biopsy can non-invasively and more accurately characterize PCa lesions as compared to the standard systematic transrectal ultrasound-guided (TRUS) biopsy [2–4]. Furthermore, the combination of the MRI-targeted biopsy with the systematic biopsy increases the overall accuracy of PCa diagnoses [5].

The mpMRI interpretation of the prostate was standardized by the Prostate Imaging Reporting and Data System (PI-RADS) v2. However, the visual interpretation by radiologists can still lead to the under-diagnosis of clinically significant PCa and the over-diagnosis of insignificant PCa [6].

A computer-aided quantitative analysis of prostate mpMRI could improve PCa detection and may help in the standardization of mpMRI interpretation [7]. Ultimately, it may contribute to improving the diagnostic chain [8], thereby reducing over- and under-diagnoses in prostate-cancer management [9]. Several methods for significant-PCa segmentation [10–12] or classification [13–15] using deep-learning networks or radiomics approaches have been reported in the literature. Both approaches offer different capabilities, challenges and difficulties [16]. Comparing the performance of both approaches based on the scientific literature can be difficult.

First, because direct comparison of the methods is often not specifically addressed or examined [17]. Second, in general, these developed methods are cohort dependent, and their performance may substantially decrease for independent test data. Differences in the data sets used, such as the selected patient population, the data-set size, the scanner type and manufacturer, the MRI protocols, the image quality, the radiologist's delineations, the pathology material (biopsies or radical prostatectomies), and the pathology reports may all substantially influence the model performance [18]. Several approaches (deep learning and radiomics) have been compared based on the same data set [19]; however, the comparison of the performances of both approaches on multiple data sets obtained at different institutions is lacking.

Therefore, the aim of this study was to perform a comparison study of a deep-learning method for significant-PCa segmentation and a radiomics-based significant-PCa-classification method on various patient data sets.

2. Materials and Methods

In this study, we compared the performance of a recently developed deep-learning network for significant-PCa segmentation [10] with a radiomics approach based on regions of interest (ROIs) that were delineated by a radiologist. Both models were trained on the same data set. Three data sets, representing various scanners, various patient populations from different hospitals, and two different ground-truth methods were used to compare the performances of both methods. PCa with an ISUP grade $\geq$ 2 (Gleason Score (GS) 3 + 4 and higher) was classified as significant [20]

2.1. Patient Data Sets

This study included four patient cohorts: an Active-Surveillance cohort (AS), a cohort of men with previous negative systematic biopsies (Prodrome), the cohort included in the Prostate Cancer Molecular Medicine study (PCMM) and the ProstateX challenge cohort, as illustrated in Figure 1. The cohort of men participating in the Active-Surveillance study and

the cohort of the previous-negative-biopsy men were acquired from prospective studies of the Erasmus MC in Rotterdam, the Netherlands [21]. The usage of the data for this study was HIPAA compliant and written informed consent with a guarantee of confidentiality was obtained from the participants. The ProstateX data were publicly available and were provided for the SPIE-AAPM-NCI ProstateX challenge [19,22,23]. The PCMM data set was acquired retrospectively from two external healthcare centers in the Netherlands [18], the data usage of this study was approved by the medical ethics review committee of the Erasmus MC under the number NL32105.078.10.

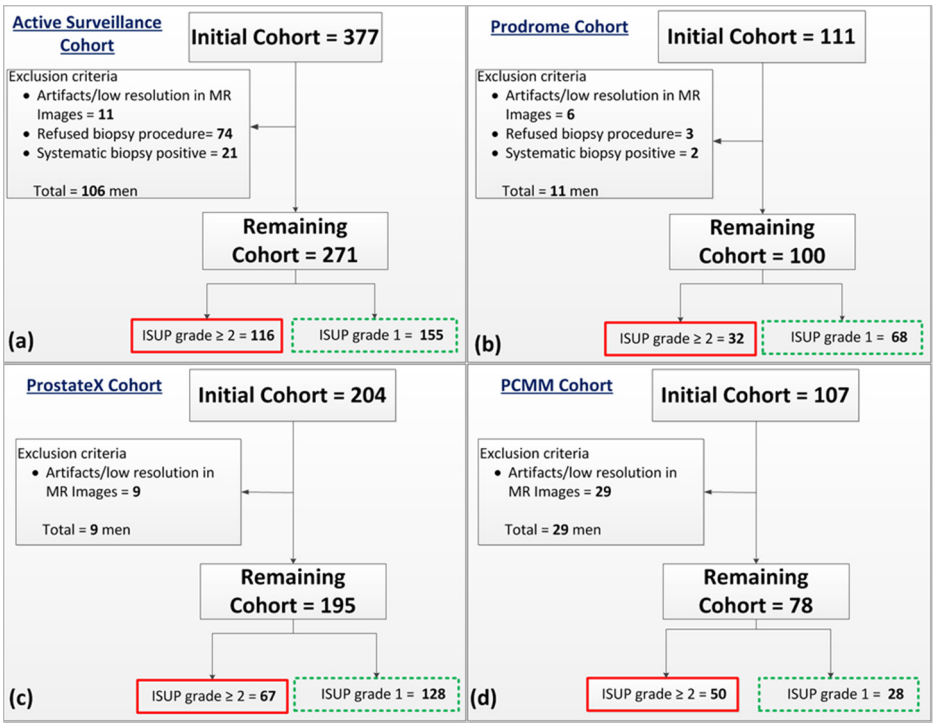

Figure 1. Flow diagram of patient exclusion and inclusion of the four cohorts used in this study: (**a**) Active Surveillance, (**b**) Prodrome, (**c**) ProstateX, and (**d**) PCMM. **ISUP**: International Society of Urological Pathology.

In total, 644 patients from all four cohorts were included in this study. The selected patients were divided into a training set (Active-Surveillance cohort) and test sets (Prodrome, ProstateX and PCMM cohorts). The general patient characteristics (prostate volume, age and prostate-specific antigen) of the sets are listed in Table 1. Prostate volume was measured on T2-weighted (T2w) images using the ellipsoid formulation [24].

For the Active Surveillance cohort, initially, 377 consecutive patients with low-risk PCa (defined as International Society of Urological Pathology "ISUP" grade 1) were prospectively enrolled in our in-house database from 2016 to 2019. All participants received a multi-parametric MRI and targeted biopsies of visible, suspicious (PI-RADS $\geq$ 3) lesions at the baseline (3 months after diagnosis). A detailed description of the data set has been published [25]. Patients who refused, who had no biopsy procedure, or whose MRI scans had artifacts were excluded from the study. Since the systematic-biopsy locations were not available, patients in whom significant PCa was found based only on systematic biopsies were also excluded. The remaining cohort (n = 271) was divided into two groups based on the pathology findings (Figure 1a). Some of the patients (n = 55) with an ISUP grade = 1 did not have an identifiable lesion on MRI, therefore no targeted biopsy was performed, and the systematic biopsy was negative.

The Prodrome cohort contained 111 consecutive patients with prior negative biopsies who were prospectively enrolled in our in-house database from 2017 to 2019. For each patient, mpMRI with blinded systematic biopsies and MRI-targeted biopsies of visible, sus-

picious (PI-RADS $\geq$ 3) lesions were performed. Patients who refused the biopsy procedure, or whose MRI scans had artifacts were excluded from the study. Patients who were found to have significant PCa based only on systematic biopsies were also excluded since the exact locations of the systematic biopsies was unknown. The remaining cohort (n = 100) was divided into two groups based on the pathology findings (Figure 1b).

Table 1. Clinical characteristics of the patients of the four cohorts (Active Surveillance, Prodrome, ProstateX and PCMM) included in this study. Tumorr volume values are presented as median (interquartile range). **PZ**: peripheral zone. **TZ**: transition zone. **ISUP**: International Society of Urological Pathology. **GSA:** Gleason Score. **PSA**: Prostate Specific Antigen. **NA**: not available. (*) The ISUP grade per lesion was not available for ProstateX challenge, the ground truth provided for this set indicated whether the lesion had ISUP grade $\geq$ 1.

	Training Cohort		**Testing Cohort**	
Patient Cohort	**Active Surveillance**	**Prodrome**	**ProstateX ***	**PCMM**
Total Number of patients	271	100	195	78
Patients with a lesion ISUP grade = 1	155	68	128	28
Patients with a lesion ISUP grade $\geq$ 2	116	32	67	50
Total number of lesions	233	104	328	156
ISUP grade 1	100	52	254	77
ISUP grade $\geq$ 2	133	52	74	79
ISUP grade 2	124	45	NA	68
ISUP grade 3	3	6	NA	8
ISUP grade 4 & 5	6	1	NA	3
Lesions in PZ	150	60	191	104
Lesions in TZ	33	41	82	49
Lesions in other zones (central, anterior stroma)	38	3	55	3
Lesion volume (mL)	0.3(0.2–0.8)	0.61 (0.3–1.0)	1.42 (1.4–3.2)	0.80 (0.2–1.1)
Prostate Volume(mL)	43.1 (30.5–76.2)	50. (33–67)	NA	NA
Age (year)	67 $\pm$ 7	68 $\pm$ 4	NA	NA
PSA(mean $\pm$ std ng/mL)	10 $\pm$ 6	12 $\pm$ 4	NA	9 $\pm$ 7

The ProstateX cohort comprises 204 patients and included suspicious-lesion coordinates and targeted-biopsy-based histopathological findings. Nine patients were excluded due to image artifacts, registration and missing DWI images at b800. The remaining cohort (n = 195) was divided into two groups based on the pathology findings (Figure 1c).

The PCMM cohort consists of 107 patients who were enrolled from 2011 to 2014 and included lesion segmentations based on delineations made by a pathologist on prostatectomy specimen photos. The MR images of these patients were correlated with the prostatectomy

photos using manual registration [18]. Twenty-nine patients were excluded due to image artifacts or low resolution (Figure 1d).

For each patient, two MRI sequences, axial T2w and diffusion-weighted images (DWIs) with their apparent-diffusion-coefficient maps (ADC) were selected. For the AS, Prodrome and ProstateX cohorts, the histopathology data from the MRI-targeted biopsies were considered as reference standards. In the case of the PCMM data set, the ground truth was obtained from pathology reports after the prostatectomy.

2.2. MR Imaging and Pre-Processing

For the Active Surveillance and Prodrome cohorts, the MRI scans were performed on a 3T system (Discovery MR750, General Electric Healthcare, Chicago, IL USA), according to the PI-RADS v2 guidelines [26]. The T2-weighted imaging (T2w) diffusion-weighted imaging (DWI, b-values 50, 400 and 800) were acquired using a 32-channel pelvic phased-array coil with $0.371 \times 0.371 \times 3.3$ mm^3 resolution. Apparent-diffusion-coefficient (ADC) maps were constructed using scanner software. All MR images were reviewed by a urogenital radiologist with over 6 years of prostate MRI experience. Individual lesions with a PI-RADS score ≥ 3 were defined as suspicious and delineated in the T2w images. The MRI-and-transrectal-ultrasound (TRUS)-fusion technique was used (UroStation™, Koelis, France) to perform targeted biopsies with a maximum of 4 cores under ultrasound guidance. One expert uropathologist reviewed the biopsy specimens according to the ISUP 2014 modified Gleason Score [27]. For every patient, binary masks were generated for our experiments based on the delineations on the T2w images with biopsy-proven significant PCa (ISUP grade ≥ 2). The DWIs with ADC values were manually and rigidly co-registered to the T2w images for every patient. Furthermore, the 3D images were cropped to the whole prostate region of interest with dimensions $128 \times 192 \times 24$ voxels along the x, y and z directions.

For the ProstateX cohort, MRI scans were acquired on one of two 3T MR systems (MAGNETRON Trio and Skyra, Siemens Medical Systems, Erlangen, Germany). The MRI protocol included T2w images acquired with 0.5 mm 2D resolution and 3.6 mm slice thickness, DWI (b-values 40, 400 and 800) series acquired using single-shot-echo-planer imaging with 2 mm 2D resolution and 3.6 mm slice thickness, and ADC maps constructed using scanner software. Since the voxel sizes in the data from the ProstateX cohort varied, all images were resampled to a uniform voxel spacing of $0.371 \times 0.371 \times 3.3$ mm^3, which was the same as the Active Surveillance and Prodrome cohorts. The DWIs with ADC values were manually and rigidly co-registered to the T2w images for every patient and cropped to the whole prostate region of interest with dimensions $128 \times 192 \times 24$ voxels along the x, y and z directions. For every patient, significant PCa (ISUP grade 00) was delineated based on each given lesion's coordinates by one investigator with 6 months experience under supervision and in consensus with a urogenital radiologist with over 6 years of prostate MRI experience.

The MRI scans of the PCMM data set were obtained from three different 3T MR systems. Two were Siemens Medical system models (MAGNETRON Trio and Skyra) with the same characteristics as the models described for the ProstateX cohort. The third scanner model was from Philips (Achieva). The MRI protocol included T2w images with 0.27 mm 2D resolution and 3.00 mm slice thickness, DWIs (b-values 100, 300, 450, 600, 750) with 1.03 mm 2D resolution and 3 mm slice thickness, and ADC maps obtained from the scanner. The images from the Philips MRI were acquired using an endorectal coil. All the original images were sampled to match the voxel spacing of the Active Surveillance cohort and cropped from the central region of the image having dimensions of $128 \times 192 \times 24$ voxels along the x, y and z directions. The lesion delineations were based on the manual registration with the pathology specimens delineated by an expert pathologist.

All the T2-weighted images in this study were pre-processed using z-scoring for pixel-intensity normalization. The MRI scans from the patient populations included in this study were from 2013 to 2019. Therefore, high b-values images (>1400) were acquired for only a

portion of the patient cohorts. Since we did not want to mix the artificially extrapolated and the acquired high b-value images, we chose to focus on the raw data.

2.3. Development of the Models

For this experiment, both models were trained on images from the Active Surveillance cohort. Whereas the deep-learning network was trained to identify significant PCa in images of the whole prostate, the radiomics model was trained to identify significant PCa based on the ROIs that were delineated by the radiologist.

2.3.1. Deep-Learning Network

A fully convolutional neural network (CNN), as described in a recently published previous work was used [10]. Three 3D MR images (T2w, DWI (b-value closest to 800) with corresponding ADC map) were used as inputs for the CNN. Each sequence was considered as a separate input channel. The network contained twelve single 3D convolution layers. Batch normalization (BN) was added after each 3D convolution to improve the convergence speed during training. In the final layer, a 3D convolution having $1 \times 1 \times 1$ kernel size was used to map the computed features to the predicted significant-PCa segmentation.

The network was trained in Python (version 3.5.3) using Keras (version 2.0.2) with Tensor Flow (version 1.0.1) as the backend. During training and prediction, a GeForce GTX TITAN Xp GPU (NVIDIA Corporation) was used. The loss function during training was the binary cross-entropy metric and was optimized using an Adam optimizer with a learning rate of 0.01. For better generalization, data augmentation was implemented in all images during training, which included rotation (0–50, along *x*,*y*,*z*-axes) and shearing (along *x*,*y*,*z*-axes) with rigid transformation and 50% probability. The output of the trained network was the binary segmentation (voxel values from 0 to 1) of clinically significant PCa lesions.

2.3.2. Radiomics Model Development

The radiomics model that was used to classify ROIs as significant versus insignificant PCa was developed with the open-source Workflow for Optimal Radiomics Classification (WORC) package for python [18,28]. Similar to the deep-learning network, the inputs for WORC were the T2w MRIs, the DWIs (b-value closest to 800) and ADC maps. Additionally, the ROIs of the lesion delineations were provided. Within the ROIs, 564 radiomic features quantifying intensity, shape, texture and orientation were extracted from the two MR images and the ADC map.

WORC performs an automated search amongst a wide variety of algorithms and their corresponding parameters in order to determine the optimal combination that maximizes the prediction performance on the training set. During each iteration, WORC generates 1000 workflows by using different combinations of methods and parameters. The internal evaluation of the model was performed by using a 100x random-split cross-validation.

At the end of each cross-validation, the 100 best-performing solutions were combined in an ensemble as a single classification model by averaging their probability predictions. The details regarding the feature computation, model selection and optimization can be found in Appendices A and B, respectively.

2.4. Methods Performance Comparison

The performances of both methods were evaluated using receiver-operating-characteristic (ROC) curves, accuracy and F1-scores that were computed at the patient level both by internal cross-validation and external validation. Internal validation was performed using a 3-fold random cross-validation, which randomly split the training set with ISUP $\geq$2 patients into three separate folds (fold 1 contained 39 patients, fold 2 contained 38 patients and fold 3 contained 39 patients). In each iteration, the models were trained on two of the three folds, and were evaluated on the third fold and an independent test set containing 155 patients with ISUP grades $\leq$ 1, more details can be found in [10]. These splits were

kept the same for the training of both models. The average over these three iterations is reported.

To compare the performance with clinical practice, the sensitivity and specificity of the visual scoring by the experienced radiologist is indicated (when available).

3. Results

3.1. Internal Cross-Validation

In Figure 2, the ROC curve of the deep-learning model (blue line) and the radiomics model (orange line) as computed by the internal cross-validation of the training set can be seen. Table 2 depicts the accuracy, sensitivity, specificity and F1-score obtained from this experiment. Overall, it can be seen that deep learning performed better on the training set compared to the radiomics model.

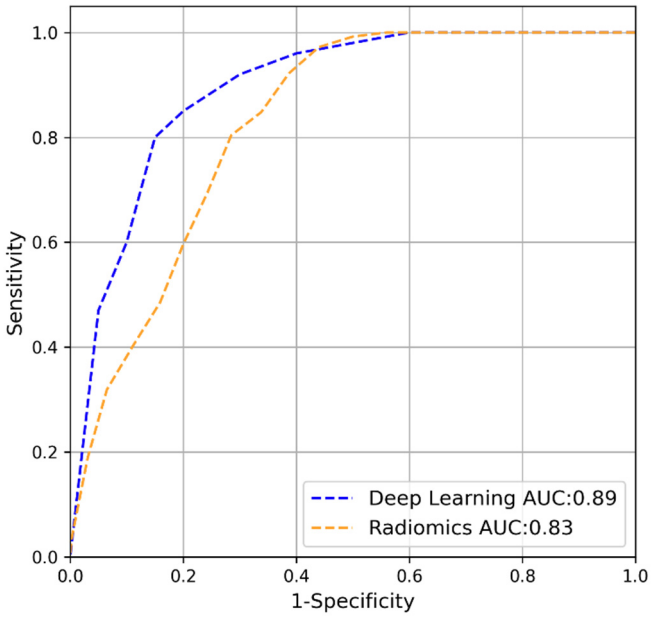

Figure 2. ROC curves of the deep-learning (blue) and radiomics (orange) models on the interval validation on Active Surveillance.

Table 2. Deep-learning- and radiomics-model performances on the training set (Active Surveillance) and on the external sets (Prodrome, ProstateX and PCMM). **DL**: deep learning. **AUC**: Area under the curve.

	Active Surveillance		**Prodrome**		**ProstateX**		**PCMM**	
Metrics	**DL**	**Radiomics**	**DL**	**Radiomics**	**DL**	**Radiomics**	**DL**	**Radiomics**
AUC	0.89	0.83	0.70	0.88	0.73	0.91	0.44	0.65
Accuracy	0.76	0.63	0.58	0.78	0.71	0.85	0.52	0.55
Sensitivity	0.85	1.00	0.72	1.00	0.70	0.72	0.70	0.44
Specificity	0.52	0.54	0.51	0.68	0.71	0.94	0.18	0.71
F1-score	0.74	0.66	0.52	0.78	0.65	0.85	0.66	0.55

3.2. External-Validation

Figure 3a–c depict the ROC curves of the deep-learning and radiomics models in the Prodrome, ProstateX and PCMM cohorts. Furthermore, the rest of the evaluation metrics can be found in Table 2. In the Prodrome cohort, the deep-learning model obtained an AUC of 0.70 versus 0.88 for the radiomics model, whereas in the ProstateX and PCMM

cohorts the AUCs were 0.73 and 0.44 for the deep-learning model and 0.91 and 0.65 for the radiomics model, respectively.

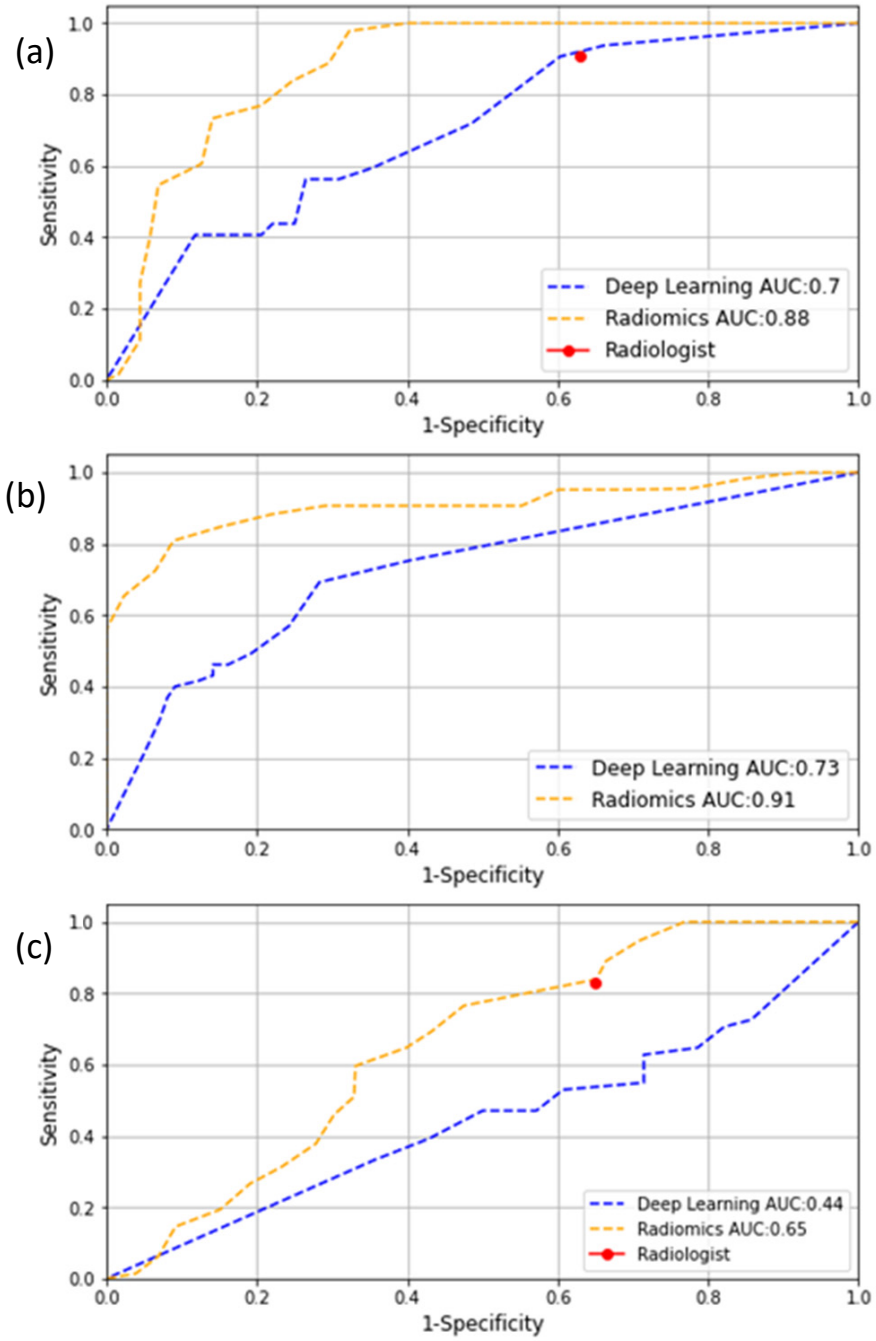

Figure 3. ROC curves of the deep-learning (blue) and radiomics (orange) models when evaluated on the test sets: (**a**) Prodrome, (**b**) ProstateX and (**c**) PCMM.

To illustrate the deep-learning-segmentation method, examples (true positive, false negative and false positive) are shown in Figure 4. In the true-positive example (Figure 4A) the network has successfully segmented both significant-PCa lesions as delineated by the radiologist and proven by targeted biopsy. In some cases, PCa segmentation was unsuccessful, leading to a false negative (Figure 4B). In the false-positive example (Figure 4C), the deep-learning network segments a lesion in the anterior stroma zone; however, the targeted biopsy found insignificant PCa (ISUP grade 1, GS 3 + 3 = 6).

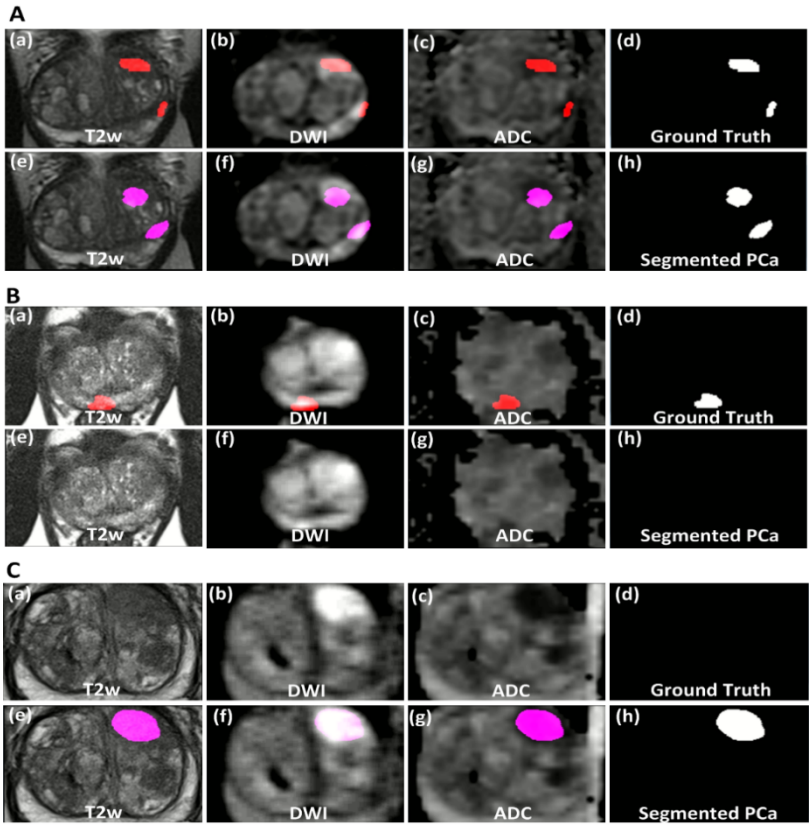

Figure 4. All images show the same axial slice as 2D view of mpMR images (**a**,**e** T2w images; **b**,**f** DWI b800; **c**,**g** ADC map) of the prostate with the reference ground truth (**d**) and the segmented PCa lesion by the deep-learning model (**h**) (**A**) Example of a true positive case (PSA = 17.6; prostate volume = 46 cc; ISUP grade = 2 (up) and 3 (down)). The ground truth is shown in overlay (red) as delineated by the radiologist and proven by targeted biopsy as significant PCa. The segmented significant-PCa lesion by the deep-learning model is shown in overlay (pink). (**B**) Example of a false negative case (ISUP grade = 2). The ground truth is shown in overlay (red) as delineated by the radiologist and proven by targeted biopsy as significant PCa. The deep-learning model has not segmented any PCa lesion. (**C**) Example of a false positive case (ISUP grade = 1). The images show no delineation due to the absence of significant PCa, the region delineated by the radiologist (not shown) proved by targeted biopsy as insignificant PCa. The lesion segmented incorrectly by the deep-learning model is shown in overlay (pink).

4. Discussion

The detection of significant PCa utilizing CAD systems that were based either on deep learning or radiomics has generated many papers in the scientific literature [29–31]. Both approaches offer different capabilities, challenges and difficulties [16]. Whereas both approaches have been compared based on the same data set [19], the direct comparison of these approaches when tested on the same unseen data is lacking in the current scientific literature. To our knowledge, this is the first study comparing a fully automated deep-learning model for significant-PCa segmentation with a radiomics approach based on segmentations that were performed by a radiologist. This study used a large patient cohort compared to the average sizes used in similar studies [29] (avg = 127 patients), and takes into account data from various patient risk groups, various MRI scanners in various hospitals, and ground truths obtained from both biopsies and prostatectomies.

The deep-learning and radiomics models obtained similar ROC curves when validated by internal cross-validation on the training set. However, when validated by external sets, the comparison showed that the radiomics model had higher AUC values than the automated deep-learning model in all data sets.

Nevertheless, when comparing, some considerations should be mentioned. First, there was a tissue-volume difference; the deep-learning model was trained to perform the segmentation of significant prostate cancer in the whole prostate, whereas the radiomics model classified the ROIs that were delineated by the radiologist. For the delineated ROIs, the ground truth was known, but when the deep-learning model segmented a lesion where no biopsy was taken, this lesion was considered to be a false positive, while in fact we did not know the ground truth.

Second, the better performance of the radiomics model could also be attributed to the set of features used by the model, which was a generic set of features that have been used for other applications [15,32]. Furthermore, the deep-learning model automatically designed features specific to the training data and it is possible that this set of features was too specific, causing the generalization issues across the testing sets. Third, the radiomics model used an ensemble of many machine-learning models. Utilizing ensembles might have granted the radiomics model a generalization ability that the deep-learning model lacked by being a model based on a single CNN [33]. This difficulty to generalize was evident in the test sets that used a different ground truth, different scanners and older acquisition protocols. Fourth, the deep-learning approach required a larger number of training examples than the radiomics approach [34]. Therefore, the limited number of patients available for training could have reduced the performance of the deep-learning model compared to the radiomics performance.

An interesting aspect of the better performance of the radiomics model is revealed when comparing it to the performance of a radiologist. For the Prodrome and PCMM cohorts, the combination of the radiologist's delineations with the radiomics model allowed for the preservation of the radiologist's high sensitivity while increasing the specificity for clinically significant PCa lesions. Adding a radiomics model to the workflow could mean a more accurate patient selection for the biopsy procedures and a reduction in overdiagnoses.

In the ProstateX challenge, several methods were compared for clinically significant-PCa classification [19]. However, the details regarding the training setup and methods were not described, thereby prohibiting a comparison with our results. The results obtained from other tumors [35–37] that were all evaluated on their own single internal validation set show that deep-learning methods outperformed radiomics, which corresponds to our findings. One study [38] that performed external validation found that radiomics obtained better results than deep learning, which also corresponds to our results.

There are some limitations to this study. For instance, the clinical variables such as age, race, and PSA level were not available for some of the patient cohorts and therefore could not be included in the models. This information is frequently taken into account by clinicians as risks factors for aggressive prostate cancer [39]. Secondly, co-existing benign prostatic diseases that can mimic PCa were not taken into account in our experiments, since no ground truth was available for these diseases. Furthermore, part of our data is related to biopsy results, which reflect the diagnostic outcome. However, biopsy results can be downgraded or upgraded in radical prostatectomy specimens, which may hamper the interpretation of the results.

Another limitation was that the delineations in this study were carried out by a single clinician; therefore, we were not able to study the effect on the feature computation or the performance with multiple evaluators. Lastly, this was a retrospective study, which precluded the comparison of the models while they were being used by a clinician.

PCa diagnosis can be improved using CAD systems that are based on radiomics or deep learning. However, there is no evidence of these methods being used in clinics in prospective studies [29]. Hence, future research should focus on studying the impact of the radiologist's decision to incorporate either deep-learning or radiomics models in the workflow.

5. Conclusions

Both deep-learning and radiomics methods provide capabilities to support PCa diagnoses. In this study, a radiomics model that was trained on segmentations provided by a radiologist resulted in a more accurate and generalizable tool for significant-PCa classification compared to a fully automated deep-learning model for significant-PCa segmentation.

Author Contributions: Conceptualization, J.M.C.T., M.A., I.G.S. and J.F.V.; data curation, J.M.C.T., M.A., C.H.B., I.G.S. and J.F.V.; formal analysis, J.M.C.T., M.A., M.P.A.S., W.J.N., C.H.B., I.G.S. and J.F.V.; funding acquisition, W.J.N. and J.F.V.; investigation, J.M.C.T., M.A., M.P.A.S., W.J.N., I.G.S. and J.F.V.; methodology, J.M.C.T., M.A., W.J.N. and J.F.V.; project administration, W.J.N., I.G.S. and J.F.V.; resources, W.J.N., C.H.B., I.G.S. and J.F.V.; software, J.M.C.T., M.A. and M.P.A.S.; supervision, W.J.N., I.G.S. and J.F.V.; validation, J.M.C.T. and M.A.; visualization, J.M.C.T. and M.A.; writing—original draft, J.M.C.T. and M.A.; writing—review & editing, M.P.A.S., W.J.N., C.H.B., I.G.S. and J.F.V. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by a grant of STW (15173) and KWF-NWO (14932), The Netherlands.

Institutional Review Board Statement: The data usage of this study was approved by the medical ethics review committee of Erasmus MC under the number NL32105.078.10. The ProstateX data is publicly available and was provided for the SPIE-AAPM-NCI ProstateX challenge.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=23691656 (accessed on 18 December 2021).

Acknowledgments: The Titan Xp's used for this research were donated by the NVIDIA Corporation.

Conflicts of Interest: Wiro Niessen is founder, shareholder and scientific lead of Quantib BV. This research grant was funded by a grant of STW(1573) and KWF-NOW (14932), The Netherlands.

Appendix A Radiomics Features Extraction

This Appendix A is similar to [32,40], but details relevant for the current study are highlighted. A total of 540 radiomics features were used in this study. All features were extracted using Workflow for Optimal Radiomics Classification (WORC) [41], which internally uses the PREDICT [42] feature extraction toolboxes. For details on the mathematical formulation of the features, we refer the reader to [43]. More details on the extracted features can be found in the documentation of the respective toolboxes, mainly the WORC documentation.

For CT scans, the images are not normalized by default as the scans already have a fixed unit and scale (i.e., Hounsfield), contrary to MRI. The images were not resampled, as this would result in interpolation errors. The code to extract the features has been open-source published.

The features can be divided into several groups. Thirteen intensity features were extracted using the histogram of all intensity values within the ROIs and included several first-order statistics such as the mean, standard deviation and kurtosis. Thirty-five shape features were extracted based only on the ROI, i.e., not using the image, and included shape descriptions such as the volume, compactness and circular variance. These describe the morphological properties of the lesion. Nine orientation features were used, describing the orientation of the ROI, i.e., not using the image. Lastly, 483 texture features were extracted using Gabor filters (144 features), Laplacian of Gaussian filters (36 features), vessel (i.e., tubular structures) filters (36 features) [44], the Gray Level Co-occurrence Matrix (144 features) [43], the Gray Level Size Zone Matrix (16 features) [43], the Gray Level Run Length Matrix (16 features) [43], the Gray Level Dependence Matrix (14 features) [43], the Neighborhood Grey Tone Difference Matrix (5 features) [43] Local Binary Patterns (18 features) [45], and local phase filters (36 features) [46]. These features describe more complex patterns within the lesion, such as heterogeneity, occurrence of blob-like structures, and presence of line patterns.

Most of the texture features include the parameters to be set for the extraction. Beforehand the values of the parameters that will result in features with the highest discriminative power for the classification at hand (e.g., high grade vs. low grade) are not known. Including these parameters in the workflow optimization, see Appendix B, would lead to repeated computation of the features, resulting in a redundant decrease in computation time. Therefore, alternatively, these features are extracted at a range of parameters as is default in WORC. The hypothesis is that the features with high discriminative power will be selected by the feature-selection methods and/or the machine learning methods as described in Appendix B.

The data set used in this study is heterogeneous in terms of acquisition protocols. Especially the variations in slice may cause feature values to be dependent on the acquisition protocol. Hence, extracting robust 3D features may be hampered by these variations, especially for low resolutions. To overcome this issue, all features were extracted per 2D axial slice and aggregated over all slices, which is default in WORC. Afterwards, several first-order statistics over the feature distributions were evaluated and used in the machine learning approach.

Appendix B Adaptive Workflow Optimization for Automatic Decision-Model Creation

This appendix is similar [18], but details relevant to the current study are highlighted. The Workflow for Optimal Radiomics Classification (WORC) toolbox [41] makes use of adaptive algorithm optimization to create the optimal performing workflow from a variety of methods. WORC defines a workflow as a sequential combination of algorithms and their respective parameters. To create a workflow, WORC includes algorithms to perform feature scaling, feature imputation, feature selection, oversampling, and machine learning. If used, as some of these steps are optional as described below, these methods are performed in the same order as described in this appendix. More details can be found in the WORC documentation. The code to use WORC for creating the differential diagnosis and molecular analysis decision models in this specific study has been published open-source.

Feature scaling was performed to make all features have the same scale, as otherwise the machine learning methods may focus only on those features with large values. This was done through z-scoring, i.e., subtracting the mean value followed by division by the standard deviation, for each individual feature. In this way, all features had a mean of zero and a variance of one. A robust version of z-scoring was used, in which outliers, i.e., values below the 5th percentile or above the 95th percentile, were excluded from computing the mean and variance.

When a feature could be computed, e.g., a lesion is too small for specific feature to be extracted or a division by zero occurs, feature imputation was used to estimate replacement values for the missing values. Strategies for imputation included (1) the mean; (2) the median; (3) the most frequent value; and (4) a nearest neighbor approach.

Feature selection was performed to eliminate features which were not useful to distinguish between the classes. These included; (1) a variance threshold, in which features with a low variance (<0.01) are removed. This method was always used, as this serves as a feature sanity check with almost zero risk of removing relevant features; (2) optionally, a group-wise search, in which specific groups of features (i.e., intensity, shape, and the subgroups of texture features as defined in Appendix B are selected or deleted. To this end, each feature group had an on/off variable which is randomly activated or deactivated, which were all included as hyperparameters in the optimization; (3) optionally, individual feature selection through univariate testing. To this end, for each feature, a Mann-Whitney U test was performed to test for significant differences in distribution between the labels. Afterwards, only features with a p-value above a certain threshold were selected. A Mann-Whitney U test was chosen as features may not be normally distributed and the samples (i.e., patients) were independent; and (4) optionally, principal component analysis (PCA), in which either only those linear combinations of features were kept which explained 95%

of the variance in the features or a limited number of components (between 10–50). These feature-selection methods may be combined by WORC, but only in the mentioned order.

Various resampling strategies can optionally be used, which can be used to overcome class imbalances and reduce overfitting on specific training samples. These included various methods from the imbalanced-learn toolbox [47]; random over-sampling, random under-sampling, near-miss resampling, the neighborhood cleaning rule, ADASYN, and SMOTE (regular, borderline, Tomek and the edited nearest neighbors).

Lastly, machine learning methods were used to determine a decision rule to distinguish the classes. These included: (1) logistic regression; (2) support vector machines; (3) random forests; (4) naive Bayes; and (5) linear and quadratic discriminant analysis.

Most of the included methods require specific settings or parameters to be set, which may have a large impact on the performance. As these parameters have to be determined before executing the workflow, these are so-called "hyperparameters". In WORC, all parameters of all mentioned methods are treated as hyperparameters, since they may all influence the decision-model creation. WORC simultaneously estimates which combination of algorithms and hyperparameters performs best. A comprehensive overview of all parameters is provided in the WORC documentation.

By default, in WORC, the performance is evaluated in a 100x random-split train-test cross-validation. In the training phase, a total of 100,000 pseudo-randomly generated workflows is created. These workflows are evaluated in a 5x random-split cross-validation on the training data set, using 80% of the data for actual training and 20% for validation of the performance. All described methods are fit on the training data sets, and only tested on the validation data sets. The workflows are ranked from best to worst based on their mean performance on the validation sets using the F1-score, which is the harmonic average of precision and recall. Due to the large number of workflows that is executed, there is a chance that the best-performing workflow is overfitting, i.e., looking at too much detail or even noise in the training data set. Hence, to create a more robust model and boost performance, WORC combines the 100 best-performing workflows into a single decision model, which is known as ensembling. These 100 best-performing workflows are re-trained using the entire training data set, and only tested on the test data sets. The ensemble is created through averaging of the probabilities, i.e., the chance of a lesion with a high grade or low grade, of these 100 workflows. A full experiment consists of executing 50 million workflows (100,000 pseudo-randomly generated workflows, times a 5x train-validation cross-validation times 100x train-test cross-validation), which can be parallelized.

References

1. Heidenreich, A.; Bastian, P.J.; Bellmunt, J.; Bolla, M.; Joniau, S.; Van Der Kwast, T.; Mason, M.; Matveev, V.; Wiegel, T.; Zattoni, F.; et al. EAU Guidelines on Prostate Cancer. Part 1: Screening, Diagnosis, and Local Treatment with Curative Intent—Update 2013. *Eur. Urol.* **2014**, *65*, 124–137. [CrossRef]
2. Drost, F.-J.H.; Osses, D.F.; Nieboer, D.; Steyerberg, E.W.; Bangma, C.H.; Roobol, M.J.; Schoots, I.G. Prostate MRI, with or without MRI-Targeted Biopsy, and Systematic Biopsy for Detecting Prostate Cancer. *Cochrane Database Syst. Rev.* **2019**, *4*, CD012663. [CrossRef]
3. Schoots, I.G.; Roobol, M.J.; Nieboer, D.; Bangma, C.H.; Steyerberg, E.W.; Hunink, M.G.M. Magnetic Resonance Imaging–Targeted Biopsy May Enhance the Diagnostic Accuracy of Significant Prostate Cancer Detection Compared to Standard Transrectal Ultrasound-Guided Biopsy: A Systematic Review and Meta-Analysis. *Eur. Urol.* **2015**, *68*, 438–450. [CrossRef] [PubMed]
4. Ahmed, H.U.; El-Shater Bosaily, A.; Brown, L.C.; Gabe, R.; Kaplan, R.; Parmar, M.K.; Collaco-Moraes, Y.; Ward, K.; Hindley, R.G.; Freeman, A.; et al. Diagnostic Accuracy of Multi-Parametric MRI and TRUS Biopsy in Prostate Cancer (PROMIS): A Paired Validating Confirmatory Study. *Lancet* **2017**, *389*, 815–822. [CrossRef]
5. Ahdoot, M.; Wilbur, A.R.; Reese, S.E.; Lebastchi, A.H.; Mehralivand, S.; Gomella, P.T.; Bloom, J.; Gurram, S.; Siddiqui, M.; Pinsky, P.; et al. MRI-Targeted, Systematic, and Combined Biopsy for Prostate Cancer Diagnosis. *N. Engl. J. Med.* **2020**, *382*, 917–928. [CrossRef] [PubMed]
6. Valerio, M.; Donaldson, I.; Emberton, M.; Ehdaie, B.; Hadaschik, B.A.; Marks, L.S.; Mozer, P.; Rastinehad, A.R.; Ahmed, H.U. Detection of Clinically Significant Prostate Cancer Using Magnetic Resonance Imaging-Ultrasound Fusion Targeted Biopsy: A Systematic Review. *Eur. Urol.* **2015**, *68*, 8–19. [CrossRef]

7. Penzkofer, T.; Padhani, A.R.; Turkbey, B.; Haider, M.A.; Huisman, H.; Walz, J.; Salomon, G.; Schoots, I.G.; Richenberg, J.; Villeirs, G.; et al. ESUR/ESUI Position Paper: Developing Artificial Intelligence for Precision Diagnosis of Prostate Cancer Using Magnetic Resonance Imaging. *Eur. Radiol.* **2021**, *31*, 9567–9578. [CrossRef]
8. Rouvière, O.; Schoots, I.G.; Mottet, N. Multiparametric Magnetic Resonance Imaging Before Prostate Biopsy: A Chain Is Only as Strong as Its Weakest Link. *Eur. Urol.* **2019**, *75*, 889–890. [CrossRef]
9. Padhani, A.R.; Barentsz, J.; Villeirs, G.; Rosenkrantz, A.B.; Margolis, D.J.; Turkbey, B.; Thoeny, H.C.; Cornud, F.; Haider, M.A.; Macura, K.J.; et al. PI-RADS Steering Committee: The PI-RADS Multiparametric MRI and MRI-Directed Biopsy Pathway. *Radiology* **2019**, *292*, 464–474. [CrossRef]
10. Arif, M.; Schoots, I.G.; Castillo Tovar, J.; Bangma, C.H.; Krestin, G.P.; Roobol, M.J.; Niessen, W.; Veenland, J.F. Clinically Significant Prostate Cancer Detection and Segmentation in Low-Risk Patients Using a Convolutional Neural Network on Multi-Parametric MRI. *Eur. Radiol.* **2020**, *30*, 6582–6592. [CrossRef]
11. Dai, Z.; Carver, E.; Liu, C.; Lee, J.; Feldman, A.; Zong, W.; Pantelic, M.; Elshaikh, M.; Wen, N. Segmentation of the Prostatic Gland and the Intraprostatic Lesions on Multiparametic Magnetic Resonance Imaging Using Mask Region-Based Convolutional Neural Networks. *Adv. Radiat. Oncol.* **2020**, *5*, 473–481. [CrossRef]
12. Pellicer-Valero, O.J.; Jiménez, J.L.M.; Gonzalez-Perez, V.; Ramón-Borja, J.L.C.; García, I.M.; Benito, M.B.; Gómez, P.P.; Rubio-Briones, J.; Rupérez, M.J.; Martín-Guerrero, J.D. Deep Learning for Fully Automatic Detection, Segmentation, and Gleason Grade Estimation of Prostate Cancer in Multiparametric Magnetic Resonance Images. *arXiv* **2021**, arXiv:2103.12650. Available online: https://arxiv.org/abs/2103 (accessed on 19 August 2021).
13. Aldoj, N.; Lukas, S.; Dewey, M.; Penzkofer, T. Semi-Automatic Classification of Prostate Cancer on Multi-Parametric MR Imaging Using a Multi-Channel 3D Convolutional Neural Network. *Eur. Radiol.* **2020**, *30*, 1243–1253. [CrossRef]
14. Schelb, P.; Kohl, S.; Radtke, J.P.; Wiesenfarth, M.; Kickingereder, P.; Bickelhaupt, S.; Kuder, T.A.; Stenzinger, A.; Hohenfellner, M.; Schlemmer, H.-P.; et al. Classification of Cancer at Prostate MRI: Deep Learning versus Clinical PI-RADS Assessment. *Radiology* **2019**, *293*, 607–617. [CrossRef]
15. Starmans, M.P.; Niessen, W.J.; Schoots, I.; Klein, S.; Veenland, J.F. Classification of Prostate Cancer: High Grade Versus Low Grade Using A Radiomics Approach. In Proceedings of the 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), Venice, Italy, 8–11 April 2019; pp. 1319–1322.
16. Starmans, M.P.A.; van der Voort, S.R.; Tovar, J.M.C.; Veenland, J.F.; Klein, S.; Niessen, W.J. Chapter 18—Radiomics: Data Mining Using Quantitative Medical Image Features. In *Handbook of Medical Image Computing and Computer Assisted Intervention*; Zhou, S.K., Rueckert, D., Fichtinger, G., Eds.; The Elsevier and MICCAI Society Book Series; Academic Press: Cambridge, MA, USA, 2020; pp. 429–456. ISBN 978-0-12-816176-0.
17. Wang, H.; Wang, L.; Lee, E.H.; Zheng, J.; Zhang, W.; Halabi, S.; Liu, C.; Deng, K.; Song, J.; Yeom, K.W. Decoding COVID-19 Pneumonia: Comparison of Deep Learning and Radiomics CT Image Signatures. *Eur. J. Nucl. Med. Mol. Imaging* **2021**, *48*, 1478–1486. [CrossRef]
18. Castillo T., J.M.; Starmans, M.P.A.; Arif, M.; Niessen, W.J.; Klein, S.; Bangma, C.H.; Schoots, I.G.; Veenland, J.F. A Multi-Center, Multi-Vendor Study to Evaluate the Generalizability of a Radiomics Model for Classifying Prostate Cancer: High Grade vs. Low Grade. *Diagnostics* **2021**, *11*, 369. [CrossRef] [PubMed]
19. Armato, S.G.; Huisman, H.; Drukker, K.; Hadjiiski, L.; Kirby, J.S.; Petrick, N.; Redmond, G.; Giger, M.L.; Cha, K.; Mamonov, A.; et al. PROSTATEx Challenges for Computerized Classification of Prostate Lesions from Multiparametric Magnetic Resonance Images. *J. Med. Imaging* **2018**, *5*, 1–9. [CrossRef] [PubMed]
20. Mottet, N.; Cornford, P.; Briers, E.; Gillessen, S.; Grummet, J.; Henry, A.M.; Lam, T.B.; Mason, M.D. Eau-Eanm-Estro-Esur-Siog Guidelines On Prostate Cancer. *Eur. Urol.* **2020**, *79*, 26.
21. Van den Bergh, R.C.N.; Roemeling, S.; Roobol, M.J.; Roobol, W.; Schröder, F.H.; Bangma, C.H. Prospective Validation of Active Surveillance in Prostate Cancer: The PRIAS Study. *Eur. Urol.* **2007**, *52*, 1560–1563. [CrossRef]
22. PROSTATEx Challenge 2017—The Cancer Imaging Archive (TCIA) Public Access—Cancer Imaging Archive Wiki. Available online: https://wiki.cancerimagingarchive.net/display/Public/PROSTATEx+Challenge+2017 (accessed on 13 July 2021).
23. Clark, K.; Vendt, B.; Smith, K.; Freymann, J.; Kirby, J.; Koppel, P.; Moore, S.; Phillips, S.; Maffitt, D.; Pringle, M.; et al. The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository. *J. Digit. Imaging* **2013**, *26*, 1045. [CrossRef]
24. Turkbey, B.; Rosenkrantz, A.B.; Haider, M.A.; Padhani, A.R.; Villeirs, G.; Macura, K.J.; Tempany, C.M.; Choyke, P.L.; Cornud, F.; Margolis, D.J.; et al. Prostate Imaging Reporting and Data System Version 2.1: 2019 Update of Prostate Imaging Reporting and Data System Version 2. *Eur. Urol.* **2019**, *76*, 340–351. [CrossRef] [PubMed]
25. Schoots, I.G.; Osses, D.F.; Drost, F.J.H.; Verbeek, J.F.M.; Remmers, S.; van Leenders, G.J.L.H.; Bangma, C.H.; Roobol, M.J. Reduction of MRI-Targeted Biopsies in Men with Low-Risk Prostate Cancer on Active Surveillance by Stratifying to PI-RADS and PSAdensity, with Different Thresholds for Significant Disease. *Transl. Androl. Urol.* **2018**, *7*, 132–144. [CrossRef]
26. Weinreb, J.C.; Barentsz, J.O.; Choyke, P.L.; Cornud, F.; Haider, M.A.; Macura, K.J.; Margolis, D.; Schnall, M.D.; Shtern, F.; Tempany, C.M.; et al. PI-RADS Prostate Imaging—Reporting and Data System: 2015, Version 2. *Eur. Urol.* **2016**, *69*, 16–40. [CrossRef] [PubMed]
27. Epstein, J.I.; Egevad, L.; Amin, M.B.; Delahunt, B.; Srigley, J.R.; Humphrey, P.A. The 2014 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason Grading of Prostatic Carcinoma: Definition of Grading Patterns and Proposal for a New Grading System. *Am. J. Surg. Pathol.* **2016**, *40*, 244–252. [CrossRef] [PubMed]

28. Starmans, M.P.A. *MStarmans91/WORC*. 2021. Available online: https://github.com/MStarmans91/WORCDatabase (accessed on 6 August 2021).
29. Castillo T., J. M.; Arif, M.; Niessen, W.J.; Schoots, I.G.; Veenland, J.F. Automated Classification of Significant Prostate Cancer on MRI: A Systematic Review on the Performance of Machine Learning Applications. *Cancers* **2020**, *12*, 1606. [CrossRef] [PubMed]
30. Castaldo, R.; Cavaliere, C.; Soricelli, A.; Salvatore, M.; Pecchia, L.; Franzese, M. Radiomic and Genomic Machine Learning Method Performance for Prostate Cancer Diagnosis: Systematic Literature Review. *J. Med. Internet. Res.* **2021**, *23*, e22394. [CrossRef]
31. Ferro, M.; de Cobelli, O.; Vartolomei, M.D.; Lucarelli, G.; Crocetto, F.; Barone, B.; Sciarra, A.; Del Giudice, F.; Muto, M.; Maggi, M.; et al. Prostate Cancer Radiogenomics—From Imaging to Molecular Characterization. *Int. J. Mol. Sci.* **2021**, *22*, 9971. [CrossRef]
32. Vos, M.; Starmans, M.P.A.; Timbergen, M.J.M.; van der Voort, S.R.; Padmos, G.A.; Kessels, W.; Niessen, W.J.; van Leenders, G.J.L.H.; Grünhagen, D.J.; Sleijfer, S.; et al. Radiomics Approach to Distinguish between Well Differentiated Liposarcomas and Lipomas on MRI. *Br. J. Surg.* **2019**, *106*, 1800–1809. [CrossRef]
33. Van der Voort, S.R.; Incekara, F.; Wijnenga, M.M.J.; Kapas, G.; Gardeniers, M.; Schouten, J.W.; Starmans, M.P.A.; Tewarie, R.N.; Lycklama, G.J.; French, P.J.; et al. Predicting the 1p/19q Codeletion Status of Presumed Low-Grade Glioma with an Externally Validated Machine Learning Algorithm. *Clin. Cancer Res.* **2019**, *25*, 7455–7462. [CrossRef]
34. Zhou, Z.-H. Ensemble Learning. In *Encyclopedia of Biometrics*; Li, S.Z., Jain, A., Eds.; Springer US: Boston, MA, USA, 2009; pp. 270–273. ISBN 978-0-387-73003-5.
35. Wang, Y.; Yue, W.; Li, X.; Liu, S.; Guo, L.; Xu, H.; Zhang, H.; Yang, G. Comparison Study of Radiomics and Deep Learning-Based Methods for Thyroid Nodules Classification Using Ultrasound Images. *IEEE Access.* **2020**, *8*, 52010–52017. [CrossRef]
36. Truhn, D.; Schrading, S.; Haarburger, C.; Schneider, H.; Merhof, D.; Kuhl, C. Radiomic versus Convolutional Neural Networks Analysis for Classification of Contrast-Enhancing Lesions at Multiparametric Breast MRI. *Radiology* **2019**, *290*, 290–297. [CrossRef]
37. Sun, Q.; Lin, X.; Zhao, Y.; Li, L.; Yan, K.; Liang, D.; Sun, D.; Li, Z.-C. Deep Learning vs. Radiomics for Predicting Axillary Lymph Node Metastasis of Breast Cancer Using Ultrasound Images: Don't Forget the Peritumoral Region. *Front. Oncol.* **2020**, *10*, 53. [CrossRef]
38. Xia, X.; Gong, J.; Hao, W.; Yang, T.; Lin, Y.; Wang, S.; Peng, W. Comparison and Fusion of Deep Learning and Radiomics Features of Ground-Glass Nodules to Predict the Invasiveness Risk of Stage-I Lung Adenocarcinomas in CT Scan. *Front. Oncol.* **2020**, *10*, 418. [CrossRef] [PubMed]
39. Perdana, N.R.; Mochtar, C.A.; Umbas, R.; Hamid, A.R.A. The Risk Factors of Prostate Cancer and Its Prevention: A Literature Review. *Acta Med. Indones* **2016**, *48*, 228–238. [PubMed]
40. Timbergen, M.J.M.; Starmans, M.P.A.; Padmos, G.A.; Grünhagen, D.J.; van Leenders, G.J.L.H.; Hanff, D.F.; Verhoef, C.; Niessen, W.J.; Sleijfer, S.; Klein, S.; et al. Differential Diagnosis and Mutation Stratification of Desmoid-Type Fibromatosis on MRI Using Radiomics. *Eur. J. Radiol.* **2020**, *131*, 109266. [CrossRef]
41. Starmans, M.P.A. WORC v3.4.5; 2021. Available online: https://worc.readthedocs.io/en/v3.3.4/ (accessed on 18 December 2021).
42. MStarmans91; Svdvoort. *Svdvoort/PREDICTFastr: V3.1.12*. 2020. Available online: https://zenodo.org/record/4045375#.Yb8HtdDMJPY (accessed on 18 December 2021).
43. Zwanenburg, A.; Vallières, M.; Abdalah, M.A.; Aerts, H.J.W.L.; Andrearczyk, V.; Apte, A.; Ashrafinia, S.; Bakas, S.; Beukinga, R.J.; Boellaard, R.; et al. The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-Based Phenotyping. *Radiology* **2020**, *295*, 328–338. [CrossRef] [PubMed]
44. Frangi, A.F.; Niessen, W.J.; Vincken, K.L.; Viergever, M.A. Multiscale Vessel Enhancement Filtering. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention—MICCAI'98; Wells, W.M., Colchester, A., Delp, S., Eds.; Springer: Berlin/Heidelberg, Germany, 1998; pp. 130–137.
45. Ojala, T.; Pietikainen, M.; Maenpaa, T. Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 971–987. [CrossRef]
46. Kovesi, P. Phase Congruency Detects Corners and Edges. In *Proceedings of the in The Australian Pattern Recognition Society Conference*; Sun, C., Talbot, H., Ourselin, S., Adriaansen, T., Eds.; Csiro Publishing: Collingwood, VIC, Australia, 2003; pp. 309–318.
47. Lemaître, G.; Nogueira, F.; Aridas, C.K. Imbalanced-Learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *J. Mach. Learn. Res.* **2017**, *18*, 1–5.