applied sciences

MDPI

Article

Non-Invasive Estimation of Gleason Score by Semantic Segmentation and Regression Tasks Using a Three-Dimensional Convolutional Neural Network

Takaaki Yoshimura [1,2,3], Keisuke Manabe [4] and Hiroyuki Sugimori [1,3,*]

[1] Faculty of Health Sciences, Hokkaido University, Sapporo 060-0812, Japan
[2] Department of Medical Physics, Hokkaido University Hospital, Sapporo 060-8648, Japan
[3] Global Center for Biomedical Science and Engineering, Faculty of Medicine, Sapporo 060-8648, Japan
[4] Graduate School of Health Sciences, Hokkaido University, Sapporo 060-0812, Japan
* Correspondence: sugimori@hs.hokudai.ac.jp

Abstract: The Gleason score (GS) is essential in categorizing prostate cancer risk using biopsy. The aim of this study was to propose a two-class GS classification (< and ≥GS 7) methodology using a three-dimensional convolutional neural network with semantic segmentation to predict GS non-invasively using multiparametric magnetic resonance images (MRIs). Four training datasets of T2-weighted images and apparent diffusion coefficient maps with and without semantic segmentation were used as test images. All images and lesion information were selected from a training cohort of the Society of Photographic Instrumentation Engineers, the American Association of Physicists in Medicine, and the National Cancer Institute (SPIE–AAPM–NCI) PROSTATEx Challenge dataset. Precision, recall, overall accuracy and area under the receiver operating characteristics curve (AUROC) were calculated from this dataset, which comprises publicly available prostate MRIs. Our data revealed that the GS ≥ 7 precision (0.73 ± 0.13) and GS < 7 recall (0.82 ± 0.06) were significantly higher using semantic segmentation ($p < 0.05$). Moreover, the AUROC in segmentation volume was higher than that in normal volume (ADCmap: 0.70 ± 0.05 and 0.69 ± 0.08, and T2WI: 0.71 ± 0.07 and 0.63 ± 0.08, respectively). However, there were no significant differences in overall accuracy between the segmentation and normal volume. This study generated a diagnostic method for non-invasive GS estimation from MRIs.

Keywords: gleason score; classification; prostate cancer; semantic segmentation; three-dimensional convolutional neural network (3D-CNN)

Citation: Yoshimura, T.; Manabe, K.; Sugimori, H. Non-Invasive Estimation of Gleason Score by Semantic Segmentation and Regression Tasks Using a Three-Dimensional Convolutional Neural Network. *Appl. Sci.* **2023**, *13*, 8028. https://doi.org/10.3390/app13148028

Academic Editor: Jan Egger

Received: 17 May 2023
Revised: 7 July 2023
Accepted: 8 July 2023
Published: 9 July 2023

1. Introduction

Prostate cancer is the most commonly diagnosed cancer in men worldwide, the most frequently diagnosed cancer in 112 countries, and the second-ranked cause of mortality following lung cancer [1]. Treatment options for prostate cancer vary according to the tumor stage and grade, patient characteristics, and personal preferences. Previously, some guidelines were published based on the current evidence in prostate cancer diagnostics, such as the National Comprehensive Cancer Network (NCCN) guidelines and the European Association of Urology–European Society for Radiotherapy and Oncology–International Society of Geriatric Oncology guidelines [2,3]. According to the NCCN guidelines, prostate cancer has been categorized into six risk groups, from very low to very high, based on prostate-specific antigen (PSA), Gleason score (GS), and TNM classification [2]. PSA is a continuous parameter of prostate cancer. In the prostate cancer grading system, the biopsy GS consists of the Gleason grade of the most extensive pattern plus the highest pattern. In the 2014 International Society of Urological Pathology, prostate cancer was classified into five levels according to the GS, which highlighted the clinical differences between GS 7 = 3 + 4 and GS 7 = 4 + 3 [4]. For the staging of prostate cancer, the use of the

2017 TNM classification of the American Joint Committee on Cancer was recommended. Initial therapies, such as active surveillance, radiation therapy, radical prostatectomy, and androgen deprivation therapy, were selected based on the risk groups and expected patient survival. Thus, the risk classification of prostate cancer is important.

A prostate cancer diagnosis is made in three steps. The latest prostate cancer screening and clinical practice guidelines highly recommend PSA screening as the first step in the algorithm classifying prostate cancer risk based on the European Randomized Study of Screening for Prostate Cancer results [3,5–7]. The Japanese Urological Association guidelines provide a PSA cut-off value of 4.0 ng/mL for further urological examination, with an age-standardized PSA cut-off value of 3.0, 3.5, and 4.0 ng/mL for men aged 50–64, 65–69, and $\geq$70 years, respectively [6]. The second step in establishing a definitive diagnosis is a prostate biopsy, with the option of two invasive approaches that have similar detection rates. One is the transrectal or transperineal ultrasound-guided biopsy, and the other is the transrectal biopsy. A transperineal approach requires anesthesia, while the transrectal approach has a higher severe infection risk [8]. Current guidelines recommend performing multiparametric magnetic resonance imaging (MRI) before biopsies to discriminate patients with prostate cancer as indolent or clinically significant. The GS is the most commonly used histopathological grading system, based on the classification of five histological patterns underlying the presence of cancer cells in the specimen. The grade group was determined according to the GS [9]. Staging was performed according to the TNM classification system published by the Union for International Cancer Control. Multiparametric MRIs, comprising T2-weighted imaging (T2WI), dynamic contrast-enhanced imaging, and diffusion-weighted imaging (DWI), are highly reliable and were used for tumor (T) staging. Current multiparametric MRI diagnostic methods follow the Prostate Imaging Reporting and Data System (PI-RADS), which involves a semiquantitative radiologist assessment of each suspicious lesion, assigning a corresponding clinically significant prostate cancer likelihood score from one to five [10]. PI-RADS scores with other parameters, such as clinical variables, family history, or PSA levels, help radiologists determine whether further investigation is needed to make a final diagnosis. For N staging, abdominal computed tomography (CT) and MRI indirectly assess nodal invasion by using lymph node diameter and morphology. In addition to imaging findings, lymph node dissection remains the most reliable method for lymph node staging. For metastasis (M) staging, multimodal imaging techniques such as CT, MRI, positron emission tomography (PET), and 99mTc-bone scans are used.

Presently, the interpretation of multiparametric MRI data is entirely performed by radiologists. Although they are competent, it is difficult for them to deal with increasing imaging demands within a limited timeframe. Also, there is significant variability between observers since performance depends on their experience [11,12]. In a report by Kohestani et al., the inter-observer variability of prostate MRI using PI-RADS was investigated outside high-volume centers [13]. Several studies have been conducted on the automation of the whole or a part of the diagnosis workflow to improve diagnostic accuracy and reduce the costs and workload of healthcare personnel. In recent years, computer-aided diagnosis (CAD) has been actively investigated based on advances in deep learning technologies, such as convolutional neural networks (CNN), which have rapidly developed in the medical imaging field with advances in artificial intelligence (AI). Because CNNs are adept at discovering complex structures in high-dimensional data, they are powerful tools for image classification and segmentation [14–16]. For example, Ozsari et al. proposed a deep learning-based approach in order to automatically diagnose temporomandibular disorder on magnetic resonance (MR) images with seven different fine-tuned, pre-trained CNNs: Xception, ResNet-101, MobileNetV2, InceptionV3, DenseNet-121, ConvNeXt, and Vision Transformer (ViT) [17]. However, CNN image classification assesses the entire image and can rely on features other than the lesion area. Consequently, even if the accuracy of the classifier is high, its reliability is reduced if the judgment is based on areas other than the lesion. A three-dimensional CNN (3D-CNN), which extends the two-

dimensional CNN (2D-CNN), which is widely used in image recognition, is used for image classification and motion recognition of 3D data. Unlike 2D-CNN, 3D-CNN can extract 3D features [18]. Semantic segmentation is a deep learning algorithm that enables pixel-level image classification and can detect irregularly shaped objects.

A comprehensive assessment of multiparametric MRIs consists of eight or more different volumetric image datasets. This is a burden for the radiologists. Even for experienced radiologists, it is difficult to detect subtle cancerous lesions expressed within multiparametric MRIs. Moreover, it is important to correlate characterized cancerous lesions in multiparametric MRIs and biopsy findings with GS for non-invasive GS prediction. Based on the various quantitative image features, various deep-learning systems have been investigated for prostate cancer detection and diagnosis using multiparametric MRI with detection and classification tasks. There were only a small number of CAD systems for prostate cancer. In related work, Firjani et al. developed a DWI-based CAD system that utilized three intensity features and a K-nearest neighbor classifier to distinguish between benign and malignant cases [19]. Niaf et al. developed the multiparametric MRI-based CAD system to detect prostate cancer in the peripheral zone using T2WI, dynamic contrast-enhanced MRI and DWI [20]. Lotjens et al. developed the prostate-segmentation technique using a combination of features such as the apparent diffusion coefficient map (ADCmap). Kiraly et al. proposed the use of multichannel image-to-image convolutional encoder–decoders to directly determine tumor malignancy without performing an invasive prostate biopsy procedure [21]. Mercaldo et al. proposed an approach focused on the automatic GS classification, which exploited a set of 18 radiomic features directly obtainable from segmented MRI [22].

Today, the performance of AI-based CAD systems is comparable to that of experienced radiologists, owing to continuous technical developments and increased dataset quantity and quality [23,24]. Winkel et al. evaluated the agreement of the diagnostic accuracy of five PI-RADS lesions as the ground truth between human readers and fully automated AI-based software and demonstrated that the AI-based software was able to identify highly suspicious lesions in image-guided prostate cancer screening [23]. Saha et al. presented a CAD system for automated localization of clinically significant prostate cancer with multiparametric MRI and achieved a 0.882 area under the receiver operating characteristics curve (AUROC) in patient-based diagnosis [24]. For the GS classification, with the recent advancement of deep learning techniques, various GS prediction models have been proposed. Cao et al. proposed a novel multi-class CNN, FocalNet, that jointly detects prostate cancer lesions and predicts their GS from multiparametric MRI and calculated an AUROC of 0.81 and 0.79 for the classifications of clinically significant prostate cancer (GS = 3 + 4) and prostate cancer (GS = 4 + 3), respectively [25].

We hypothesized that if the GS is estimated from MRI with high accuracy by 3D-CNN using semantic segmentation to crop the image and restrict the evaluation range to the prostate, the grade group can be determined without a highly invasive prostate biopsy, and as a result, the physical burden on the patient can be reduced. Furthermore, there is significant variability among observers depending on their experience in processing and interpreting prostate cancer diagnostics [11,12,26]. By using the quantitative features from MRI, an AI-based GS estimation system may not only automate and support the radiologist's workflow but also alleviate the workload of medical staff. Therefore, this study aimed to construct a system to estimate GS non-invasively from diagnostic MRI images.

The present study is organized as follows: in Section 2, we describe MRI data and the technical framework for this study. Section 3 presents the results of precision, recall, overall accuracy and AUROC. In Section 4, we discuss the potential implications and extensions of this study, followed by concluding remarks.

2. Materials and Methods

2.1. Subjects

As we used images from a public database, no ethical approval was required for the implementation of the present study. A training cohort of 204 subjects was enrolled in this study, with available prostate MRI data at the cancer image archive from the Society of Photographic Instrumentation Engineers, American Association of Physicists in Medicine, and National Cancer Institute (SPIE–AAPM–NCI) PROSTATEx Challenge occurring from 21 November 2016 to 16 February 2017 [27]. This dataset was included in the Prostate Imaging-Cancer Artificial Intelligence (PI-CAI) Public Training and Development dataset established in conjunction with an international multidisciplinary scientific advisory board. The board consisted of 16 experts in prostate AI, radiology, and urology who curated prostate MRI examinations to validate modern AI algorithms and estimated the radiologists' performance in the detection and diagnosis of clinically significant prostate cancer [27–29].

This study included a public dataset of T2WI and ADCmap, which included the defined lesion information. The lesions were annotated with a GS of $\geq$7 or <7. Of the 204 patients in the database, 134 and 70 had a GS of <7 and $\geq$7, respectively.

2.2. Prostate Segmentation

A segmentation model was created for prostate region extraction using deeplabv3+, a CNN for semantic segmentation with T2WI, and published labeled images indicating the prostate regions in the same database. Table 1 shows the division of the training dataset into five subsets, labeled as subsets (a–e), for training purposes. The evaluation of the segmentation model was conducted through five-fold cross-validation, wherein the training dataset was further split into training and test data subsets (Figure 1).

Table 1. Subset data of segmentation for the five-fold cross-validation.

Number of Subjects (Images) per Subset					
a	**b**	**c**	**d**	**e**	**Total**
12 (1505)	12 (1393)	12 (1417)	12 (1243)	12 (1413)	60 (1876)

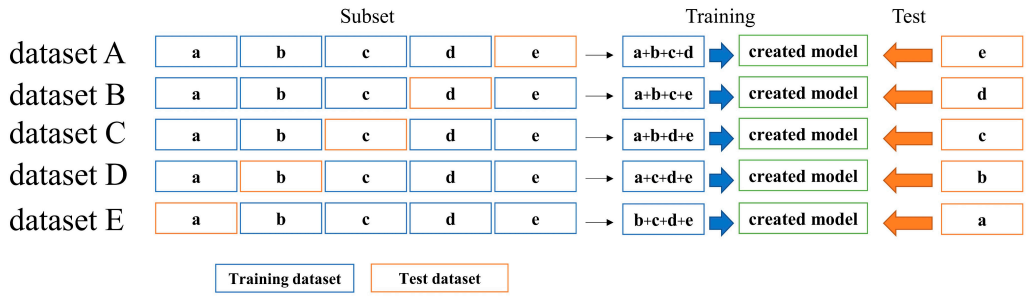

Figure 1. Dataset for training and testing.

To enhance the training process, data augmentation was applied to the training images. This involved rotating the images by ±10 degrees with increments of 5 degrees, resulting in an expanded dataset containing five times the original number of images.

The software for the deep learning technique was developed in-house using the MATLAB software version 2022b (The MathWorks, Inc., Natick, MA, USA) and a desktop computer with two NVIDIA RTX A6000 graphics cards (Nvidia Corporation, Santa Clara, CA, USA), with 38.7 TFlops of single-precision performance, 768 GB of memory bandwidth, and 48 GB of memory per board. A deeplabv3+ was used as a CNN for prostate segmentation. The input was 384 × 384, and the parameters were trained by loading the Neuroimaging Informatics Technology Initiative files. The optimizer was stochastic

gradient descent and momentum optimization. Regarding the training parameters, the batch size for the number of training samples was 128, the number of epochs was 10, and the initial learning rate was 0.001. The learning rate drop factor was 0.3, the learning rate drop period was 1, the L2 regularization was 0.005, and the momentum was 0.9.

The calculation of the Dice Similarity Coefficient (DSC) was performed using the following formula when the supervised images were designated as A and the predicted images were designated as B. The DSC serves as an index to assess the level of agreement between the images, with a value approaching 1 indicating a higher degree of agreement. The highest DSC model was used to extract the prostate after the next section.

$$DSC = \frac{2 \times |A \cap B|}{|A| + |B|}$$

2.3. Image Preprocessing

Images were acquired using two whole-body Siemens 3T MR scanners (MAGNETOM and Skyra). Image acquisition details have been previously reported [27]. This dataset was collected and curated for research on prostate MRI CAD at the Prostate MR Reference Center, Radboud University Medical Center, Nijmegen, Netherlands. T2WI was acquired using a turbo spin echo sequence and with a resolution of around 0.5 and 3.6 mm for plane and slice thickness, respectively. DWI was acquired using a single-shot echo planar imaging sequence comprising diffusion-encoding gradients in three directions with a resolution and slice thickness of 2.0 and 3.6 mm, respectively. Three b-values were acquired (50, 400, and 800), and scanner software calculated the ADCmap. All images were acquired without an endorectal coil.

The T2WI and ADCmap in the dataset had different fields of view, spatial image resolution, and slice numbers at imaging time. Each subject's slice number differed, with most ADCmaps having fewer slices than T2WI. Therefore, T2WI positions exceeding those in the ADCmap were deleted by referencing the ADCmap head and tail slice positional information retrieved from the digital imaging and communications in medicine tag information. The T2WI was processed using the prostate extraction obtained from the semantic segmentation model, resulting in an image containing only the prostate. The same coordinate information was used to generate an ADCmap, which also contained only the prostate. The ADCmap and T2WI with and without semantic segmentation were resized to 224 × 224 for 3D-CNN input. The image data were saved in Neuroimaging Informatics Technology Initiative format (Figure 2). A segmentation volume was defined as a file group with segmentation, while a normal volume was defined as one without segmentation.

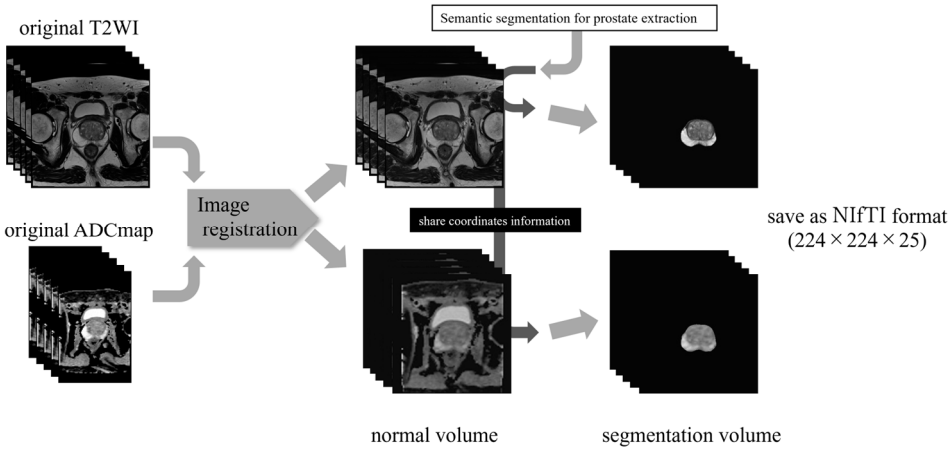

Figure 2. Schematics of image preprocessing. T2WI, T2-weighted images; ADCmap, apparent diffusion coefficient map; NIfTI, Neuroimaging Informatics Technology Initiative.

2.4. Network Architecture

The network architecture used in this study is shown in Figure 3. It is a network consisting of 177 layers in total. This network architecture is 3D-ResNet50, based on ResNet50 pre-trained by ImageNet, and has a structure in which the input size is changed to 224 × 224 × 25 to allow 25 image slices of 224 × 224 to be input in one input section [30]. This network consists of several layers, including 3D convolutional, activation, batch normalization, pooling, fully connected, and softmax layers. The model has 48 million learnable parameters. The final layer is the classification layer, which can output the classification results of multiple slices input in 3D using the features trained based on the GS in this study.

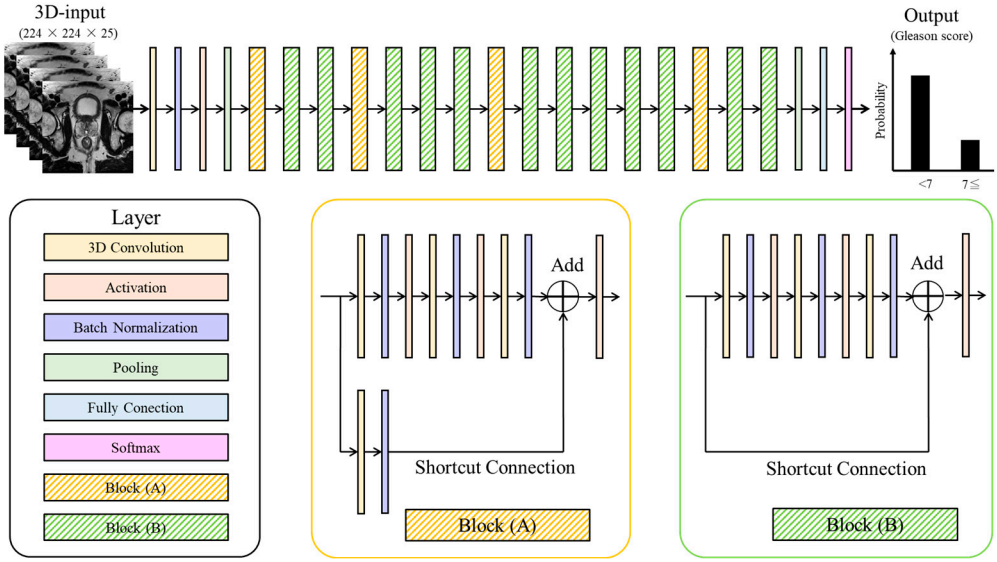

Figure 3. Network structure of 3D-ResNet50.

2.5. Experimental Setup

Table 2 demonstrates the subdivision of five subsets (a–e) for both GS ≥ 7 and <7 as the subject number differed between the two groups. The classifiers were evaluated by five-fold cross-validation using the training dataset, divided into training and test data (Figure 1). The dataset numbers for GS ≥ 7 were half compared to those <7. Therefore, the rotated datasets above GS 7 were doubled using left-right flipping to align the dataset numbers above and below GS 7.

Table 2. Subset data of classification for the five-fold cross-validation.

		Number of Subjects (Images) per Subset					
		a	**b**	**c**	**d**	**e**	**Total**
GS [a]	≥7	14 (350)	14 (350)	14 (350)	14 (350)	14 (350)	70 (1750)
	<7	27 (378)	27 (378)	27 (378)	27 (378)	26 (364)	134 (1876)

[a] Gleason score.

The deep learning technique was implemented using the same software environment as described in Section 2.2. A 3D-ResNet was used as a CNN for data classification into GS ≥ 7 and GS < 7. The input was changed to 224 × 224 × 25, and the parameters were trained by loading the Neuroimaging Informatics Technology Initiative files. The optimizer was stochastic gradient descent and momentum optimization. Regarding the training

parameters, the batch size for the number of training samples was 128, the number of epochs was 15, and the initial learning rate was 0.0001. The learning rate drop factor was 0.1, the learning rate drop period was 10, the L2 regularization was 0.0001, and the momentum was 0.9.

2.6. Evaluation of the Classifier and Statistics

Four training datasets comprising the T2WI and ADCmap with and without semantic segmentation were used to evaluate model performance. These datasets were applied to the test images, with precision, recall, and overall accuracy calculated using the following equations (Equations (1)–(3)).

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \quad (1)$$

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \quad (2)$$

$$Overall\ accuracy = \frac{True\ Positives + True\ Negatives}{True\ Positives + True\ Negatives + False\ Positives + False\ Negatives} \quad (3)$$

The AUROC was calculated as an additional indicator of the model's effectiveness; AUROC values ranged from 0.5 to 1, with larger values indicating the better generalization and prediction performance of the model.

The evaluation indices for each dataset were expressed as the mean ± standard deviation. The Wilcoxon signed-rank test was used for precision and recall with a significance level of less than 5% to determine the semantic segmentation effectiveness. Statistical analyses were performed using JMP Pro version 16.2.0 (SAS Institute Inc., Cary, NC, USA).

3. Results

Table 3 demonstrates the DSC for segmentation. The mean DSC of five folds was 0.7528. The segmentation model created in Fold 3 was used to extract the prostate.

Table 3. DSC for segmentation.

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean
DSC	0.6981	0.7462	0.7897	0.7421	0.7879	0.7528

Table 4 demonstrates the overall ADCmap and T2WI for segmentation and normal volume accuracy. There was no significant difference between overall segmentation accuracy and normal volumes or between T2WI and ADCmap images. Figure 4 shows the AUROC for each dataset.

Table 4. Overall accuracy and AUROC of the classification.

		Overall Accuracy (%)	AUROC [c]
Segmentation volume	ADCmap [a]	64.68 ± 3.97	0.70 ± 0.05
	T2WI [b]	66.64 ± 6.72	0.71 ± 0.07
Normal volume	ADCmap	69.60 ± 2.89	0.69 ± 0.08
	T2WI	64.72 ± 8.16	0.63 ± 0.08

[a] apparent diffusion coefficient map; [b] T2-weighted images; [c] area under the receiver operating characteristic curve.

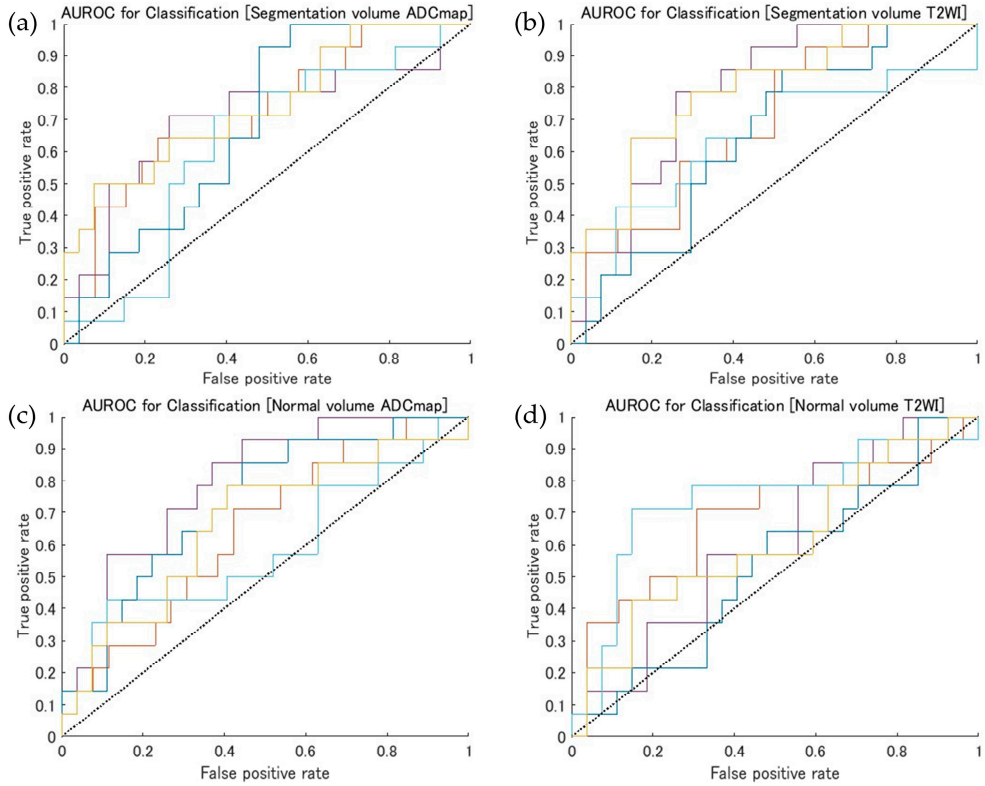

Figure 4. Area under the receiver operating characteristic curve (AUROC) for each dataset: dataset A (orange), dataset B (purple), dataset C (light blue), dataset D (blue) and dataset E (yellow). (**a**) ADCmap for segmentation volume; (**b**) T2WI for segmentation volume; (**c**) ADCmap for normal volume; (**d**) T2WI for normal volume.

Table 5 displays the precision and recall values for the ADCmap and T2WI in segmentation and normal volume. The ADCmap exhibited the highest values for precision for GS $\geq$ 7 and recall for GS < 7 in the segmentation volume. The ADCmap demonstrated the highest values for GS $\geq$ 7 and GS < 7 compared to the ADCmap and T2WI.

Table 5. Precision and recall for classification. GS, Gleason score; T2WI, T2-weighted images; ADCmap, apparent diffusion coefficient map diffusion.

		Precision		**Recall**	
		GS $\geq$ 7	**GS < 7**	**GS $\geq$ 7**	**GS < 7**
Segmentation volume	ADCmap	0.73 ± 0.13	0.60 ± 0.11	0.50 ± 0.04	0.82 ± 0.06
	T2WI	0.57 ± 0.13	0.72 ± 0.13	0.53 ± 0.09	0.76 ± 0.04
Normal volume	ADCmap	0.34 ± 0.06	0.88 ± 0.04	0.61 ± 0.07	0.72 ± 0.02
	T2WI	0.49 ± 0.12	0.73 ± 0.11	0.50 ± 0.12	0.73 ± 0.05
p-value		0.0015	0.0137	0.2097	0.008

4. Discussion

The AI-based CAD system is drawing significant attention due to its potential to revolutionize the diagnostic workflow, improve diagnostic accuracy among observers, reduce costs, and decrease the workload of healthcare personnel. The role of AI is not merely to replicate human cognitive processes but to enhance and augment them, facilitating a higher level of diagnostic precision that could lead to more effective treatment plans. This shift toward AI-centered diagnostic methods could also redefine the role of healthcare

professionals, with an increased emphasis on interpreting and implementing AI-generated data in patient care.

The present study proposed a non-invasive GS classification method, classifying GS as either ≥7 or <7, utilizing semantic segmentation and a 3D-CNN. This approach aimed to address the challenges and limitations associated with current diagnostic methods. In prostate outcome studies, GS and its associated Gleason Grade Group have consistently remained the most powerful prognostic predictors, often influencing treatment decisions. However, the GS is currently determined based on pathological diagnoses that require invasive biopsies and has been observed to demonstrate poor reproducibility.

There is an inherent discordance between the inter- and intra-observer variability of pathologists, primarily due to the subjectivity of the GS. This variability can lead to discrepancies in diagnosis and treatment, potentially impacting patient outcomes. This observation underscores the need for more objective and reproducible methods of GS determination. In response to this, Nagpal et al. proposed a deep learning system for GSs of whole-slide prostatectomy images [31]. Their results revealed a significantly higher diagnostic accuracy in deep learning (0.70), which showed a trend toward improved patient risk stratification in correlation with clinical follow-up data.

The potential of deep learning system applications to improve the accuracy of GSs without observer dependence is compelling. The ability of such systems to analyze vast amounts of data and identify subtle patterns that may be missed by human observers could revolutionize prostate cancer diagnosis. However, the implementation of these systems is not without challenges and considerations. The integration of AI into clinical practice must be performed thoughtfully, with a clear understanding of its potential impact on patient care, professional roles, and healthcare systems more broadly.

Several groups have attempted to use AI-based approaches for non-invasive GS estimation of prostate cancer. Cao et al. proposed multi-class CNN (FocalNet) to jointly detect prostate cancer lesions and predict their aggressiveness using GS and evaluated the GS classification by AUROC for clinically significant lesions (GS 7) (AUROC = 0.81 ± 0.01) in comparison to U-Net-Mult and Deeplab (AUROC = 0.72 ± 0.01 and 0.71 ± 0.02, respectively) [25]. Since they used a publicly unavailable dataset, it is difficult to compare their study directly with ours. In the previously reported grand challenges by Armato et al. using a publicly available dataset, the AUROC for the PROSTATEx Challenge task for differentiating between lesions that are and are not clinically significant ranged from 0.45 to 0.87 [32]. Although the AUROC in our results was not state-of-the-art, the AUROC in segmentation volume was higher than that in normal volume (Table 2). As shown in our result in Figure 4, all datasets outperformed random guessing (AUROC = 0.5). Moreover, the PROSTATEx-2 Challenges demonstrated the five-point Gleason Grade Group classification task. For further reduction of unnecessary biopsies, we will focus on the discrimination of the five-point Gleason Grade Group classification.

This study has several limitations. First, the prostate volume on the MRI was small in the overall image. The same image matrix size was used when performing 3D-CNN, regardless of prostate semantic segmentation use. This suggests that the amount of relevant data available for AI to analyze was limited, potentially affecting learning accuracy. This implies that the AI system may require more extensive data to improve its performance, which could involve utilizing larger MRI images or incorporating additional imaging modalities.

Secondly, the training data used in this study were scarce, with only 134 and 70 patients with GS < 7 and ≥7, respectively, available in the public database. This limitation emphasizes the need for larger, more diverse datasets to train the AI system efficiently. The NCCN prostate cancer guidelines currently select the initial therapy based on initial risk stratification and staging workup for clinically localized diseases. The risk group is categorized according to clinical or pathological features, which are characterized by the Gleason pattern or Gleason Grade Group. Based on the GS division, these are very low, low, favorable, unfavorable intermediate, high, and very high-risk groups.

The Gleason Grade Group was categorized into five groups based on combinations of GS and Gleason patterns, including $\leq$6 and $\leq$3 + 3; 7 and 3 + 4; 7 and 4 + 3; 8 and 4 + 4, 3 + 5, 5 + 3; 9 or 10 and 4 + 5, 5 + 4, or 5 + 5. However, the SPIE–AAPM–NCI PROSTATEx Challenge dataset training cohort was labeled as GS $\geq$ 7 for clinically significant prostate cancer. Therefore, the proposed method represents a two-class classification of a GS $\geq$ 7 or GS < 7 using semantic segmentation and 3D-CNN.

This highlights the need for more nuanced classification systems, which could potentially be achieved through the use of larger and more diverse datasets. Moreover, further research is necessary to utilize more detailed GS and Gleason pattern data for a five-class classification, which could offer a more precise and personalized approach to prostate cancer diagnosis and treatment.

Third, there is a shortage of publicly available, high-quality prostate MRI datasets. The SPIE–AAPM–NCI PROSTATEx Challenge is a popular and publicly available prostate MRI dataset that was used in this study [29]. However, despite its focus on quantitative image analysis methods for diagnostic, clinically significant prostate cancer classification, and prostate AI, radiology, and urology expert involvement, this study contained a small, single-center, and multivendor dataset.

The need for well-curated, larger datasets with diverse, multi-center, and multivendor data is clear. These datasets are essential for training the AI-based CAD system effectively and ensuring meaningful validation, robust performance, and a generalized model [33]. Unfortunately, most public datasets are too small, and the quality of annotations provided per dataset varies significantly.

Sunoqrot et al. reviewed 17 public prostate MRIs [34]. The SPIE–AAPM–NCI PROSTATEx Challenge, including the SPIE–AAPM–NCI PROSTATEx-2 Challenges running from 15 May 2017, to 23 June 2017, presents additional difficulties, such as the focus on quantitative multiparametric MRI biomarker development to determine the Gleason Grade Group in prostate cancer [27–29]. As of 5 May 2022, the PI-CAI challenge has publicly released 1500 anonymized multiparametric prostate MRIs from 1476 patients at multi-centers in The Netherlands, acquired with a multivendor between 2012 and 2021 [35]. This challenge aimed to validate the diagnostic performance of AI and radiologists at clinically significant prostate cancer detection or diagnosis in MRI with histopathology and over three years of follow-up as the reference standard. Therefore, datasets should be added cautiously, ensuring that they provide unique and high-quality data. Another primary limitation of our study is that we have not delved into the investigation of hyperparameters in deep learning model training. In machine learning, and more specifically in deep learning, hyperparameters are crucial elements that can significantly influence the accuracy, efficiency, and computational cost of models. These variables, set before the model training process begins, control the overall behavior of a learning algorithm and can substantially affect the performance of the model. Although many public prostate MRI data may be available, it is difficult to consolidate multiple public datasets using the definition and annotation quality provided per dataset with missing information across images and cohort distributions. Also, data overlap in public datasets should be avoided, as multiple public datasets may contain identical cases [34]. Sunoqrot et al. reviewed data overlap as an example of how The National Cancer Institute's Cancer Imaging Program, in collaboration with the International Society for Biomedical Imaging (NCI–ISBI 2013) dataset [36], combined the Prostate-3T [37] and PROSTATE-DIAGNOSIS dataset [38], and the PROSTATEx [27] and Prostate-3T datasets were included in the PI-CAI dataset [34,35]. Therefore, datasets should be added cautiously.

Although this study was ensured to use well-established, robust, and commonly accepted values for hyperparameters in our models, an exhaustive hyperparameter tuning process was not implemented. There could potentially exist a different set of hyperparameters that could yield better performance or more accurate results for the same models applied to the same dataset. In addition, further study of the segmentation model for prostate extraction is needed to determine changes in extraction accuracy and classification

results. It is important to consider common factors between segmentation models and classification models, as well as the extent of obtaining large-scale data, to ensure accuracy. To address this, we believe it is crucial to thoroughly investigate previous research papers and refer to them in order to understand how much data is needed to guarantee accuracy. It is also important to train models using not only publicly available databases but also real-world datasets that can be obtained through ethical approval. This approach will contribute to refining and enhancing the performance of the models.

Despite the limitations of the current study, its findings underline the potential of AI-based systems for advancing prostate cancer diagnosis. There is a clear need for more detailed classification systems in the clinical guidelines for initial therapy, and further research is required to non-invasively estimate the GS using multiparametric MRI.

Public prostate MRI datasets are often small and feature different cohorts and annotation qualities, underscoring the need for larger and more diverse datasets. More studies are required to provide independent validation, build trust in non-invasive GS prostate cancer prediction, and ultimately enhance the effectiveness of AI-based CAD systems in improving diagnostic accuracy and patient outcomes. The future of AI in healthcare is promising, with its potential to revolutionize diagnostic processes, improve patient outcomes, and reshape healthcare systems. However, realizing this potential requires concerted and collaborative efforts to overcome the existing challenges and limitations.

In future work, we plan to improve the network models and evaluate them with external validation data. In this study, we used 3D-ResNet as a CNN for data classification into GS < 7 and ≥ 7. Today, new models are constantly being developed. For example, ViT, an architecture with a transformer that was originally used predominantly in the natural language processing field, has now started to find a place in the field of medical image analysis for the segmentation and classification of medical images [39]. Utilizing such new models is expected to improve GS estimation accuracy. Also, we used five-fold cross-validation using the training dataset divided into training and test data from public data. Using the external validation data, which completely isolated patient cohorts from training datasets, will achieve an assessment of external validity.

5. Conclusions

The objective of this study was to investigate the possibility that a highly accurate estimation of GS from MRI images using 3D-CNN, which uses semantic segmentation to isolate images and limit the evaluation range to the prostate, would allow grade group determination without invasive prostate biopsy, thereby reducing the physical burden on the patient. Based on this hypothesis, this study aimed to construct a system to estimate GS noninvasively from diagnostic MRI images. A methodology for GS classification was proposed using a 3D-CNN with semantic segmentation. The precision for GS ≥ 7 and recall for GS < 7 were significantly higher using semantic segmentation. Using quantitative MRI features, an AI-based GS estimation system may not only automate and support the workflow of radiologists but also reduce the workload of medical staff.

Author Contributions: T.Y. contributed to the data analysis, algorithm construction, writing, and editing of the manuscript. K.M. contributed to the data analysis and writing of the manuscript. H.S. proposed the idea, contributed to the data acquisition, performed supervision and project administration, and reviewed and edited the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported in part by the Japan Society for the Promotion of Science (JSPS), KAKENHI (grant number: JP22K15797), and Grants-in-Aid for the Regional R&D Proposal-Based Program from the Northern Advancement Center for Science & Technology of Hokkaido, Japan (grant number: T-1-8).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The models created in this study are available on request from the corresponding author. The source code of this study is available at https://github.com/MIA-laboratory/ProstateGSpred (accessed on 16 May 2023).

Acknowledgments: The authors would like to thank the laboratory members of the Medical Image Analysis Laboratory and the Yoshimura Laboratory for their help.

Conflicts of Interest: The authors declare that they have no competing financial interests. All authors certify that they have no affiliations with or involvement in any organization or entity with any financial or non-financial interest in the subject matter or materials discussed in this manuscript.

References

1. Sung, H.; Ferlay, J.; Siegel, R.L.; Laversanne, M.; Soerjomataram, I.; Jemal, A.; Bray, F. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J. Clin.* **2021**, *71*, 209–249. [CrossRef]
2. National Comprehensive Cancer Network (NCCN). *NCCN Clinical Practice Guidelines in Oncology (NCCN Guidelines): Prostate Cancer*; Version 1. 2023; NCNN: Cleveland, OH, USA, 2022.
3. Mottet, N.; van den Bergh, R.C.N.; Briers, E.; Van den Broeck, T.; Cumberbatch, M.G.; De Santis, M.; Fanti, S.; Fossati, N.; Gandaglia, G.; Gillessen, S.; et al. EAU-EANM-ESTRO-ESUR-SIOG Guidelines on Prostate Cancer-2020 Update. Part 1: Screening, Diagnosis, and Local Treatment with Curative Intent. *Eur. Urol.* **2021**, *79*, 243–262. [CrossRef] [PubMed]
4. Epstein, J.I.; Egevad, L.; Amin, M.B.; Delahunt, B.; Srigley, J.R.; Humphrey, P.A.; Grading Committee. 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]
5. Kakehi, Y.; Sugimoto, M.; Taoka, R.; Committee for Establishment of the Evidenced-Based Clinical Practice Guideline for Prostate Cancer of the Japanese Urological Association. Evidenced-based clinical practice guideline for prostate cancer (summary: Japanese Urological Association, 2016 edition). *Int. J. Urol.* **2017**, *24*, 648–666. [CrossRef] [PubMed]
6. Ito, K.; Oki, R.; Sekine, Y.; Arai, S.; Miyazawa, Y.; Shibata, Y.; Suzuki, K.; Kurosawa, I. Screening for prostate cancer: History, evidence, controversies and future perspectives toward individualized screening. *Int. J. Urol.* **2019**, *26*, 956–970. [CrossRef] [PubMed]
7. Schröder, F.H.; Hugosson, J.; Roobol, M.J.; Tammela, T.L.; Zappa, M.; Nelen, V.; Kwiatkowski, M.; Lujan, M.; Määttänen, L.; Lilja, H.; et al. Screening and prostate cancer mortality: Results of the European Randomised Study of Screening for Prostate Cancer (ERSPC) at 13 years of follow-up. *Lancet* **2014**, *384*, 2027–2035. [CrossRef]
8. Shen, P.F.; Zhu, Y.C.; Wei, W.R.; Li, Y.Z.; Yang, J.; Li, Y.T.; Li, D.M.; Wang, J.; Zeng, H. The results of transperineal versus transrectal prostate biopsy: A systematic review and meta-analysis. *Asian J. Androl.* **2012**, *14*, 310–315. [CrossRef]
9. Humphrey, P.A.; Moch, H.; Cubilla, A.L.; Ulbright, T.M.; Reuter, V.E. The 2016 WHO Classification of Tumours of the Urinary System and Male Genital Organs-Part B: Prostate and Bladder Tumours. *Eur. Urol.* **2016**, *70*, 106–119. [CrossRef]
10. Barrett, T.; Rajesh, A.; Rosenkrantz, A.B.; Choyke, P.L.; Turkbey, B. PI-RADS version 2.1: One small step for prostate MRI. *Clin. Radiol.* **2019**, *74*, 841–852. [CrossRef]
11. Rosenkrantz, A.B.; Ayoola, A.; Hoffman, D.; Khasgiwala, A.; Prabhu, V.; Smereka, P.; Somberg, M.; Taneja, S.S. The Learning Curve in Prostate MRI Interpretation: Self-Directed Learning Versus Continual Reader Feedback. *Am. J. Roentgenol.* **2017**, *208*, W92–W100. [CrossRef]
12. Greer, M.D.; Shih, J.H.; Lay, N.; Barrett, T.; Bittencourt, L.; Borofsky, S.; Kabakus, I.; Law, Y.M.; Marko, J.; Shebel, H.; et al. Interreader Variability of Prostate Imaging Reporting and Data System Version 2 in Detecting and Assessing Prostate Cancer Lesions at Prostate MRI. *Am. J. Roentgenol.* **2019**, *212*, 1197–1205. [CrossRef] [PubMed]
13. Kohestani, K.; Wallström, J.; Dehlfors, N.; Sponga, O.M.; Månsson, M.; Josefsson, A.; Carlsson, S.; Hellström, M.; Hugosson, J. Performance and inter-observer variability of prostate MRI (PI-RADS version 2) outside high-volume centres. *Scand J. Urol.* **2019**, *53*, 304–311. [CrossRef]
14. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural. Inf. Process. Syst.* **2012**, *25*, 1097–1105. [CrossRef]
15. Shelhamer, E.; Long, J.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
16. Li, S.; Xiong, H.; Diao, X. Pre-Impact Fall Detection Using 3D Convolutional Neural Network. *IEEE Int. Conf. Rehabil. Robot.* **2019**, *2019*, 1173–1178. [CrossRef] [PubMed]
17. Ozsari, S.; Yapicioglu, F.R.; Yilmaz, D.; Kamburoglu, K.; Guzel, M.S.; Bostanci, G.E.; Acici, K.; Asuroglu, T. Interpretation of Magnetic Resonance Images of Temporomandibular Joint Disorders by Using Deep Learning. *IEEE Access* **2023**, *11*, 49102–49113. [CrossRef]
18. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning Spatiotemporal Features with 3D Convolutional Networks. *arXiv* **2014**. [CrossRef]
19. Firjani, A.; Elnakib, A.; Khalifa, F.; Gimel'farb, G.; El-Ghar, M.A.; Elmaghraby, A.; El-Baz, A. A diffusion-weighted imaging based diagnostic system for early detection of prostate cancer. *J. Biomed. Sci. Eng.* **2013**, *06*, 346–356. [CrossRef]

20. Niaf, E.; Rouvière, O.; Mège-Lechevallier, F.; Bratan, F.; Lartizien, C. Computer-aided diagnosis of prostate cancer in the peripheral zone using multiparametric MRI. *Phys. Med. Biol.* **2012**, *57*, 3833–3851. [CrossRef]
21. Kiraly, A.P.; Nader, C.A.; Tuysuzoglu, A.; Grimm, R.; Kiefer, B.; El-Zehiry, N.; Kamen, A. Deep Convolutional Encoder-Decoders for Prostate Cancer Detection and Classification. In Proceedings of the Medical Image Computing and Computer Assisted Intervention−MICCAI 2017, Cham, Switzerland, 11–13 September 2017; pp. 489–497.
22. Mercaldo, F.; Brunese, M.C.; Merolla, F.; Rocca, A.; Zappia, M.; Santone, A. Prostate Gleason Score Detection by Calibrated Machine Learning Classification through Radiomic Features. *Appl. Sci.* **2022**, *12*, 11900. [CrossRef]
23. Winkel, D.J.; Wetterauer, C.; Matthias, M.O.; Lou, B.; Shi, B.; Kamen, A.; Comaniciu, D.; Seifert, H.H.; Rentsch, C.A.; Boll, D.T. Autonomous Detection and Classification of PI-RADS Lesions in an MRI Screening Population Incorporating Multicenter-Labeled Deep Learning and Biparametric Imaging: Proof of Concept. *Diagnostics* **2020**, *10*, 951. [CrossRef]
24. Saha, A.; Hosseinzadeh, M.; Huisman, H. End-to-end prostate cancer detection in bpMRI via 3D CNNs: Effects of attention mechanisms, clinical priori and decoupled false positive reduction. *Med. Image Anal.* **2021**, *73*, 102155. [CrossRef]
25. Cao, R.; Mohammadian Bajgiran, A.; Afshari Mirak, S.; Shakeri, S.; Zhong, X.; Enzmann, D.; Raman, S.; Sung, K. Joint Prostate Cancer Detection and Gleason Score Prediction in mp-MRI via FocalNet. *IEEE Trans. Med. Imaging* **2019**, *38*, 2496–2506. [CrossRef]
26. Gatti, M.; Faletti, R.; Calleris, G.; Giglio, J.; Berzovini, C.; Gentile, F.; Marra, G.; Misischi, F.; Molinaro, L.; Bergamasco, L.; et al. Prostate cancer detection with biparametric magnetic resonance imaging (bpMRI) by readers with different experience: Performance and comparison with multiparametric (mpMRI). *Abdom. Radiol.* **2019**, *44*, 1883–1893. [CrossRef] [PubMed]
27. Litjens, G.D.O.; Barentsz, J.; Karssemeijer, N.; Huisman, H. ProstateX Challenge data. *Cancer Imaging Arch.* **2017**. [CrossRef]
28. 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*, 1045–1057. [CrossRef]
29. Litjens, G.; Debats, O.; Barentsz, J.; Karssemeijer, N.; Huisman, H. Computer-aided detection of prostate cancer in MRI. *IEEE Trans. Med. Imaging* **2014**, *33*, 1083–1092. [CrossRef] [PubMed]
30. Ebrahimi, A.; Luo, S. Alzheimer's Disease Neuroimaging Initiative. Convolutional neural networks for Alzheimer's disease detection on MRI images. *J. Med. Imaging* **2021**, *8*, 024503. [CrossRef]
31. Nagpal, K.; Foote, D.; Liu, Y.; Chen, P.C.; Wulczyn, E.; Tan, F.; Olson, N.; Smith, J.L.; Mohtashamian, A.; Wren, J.H.; et al. Development and validation of a deep learning algorithm for improving Gleason scoring of prostate cancer. *NPJ Digit. Med.* **2019**, *2*, 48. [CrossRef]
32. Armato, S.G., 3rd; 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*, 044501. [CrossRef]
33. Hosseinzadeh, M.; Saha, A.; Brand, P.; Slootweg, I.; de Rooij, M.; Huisman, H. Deep learning-assisted prostate cancer detection on bi-parametric MRI: Minimum training data size requirements and effect of prior knowledge. *Eur. Radiol.* **2022**, *32*, 2224–2234. [CrossRef]
34. Sunoqrot, M.R.S.; Saha, A.; Hosseinzadeh, M.; Elschot, M.; Huisman, H. Artificial intelligence for prostate MRI: Open datasets, available applications, and grand challenges. *Eur. Radiol. Exp.* **2022**, *6*, 35. [CrossRef] [PubMed]
35. Saha, A.; Twilt, J.J.; Bosma, J.S.; van Ginneken, B.; Yakar, D.; Elschot, M.; Veltman, J.; Fütterer, J.; de Rooij, M.; Huisman, H. The PI-CAI Challenge: Public Training and Development Dataset (v2.0). 2022. Available online: https://doi.org/10.5281/zenodo.6624726 (accessed on 7 July 2023).
36. Bloch, N.M.A.; Huisman, H.; Freymann, J.; Kirby, J.; Grauer, M.; Enquobahrie, A.; Jaffe, C.; Clarke, L.; Farahani, K. NCI-ISBI 2013 Challenge: Automated Segmentation of Prostate Structures. *Cancer Imaging Arch.* **2015**. [CrossRef]
37. Litjens, G. Futterer, Jurgen, Huisman, Henkjan. Data From Prostate-3T. *Cancer Imaging Arch.* **2015**. [CrossRef]
38. Bloch, B.N.; Jain, A.; Jaffe, C.C. Data from PROSTATE-DIAGNOSIS [Dataset]. *Cancer Imaging Arch.* **2015**. [CrossRef]
39. Valanarasu, J.M.J.; Oza, P.; Hacihaliloglu, I.; Patel, V.M. Medical Transformer: Gated Axial-Attention for Medical Image Segmentation. *arXiv* **2021**, arXiv:2102.10662. [CrossRef]