Segmentation uncertainty with statistical guarantees in prostate MRI

Kevin Mekhaphan Nguyen[*], Alvaro Fernandez-Quilez[*][||]

[*]Department of Electrical Engineering and Computer Science, University of Stavanger, Stavanger, Norway.
[||]Stavanger Medical Imaging Laboratory (SMIL), Department of Radiology, Stavanger University Hospital, Stavanger, Norway.
Corresponding author email: alvaro.f.quilez@uis.no

***Abstract*—Prostate volume (PV) is an important factor in prostate cancer (PC) patient management and diagnostic pathway. Over the past years, efforts have been made to develop artificial intelligence (AI) systems able to standardize and reduce inter-reader variability in prostate segmentation and subsequent PV estimation. In spite of the remarkable results of AI segmentation architectures such as nnU-Net, most benchmarks and prostate segmentation results neglect the uncertainty of the AI segmentation system as part of their evaluation protocol. In this study, we use conformal prediction (CP), a model-agnostic uncertainty quantification method that provides strong statistical guarantees to keep the error of a nnU-Net for prostate whole gland (WG) segmentation bounded by a pre-specified level. Our results show that nnU-Net coupled with CP and a confidence level of 95.00% is able to significantly improve segmentation results in terms of Volume Difference (VD) when compared to nnU-Net without CP for two independent data cohorts (VD$_1$ = 1.67±0.99 p<.001 and VD$_2$ = 2.86±0.28 p<.001)**

***Index Terms*—uncertainty, conformal prediction, MRI, segmentation, prostate cancer**

I. Introduction

At present, one of the main barriers in the diagnostic pathway of prostate cancer (PC) is the high rate of overdiagnosis, overtreatment and excessive invasive testing in the form of biopsies [1]. Typically, decisions to refer subjects to confirmatory tests are based on known PC risk factors such as the prostate volume (PV), prostate specific antigen (PSA) or age of the subject [2]. In some cases, PSA density (PSAD), a combination of PV and PSA, might also be considered [3]. Despite the relevance of an accurate characterization of the PV for PC diagnosis and management, typical PC diagnostic workflows consider manual calculation of the PV from magnetic resonance imaging (MRI) [4]. Whilst the process can be relatively accurate, it has been shown to be prone to being reader-dependant and suffering from errors due to anatomical challenges of the prostate [4].

Over the past years, there has been a proliferation of artificial intelligence (AI) tools aiming to automatize and standardize prostate whole gland (WG) segmentation [5], [6]. Among the proposed AI solutions, nnU-Net is commonly a top performer in prostate WG challenges and considered a *de facto* choice for the task [5], [7], [8]. In spite of the progress in deep learning (DL) networks in prostate WG segmentation, there is a growing evidence suggesting that AI models are poorly calibrated and present overconfident predictions [9], [10]. Nevertheless, most segmentation benchmarks present systems with point predictions and neglect any assessment of the *uncertainty* of the AI systems [7], [11].

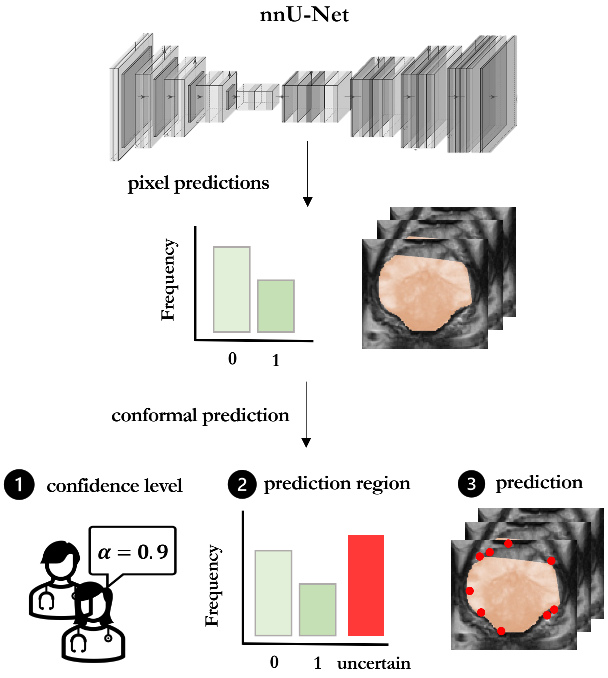

Fig. 1: Breakdown of the steps included in our uncertainty quantification framework based on conformal prediction (CP) and applied to nnU-Net.

In light of the significance of uncertainty quantification (UQ), there has been an increase in the popularity of methods to quantify it in DL architectures [12], [13]. Typical UQ methods include confidence-based, Bayesian, and ensemble methods [12]. In spite of their wide use, confidence-based methods can be sensitive to poor calibration. Further, Bayesian methods are built on subjective beliefs and assumptions [14], and ensemble methods usually suffer from high computational complexity, limiting their usage to light models [15].

Conformal prediction (CP) is a distribution-free and model-agnostic UQ method [16], [17]. Compared to other UQ methods, CP provides strong statistical guarantees to keep the error rate of a given AI system bounded by a pre-specified level

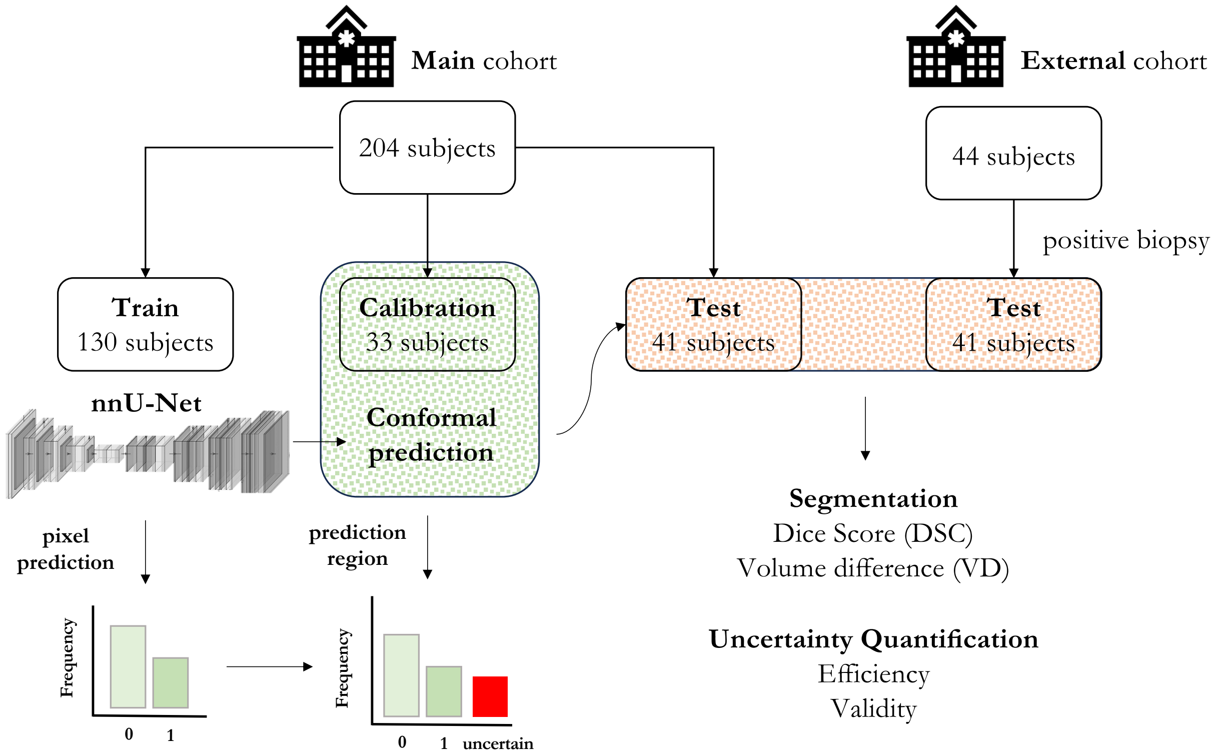

Fig. 2: Technical approach to the project, depicting the main cohort used to train the nnU-Net, the external cohort and the different splits used to calibrate the conformal predictor and to test the uncertainty quantification effect.

[18]. With CP, the end-user can set a desired confidence level α such that the conformal predictor will then provide a region around the point prediction that contains the true label with a confidence level % probability. In the event were the prediction does not reach the required α level, an empty prediction $\{\emptyset\}$ is returned-flagging the prediction as too *uncertain* to be reported.

In this study, we develop conformal predictors for AI-assisted prostate WG segmentation (Figure 1). We show how these predictors can be used for UQ to detect unreliable pixel predictions in the WG segmentation. In particular, we make the following contributions: (*i*) Systematic analysis of a nnU-Net performance for WG segmentation in an *internal* and *external* cohorts (*ii*) quantitative analysis of the performance of the model *without* conformal prediction for UQ (*iii*) quantitative analysis of the effect of CP and different α levels for UQ. Code and model weights will be shared upon request on GitHub.

II. Materials and Methods

A. Study cohorts

*1) **Main cohort**:* The ProstateX challenge data (Radboud University, Netherlands) [7] is a collection of *open-access* and retrospectively collected prostate MRI exams to validate modern AI algorithms for the diagnostic classification of PC. Subjects included in the ProstateX challenge were recruited on the basis of suspicion of PC based on high PSA levels. Following, PC diagnosis was confirmed through an MRI-guided biopsy [19].

The ProstateX challenge cohort consisted of 204 subjects (median age 66 years [range, 48-83]) with available prostate volume (mL) and Gleason Score (GS) obtained from the biopsy. In addition, the cohort had pixel-level annotations for the WG obtained by two experienced board-certified radiologists with > 5 years of experience [20]. This cohort served as the primary source to train the models and for in-distribution (internal) testing.

*2) **External cohort**:* Data from Stavanger University Hospital (SUS, Norway) [21] was collected to assess the replicability and generalisability of our proposed methods. The external cohort consisted of 48 subjects (median age 68 years, [range, 49-83]) that were recruited under the basis of PC suspicion based on high PSA levels. All diagnosis of all subjects were confirmed by biopsies.

Clinical and demographic available data included prostate volume (mL) and Gleason Score (GS) obtained from the biopsy. Manual annotations of the WG were also available for the external cohort, obtained by a radiologist in training with < 2 years of experience. All annotations were obtained with ITK-SNAP v.380 software (http://www.itksnap.org/).

B. Magnetic resonance Imaging

*1) **Acquisition**:* We used 3.0 Tesla (T) axial T2-weighted (T2w) spin sequences together with their paired prostate WG gland masks for model training and testing. Images were acquired with either Siemens (Siemens Health Engineerns, Erlangen, Germany) or Philips (Philips & Co, Endhoven, The Netherlands). Additionally, the T2w MRI exams had an in-plane resolution of [0.5-0.562mm]x[0.5-0.562mm]x[3.0-3.15mm] and 0.50mm x 0.50mm x 3.0mm for the main and external cohorts, respectively.

Inclusion criteria limited the cohort to those with T2w MRI exams with a biopsy-confirmed (systematic, MRI-guided or both) diagnosis consisting of low-grade PC (GS <= 6) or high-grade PC (GS >= 7). After application of the inclusion criteria, 204 participants from the main cohort and 41 from the external cohort were included in the study (Figure 2).

*2) **Data splitting**:* We split the main cohort in 80/20%, resulting in 163 subjects used to train the DL model and for UQ and 41 subjects used to test. The splitting was performed at the subject level to avoid cross-contamination. We keep the entire external cohort (41 subjects) for external testing purposes. Data characteristics of the whole cohort and the resulting splits are depicted in Table I and in Figure 2.

TABLE I: Main and external cohorts characteristics after matching by diagnostic criteria.

	Main	*Main (train)*	*Main (calib.)*	*Main (test)*	**External**
Subjects	204	130(63.70)	33(16.17)	41(20.13)	41
GS <= 6	65	42(32.31)	10(30.31)	13(31.70)	14(34.15)
GS >= 7	139	88(67.69)	23(69.69)	28(68.30)	27(65.85)
Volume (mL)	89.52±49.99	88.08±22.37	90.08±43.99	86.08±35.05	73.46±46.30

† GS = Gleason Score, Calib. = Calibration set.

C. Conformal Prediction

We applied a CP framework for the segmentation of prostate WG from prostate MRI images. A description of the DL system that provides the prostate masks is provided in the next section. Conformal predictions can be constructed in different ways, and for the purpose of the study, we implemented a Mondrian type of CP which guarantees the error rate per class [16].

In order to train the CP framework, we split the previously obtained training data in 80% and a CP calibration set of 20%, resulting in 130 subjects used to train the DL system and 33 subjects used to calibrate the CP framework (Figure 2). Once the CP framework is trained, we explore different α confidence levels (75%, 90% and 95%) and their effect on the UQ performance.

Given the high dimension nature of the data (320x320 pixels per slice), we reduce the amount of pixels used in training the CP by extracting the region of interest of the calibration set. In particular, we crop around the prostate gland with a 20 pixel margin to ensure that there is enough representation of pixels assigned to class 0 and pixels assigned to class 1. At inference time, we use the entire prostate slice.

D. Deep learning for whole gland segmentation

We choose nnU-Net based on its wide adoption and positive results in previous WG segmentation challenges[19], [21]. Briefly, nnU-Net provides a framework that includes automatic pre-processing, automatic configuration and training of U-Net architecture and post-processing of the results.

As part of the pre-processing, the pixel intensity range of the images is normalized followed by a center cropping, re-ordering of the axis and re-sampling of the T2w MRI sequence. We select a 2D nnU-Net as the base architecture, based on experimentation with the other variants and omit post-processing options. More details about post-processing options or data augmentation techniques can be found in the original article [5].

TABLE II: Effect of different confidence levels in nnU-Net segmentation performance and conformal predictor uncertainty quantification.

confidence level	DSC↑	VD(%) ↓	Efficiency(%) ↑	Validity(%) ↑
α = 0.75	0.89±0.15	10.70±4.60	99.33±1.68	99.34±1.11
α = 0.90	0.96±0.05	2.98±1.65	99.60±1.13	99.61±0.97
α = 0.95	**0.98±0.01**	**1.67±0.99**	**99.78±0.83**	**99.80±0.59**

† DSC = Dice Score Coefficient, VD = Volume difference.

*1) **Training**:* Training is performed with the default configuration of the 2D nnU-Net, as exploration of architecture modfications is considered out of the scope of the work. We train the architecture to minimize a combination of dice loss and cross-entropy. The architecture is trained with an SGD optimizer and for 1000 epochs for every fold. The validation loss is monitored for every fold, and we keep the weights for the epoch were the minimum is reached. As part of nnU-Net framework, an automatic selection of data augmentation techniques are applied on the fly [5]. Model training and evaluation were carried out on an NVIDIA A100-80G GPU (NVIDIA Corporation, Santa Clara, USA).

*2) **Evaluation**:* At test time, we consider nnU-Net without UQ application as the baseline of the work. Following, we applied the CP framework to the baseline nnU-Net, and characterize the segmentation results. We compare the results in terms of segmentation and UQ performance.

In terms of metrics, we resort to VD (%) and dice score coefficient (DSC) for segmentation purposes. Volume difference (VD) is defined as the coefficient between the estimated prostate volume obtained from the segmentation result, and the original prostate volume from the ground truth. In the UQ case, we employ efficiency (%) and validity (%). Efficiency is defined as the *amount of pixels assigned to a single class* whilst validity is defined as *the fraction of correct pixel predictions*. Calculation of DSC and VD is based on pixel predictions that are **not** flagged as uncertain, under the assumption that those pixels predictions would be flagged by the system and a posterior human intervention would correct them.

E. Statistical testing

All analyses were performed in Python 3 (www.python.org/downloads) with the open-sourced statsmodels 0.14.0 module (www.statsmodels.org). We reported continuos variables as mean and standard deviation (mean ± SD) and categorical variables as number of ocurrences and percentage (N[%]). We performed unpaired t-test and Mann-Whitney U tests where appropriate to assess the differences between nnU-Net with and without CP for UQ and for the different α confidence levels (75%, 90% and 95%). A P value <0.05 was considered statistically significant.

TABLE III: nnU-Net results without uncertainty quantification and after applying conformal prediction with a confidence level of 95%, for the main and external cohort test set.

cohort	*method*	DSC ↑	VD(%) ↓	*p value*$_{VD}$	Efficiency(%) ↑	Validity(%) ↑
main	nnU-Net	0.92±0.48	2.46±1.20	<.001	-	-
	nnU-Net w/ CP (α = 0.95)	**0.98±0.01**	**1.67±0.99**		**99.78±0.83**	**99.80±0.59**
external	nnU-Net	0.89±0.65	3.47±1.81	<.001	-	-
	nnU-Net w/ CP (α = 0.95)	**0.94±0.08**	**2.86±0.28**		**99.46±0.43**	**99.51±0.46**

† DSC = Dice Score Coefficient, VD = Volume difference, w/ CP = with Conformal Prediction.

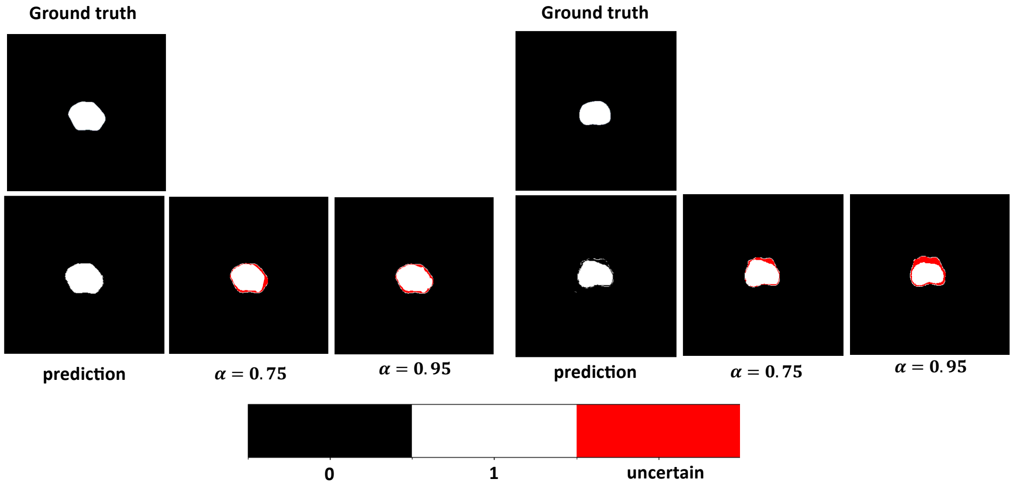

Fig. 3: Effect of applying conformal prediction with $\alpha = 0.75$ and $\alpha = 0.95$ to two different prostate segmentation slices.

III. Results

A total of 130 participants of the main cohort were considered for nnU-Net model training, whilst 33 participants were used to calibrate the conformal predictors. Table I depicts a summary of some of the available participants characteristics. At test time, we considered 41 participants from the main cohort and 41 participants from the external cohort with biopsy-confirmed PC who were not included in the model training stage. The external cohort participants were the result of matching the original cohort to the main cohort based on PC diagnosis. As depicted in Table I, both the different splits and cohorts had a similar distribution in terms of PC diagnosis and prostate WG volume. Specifically, the external cohort presented a smaller WG when compared to the main cohort.

As part of our sensitivity analysis, we compared the effect of CP different confidence levels (α) in the segmentation results of the main and external cohorts. As shown in Table II, CP($\alpha = 0.95$) largely improves the DSC(0.98±0.01), VD(1.67±0.99%) and provides a large efficiency(99.78±0.83%) and validity(99.80±0.59%). Hereby, we considered CP($\alpha = 0.95$) for the rest of the evaluation.

Table III shows the results of the nnU-Net architecture with and without CP($\alpha = 0.95$). When using CP($\alpha = 0.95$), we observe a statistically significant reduction in VD for the main (1.67±0.99%) and external cohorts test set (2.86±0.28%). Furthermore, we also observe a significant improvement in terms of DSC for the main (0.98±0.01) and external cohorts (0.94±0.08). In that regard, figure 3 shows the effect of CP in a qualitative way in the predicted segmentation of two different slices of different subjects.

IV. Discussion

In our retrospective study, we proposed CP as a UQ method for prostate WG segmentation. Our study provides some degree of evidence of the positive effect of using CP to flag uncertain pixel predictions in terms of VD and DSC in the main cohort. Furthermore, our results in the external cohort matched by diagnosis support the results observed in the main cohort and the improvement in terms of DSC and VD when incorporating CP as an UQ method in prostate WG segmentation.

When compared to other UQ methods, our approach leveraging CP presents the advantage of being model-agnostic, distribution-free and as depicted by our results, strong statistical guarantees that ensure the error of the AI system is bounded by a pre-specified confidence level (α). As depicted in other works [6], [19], prostate segmentation is a crucial aspect in advanced computer aided detection (CAD) pipelines and the flexibility of the CP framework used in our work allows for integration in different stages of CAD prostate systems.

The results presented in our study have some limitations. First, our study was limited by its retrospective nature and small sample size of the test sets of both cohorts. Future studies including a larger sample size are required to verify our findings. Furthermore, quantification of the calibration of the architecture with and without CP is not included. Finally, some of the annotations used for the purpose of the study were obtained by one expert. Future studies should reflect the variability present in the annotations when multiple experts are accounted for.

V. Conclusion

We present a prostate segmentation framework that accounts for uncertainty quantification with a model-agnostic, distribution-free and with strong statistical guarantees that keep the error of the system bounded. We show that our approach can improve prostate segmentation results by flagging uncertain pixel predictions. Our approach could potentially serve to mark pixel predictions that require human-supervisions, effectively leading to a human-AI collaboration and reduction of time spent on the task by experts.

References

[1] M. Eklund, F. Jäderling, A. Discacciati, *et al.*, “Mri-targeted or standard biopsy in prostate cancer screening,” *New England Journal of Medicine*, vol. 385, no. 10, pp. 908–920, 2021.

[2] A. Fernandez-Quilez, T. Nordström, F. Jäderling, S. R. Kjosavik, and M. Eklund, “Prostate age gap: An mri surrogate marker of aging for prostate cancer detection,” *Journal of Magnetic Resonance Imaging*, 2023.

[3] T. Nordström, O. Akre, M. Aly, H. Grönberg, and M. Eklund, “Prostate-specific antigen (psa) density in the diagnostic algorithm of prostate cancer,” *Prostate cancer and prostatic diseases*, vol. 21, no. 1, pp. 57–63, 2018.

[4] E. Thimansson, J. Bengtsson, E. Baubeta, *et al.*, “Deep learning algorithm performs similarly to radiologists in the assessment of prostate volume on mri,” *European Radiology*, vol. 33, no. 4, pp. 2519–2528, 2023.

[5] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, “Nnu-net: A self-configuring method for deep learning-based biomedical image segmentation,” *Nature methods*, vol. 18, no. 2, pp. 203–211, 2021.

[6] E. S. Rolfsnes, P. Thangngat, T. Eftestøl, *et al.*, “Reconsidering evaluation practices in modular systems: On the propagation of errors in mri prostate cancer detection,” *arXiv preprint arXiv:2309.08381*, 2023.

[7] G. Litjens, R. Toth, W. Van De Ven, *et al.*, “Evaluation of prostate segmentation algorithms for mri: The promise12 challenge,” *Medical image analysis*, vol. 18, no. 2, pp. 359–373, 2014.

[8] T. N. Lindeijer, T. M. Ytredal, T. Eftestøl, *et al.*, “Leveraging multi-view data without annotations for prostate mri segmentation: A contrastive approach,” *arXiv preprint arXiv:2308.06477*, 2023.

[9] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, “On calibration of modern neural networks,” in *International conference on machine learning*, PMLR, 2017, pp. 1321–1330.

[10] A. Kurbatskaya, A. Jaramillo-Jimenez, J. F. Ochoa-Gomez, K. Brønnick, and A. Fernandez-Quilez, “Assessing gender fairness in eeg-based machine learning detection of parkinson’s disease: A multi-center study,” in *2023 31st European Signal Processing Conference (EUSIPCO)*, 2023, pp. 1020–1024. DOI: 10.23919/EUSIPCO58844.2023.10289837.

[11] A. Kurbatskaya, A. Jaramillo-Jimenez, J. F. Ochoa-Gomez, K. Brønnick, and A. Fernandez-Quilez, “Machine learning-based detection of parkinson’s disease from resting-state eeg: A multi-center study,” in *2023 45th Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, 2023, pp. 1–4. DOI: 10.1109/EMBC40787.2023.10340700.

[12] M. Abdar, F. Pourpanah, S. Hussain, *et al.*, “A review of uncertainty quantification in deep learning: Techniques, applications and challenges,” *Information fusion*, vol. 76, pp. 243–297, 2021.

[13] A. Fernandez-Quilez, “Deep learning in radiology: Ethics of data and on the value of algorithm transparency, interpretability and explainability,” *AI and Ethics*, vol. 3, no. 1, pp. 257–265, 2023.

[14] E. Fong and C. C. Holmes, “Conformal bayesian computation,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 18 268–18 279, 2021.

[15] J. Gawlikowski, C. R. N. Tassi, M. Ali, *et al.*, “A survey of uncertainty in deep neural networks,” *Artificial Intelligence Review*, vol. 56, no. Suppl 1, pp. 1513–1589, 2023.

[16] A. N. Angelopoulos and S. Bates, “A gentle introduction to conformal prediction and distribution-free uncertainty quantification,” *arXiv preprint arXiv:2107.07511*, 2021.

[17] A. Angelopoulos, S. Bates, J. Malik, and M. I. Jordan, “Uncertainty sets for image classifiers using conformal prediction,” *arXiv preprint arXiv:2009.14193*, 2020.

[18] V. Vovk, A. Gammerman, and G. Shafer, *Algorithmic learning in a random world*. Springer, 2005, vol. 29.

[19] G. Litjens, O. Debats, J. Barentsz, N. Karssemeijer, and H. Huisman, “Computer-aided detection of prostate cancer in mri,” *IEEE transactions on medical imaging*, vol. 33, no. 5, pp. 1083–1092, 2014.

[20] R. Cuocolo, A. Comelli, A. Stefano, *et al.*, “Deep learning whole-gland and zonal prostate segmentation on a public mri dataset,” *Journal of Magnetic Resonance Imaging*, vol. 54, no. 2, pp. 452–459, 2021.

[21] A. Fernandez-Quilez, T. Nordstom, T. Eftestol, *et al.*, “Revisiting prostate segmentation in magnetic resonance imaging (mri): On model transferability, degradation and pi-rads adherence,” *medRxiv*, pp. 2023–08, 2023.