applied sciences

MDPI

Article

Enhancing Reliable Prostate Lesion Detection: Integrating Multi-Expert Annotations and Tailored nnU-Net Ensemble Learning Strategies

Rafal Jozwiak [1,2,*], Michal Gonet [1], Jan Mycka [1], Ihor Mykhalevych [1], Dariusz S. Radomski [3,4], Krzysztof Tupikowski [5,6], Tomasz Lorenc [7], Joanna Dolowy [8] and Anna Zacharzewska-Gondek [9]

1 Innovation Centre for Digital Medicine, National Information Processing Institute, 00-608 Warsaw, Poland
2 Faculty of Mathematics and Information Science, Warsaw University of Technology, 00-662 Warsaw, Poland
3 Institute of Radioelectronics and Multimedia Techniques, Warsaw University of Technology, 00-665 Warsaw, Poland; dariusz.radomski@pw.edu.pl
4 RadDarMed, 00-653 Warsaw, Poland
5 Subdivision of Urology, Lower Silesian Oncology, Pulmonology and Hematology Center, 53-413 Wroclaw, Poland
6 Departament of Oncologic Urology, Medical Faculty, Wrocław University of Science and Technology, 58-376 Wroclaw, Poland
7 1st Department of Clinical Radiology, Medical University of Warsaw, 02-004 Warsaw, Poland
8 Department of Radiology and Diagnostic Imaging, Lower Silesian Oncology, Pulmonology and Hematology Center, 53-413 Wroclaw, Poland
9 Independent Researcher, 52-234 Wroclaw, Poland
* Correspondence: rafal.jozwiak@opi.org.pl

Abstract

Accurate detection of prostate cancer suspicious areas in biparametric MRI (bpMRI) remains challenging because of severe lesion-to-background imbalance, limited lesion contrast, and inter-reader variability in lesion delineation. Unlike prior approaches that collapse inter-reader disagreement into a single consensus label, this study makes three contributions: (1) an adapted nnU-Net framework with prostate-centered preprocessing to reduce voxel-level class imbalance; (2) a class-imbalance-aware composite loss combining Dice, binary cross-entropy, and tailored focal loss to improve sensitivity to small and low-contrast lesions; and (3) a multi-expert learning strategy that preserves reader-specific annotations as separate supervision targets and aggregates predictions at the ensemble level. The method was developed on a single-center dataset of 378 bpMRI studies independently annotated by three board-certified radiologists. Of these, 323 studies were used for model development with patient-level 5-fold cross-validation, and 55 studies were reserved as a fixed independent test set. Compared with our previously published U-Net baseline, the proposed consensus-based nnU-Net improved Average Precision (AP) from 0.69 to 0.75, AUROC from 0.92 to 0.96, and the PI-CAI score from 0.81 to 0.85 on the independent test set. In addition, the multi-expert approach further improved AP to 0.81 versus 0.76 (+6.6%, $p < 0.01$), AUROC to 0.99 versus 0.95 (+4.2%, $p < 0.01$), and the PI-CAI score to 0.90 versus 0.86 (+4.7%). These findings demonstrate that explicitly preserving expert disagreement as a training signal, combined with anatomically targeted preprocessing and tailored loss design, substantially improves prostate lesion detection in bpMRI, providing a strong basis for future multi center external validation.

Check for updates

Academic Editor: Thomas Lindner

Received: 20 February 2026
Revised: 8 April 2026
Accepted: 15 April 2026
Published: 18 April 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

Keywords: prostate cancer detection; multi-parametric MRI (mpMRI); nnU-Net; multi-expert annotations; lesion detection; radiology AI; deep learning

1. Introduction

Prostate cancer is the most common malignancy in men worldwide and remains a major cause of morbidity and mortality [1,2]. Multi-parametric magnetic resonance imaging (mpMRI) is currently the preferred non-invasive method to detect and localize clinically significant prostate cancer [3]. It combines anatomical and functional information from T2-weighted imaging, diffusion-weighted imaging (DWI), apparent diffusion coefficient (ADC) maps, and, in the full clinical mpMRI protocol, dynamic contrast-enhanced imaging. However, mpMRI interpretation is time-consuming, operator-dependent, and subject to inter-reader variability, particularly in delineating small or low-contrast lesions [4–6]. These limitations affect diagnostic reproducibility, reducing the consistency of Prostate Imaging Reporting and Data System (PI-RADS) assessments between institutions. Reliable and standardized lesion detection is particularly important, as it can help reduce unnecessary biopsies, stabilize active surveillance and watchful waiting, and improve patient care. In addition, these limitations complicate the development of robust training datasets for artificial intelligence (AI) models, because lesion conspicuity and boundary delineation may vary substantially between experts. By enhancing reproducibility, AI-based tools may also help bridge the gap between expert and non-expert readers, thereby enabling broader access to high-quality prostate cancer diagnostics. An AI-assisted detection model has the potential to mitigate these inequities by offering consistent performance across different settings, supporting both high-volume centers and resource-limited clinics. At the same time, bi-parametric MRI (bpMRI), based on T2-weighted imaging, DWI, and ADC without DCE, has emerged as a practically attractive input configuration for AI development; accordingly, in the present study, model development and evaluation were restricted to the bpMRI subset of sequences [7].

1.1. Related Work

Deep learning approaches have shown promise in automating prostate image analysis, with convolutional neural networks achieving strong results in prostate gland segmentation and lesion detection [8–13]. Bardis et al. [9] and Aldoj et al. [10] demonstrated that U-Net-based architectures can achieve Dice coefficients that exceed 0.90 for whole-gland segmentation. However, the performance of lesion segmentation remains more modest, with AP scores typically below 0.70 [11,12,14], underscoring the difficulty of detecting small, low-contrast lesions associated with clinically significant disease. The introduction of nnU-Net [8] changed the focus from manual network engineering to automated data-driven configuration. nnU-Net dynamically adjusts preprocessing, patch sizes, network depth, and hyperparameters to the dataset and has achieved state-of-the-art performance in many medical image segmentation challenges, including prostate imaging [12]. However, most applications rely on single-expert annotations and standard loss functions, leaving variability and class imbalance insufficiently addressed.

Recently, the use of multiple expert annotations has emerged as a powerful strategy to improve robustness and generalization [14–18]. Rather than treating inter-expert disagreement as noise, these approaches incorporate it as a valuable signal reflecting diagnostic uncertainty. In prostate MRI, differences between expert annotations may reflect clinically plausible variability in lesion conspicuity and boundary definition rather than random labeling error alone. Preserving this variability during training may

therefore reduce overfitting to the contouring style of a single reader and promote lesion representations that are more stable across experts. Recent studies have also expanded the field beyond conventional supervised convolutional pipelines. Report-guided semi-supervised learning has shown that annotation-efficient training can improve lesion detection in bpMRI, while interactive and large-scale validation studies have demonstrated clinically meaningful AI performance in prostate MRI interpretation [7,14,19–21]. At the same time, recent literature has emphasized that robustness across scanners and institutions, as well as transparent development and reporting standards, remain essential for successful clinical translation [22–24].

In our preliminary research [25], we showed that the incorporation of annotations from various domain specialists into U-Net architectures improves model generalization and reduces the impact of inter-annotator variability, resulting in an average precision (AP) of 0.76 and an area under the receiver operating characteristic curve (AUROC) of 0.95. However, that earlier method relied on a standard U-Net, lacked data-driven adaptation, and used a basic preprocessing pipeline. Moreover, inter-expert variability was not yet operationalized as a dedicated multi-expert ensemble learning strategy within an nnU-Net framework. In this study, we aim to systematically improve prostate lesion detection by implementing targeted modifications to the nnU-Net framework.

1.2. Aim of This Work

This paper presents an adapted nnU-Net-based framework for prostate lesion detection, with the primary aim of comparing two training strategies and evaluating whether a multi-reader annotation-based approach improves performance relative to training based on a single aggregated reference label. In the proposed strategy, reader-specific annotations are preserved as separate supervision targets and combined through ensemble-based prediction aggregation, thereby treating inter-reader variability as inherent label uncertainty rather than simple annotation noise.

The framework additionally incorporates prostate-centered preprocessing to increase anatomical focus and a composite loss formulation to address severe lesion-to-background imbalance in prostate MRI. These elements were included to support the overall detection pipeline and were motivated by prior literature and task-specific considerations, but the study was not designed to isolate their individual effects in a dedicated ablation analysis.

Building on our previous U-Net-based study [25], the present work primarily aims to determine whether multi-reader supervision can improve the robustness and reliability of prostate lesion detection within the nnU-Net framework, while providing a basis for future external multicenter validation toward clinical translation [22,24].

2. Materials and Methods

2.1. Study Sample

This single-center study, which required informed consent for data analysis, was approved by the Bioethics Committee of the Hirszfeld Institute of Immunology and Experimental Therapy (KB 4/2022). The dataset, collected as part of the INFOSTRATEG research project I/0036/2021 "AI augmented radiology detection, reporting and clinical decision making in prostate cancer diagnosis", represents a heterogeneous clinical population, including healthy controls, negative mpMRI studies, and biopsy-confirmed cancers with varying Gleason grades.

2.2. Image Acquisition

The original clinical examinations were acquired as multi-parametric MRI including T2-weighted, diffusion-weighted imaging with multiple b-values, dynamic contrast-enhanced imaging and apparent diffusion coefficient. All studies in the present cohort were acquired at a single institution on the same scanner platform (1.5 T Siemens MAGNETOM Avanto Fit, Siemens Healthcare, Erlangen, Germany); therefore, scanner distribution was homogeneous across the dataset. MRI examinations were performed using a standardized acquisition protocol in the axial plane.

For T2-weighted imaging, the acquisition parameters were as follows: repetition time (TR) 5810 ms, echo time (TE) 119 ms, slice thickness 4 mm, field of view (FOV) 25×25 cm^2, and matrix 320×320. For diffusion-weighted imaging, the acquisition parameters were: TR 3700 ms, TE 63 ms, slice thickness 4 mm, FOV 36×28 cm^2, and matrix 162×126, with b-values of 50, 800, 1200 s/mm^2. ADC maps were reconstructed automatically on the scanner workstation using a mono-exponential decay model.

2.3. Multi-Reader Annotation

Manual segmentation of MRI images was performed on the MD.ai, a web-based platform for collaborative medical image annotation (MD.ai, Inc., New York, NY, USA https://md.ai/ (accessed on 20 February 2026)). Annotations were provided independently by three board-certified radiologists with at least three years of experience (>3000 prostate mpMRI interpreted), who delineated the prostate gland, anatomical zones, and PI-RADS ≥ 3 lesions. Additionally, for every case, a structured report was created in accordance with the PI-RADS 2.1 SR template using the prostate cancer module available on our in-house AI4AR structured reporting platform, which included prostate volume measurements, zonal characterization, and suspicious lesion reporting. The annotation processes were conducted blinded to each other's annotations and to biopsy outcomes. For each case, lesion annotations from three independent experts were reviewed and compared. The primary criterion for lesion acceptance was identification by at least 2 of the 3 radiologists. Lesions meeting this criterion were then verified against biopsy results and clinical records from the referring institution. Any uncertainties or ambiguities related to the historical or clinical data were subsequently resolved through expert panel discussion. All segmentation results were finally exported as NIfTI/mha files or, alternatively, as DICOM SEG/DICOM files.

2.4. Dataset Characteristics

As a result of the annotation process, 378 cases with complete clinical records and histopathologically confirmed or clinically documented diagnosis were included. The cohort represented a heterogeneous clinical population. The median (IQR) age was 68 (63–73) years; mean (SD) PSA level was 10.47 (10.37) ng/mL, and mean (SD) prostate volume was 56.1 (29.7) mL. A fixed independent test set of 55 (14.55%) studies was reserved for the final evaluation of the model. The remaining 323 (85.45%) studies constituted the training cohort. Detailed characteristics of the dataset are presented in Table 1.

2.5. Training Strategy

In our research, we used a subset of bpMRI sequences (T2W, DWI, ADC), which omits dynamic contrast-enhanced sequences, reducing acquisition time and eliminating the need for gadolinium administration. All models were trained using 5-fold cross-validation at the patient level. In each fold, 260 studies were used for training and 63 for internal validation. All scans associated with a given patient were assigned to a single fold, thereby preventing information leakage between the training and validation sets.

Two training strategies were explored and compared: a consensus-based approach and a multi-expert approach (Figure 1). In the consensus-based approach, lesion masks were constructed as the intersection of expert annotations, representing regions consistently identified by multiple readers. In the multi-expert training approach, each reader's annotations were retained as separate supervision targets and not merged into a single consensus label. For each expert, a dedicated 5-fold ensemble was trained using the same patient-level data partitioning, resulting in 15 models in total. The final system output was obtained by aggregating the predictions of the expert-specific models across all folds.

Table 1. Characteristics of the dataset used in the study. The dataset was organized into two stages: (1) a development subset (n = 323), used with patient-level 5-fold cross-validation, and (2) a fixed independent test set (n = 55) reserved for final evaluation.

	Total Patients	**Development Subset**	**Test Subset**
Characteristics	**(n = 378, 100%)**	**(n = 323, 85.45%)**	**(n = 55, 14.55%)**
Age (years), median (IQR)	68 (63–73)	68 (64–73)	69 (62–71)
PSA (ng/mL), mean (SD)	10.47 (10.37)	10.72 (10.76)	9.03 (7.66)
Prostate volume (mL) mean (SD)	56.1 (29.7)	55.1 (30.0)	62.1 (27.5)
ISUP score, case (%)			
0	197 (52.1%)	168 (44.4%)	29 (7.7%)
1	80 (21.2%)	68 (18.0%)	12 (3.2%)
2	53 (14%)	46 (12.2%)	7 (1.9%)
3	35 (9.3%)	29 (7.7%)	6 (1.6%)
4	8 (2.1%)	7 (1.9%)	1 (0.3%)
5	5 (1.3%)	5 (1.3%)	0 (0%)

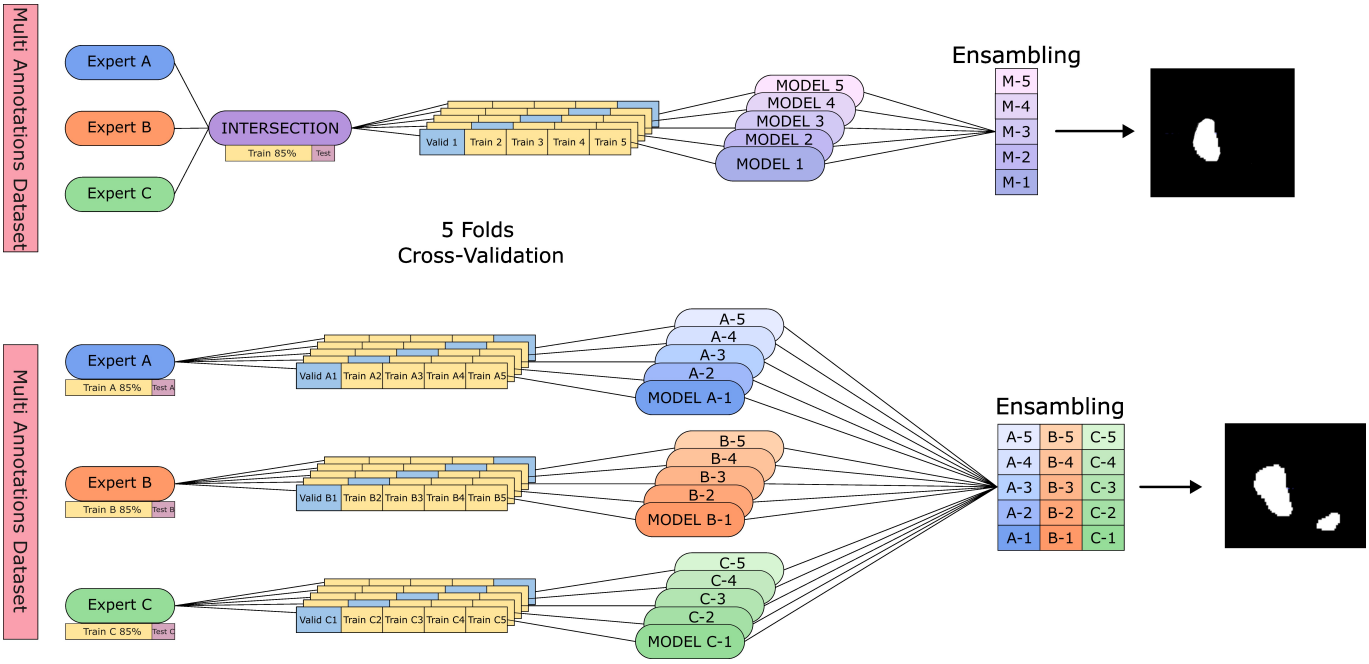


Figure 1. Overview of the two evaluated learning strategies. In the consensus-based strategy, a single lesion reference mask was constructed according to the predefined consensus rule and used for 5-fold ensemble training. In the multi-expert strategy, reader-specific annotations were retained as separate supervision targets and used to train three expert-specific 5-fold ensembles, resulting in 15 internal models in total. The final output is a binary mask representing lesion candidates derived from the predicted probability map.

This multi-expert strategy was designed to preserve inter-reader variability rather than collapse it into a single deterministic target. In prostate MRI, differences between expert annotations may reflect inherent variability in lesion conspicuity and boundary definition rather than solely random labeling noise. Exposing the model to multiple valid delineations of the same lesion may therefore reduce overfitting to the contouring style of any single reader and encourage learning of lesion representations that are more stable across experts.

2.6. Deep Learning Model Architecture and Training

The proposed framework combines several task-specific components, including prostate-centered preprocessing, a class-imbalance-aware composite loss, multi-expert supervision, and ensemble-based prediction aggregation. In the present work, these components were evaluated as an integrated design rather than through isolated ablation experiments. The segmentation model was implemented in the nnU-Net v2 framework [8] (https://github.com/MIC-DKFZ/nnUNet (accessed on 20 February 2026)) using a Residual Encoder U-Net backbone. The network input consisted of four stacked 3D channels provided in a fixed order across all experiments: T2-weighted imaging, high b-value diffusion-weighted imaging, ADC maps, and a binary prostate gland mask. This input composition was used consistently in all folds and training configurations.

2.6.1. Preprocessing

Preprocessing was designed to maximize anatomical focus and mitigate class imbalance between the surrounding background tissue and the foreground prostate gland. To reach these objectives, an expert-derived prostate gland mask was incorporated as an additional 3D input channel to explicitly provide anatomical context relevant to clinically significant prostate cancer (PCa) detection. The gland was then localized, and images were cropped around the prostate bounding box with a 20-voxel margin. In the present study, prostate-centered cropping was performed using expert-derived masks available in the dataset. In a fully automated deployment scenario, these masks would be generated by a dedicated upstream prostate gland segmentation model.

All MRI sequences were registered to the T2-weighted reference space using a rigid registration algorithm implemented in the SimpleITK library [26]. Following registration, all image volumes were resampled to a uniform voxel spacing of $0.5 \times 0.5 \times 3.0$ mm using third-order B-spline interpolation to preserve image detail and spatial smoothness. For binary masks—including prostate gland segmentations and lesion annotations—nearest-neighbor interpolation was applied to maintain label integrity and prevent partial-volume artifacts.

Because the dataset was single-center and single-scanner, no inter-scanner harmonization was performed. Intensity standardization was instead achieved through per-patient Z-score normalization for T2W and DWI sequences, and global normalization for ADC maps, with parameters computed separately for the training and test subsets to prevent data leakage. All preprocessing steps were applied uniformly across the dataset. The overall preprocessing pipeline is illustrated in Figure 2.

2.6.2. Network Architecture and Loss Function

The proposed model was implemented in the nnU-Net v2 framework using a Residual Encoder U-Net backbone. The input patch size was $16 \times 80 \times 96$. The network comprised three encoder stages and two decoder stages, with a feature-map progression of 16, 32, and 64 channels. Downsampling was performed using strided convolutions with strides of [1, 1, 1], [1, 2, 2], and [2, 2, 2]. Deep supervision was enabled during training at all decoder resolution levels. Dropout was set to 0.2 to improve regularization.

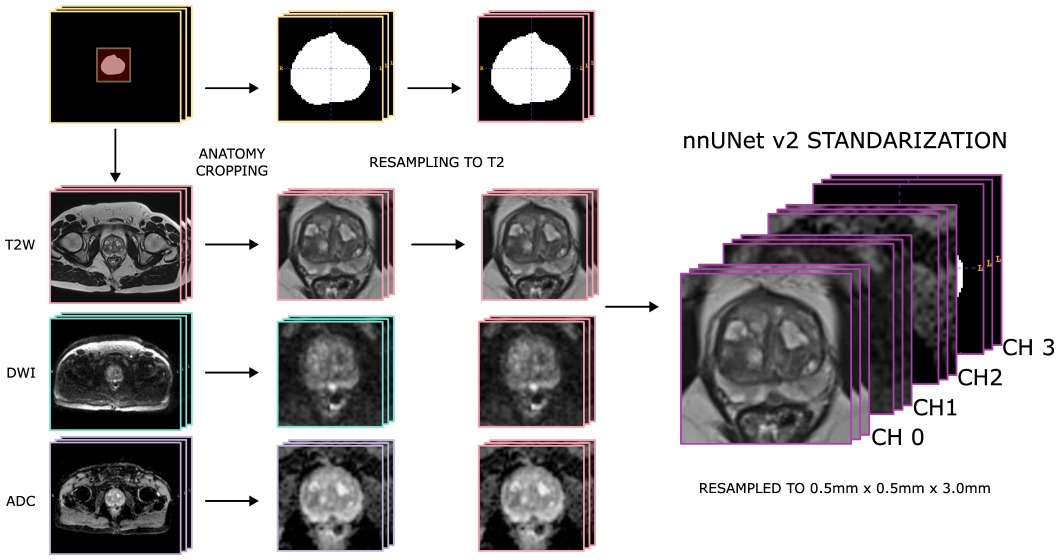


Figure 2. Overview of the data preprocessing. The network input consisted of four stacked 3D channels provided in a fixed order across all experiments: T2-weighted imaging, ADC maps, high b-value diffusion-weighted imaging, and a binary prostate gland mask. Preprocessing steps include prostate localization with prostate-centered cropping, rigid registration of all sequences to the T2-weighted reference space, and subsequent resampling of each channel to a resolution of 0.5 × 0.5 × 3.0 mm. Finally, all modalities and masks are stacked in accordance with the nnU-Net standard preprocessing.

The difficulties in accurately and precisely segmenting PCa lesions arise from the significantly unbalanced and transient characteristics of the image data across the entire dataset—the prostate tumor sometimes occupies only a limited area within the anatomical boundaries of the prostate gland, and in some cases, it may not appear at all. Traditional loss functions, while successful in delineating large anatomical structures, do not meet the demands imposed by the subtle complexities of pathological changes. To address this problem, we propose a composite loss function consisting of three complementary components: Focal Loss, Dice Loss, and Binary Cross-Entropy (BCE). Focal Loss was incorporated to address the pronounced class imbalance and to emphasize hard-to-classify voxels. It is defined as follows:

$$\mathcal{L}_{\text{Focal}}(p_t) = -\alpha(1 - p_t)^{\gamma} \log(p_t) \tag{1}$$

where pt denotes the predicted probability assigned by the model to the ground-truth class. For the Focal Loss component, the balancing parameter was set to $\alpha = 0.35$ and the focusing parameter was set to $\gamma = 2.0$. The parameter α controls the relative importance of the positive class and was used to assign lower weight to well-recognized large structures, while emphasizing smaller and less frequent lesion regions. The focusing parameter γ reduces the contribution of easily classified samples and forces the model to concentrate on difficult cases. This formulation helps reduce false negative predictions and improves the detection of small prostate cancer lesions.

The overall loss function was defined as:

$$\mathcal{L} = a\mathcal{L}_{\text{Dice}} + b\mathcal{L}_{\text{BCE}} + c\mathcal{L}_{\text{Focal}}, \quad a + b + c = 1 \tag{2}$$

The weighting coefficients a, b, and c were determined empirically based on preliminary experiments conducted on the training set. The best overall performance was achieved for $a = 0.5$, $b = 0.25$, and $c = 0.25$, which were therefore used in all subsequent experiments. This combination balanced the often-dominant anatomical context while also improving the model's ability to detect small, hard-to-find tumor lesions. The purpose of

this study is to evaluate the performance of the proposed nnU-Net pipeline in a clinically motivated application setting. Accordingly, the final loss formulation and the task-specific modifications introduced in this study were selected empirically during preliminary model development on the training data and were then kept fixed for all reported experiments.

2.6.3. Training Parameters

Model training was performed using the Adam optimizer with an initial learning rate of 0.01 and a polynomial learning-rate schedule. The batch size was set to 32, and training was performed for 1000 epochs. No early stopping was applied; the final model was selected based on the checkpoint with the highest moving average of the pseudo-Dice score on the validation set. During training, we used the standard nnU-Net augmentation pipeline applied on-the-fly to increase robustness to moderate variability in image appearance and geometry. These transformations were used to regularize the model. All experiments were implemented in PyTorch 2.2 with CUDA 12.2 on Ubuntu 22.04 using nnU-Net version 2.0. Training was performed on a workstation with four NVIDIA T4 GPUs (16 GB VRAM each), with each fold requiring approximately 15 h.

2.7. Evaluation

For evaluation purposes, all models were assessed against consensus-based reference masks. These masks were constructed according to the predefined agreement rule between experts and served as the unified ground truth for both training strategy comparisons. Importantly, while individual expert annotations were used as separate supervision targets during model development in the multi-expert strategy, all final evaluations—regardless of training strategy—were performed against the same consensus-based reference masks. This ensured consistent and fair evaluation across the consensus-based and multi-expert approaches.

Model evaluation was performed using the official evaluation pipeline of the PI-CAI challenge [27], an open-source framework specifically developed for standardized assessment of prostate cancer detection algorithms [14]. The tool provides lesion-level and patient-level metrics, including Average Precision (AP), Area Under the ROC Curve (AUROC), and the PICAI score, enabling consistent benchmarking against other approaches. Its use ensures reproducibility and comparability of results across studies and model configurations. The Dice coefficient was additionally calculated to quantify spatial overlap between predictions and annotations. Lesion matching was performed using the standard PI-CAI hit criterion, i.e., a predicted lesion was considered a true positive when its overlap with a reference lesion reached the minimum required IoU = 0.10. In split/merge cases, the candidate with the highest overlap was selected, in accordance with the default PI-CAI rule. Both training strategy evaluations were performed against the consensus reference masks. All performance metrics are reported together with 95% confidence intervals and were calculated on the actual independent held-out test set. A statistical analysis showing the p-values has been conducted to determine how our proposed method provides statistically significant improvements. We calculated p-values using a t-test to compare the mean performance scores of the training strategies examined with those of the baseline models. Results were considered statistically significant when $p < 0.01$. All analyses were performed in Python 3.10, using NumPy 2.1.2 for numerical computations, SimpleITK 2.5.0 for medical image input and processing, and scikit-learn 1.6.1 for performance evaluation.

3. Results

The proposed nnU-Net pipeline demonstrated improvements across the primary performance metrics compared with the baseline U-Net from our previous study [25].

For the multi-expert annotation learning strategy, AP increased from 0.76 to 0.81 (+6.6%), AUROC improved from 0.95 to 0.99 (+4.2%), Dice coefficient increased from 0.51 to 0.56 (+9.8%), and the PI-CAI score rose from 0.86 to 0.90 (+4.7%). The corresponding exact p-values were 0.004 for AP, 0.009 for AUROC, 0.006 for Dice, and 0.018 for the PI-CAI score. Improvements across the evaluated metrics were also observed for the consensus-based annotation learning strategy. For the consensus-based strategy, AP increased from 0.69 to 0.75 (+8.7%; $p = 0.012$), AUROC from 0.92 to 0.96 (+4.3%; $p = 0.021$), Dice coefficient from 0.51 to 0.52 (+1.9%; $p = 0.26$), and PI-CAI score from 0.81 to 0.85 (+4.9%; $p = 0.047$).

Detailed quantitative results are summarized in Table 2. All values reported in this table correspond to actual model performance obtained on the independent held-out test set. The primary endpoints were the official PI-CAI metrics, namely lesion-level Average Precision, patient-level AUROC, and the PI-CAI score, while Dice was reported as a complementary overlap-based metric. All primary metrics are presented together with 95% confidence intervals derived from the actual held-out test set using the PI-CAI evaluation framework. Representative qualitative examples of these improvements are shown in Figure 3.

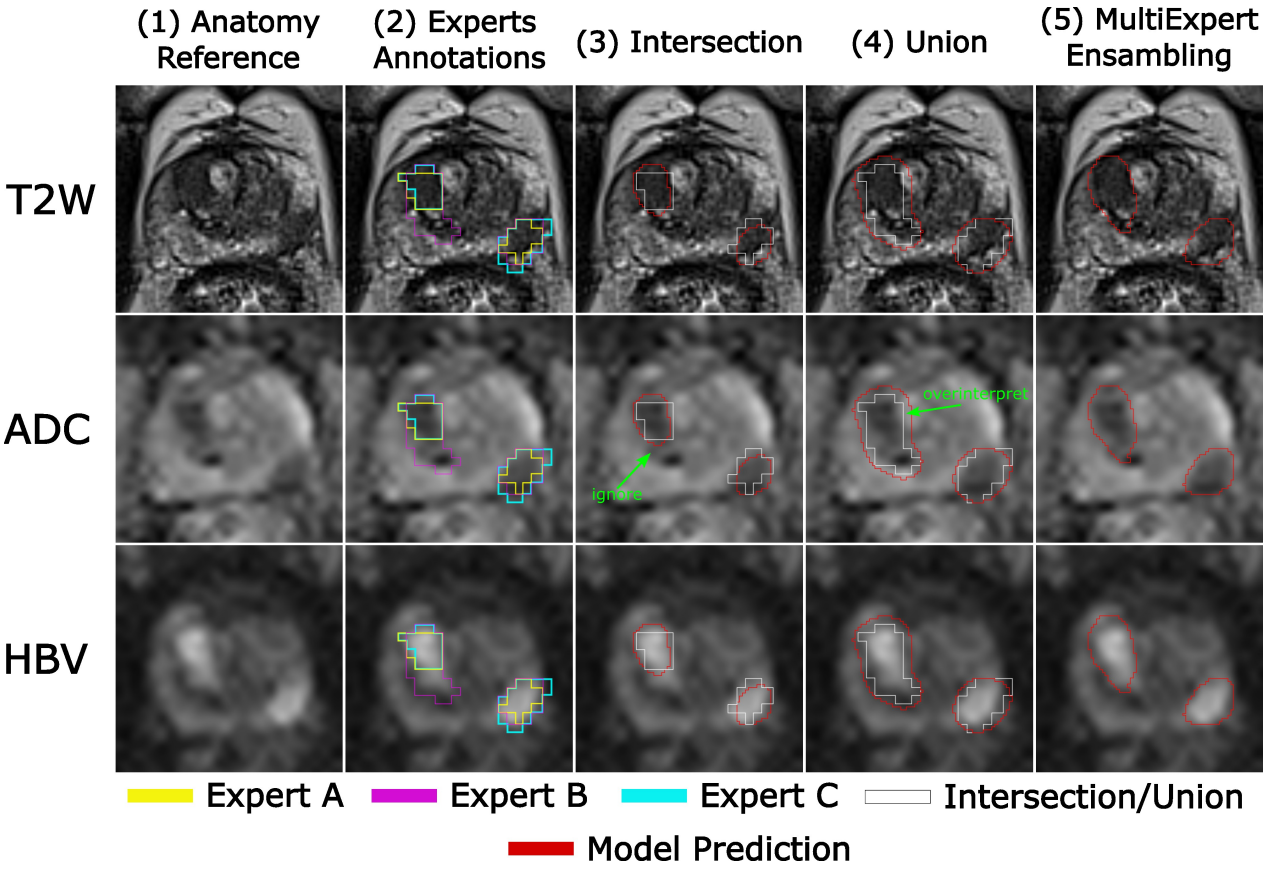


Figure 3. Qualitative comparison of prostate lesion detection between the baseline U-Net and the proposed nnU-Net pipeline. Example cases illustrate T2W and ADC inputs, predicted segmentations, and ground truth. The nnU-Net predictions show improved lesion localization and boundary delineation compared to the baseline U-Net, particularly for small or low-contrast lesions.

Table 2. Quantitative comparison between the baseline U-Net and the proposed nnU-Net under the evaluated annotation-learning strategies. Reported values correspond to actual model performance on the independent held-out test set. The primary endpoints are lesion-level Average Precision (AP), patient-level AUROC, and the PI-CAI score; Dice is reported as a complementary overlap-based metric. Confidence intervals are shown as 95% confidence intervals. Δ (%) denotes the relative improvement of nnU-Net over the baseline.

Metric	U-Net (Baseline)	nnU-Net (Proposed)	Δ (%)	*p*-Value	95% CI
Consensus-based Annotation Learning Strategy					
Average Precision (AP)	0.69	0.75	+8.7%	0.012	0.70–0.78
AUROC	0.92	0.96	+4.3%	0.021	0.93–0.97
Dice Coefficient	0.51	0.52	+1.9%	0.26	0.49–0.54
PICAI Score	0.81	0.85	+4.9%	0.047	0.81–0.87
Multi-expert Annotation Learning Strategy					
Average Precision (AP)	0.76	0.81	+6.6%	<0.01	0.77–0.83
AUROC	0.95	0.99	+4.2%	<0.01	0.97–1.00
Dice Coefficient	0.51	0.56	+9.8%	<0.01	0.52–0.58
PICAI Score	0.86	0.90	+4.7%	0.018	0.87–0.92

4. Discussion

4.1. Clinical and Technical Relevance of Multi-Expert Training

Our study highlights the added value of a multi-expert annotation learning strategy, consistent with our previous findings [25]. Rather than treating inter-reader disagreement as annotation noise, the present framework considers it a clinically meaningful source of label uncertainty, exposing the network to a broader spectrum of diagnostic interpretations and enhancing generalization. In prostate MRI, annotation differences between experts may reflect genuine variation in lesion conspicuity and boundary definition. Exposure to multiple valid delineations may therefore reduce overfitting to any single reader's contouring style and promote more stable lesion representations, consistent with recent evidence supporting diverse supervision signals in prostate MRI AI [7,14,19–21,25].

4.2. Impact of Focal Loss and Anatomical Preprocessing

The observed performance gains across all evaluated strategies underscore the contribution of the introduced preprocessing and loss-function modifications. Prostate-centered cropping reduced voxel-level class imbalance, while the tailored loss increased sensitivity to subtle lesions—together amplifying the benefits of multi-expert training without additional architectural complexity. As the pipeline was evaluated as an integrated framework, the results should not be interpreted as formal component-wise attribution of effect [11,12,20].

4.3. Broader Methodological Implications and Clinical Translation

Most prior studies on automated prostate lesion detection have relied on single-expert labels or simplified fusion strategies, which may introduce bias and underestimate diagnostic variability [11,12,14]. By preserving reader-specific supervision targets and aggregating them only at the ensemble level, our approach addresses this variability in a more explicit and clinically grounded manner. Importantly, inter-expert variability in the present study was not modeled using an explicit probabilistic label framework; instead, it was addressed operationally through two alternative supervision strategies: construction of a single consensus target and preservation of separate expert-specific targets with subsequent ensemble fusion. Although this design is simpler than formal probabilistic uncertainty modeling, it allowed us to examine the practical effect of expert variability on training and

performance. Our findings should also be interpreted in the context of recent advances in prostate MRI AI. Although strong performance has been reported in large-scale confirmatory and fully automated detection studies [14,21], recent multicentre evidence indicates that robustness across scanners and institutions remains a critical requirement for clinical deployment [22]. Similarly, recent recommendations from the PI-RADS Steering Committee emphasize detailed dataset characterization, transparent validation design, and clearly defined clinical use cases for AI development in prostate MRI [23]. Recent explainable-AI studies further suggest that interpretable outputs and confidence-aware decision support may be important for clinical adoption [19], while robustness-oriented reviews continue to highlight the need for systematic evaluation of vendor- and protocol-related variability [20]. The proposed framework demonstrated strong internal performance; however the present study remains limited to a single-center, single-vendor dataset. Standard training-time augmentation was used to improve robustness to moderate image variability; however, such augmentation should be interpreted as a regularization strategy rather than a substitute for external validation. In line with structured validation principles used in biomedical engineering [28] and broader guidance for trustworthy and deployable healthcare AI [24], the appropriate next step for this methodology is a staged evaluation pathway including perturbation-based robustness testing, retrospective external validation on multicenter and multi-vendor cohorts, and prospective clinical assessment before real-world deployment claims can be made. Taken together, the three core components of our approach—multi-expert training, focal loss, and prostate-centered preprocessing—acted synergistically to improve both technical performance and the clinical relevance of lesion detection. For this reason, we frame the present approach as a clinically motivated and internally validated pipeline that provides a basis for further external validation toward clinical deployment, rather than as a system already ready for routine real-world implementation.

4.4. Study Limitations and Fairness Considerations

Several limitations should be acknowledged. The dataset was sourced exclusively from a single institution and acquired on a single scanner platform from a single vendor; therefore, scanner-related variability, harmonization strategies, and multicenter generalizability could not be assessed in the present study. Although this study did not explicitly stratify by demographic groups, the dataset included a representative range of patient characteristics, and all preprocessing steps were applied uniformly. Future work should explore external validation on demographically and institutionally diverse datasets to examine fairness and generalization [22–24]. A limitation of the present study is the absence of a dedicated ablation study assessing the individual contributions of the proposed framework components. Consequently, the observed performance improvements cannot be formally attributed to any single component. A further limitation of the present study concerns the fact that the robustness of prostate-centered cropping to failures in automatic gland localization was not assessed independently, since cropping relied on expert-derived gland masks. In a fully automated pipeline, inaccuracies in gland segmentation may propagate to the cropping stage, thereby adversely affecting lesion detection performance. Future work should therefore quantify the failure rates of gland localization and evaluate suitable quality-control procedures or fallback strategies. In addition, we did not perform a formal benchmark of inference time or computational footprint, and we have not demonstrated direct integration with clinical PACS environments. These deployment-oriented aspects should be evaluated explicitly in future translational studies [23,24]. Finally, the present work did not include formal subgroup analysis of lesion-level performance by PI-RADS category, and it was not designed as an exhaustive benchmark against all contemporary backbone families, such as transformer-based architectures.

5. Conclusions

We present a streamlined and effective extension of nnU-Net for automated prostate lesion detection, integrating a tailored loss function, anatomically focused preprocessing, and multi-expert training. These targeted and clinically informed modifications resulted in substantial performance gains, with AP increasing by 6.6%, AUROC by 4.2%, and Dice coefficient by 9.8% compared to a U-Net baseline on the independent held-out test set. By combining an adaptive network architecture with clinically informed training strategies and a quantitative understanding of inter-expert variability, our approach demonstrates promising internal performance for reliable AI-assisted prostate cancer detection in bpMRI. Given that the present study was limited to a single-center, single-vendor dataset, these results should be interpreted as internal validation rather than evidence of immediate multicenter generalizability or routine real-world deployment. Future work will focus on multi-center external validation, integration with clinical data, fairness assessment, robustness to scanner- and protocol-related variability, and dedicated explainability and uncertainty analyses to support safe and trustworthy clinical translation.

Author Contributions: Conceptualization, R.J.; methodology, R.J., M.G. and K.T.; software, M.G., J.M. and I.M.; validation, M.G., J.M., D.S.R., J.D. and A.Z.-G.; formal analysis, R.J., J.M. and K.T.; investigation, R.J. and I.M.; resources, K.T. and T.L.; data curation, T.L., J.D. and A.Z.-G.; writing—original draft preparation, R.J., M.G. and T.L.; writing—review and editing, R.J. and T.L.; visualization, R.J. and M.G.; supervision, D.S.R.; project administration, R.J. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Polish National Centre for Research and Development under the program INFOSTRATEG I, project INFOSTRATEG-I/0036/2021 "AI-augmented radiology—detection, reporting and clinical decision making in prostate cancer diagnosis".

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Bioethics Committee of the Hirszfeld Institute of Immunology and Experimental Therapy (approval number: KB 4/2022).

Informed Consent Statement: Written informed consent was obtained from all subjects for database creation and for the use of their data for research purposes.

Data Availability Statement: The AI4AR dataset that supports the findings of this study is available from the corresponding author upon reasonable request. The AI4AR dataset is currently being published and will soon be available to the public.

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

References

1. Siegel, R.L.; Miller, K.D.; Wagle, N.S.; Jemal, A. Cancer statistics, 2023. *CA Cancer J. Clin.* **2023**, *73*, 17–48. [CrossRef]
2. Zhu, Y.; Wang, H.K.; Qu, Y.Y.; Ye, D.W. Prostate cancer in East Asia: Evolving trend over the last decade. *Asian J. Androl.* **2015**, *17*, 48–57. [CrossRef]
3. Padhani, A.R.; Weinreb, J.; Rosenkrantz, A.B.; Villeirs, G.; Turkbey, B.; Barentsz, J. PI-RADS v2 status update and future directions. *Eur. Urol.* **2019**, *75*, 385–396. [CrossRef] [PubMed]
4. 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. *AJR Am. J. Roentgenol.* **2019**, *212*, 1197–1205. [CrossRef] [PubMed]
5. Sonn, G.A.; Fan, R.E.; Ghanouni, P.; Wang, N.N.; Brooks, J.D.; Loening, A.M.; Daniel, B.L.; To'o, K.J.; Thong, A.E.; Leppert, J.T. Prostate magnetic resonance imaging interpretation varies substantially across radiologists. *Eur. Urol. Focus* **2019**, *5*, 592–599. [CrossRef] [PubMed]

6. Jóźwiak, R.; Sobecki, P.; Lorenc, T. Intraobserver and Interobserver Agreement between Six Radiologists Describing mpMRI Features of Prostate Cancer Using a PI-RADS 2.1 Structured Reporting Scheme. *Life* **2023**, *13*, 580 . [CrossRef]
7. Bosma, J.S.; Saha, A.; Hosseinzadeh, M.; Slootweg, I.; de Rooij, M.; Huisman, H. Semisupervised Learning with Report-guided Pseudo Labels for Deep Learning-based Prostate Cancer Detection Using Biparametric MRI. *Radiol. Artif. Intell.* **2023**, *5*, e230031. [CrossRef]
8. Isensee, F.; Jaeger, P.F.; Kohl, S.A.A.; Petersen, J.; Maier-Hein, K.H. nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods* **2021**, *18*, 203–211. [CrossRef]
9. Bardis, M.; Houshyar, R.; Chantaduly, C.; Tran-Harding, K.; Ushinsky, A.; Chahine, C.; Rupasinghe, M.; Chow, D.; Chang, P. Segmentation of the prostate transition zone and peripheral zone on MR images with deep learning. *Radiol. Imaging Cancer* **2021**, *3*, e200024. [CrossRef]
10. Aldoj, N.; Biavati, F.; Michallek, F.; Stober, S.; Dewey, M. Automatic prostate and prostate zones segmentation of magnetic resonance images using DenseNet-like U-net. *Sci. Rep.* **2020**, *10*, 14315. [CrossRef]
11. Alzate-Grisales, J.A.; Mora-Rubio, A.; García-García, F.; Tabares-Soto, R.; Iglesia-Vayá, M.D.L. SAM-UNETR: Clinically significant prostate cancer segmentation using transfer learning from large model. *IEEE Access* **2023**, *11*, 118217–118232. [CrossRef]
12. Song, E.; Long, J.; Ma, G.; Liu, H.; Hung, C.C.; Jin, R.; Wang, P.; Wang, W. Prostate lesion segmentation based on a 3D end-to-end convolutional neural network with deep multi-scale attention. *Magn. Reson. Imaging* **2023**, *99*, 98–109. [CrossRef]
13. Mitura, J.; Jóźwiak, R.; Mycka, J.; Mykhalevych, I.; Gonet, M.; Sobecki, P.; Lorenc, T.; Tupikowski, K. Ensemble Deep Learning Models for Segmentation of Prostate Zonal Anatomy and Pathologically Suspicious Areas. In *Proceedings of the Medical Image Understanding and Analysis*; Yap, M.H., Kendrick, C., Behera, A., Cootes, T., Zwiggelaar, R., Eds.; Springer: Cham, Switzerland, 2024; pp. 217–231.
14. Saha, A.; Bosma, J.S.; Twilt, J.J.; Van Ginneken, B.; Bjartell, A.; Padhani, A.R.; Bonekamp, D.; Villeirs, G.; Salomon, G.; Giannarini, G.; et al. Artificial intelligence and radiologists in prostate cancer detection on MRI (PI-CAI): An international, paired, non-inferiority, confirmatory study. *Lancet Oncol.* **2024**, *25*, 879–887. [CrossRef] [PubMed]
15. Adams, L.C.; Makowski, M.R.; Engel, G.; Rattunde, M.; Busch, F.; Asbach, P.; Niehues, S.M.; Vinayahalingam, S.; van Ginneken, B.; Litjens, G.; et al. Prostate158—An expert-annotated 3T MRI dataset and algorithm for prostate cancer detection. *Comput. Biol. Med.* **2022**, *148*, 105817. [CrossRef] [PubMed]
16. Warfield, S.K.; Zou, K.H.; Wells, W.M. Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation. *IEEE Trans. Med. Imaging* **2004**, *23*, 903–921. [CrossRef]
17. Le, K.H.; Tran, T.V.; Pham, H.H.; Nguyen, H.T.; Le, T.T.; Nguyen, H.Q. Learning from multiple expert annotators for enhancing anomaly detection in medical image analysis. *arXiv* **2022**, arXiv:2203.10611. [CrossRef]
18. Mirikharaji, Z.; Abhishek, K.; Izadi, S.; Hamarneh, G. D-LEMA: Deep learning ensembles from multiple annotations—Application to skin lesion segmentation. *arXiv* **2021**, arXiv:2012.07206. [CrossRef]
19. Hamm, C.A.; Baumgärtner, G.L.; Biessmann, F.; Beetz, N.L.; Hartenstein, A.; Savic, L.J.; Froböse, K.; Dräger, F.; Schallenberg, S.; Rudolph, M.; et al. Interactive Explainable Deep Learning Model Informs Prostate Cancer Diagnosis at MRI. *Radiology* **2023**, *307*, e222276. [CrossRef]
20. Fassia, M.K.; Balasubramanian, A.; Woo, S.; Vargas, H.A.; Hricak, H.; Konukoglu, E.; Becker, A.S. Deep Learning Prostate MRI Segmentation Accuracy and Robustness: A Systematic Review. *Radiol. Artif. Intell.* **2024**, *6*, e230138. [CrossRef]
21. Cai, J.C.; Nakai, H.; Kuanar, S.; Froemming, A.T.; Bolan, C.W.; Kawashima, A.; Takahashi, H.; Mynderse, L.A.; Dora, C.D.; Humphreys, M.R.; et al. Fully Automated Deep Learning Model to Detect Clinically Significant Prostate Cancer at MRI. *Radiology* **2024**, *312*, e232635. [CrossRef]
22. Giganti, F.; Moreira da Silva, N.; Yeung, M.; Davies, L.; Frary, A.; Ferrer Rodriguez, M.; Sushentsev, N.; Ashley, N.; Andreou, A.; Bradley, A.; et al. AI-powered Prostate Cancer Detection: A Multi-centre, Multi-scanner Validation Study. *Eur. Radiol.* **2025**, *35*, 4915–4924. [CrossRef] [PubMed]
23. Turkbey, B.; Huisman, H.; Fedorov, A.; Macura, K.J.; Margolis, D.J.; Panebianco, V.; Oto, A.; Schoots, I.G.; Siddiqui, M.M.; Moore, C.M.; et al. Requirements for AI Development and Reporting for MRI Prostate Cancer Detection in Biopsy-Naive Men: PI-RADS Steering Committee, Version 1.0. *Radiology* **2025**, *315*, e240140. [CrossRef]
24. Lekadir, K.; Frangi, A.F.; Porras, A.R.; Glocker, B.; Cintas, C.; Langlotz, C.P.; Weicken, E.; Asselbergs, F.W.; Prior, F.; Collins, G.S.; et al. FUTURE-AI: International Consensus Guideline for Trustworthy and Deployable Artificial Intelligence in Healthcare. *BMJ* **2025**, *388*, e081554. [CrossRef]
25. Gonet, M.; Majorek, S.; Mycka, J.; Mykhalevych, I.; Jozwiak, R. Improving prostate lesion detection with multiple annotations and ensemble techniques. In Proceedings of the MIDI, Warsaw, Poland, 12 December 2024. [CrossRef]
26. Lowekamp, B.C.; Chen, D.T.; Ibáñez, L.; Blezek, D. The Design of SimpleITK. *Front. Neuroinform.* **2013**, *7*, 45. [CrossRef] [PubMed]

27. Pical Evaluator. Available online: https://github.com/DIAGNijmegen/picai_eval (accessed on 15 October 2025).
28. Laganà, F. Design and Simulation-Based Validation of an Embedded Acquisition Architecture for In Situ PCB Integrity Monitoring in Biomedical Devices. *Electronics* **2026**, *15*, 833. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.