PAPER • **OPEN ACCESS**

Advancements in prostate zone segmentation: integrating attention mechanisms into the nnU-Net framework

To cite this article: Aleksas Vaitulevičius *et al* 2024 *Mach. Learn.: Sci. Technol.* **5** 045003

IOP Publishing Mach. Learn.: Sci. Technol. 5 (2024) 045003 https://doi.org/10.1088/2632-2153/ad7f24

MACHINE LEARNING
Science and Technology

OPEN ACCESS

RECEIVED
5 July 2024

REVISED
27 August 2024

ACCEPTED FOR PUBLICATION
24 September 2024

PUBLISHED
7 October 2024

PAPER

Advancements in prostate zone segmentation: integrating attention mechanisms into the nnU-Net framework

Aleksas Vaitulevičius[1,*], Jolita Bernatavičienė[1], Jurgita Markevičiutė[2], Ieva Naruševičiūtė[3], Mantas Trakymas[3] and Povilas Treigys[1]

[1] Institute of Data Science and Digital Technologies, Vilnius University, Akademijos st. 4, Vilnius LT-0841, Lithuania
[2] Institute of Applied Mathematics, Vilnius University, Naugarduko g. 24, Vilnius LT-03225, Lithuania
[3] National Cancer Institute, Vilnius University, Santariškiu st. 1, Vilnius LT–08406, Lithuania
[*] Author to whom any correspondence should be addressed.

E-mail: aleksas.vaitulevicius@mif.stud.vu.lt, jolita.bernataviciene@mif.vu.lt, jurgita.markeviciute@mif.vu.lt, ieva.naruseviciute@nvi.lt, mantas.trakymas@nvi.lt and povilas.treigys@mif.vu.lt

Keywords: mpMRI, deep learning, artificial intelligence, prostate zone segmentation, computer vision

Abstract

Prostate cancer is one of the most lethal cancers in the world. Early diagnosis is essential for successful treatment of prostate cancer. Segmentation of prostate zones in magnetic resonance images is an important task in the diagnosis of prostate cancer. Currently, the state-of-the-art method for this task is no-new U-Net. In this paper, a method to incorporate the attention U-Net architecture into no-new U-Net is proposed and compared with a classical U-net architecture as research. The experimental results indicate that there is no significant statistical difference between the proposed modification of no-new U-Net with the generalizability of the attention mechanism or the ability to achieve more accurate results. Moreover, two novel workflows are proposed for prostate segmentation, transitional zone segmentation and peripheral zone calculation workflow, and separate models for peripheral zone and transitional zone segmentation workflow. These workflows are compared with a baseline single peripheral zone and transitional zone segmentation model workflow. The experimental results indicate that separate models for peripheral zone and transitional zone segmentation workflow generalizes better than the baseline between data sets of different sources. In peripheral zone segmentation separate models for peripheral zone and transitional zone segmentation workflow achieves 1.9% higher median Dice score coefficient than the baseline workflow when using the attention U-Net architecture and 5.6% higher median Dice score coefficient when using U-Net architecture. Moreover, in transitional zone segmentation separate models for peripheral zone and transitional zone segmentation workflow achieves 0.4% higher median Dice score coefficient than the baseline workflow when using the attention U-Net architecture and 0.7% higher median Dice score coefficient when using U-Net architecture. Meanwhile, prostate segmentation, transitional zone segmentation and peripheral zone calculation workflow generalizes worse than the baseline. In peripheral zone segmentation prostate segmentation, transitional zone segmentation and peripheral zone calculation workflow achieves 4.6% lower median Dice score coefficient than the baseline workflow when using the attention U-Net architecture and 3.6% lower median Dice score coefficient when using U-Net architecture. In transitional zone segmentation prostate segmentation, transitional zone segmentation and peripheral zone calculation workflow achieves a similar median Dice score coefficient to the baseline workflow.

1. Introduction

Prostate cancer is one of the most lethal cancers in the world. Research by Bray *et al* in [1] has determined that it is the second most common cancer in men after lung cancer. Early diagnosis is essential for successful treatment of prostate cancer.

One way to screen for prostate cancer is a multiparametric (mp) prostate magnetic resonance imaging (MRI). mpMRI is an accurate way to detect prostate cancer. It uses several different modalities of MRI images to evaluate the prostate gland for cancer. Such as Dynamic Contrast-Enhanced (DCE), Apparent Diffusion Coefficients (ADC) images, diffusion-weighted images (DWI) and T2-weighted images (T2W). The Prostate Imaging Reporting and Data System (PI-RADS) is a structured reporting system for mpMRI, which was introduced in the article [2]. It helps radiologists interpret mpMRI scans and make recommendations for further testing.

Radiologists who use PI-RADS first need to perform a prostate zone segmentation task. They segment the prostate into transitional (TZ) and peripheral (PZ) zones, as different modalities are inspected in different zones. According to the authors of the article [2] the principal modalities for PZ are DWI and ADC while for TZ—T2W modality. The segmentation task of prostate zones is mostly performed in the T2W modality only as T2W is capable to distinguish not only prostate tissue but also the internal anatomy of the prostate, which is difficult to achieve with any other sequence. Therefore, in the experiment provided in this paper only this modality is used. T2W modality is great for visualizing soft tissues as it is very sensitive to the amount of water in the tissue. The data set, which consists of the T2W modality, is a collection of 3D anisotropic images, where each 3D image belongs to a different case.

In general, the method that automates the prostate segmentation task would greatly assist radiologists in detecting prostate cancer. In addition, it has the potential to improve the accuracy of the method used to automate prostate cancer segmentation. Thus, this paper proposes several approaches to automate the prostate zone segmentation task and compares them. The most commonly used approach for this task is the architecture of Deep Neural Network (DNN), U-Net, which was proposed in paper [3] by Ronneberger *et al* and its variations. There are many publications, such as [4] and [5], which provide research on U-Net architecture or it is variation use to solve segmentation tasks. This indicates that the U-Net architecture is currently the state-of-the-art approach to segmentation tasks. Therefore, this paper contains research limited to this architecture and its variations.

The research, provided in the survey [6], indicates that the attention mechanism is an efficient improvement for various computer vision tasks. Therefore, in this paper, the variation of the U-Net, which uses the attention mechanism, is compared with a classical U-Net architecture, The variation is Single Attention U-Net, which was introduced in papers [7, 8] by Oktay *et al* and Schlemper J and *et al*

The specific inspected method is the state-of-the-art method, no-new U-Net (nnU-Net), which was introduced in the paper by Isensee *et al* [9]. In this paper, the modification of this method is proposed. The modification is performed by adding an attention mechanism, creating the aforementioned U-Net variations with hyperparameters, preprocessing, and postprocessing of nnU-Net. Furthermore, research conducted in this article also uses an additional preprocessing step introduced by Jucevicius *et al* [10]. This step converts anisotropic images to isotropic. The experiment results indicate that nnU-Net is more accurate when segmenting cancerous regions with isotropic images.

Lastly, in this paper two different workflows for this task are proposed and compared: separate models for PZ and TZ segmentation and prostate segmentation, TZ segmentation, and PZ calculation. The comparison is also made with the third workflow—single PZ and TZ segmentation model, which is currently used for prostate zone segmentation task.

Therefore, in this paper:

- Two workflows are proposed: separate models for PZ and TZ segmentation, and the prostate segmentation, TZ segmentation and PZ calculation. These workflows are compared with the baseline workflow—single PZ and TZ segmentation model. The results of the experiment presented in this paper show that the separate models for PZ and TZ segmentation workflow generalizes better between data sets of different sources than the others.
- A method to incorporate the Attention U-Net architecture into nnU-Net is proposed. The results of the experiment presented in this paper indicate that there is no significant statistical difference between the proposed modification of nnU-Net and the original nnU-Net.

The remainder of this paper is organized as follows. Section 2 contains related works. Then compared approaches are described in detail in section 3. Section 4 presents the implementation details and experimental results and section 5 concludes.

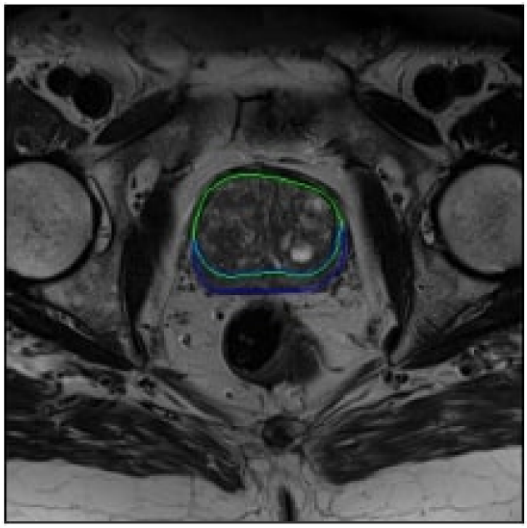

Figure 1. Example of a single slice of T2W modality where area surrounded by green perimeter marks TZ and blue perimeter—PZ.

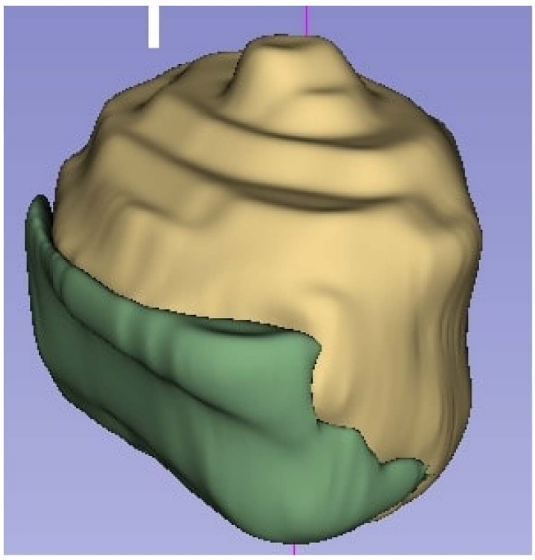

Figure 2. Example of a volumetric segmentation of peripheral (blue) and transitional (yellow) zones.

2. Related works

This section reviews works related to prostate zone segmentation task and papers indicating possible new approaches to this task.

As mentioned, the prostate zone segmentation task consists of segmenting mpMRI into PZ and TZ. Usually, it is done by segmenting only the T2W modality. For example, experiment provided in the paper [5] segments only the T2W modality into prostate zones. The overlay of PZ and TZ over T2W modality is visualized in figure 1. T2W modality is a volumetric data. Thus, the PZ and TZ segmentation data is also volumetric. The example of such segmentation is shown in the figure 2.

A very important indication for prostate zone segmentation task is made in the research provided in paper [11]. This research examines the data variability of TZ and PZ masks in the T2W modality. The experiment in this research indicates high data heterogeneity in the PZ segmentation task, while TZ segmentation is relatively homogeneous. Therefore, this research highlights the need to include representative clinical samples with morphological variation in image databases.

Much research has been done on testing different variations of U-Net in prostate zone segmentation task. One of such researches is described in the paper [5]. The authors of the paper [5] propose a new neural

network architecture, dense-2 U-Net, for segmentation of T2W modality into prostate region and then prostate region into PZ and TZ. The experiment provided in the paper [5] indicates that proposed neural network architecture achieves greater accuracy than classical U-Net architecture. For prostate region segmentation task the proposed dense-2 U-Net architecture achieves 1.6% higher median dice score coefficient than the classical U-Net architecture. Meanwhile, for PZ segmentation task the dense-2 U-Net architecture achieves 2% higher median dice score coefficient than the classical U-Net architecture. Lastly, for TZ segmentation task the dense-2 U-Net architecture achieves 2.2% higher median dice score coefficient than the classical U-Net architecture.

The other research, provided in the paper [12], proposes to use Efficient neural network (ENet) or Efficient Residual Factorized convolution network (ERFNet) architectures for T2W segmentation into prostate region. These architectures have fewer parameters than U-Net, and hence the training and prediction time are much faster. Provided experiments in the paper show that ENet achieves dice score coefficient greater than ERFNet by 0.75%. In addition, the experiments also indicate that the difference between the accuracies achieved by ENet and U-Net is not significant.

Reviews of different researches is provided in the paper [13]. These investigations provide experiments on different data sets available online of T2W modality segmentation into PZ and TZ.

The paper [4] provides an overview of the participants in the Decathlon challenge and their submitted solutions. The goal of this challenge was to create a machine learning (ML) method for segmentation of T2W and ADC modalities into PZ and TZ. The winners of this challenge were the authors of the paper [9]. The paper [9] proposes a method, which in Decathlon challenge achieved mean dice coefficient higher than the competitor who took the second place by 3% for TZ segmentation task and 5% for PZ segmentation task. Therefore, this paper contains the proposed extension of the method provided in the paper [9].

Several analyses have been conducted on the mpMRI data and the prostate cancer detection task. Prostate cancer detection task is a segmentation task which uses mpMRI data, just as prostate zone segmentation task. Therefore, the approaches to solving the prostate cancer detection task are possibly applicable to the prostate zone segmentation task. One of such approaches is attention mechanism. Therefore, this paper contains related works that contain research on prostate cancer detection and the mechanism of attention.

One of such researches is provided in the paper [14], which focuses on segmenting PZ based on the Gleason score (GS). MRI modalities used in the research are ADC and T2W.

The paper [14] proposes using the ProstAttention-Net neural network architecture for this task. This architecture is a U-Net variation that uses an attention mechanism. The main difference from classical U-Net is that it has 2 decoders and 1 encoder. The encoder takes T2W and ADC modalities as input channel wise. The first decoder's output is PZ segmentation, while the second decoder's output is segmentation of PZ into regions with different GS. Each layer of the second decoder has an attention mechanism which uses the corresponding output of the first decoder's layer as a gating signal. The author of the paper [14] compares the mean quadratic weighted Cohen kappa (mean κ) achieved by the ProstAttention-Net architecture with the classical u-Net architecture. The result of the experiment shows that ProstAttention-Net architecture achieves higher mean κ than classical U-Net by 4%.

Moreover, the ProstAttention-Net architecture is further explored in the paper [15]. In the research provided in this article, the first decoder is fit to segment MRI modalities into prostate region and background. The modalities used in this research are also T2W and ADC. The second decoder segments the whole prostate regions into regions with different GS. The author's of the paper [15] compares the mean κ achieved by ProstAttention-Net architecture with other DNN architectures. The experiment presented in this paper shows that ProstAttention-Net achieves higher mean κ than classical U-Net by 9.5%, higher mean κ than Attention U-Net by 7.3%, higher mean κ than E-Net by 0.8% and higher mean κ than DeepLabv3+ by 10%.

The other research whose proposed solution uses an attention mechanism is provided in the paper [16]. The authors of this paper propose a workflow as a solution. This workflow uses a U-Net variation, Dual-Attention U-Net. Moreover, the research also contains a comparison of their proposed workflow with Attention U-Net, nnU-Net, U-Net++, U-SEResNet, and Dual-Attention U-Net architectures on two different test sets. The comparison is made with Area Under the Curve (AUC) metric. The comparison show that their proposed workflow achieves the highest AUC. On the first test set the proposed Dual-Attention U-Net achieves higher AUC than Attention U-Net by 2.4%, higher AUC than nnU-Net by 1.3%, higher AUC than U-Net++ by 3.5%, higher AUC than U-SEResNet by 2.9% and higher AUC than Dual-Attention U-Net by 1.2%. On the second test set the proposed Dual-Attention U-Net achieves higher AUC than Attention U-Net by 5.7%, higher AUC than nnU-Net by 2.1%, higher AUC than U-Net++ by 7.7%, higher AUC than U-SEResNet by 10.4% and higher AUC than Dual-Attention U-Net by 4.2%.

In conclusion, this literature survey indicates that the attention mechanism may improve the accuracy of the model. Therefore, in this paper a modification of nnU-Net with attention mechanism is proposed and experiment results on this modifications is provided.

3. Methodology

3.1. Attention mechanism

As mentioned before in this paper a modified nnU-Net method is proposed where modification is performed by replacing a classical U-Net architecture with U-Net variation which uses attention mechanism. The replaced U-Net variation is Attention U-Net. This architecture uses grid attention mechanism which was proposed in papers [7, 8] by Oktay and *et al* and Schlemper J and *et al* respectively. These authors proposed to add grid-based gating to the attention mechanism proposed in the paper [17] by Jetley *et al*. This gating allows attention gates to be more specific to local regions. The resulting grid attention mechanism schematic is displayed in the figure 3, where:

- x is an input signal, g is a gating signal and y is a result of the attention mechanism.
- $Conv(n, size, stride$ are convolution layers where n is a number of kernels, $size$ is a kernel size and $stride$—is a stride.
- n_i is the hyper parameter of the grid attention mechanism called the number of kernels in intermediate convolution layers.
- sf is a hyper parameter of the grid attention mechanism called the subsample factor.
- n_x is equal to the number of kernels in the layer whose output is the input signal of the grid attention mechanism.
- ReLU and Sigmoid are rectified linear unit and sigmoid activation functions respectively.
- Resampling is a resampling using trilinear interpolation operation.

This mechanism is incorporated into the U-Net architecture, which is used in the nnU-Net method, using it at each skip connection. The input signal for the grid attention mechanism is the skip connection, while the gating signal is the output of the previous upsample block.

This incorporation does not replace or remove any layers in classical U-Net architecture but only adds new layers resulting in additional 2 hyper parameters per grid attention mechanism. The first hyper parameters are subsample factors. The value for these hyper parameters are taken from experiments conducted in papers [7, 8] and for every grid attention mechanism they are equal to 2. The other hyper parameters are number of kernels in intermediate convolution layers. In the experiments performed in [7, 8] the value of this hyper parameter is equal to the number of kernels in the layer whose output is the input signal of the grid attention mechanism. In the research provided in this paper these hyper parameters are adopted in the same manner. The rest of the hyperparameters of the U-Net architecture, which is used in the nnU-Net method, are not changed.

3.2. Workflows for comparison

This paper contains an experiment on 3 different workflows to solve the prostate zone segmentation task:

- Single PZ and TZ segmentation model (SPZTZSM) workflow.
- Separate models for the TZ and PZ segmentation (SM4TZPZS) workflow.
- Prostate Segmentation, TZ Segmentation, and PZ Calculation (PSTZSPZC).

The first workflow is SPZTZSM, which uses a single DNN model to segment PZ and TZ. This is the baseline workflow in the compared workflows. The preprocessing of training data for this workflow takes the prostate region mask and TZ mask as input and then calculates the PZ mask by subtracting the TZ mask area from the prostate region mask. Lastly, PZ and TZ masks are combined into multi-label mask. Training results in a single model that segments T2W into PZ and TZ. This workflow is visualized in the figure 4.

In this paper, the second workflow is proposed, SM4TZPZS. It consists of segmenting PZ and TZ separately. In this workflow, preprocessing of training data is performed for PZ and TZ segmentation tasks. Preprocessing for the TZ segmentation task does not take any additional steps, while for the PZ segmentation task, the PZ mask is calculated. The calculation as usual consists of subtracting TZ mask area from prostate region mask. Two models are then trained—one for PZ segmentation task and the other for TZ segmentation task. The result is 2 models, one of them segments T2W into PZ and the other—to TZ. This workflow is visualized in figure 5.

The other workflow proposed in this article is PSTZSPZC. This workflow consists of consecutive steps: segmenting the prostate region, segmenting TZ in the prostate region, and calculating PZ. The first step does

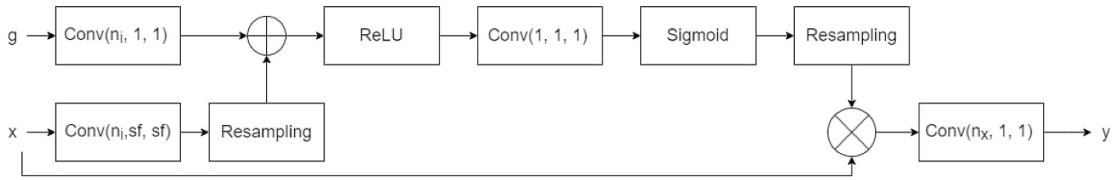

Figure 3. Schematic of the attention mechanism.

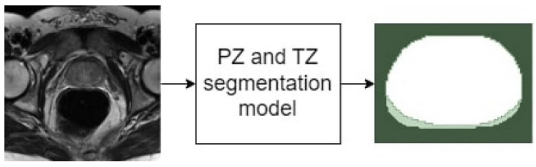

Figure 4. Prostate zone segmentation with a SPZTZSM workflow.

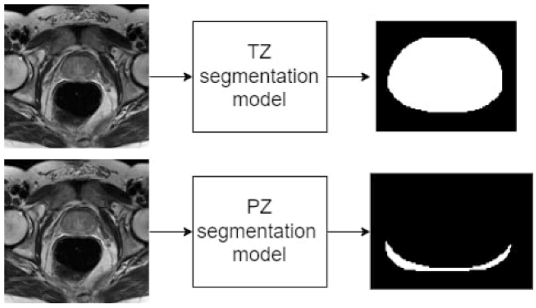

Figure 5. Prostate zone segmentation with SM4TZPZS workflow.

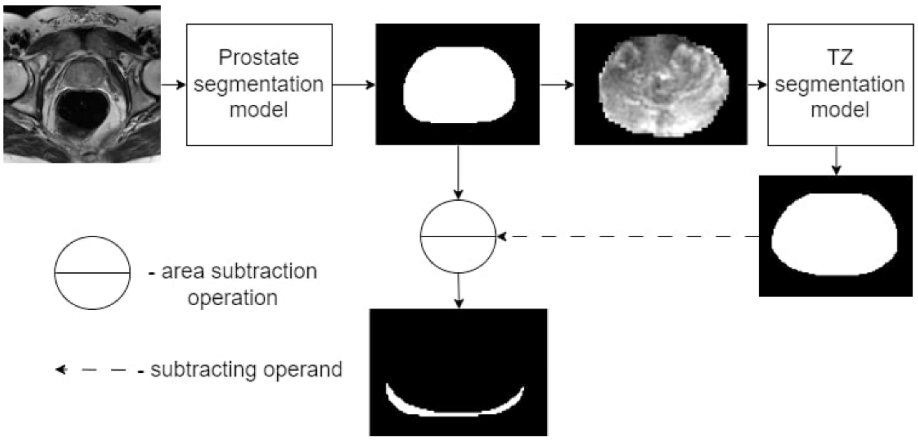

Figure 6. Prostate zone segmentation with PSTZSPZC workflow.

not perform any additional transformations in the preprocessing and applies nnU-Net method to segment T2W into prostate region and background. The second step consists of setting the T2W voxel values to zero if they do not overlap the predicted prostate region. During training, the TZ mask is adjusted by calculating the union between the TZ and the segmented prostate region. Lastly, the second step applies the nnU-Net method to segment T2W into TZ. The third step calculates PZ by subtracting the predicted TZ area from the predicted prostate region. The PSTZSPZC models workflow consists of two models, one of them segments T2W into the prostate region and the other to TZ. This workflow is visualized in the figure 6.

3.3. Performance measures

Ensuring the reliability of a model hinges on the scrutiny of its performance metrics. One of such metrics is the dice similarity coefficient (DSC) [18]. This metric is employed in many experiments on prostate zone segmentation task. Such as experiments presented in the papers [4, 5, 12]. Moreover, DSC is incorporated into the loss function of the nnU-net method. Therefore, this metric is used in the experiment presented in this paper. The segmented data is volumetric, therefore the calculation of DSC is performed on volumetric data. DSC is calculated separately for PZ and TZ. Dice similarity coefficient is expressed by the formula displayed in the equation (1). In this equation TP is the number of voxels, which are correctly assigned to either PZ or TZ. Further, FN represents the number of voxels which are not assigned to either PZ or TZ while the ground truth indicates that those voxels do belong to the PZ or TZ. Lastly, FP is the number of voxels identified as either PZ or TZ, while ground truth indicates that these voxels do not belong to PZ or TZ

$$\mathrm{DSC} = \frac{2 \times \mathrm{TP}}{2 \times \mathrm{TP} + \mathrm{FP} + \mathrm{FN}}. \tag{1}$$

It is important to note that in the research of the articles [5, 12], not only DSC was used to evaluate the results. In the article [5] mean relative absolute volume difference (MRAVD) and mean Hausdorff distance were used. Meanwhile, sensitivity, positive predictive value, volume overlap error, and volumetric difference were used in the article [12]. In addition, the research results provided in the article [19] are evaluated not only with DSC, but also with volumetric difference. However, the authors of both articles [5, 12] emphasize DSC over the other performance metrics. Therefore, the results of the experiment, provided in this paper, are evaluated only on DSC.

In addition, two-sided Wilcoxon signed rank and Friedman tests are used for related or paired groups to test statistical significance between compared approaches. Two-sided Wilcoxon signed rank test was introduced in the article [20], while Friedman test was introduced by Friedmann in the paper [21] Both are non-parametric tests, so they are appropriate to assess whether the mean ranks of the group population differ for small and non-normally distributed samples. Two-sided Wilcoxon signed rank test is used for two sample tests, while the Friedman test is an extension for more than two sample tests. The two-sided Wilcoxon test ranks the differences between paired observations and then assesses whether the ranks of positive and negative differences differ significantly. Meanwhile, Friedman's test assesses whether there are significant differences in the rankings across the groups, accounting for the repeated measures design.

4. Experiments

4.1. Data set

The data set, used in the experiment presented in this paper, is a combination of two data sets that come from two different sources. The first source is the Decathlon challenge, the results of this challenge is described in the paper [4]. It was provided by Radboud University and Nijmegen Medical Centre. This data set is publicly available and it consists of 31 cases with labeled TZ and PZ masks. The prostate region mask is acquired by calculating union of PZ and TZ areas. The data set is upside-down on the second axis. Therefore, T2W modality and masks were flipped by performing direct discrete Fourier transformation.

The other source is ProstateX challenge, provided in the [22]. This data set is publicly available and consists of 98 cases with TZ, PZ, fibromuscular stroma, and the distal prostatic urethra labeled. The labeling was performed by a medical student experienced in prostate segmentation, under the guidance of an expert urologist who provided instruction and conducted a final review of the labels.

Each case's T2W modality of the data set is saved as separate files, where each file contains a slice of the T2W modality. Therefore, the first action performed on this data set is constructing volumetric data from those files. Meanwhile, all the masks in this data set are saved as a single file by concatenating them on a third dimension. The concatenation is performed in the following order: TZ, PZ, prostatic urethra, and fibromuscular stroma. Thus, masks are separated into different files. Further the prostate region mask is acquired by calculating union of all available masks. Meanwhile, the TZ mask is acquired by performing the following steps. In the first step, slices are determined that contain TZ and prostatic urethra masks while not PZ. In the second step, slices are determined, which contain TZ, PZ and prostatic urethra masks. In the third step TZ mask is dilated with ball shaped kernel of radius equal to 4. The dilation operation is described in the paper [23]. In the fourth step the intersection of the second step's result and third step's result is calculated. Further, the union is calculated of the first step's result, the fourth step's result, original TZ mask and fibromuscular stroma's mask.

Eighty-eight cases from the ProstateX source data set are randomly selected for the training data set. Meanwhile, the remaining 10 cases are used for the testing. Lastly, all 31 cases from the Decathlon source data set are used in the test set.

Several augmentation steps are then performed on the data set. Firstly, the data are converted from anisotropic to isotropic data by applying the step proposed by Jucevicius *et al* in the paper [10]. Then the preprocessing part of nnU-Net is performed. It begins by processing the training data to extract the dataset fingerprint, which encompasses various statistical details such as image intensity values, spacing information, image sizes, and class ratios. This fingerprint is then utilized to determine inferred parameters like batch and patch sizes, target image resampling, target voxel spacings, and the image normalization technique, all based on predefined heuristic rules. These parameters are used in training and prediction pipelines of nnU-Net method.

4.2. Experiment setup

As mentioned in the previous subsection, the data set comes from 2 different data sources. 88 patients from the ProstateX source are used for training and 10 patients for testing. Meanwhile all 31 patients from Decathlon source are used only for testing. In total 6 approaches are compared: all 3 workflows described in the 3.2 section by using nnU-Net with classical U-Net architecture and the same workflows using nnU-Net with Attention U-Net architecture. The results of each approach are evaluated by calculating Dice similarity coefficients on the test sets. The evaluations of TZ and PZ segmentation are performed separately.

In this paper, the evaluation of the results on the test set of ProstateX and Decathlon sources is performed separately. The evaluation on ProstateX source test set allows one to determine if there is a statistically significant difference between the compared approaches and which approach achieves the highest accuracy. The evaluation on the Decathlon source test set allows one to determine if any of the compared approaches generalizes better than the others by achieving statistically significant higher accuracy on the test set of a different source.

5. Results

5.1. Results on ProstateX source test set

Firstly, statistical tests are performed on the results of the ProstateX source test set. The compared approaches consist of two variables, DNN architecture and workflow, tests on each workflow are conducted separately and then on each DNN architecture separately. The tests on each workflow allows to compare DNN architectures, while tests on each DNN architecture—to compare workflows. As there are only 2 DNN architectures compared, the statistical test chosen to compare DNN architectures is the two-sided Wilcoxon signed rank test. Meanwhile, as there are more than two compared workflows, the chosen statistical test for comparison of the workflows is Friedmann test. The confidence level chosen for both of these tests is 95%. Therefore, the significance level is 0.05. The statistical tests allows to determine the correct hypothesis.

Table 1 contains the Friedmann and Wilcoxon signed rank test results. Friedmann results are acquired by performing tests on each DNN architecture, while Wilcoxon signed rank test results are performed on each workflow. The column test represents which test, Friedmann or Wilcoxon signed rank, is performed. Furthermore, the column prostate zone indicates on which zone segmentation results are acquired. Lastly, the column DNN architecture/workflow represents the DNN architecture on which Friedmann test is conducted and the workflow on which Wilcoxon signed rank test is performed.

As none of the statistical test results have p-value lower than the chosen significance level, the hypothesis H_0 is not rejected. Therefore, there are no statistically significant differences between the compared approaches.

Furthermore, the aggregated test results of ProstateX source test set are provided in the tables 2 and 3. Aggregation is performed by calculating the mean, standard deviation, median, and Median Absolute Deviation (MAD). All of these values are shown in both tables. Moreover, column workflow indicates compared workflow, while column DNN architecture—compared DNN architecture. Table 3 contains the results of the TZ segmentation while table 2—the results of the PZ segmentation. Just as statistical tests indicated, both tables show no significant difference between compared approaches.

5.2. Results on Decathlon source test set

Just like with the results of the ProstateX source test set, first of all statistical tests are conducted on the results of the Decathlon source test set. The same confidence level of 95% is used, and the tests on each DNN architecture and workflow are conducted separately. Table 4 contains the Friedmann and Wilcoxon signed rank test results. Friedmann results are acquired by performing tests on each DNN architecture, while Wilcoxon signed rank test results are performed on each workflow. Just as in table 1 the column test represents which test is performed, the column prostate zone indicates on which zone segmentation results are acquired and the column DNN architecture/workflow represents the DNN architecture on which Friedmann test is conducted and the workflow on which Wilcoxon signed rank test is performed.

Table 1. Friedmann and Wilcoxon signed rank test results. Friedmann tests are used to compare workflows. Meanwhile, Wilcoxon signed rank tests are used to compare DNN architectures. The compared results are acquired on ProstateX source test set.

Test	Prostate zone	DNN architecture/workflow	p-value
Friedmann	TZ	Attention U-Net	0.272 532
		U-Net	0.272 532
	PZ	Attention U-Net	0.406 570
		U-Net	0.904 837
Wilcoxon signed rank	TZ	SM4TZPZS	0.275 391
		SPZTZSM	0.130 859
		PSTZSPZC	0.232 422
	PZ	SM4TZPZS	0.130 859
		SPZTZSM	0.322 266
		PSTZSPZC	0.556 641

Table 2. Mean, standard deviation, median and MAD of PZ segmentation results on ProstateX source test set. Just as statistical tests indicated, there is no significant difference between compared approaches on PZ segmentation. The approaches are compared independently for PZ and TZ segmentation.

DNN architecture	workflow	Mean+/−std	Median+/−MAD
Attention U-Net	SM4TZPZS	0.819+/−0.060	0.823+/−0.022
	SPZTZSM	0.825+/−0.052	0.815+/−0.025
	PSTZSPZC	0.819+/−0.055	0.815+/−0.034
U-Net	SM4TZPZS	0.821+/−0.060	0.822+/−0.026
	SPZTZSM	0.827+/−0.050	0.813+/−0.026
	PSTZSPZC	0.821+/−0.058	0.824+/−0.012

Table 3. Mean, standard deviation, median and MAD of TZ segmentation results on ProstateX source test set. Just as statistical tests indicated, there is no significant difference between compared approaches on TZ segmentation. The approaches are compared independently for PZ and TZ segmentation.

DNN architecture	workflow	Mean+/−std	Median+/−MAD
Attention U-Net	SM4TZPZS	0.919+/−0.037	0.928+/−0.041
	SPZTZSM	0.914+/−0.039	0.923+/−0.041
	PSTZSPZC	0.918+/−0.032	0.922+/−0.024
U-Net	SM4TZPZS	0.918+/−0.037	0.926+/−0.035
	SPZTZSM	0.918+/−0.036	0.927+/−0.036
	PSTZSPZC	0.919+/−0.032	0.925+/−0.029

Table 4. Friedmann and Wilcoxon signed rank test results. Friedmann tests are used to compare workflows. Meanwhile, Wilcoxon signed rank tests are used to compare DNN architectures. The compared results are acquired on Decathlon source test set.

Test	Prostate zone	DNN architecture/workflow	p-value
Friedmann	TZ	Attention U-Net	0.014 504
		U-Net	0.272 532
	PZ	Attention U-Net	0.000 002
		U-Net	0.000 041
Wilcoxon signed rank	TZ	SM4TZPZS	0.004 682
		SPZTZSM	0.036 826
		PSTZSPZC	0.158 855
	PZ	SM4TZPZS	0.543 524
		SPZTZSM	0.709 644
		PSTZSPZC	0.108 073

The results of the Friedmann test shown in table 4 indicate that there is a statistically significant difference between the compared workflows on the segmentation of PZ, as the p-values of these results are lower than the chosen significance level and, therefore, hypothesis H_0 is rejected. Meanwhile, it is not clear if there is a statistically significant difference in TZ segmentation as hypothesis H_0 is only rejected when the Friedmann test on Attention U-Net DNN architecture. Therefore, further evaluation of workflows is conducted with an emphasis on PZ segmentation.

Table 5. Posthoc Wilcoxon signed rank test results when comparing workflows on Decathlon source test set PZ segmentation.

DNN architecture	Comparison	p-value
Attention U-Net	SPZTZSM—PSTZSPZC	0.000 289
	SM4TZPZS—PSTZSPZC	0.000 009
	SPZTZSM—SM4TZPZS	0.068 382
U-Net	SPZTZSM—PSTZSPZC	0.000 182
	SM4TZPZS—PSTZSPZC	0.000 082
	SPZTZSM—SM4TZPZS	0.031 113

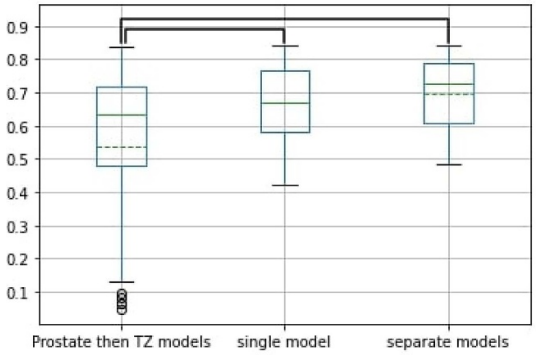

Figure 7. Boxplots of Dice similarity coefficients evaluated on the test set, PZ segmentation. Evaluated DNN architecture—U-Net.

As Friedmann test results, displayed in the table 4, indicate that there is a significant statistical difference between compared workflows on PZ segmentation, posthoc Wilcoxon signed rank tests are conducted on Decathlon source test set PZ segmentation in order to determine the cause of difference between workflows. As there are three compared workflows, the confidence level is increased to 96.667%. Table 5 contains the results of this statistical test where the column DNN architecture represents the compared DNN architecture, the column DNN architecture represents the compared DNN architecture, the column comparison indicates which workflows are compared and the column p-value—the p-value of the test.

The results of these post hoc Wilcoxon tests indicate that there is a statistically significant difference between the PSTZSPZC workflow and the other workflows, as the p-values of these results are lower than the chosen significance level and, therefore, hypothesis H_0 is rejected. Meanwhile, the results of these post hoc Wilcoxon tests indicate that there is no statistically significant difference between the workflows of an SPZTZSM and an SM4TZPZS, as the p-values of this result are higher than the chosen significance level, and therefore the hypothesis H_0 is not rejected.

Furthermore, since the Wilcoxon test results for PZ segmentation displayed in table 4 do not have a p-value lower than the chosen significance level, the hypothesis H_0 is not rejected. Thus, there are no statistically significant differences between the compared DNN architectures when segmenting PZ. Meanwhile, it is not clear if there is a statistically significant difference in TZ segmentation as hypothesis H_0 is only rejected when conducting a Wilcoxon test on the PSTZSPZC workflow. Therefore, further inspection of test results is focused on comparing workflows.

5.3. Results on Decathlon source test set: comparing workflows

As the Friedmann test results provided in table 4 indicate a significant statistical difference between the workflows on PZ segmentation, the PZ segmentation test results are visualized as boxplots in figures 7 and 8. Both of the boxplots are calculated from results of Decathlon source test set. The figure 7 contains boxplots calculated from test results acquired with U-Net architecture, while figure 8—Attention U-Net architecture. The horizontal axis contains workflow used in the experiment, while vertical one—aggregated DSC. The dashed line represents the mean DSCs, whereas the solid ones represent the median DSCs. The black lines connecting the box plots indicate the workflows between which there is a statistically significant difference. Both of these figures indicate that the SM4TZPZS workflow achieves the highest DSC while PSTZSPZC achieves the lowest.

Furthermore, the aggregated test results of Decathlon source test set are provided in the tables 6 and 7. Similarly as in tables 2 and 3, the aggregated results are mean, standard deviation, median, and MAD. Moreover, column workflow indicates compared workflow, while column DNN architecture—compared

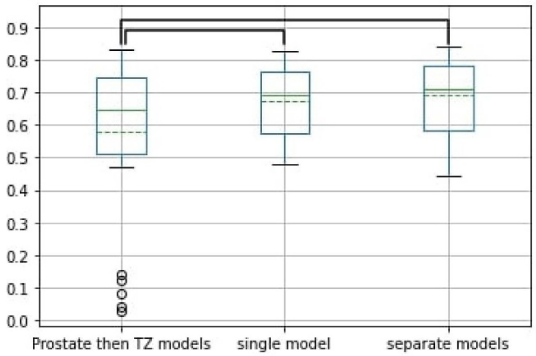

Figure 8. Boxplots of Dice similarity coefficients evaluated on the test set, PZ segmentation. Evaluated DNN architecture—Attention U-Net.

Table 6. Mean, standard deviation, median and MAD of PZ segmentation results on Decathlon source test set. The approaches are compared independently for PZ and TZ segmentation. Aggregated results written in bold indicate the highest achieved DSC for each DNN architecture. For both DNN architectures the highest achieved DSC is with SM4TZPZS workflow.

DNN architecture	Workflow	Mean+/−std	Median+/−MAD
Attention U-Net	SM4TZPZS	**0.691+/−0.113**	**0.709+/−0.143**
	SPZTZSM	0.675+/−0.106	0.690+/−0.148
	PSTZSPZC	0.578+/−0.244	0.644+/−0.118
U-Net	SM4TZPZS	**0.698+/−0.108**	**0.725+/−0.177**
	SPZTZSM	0.668+/−0.117	0.669+/−0.121
	PSTZSPZC	0.538+/−0.264	0.633+/−0.099

Table 7. Mean, standard deviation, median and MAD of TZ segmentation results on Decathlon source test set. The approaches are compared independently for PZ and TZ segmentation. Aggregated results written in bold indicate the highest achieved DSC for each DNN architecture. For both DNN architectures the highest achieved DSC is with SM4TZPZS workflow.

DNN architecture	Workflow	Mean+/−std	Median+/−MAD
Attention U-Net	SM4TZPZS	**0.806+/−0.233**	**0.887+/−0.000**
	SPZTZSM	0.803+/−0.215	0.883+/−0.000
	PSTZSPZC	0.771+/−0.268	0.883+/−0.003
U-Net	SM4TZPZS	**0.783+/−0.250**	**0.886+/−0.001**
	SPZTZSM	0.773+/−0.258	0.879+/−0.000
	PSTZSPZC	0.782+/−0.251	0.882+/−0.001

Table 8. Mean, standard deviation, median and MAD of segmentation results with SM4TZPZS workflow on Decathlon source test set. Aggregated results written in bold indicate the highest achieved DSC for each prostate zone. For TZ the highest DSC is with Attention U-Net DNN architecture, while for PZ - U-Net architecture.

Prostate zone	DNN architecture	Mean+/−std	Median+/−MAD
TZ	Attention U-Net	**0.806+/−0.233**	**0.887+/−0.000**
	U-Net	0.783+/−0.250	0.886+/−0.001
PZ	Attention U-Net	0.691+/−0.113	0.709+/−0.143
	U-Net	**0.698+/−0.108**	**0.725+/−0.177**

DNN architecture. Table 7 contains the results of the TZ segmentation while table 6 contains the results of the PZ segmentation. Both tables show that the highest DSC is achieved by SM4TZPZS workflow. Therefore, a further evaluation of the test results is performed only on this workflow.

5.4. Results on Decathlon source test set: comparing DNN architectures

As tables 6 and 7 indicate that the separate model workflow achieves the highest DSC, the aggregated test results with the separate model workflow are provided in table 8. This table contains the results of the Decathlon source test set. The column DNN architecture represents the compared DNN architecture. Meanwhile, the column prostate zone indicates on which zone segmentation results are acquired. The

performed aggregation is the same as in the tables 6 and 7. As statistical tests indicated, the difference in DSC achieved between the compared DNN architectures is very small. However, the table 8 indicates that Attention U-Net achieves higher DSC on TZ segmentation while U-Net achieves higher DSC on PZ segmentation.

6. Conclusion

The results of the experiment, presented in this paper, evaluate possible improvements to the method to solve the prostate zone segmentation task. The accurate method to solve this task is important because it is necessary to differentiate between zones for an accurate diagnosis of prostate cancer due to the different views of the tumor in different MRI modalities. This method could then be used to automate the prostate segmentation task, significantly aiding radiologists in detecting prostate cancer. Moreover, it has the potential to enhance the accuracy of automated prostate cancer segmentation methods.

In this paper, two workflows for the prostate zone segmentation task, SM4TZPZS and PSTZSPZC, are proposed. Furthermore, these workflows are compared with a baseline workflow, SPZTZSM workflow.

In every single comparison, SM4TZPZS workflow achieves higher mean and median DSC on Decathlon source data set than the baseline, SPZTZSM workflow. In PZ segmentation results when using the attention U-Net architecture, the mean DSC is higher by 1.6%, while the median DSC is higher by 1.9%. Furthermore, when using the U-Net architecture, the mean DSC is higher by 3%, while the median DSC is higher by 5.6%. Meanwhile in TZ segmentation results when using attention U-Net architecture mean DSC is higher by 0.3%, while median DSC is higher by 0.4%. Moreover, when using U-Net architecture mean DSC is higher by 1%, while median DSC is higher by 0.7%. Although the statistical test indicated no significance between these two workflows, the persistence of the results indicates that SM4TZPZS generalizes between data sets from different sources better than the baseline SPZTZSM workflow.

Meanwhile, PSTZSPZC workflow in PZ segmentation, every single comparison achieves lower mean and median DSC than the SPZTZSM workflow on Decathlon source data set. Firstly, when using attention U-Net architecture mean DSC is lower by 9.7%, while median DSC is lower by 4.6%. Furthermore, when using U-Net architecture mean DSC is lower by 13%, while median DSC is lower by 3.6%. Meanwhile, when comparing the PSTZSPZC workflow and SPZTZSM in TZ segmentation, the results are very similar. Moreover, the statistical tests indicate that comparisons made on PZ segmentation are statistically significant while on TZ—not significant. This experiment indicates that the PSTZSPZC workflow generalizes between data sets of different sources worse than a baseline.

Furthermore, in this paper, a method is proposed to incorporate the Attention U-Net DNN architecture into the nnU-Net method and compared to the original nnU-Net. The statistical tests performed indicate that there is no significant statistical difference between the proposed modification of nnU-Net and the original nnU-Net in neither generalizability nor capability to achieve more accurate results.

Lastly, the performed statistical tests indicate that comparing our proposed approaches on the test set, whose source is the same as the training set, bears no statistically significant difference between them.

The results obtained show potential further research directions on the PZ and TZ segmentation task. One of them is testing incorporation of other attention mechanism based architectures into nnU-Net such as Dual-Attention U-Net, Grid Attention U-Net and Multi Dimensional Attention U-Net. The other possible research direction is to further test the effects of using different data sources on the PZ and TZ segmentation task. Another research direction is to use Generative Artificial Networks (GANs) to improve the accuracy of the proposed methods for the PZ and TZ segmentation task. Lastly, the performance metric used in the experiment provided in this paper, DSC, has some known limitations. For instance, DSC may not be suitable for small structures because even a single pixel discrepancy can significantly affect the metric. Another limitation could be the lack of sensitivity to shape differences. Therefore, the further research direction can include evaluating the experiment results with different metrics such as Hausdorff distance, intersection over union or change in center of mass.

Data availability statement

No new data were created or analyzed in this study.

Acknowledgments

Publication/Research is funded by Research Council of Lithuania under the Programme ‘University Excellence Initiatives’ of the Ministry of Education, Science and Sports of the Republic of Lithuania (Measure No. 12-001-01-01-01 ‘Improving the Research and Study Environment’). Project No.: S-A-UEI-23-11.

The authors are thankful for the high performance computing resources provided by the Information Technology Research Center of Vilnius University.

ORCID iD

Aleksas Vaitulevičius https://orcid.org/0000-0003-1323-5098

References

[1] Sung H, Ferlay J, Siegel R L, Laversanne M, Soerjomataram I, Jemal A and Bray F 2021 Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries *CA Cancer J. Clin.* **71** 209–49
[2] Alqahtani S, Wei C, Zhang Y, Szewczyk-Bieda M, Wilson J, Huang Z and Nabi G 2020 Prediction of prostate cancer Gleason score upgrading from biopsy to radical prostatectomy using pre-biopsy multiparametric MRI PIRADS scoring system *Sci. Rep.* **10** 7722
[3] Ronneberger O, Fischer P and Brox T 2015 U-net: convolutional networks for biomedical image segmentation *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th Int. Conf. (Munich, Germany, 5–9 October 2015), Proc. Part III* vol 18 (Springer) pp 234–41
[4] Antonelli M *et al* 2022 The medical segmentation decathlon *Nat. Commun.* **13** 4128
[5] Aldoj N, Biavati F, Michallek F, Stober S and Dewey M 2020 Automatic prostate and prostate zones segmentation of magnetic resonance images using DenseNet-like U-net *Sci. Rep.* **10** 14315
[6] Guo M H, Xu T X, Liu J J, Liu Z N, Jiang P T, Mu T J, Zhang S H, Martin R R, Cheng M M and Hu S M 2022 Attention mechanisms in computer vision: a survey *Comput. Vis. Media* **8** 331–68
[7] Oktay O *et al* 2018 Attention U-Net: learning where to look for the pancreas (arXiv:1804.03999)
[8] Schlemper J, Oktay O, Schaap M, Heinrich M, Kainz B, Glocker B and Rueckert D 2019 Attention gated networks: learning to leverage salient regions in medical images *Med. Image Anal.* **53** 197–207
[9] Isensee F, Jaeger P F, Kohl S A, Petersen J and Maier-Hein K H 2021 nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation *Nat. Methods* **18** 203–11
[10] Jucevičius J, Treigys P, Bernatavičienė J, Trakymas M and Naruševičiūtė I 2023 Usage of isotropic MRI images improves prostate cancer localization results *Balt. J Mod. Comput.* **11** 703–25
[11] Montagne S, Hamzaoui D, Allera A, Ezziane M, Luzurier A, Quint R, Kalai M, Ayache N, Delingette H and Renard-Penna R 2021 Challenge of prostate MRI segmentation on T2-weighted images: inter-observer variability and impact of prostate morphology *Insights Imaging* **12** 1–12
[12] Comelli A, Dahiya N, Stefano A, Vernuccio F, Portoghese M, Cutaia G, Bruno A, Salvaggio G and Yezzi A 2021 Deep learning-based methods for prostate segmentation in magnetic resonance imaging *Appl. Sci.* **11** 782
[13] Gillespie D, Kendrick C, Boon I, Boon C, Rattay T and Yap M H 2020 Deep learning in magnetic resonance prostate segmentation: a review and a new perspective (arXiv:2011.07795)
[14] Duran A, Jodoin P M and Lartizien C 2020 Prostate cancer semantic segmentation by Gleason score group in bi-parametric MRI with self attention model on the peripheral zone *Medical Imaging With Deep Learning* (PMLR) pp 193–204
[15] Duran A, Dussert G, Rouvière O, Jaouen T, Jodoin P M and Lartizien C 2022 ProstAttention-Net: a deep attention model for prostate cancer segmentation by aggressiveness in MRI scans *Med. Image Anal.* **77** 102347
[16] Saha A, Hosseinzadeh M and Huisman H 2021 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.* **73** 102155
[17] Jetley S, Lord N A, Lee N and Torr P H 2018 Learn to pay attention (arXiv:1804.02391)
[18] Klein S, Van Der Heide U A, Lips I M, Van Vulpen M, Staring M and Pluim J P 2008 Automatic segmentation of the prostate in 3D MR images by atlas matching using localized mutual information *Med. Phys.* **35** 1407–17
[19] Joskowicz L, Cohen D, Caplan N and Sosna J 2019 Inter-observer variability of manual contour delineation of structures in CT *Eur. Radiol.* **29** 1391–9
[20] Wilcoxon F 1947 Probability tables for individual comparisons by ranking methods *Biometrics* **3** 119–22
[21] Friedman M 1940 A comparison of alternative tests of significance for the problem of m rankings *Ann. Math. Stat.* **11** 86–92
[22] Meyer A, Schindele D, Von Reibnitz D, Rak M, Schostak M and Hansen C 2020 Prostatex zone segmentations *The Cancer Imaging Archive* (https://doi.org/10.7937/TCIA.NBB4-4655)
[23] Vincent L 1991 Morphological transformations of binary images with arbitrary structuring elements *Signal Process.* **22** 3–23