A multi-center, multi-vendor study to evaluate the generalizability of a radiomics model for classifying prostate cancer

High grade vs. low grade

Castillo T., Jose M.; Starmans, Martijn P.A.; Arif, Muhammad; Niessen, Wiro J.; Klein, Stefan; Bangma, Chris H.; Schoots, Ivo G.; Veenland, Jifke F.

DOI
10.3390/diagnostics11020369

Publication date
2021

Document Version
Final published version

Published in
Diagnostics

diagnostics

MDPI

Article

A Multi-Center, Multi-Vendor Study to Evaluate the Generalizability of a Radiomics Model for Classifying Prostate cancer: High Grade vs. Low Grade

Jose M. Castillo T. [1,*], Martijn P. A. Starmans [1], Muhammad Arif [1], Wiro J. Niessen [1,2], Stefan Klein [1], 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; m.starmans@erasmusmc.nl (M.P.A.S.); a.muhammad@erasmusmc.nl (M.A.); w.niessen@erasmusmc.nl (W.J.N.); s.klein@earasmusmc.nl (S.K.); i.schoots@erasmusmc.nl (I.G.S.); j.veenland@erasmusmc.nl (J.F.V.)
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.castillotovar@erasmusmc.nl

Abstract: Radiomics applied in MRI has shown promising results in classifying prostate cancer lesions. However, many papers describe single-center studies without external validation. The issues of using radiomics models on unseen data have not yet been sufficiently addressed. The aim of this study is to evaluate the generalizability of radiomics models for prostate cancer classification and to compare the performance of these models to the performance of radiologists. Multiparametric MRI, photographs and histology of radical prostatectomy specimens, and pathology reports of 107 patients were obtained from three healthcare centers in the Netherlands. By spatially correlating the MRI with histology, 204 lesions were identified. For each lesion, radiomics features were extracted from the MRI data. Radiomics models for discriminating high-grade (Gleason score ≥ 7) versus low-grade lesions were automatically generated using open-source machine learning software. The performance was tested both in a single-center setting through cross-validation and in a multi-center setting using the two unseen datasets as external validation. For comparison with clinical practice, a multi-center classifier was tested and compared with the Prostate Imaging Reporting and Data System version 2 (PIRADS v2) scoring performed by two expert radiologists. The three single-center models obtained a mean AUC of 0.75, which decreased to 0.54 when the model was applied to the external data, the radiologists obtained a mean AUC of 0.46. In the multi-center setting, the radiomics model obtained a mean AUC of 0.75 while the radiologists obtained a mean AUC of 0.47 on the same subset. While radiomics models have a decent performance when tested on data from the same center(s), they may show a significant drop in performance when applied to external data. On a multi-center dataset our radiomics model outperformed the radiologists, and thus, may represent a more accurate alternative for malignancy prediction.

Keywords: prostate carcinoma; radiomics; machine learning; MRI

Citation: 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. https://doi.org/10.3390/diagnostics11020369

Academic Editors: Martin Andreas Røder and John Thomas Helgstrand

Received: 15 December 2020
Accepted: 19 February 2021
Published: 22 February 2021

1. Introduction

Prostate cancer (PCa) is the most common malignancy and second leading cause of cancer-related death in men [1]. From all patients diagnosed with PCa, those with low-grade lesions might be candidates for active surveillance, whereas patients with high-grade PCa require treatment [2]. The gold standard for PCa assessment in current clinical practice is histopathological verification of biopsy cores [2]. These cores are evaluated by a pathologist and assigned a grade using the Gleason score (GS). However, this procedure has shown to be susceptible to under-diagnosis of high-grade PCa and over-diagnosis of low grade PCa [3].

Multi-parametric magnetic resonance imaging (mpMRI) has received increasing interest for diagnosing, monitoring and treatment follow up for PCa. MpMRI allows non-invasive visualization of the whole prostatic tissue and extraction of quantitative parameters such as tissue density and permeability. To evaluate mpMRI, radiologists use the Prostate Imaging Reporting and Data System (PIRADS) v2, with a grading scale from one (highly unlikely to be clinically significant prostate cancer) to five (highly likely to be clinically significant prostate cancer) [4]. Nevertheless, mpMRI interpretation is challenging and prone to inter- and intra-reader variability among expert radiologists [3].

By extracting multiple imaging features, radiomics has the potential to evaluate the mpMRI data in a more objective way. In the context of PCa, the literature has shown evidence of the potential of radiomics in classifying PCa lesions [5–8], with promising performances in terms of sensitivity and specificity [9]. Nevertheless, current studies on prostate MRI radiomics still lack the quality required to allow their introduction in clinical practice [9,10]. This is due to the fact that most of the radiomics studies validated their approach by splitting their original dataset in training and validation subsets, while only a few studies performed a validation using an external set [11–13]. The latter evaluation is more relevant for a clinical context, where new data can present variations that were not taken into account when the original model was created. Three sources of variations can be identified: at the patient level, at the level of the MRI scanner, and at the level of the clinician. At the patient level: a model created with patient data collected in a specialized treatment centre, will differ from a model based on data collected in a hospital with a surveillance function. Magnetic resonance (MR) images vary between vendors and between scanner types from the same vendor, even if the same acquisition parameters are used. Current evidence shows that is possible to overcome these differences by applying feature harmonization techniques [14]. These techniques aim to estimate the statistical differences between imaging features computed from different data sets and apply a correction for it. To our knowledge there is no scientific evidence reporting the usage of feature harmonization in the context of PCa classification. At the clinician level: the pathologist reports, which are used as ground truth for the model, are based on the visual Gleason grading of pathologists, who are prone to considerable inter-observer variation [15,16]. Therefore, the question arises what performance can be expected when testing radiomics models on unseen multi-center-multivendor data: how generalizable are radiomics model in the context of PCa? The number of studies addressing generalizability is limited. To our knowledge, few studies tested their model's generalizability for PCa detection regarding tumor aggressiveness using multiple scanners [17–19]. Only a few studies have validated their methods using external datasets for PCa tumor grade prediction [9]. When radiomics models are being considered as decision support tools for clinical practice, the generalizability issue should be addressed.

The main contribution of this study is two-fold. First, we assessed the generalizability of a radiomics approach for classifying PCa in a multi-center, multivendor setting. Second, in the same setting we compared the classification performance of radiologists to the performance of our radiomics model.

2. Materials and Methods

Our patient cohort was obtained from three healthcare centers in the Netherlands in the context of the Prostate Cancer Molecular Medicine project (PCMM), in Table 1 some of the clinical variables of this set are summarized. A Kruskal–Wallis test was performed to check whether the median of the GS distribution, volume, and prostatic specific antigen (PSA) of the included data sets were comparable.

Table 1. Prostate Cancer Molecular Medicine (PCMM) data set clinical variables and lesions characteristics. PIRADS grading performed by radiologist 1 (R1) and 2(R2). Age of patients for data sets B and C was not available (NA). PZ: Peripheral zone. TZ: transition zone. AFS: anterior fibromuscular stroma. IQR: interquartile range.

Prostate Cancer Molecular Medicine Data set Clinical Variables			
Center	**A**	**B**	**C**
Number of Patients	29	38	40
Age at Diagnosis (mean ± std years)	64 ± 7	NA	NA
PSA before treatment (mean ± std ng/mL)	12 ± 10	9 ± 5	10 ± 8
Lesions Characteristics			
Number of lesions		204	
Lesion location			
PZ	33	59	45
TZ	15	23	26
AFS	NA	2	1
Lesion volume (median and IQR mL)	1.6 (0.2–1.8)	1.4 (0.1–1.5)	0.8 (0.2–1.1)
Radiologist PIRADS grading		**R1**	**R2**
	I	0	4
	II	16	9
	III	21	36
	IV	33	34
	V	43	61
	Total	113	144

The data usage of this study was approved by the medical ethics review committee of Erasmus MC under the number NL32105.078.10. In this PCMM-project, the mpMRI and pathology data of men with localized PCa who were scheduled for prostatectomy were prospectively collected from 2011 to 2014. In this study, we will refer to the data from the respective centers as data set A, B and C. The data of each center were visually graded by a radiologist and a pathologist working at that center. In total we included 107 patients for whom MRI, pathology images and reports were available. The distribution was as follows: A = 29, B = 38 and C = 40, the details regarding the MRI scanners and acquisition parameters of each set are described in Appendix A. The dataset shows considerable variability, with images acquired with scanners from three different vendors, using various voxel sizes and b values for the diffusion weighted sequences. In deriving our radiomics models we included the T2-weighted (T2w) and the diffusion weighted imaging (DWI) sequences and the apparent diffusion coefficient maps (ADC) derived from the DWI images.

All 107 patients had their prostate surgically removed. After the prostatectomy, the prostate was cut into 3 mm thick slices. Of the top of each slice, a photograph was taken, and 4µm coupes were cut and stained with H&E. Based on the H&E, the pathologist marked the areas with cancerous tissue on the photographs and assigned a GS to each tumor region. In Figure 1 the number of lesions per GS found in each set is summarized. We grouped lesions with a GS $\leq$ 6 as low-grade tumors and lesions with a GS $\geq$ 7 as high-grade tumors. Out of the 107 patients, 204 lesions in total were processed, 92 (45%) low-grade and 112 (55%) high-grade. The methods used to correlate the lesions found in the pathology with MRI are explained in the following section.

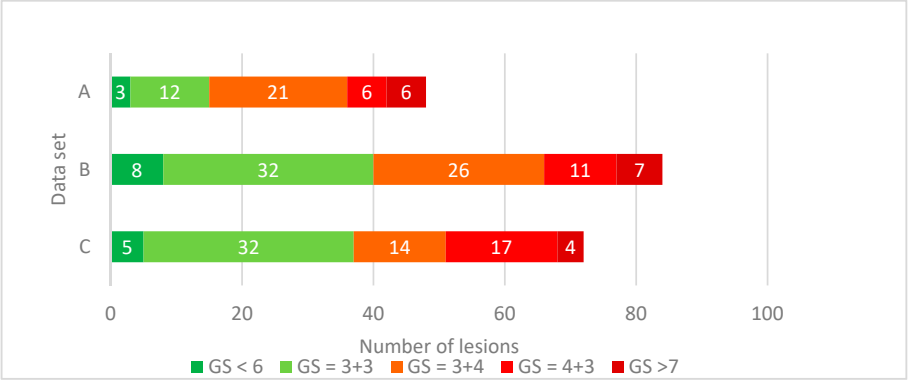

Figure 1. Distribution of Gleason grading of identified lesions at radical prostatectomy specimen of three different centers. The number of lesions per group is shown in white.

2.1. Ground Truth Construction: Pathology-MRI Correlation

A mask of identified lesions based on microscopy analysis (H&E staining) was manually drawn by a pathologist on the prostatectomy specimens' photos. Using in house software implemented in Mevislab (v-2.2.1, Germany) [20], the macroscopy images of the prostatectomy specimen were manually registered and stacked to generate a prostate volume to enable the registration with MRI. Then, based on the prostate borders, prostate masks were manually drawn on the MR and macroscopy images. Afterwards, these two masks were manually aligned in 3D by rotation, translation, and scaling of the pathology volume. Subsequently, the translation in slice-direction was fine-tuned while inspecting the pathology and the corresponding T2w slices. As the last step, the lesion segmentation from the pathology volume was overlaid on the T2w volume.

2.2. Image Pre-Processing.

In order to address the variation in image resolution between and within data sets, the MR images were resampled to a voxel grid of 0.27 mm × 0.27 mm × 3 mm, which was the spacing used in the largest proportion (36%) of the T2w images.

2.3. Radiomics Generalizability Evaluation

To assess the generalizability of our radiomics models, we used the experimental setup as shown in Figure 2. Image data from a single center was used to train a radiomics classifier for each center. On this training set, an 100× internal random-split cross-validation was used to assess the single center performance. Finally, the model was evaluated using the other two sets to assess the generalizability; this procedure was repeated with each set. The details regarding the development of the radiomics classifiers are explained in the following section.

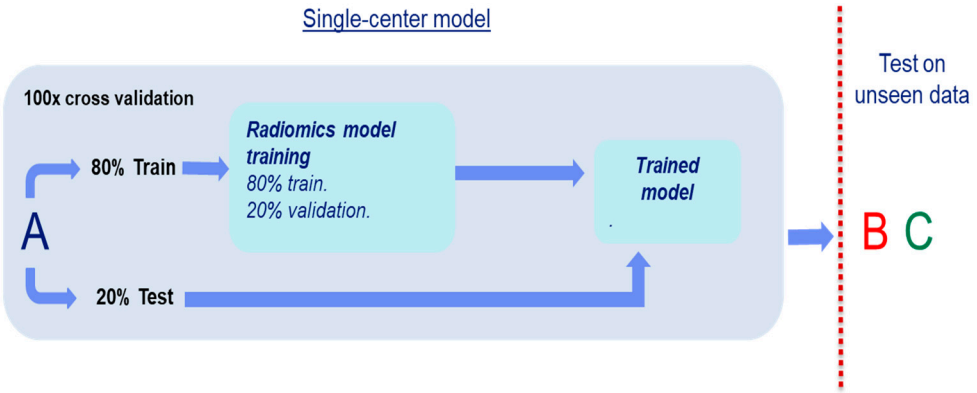

Figure 2. Scheme of the generalization experiment setting. In this example dataset A is used to develop a model. The model is tested on the other two sets (B and C).

2.4. Radiomics Model Development

To generate the radiomics classifiers for each data set, we used the open-source Workflow for Optimal Radiomics Classification (v-3.3.2, Rotterdam, The Netherlands,) platform (WORC) with the default settings [21] and another setting including feature harmonization with ComBat [22]. WORC performs an automatic search amongst a wide variety of algorithms and their corresponding parameters to determine the optimal combination that maximizes the prediction performance on the training set, a schematic overview of the method is shown in Figure 3**.** The workflow starts with the user defining a region of interest (ROI) from the image, which in our case was the delineation obtained by the pathology–MRI correlation. Within these tumor masks, features quantifying intensity, shape, texture and orientation were extracted from the T2w, ADC and the highest b-value image available from the DWI images. Following feature extraction, a decision model was created, which in WORC consist of several steps, such as feature selection, oversampling and machine learning methods. WORC automatically optimizes the radiomics pipeline: during each iteration WORC generates 100,000 workflows by using different combinations of methods and parameters. At the end of each cross validation, the 50 best performing solutions were combined in an ensemble as a single classification model. The final ensemble of 50 classifiers is the resulting radiomics model, the performance of which is evaluated on the independent test set (external evaluation). Feature selection was done to select the most predictive features through enabling/disabling entire families of features (e.g., shape, local binary patterns, texture based on grey-level co-occurrence matrices). The code utilized for these experiments is available online in a GitHub repository [23].

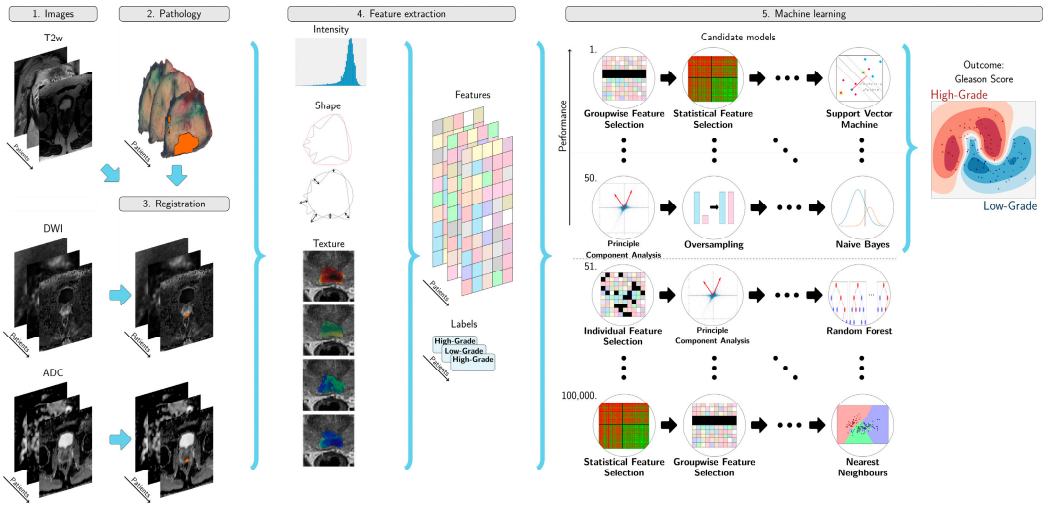

Figure 3. (1) The magnetic resonance sequences to be used in the model are defined. (2) The lesions from the pathology are copied and registered to the T2w sequence. (3) The diffusion weighted imaging (DWI) and apparent diffusion coefficient (ADC) are resampled and registered to the T2w. (4) Features are extracted from the T2w, DWI and ADC. (5) A radiomics model is created from the features, using an ensemble of the best 50 workflows from 100,000 candidate workflows, where the workflows are different combinations of the different classifiers.

2.5. Radiomics Classifier Evaluation

The internal evaluation of the model was performed by using a 100× random-split cross validation: First, the data set was split into 80% for training and 20% for testing. After this, 20% of the training set was used as validation set. This validation set was used in each training iteration to select the best parameters in order to optimize the prediction accuracy. The remaining 20% was used for performance evaluation: area under the curve (AUC), receiver operating characteristic (ROC) curve, sensitivity, and specificity. The high-grade tumors were considered the positive class. To compute the 95% confidence intervals (CI)

in the cross-validation experiment, we used the corrected resampled t-test [24]. ROC confidence bands were constructed using fixed-width bands [25].

To analyze the impact of having multiple lesions from the same patient, we performed the external evaluation both at the lesion and patient level. At the patient level, for each patient only the highest grade lesion was taken into account.

2.6. Comparison of Our Radiomics Model with the Clinical Assessment using PIRADS v2

To compare the classification performance of a multi-center radiomics model with the clinical assessment using the PIRADS v2 score, a test set was evaluated by both radiomics and the radiologist, see Figure 4. The PIRADS scoring of the lesions was done by two radiologists with 4 years and 10 years of experience, respectively, from of the partaking centers A and B, fully blinded from histopathology results. The lesions graded as having a PIRADS ≥ 3 were considered positive for high-grade PCa and the lesions with a score ≤ 2 as negative for high-grade PCa.

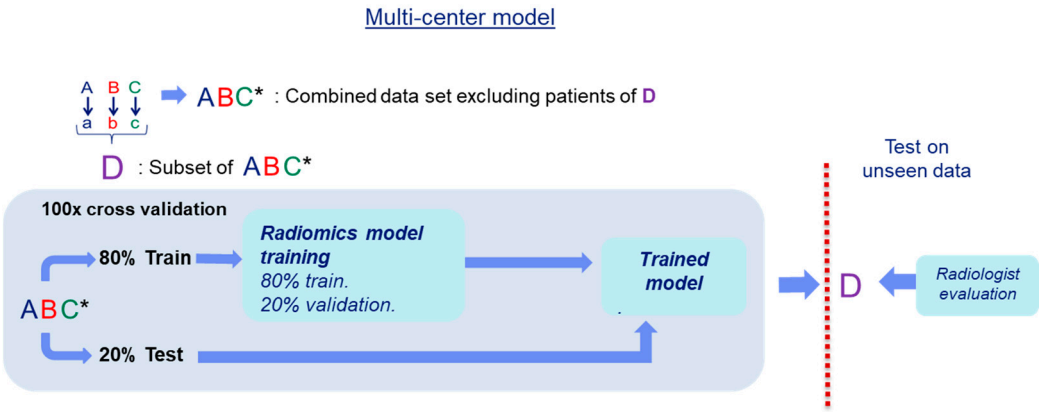

Figure 4. Scheme of the comparison experiment of our multi-center radiomics model with the evaluation by the radiologist. A randomly selected set of patients in ABC was set apart as test set (D), the rest of the data (ABC*) was used to develop the multi-center radiomics model.

For this experiment, in order to avoid a bias towards a single center, we created a test set (D) by randomly selecting 20% of the data from each of the three centers. From this set, the lesions that were not detected by one of the two radiologists were removed from the study since our goal was to compare the classification performance, not the detection rate. Subsequently, the remaining patient data (ABC*) was used to train a radiomics model to classify the patients in set D. The end performance for either radiologist and the radiomics model was computed on patient level classification.

3. Results

Statistical analysis of clinical variables:

The median of the Gleason Score (H = 4.63, $p = 0.09$), the lesion volume (H = 5.85, $p = 0.06$) and PSA (H = 1.99, $p = 0.36$) were similar for the three data sets.

Radiomics model generalizability:

Table 2 shows the results for the generalizability test. Overall, it can be seen that even though reasonable performances in terms of AUC (mean = 0.75) were obtained from the internal cross-validations, when the models were tested on the other data sets, the performances dropped considerably (mean AUC = 0.54). The inclusion of feature harmonization with ComBat did not improve the performance of the radiomics models. The performance metrics on the external validation sets were comparable when evaluated lesion and patient wise. Meanwhile, radiologists' performance (mean AUC = 0.47) shows high sensitivity with a low specificity.

Table 2. Generalization study results. Internal: internal evaluation was performed using a 100× random-split cross-validation, reported with confidence interval. External: by training in one dataset, testing on the two remaining datasets. LC: lesion level classification. PC: patient level classification. AUC: area under the curve. CH: Test result using ComBat feature harmonization. R1 and R2: radiologist 1 and 2.

Model	Internal	External LC	External CH	External PC	R1 and R2
Trained on A	**A**	**B and C**			
AUC	0.75 (0.58–0.92)	0.43	0.49	0.55	0.44
Sensitivity	0.91 (0.82–1.00)	0.80	0.78	0.81	0.80
Specificity	0.30 (0.03–0.55)	0.22	0.27	0.21	0.06
Trained on B	**B**	**A and C**			
AUC	0.69 (0.57–0.81)	0.60	0.57	0.55	0.50
Sensitivity	0.64 (0.47–0.80)	0.43	0.74	0.86	0.88
Specificity	0.67 (0.50–0.83)	0.62	0.38	0.25	0.13
Trained on C	**C**	**A and B**			
AUC	0.80 (0.68–0.92)	0.60	0.62	0.65	0.44
Sensitivity	0.74 (0.66–0.86)	0.52	0.51	0.48	0.69
Specificity	0.66 (0.50–0.82)	0.63	0.69	0.63	0.19

Comparison of Our Radiomics Model with the Clinical Assessment using PIRADS v2

The resulting test set was composed of 16 patients with high-grade lesions and eight patients with low-grade lesions. Table 3 presents the results of the classification performance for the internal cross-validation and the performance on the test set (ABC*) for the model and the two radiologists. It can be seen that the radiomics model outperformed (AUC = 0.75) the radiologist classification with the PIRADS score (AUC of 0.50 and 0.44). Radiologists achieved a decent sensitivity (0.76 and 0.88), but near-zero specificity (0.25 and 0.0), whereas the radiomics model achieved a sensitivity of 0.88 and a specificity of 0.63.

Table 3. Performance comparison of the multi-center radiomics model with the PIRADS score performed by two radiologists. Internal: Internal cross validation results reported with confidence intervals. AUC: area under the curve. Model: results from the multi-center model for the unseen data. R1 and R2: radiologist 1 and 2, respectively.

Metrics	Internal	Model	R1	R2
AUC	0.72 (0.64–0.79)	0.75	0.50	0.44
Sensitivity	0.76 (0.66–0.89)	0.88	0.76	0.88
Specificity	0.55 (0.44–0.66)	0.63	0.25	0.00

4. Discussion

The expanding usage of prostate MRI for PCa diagnosis has brought an increased interest in radiomics research for tumor classification. As a result, many approaches have been proposed, and promising results have been presented, thus raising the opportunity of using these models in daily clinical workflow. However, there is limited evidence regarding the performance of these models with unseen data in a new clinical contexts, for instance with MR scanners from different vendors and/or grading by different pathologists and/or

different patient profiles. Investigating how these changes affect radiomics performance is required prior to applying these models in a clinical setting.

In this study we developed radiomics classifiers starting from three independent sets and evaluated the performance on the unseen data of the other centers. To compensate for the differences between data sets and reduce the negative effects on performance that these differences might have, resampled all the images in our experiments to the same voxel size, and used the same method to correlate the pathology data to the MR data. Furthermore, we applied techniques such as normalization and class unbalance correction. While obtaining a decent performance working with data from a single center, our results showed a substantial decline in performance when evaluating the radiomics models on external data. Thus, since an internal validation on a single-center dataset is not representative of external performance, it is advisable to carry out external validations to have a realistic estimation of predictive power.

The decline in performance is most probably related to several factors. One important factor affecting the feature computation is the dependency of the radiomics features on MR scanning parameters [26]. It has been shown that image normalization applied with variety of approaches or pre-filtering cannot overcome the scan-feature dependency problem [27]. Recent literature shows evidence that it is possible to overcome the scanner-feature dependency issue by applying feature harmonization techniques such as ComBat [22]. In our experiments, we applied feature harmonization using ComBat, however the inclusion of this technique did not improve our results while testing on the external sets.

Another factor is that the delineations on the pathology data were carried out by different pathologists working at the different centers. These delineations were transferred to the MRI, but the delineation is a factor that influences the feature computation [28], compromising the likeness of the features computed from different datasets. In clinical practice, the delineation of lesions in MRI is mostly performed by a single clinician, which makes it unfeasible to test feature robustness for several delineations. Furthermore, manual delineation by specialists is time consuming and potentially subject to observer variability. Utilizing either assisted or fully automatic segmentation methods available [29,30] for the prostate and PCa lesions could improve feature computation consistency, important for radiomics approaches, and positively impact the model generalizability.

Various studies have assessed the use of radiomics in PCa classification on mpMRI [9]. To our knowledge, this is the first study to specifically address the generalizability of radiomics models in the context of PCa classification. Our study consisted of multi-centric data sets: image data from multiple vendors and multiple scanners from the same vendor, two different radiologists diagnosing the patients, three different pathology departments grading histology slices of prostatectomies as ground truth. There are studies in which one factor is varied, e.g., the study published by Dinh et al. [31]. In their study they developed a model specifically for peripheral zone PCa detection, maintaining the model's performance between two MR scanners belonging to different vendors. However, in their experiments the data were acquired from the same center, evaluated, and processed by the same radiologists and pathologists. This might have affected positively the performance of their method.

When comparing our radiomics model to the PIRADS v2 scoring by radiologists, our results show that the radiologists achieved high sensitivity at the cost of a low specificity, while our model increased specificity substantially. This high sensitivity with PIRADS v2 may translate in clinical practice in overdiagnosis and overtreatment. A radiomics model may not only provide a more objective quantitative support tool to recommend surveillance for those cases where treatment may not instantly be required, but should also maintain a high sensitivity for those cases with aggressive PCa. However, it is important to take into account the data that the radiomics model was developed on, and the setting the model will be applied in. In other words, the safe utilization of a radiomics model in the clinic is feasible, as long as the population on which it is applied, holds similar characteristics to the population used to develop the model.

Our study has some limitations. First, our ground truth tumor grading is based on one pathologist per center, which can cause discrepancies in lesion delineations and grading. Having a consensus ground truth could have positively impacted our performance. However, this limitation represents current clinical practice, where the reader agreement between pathologists is between 70–80% [15,16].

Secondly, the number of patients included per medical center is limited. However, the total number of patients in our study is higher than the average value of 80 patients found in similar radiomics studies [9]. Thirdly, the clinical assessment was performed using the PIRADS classification v2.0 because v2.1 was not available at the moment of the readings.

Finally, we did not include clinical variables or epidemiological factors in our model. This information plays a role in clinical decision making, therefore, including this information may have a positive impact on the end performance in a multi-center and multi-vendor setting. Although, clinical patient information such as the level of PSA, the patient risk group and the outcome of the digital rectal examination were not available for a substantial number of patients which represented an obstacle to include these variables.

Despite the previous limitations, our study contributes to the field of PCa classification using radiomics by: (1) being the first study with the generalizability of PCa classification radiomics models as main focus; (2) making our scientific code available in a public repository. As regards this last point, we would like to invite the scientific community to test this code on their own data sets and so promote discussions and future collaborations.

Additionally, we would like to make some recommendations for future work: when developing a generalizable radiomics model for PCa classification the data should represent the variation present in the clinical practice with data of several centers with various pathologists and radiologists, and multiple MRI scanners from multiple vendors. The validation of the model should be performed in a prospective cohort.

5. Conclusions

In this paper we assessed the generalizability of radiomics models in the context of PCa grading. When limited to a specific center or, e.g., to a specific scanner or specific setting, these models perform well and may represent a valuable tool to differentiate low-grade from high grade tumors. However, when applying radiomics on data from different centers and/or scanners, a considerable drop in performance can be expected, making these models less reliable in this context.

To become clinical viable and support clinical decision making, training and validation of radiomics models should be performed in multi-center scenarios with data representative of the population on which the model will be applied.

Author Contributions: Conceptualization, J.M.C.T., W.J.N., I.G.S. and J.F.V.; data curation, J.M.C.T., C.H.B., J.F.V.; formal analysis, J.M.C.T., M.P.A.S., M.A., W.J.N., S.K. and J.F.V.; funding acquisition, W.J.N. and J.F.V.; investigation, J.M.C.T., M.P.A.S., M.A., W.J.N., S.K., I.G.S. and J.F.V.; methodology, J.M.C.T., M.P.A.S., M.A., W.J.N., S.K., I.G.S. and J.F.V.; project administration, W.J.N. and J.F.V.; resources, J.M.C.T., W.J.N., C.H.B., I.G.S. and J.F.V.; software, J.M.C.T. and M.P.A.S.; supervision, W.J.N., I.G.S. and J.F.V.; validation, J.M.C.T., M.P.A.S., S.K., I.G.S. and J.F.V.; visualization, J.M.C.T., M.P.A.S. and J.F.V.; writing—original draft, J.M.C.T. and J.F.V.; writing—review and editing, M.P.A.S., M.A., W.J.N., S.K., 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 work is part of the research program strategy with project numbers 14929, 14930, and 14932, which is (partly) financed by the Netherlands organization for scientific research (NWO).

Data Availability Statement: Please refer to suggested Data Availability Statements in section "MDPI Research Data Policies" at https://www.mdpi.com/ethics.data managing tasks related to the PCMM data set.

Acknowledgments: We would like to specially acknowledge the support given by Tim Hulsen with.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Table 1. Table describing scanners characteristics at the three clinical sites. GE: General Electric. T: Tesla. T2: T2-weighted sequence DWI: Diffusion weighted imaging.

Center.	Vendor	Model	Magnetic Field (Tesla).	#Patients	Sequence	Voxel Size (mm)	B-Values	Endorectal Coil
A.	GE Medical Systems.	MR750.	3T.	21.	T2.	0.37 × 0.37 × 3.00 ...		No ...
					DWI	1.09 × 1.09 × 4.00	50/400/800	
	GE Medical Systems	MR450	1.5T	3	T2	0.47 × 0.47 × 3.00		No
					DWI	1.25 × 1.25 × 4.00	100/500/1000	
	SIEMENS	Avanto	1.5T	5	T2	0.70 × 0.70 × 3.00		No
					DWI	1.85 × 1.85 × 6.00	50/400/600	
B	Philips Healthcare	Achieva	3T	38	T2	0.27 × 0.27 × 3.00		Yes
					DWI	1.03 × 1.03 × 3.00	150/300/450/600/750	
C	SIEMENS	TrioTim	3T	17	T2	0.63 × 0.63 × 3.00		
					DWI	2.00 × 2.00 × 4.00	50/500/800	No
		Skyra	3T	23				
					T2	0.60 × 0.60 × 3.00		
					DWI	2.00 × 2.00 × 4.00	50/500/800	No

Appendix B. Radiomics Features Extraction

This supplemental material is similar to (Timbergen et al., 2020; Vos et al., 2019), 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) (Starmans, Van der Voort, Phil, & Klein, 2018), which internally uses the PREDICT (van der Voort & Starmans, 2018) and PyRadiomics (Van Griethuysen et al., 2017) feature extraction toolboxes. For details on the mathematical formulation of the features, we refer the reader to (Zwanenburg et al., 2020). More details on the extracted features can be found in the documentation of the respective toolboxes, mainly the WORC documentation (Starmans, 2018).

For CT scans, the images are by default not normalized 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 published open-source (Starmans, 2020).

The features can be divided in 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 these 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) (Frangi, Niessen, Vincken, & Viergever, 1998), the Gray Level Co-occurrence Matrix (144 features) (Zwanenburg et al., 2020), the Gray Level Size Zone Matrix (16 features) (Zwanenburg et al., 2020), the Gray Level Run Length Matrix (16 features) (Zwanenburg et al., 2020), the Gray Level Dependence Matrix (14 features) (Zwanenburg et al., 2020), the Neighbourhood Grey Tone Difference Matrix (five features) (Zwanenburg et al., 2020), Local Binary Patterns (18 features) (Ojala, Pietikainen, & Maenpaa, 2002), and local phase filters (36 features) (Kovesi, 1997, 2003).

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 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 C, 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 C.

The dataset 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 C. Adaptive Workflow Optimization for Automatic Decision Model Creation

This appendix is similar to (Timbergen et al., 2020; Vos et al., 2019), but details relevant for the current study are highlighted. The Workflow for Optimal Radiomics Classification (WORC) toolbox (Starmans et al., 2018) 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 (Starmans, 2018). The code to use WORC for creating the differential diagnosis and molecular analysis decision models in this specific study has been published open-source (Starmans, 2020).

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 fifth 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 a 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 (Lemaitre, Nogueira, & Aridas, 2017); 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 (Starmans, 2018).

By default, in WORC, the performance is evaluated in a 100× 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 5× random-split cross-validation on the training dataset, using 80% of the data for actual training and 20% for validation of the performance. All described methods are fit on the training datasets, and only tested on the validation datasets. 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 dataset. Hence, to create a more robust model and boost performance, WORC combines the 50 best performing workflows into a single decision model, which is known as ensembling. These 50 best performing workflows are re-trained using the entire training dataset, and only tested on the test datasets. The ensemble is created through averaging of the probabilities, i.e., the chance of lesion with high grade or low grade, of these 50 workflows. A full experiment consists of executing 50 million workflows (100,000 pseudo-randomly generated workflows, times a 5× train-validation cross-validation times 100× train-test cross-validation), which can be parallelized.

Appendix D. Appendix References

1. Frangi, A.F.; Niessen, W.J.; Vincken, K.L.; Viergever, M.A. Multiscale Vessel Enhancement Filtering. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI'98*; MICCAI 1998, Lecture Notes in Computer Science; Wells, W.M., Colchester, A., Delp, S., Eds.; Springer: Berlin/Heidelberg, Germany, 1998; Volume 1496, doi:10.1007/BFb0056195.
2. Kovesi, P. Symmetry and Asymmetry from Local Phase. In Proceedings of the Roceedings of the 10th Australian Joint Conference on Artificial Intelligence: Advanced Topics in Artificial Intelligence, Perth, Australia, 1997, pp. 185–190.
3. Kovesi, P. Phase Congruency Detects Corners and Edges. In Proceedings of the VIIth Digital Image Computing: Techniques and Applications, Sydney, Australia, 10−12 December 2003.
4. Lemaitre, 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.

5. 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, doi:10.1109/TPAMI.2002.1017623.
6. Starmans, M.P.A. Workflow for Optimal Radiomics Classification (WORC) Documentation. Available online: https://worc.readthedocs.io (accessed on 16 February 2021); doi:10.5281/zenodo.3840534.
7. Starmans, M.P.A. GISTRadiomics. Available online: https://github.com/MStarmans91/GISTRadiomics (accessed on 16 February 2021); doi:10.5281/zenodo.3839323.
8. Starmans, M.P.A.; Van der Voort, S.R.; Phil, T.; Klein, S. Workflow for Optimal Radiomics Classification (WORC). Available online: https://github.com/MStarmans91/WORC (accessed on 16 February 2021); doi:10.5281/zenodo.3840534.
9. Timbergen, M.J.M.; Starmans, M.P.A.; Padmos, G.A.; Grünhagen, D.J.; van Leenders, G.J.L.H.; Hanff, D.; Niessen, W.J.; Sleijfer, S.; Klein, S.; Visser, J.J. Differential diagnosis and mutation stratification of desmoid-type fibromatosis on MRI using radiomics. *Eur. J. Radiol.* 2020, 109266, doi:10.1016/j.ejrad.2020.109266.
10. Van der Voort, S.R.; Starmans, M.P.A. Predict a Radiomics Extensive Differentiable Interchangable Classification Toolkit (P ICT). Available online: https://github.com/Svdvoort/PREDICTFastr (accessed on 16 February 2021); doi:10.5281/zenodo.3854839.
11. Van Griethuysen, J.J.; Fedorov, A.; Parmar, C.; Hosny, A.; Aucoin, N.; Narayan, V.; Beets-Tan, R.G.H.; Fillion-Robin,, J.-C.; Pieper, S.; Aerts, H.J. Computational radiomics system to decode the radiographic phenotype. *Cancer Res.* 2017, *77*, e104–e107, doi:10.1158/0008-5472.CAN-17-0339.
12. Vos, M.; Starmans, M.P.A.; Timbergen, M.J.M.; van der Voort, S.R.; Padmos, G.A.; Kessels, W.; Visser, J.J.; 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, doi:10.1002/bjs.11410.
13. Zwanenburg, A.; Vallières, M.; Abdalah, M.; Aerts, H.; 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*, 191145, doi:10.1148/radiol.2020191145.

References

1. Rawla, P. Epidemiology of Prostate Cancer. *Rev. World J. Oncol.* **2019**, *10*, 63–89. [CrossRef] [PubMed]
2. Mottet, N.; van den Bergh, R.C.N.; Briers, E.; Cornford, P.; De Santis, M.; Fanti, S.; Gillessen, S.; Grummet, J.; Henry, A.M.; Lam, T.B.; et al. European Association of Urology: Prostate Cancer Guidelines. Available online: https://uroweb.org/wp-content/uploads/Prostate-Cancer-2018-pocket.pdf (accessed on 15 June 2019).
3. 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]
4. 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]
5. Min, X.; Li, M.; Dong, D.; Feng, Z.; Zhang, P.; Ke, Z.; You, H.; Han, F.; Ma, H.; Tian, J.; et al. Multi-parametric MRI-based radiomics signature for discriminating between clinically significant and insignificant prostate cancer: Cross-validation of a machine learning method. *Eur. J. Radiol.* **2019**, *115*, 16–21. [CrossRef]
6. Hoang Dinh, A.; Souchon, R.; Melodelima, C.; Bratan, F.; Mège-Lechevallier, F.; Colombel, M.; Rouvière, O. Characterization of prostate cancer using T2 mapping at 3 T: A multi-scanner study. *Diagn. Interv. Imaging* **2015**, *96*, 365–372. [CrossRef] [PubMed]
7. Chaddad, A.; Kucharczyk, M.J.; Niazi, T. Multimodal radiomic features for the predicting gleason score of prostate cancer. *Cancers* **2018**, *10*, 249. [CrossRef] [PubMed]
8. Castillo, T.J.M.; Starmans, M.P.A.; 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(New York, USA) 16th International Symposium on Biomedical Imaging (ISBI 2019), Venice, Italy, 8–11 April 2019; pp. 1319–1322.
9. 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]
10. Stanzione, A.; Gambardella, M.; Cuocolo, R.; Ponsiglione, A.; Romeo, V.; Imbriaco, M. Prostate MRI radiomics: A systematic review and radiomic quality score assessment. *Eur. J. Radiol.* **2020**, *129*, 109095. [CrossRef] [PubMed]

11. Transin, S.; Souchon, R.; Gonindard-Melodelima, C.; de Rozario, R.; Walker, P.; Funes de la Vega, M.; Loffroy, R.; Cormier, L.; Rouvière, O. Computer-aided diagnosis system for characterizing ISUP grade ≥2 prostate cancers at multiparametric MRI: A cross-vendor evaluation. *Diagn. Interv. Imaging* **2019**, *100*, 801–811. [CrossRef]
12. Penzias, G.; Singanamalli, A.; Elliott, R.; Gollamudi, J.; Shih, N.; Feldman, M.; Stricker, P.D.; Delprado, W.; Tiwari, S.; Böhm, M.; et al. Identifying the morphologic basis for radiomic features in distinguishing different Gleason grades of prostate cancer on MRI: Preliminary findings. *PLoS ONE* **2018**, *13*. [CrossRef] [PubMed]
13. Dinh, A.H.; Melodelima, C.; Souchon, R.; Moldovan, P.C.; Bratan, F.; Pagnoux, G.; Mège-Lechevallier, F.; Ruffion, A.; Crouzet, S.; Colombel, M.; et al. Characterization of Prostate Cancer with Gleason Score of at Least 7 by Using Quantitative Multiparametric MR Imaging: Validation of a Computer-aided Diagnosis System in Patients Referred for Prostate Biopsy. *Radiology* **2018**, *287*, 525–533. [CrossRef] [PubMed]
14. Orlhac, F.; Boughdad, S.; Philippe, C.; Stalla-Bourdillon, H.; Nioche, C.; Champion, L.; Soussan, M.; Frouin, F.; Frouin, V.; Buvat, I. A postreconstruction harmonization method for multicenter radiomic studies in PET. *J. Nucl. Med.* **2018**, *59*, 1321–1328. [CrossRef] [PubMed]
15. Ozkan, T.A.; Eruyar, A.T.; Cebeci, O.O.; Memik, O.; Ozcan, L.; Kuskonmaz, I. Interobserver variability in Gleason histological grading of prostate cancer. *Scand. J. Urol.* **2016**, *50*, 420–424. [CrossRef]
16. Nilsson, B.; Egevad, L.; Sundelin, B.; Glaessgen, A.; Hamberg, H.; Pihl, C.-G. Interobserver reproducibility of modified Gleason score in radical prostatectomy specimens. *Virchows Arch.* **2004**, *1*, 17–21. [CrossRef] [PubMed]
17. Viswanath, S.E.; Chirra, P.V.; Yim, M.C.; Rofsky, N.M.; Purysko, A.S.; Rosen, M.A.; Bloch, B.N.; Madabhushi, A. Comparing radiomic classifiers and classifier ensembles for detection of peripheral zone prostate tumors on T2-weighted MRI: A multi-site study. *BMC Med. Imaging* **2019**, *19*, 22. [CrossRef]
18. Artan, Y.; Oto, A.; Yetik, I.S. Cross-Device Automated Prostate Cancer Localization With Multiparametric MRI. *IEEE Trans. Image Process.* **2013**, *22*, 5385–5394. [CrossRef]
19. Peng, Y.; Jiang, Y.; Antic, T.; Giger, M.L.; Eggener, S.E.; Oto, A. Validation of Quantitative Analysis of Multiparametric Prostate MR Images for Prostate Cancer Detection and Aggressiveness Assessment: A Cross-Imager Study. *Radiology* **2014**, *271*, 461–471. [CrossRef] [PubMed]
20. MeVisLab: MeVisLab. Available online: https://www.mevislab.de/ (accessed on 13 August 2020).
21. Starmans MPA GitHub—MStarmans91/WORC: Workflow for Optimal Radiomics Classification. Available online: https://github.com/MStarmans91/WORC (accessed on 17 October 2019).
22. Fortin, J.P.; Parker, D.; Tunç, B.; Watanabe, T.; Elliott, M.A.; Ruparel, K.; Roalf, D.R.; Satterthwaite, T.D.; Gur, R.C.; Gur, R.E.; et al. Harmonization of multi-site diffusion tensor imaging data. *Neuroimage* **2017**, *161*, 149–170. [CrossRef] [PubMed]
23. Josemanuel097/PCa_classification_generalizability. Available online: https://github.com/josemanuel097/PCa_classification_generalizability (accessed on 11 February 2021).
24. Nadeau, C.; Bengio, Y. Inference for the Generalization Error. *Mach Learn* **2003**, *52*, 239–281. [CrossRef]
25. Macskassy, S.A.; Provost, F.; Rosset, S. ROC Confidence Bands: An Empirical Evaluation. In Proceedings of the 22nd International Conference on Machine Learning, Bonn, Germany, 7–11 August 2005; Association for Computing Machinery: New York, NY, USA, 2005; pp. 537–544.
26. Buch, K.; Kuno, H.; Qureshi, M.M.; Li, B.; Sakai, O. Quantitative variations in texture analysis features dependent on MRI scanning parameters: A phantom model. *J. Appl. Clin. Med. Phys.* **2018**, *19*, 253–264. [CrossRef]
27. Schwier, M.; van Griethuysen, J.; Vangel, M.G.; Pieper, S.; Peled, S.; Tempany, C.; Aerts, H.J.W.L.; Kikinis, R.; Fennessy, F.M.; Fedorov, A. Repeatability of Multiparametric Prostate MRI Radiomics Features. *Sci. Rep.* **2019**, *9*, 9441. [CrossRef] [PubMed]
28. Starmans, M.P.A.; van der Voort, S.R.; Castillo Tovar, J.M.; Veenland, J.F.; Klein, S.; Niessen, W.J. Radiomics: Data mining using quantitative medical image features. In *Fichtinger GBT-H of MIC and CAI*; Zhou, S.K., Rueckert, D., Eds.; Academic Press: Cambridge, MA, USA, 2020; pp. 429–456.
29. Rundo, L.; Militello, C.; Russo, G.; Garufi, A.; Vitabile, S.; Gilardi, M.C.; Mauri, G. Automated Prostate Gland Segmentation Based on an Unsupervised Fuzzy C-Means Clustering Technique Using Multispectral T1w and T2w MR Imaging. *Information* **2017**, *8*, 49. [CrossRef]
30. Arif, M.; Schoots, I.G.; Castillo, T.J.M.; 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**, 1–11. [CrossRef]
31. Hoang Dinh, A.; Melodelima, C.; Souchon, R.; Lehaire, J.; Bratan, F.; Mège-Lechevallier, F.; Ruffion, A.; Crouzet, S.; Colombel, M.; Rouvière, O. Quantitative Analysis of Prostate Multiparametric MR Images for Detection of Aggressive Prostate Cancer in the Peripheral Zone: A Multiple Imager Study. *Radiology* **2016**, *280*, 117–127. [CrossRef] [PubMed]