https://doi.org/10.15388/vu.thesis.975
https://orcid.org/0000-0003-1323-5098

VILNIUS UNIVERSITY

Aleksas Vaitulevičius

Research and Modeling of Digital Signals for the Detection of Cancerous Areas of Tissue in Different MRI Modalities

DOCTORAL DISSERTATION

Natural Sciences,
Informatics (N 009)

VILNIUS 2026

The dissertation was prepared between 2021 and 2025 at Vilnius University.

Academic supervisor – Prof. Dr. Povilas Treigys (Vilnius University, Natural Sciences, Informatics – N 009).

This doctoral dissertation will be defended at a public meeting of the Dissertation Defence Panel:

Chairman – Prof. Dr. Remigijus Paulavičius (Vilnius University, Natural Sciences, Informatics – N 009).
Members:
Assoc. Prof. Dr. Viktor Medvedev (Vilnius University, Natural Sciences, Informatics – N 009),
Assoc. Prof. Dr. Pavel Stefanovič (Vilnius Gediminas Technical University, Natural Sciences, Informatics – N 009),
Assoc. Prof. Dr. Stepas Toliautas (Vilnius University, Natural Sciences, Physics – N 002),
Dr. Paulina Trybek (The University of Silesia in Katowice, Poland, Natural Sciences, Biophysics – N 011).

The dissertation shall be defended at a public meeting of the Dissertation Defense Panel at 1:00 p.m. on 21st of September 2026 in room 203 of the Institute of Data Science and Digital Technologies of Vilnius University.
Address: Akademijos g. 4, LT-04812, Vilnius, Lithuania
Tel. +370 5 210 9300; e-mail: info@mii.vu.lt

https://doi.org/10.15388/vu.thesis.975
https://orcid.org/0000-0003-1323-5098

VILNIAUS UNIVERSITETAS

Aleksas Vaitulevičius

Skaitmeninių signalų tyrimas ir modeliavimas audinio vėžinėms zonoms aptikti skirtinguose MRT modalumuose

DAKTARO DISERTACIJA

Gamtos mokslai,
Informatika (N 009)

VILNIUS 2026

Disertacija rengta 2021–2025 metais Vilniaus universitete.

Mokslinis vadovas – prof. dr. Povilas Treigys (Vilniaus universitetas, gamtos mokslai, informatika – N 009).

Gynimo taryba:
Pirmininkas – prof. dr. Remigijus Paulavičius (Vilniaus universitetas, gamtos mokslai, informatika – N 009).
Nariai:
doc. dr. Viktor Medvedev (Vilniaus universitetas, gamtos mokslai, informatika – N 009),
doc. dr. Pavel Stefanovič (Vilniaus Gedimino technikos universitetas, gamtos mokslai, informatika – N 009),
doc. dr. Stepas Toliautas (Vilniaus universitetas, gamtos mokslai, fizika – N 002),
dr. Paulina Trybek (Katovicų Silezijos universitetas, Lenkija, gamtos mokslai, biofizika – N 011).

Disertacija ginama viešame Gynimo tarybos posėdyje 2026 m. rugsėjo 21 d. 13:00 val. Vilniaus universiteto Duomenų mokslo ir skaitmeninių technologijų instituto 203 auditorijoje. Adresas: Akademijos g. 4, LT-04812, Vilnius, Lietuva, tel. +370 5 210 9300; el. paštas: info@mii.vu.lt.

ABSTRACT

Prostate Cancer (PCa) is one of the most prevalent malignancies worldwide, as reflected in recent global cancer statistics. A critical step in its diagnosis is the assessment of multiparametric Magnetic Resonance Imaging (mpMRI). In recent years, various deep learning approaches have been investigated to enhance the detection and segmentation of PCa using mpMRI. Despite these advances, prostate cancer segmentation remains a challenging task. This thesis investigates PCa segmentation ensembles to further improve the segmentation performance and robustness. In particular, it focuses on the state-of-the-art medical image segmentation method no-new U-Net (nnU-Net).

A substantial body of research has demonstrated that integrating prostate zonal information into prostate cancer segmentation approaches improves the segmentation and detection accuracy. Therefore, this thesis also investigates prostate zone segmentation workflows to further improve the accuracy of the prostate zone segmentation. It proposes two novel prostate zone segmentation workflows, both of which employ Deep Neural Network (DNN) models trained with the nnU-Net method. The empirical comparison indicates that novel Separate Models for the Transition Zone, Peripheral Zone Segmentation workflow achieves the most accurate prostate zone segmentation results. Moreover, the thesis investigates the impact of substituting the original U-Net architecture used in the nnU-Net method with U-Net variation employing attention mechanisms on the prostate zone segmentation. These variations include grid attention U-Net, single attention U-Net, multidimensional attention U-Net and dual-attention U-Net. The empirical experiment did not indicate a statistical significance of the difference between prostate zone segmentation results of compared DNN architectures. The dataset used for the comparative experiment on the prostate zone segmentation task consists of two publicly available datasets, ProstateX and Decathlon challenge datasets.

This thesis proposes four novel prostate cancer segmentation ensembles which employ DNN models trained with the nnU-Net method. All of these ensembles are structured to utilize prostate zone masks. The empirical experiment indicates that the most accurate results are achieved by the ensemble which combines the results of the DNN model specialized in segmenting lesions in the peripheral zone, and the model

which segments PCa in the whole prostate gland and uses prostate zone masks as a separate input channel. Additionally, this thesis proposes the modification of the Dice Score Coefficient (DSC) metric, tolerant Dice Score Coefficient, DSC_τ. This metric is designed to account for a high inter-observer variability which is present in the ground truth of PCa lesion segmentations. The empirical comparison of DSC and the novel metric shows that DSC_τ consistently yields higher scores than the conventional DSC, indicating that the proposed metric accounts for tolerable boundary discrepancies. The dataset used for the empirical experiments on the prostate cancer segmentation was provided by the National Cancer Institute and is not publicly available.

DSC_τ	tolerant Dice Score Coefficient.
ADC	Apparent Diffusion Coefficient.
AUROC	Area Under the ROC curve.
CNN	Convolutional Neural Network.
CT	Computer Tomography.
DCE	Dynamic Contrast-Enhanced.
DL	Deep Learning.
DNN	Deep Neural Network.
DSC	Dice Score Coefficient.
DWI	Diffusion-Weighted Imaging.
ENET	Efficient Neural Network.
ERFNet	Efficient Residual Factorized convolution Network.
FP	False Positive.
FROC	Free-Response ROC.
GS	Gleason Score.
MAD	Median Absolute Deviation.
mpMRI	multi-parametric Magnetic Resonance Imaging.
MRI	Magnetic Resonance Imaging.
nnU-Net	no-new U-Net.
NVI	National Cancer Institute.
PCa	Prostate Cancer.
PI-RADS	Prostate Imaging Reporting and Data System.
PSA	Prostate Specific Antigen.
PSTZSPZC	Prostate Segmentation, Transitional Zone Segmentation, and Peripheral Zone Calculation.
PZ	Peripheral Zone.
ReLU	Rectified Linear Unit.
ROC	Receiver Operating Characteristics.
SE	Squeeze-and-Excitation.
SM4TZPZS	Separate Models for the Transitional Zone and Peripheral Zone Segmentation.
SPZTZSM	Single Peripheral Zone and Transitional Zone Segmentation Model.

std	standard deviation.
T2W	T2-Weighted Imaging.
TZ	Transitional Zone.

TABLE OF CONTENTS

ABSTRACT 5
ACRONYMS AND ABBREVIATIONS 7
INTRODUCTION 19
Actuality 19
Research Problem 20
Research Object 22
Research Aim and Objectives 23
Scientific Novelty 24
Statements to be Defended 25
1. RELATED STUDIES ON THE PROSTATE ZONE AND PROSTATE CANCER SEGMENTATION APPROACHES 28
1.1. Global Incidence of Prostate Cancer 28
1.2. Related Studies on the Prostate Cancer Segmentation Task 30
1.2.1. Prostate Cancer Segmentation Challenges 30
1.2.2. Deep Learning Approaches for the Prostate Cancer Segmentation Task Employing Prostate Zonal Information 34
1.2.3. Deep Learning Approaches for the Prostate Cancer Segmentation Task Employing Attention Mechanisms and Prostate Zonal Information 39
1.3. Related Studies on the Prostate Zone Segmentation Task 46
1.3.1. Prostate Zone Segmentation Challenges 46
1.3.2. Deep Learning Approaches for the Prostate Zone Segmentation Task 47
1.4. Summary of the Chapter 50
2. METHODOLOGY 53
2.1. No-new U-Net Method 53
2.2. Compared Prostate Cancer Segmentation Ensembles and Baseline Model 58

2.3. Compared Prostate Zone Segmentation Workflows . . . 63
2.4. Compared Deep Neural Network Architectures for the Prostate Zone Segmentation Task . . . 65
2.4.1. Grid Attention U-Net . . . 66
2.4.2. Single Attention U-Net . . . 68
2.4.3. Multidimensional Attention U-Net . . . 68
2.4.4. Dual-Attention Attention U-Net . . . 70
2.5. Performance Measures . . . 71
2.6. Summary of the Chapter . . . 76
3. DATASET AND EXPERIMENT SETUP . . . 78
3.1. Dataset of the Prostate Cancerous Region Segmentation Task . . . 78
3.2. Experiment Setup of the Prostate Cancer Segmentation Task . . . 83
3.3. Dataset of Prostate Zone Segmentation . . . 86
3.4. Experiment Setup of the Prostate Zone Segmentation Task 89
3.5. Summary of the Chapter . . . 91
4. RESULTS . . . 93
4.1. Results of Experiment on the Prostate Zone Segmentation Task . . . 94
4.1.1. Results on the ProstateX Test Set . . . 94
4.1.2. Results on the Decathlon Test Set . . . 97
4.2. Results of Experiment on the Prostate Cancer Segmentation Task . . . 107
4.3. Summary of the Chapter . . . 118
GENERAL CONCLUSIONS . . . 120
BIBLIOGRAPHY . . . 122
LIST OF AUTHOR PUBLICATIONS . . . 132
CURRICULUM VITAE . . . 135

SANTRAUKA 136
ĮVADAS 138
Tyrimo problema 139
Tyrimo objektas 141
Tyrimo tikslas ir uždaviniai 141
Mokslinis naujumas ir praktinė vertė 142
Ginamieji teiginiai 143
S.1. Literatūros analizė 145
S.2. Metodų taikymas 146
S.2.1. Lyginti prostatos vėžio segmentavimo ansambliai 146
S.2.2. Lyginti prostatos zonų segmentavimo sprendimai 146
S.2.3. Segmentavimo tikslumo vertinimo matai 147
S.3. Duomenų ir eksperimento eiga 148
S.4. Rezultatai 149
S.4.1. Prostatos zonų segmentavimo rezultatai 149
S.4.2. Prostatos vėžio segmentavimo rezultatai 151
BENDROSIOS IŠVADOS 155

LIST OF TABLES

1.1 Overview of the reviewed prostate cancer segmentation methods. 45

3.1 Statistical summary of the echo time and repetition time parameters in the acquired T2-Weighted Imaging and Diffusion-Weighted Imaging modalities of the NVI dataset (in ms). 79

3.2 Gleason score distribution. 80

3.3 Distribution of lesion counts and lesion volume statistics across prostate zones, with volumes reported in mm^3. . . 81

3.4 Processed dataset size statistics. 83

3.5 Statistical summary of the echo time and repetition time parameters in the acquired T2W modalities of the ProstateX dataset in ms. 88

4.1 Friedman test results. These tests determine the presence of significant difference between compared deep neural network architectures. The test results are calculated on the ProstateX test set. 94

4.2 Friedman test results. These tests determine the presence of significant difference between compared workflows. The test results are calculated on the ProstateX test set. . 95

4.3 Mean, standard deviation, median and MAD of transitional zone segmentation results calculated on the ProstateX test set. 97

4.4 Mean, standard deviation, median and MAD of peripheral zone segmentation results calculated on the ProstateX test set. 98

4.5 Friedman test results. These tests determine the presence of significant difference between the compared deep neural network architectures. The test results are calculated on the Decathlon test set. 98

4.6 Conover post hoc test for unreplicated block design results. These tests indicate which deep neural network architectures are responsible for the observed statistical difference in the transitional zone segmentaion results. The segmentations are performed on the Decathlon test set. 100

4.7 Conover post hoc test for unreplicated block design results. These tests indicate which deep neural network architectures are responsible for the observed statistical difference in the peripheral zone segmentaion results. The segmentations are performed on the Decathlon test set. . 101

4.8 Friedman test results. These tests determine the presence of significant difference between the compared workflows. The test results are calculated on the Decathlon test set. . 103

4.9 Post hoc Wilcoxon signed-rank test results. These tests indicate which workflows are responsible for the observed statistical difference in the peripheral zone segmentation results. 104

4.10 Mean, standard deviation, median and MAD of transitional zone segmentation results calculated on the Decathlon test set. 105

4.11 Mean, standard deviation, median and MAD of peripheral zone segmentation results calculated on the Decathlon test set. 106

4.12 Mean, standard deviation, median and MAD of the Decathlon test set peripheral zone segmentation results obtained with Separate Models for the Transitional Zone and Peripheral Zone Segmentation. 107

4.13 Mean and std of prostate cancer segmentation results. These values are calculated based on the collected metrics DSC_τ and DSC. 109

4.14 Median and MAD of prostate cancer segmentation results. These values are calculated based on the collected metrics DSC_τ and DSC. 109

4.15 Conover post hoc test results on the prostate cancer segmentation task. The tests were conducted on the collected metrics DSC_τ and DSC. The metric corresponding to each reported p-value is indicated in parentheses. 111

4.16 False positive patients calculated on the results of the experiment on the prostate cancer segmentation task. . . . 112

4.17 Results of Wilcoxon signed-rank test performed on the prostate cancer segmentation results. These tests are conducted to determine the statistical significance of difference between the DSC metric and the novel metric, DSC_τ. 112

4.18 Distribution of predicted lesion counts across patients. Each distribution represents the number of patients for whom a given ensemble or the baseline model predicted the corresponding lesion count. 113

4.19 Distribution of missed lesions across the patients (42 cases with at least one lesion). Each distribution shows the number and percentage of patients for whom the ensemble or the baseline model failed to detect the corresponding number of lesions. 114

4.20 Distribution of spurious foci across patients. Each distribution shows the number of patients for whom a given ensemble or the baseline model produced the corresponding number of spurious lesions. 117

4.21 Distribution of spurious foci across all 119 patients. Each distribution shows the number and percentage of patients for whom a given ensemble or the baseline model produced the corresponding number of spurious lesions. 117

S.1 Friedmano testų rezultatų p-reikšmės. Šie testai vertina, ar tarp lyginamų DNN architektūrų yra statistiškai reikšmingų skirtumų. 149

S.2 Friedmano testų rezultatų p-reikšmės. Šie testai naudojami įvertinti, ar tarp lyginamų prostatos zonų segmentavimo veiksmų sekų yra statistiškai reikšmingų skirtumų. 150

S.3 Wilcoxono suderintų rangų post hoc testų rezultatai. Šie testai rodo, kurios prostatos zonų segmentavimo veiksmų sekos lemia statistiškai reikšmingus skirtumus segmentuojant periferinę zoną. Rezultatai apskaičiuoti naudojant Decathlono testavimo rinkinio rezultatus. 151

S.4 Aritmetinis vidurkis, standartinis nuokrypis, mediana ir MAD reikšmės apskaičiuotos iš periferinės zonos segmentavimo rezultatų, gautų naudojant Decathlono testavimo rinkinį. 152

S.5 Prostatos vėžio segmentavimo rezultatų vidurkis ir standartinis nuokrypis (DSC_τ ir DSC). 152

S.6 Conoverio post hoc testų rezultatai prostatos vėžio segmentavimo užduotyje. Kiekvienai pateiktai p-reikšmei naudotas matas nurodytas skliaustuose. 153

LIST OF FIGURES

1 Example of a single slice of each multi-parametric magnetic resonance imaging modality. The modality is specified in each caption of each image. A slice is a single planar cross-sectional image extracted from a volumetric image, representing a specific spatial plane within the volume. 23

2 Example of the cancerous lesion segmentation mask. The red volume is a cancerous lesion segmentation mask while transparent green region is a prostate mask. 23

3 Research workflow diagram. 27

1.1 Example of the transitional zone (yellow contour) and peripheral zone (green contour) within a single slice. The anatomical regions are overlaid on a T2-Weighted Imaging (T2W) image. 32

1.2 Schematics of the Squeeze-and-Excitation residual block. 43

2.1 The workflow the of the no-new U-Net method. 54

2.2 U-Net architecture with hyperparameters used in the no-new U-Net method. 57

2.3 The upsample and downsample block used in the no new U-Net where n_k is the number of kernels within convolution layers. 57

2.4 Ensemble 1 workflow diagram. 60

2.5 Ensemble 2 workflow diagram. 61

2.6 Ensemble 3 workflow diagram. 62

2.7 Ensemble 4 workflow diagram. 63

2.8 The prostate zone segmentation workflow: Separate Models for the Transitional Zone and Peripheral Zone Segmentation. 64

2.9 The prostate zone segmentation workflow: Prostate Segmentation, Transitional Zone Segmentation, and Peripheral Zone Calculation. 65
2.10 Diagram of the grid attention mechanism 66
2.11 Diagram of the grid attention U-Net architecture 68
2.12 Diagram of the single attention U-Net architecture 69
2.13 Diagram of the multidimensional attention mechanism . 70
2.14 Diagram of the dual-attention U-Net 71
2.15 Venn diagrams illustrating the calculation of FN_τ and FP_τ. 74

3.1 Flow diagram illustrating the patient selection and exclusion process. 80
3.2 Experiment schema for the prostate cancer segmentation task. 85
3.3 (**a**) A single slice example of T2W modality. The yellow contour marks the transitional zone while the green one is the peripheral zone. (**b**) A volumetric mask of transitional (yellow) and peripheral (green) zones. 86
3.4 Experimental schema for the prostate zone segmentation task. 90

4.1 Single slice examples of prostate zone segmentations on the ProstateX test set. The segmentations are performed with Separate Models for the Transitional Zone and Peripheral Zone Segmentation. The deep neural network architecture used for these segmentations is U-Net. . . . 99
4.2 Single slice examples of prostate zone segmentations on the Decathlon dataset. The segmentations are performed with Separate Models for the Transitional Zone and Peripheral Zone Segmentation. The deep neural network architecture used for these segmentations is U-Net. . . . 108

4.3 Single slice segmentation examples from a single patient for each ensemble and the baseline model, showing regions where predicted lesions overlap with the ground truth. Segmentations are overlaid on the T2W image. . . 115

4.4 Single slice examples of patients in whom none of the ensembles or the baseline model successfully localized the lesions. The ground truth is overlaid on the T2W image. 116

4.5 Single slice examples of spurious foci generated by each ensemble. The predicted spurious lesions are overlaid on the T2W image. 118

Actuality

Prostate Cancer (PCa) represents one of the most prevalent malignancies worldwide. The most recent comprehensive assessment of the global cancer burden, reported by Bray et al. [11], analyses data from 2022 encompassing 36 cancer types across 185 countries. According to this study, PCa is the second most frequently diagnosed cancer among men globally, accounting for 14.2% of all newly reported male cancer cases in 2022. It is exceeded only by lung cancer, which constitutes 15.3% of incident cases. In addition to its high incidence, PCa contributes substantially to cancer related mortality. In 2022, it was responsible for 7.3% of all cancer deaths among men, ranking fifth among the leading causes of male cancer mortality worldwide. Given its high prevalence and significant mortality, precise, fast and large scale delineation of malignant lesions are critical.

Recently, multi-parametric Magnetic Resonance Imaging (mpMRI) has emerged as one of the principal imaging modalities for the detection and diagnostic assessment of prostate cancer. Among the structured reporting systems developed to support mpMRI evaluation, one of the most widely adopted is the Prostate Imaging Reporting and Data System (PI-RADS). This framework was originally introduced by Barentsz et al. [6] with the aim of standardizing the interpretation and reporting of mpMRI findings within the clinical diagnostic pathway of PCa, thereby improving consistency, reproducibility, and communication across radiologists and institutions. Since its introduction, PI-RADS has evolved into a globally adopted reporting system, providing a structured and reproducible approach to the interpretation and communication of prostate mpMRI findings. The most recent version, PI-RADS v2.1, was presented by Turkbey et al. [75], refining prior recommendations and clarifying sequence specific assessment criteria. Importantly, PI-RADS v2.1 specifies zone dependent assessment criteria [75], reflecting the differential diagnostic performance of imaging sequences across anatomical regions of the prostate. Evaluation is primarily conducted in the Transitional Zone (TZ) and Peripheral Zone (PZ), as these regions account for the majority of prostate cancer cases.

Despite the widespread adoption of mpMRI and structured report-

ing systems, accurate localization and segmentation of PCa lesions remain technically demanding, even for experienced radiologists. Segmentation is a task that involves partitioning an image into meaningful regions by assigning a label to each pixel. The goal of this task is to identify and delineate objects or structures within an image. Several studies have highlighted the particular difficulty of identifying malignant regions within the transitional zone. As reported by Oto et al. [57], Rosenkrantz et al. [64], and Yu et al. [85], tumours arising in the transitional zone are frequently challenging to differentiate from benign prostatic hyperplasia nodules, which may exhibit overlapping imaging characteristics. Rosenkrantz et al. [64] further emphasize that the intrinsic diffuse background heterogeneity and multinodular architecture of the transitional zone substantially complicate lesion delineation in this region. Similarly, Yu et al. [85] report that stromal asymmetry within the transitional zone can mimic malignancy, while chronic prostatitis may generate imaging patterns in the peripheral zone that resemble cancerous lesions, thereby increasing the risk of false positive interpretation.

Consequently, there is a clear need for novel, robust, and scalable methodologies capable of accurately segmenting and localizing cancerous regions within the prostate in order to enhance diagnostic performance. The increasing availability of curated historical imaging datasets further supports the development of data driven approaches, particularly those based on Deep Learning (DL). DL techniques have been extensively applied to a wide range of medical image segmentation tasks, demonstrating state-of-the-art performance across multiple imaging modalities. In the context of prostate mpMRI, Deep Neural Network (DNN) architectures are capable of learning complex, hierarchical feature representations directly from volumetric data. Such models can effectively capture subtle spatial and intensity-based patterns within mpMRI scans, enabling precise delineation and localization of prostate cancer lesions.

Research Problem

The mpMRI integrates multiple imaging sequences, each providing complementary anatomical and functional information for the prostate cancer evaluation in the specific prostate zone. In PI-RADS v2.1, the prin-

cipal sequences include T2W, Apparent Diffusion Coefficient (ADC), Diffusion-Weighted Imaging (DWI) and Dynamic Contrast-Enhanced (DCE) imaging modalities. DWI and ADC are comprehensively described by Baliyan et al. [5]. DWI is directly acquired using an Magnetic Resonance Imaging (MRI) scanner, whereas ADC maps are computationally derived from multiple DWI acquisitions obtained at different diffusion weightings. Both of these sequences constitute the principal modalities for evaluating cancerous lesions in the PZ. The T2W sequence, detailed by Chavhan et al. [12], provides high resolution anatomical visualization and is similarly acquired directly from the scanner. This modality serves as the dominant sequence for lesion assessment in the TZ. DCE imaging, described by Yankeelov and Gore [83], involves repeated volumetric acquisitions following intravenous contrast administration to characterize enhancement kinetics over time. However, according to Bass et al. [7], and Turkbey et al. [75], the contribution of DCE to PCa evaluation, particularly within the PZ, is considered secondary. Its incremental diagnostic value is limited when compared to DWI and ADC modalities, and it primarily serves a supportive role rather than acting as a dominant determinant for lesion assessment. All of these sequences are volumetric and typically exhibit anisotropic spatial resolution.

Previous studies have investigated prostate cancer segmentation that employ prostate zonal information, including works such as Duran et al. [25], Hosseinzadeh et al. [33], and Saha et al. [66]. These approaches have demonstrated that leveraging zonal anatomy can improve cancer segmentation performance. However, existing research has not explicitly aligned DNN models with the PI-RADS framework, which provides a clinically established guideline for mpMRI interpretation. In particular, there is a notable gap in the literature regarding DL methods that integrate zone specific mpMRI sequences according to PI-RADS recommendations. Addressing this gap, the present dissertation systematically explores the prostate cancer segmentation incorporating prostate zonal information, with an emphasis on aligning model design and sequence selection with the PI-RADS framework to enhance both clinical relevance and segmentation accuracy.

Moreover, prostate cancer segmentation approaches that leverage prostate zonal information inherently require accurate delineation of

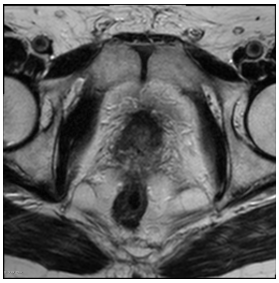
the prostate zones themselves. Considerable research has been devoted to this foundational task, including studies such as Aldoj et al. [2], Antonelli et al. [3], and Comelli et al. [19], which have proposed various DL-based approaches for automated prostate zone segmentation. In this dissertation, this line of research is further extended and the results are directly integrated into the primary focus of study, prostate cancer segmentation. The integration enables zone informed cancer segmentation, aligning the predictions of deep learning approaches with the anatomical and clinical context provided by the PI-RADS framework, and aims to improve the precision of the segmentation outcomes.

Finally, the experimental findings reported by Chen et al. [16] demonstrate that prostate cancer segmentation remains an intrinsically challenging task. Notably, substantial variability has been observed even among experienced radiologists, underscoring the subjective complexity of lesion delineation in mpMRI. This pronounced inter-observer variability directly affects the reliability of reference annotations, thereby complicating both the training and the objective evaluation of automated segmentation methods. In response to this limitation, the present dissertation investigates alternative and more robust evaluation strategies for PCa segmentation.

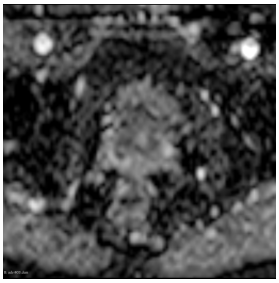
In summary, this dissertation develops a unified workflow for automated prostate cancer segmentation. The framework incorporates a dedicated prostate zone segmentation component, whose outputs are subsequently utilized within the PCa segmentation stage to enhance the accuracy and anatomical consistency of cancerous lesion delineation. The central component of the workflow, prostate cancer segmentation, is evaluated using a strategy specifically tailored to the challenges of this task, including considerations related to lesion heterogeneity and inter-observer variability. This integrated design ensures methodological coherence while maintaining strong clinical relevance.

Research Object

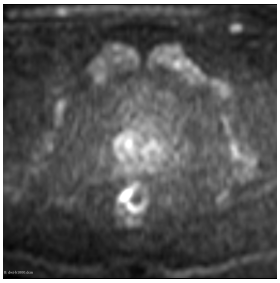
The research object is volumetric mpMRI data used for the segmentation of prostate cancer lesions. The example of a single slice of each modality is displayed in Figure 1. The example of the cancerous lesion segmentation is displayed in Figure 2.



(a) T2W



(b) ADC



(c) DWI

Figure 1. Example of a single slice of each multi-parametric magnetic resonance imaging modality. The modality is specified in each caption of each image. A slice is a single planar cross-sectional image extracted from a volumetric image, representing a specific spatial plane within the volume.

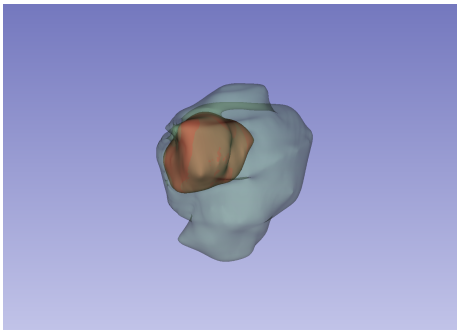


Figure 2. Example of the cancerous lesion segmentation mask. The red volume is a cancerous lesion segmentation mask while transparent green region is a prostate mask.

Research Aim and Objectives

The aim of this research is to develop and evaluate segmentation accuracy of deep neural network-based ensembles and workflows for the prostate anatomical zones and prostate cancer segmentation in the mpMRI data.

To achieve this aim, the following objectives are set:

1. Conduct a comprehensive literature review of contemporary deep learning-based approaches for prostate zonal and prostate cancer

segmentation, with particular attention to the diversity of deep neural network architectures, data preprocessing strategies, evaluation methodologies, and their reported applications in medical image segmentation.

2. Design deep learning–based ensembles that incorporate prostate zonal information for prostate cancer segmentation task, and empirically evaluate their segmentation performance to determine the most accurate ensemble.

3. Propose and investigate DL workflows for prostate zone segmentation, focusing on empirically comparing deep neural network architectures and workflow designs to identify the most effective approach for accurate zone delineation.

4. Propose an evaluation method for prostate cancer segmentation that accounts for inter-observer variability in reference annotations, and empirically assess its advantages by benchmarking it against standard evaluation approaches.

Scientific Novelty

This research introduces and systematically evaluates novel contributions that advance both the prostate cancer segmentation task and the prostate zonal segmentation task. The main contributions of this work are as follows:

1. No statistically significant performance differences were observed between attention-based U-Net variants (single attention U-Net, grid attention U-Net, multidimensional attention U-Net, and dual-attention U-Net) and the standard U-Net for prostate zonal segmentation.

2. Identification of the novel two-model, each dedicated to the specific zone, prostate zonal segmentation workflow as the most effective among all evaluated workflows as it yields the strongest generalization between datasets of different source.

3. Identification of the most effective prostate cancer segmentation ensemble, consisting of the baseline model and a model trained

on all mpMRI sequences for cancer segmentation within PZ as it achieves similar segmentation accuracy as the most accurate ensemble as well as significantly lower false positive patients than the aforementioned ensemble.

4. Introduction of a tolerance-based Dice Score Coefficient for prostate cancer segmentation that accounts for inter-observer variability in annotations, yielding higher scores than the conventional Dice metric by reducing penalties for minor boundary deviations within a defined tolerance.

Statements to be Defended

The following research-based claims are defended in this thesis:

1. Extending the U-Net architecture with the attention-based mechanisms does not improve the prostate zone segmentation accuracy.

2. Using two separate deep neural network models, each dedicated to segmenting a specific prostate zone, results in a segmentation accuracy improvement.

3. Prostate cancer segmentation ensembles composed of zone specific DNN models achieve higher segmentation accuracy than a single PCa segmentation model.

4. Incorporating a tolerance into the Dice Score Coefficient accounts for the high inter-observer variability in manual delineation of prostate cancer lesions.

The structure of the dissertation is presented as follows. The introduction presents the general situation of the prostate cancer segmentation, the relevance of the topic, the aim of the research, the research problem, the thesis statement and the workflow diagram of the whole study. Chapter 1 contains a literature review of the subject area, focusing on contemporary deep learning-based approaches for prostate zone and prostate cancer segmentation. Particular emphasis is placed on methods that integrate prostate zonal information into prostate cancer segmentation ensembles. Chapter 2 provides the methodology employed in this research, including the no-new U-Net approach, which

is used in the experiments as the basis for all trained models, the proposed prostate zone segmentation workflows, and DNN architectures incorporating attention mechanisms. It further details the integration of these architectures into the no-new U-Net, the proposed prostate cancer segmentation ensembles, the performance evaluation metrics used, and the introduction of the tolerance-based Dice Score Coefficient. Chapter 3 describes the research data, data preprocessing and the setup of the empirical experiment. Chapter 4 reports the results, and the general conclusions are provided in the final section.

The thesis consists of 158 pages, with the summary in Lithuanian starting from page 136, 27 tables, 29 figures, and 89 references.

The research process is divided into these main parts: 1. Introduction of the data sources and datasets, including a description of the imaging modalities and characteristics relevant to prostate zone and prostate cancer segmentation; 2. Data preprocessing during which dataset-specific transformations and harmonization steps are applied to ensure compatibility, standardization, and suitability for subsequent deep neural network training; 3. Model training where deep neural network architectures are developed and trained for use in the proposed prostate zone segmentation workflows and prostate cancer segmentation ensembles; 4. Evaluation involving the assessment of prostate zone segmentation workflows and prostate cancer segmentation ensembles using the Dice Score Coefficient, with the prostate cancer ensembles additionally evaluated using the proposed tolerant Dice Score Coefficient. The research workflow is visualized as the diagram in Figure 3 where notation are defined as follows:

- SM4TZPZS refers to the proposed prostate zone segmentation workflow, which employs two separate models each dedicated to segmenting a specific prostate zone.
- PSTZSPZC denotes the proposed prostate zone segmentation workflow that utilizes dedicated prostate and transitional zone segmentation models, with the peripheral zone obtained by subtracting the segmented transitional zone from the segmented prostate mask.
- TZ model refers to the transitional zone segmentation model, whereas the PZ model denotes the peripheral zone segmentation

model.

- Ensemble 1, Ensemble 2, Ensemble 3 and Ensemble 4 denote the individual proposed prostate cancer segmentation ensembles.
- DSC and DSC_τ denote the Dice Score Coefficient and the proposed metric, tolerant Dice Score Coefficient, respectively.

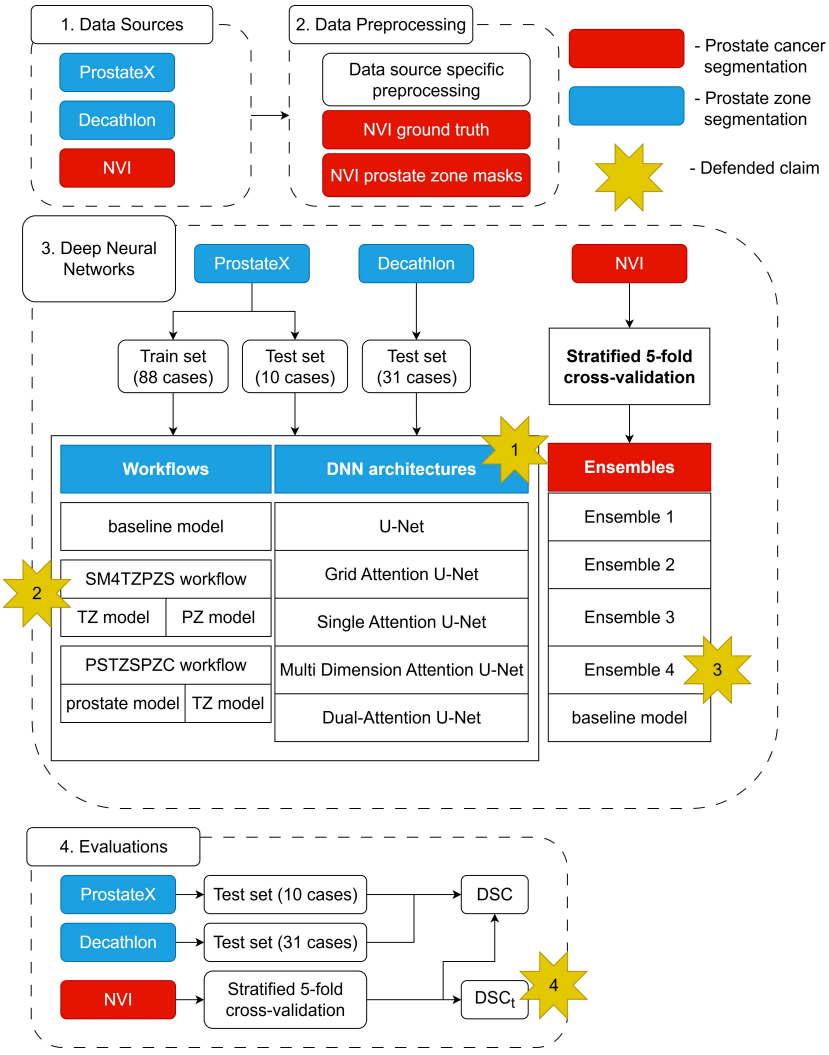


Figure 3. Research workflow diagram.

1. RELATED STUDIES ON THE PROSTATE ZONE AND PROSTATE CANCER SEGMENTATION APPROACHES

This chapter reviews scientific literature on PCa and prostate zone segmentation tasks. In particular it examines various DL approaches to these tasks. This chapter explores preprocessing steps employed for both tasks and various DNN architectures. These architectures include U-Net, dual-attention U-Net, attention U-Net and its variants. The theoretical foundation of these approaches together with the practical insights of empirical experiment results is provided in this chapter. Moreover, this section is a literature review on the empirical studies on the tasks themselves. It explores the limitations and challenges present in the prostate zone and prostate cancer segmentation task.

1.1. Global Incidence of Prostate Cancer

PCa is one of the most frequent cancers in the world. The most recent study on global cancer statistics is presented by Bray et al. [11]. This paper analyses cancer statistics derived from 2022 data covering 36 types of cancer in 185 countries. This article was published in 2024, indicating a two year delay in the reporting and dissemination of the presented data. According to this publication, prostate cancer is the second most common cancer in men globally. It accounted for 14.2% of all recorded cancer incidents in males in 2022 and it was only surpassed by lung cancer which accounted for 15.3% of all recorded cancer cases. Moreover, PCa accounts for a substantial proportion of cancer related deaths. It accounted for 7.3% of all cancer related deaths in men in 2022. Prostate cancer ranked fifth among the deadliest cancers affecting men worldwide. It was surpassed by lung, liver, colorectal and stomach cancers, which accounted for 22.7%, 9.6%, 9.2% and 7.9% respectively.

The same trend was observed in the previous study presented by Sung et al. [74]. This paper provides the statistics calculated on the data of the same cancer types. The data, on which these statistics are calculated, was gathered in the same 185 countries. The statistics presented in the study [74] show that PCa accounted for 14.1% of all recorded cancer incidents in men in 2020. It once again was surpassed only by lung cancer which accounted for 21.5% of all recorded cancer cases in men. The trend in cancer-related deaths seen in 2022 persisted from

2020. Prostate cancer accounted for 6.8% of cancer related deaths in men. Once again it was surpassed by lung, liver, colorectal and stomach cancers, which accounted for 21.5%, 10.5%, 9.3% and 9.1% respectively. The prevalence of this trend, reported in the publication [74], affirms the forecast made by Chhikara et al. [17].

Furthermore, Siegel et al. [70] reports the forecasted cancer trends in the US, 2023. The authors employ the data gathered from 1975 to 2019 in the US. They observed that the number of PCa cases increased by 3% annually from 2014 through 2019. Thus, the authors of [70] estimated that in 2023 there would be approximately 288,300 new cases of prostate cancer, accounting for 29% of all cancer diagnoses in men making it the most frequently diagnosed cancer in males in the US in 2023. The authors also projected that PCa would be one of the leading causes of cancer related mortality in men in the US in 2023, with an estimated 34,700 deaths, accounting for 11% of all cancer related deaths in men. This number is only surpassed by lung and bronchial cancer which is 67,160 (21%).

The publication [54] provides statistical data on cancer incidence and mortality in Lithuania. The latest data show that in 2023, 1,942 new cases of prostate cancer were registered, accounting for 22% of all cancer diagnoses in men. Moreover, in 2023, 543 deaths related to prostate cancer were recorded, accounting for 13% of all cancer related deaths in men.

However, it has to be noted that prostate cancer incident rate might be inflated due to unregulated and opportunistic Prostate Specific Antigen (PSA) testing. Vaccarella et al. [77] report the PCa trends in 26 European countries during 1980–2017. Moreover, they report PSA testing trends for 12 countries, which represent a subset of aforementioned 26 European countries. The authors observed a correlation between the direction and rate of change in prostate cancer incidence relative to the prostate specific antigen testing. Furthermore, they also note that across all 26 countries, excluding Croatia, Latvia and several countries in eastern Europe, there is a visible divergence between incidence and mortality rates. The reported incidence rate grew during the observed period, while the mortality rate remained stable. As a result, the standardised rate ratio between incidence and mortality has almost doubled. These observations strongly suggest that PCa is overdiagnosed

and a more reliable diagnosis method is required.

Overall, the evidence strongly suggests that prostate cancer is one of the most common cancers in men. Therefore, accurate diagnostic approach at large scale of PCa is essential. Accordingly, the primary objective of this thesis is to investigate novel computer aided approaches for prostate cancer segmentation. Specifically, this work focuses on the development of an automated workflow that incorporates prostate zonal information to improve PCa segmentation. In this context, potential advancements in prostate zone segmentation are also explored alongside prostate cancer segmentation. This section presents an overview of the scientific literature on DL-based methods for prostate zone segmentation, as well as approaches that leverage zonal information for prostate cancer segmentation. Collectively, these studies support the rationale for investigating a PCa segmentation workflow that integrates prostate zonal information.

1.2. Related Studies on the Prostate Cancer Segmentation Task

This section explores various studies conducted on the PCa segmentation task. A substantial amount of research has already been conducted on machine learning methods for this task. Consequently, several review articles summarise this body of work, including the review published by Cuocolo et al. [21] in 2019. Since then, research on prostate cancer segmentation has increasingly focused on DL approaches. In particular, many studies emphasise deep learning-based methods that incorporate prostate zone information. Lastly, the majority of these studies rely on the U-Net DNN architecture and its variants. This architecture, introduced by Ronneberger et al. [63], along with its variants, remains the state-of-the-art deep neural network architecture for segmentation to this day. As a result, the capabilities of this architecture and its variants have been widely explored in medical segmentation tasks.

1.2.1. Prostate Cancer Segmentation Challenges

One of the greatest challenges present in the automation of the PCa segmentation task is a high inter-observer variability. The research provided by Chen et al. [16] indicates that even among experienced ra-

diologists, the manual prostate cancer segmentation contains a lot of variance. In their study, the authors report the results of the experiment in which they tasked four radiologists, 4 urologists, 4 trainees and 4 non-clinician segmentation scientists to segment PCa. The participants segmented two slices of two different cases. One case contained a cancerous lesion in the anterior zone and another case in the PZ. The mpMRI modalities consisted of T2W, ADC and DWI. Moreover, participants were required to segment kidney tumours within a single computed tomography scan of a single kidney. The authors of article [16] report root mean square difference, Dice Score Coefficient (DSC), Jaccard index, Matthews correlation coefficient, specificity and sensitivity. The root-mean-square difference is calculated using a point-by-point comparison method. Each radiologist's segmentation was treated independently as a separate ground truth. As a result, four scores were computed for each metric and segmentation result for the participant who is not a radiologist. Meanwhile, each radiologist's results were compared with those of the other radiologists.

The authors of the study [16] mainly focuses to the DSC metric. Overall, the variability among radiologists in prostate cancer segmentation was very high, with a mean DSC of 0.81 for segmentations of the lesion in the peripheral zone and 0.58 for segmentations of the lesion in the anterior zone. The variance is especially high relative to the results of the kidney tumour segmentations, as the mean DSC among radiologists in the kidney segmentation task was 0.96. Among all participants, the PCa segmentation results of urologists were the most similar to the segmentations of the radiologists. They achieved a mean DSC of 0.73 for the segmentations of the lesion located in the PZ and 0.59 for the segmentations of the lesion located in the anterior zone. The trainees achieved similar results (mean DSC of 0.7 for the segmentations of the lesion located in the PZ and 0.54 for the segmentations of the lesion located in the anterior zone). The segmentations of the non-clinicians were the least similar (mean DSC of 0.6 for the segmentations of the lesion located in the PZ and 0.47 for the segmentations of the lesion located in the anterior zone). Meanwhile, kidney tumour segmentations of all participants were very similar to the segmentation results of the radiologists (0.96, 0.95 and 0.95 mean DSC calculated for the segmentations of urologists, trainees and non-clinicians, respectively). In

conclusion, due to the nature of the prostate cancer segmentation task, there is a high inter-observer variability present in the datasets of PCa segmentations.

Furthermore, extensive research on manual PCa segmentation of mpMRI scans indicates that lesion segmentation is significantly more difficult in the transitional zone than in the peripheral zone. A single slice example illustrating peripheral and transitional zones is displayed in Figure 1.1. The research presented in the article by Oto et al. [57] indicates that the diffuse background heterogeneity and multi-modularity of the TZ complicate the prostate cancer segmentation. This research involved three radiologists independently segmenting PCa lesions within the TZ for 106 PCa patients. The radiologists segmented each case multiple times. First, they used only the T2W images. Next, they added ADC and DWI. After that, they included an additional DWI sequence acquired with a different parameter b. The parameter b characterizes the strength and timing of the gradients employed to generate DWI. It determines the sensitivity of the image to the diffusion of water molecules, with higher values resulting in increased diffusion weighting and greater sensitivity to diffusion effects. Finally, they added the DCE modality to the segmentation process. The results were compared with the prostatectomy findings as a reference standard.

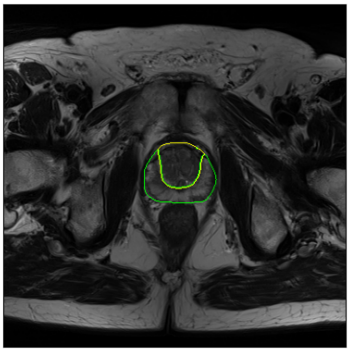


Figure 1.1. Example of the transitional zone (yellow contour) and peripheral zone (green contour) within a single slice. The anatomical regions are overlaid on a T2W image.

The authors of the research reported in [57] found that segmentation accuracy improved significantly after incorporating DWI ($b =$

$1000s/mm^2$) and ADC modalities. The sensitivity increased by 20.8%, 30.6% and 30.6% for each radiologist, respectively. Meanwhile, the specificity changed marginally by 0.3%, 0.8% and -3.7% for each radiologist, respectively. Furthermore, segmentation accuracy improved further after adding DWI modality images with a higher b parameter ($b = 2000s/mm^2$). The sensitivity increased by 19.4%, 8.3% and 6.5% for each radiologist, respectively. Meanwhile, the specificity changed marginally by 0.6%, -0.2% and -1.2% for each radiologist, respectively. Lastly, the results indicate that DCE does not significantly improve manual PCa segmentation. The sensitivity changed by 0%, -8.1% and 4.8% for each radiologist, respectively, while the specificity changed by -0.7%, 0.4% and -0.9% for each radiologist, respectively. Overall, the most performative results were achieved with T2W, ADC and DWI images with $b = 1000s/mm^2$ and $b = 2000s/mm^2$. However, even though the authors emphasise that DWI images assist in segmenting prostate cancer by reflecting changes in cellularity and providing complementary information for segmentation, the PCa segmentation in the transitional zone remains challenging.

Furthermore, the research conducted by Rosenkrantz et al. [64] indicates that stromal asymmetry may mimic PCa within the TZ, thereby making prostate cancer segmentation in this zone particularly challenging. In contrast, although chronic prostatitis typically occurs in the PZ and may also resemble cancerous tissue, the resulting ambiguity is generally less pronounced than in the TZ. In this research, the authors provide a detailed explanation of these phenomena, along with case studies as examples.

Lastly, the research presented by Yu et al. [85] provides a literature review of PCa lesions detected in each prostate zone. The authors reviewed the research over the past two decades, summarised the results, and identified several key issues arising in prostate cancer segmentation. The results indicate that PCa incidents are more frequent in the peripheral zone than the transitional zone which further complicates the collection of a representative dataset of prostate cancer segmentations in the TZ relative to PCa segmentations in the PZ. Therefore, in this thesis, we address this issue by proposing a novel metric which accounts for a high inter-observer variance.

1.2.2. Deep Learning Approaches for the Prostate Cancer Segmentation Task Employing Prostate Zonal Information

Extensive research has already explored the use of prostate zonal information and deep neural networks for segmenting PCa. One of such studies is presented by Hosseinzadeh et al. [32], and Hosseinzadeh et al. [33]. This research demonstrates the advantages of employing probabilistic prostate zone segmentation within a 2D computer-aided detection system. Authors in their study [32] report the results of the experiment in which they trained three DNN models of U-Net architecture. The first model is a simple 2D U-Net which segments lesions. This model is used as the baseline. The second model is the same 2D U-Net, with one additional input channel for probabilistic prostate zonal segmentation. The authors refer to this model as an early fusion model. The last model is once again 2D U-Net model. The probabilistic zonal segmentation is incorporated into this model by combining the feature map with it before the model's last convolutional layer. The authors of the studies [32, 33] refer to this model as a late fusion model. Additionally, the authors report results from the early fusion model and the late fusion model, both incorporating deterministic prostate zonal segmentations rather than probabilistic ones.

The dataset used in the study [32] consists of 848 cases. Each case consists of mpMRI, which includes T2W, ADC and DWI modalities. The prostate cancer in this dataset is manually segmented by expert radiologists by employing PI-RADS v2. In total, 319 patients have a PCa region. In this study, all images were resampled and cropped around the centre. The prostate zones are obtained using a 3D U-Net model trained on 53 T2W images. This model is taken directly from the experiment reported by Mooij et al. [52]. Lastly, the dataset was split into training, validation, and test sets by randomly allocating cases in a 3:1:1 ratio via stratified sampling. The authors report the average test-set sensitivity for each deep neural network model.

The results reported by Hosseinzadeh et al. [32] indicate that the highest mean sensitivity is achieved by an early fusion model with incorporated probabilistic prostate zonal segmentations. This model achieved a mean sensitivity of 0.795 at a 0.5 False Positive (FP) rate, 0.887 at 1 and 2 FP rates, averaging to 0.856. Meanwhile, the late fusion model with incorporated probabilistic prostate zonal segmenta-

tion achieved mean sensitivities of 0.774, 0.873, and 0.873 at 0.5, 1, and 2 FP rates, respectively, averaging to 0.84. The baseline model achieved the least accurate results, indicating the positive impact of prostate zone segmentations on PCa segmentation task. This model achieved mean sensitivity of 0.76 at FP rate of 0.5, 0.825 mean sensitivity at 1 and 2 FP rates with average of 0.803. Moreover, both early and late fusion models performed better with probabilistic prostate zone segmentations rather than deterministic ones. The early fusion model achieved higher mean sensitivity by 2.1%, 8.5%, and 8.5% at FP ratios of 0.5, 1, and 2, respectively, for an overall improvement in sensitivity of 6.3%. Meanwhile, the late fusion model achieved higher sensitivity by 5.7% at 1 and 2 FP ratios, while no improvement was observed at a FP ratio of 0.5, resulting in an overall improvement of 3.8% in sensitivity. Overall, the early fusion model incorporating probabilistic prostatic zonal information is the most accurate prostate cancer segmentation approach among the compared approaches.

The authors further explore the early fusion model and report their findings in their next article [33]. In this follow-up experiment, they use a much larger dataset which contains patients from two centres. The Centre 1 (Radboudumc, Nijmegen, The Netherlands) provided a cohort of 2,438 cases, while the Centre 2 (Ziekenhuis Groep Twente, Almelo, The Netherlands) supplied a dataset consisting of 296 cases. This dataset is annotated by two investigators under the supervision of a radiologist with seven years of experience in mpMRI and prostate imaging. The annotators used the size and location information provided in the clinical reports from centres. These reports were compiled by employing the PI-RADS v2 framework. Additionally, Centre 1 provided biopsy results in these reports. The authors considered manually segmented region in cases of dataset provided by Centre 1 as the ground truth if it scored 4 or above in PI-RADS framework. In total, 898 patients had at least one ground truth region in the dataset provided by Centre 1, leaving 1,540 cases with no ground truth regions. Meanwhile, manually segmented regions in cases from the dataset provided by Centre 2 were considered ground truth if at least one biopsy within the case satisfied the condition: $GS \geq 3 + 4$ where GS is a Gleason Score (GS). The Gleason score is a histopathological grading system used to assess the aggressiveness of prostate cancer. This score

is denoted as the sum of the primary and secondary histological pattern grades observed in the biopsy, each reflecting the aggressiveness of the cancer. Out of 296 cases of the dataset provided by Centre 2, 86 patients have at least a single ground truth region, while 210 patients do not. The prostate zones are acquired by employing the DNN model developed in the research presented by Riepe et al. [61]. The authors of the study [33] compare an early-fusion model incorporating probabilistic prostate zone segmentations with a deep neural network model that does not employ prostate zonal information. Firstly, they split the dataset provided by Centre 1 into training, validation and test sets through random stratified sampling. The test set size is 20% of the whole dataset, the validation set size is 15% of the whole dataset, and the remaining cases are used in the training set. In the paper [33], the authors report Receiver Operating Characteristics (ROC) and the Area Under the ROC curve (AUROC) to compare patient-based accuracy. Furthermore, the study includes Free-Response ROC (FROC) with sensitivities at 0.5, 1 and 2 FP ratios for lesion-based analysis. Lastly, the authors used bootstrapping with 2,000 bootstrap samples to determine statistical significance (confidence interval is 95% and the significance level is 0.05).

All collected metrics indicated that incorporating prostate zonal information into a deep learning approach to solve the PCa segmentation task improves the solution's accuracy. All statistical tests indicate that the difference is statistically significant. The incorporation of prostate zone segmentation into a DNN model resulted in AUROC increase by 2.8%. Furthermore, adding prostate zone segmentation to the model increased the sensitivity on the dataset supplied by Centre 1 by 6.3%, 4%, and 3.4% at FP ratios of 0.5, 1, and 2, respectively. In conclusion, the research conducted by Hosseinzadeh et al. [32], and Hosseinzadeh et al. [33] shows that incorporating prostate zonal information into DL approaches improves the accuracy of prostate cancer segmentation.

Another study on PCa segmentation using DL and prostate zone information is presented by Zheng et al. [88]. In this research, a new DNN architecture, AtPCa-Net, is proposed. In this architecture, the input is mpMRI modalities, a binary mask of the transitional zone and a binary mask of the peripheral zone. The input is duplicated, with one copy mirrored along the vertical axis. The architecture itself is an ex-

tended U-Net. It consists of two encoders and a single decoder. One encoder encodes the original input, while the other encodes the mirrored input. The architecture's output contains two channels. The results of these channels are used together with the prostate zone segmentations in the proposed novel loss function. This function is a modified cross-entropy loss function that is anatomically aware, adjusting the result based on the prostate zone where the lesion is detected. This design of the loss function suppresses prediction vector patterns that should not exist. Additionally, focal loss is incorporated into the proposed loss function to account for the relatively small volumes of PCa lesions compared to the volume of the scans themselves. The hyperparameters of U-Net used in the no-new U-Net (nnU-Net) method, proposed by Isensee et al. [35], are adopted for the hyperparameters of the AtPCa-net.

The authors empirically compare the novel DNN architecture, AtPCa-Net, with current state-of-the-art methods such as the nnU-Net proposed by Isensee et al. [35], SEResU-net proposed by Hu et al. [34], ResidualU-Net proposed by Bhalerao and Thakur [8], VNet proposed by Chen et al. [13], Attention U-Net proposed by Oktay et al. [56], VoxResNet proposed by Chen et al. [14] and UNETR proposed by Hatamizadeh et al. [30]. The dataset used in the experiment was provided by the authors' affiliation. It consists of 652 patients, of whom 220 have a prostate cancer lesion in the ground truth, while 432 do not. The authors of the article [88] used five-fold cross-validation to validate and evaluate the performance of the compared deep learning approaches. The cases with at least one lesion in the ground truth are distributed equally across folds.

The authors of the study [88] test the accuracy improvement achieved by using zonal loss instead of loss used in the nnU-Net method and using the model of their proposed DNN architecture, AtPCa-Net, instead of the nnU-Net method. They report that using zonal loss instead of the loss function used in the nnU-Net method improves PCa segmentation sensitivity at FROC measurements of 0.5, 1, 1.5, 2 and 2.5 FP predictions per patient by 0.8%, 0%, 0.8%, 2.1% and 0.4% respectively. Meanwhile, the AtPCa-net model achieves higher sensitivity than the nnU-Net method at the same measurements by 1.6%, 0%, 2%, 1.5% and 0% respectively. Lastly, using AtPCa-net with zonal loss

outperforms the model trained with the nnU-Net method at the same measurements by 4.1%, 2.1%, 3.6%, 3.3% and 2% respectively.

Moreover, the proposed deep neural network architecture, AtPCa-Net, achieved more accurate results than all other compared DNN architectures. AtPCa-Net achieved higher PCa lesion segmentation sensitivity at 0.5 FP predictions per patient by 8%, 8.1%, 7.7%, 12.2%, 8.1% and 7.7% than ResidualU-net, VNet, SEResU-net, VoxResNet, Attention U-Net and UNETR, respectively. Meanwhile, at 1 FP predictions per patient sensitivity was higher by 7.8%, 6.1%, 5.7%, 9.4%, 4.5% and 4.9% than the same compared DNN architectures. Further, at 1.5 FP predictions per patient sensitivity was higher by 4.5%, 6.5%, 5.2%, 9.7%, 3.6% and 5.7% than the same compared DNN architectures. At 2 FP predictions per patient sensitivity was higher by 2.1%, 6.5%, 4.1%, 9.4%, 3.3% and 5.3% than the same compared deep neural network architectures. Lastly, at 2.5 FP predictions per patient sensitivity was higher by 1.2%, 5.7%, 2.4%, 6.5%, 2.8% and 4% than the same compared DNN architectures. It is also worth mentioning that the model trained with the nnU-Net method can be considered the second best DL approach for prostate cancer segmentation task in this experiment. This model achieved worse sensitivity than ResidualU-net by 1.2% at 2 FP predictions per patient and by 0.8% at 2.5 FP predictions per patient.

Another research on prostate zonal information incorporation to DL approach for PCa segmentation is provided by Mehta et al. [48] This research proposes a general framework for applying a deep learning approach for prostate cancer segmentation. This framework consists of prostate zone segmentation, prostate cancer segmentation and report generation modules.

Prostate zone segmentation and PCa segmentation modules are ensembles, which were constructed by employing the nnU-Net method. These ensembles are made by performing k-fold cross-validation. During the inference, the segmentations of the resulting k models are combined into a single probability map. In the experiment described in the study [48] the number of folds for each ensemble is ten. Each ensemble consists of DNN models trained with the nnU-Net method with task-specific hyperparameter modifications. The input to these models is 2D tensors with different numbers of channels. The prostate zone segmentation module ensemble is referred to by the authors as Zone-U-Net-E,

and the models in this ensemble are referred to as Zone-Nets. This ensemble process segments prostate zones from T2W image slices. The segmentation output is probability maps of PZ and TZ. Meanwhile, the prostate cancer segmentation module is the ensemble, CSPCa-U-Net-E, and the models used by this ensemble are called CSPCa-U-Nets. The segmentation output of this ensemble is a probability map of PCa lesions. The last module is automatic result reporting, which is out of scope of this thesis.

The authors of the article [48] train the models of their framework on the publicly available ProstateX challenge of 2017 dataset, which is described by Litjens et al. [44]. The researchers used 204 cases of this dataset. Namely, they employed T2W, ADC, DWI, prostate zone mask and PCa mask of each case. Furthermore, the authors evaluate the resulting framework on another publicly available PICTURE dataset, previously reported by Simmons et al. [71], and Simmons et al. [72]. In total of 80 patients within this dataset contain prostate zones. Thus, the prostate zone segmentation model was evaluated on the dataset of size 80. Meanwhile, 247 cases were used to evaluate the prostate cancer segmentation ensemble.

The Zone-U-Net-E achieved mean DSC of 0.75 and 0.8, median DSC of 0.75 and 0.81 for peripheral zone and transitional zone segmentation respectively on test set. Furthermore, CSPCa-Net-E achieved mean DSC of 0.46 and median DSC of 0.58 for PCa segmentation on test set.

Overall, the incorporation of prostate zonal information into prostate a cancer segmentation task is a highly researched field. Thus, in this thesis we focus on expanding in this direction by proposing novel ensembles which employ prostate zonal information in the PCa segmentation task.

1.2.3. Deep Learning Approaches for the Prostate Cancer Segmentation Task Employing Attention Mechanisms and Prostate Zonal Information

Since the introduction of the transformer architecture by Vaswani et al. [78], attention mechanisms, originally proposed by Bahdanau [4], have been extensively explored across a wide range of DL tasks, including applications in medical computer vision. In particular, attention-based

models have demonstrated significant potential in improving performance on complex image analysis problems. Accordingly, this section presents a literature review of PCa segmentation methods that incorporate both prostate zonal information and various attention mechanisms.

One of the studies on prostate cancer segmentation, employing prostate zonal information and U-Net architecture with incorporated attention mechanism, is presented by Duran et al. [25], and Duran et al. [26]. The explored U-Net variation in this research employs attention mechanism introduced by Oktay et al. [56], and Schlemper et al. [69]. This variation consists of one encoder and two consecutive decoders. The encoder and first decoder together comprise a standard U-Net architecture. However, the second decoder employs attention mechanism in every upsample block. The gating signal of each attention mechanism is the output of the first decoder while the input signal comes from the last layer of the corresponding upsample block. The first decoder segments the location within the prostate in which the second decoder segments the PCa regions. The authors of the papers [25, 26] refer to this architecture as ProstAttention-Net.

The authors of the study presented in [25] compare their proposed deep neural network architecture with U-Net architecture. They train and evaluate the models of compared architectures on the dataset which is collected and annotated by partner clinical center of authors affiliation. This dataset consists of 98 patients and two mpMRI modalities – T2W and ADC. The training and evaluation is performed by employing five-fold cross-validation. Each fold is balanced regarding lesion classes and the type of MRI scanner. The authors feed these modalities channel wise to all compared architectures. The annotations include PZ masks and PCa lesion masks. The PCa lesion masks are multiclass where each class represents different GS grade. These grades consist of $GS \geq 9$, $GS = 8$, $GS = 4 + 3$, $GS = 3 + 4$ and $GS = 3 + 3$, totalling in five classes. Thus, the authors of the paper [25] train ProstAttention-Net and U-Net models to segment PZ and PCa lesions of five classes in the PZ.

The authors of the paper [25] provide prostate cancer detection sensitivity at 1 and 1.5 FP per patient and Cohen's quadratic weighted kappa coefficients for each compared DNN architecture. ProstAttention-

Net achieved higher PCa detection sensitivity for each GS at each FP per patient than U-Net except for $GS = 8$. Moreover, ProstAttention-Net achieved higher Cohen's kappa coefficient by 4% than U-Net. Furthermore, in order to compare their results to the results published by de Vente et al. [23], the authors also calculate Cohen's kappa coefficient for PCa segmentation only by excluding peripheral zone segmentation results. The authors conclude that ProstAttention-Net is a more preferred solution than the most accurate model compared by de Vente et al. [23], SLOR.

In their next study, published in the paper [26], the authors compare ProstAttention-Net architecture with Attention U-Net [56, 69], U-Net [63], ENet [58] and DeepLabv3+ with XceptionNet as backbone [15] architectures. In this paper, the authors experiment with a much larger dataset which consists of 219 cases. The training and evaluation is performed by employing five-fold cross-validation with each fold being balanced regarding lesion classes and the type of MRI scanner. Once again this dataset consists of T2W and ADC modalities. Meanwhile, the annotations contain PCa segmentations and prostate region segmentations instead of the PZ. Additionally, instead of five GS classes four used by merging classes $GS \geq 9$ and $GS = 8$ into a single class $GS \geq 8$. Thus, the authors train the compared models to segment prostate region and prostate cancer lesions of four classes in the whole prostate gland.

Duran et al. [26] provide FROC values, namely, mean PCa detection sensitivities at 1 and 1.5 FP per patient. Furthermore, the authors provide Cohen's quadratic weighted kappa coefficients. Both metrics indicate that ProstAttention-Net achieves the most accurate segmentation results. ProstAttention-Net was only outperformed by ENet at prostate cancer of class $GS = 3 + 3$ detection sensitivities where ENet achieved higher sensitivity by 4% and 6% at 1 and 1.5 FP per patient. Moreover, they conduct one-sided Wilcoxon signed-rank statistical tests on collected metrics. The test results on the collected FROC values showed that there is no significant difference between the results achieved by ProstAttention-Net and the results achieved by Attention U-Net and ENet. Furthermore, statistical test conducted on collected Cohen's kappa coefficients did not indicate a statistically significant difference between the compared DNN architectures. Meanwhile, these results indicate

that there is a statistically significant difference between the FROC values achieved by ProstAttention-Net and the FROC values achieved by DeepLabv3+ and U-Net. ProstAttention-Net achieved higher Cohen's kappa coefficient by 0.4%, 7.3%, 9.5% and 10% than ENet, Attention U-Net, U-Net and DeepLabv3+.

Lastly, Duran et al. [26] evaluates ProstAttention-Net on external dataset PROSTATEx-2, which is publicly available and is described by Litjens et al. [42]. The authors use the best performing ProstAttention-Net model acquired by conducting five-fold cross-validation. They compare its achieved Cohen's kappa coefficient with the Cohen's kappa coefficient achieved by the solution provided by De Vente et al. [24]. The authors of the paper [26] observe that their proposed DNN architecture achieves less accurate results. However, they also acknowledge that a more reliable comparison is required as De Vente et al. [24] solution is trained on a different dataset than the ProstAttention-Net model.

Another study on prostate cancer segmentation task, employing attention mechanisms and prostate zonal information, is published by Saha et al. [66]. In this paper a novel end-to-end framework is proposed for PCa detection. This framework uses T2W, ADC and DWI mpMRI modalities as an input. The output of the framework is PCa segmentation. The framework itself consists of two parallel DNN models – M_1 and M_2, followed by a decision fusion node.

The deep neural network model, M_2, is a model, which takes patches of mpMRI scan and assigns score in the interval $[0, 1]$. This score represents the probability of the presence of the PCa lesion within the patch. These scores are later used to reduce the FP of the output of the M_1 model in decision fusion node. The authors divide the input of the framework, mpMRI scan, into eight uniformaly divided patches. The authors of the study [66] use Convolutional Neural Network (CNN) architectures based on residual learning as these DNN architectures are state-of-the-art classification models in medical domain as supported by Jiang et al. [37], McKinney et al. [46], and Yoo et al. [84].

The other model, M_1 is the principal component of the framework. This model employs prostate zonal information as a separate channel in addition to the mpMRI modalities. The authors support this decision by providing studies on various medical segmentation tasks where the use of clinical priori in CNN architectures achieved a higher accuracy.

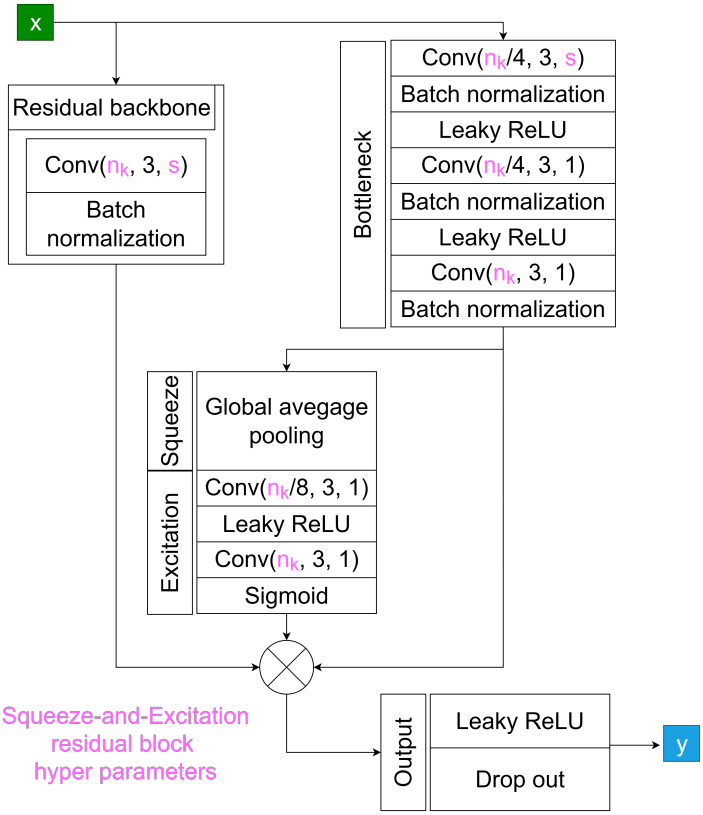


Figure 1.2. Schematics of the Squeeze-and-Excitation residual block.

These studies include Dalca et al. [22], Gibson et al. [28], and Wachinger et al. [80]. The prostate zonal information is encoded into a probabilistic population prior as introduced by the authors in their previous work [65].

The DNN architecture of M_1 model is dual-attention U-Net. Instead of being a variation of U-Net itself, this architecture is a variation of the UNet++ architecture, which was introduced by Zhou et al. [89]. Dual-attention U-Net uses the grid attention mechanism proposed by Oktay et al. [56], and Schlemper et al. [69]. In addition, dual-attention U-Net utilizes the Squeeze-and-Excitation (SE) residual blocks. This block is visualized in the schematics of Figure 1.2.

The SE residual block consists of the following components:

- Bottleneck represented by block *Bottleneck* in Figure 1.2.

- Pre-activation residual backbone proposed by He et al. [31]. This component is represented by $Residualbackbone$ block in Figure 1.2.
- Channel-wise attention mechanism, SE, proposed by Hu et al. [34]. This component of the SE residual block is represented by blocks $Squeeze$ and $Excitation$.
- Output component which is represented by block $Output$ in Figure 1.2.

The authors of the paper [66] compare their proposed framework with Attention U-Net [56], nnU-Net [35], SEResU-net [34] and UNet++ [89] architectures on two different test sets. In their experiment they employ a dataset provided by the Radboud University Medical Centre which consists of 2436 cases. The authors split this dataset to train, validation and test sets each consisting of 65%, 15% and 20% of the whole dataset respectively via double-stratified sampling preserving the same class balance. They report the results acquired only on the test set and refer to it as TS1. The other dataset is provided by Ziekenhuisgroep Twente which consists of 296 patients. The authors refer to this dataset as TS2. Both datasets contain T2W, ADC and DWI mpMRI modalities together with manual prostate cancer segmentation masks. Lastly, the prostate zone masks were acquired by employing anisotropic 3D U-Net model from a separate study presented by Riepe et al. [61].

The authors of the study [66] report the highest achieved sensitivity and corresponding FP per patient by each compared approach. They report the results achieved on each test set separately. Based on the reported results, the authors conclude that the proposed framework achieved the highest results among all of the compared approaches on both test sets. Furthermore, the stand-alone M_1 model of dual-attention U-Net architecture achieved more accurate results than the compared stand-alone models of the other DNN architectures. Overall, all stand-alone models employing attention mechanism, dual-attention U-Net and attention U-Net, achieved more accurate results than the stand-alone models which do not employ attention U-Net, U-SEResNet, nnU-Net, UNet++.

Overall, the literature evidence strongly suggests that the use of attention mechanism in the medical image segmentation tasks can improve the segmentation accuracy of the solutions. In this thesis, we do

not further investigate the impact of using attention mechanisms for prostate cancer segmentation, as extensive prior work has already explored attention-based deep learning approaches for this task. Instead we inspect the effect of the attention mechanisms on prostate zone segmentation task.

Table 1.1 provides an overview of the reviewed prostate cancer segmentation methods. The Method column lists the segmentation methods, Employed Zones indicates whether the method targets the peripheral zone, transitional zone, or both, With Attention Mechanism specifies whether an attention mechanism is used, and Dataset Sources identifies the datasets on which each solution was evaluated. This table shows that nearly all reviewed methods consider both prostate zones. In addition, attention mechanisms are used in roughly the same number of methods as those that do not use them. Finally, a systematic comparison between methods is not feasible because each approach is evaluated on different datasets.

Table 1.1. Overview of the reviewed prostate cancer segmentation methods.

Method	Employed Zones	With attention mechanism	Dataset Sources
Early / late fusion [32, 33]	Both	No	Private
AtPCa-Net [88]	Both	No	Private
Autoprostate [48]	Both	No	ProstateX [44], PICTURE [71, 72]
ProstAttention-Net [25, 26]	PZ only	Yes	Private, Prostatex-2 [42]
End-to-end PCa segmentation framework [66]	Both	Yes	Private

1.3. Related Studies on the Prostate Zone Segmentation Task

This section inspects various research conducted on the prostate zones segmentation task. Namely, the segmentation of the PZ, TZ and prostate region itself.

1.3.1. Prostate Zone Segmentation Challenges

The research by Montagne et al. [51] highlights a significant challenge posed by the prostate zone segmentation task. This research investigates the data variability in the T2W modality, specifically in the TZ and PZ regions. The research reported in the study [51] consists of seven radiologists performing manual segmentation of the prostate zones. These zones included the whole prostate region and the transitional zone. Meanwhile, the peripheral zone was calculated by subtracting TZ from the prostate region segmentations. The segmentation is performed on the compiled cohort of 40 patients. The dataset used in this experiment was provided by Health Data Center of the Assistance Publique-Hopitaux de Paris.

The authors report DSC, Hausdorff distance and average Hausdorff distance. They calculate these metrics by employing pairwise comparison and comparison with consensus, which is calculated by employing STAPLE algorithm. The authors of the study, Montagne et al. [51], focus on the DSC when interpreting the results. They also use the paired Wilcoxon signed-rank test and the Mann-Whitney U test for related and independent samples, respectively, to determine whether a statistically significant difference is present. All statistical tests are two-sided. The research found out that prostate zone segmentation inter-observer variability is significant in:

- Extreme parts of the prostate gland. The authors of the paper [51] report that pairwise comparison yields a mean DSC of 0.90 for the apex segmentation of the prostate gland and a mean DSC of 0.87 for the base. Meanwhile, the consensus comparison shows mean DSC of 0.90 for the apex segmentation of the prostate gland and mean DSC of 0.91 for the base. Considering that the base is located in the PZ, the inter-observer variability affects the segmentation of this zone more than the TZ.

- Relatively smaller prostate. The authors use the Spearman correlation coefficient ρ to assess the association between prostate volume and DSC. The p-values are corrected by employing the Holm-Bonferroni method. The result of this test is $\rho > 0.8$ and p-value < 0.001.

Therefore, this research underscores the importance of incorporating representative clinical samples that capture morphological variation into the prostate zone segmentation dataset. In this thesis, we address this issue by employing two publicly available datasets in the experiment on the prostate zone segmentation task. Both of these datasets are described in Section 3.3.

1.3.2. Deep Learning Approaches for the Prostate Zone Segmentation Task

A substantial amount of research has already been conducted on the prostate zone segmentation task. As a result, several review articles cover this research, such as the review published by Gillespie et al. [29]. In this subsection, we inspect a more recent development on the prostate zone segmentation task.

Just as the research on prostate cancer segmentation, most of the research on the prostate zone segmentation task investigates U-Net DNN architecture and its variations.

These studies include the research reported by Comelli et al. [19]. The authors of this paper propose to employ Efficient Neural Network (ENET) or Efficient Residual Factorized convolution Network (ERFNet) DNN architectures for the prostate region segmentation task. These DNN architectures are proposed by Paszke et al. [58], and Romera et al. [62], respectively. These architectures are designed to be less complex than U-Net, as they do not employ skip connections. As a result, the training of ENET and ERFNet is faster relative to the training of U-Net. The authors of the article [19] compare the performance of these architectures with the performance of U-Net by conducting the empirical experiment. The comparison includes both accuracy and segmentation speed. The dataset used for this experiment was provided by the Radiology Department at the author's affiliation and is not publicly disclosed. This dataset consists of 85 cases. The cases in this dataset

consist of mpMRI modality, T2W volumetric image, and a manually segmented prostate region.

The results of the experiment, reported in the paper [19], show that ENET and ERFNet achieves slightly lower segmentation accuracy than U-Net. ENET achieves 0.75% lower mean DSC than U-Net. Meanwhile, ERFNet achieves 2.94% lower mean DSC than U-Net. Moreover, the results of this experiment show that the segmentation of ENET on the CPU is approximately 6.8 times faster than that of U-Net. Meanwhile, the segmentation of ERFNet on the CPU is ~ 4.89 times quicker than the segmentation of U-Net. However, the difference is much lower when the segmentation is accelerated with the GPU. ERFNet is faster ~ 1.5 times than U-Net, while ENET is faster ~ 1.47 times. Overall, the authors of the study [19] conclude that, among the compared deep neural network architectures, ENET is the most preferred architecture for prostate gland segmentation when GPUs are not available. Otherwise, U-Net remains the most suitable DNN architecture for the prostate region segmentation.

Another research is conducted by Aldoj et al. [2]. In this research, the authors propose the novel DNN architecture, Dense-2 U-Net for prostate zone segmentation. This architecture is a variant of U-Net that employs dense blocks rather than a conventional stack of convolutional layers. The authors compare the accuracy of the Dense-2 U-Net with that of U-Net through an empirical experiment. The empirical experiment reported by Aldoj et al. [2] is conducted on the publicly available ProstateX dataset. The cases in this dataset include a single mpMRI modality, a T2W volumetric image, and transitional and peripheral zone masks. The prostate mask in the research is calculated by combining both prostate zones.

The results of the empirical experiments reported by the authors of the article [2] indicate that the proposed architecture achieves higher DSC than the classical U-Net. In the prostate region segmentation task, the experimental results report a 1.6% higher median DSC achieved by the Dense-2 U-Net than by U-Net. Furthermore, the results of PZ segmentation task experiment show that the Dense-2 U-Net achieves higher median DSC than U-Net by 2% percent. Finally, the results reported by the authors show that the Dense-2 U-Net achieves a 2.2% higher median DSC than U-Net on the TZ segmentation task.

Moreover, the article by Antonelli et al. [3] contains the results of the Decathlon challenge. This challenge focuses on various medical image segmentation tasks. One of them is prostate zone segmentation. The organisers of the challenge provide a publicly available dataset. This dataset consists of 32 cases, each comprising mpMRI modalities, T2W and ADC volumetric images, and TZ and PZ masks. The most accurate results of this task are achieved by the method proposed by Isensee et al. [35]. This method achieves 3% higher mean DSC than the second-place method in transitional zone segmentation and 5% higher mean DSC in peripheral zone segmentation.

Isensee et al. [35] propose a method for various medical image segmentation tasks – nnU-Net. Since then, this method has been widely employed in experiments on various medical segmentation tasks such as the once presented by Akinci D'Antonoli et al. [1], Jucevičius et al. [40], Ma et al. [45], Wasserthal et al. [81], and Zhang et al. [87].

Furthermore, the prostate gland segmentation performance of the nnU-Net method was compared with U-Net and its variants by Bhandary et al. [9]. The U-Net variants consist of Attention U-Net, V-Net proposed by Milletari et al. [50], SegResNet proposed by Myronenko [53], and U-Net++. The evaluation was conducted using four datasets: an internal Computer Tomography (CT) dataset, an internal MRI dataset, the MSD-Prostate dataset published by Simpson et al. [73], and the PROMISE12 dataset published by Litjens et al. [43]. The internal CT dataset was collected from three medical centres: the Medical Center – University of Freiburg, the Hannover Medical School, and the Affiliated Drum Tower Hospital of Nanjing University Medical School. Data collected from the Affiliated Drum Tower Hospital of Nanjing University Medical School were designated as the independent test set, whereas data obtained from all the remaining participating centres were used for model training. Manual prostate delineations for all patients were performed by two experts from the Medical Center – University of Freiburg. The internal MRI dataset was also acquired at the Medical Center – University of Freiburg, where prostate delineation was performed by experienced physicians. For all experiments involving MRI data, only the T2W modality was utilised. The internal CT, internal MRI, MSD-Prostate, and PROMISE12 datasets contained 184, 209, 32, and 50 cases, respectively. The internal CT training cohort

comprised 135 cases, whereas an independent test cohort consisted of 49 cases. Furthermore, the performance of the DNN models on the remaining datasets was evaluated using a five-fold cross-validation strategy.

The segmentation performance of each model was assessed using the DSC. Across all evaluated datasets, the nnU-Net achieved the highest DSC values. Compared with the second-best performing model, Attention U-Net, the nnU-Net improved the DSC by approximately 0.4%, 0.2%, 0.3%, and 1.0% on the internal CT, internal MRI, MSD-Prostate, and PROMISE12 datasets, respectively. However, statistical analysis indicated that these improvements were not statistically significant for most datasets, with the exception of the PROMISE12 dataset, where the performance difference was reported as statistically significant.

Given its prevalence, we use this method to further develop prostate zone segmentation and PCa segmentation solutions. This method consists of data preprocessing, DNN model with the U-Net architecture, adapted for volumetric imaging, post-processing and training procedure. Each component of the method is described in detail in Chapter 2.

In conclusion, extensive research has already been conducted on prostate zone segmentation using U-Net architecture and its variants. We further expand this direction in this thesis by proposing to employ other U-Net variations and proposing novel workflows for the prostate segmentation task.

1.4. Summary of the Chapter

This chapter reviews literature on deep learning approaches for PCa and prostate zone segmentation tasks, with a particular emphasis on DL approaches, the role of prostate zonal information, and the use of attention mechanisms. Collectively, the reviewed studies demonstrate that both tasks remain challenging due to substantial inter-observer variability, anatomical complexity, and dataset heterogeneity, yet they also show consistent progress driven by advances in deep neural network architectures and learning strategies.

In the context of PCa segmentation, the literature consistently reports high inter-observer variability particularly for lesions located in the transitional zone. The dominant DNN architecture in the prostate

cancer segmentation research is U-Net and its variants. Furthermore, mpMRI, especially the combination of T2W, ADC, and DWI, has been shown to be essential for improving lesion delineation, yet intrinsic ambiguity in lesion boundaries persists. Numerous studies demonstrate that incorporating prostate zonal information as an additional input channel, via a probabilistic prior, or through anatomically informed loss functions significantly improves lesion detection and segmentation performance. These improvements are observed across different datasets, evaluation protocols, and learning paradigms, underscoring the clinical relevance of prostate zonal information for PCa segmentation.

Moreover, recent work integrating attention mechanisms with prostate zonal information suggests further gains in segmentation accuracy by enabling models to focus on anatomically and diagnostically relevant regions. While these approaches often outperform classical U-Net baselines, the literature also indicates diminishing returns and increasing model complexity, motivating careful consideration of trade off between accuracy and computational cost.

In the context of prostate zone segmentation, the reviewed literature highlights a high morphological variation in the apex and base of the prostate and in smaller prostate glands. Just as in prostate cancer segmentation, U-Net and its variants have emerged as the dominant and most effective DNN architectures for prostate zone segmentation. Incremental DNN architectural modifications, such as dense connections or task adaptive configurations exemplified by the nnU-Net, have been shown to yield measurable improvements in segmentation accuracy. Public challenges and benchmarks further confirm that carefully designed preprocessing, training, and post processing pipelines often contribute as much to performance gains as DNN architectural novelty. These findings motivate the use of multiple datasets and robust evaluation strategies to ensure generalizability, a direction adopted in this thesis.

Unfortunately, a direct validation and cross comparison of the results reported across the reviewed studies is not feasible, as each experiment was conducted using a different dataset. For instance, the study published by Isensee et al. [35] employed the Decathlon dataset, whereas the work presented by Mehta et al. [48] utilised the ProstateX

and PICTURE datasets. Similarly, the study by Duran et al. [26] was conducted using the ProstateX-2 dataset, while Aldoj et al. [2] evaluated their approach on the ProstateX dataset. Furthermore, the majority of these studies rely on datasets that are not publicly accessible, which limits reproducibility and hinders independent verification of their findings. These studies include [19, 25, 26, 32, 33, 66, 88].

In summary, existing research establishes U-Net and its variations as strong baselines for both prostate cancer and prostate zone segmentation. The evidence clearly supports the inclusion of prostate zonal information to improve PCa segmentation accuracy. Moreover, the consistently observed improvement in segmentation accuracy achieved through the incorporation of attention mechanisms in the PCa segmentation task suggests a promising research direction for prostate zone segmentation, namely the integration of attention-based U-Net variations. Building on these insights, this thesis extends prior work by exploring alternative attention-based U-Net variations and novel workflow designs for advancements in the prostate zone segmentation task as well as novel ensembles and evaluation metrics that explicitly account for inter-observer variability for advancements in prostate cancer segmentation task.

2. METHODOLOGY

This chapter describes proposed novel approaches for prostate cancer and the prostate zone segmentation task as well as employed evaluation methods in the thesis. The key innovations of this research are:

- Four novel ensembles for the PCa segmentation task.
- Novel workflows for prostate zone segmentation: Separate Models for the Transitional Zone and Peripheral Zone Segmentation (SM4TZPZS) and Prostate Segmentation, Transitional Zone Segmentation, and Peripheral Zone Calculation (PSTZSPZC). The result of the prostate zone segmentation is used to improve the solutions of the PCa segmentation task.
- A method to incorporate the attention mechanism into the nnU-Net method. This incorporation consists of several deep neural network architectures, which employ attention mechanism, into the nnU-Net: grid attention U-Net, single attention U-Net, multidimensional attention U-Net and dual-Attention U-Net. This modification of the nnU-Net method is inspected on the prostate zone segmentation task.
- A novel metric, tolerant Dice Score Coefficient (DSC_τ) to evaluate PCa segmentation results. This metric is a modified DSC. The modification introduces tolerance in order to address a high inter-observer variability among radiologists.

These methods were presented in the main papers by Vaitulevičius et al. [A.1], [A.2] and [B.4]. Additionally, this chapter describes the nnU-Net method in detail as this method is widely explored throughout our research.

2.1. No-new U-Net Method

As mentioned in Chapter 1, the no-new U-Net method consists of data preprocessing, deep neural network model with the U-Net architecture, adapted for volumetric imaging, post processing and training procedure. Each of these components are described in detail in this section. As an input the nnU-Net method takes volumetric images to segment and optional masks of regions, in which segmentation needs to

be performed. Each image within the input is provided to the model in smaller patches. During training the label masks are supplied as well. The output is a volumetric multiclass image. The number of classes is supplied during training. Furthermore, the nnU-Net method contains predefined parameters, heuristic rules and blueprint parameters. In addition, this method employs parameters derived from the training set, data fingerprints. Both predefined and derived parameters are used within the workflow of the nnU-Net method, which is visualized in Figure 2.1. The use of derived and predefined parameters is explained in this section.

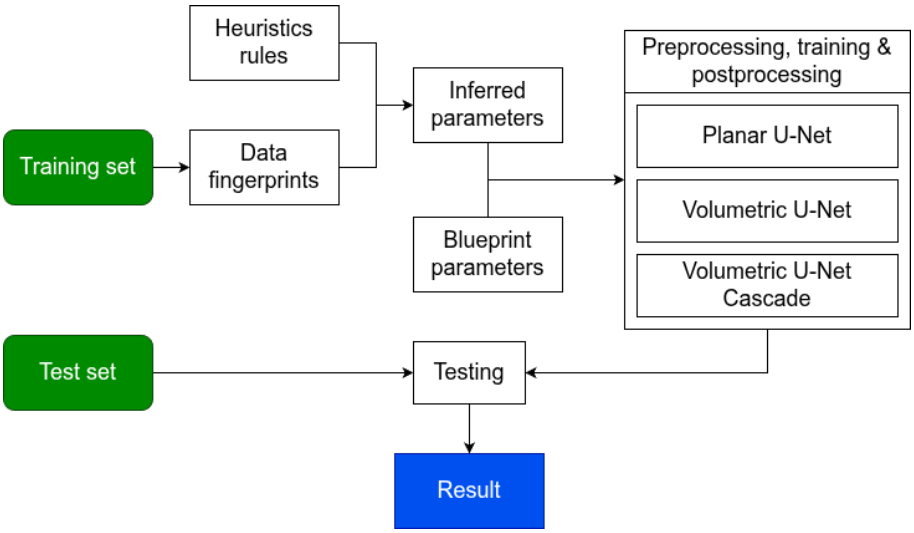


Figure 2.1. The workflow the of the no-new U-Net method.

The preprocessing consists of cropping, resampling and normalization steps. Firstly, the volumetric images are cropped to the non-zero values of the region masks, if the masks were provided. Secondly, the volumetric images are resampled. Almost any medical volumetric image contains voxel spacing information and most medical images are anisotropic. However, this information cannot simply be provided to the U-Net model. Thus, the anisotropic volumetric images are resampled to the median voxel spacing of the whole dataset, which is one of the dataset fingerprint parameters. The third order spline interpolation is used for images themselves, while the nearest neighbour interpolation is used for the label masks.

The novel label mask resampling algorithm is proposed by Jucevičius et al. [39]. In the follow up research Jucevičius et al. [40] investigate the impact of the novel label mask resampling algorithm on the prostate cancer segmentation with the nnU-Net method. In their experiment, the authors train several PCa segmentation models by employing the nnU-Net method with different the nnU-Net method configurations. Namely, they inspect the results by employing planar and volumetric images. They inspect the impact of the proposed resampling algorithm by training and evaluating the models with data preprocessed with the aforementioned algorithm and the raw data. They refer to the images preprocessed with the proposed resampling algorithm as isotropic, while the raw images are referred to as anisotropic. Isotropic input images performed better than anisotropic on both configurations, planar and volumetric images. On planar images isotropic images achieved 5.54% higher mean DSC than anisotropic ones and on volumetric images isotropic images achieved 18.07% higher mean DSC. Thus, in this thesis we employ this preprocessing step in all of our models trained with the nnU-Net method.

The last preprocessing step is normalization, which is a simple z-score normalization applied to each patient individually. The normalization step of the nnU-Net preprocessing differs for computer tomography scans. However, as it is out of scope of our research, we do not provide a detailed explanation for it.

The core component of the nnU-Net method is the training of the U-Net model. In their paper [35], the authors provide empirical evidence across several medical image segmentation tasks showing that planar images are segmented more accurately than volumetric images when the ratio between the maximum and minimum voxel spacing across all axes exceeds three. Hence, the nnU-Net method infers which images should be segmented by employing the aforementioned heuristic rule and the dataset fingerprint – voxel spacing. In our research we omit this rule and use volumetric images as label mask resampling, proposed by Jucevičius et al. [39], together with third order spline interpolation transform anisotropic images into isotropic ones. Thus, none of the volumetric images within our dataset have a ratio between the maximum and minimum voxel spacing across all axes that is greater than three.

Moreover, the no-new U-Net method additionally trains a volumetric U-Net Cascade if resampled images are four times larger than the patch size with a batch size of two. This heuristic rule addresses high computational resource requirement for the large images. The authors of the study [35] determined that this rule does not apply for prostate images as they are not large enough. Thus, in our research we exclude this heuristic rule as well and do not provide a detailed explanation of a volumetric U-Net Cascade configuration of the nnU-Net method.

Furthermore, the nnU-Net infers the batch size, S_B, and patch size, (x_p, y_p, z_p), from the training set. It employs the following heuristic rule: the nnU-Net adjusts the aspect ratio of the patch volume to the median number of voxels. If the median number of voxels of the images within the dataset is smaller than 128^3 voxels then the median shape is used as the patch size. Otherwise, the patch shape is $128 \times 128 \times 128$. The batch size is calculated with the following equation:

$$S_B = \frac{2 \times 128^3}{x_p \times y_p \times z_p} \tag{2.1}$$

The nnU-Net method contains the following blueprint parameters: the hyperparameters of the U-Net model, loss function, the training schedule and data augmentation. The hyperparameter tuning of the U-Net were performed in the research presented in the paper [35] and validated by performing five-fold cross-validation. The hyperparameters were selected manually based on prior empirical experience obtained from a range of medical image segmentation tasks. The resulting DNN architecture and training configuration were designed to be broadly applicable to medical image segmentation problems, rather than optimized for any single specific task. Moreover, the authors substituted the Rectified Linear Unit (ReLU) used in the original U-Net with Leaky ReLU whose negative slope is 0.01. The resulting topology of the trained U-Net model is displayed in Figure 2.2 where n_k represents number of convolution kernels, s - stride and kernel size. $Conv^T$ stands for transposed convolution. The $Upsample(n_k)$ and $Downsample(n_k)$ blocks are presented in Figure 2.3. The loss function is the sum of inverted DSC and cross-entropy loss. The training schedule uses stochastic gradient descent optimizer with an initial learning rate of 1×10^{-2} and Nesterov momentum ($\mu = 0.99$). The learning rate

is updated at the start of every epoch according to a polynomial decay schedule, defined in Equation 2.2. In this formulation, lr_0 denotes the initial learning rate, e represents the current epoch, E the total number of training epochs, and lr_e the learning rate computed for epoch e.

$$lr_e = lr_0 \times (1 - \frac{e}{E})^{exp} \tag{2.2}$$

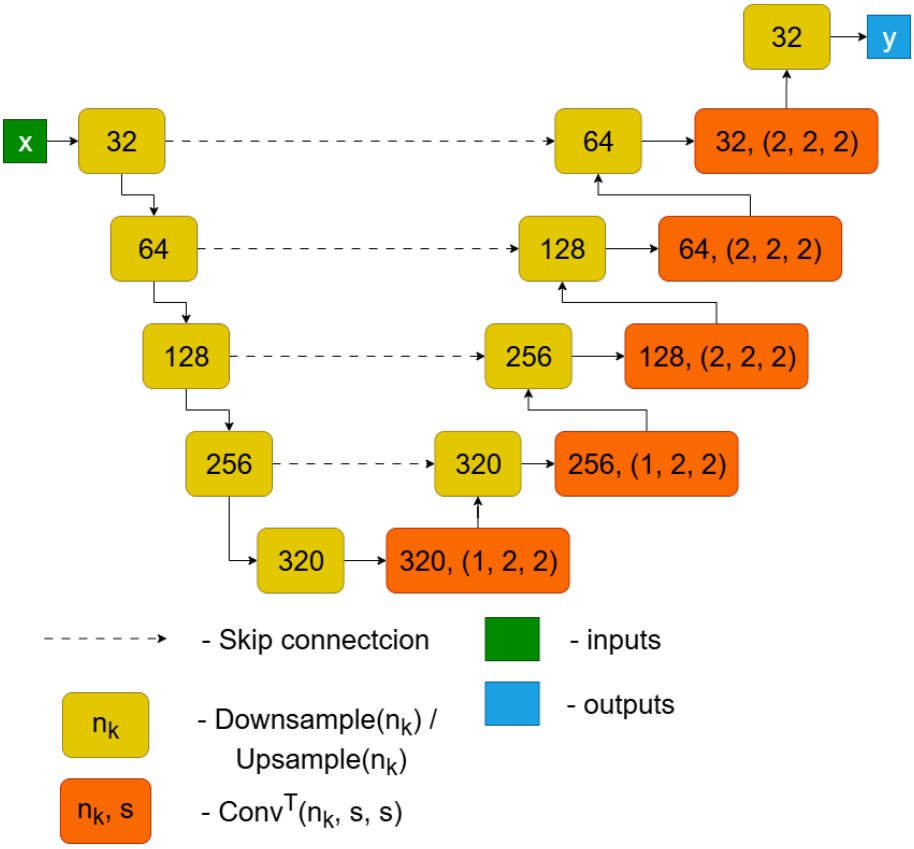


Figure 2.2. U-Net architecture with hyperparameters used in the no-new U-Net method.



Figure 2.3. The upsample and downsample block used in the no new U-Net where n_k is the number of kernels within convolution layers.

The training is stopped if an exponential moving average of the training losses did not improve by at least 5×10^{-3} within the last 60 epochs, but not before the learning rate becomes smaller than 10^{-6}. The maximum number of epochs is 1,000. Furthermore, training employs deep supervision. Deep supervision is a training strategy where loss function is applied to intermediate layers in addition to the final output layer. This encourages earlier network representations to be more discriminative and can improve gradient flow, convergence speed, and overall segmentation performance. The data augmentation consists of random rotations, random scaling, Gaussian noise, Gaussian blur, gamma correction augmentation and mirroring. Lastly, nnU-Net enforces that at least a third of the samples within each batch would contain at least one randomly chosen foreground class.

The post processing of nnU-Net is a single step. A connected component analysis is performed on the ground truth of the training set. All classes, that are contained within a single connected component in every patient, are considered to exhibit the same behaviour in the whole population. Thus, only the largest connected component is kept for these classes within the predictions.

Lastly, as mentioned before, two models are trained with different configurations, planar and volumetric U-Net models, as well as optional volumetric U-Net Cascade. The original nnU-Net method creates ensembles of pairs of these models. As in our research we only inspect volumetric U-Net configuration, these ensembles are not investigated. Instead, in this thesis we propose novel workflows for prostate zone segmentation and novel ensembles for prostate cancer segmentation.

2.2. Compared Prostate Cancer Segmentation Ensembles and Baseline Model

We propose four novel ensembles. All ensembles utilize deep neural network models trained with the no-new U-Net method. The data preprocessing and training of all models are performed independently from one another. These models are: $PCa_{seg}(T2W \times TZ)$, $PCa_{seg}(All \times TZ)$, $PCa_{seg}((ADC, DWI) \times PZ)$, $PCa_{seg}(All \times PZ)$ and $PCa_{seg}((All, Zones) \times prostate)$. The difference between them lies in the inputs and the region in which it segments cancerous region. To segment prostate cancer within a specific anatomical region, a prostate

zone mask is applied to the input of the deep neural network. Voxels that do not overlap with the mask are set to zero, thereby restricting the network's focus to the region of interest. This masking strategy is an integral component of the nnU-Net method and was therefore adopted in this work, rather than alternative approaches such as soft masking. The following DNN models are trained for use within the proposed ensembles:

- $PCa_{seg}(T2W \times TZ)$ model's input is only T2W modality. The transitional zone mask is applied on the input by setting all voxels outside the mask to zero. Thus, this model segments prostate cancer only in the TZ region.
- $PCa_{seg}((ADC, DWI) \times PZ)$ model's input are ADC and DWI modalities. The modalities are provided channel-wise as a model's input. This model segments PCa only in the peripheral zone.
- $PCa_{seg}(All \times TZ)$ model's input are all modalities. The voxels of the input of this model are set to zero if they do not overlap with the TZ mask.
- $PCa_{seg}(All \times PZ)$ model's input are all modalities as well. This model segments PCa only in the PZ region.
- $PCa_{seg}((All, Zones) \times prostate)$ model's input are all modalities as well as prostate zone masks as separate channel of the model's input. Therefore, the input tensor of this model has these channels: T2W, ADC, DWI and prostate zones. In our research the prostate zones masks contain values of 1, which represent PZ, and 2, which represent TZ region. The voxels of the input of $PCa_{seg}((All, Zones) \times prostate)$ model are set to zero if they do not overlap with the prostate gland mask. This mask is a disjunction of both prostate zone masks. Currently, $PCa_{seg}((All, Zones) \times prostate)$ model is used as a state-of-the-art method to employ prostate zone information in prostate cancer segmentation task. This model is used many experiments such as the ones presented in the papers [32, 48, 66]. Thus, $PCa_{seg}((All, Zones) \times prostate)$ model is the baseline model of our research.

The first proposed and compared ensemble is denoted as Ensemble 1. Different mpMRI modalities vary in their effectiveness at detecting prostate cancer within specific prostate zones. The main modality used for PCa segmentation in the transitional zone is T2W. Meanwhile, the principal modalities for PCa segmentation in peripheral zone are ADC and DWI. Thus, the result of Ensemble 1 is a disjunction of the results of the $PCa_{seg}(T2W \times TZ)$ and $PCa_{seg}((ADC, DWI) \times PZ)$ models. In other words $Ensemble1 = PCa_{seg}(T2W \times TZ) \vee PCa_{seg}((ADC, DWI) \times PZ)$. The workflow of this ensemble is visualized in Figure 2.4.

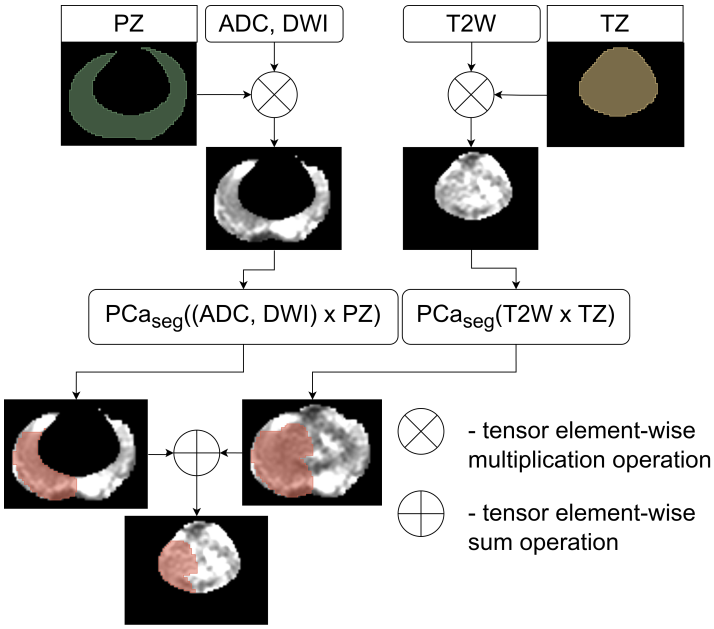


Figure 2.4. Ensemble 1 workflow diagram.

Although radiologists tend to prioritize mpMRI modalities that are most informative for a given prostate zone, all modalities are ultimately reviewed for each zone, thereby maximizing their confidence in the segmentation. Thus, the second proposed ensemble is the disjunction of the results of $PCa_{seg}(All \times TZ)$ model and $PCa_{seg}(All \times PZ)$ model. Formally, $Ensemble2 = PCa_{seg}(All \times TZ) \vee PCa_{seg}(All \times PZ)$. This ensemble is denoted as Ensemble 2 and its workflow is provided in Figure 2.5. However, Ensemble 2 is more computationally demanding than Ensemble 1 as the input of the models used by Ensemble 2 is larger than the ones used by Ensemble 2. Firstly, $PCa_{seg}(All \times TZ)$

is three times larger than the $PCa_{seg}(T2W \times TZ)$ model's. Secondly, $PCa_{seg}(All \times PZ)$ is 1/3 times larger than the $PCa_{seg}((ADC, DWI) \times PZ)$ model's. Thus, if both ensembles achieve similar results then Ensemble 1 is a preferred solution to the prostate cancer segmentation task.

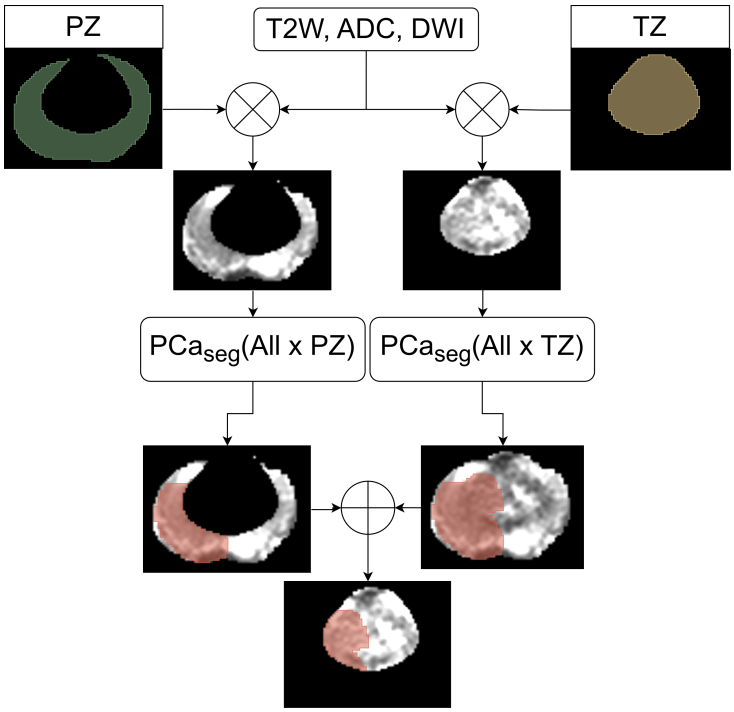


Figure 2.5. Ensemble 2 workflow diagram.

The third proposed ensemble is denoted as Ensemble 3. This ensemble consists of disjunction of model $PCa_{seg}(All \times TZ)$, model $PCa_{seg}(All \times PZ)$ and the baseline model, $PCa_{seg}((All, Zones) \times prostate)$. Thus, the equation of Ensemble 3 is $Ensemble\,3 = PCa_{seg}(All \times TZ) \vee PCa_{seg}(All \times PZ) \vee PCa_{seg}((All, Zones) \times prostate)$. The input of the models used by this ensemble contains more information than the models of Ensemble 2. Therefore, assuming that models $PCa_{seg}(All \times TZ)$, $PCa_{seg}(All \times PZ)$ and $PCa_{seg}((All, Zones) \times prostate)$ are complementary to each other, it is more likely that Ensemble 3 will achieve higher accuracy than Ensemble 2. Ensemble 3 workflow is visualized in Figure 2.6.

Literature analysis of Yu et al. [86] provides an overview of the epi-

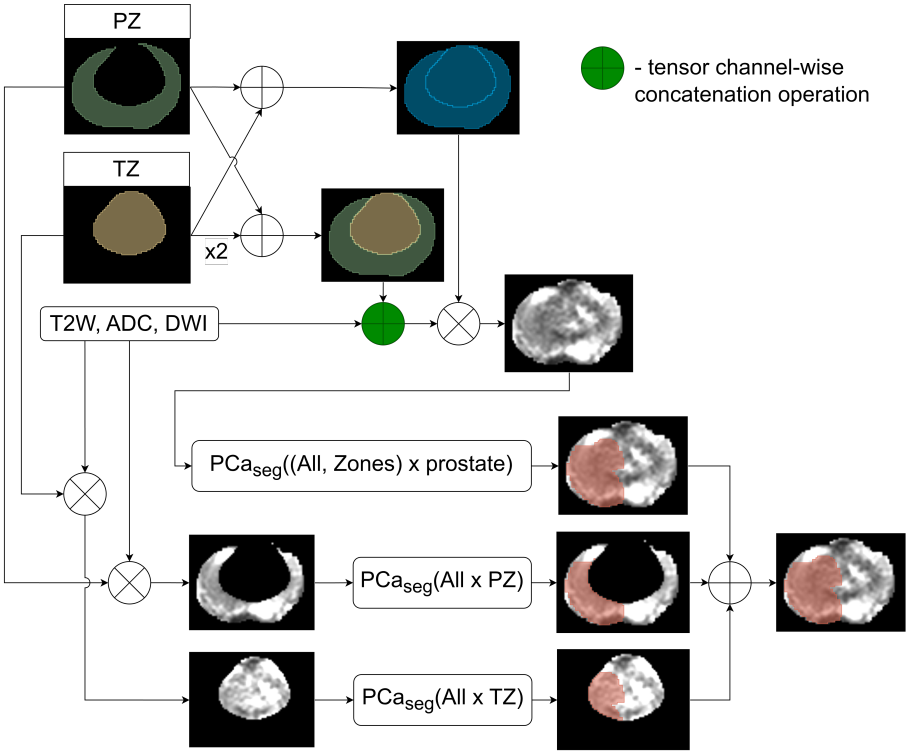


Figure 2.6. Ensemble 3 workflow diagram.

demiological literature on the distribution of prostate cancer instances in prostate zones. This overview indicates that only $\sim 20\%$ of PCa cases are found in TZ. Moreover, research by Oto et al. [57], Rosenkrantz et al. [64], and Yu et al. [85] indicates that PCa segmentation in the TZ is generally more challenging than in the PZ. Therefore, training of models $PCa_{seg}(T2W \times TZ)$ and $PCa_{seg}(All \times TZ)$ is complicated as these models segment PCa in the TZ region. The dataset contains less instances with PCa and it is more difficult to detect a lesion.

Hence, the last proposed ensemble is the disjunction of models $PCa_{seg}(All \times PZ)$ and the baseline model, $PCa_{seg}((All, Zones) \times prostate)$. Formally, $Ensemble\ 4 = PCa_{seg}(All \times PZ) \vee PCa_{seg}((All, Zones) \times prostate)$. This ensemble is denoted as Ensemble 4. Differently from Ensemble 3 it does not have a model which segments prostate cancer in the transitional zone. Therefore, if Ensemble 4 outperforms or achieves similar results to the results of Ensemble 3 then it will indicate that the models, who segment prostate cancer in the TZ, require additional ac-

cumulation of the dataset. The workflow diagram of the last proposed ensemble is displayed in Figure 2.7.

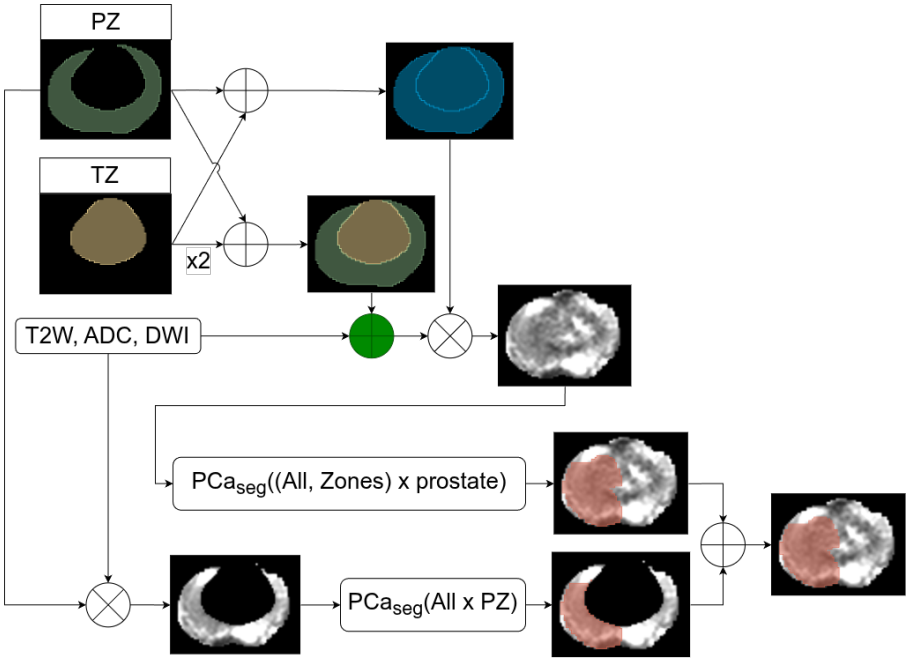


Figure 2.7. Ensemble 4 workflow diagram.

2.3. Compared Prostate Zone Segmentation Workflows

We propose two novel prostate zone segmentation workflows: Separate Models for the Transitional Zone and Peripheral Zone Segmentation and Prostate Segmentation, Transitional Zone Segmentation, and Peripheral Zone Calculation. Both of the workflows utilize deep neural network models trained with the nnU-Net method. The data preprocessing and training of all models are performed independently from one another. The preprocessing and training of each model is analogous. The only difference is the ground truth of the dataset. Furthermore, the input of both models is the T2W modality of mpMRI. The T2W modality is highly sensitive to the water content within tissues, allowing it to provide excellent visualization of soft tissue structures.

The first workflow, SM4TZPZS, consists of two models: one for segmenting the transitional zone and the other for segmenting the peripheral zone. The ground truth of the first model, denoted as $TZ_{seg}(T2W)$,

is TZ, while for the second one, denoted as $PZ_{seg}(T2W)$ is PZ. The SM4TZPZS is visualized in Figure 2.8.

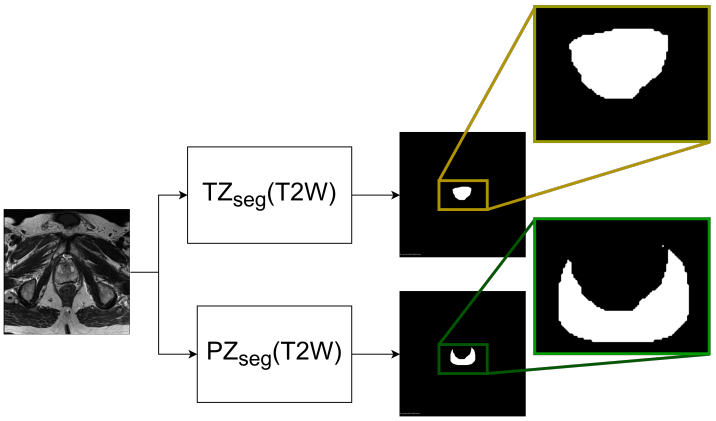


Figure 2.8. The prostate zone segmentation workflow: Separate Models for the Transitional Zone and Peripheral Zone Segmentation.

The second workflow, PSTZSPZC, consists of two models: one for segmenting prostate gland and the other for segmenting TZ. The PZ segmentation of this workflow is calculated from the results of those two models. Overall, PSTZSPZC workflow consists of the following consecutive steps:

1. Segmenting prostate gland by employing a DNN model trained with the nnU-Net method. This model is denoted as $prostate_{seg}(T2W)$. The ground truth of the $prostate_{seg}(T2W)$ model is the prostate mask. This mask is a disjunction of prostate zone masks.

2. Voxels in the T2W image that do not overlap with the predicted prostate gland region are set to zero.

3. Segmenting TZ by employing a DNN model trained with the nnU-Net method. This model is denoted as $TZ_{seg}(T2W \times prostate)$. $TZ_{seg}(T2W \times prostate)$ model's training data preprocessing includes one additional step – the voxels of T2W images which do not intersect with the segmented prostate gland regions are set to zero.

4. The peripheral zone is calculated by subtracting the segmented transitional zone from the segmented prostate gland region.

The PSTZSPZC workflow is visualized in Figure 2.9.

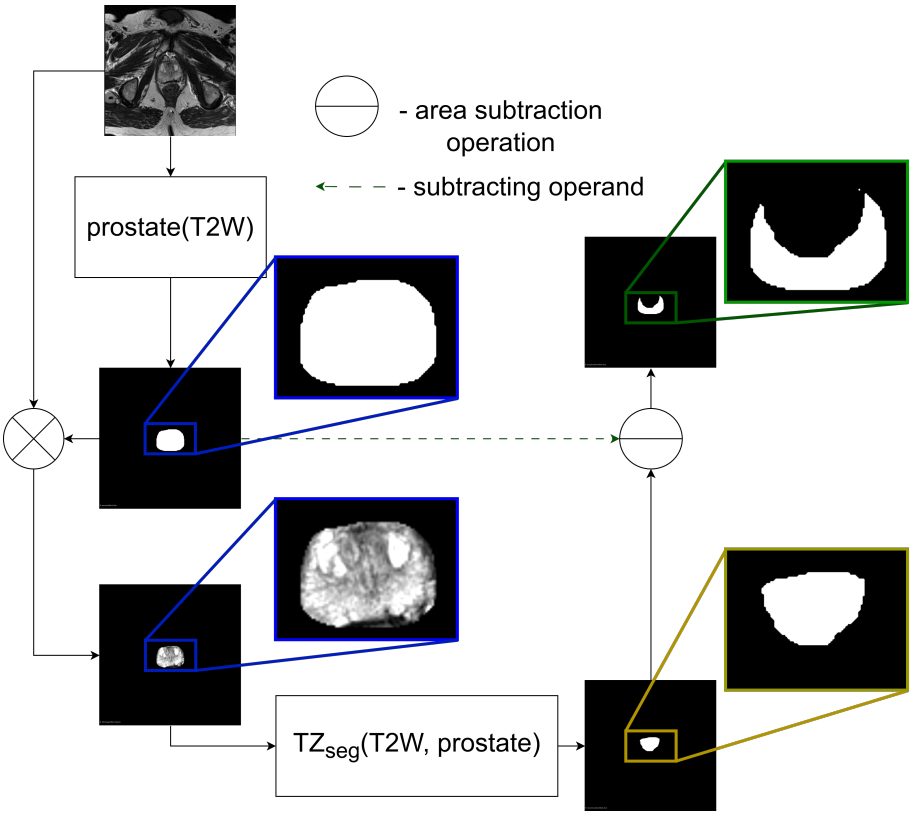


Figure 2.9. The prostate zone segmentation workflow: Prostate Segmentation, Transitional Zone Segmentation, and Peripheral Zone Calculation.

2.4. Compared Deep Neural Network Architectures for the Prostate Zone Segmentation Task

This section provides a description of the proposed method to incorporate attention mechanism into the no-new U-Net method for the prostate zone segmentation task. This incorporation consist of replacing the original U-Net architecture used in the nnU-Net method with the following deep neural network architectures:

- Grid attention U-Net
- Single attention U-Net

- Multidimensional attention U-Net
- Dual-attention U-Net

2.4.1. Grid Attention U-Net

The first incorporated DNN architecture into the nnU-Net method is grid attention U-Net. This architecture employs the grid attention mechanism proposed by Oktay et al. [56], and Schlemper et al. [69]. This mechanism uses grid-based gating, which was introduced by Jetley et al. [36]. The attention mechanism, which employs grid-based gating, is capable of being more specific to local regions according to the authors of the paper [36]. The diagram of this mechanism is visualized in Figure 2.10.

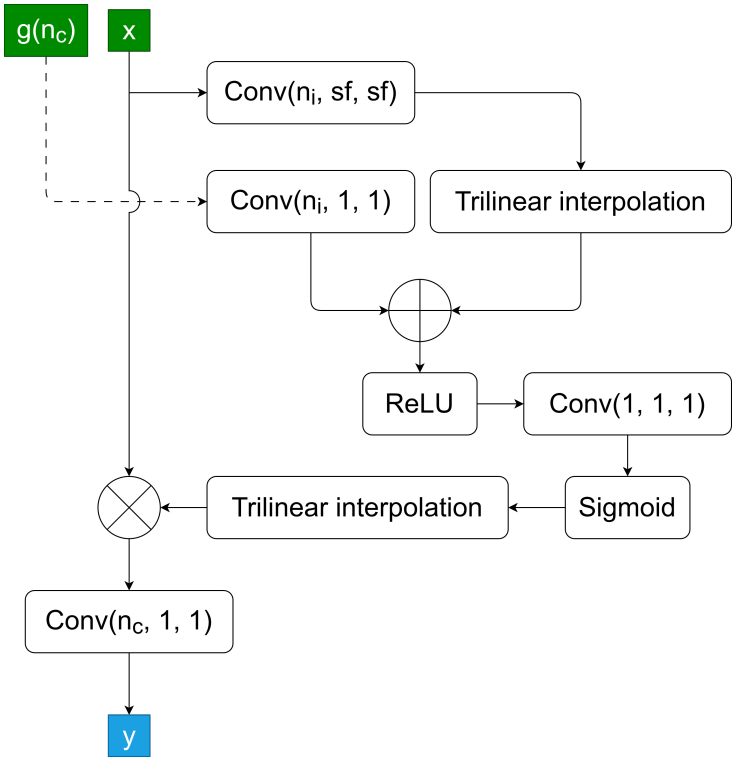


Figure 2.10. Diagram of the grid attention mechanism

In Figure 2.10 x is an input signal and $g(n_c)$ is a gating signal, where n_c is the number of channels. Both signals are the inputs of the

attention mechanism. Meanwhile, y is the output of the attention mechanism. Furthermore, squares $Conv(n, size, stride)$ within the diagram indicate convolution layers where n is the number of kernels, $size$ is the kernel size and $stride$ is a stride. Moreover, squares $ReLU$ and $Sigmoid$ represent activation layers, ReLU and sigmoid respectively. $Trilinear interpolation$ is a trilinear interpolation operation. This operation is used to perform resampling within a grid attention mechanism.

Grid attention mechanism has two hyperparameters. In our experiments, we use the same values for these hyperparameters as those reported in the studies reported in the papers [56, 69].

- The subsample factor – the value of this hyperparameter – two for all used grid attention mechanisms. In Figure 2.10 subsample factor is represented by sf.
- The number of the kernels in the intermediate convolution layer – the value of this hyperparameter is equal to the number of the channels of the input signal. The number of the kernels of the intermediate convolution layer is indicated as n_i in the diagram of the grid attention mechanism.

The other hyperparameters used in the U-Net architecture of the nnU-Net method are not changed. This mechanism is used at each skip connection in the grid attention U-Net. The input signal of each grid attention mechanism is the skip connection itself. Furthermore, the output of the grid attention mechanism is concatenated together with the output of the previous layer on the channel axis. Meanwhile, the gating signal is the output of the gating block. The gating block consists of the convolution layer, batch normalization layer and ReLU activation layer. The hyperparameters of the convolution layer are:

- The kernel count equals the bottleneck kernel count.
- Both kernel size and stride are equal to 1.

The input signal of the gating block is the output of the bottleneck of the architecture. The incorporated architecture of the grid attention U-Net is displayed in Figure 2.11 where the block gates is a gating block.

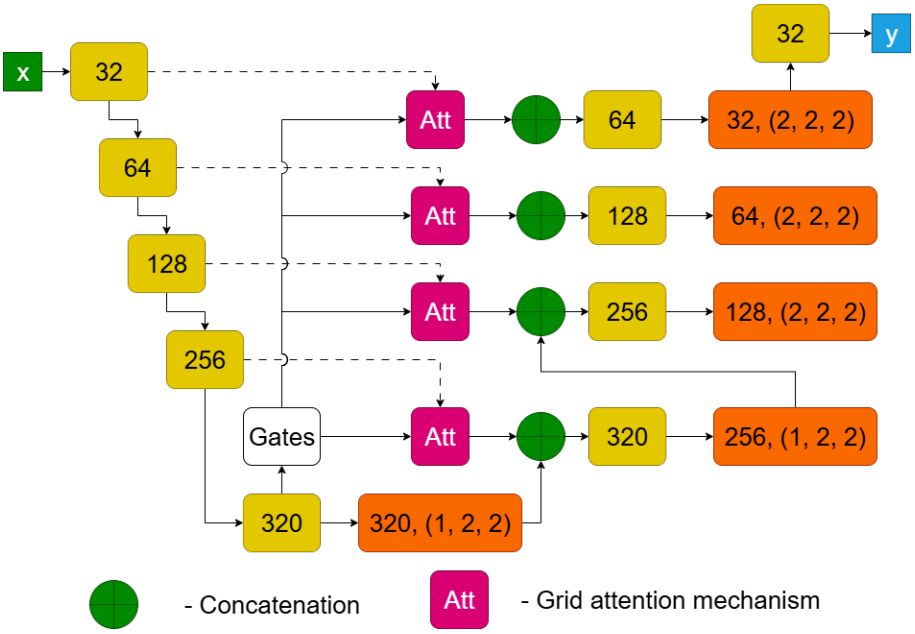


Figure 2.11. Diagram of the grid attention U-Net architecture

2.4.2. Single Attention U-Net

The other incorporated U-Net variation which uses attention mechanism is single attention U-Net. Just as a grid attention U-Net this architecture was introduced by Oktay et al. [56], and Schlemper et al. [69]. This architecture employs the same grid attention mechanism, which is described in the previous section. Differently from the grid attention U-Net, the gating signal of each attention mechanism is the output of the previous layer in the single attention U-Net architecture. Furthermore, the single attention U-Net architecture does not employ a gating block. The other hyperparameters of DNN are taken from the nnU-Net method. The incorporated architecture of the single attention U-Net is displayed in Figure 2.12.

2.4.3. Multidimensional Attention U-Net

The third incorporated deep neural network architecture is multidimensional attention U-Net. This architecture was introduced by Oktay et al. [56], and Schlemper et al. [69] as well. Differently from a grid attention U-Net and single attention U-Net, a multidimensional attention

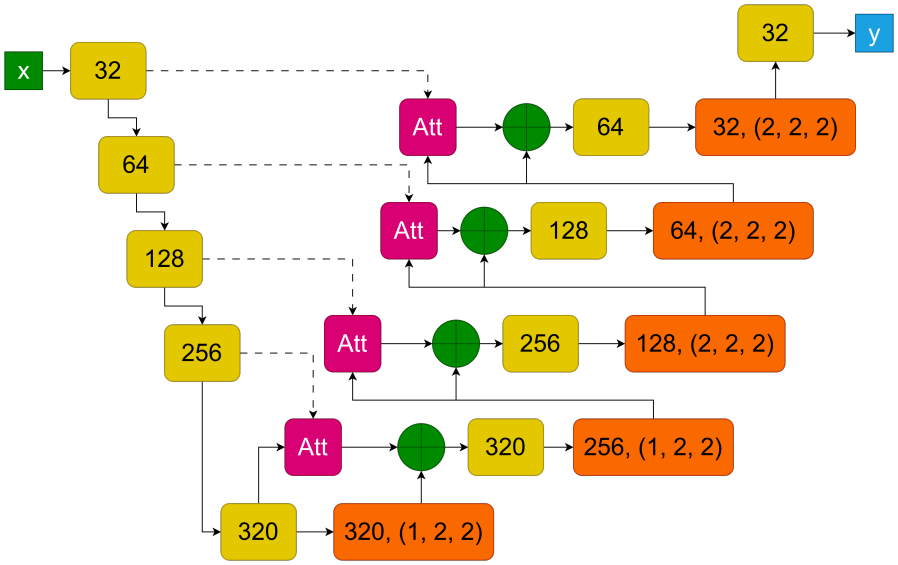


Figure 2.12. Diagram of the single attention U-Net architecture

U-Net employs an extended grid attention mechanism, a multidimensional attention mechanism. The multidimensional attention mechanism uses multiple parallel grid attention mechanisms. Multiple grid attention mechanisms enables the extraction of complementary information from the input signal. The outputs of the grid attention mechanisms are concatenated on the channel axis and processed by a convolution layer with hyperparameters:

- The number of kernels is equal to the number of channels of the input signal.
- Both kernel size and stride are equal to 1.

Finally, batch normalization and ReLU activation function is performed on the results. The multidimensional attention mechanism is visualized in Figure 2.13. The number of grid attention mechanisms used is a hyperparameter of this architecture, referred to as the number of dimensions. In the experiment reported in the paper [69] the authors use the number of dimensions equal to the number of positive classes. The other hyperparameters of the DNN are taken from the nnU-Net method. In our experiments the value of this hyperparameter is 2 as the positive classes are the PZ and the TZ.

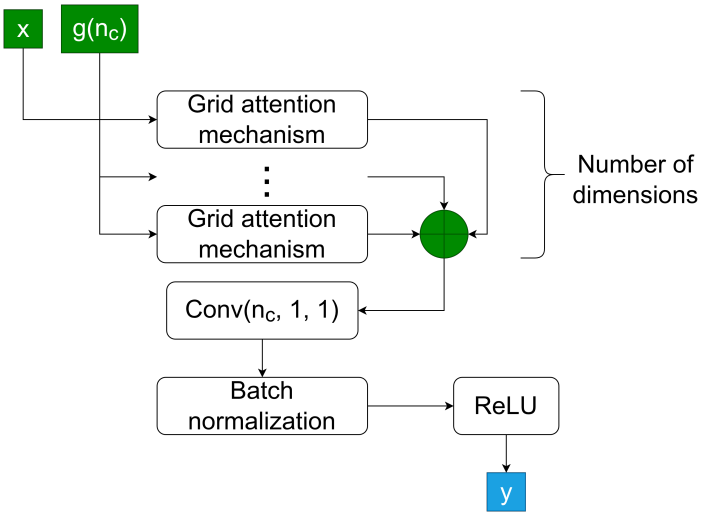


Figure 2.13. Diagram of the multidimensional attention mechanism

The architecture of the multidimensional attention U-Net is analogous to the grid attention U-Net. The only difference is in the type of attention mechanism.

2.4.4. Dual-Attention Attention U-Net

The final incorporated DNN architecture is the dual-attention U-Net, which was introduced by Saha et al. [66], which was described in Chapter 1. We replaced ReLU activation function, used by Saha et al. [66], with Leaky ReLU as it is used in the U-Net architecture of the nnU-Net method. Squeeze-and-Excitation residual block has two hyperparameters: the number of kernels (represented by n_k in Figure 1.2) and stride size (represented by s in Figure 1.2) of the convolution layers located within the block. Furthermore, dual-attention U-Net employs grid attention mechanism. The hyperparameters of these mechanisms are analogous to the hyperparameters of the incorporated grid attention U-Net. Lastly, the hyperparameters of downsampling and upsampling blocks are applied to the SE residual blocks. The resulting architecture of the incorporated dual-attention U-Net is visualized in Figure 2.14.

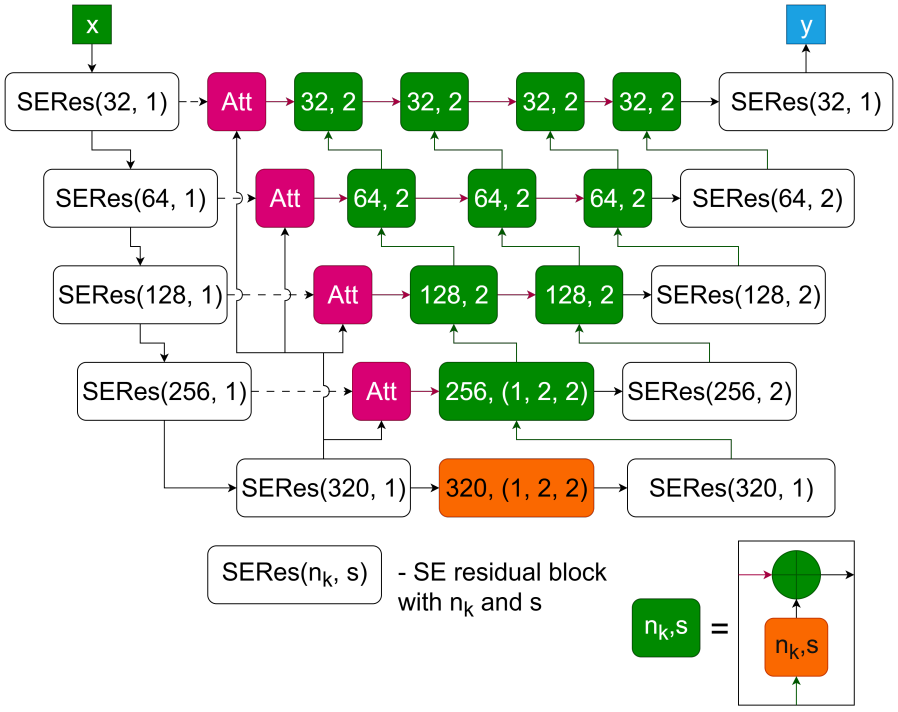


Figure 2.14. Diagram of the dual-attention U-Net

2.5. Performance Measures

In order to ensure the reliability of the proposed approaches we evaluate them by calculating the performance metrics. The particular metric used is DSC described by Klein et al. [41]. This metric is extensively used in the experiments on prostate zone and prostate cancer segmentation tasks. The examples of experiments on the prostate zones segmentation task are presented by Aldoj et al. [2], Antonelli et al. [3], and Comelli et al. [19]. Meanwhile, the examples of experiments on the PCa segmentation task are provided by Mehralivand et al. [47], Mehta et al. [48], and Saha et al. [66]. Furthermore, DSC is incorporated into the loss function of the no-new U-Net method and is calculated with the following equation:

$$DSC = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{2.3}$$

where TP is true positives, which is the number of voxels correctly

assigned to a positive class. In our experiments these classes are either cancerous region, transitional or peripheral zones. Meanwhile, FP is false positives, which is the number of voxels that are not assigned correctly to a positive class. Finally, FN is false negatives, which is the number of voxels incorrectly assigned a positive class.

It is important to note that in the studies presented in [2] and [19], the evaluation of results was not limited to DSC. In [2], the mean relative absolute volume difference and mean Hausdorff distance were also employed as evaluation metrics. In contrast, [19] utilized sensitivity, positive predictive value, volume overlap error, and volumetric difference. Similarly, the research conducted by Joskowicz et al. [38] assessed performance using both DSC and volumetric difference. Despite the inclusion of these additional metrics, Aldoj et al. [2], and Comelli et al. [19] emphasized DSC as the primary performance measure. Therefore, in our research on the prostate zone segmentation task, we use only DSC metric. Each prostate zone is evaluated independently from one another.

Furthermore, metrics such as average precision and the area under the receiver operating characteristic curve, as employed in the PI-CAI challenge described by Saha et al. [67], are well suited for lesion-level detection and case-level clinical decision support. These metrics have also been used in previously published work [66]. In contrast, we focus on voxel wise segmentation of prostate cancer lesions, where spatial overlap and boundary accuracy are of primary interest. Therefore, Dice Similarity Coefficient is selected as the principal evaluation metric.

Evaluating the accuracy of prostate cancer segmentation remains a challenging task, largely due to pronounced inter-observer variability in tumour boundary annotations, even for experienced radiologists and pathologists. Such variability is driven by poorly defined lesion borders, heterogeneous tumour appearance on multiparametric MRI, and partial volume effects. Consequently, segmentation results may differ slightly at the boundaries in ways that are clinically negligible, yet these differences are disproportionately penalized by standard overlap-based evaluation measures, including DSC. In response to this limitation, this work introduces a new evaluation measure, termed DSC_τ designed to better account for clinically acceptable boundary uncertainty. This metric is a modified Dice Similarity Coefficient. The pro-

posed metric, DSC_τ introduces a spatial tolerance that reduces the impact of false positive and false negative predictions occurring within a predefined margin, while retaining the conventional definition of true positives. This approach is motivated by clinical practice, in which small deviations along lesion boundaries are commonly attributable to annotation uncertainty and should not be regarded as genuine segmentation errors. The DSC_τ is calculated by using the following equation:

$$DSC_\tau = \frac{2TP}{2TP + FN_\tau + FP_\tau} \tag{2.4}$$

In this equation, FN_τ represents the number of tolerant false negatives, and FP_τ represents the number of tolerant false positives. These tolerant metrics account for minor segmentation errors in the ground truth. Specifically, FN_τ is computed by dilating the predicted volume, whereas FP_τ is computed by dilating the ground truth volume. As described by Vincent [79], dilation is a morphological operation that enlarges foreground regions in a binary image by adding pixels along object boundaries. The operation is performed by sliding a structuring element across the image and assigning each pixel the maximum value within the neighbourhood defined by the structuring element. Dilation is commonly employed to fill small gaps, connect nearby structures, and enhance the continuity of object boundaries. The dilation operation of FN_τ and FP_τ employs an annulus-shaped kernel. The kernel radius is adjusted so that the dilation expands the regions by a fixed distance, referred to as the tolerance. Novel evaluation metrics, which use tolerance, were also employed in the experiments described by Piert et al. [59], and Zheng et al. [88]. The calculation of metric, introduced by Piert et al. [59], expands the PCa segmentation by 5 mm. Similarly, the authors of [88] define predicted local maxima in prostate cancer probability maps as true positives if they lie within 5 mm of any cancer region. Moreover, 5 mm tolerance is further supported by expert consensus as a clinically appropriate spatial threshold. Hence, a tolerance of 5 mm was adopted for the calculation of DSC_τ. In other words, the radius of the kernel, used in dilation operations, is calculated so that it represents a physical size of 5 mm, based on the pixel spacing information provided in the image metadata. Due to extensive evidence in the literature supporting a fixed tolerance value of 5 mm, the sensitivity of this metric to alternative tolerance values is not explored in this the-

sis. However, further analysis using smaller tolerance values could be undertaken in future research. Figure 2.15 presents the Venn diagram illustrating the calculation of FN_τ and FP_τ.

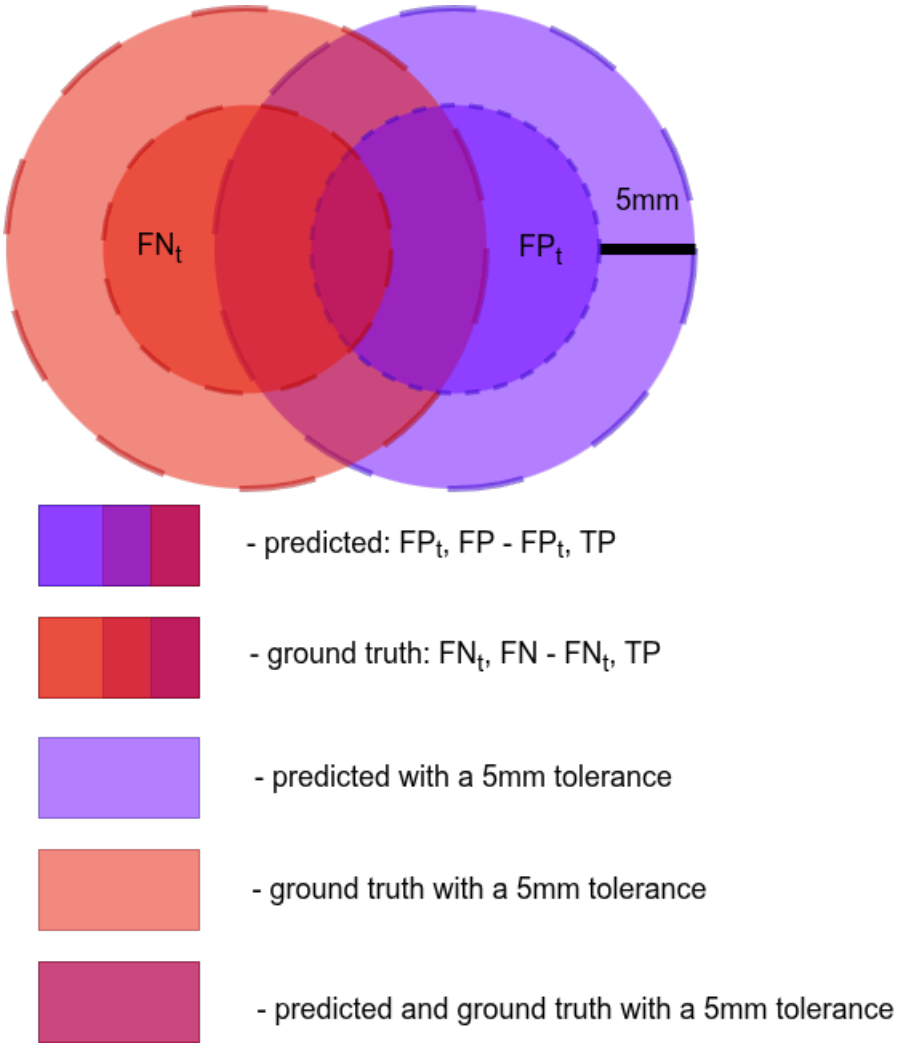


Figure 2.15. Venn diagrams illustrating the calculation of FN_τ and FP_τ.

Several tolerant variants of Dice Similarity Coefficient have been proposed previously. One such metric, introduced by Colonna and Scarpa [18], modifies the calculation of true positives by incorporating a tolerance margin. In contrast, DSC_τ proposed in this study preserves a strict definition of true positives, maintaining sensitivity to clinically meaningful lesion detection while reducing undue penaliza-

tion of ambiguous boundary regions. Other related metrics include Surface DSC [55], Weighted DSC (WDC) [60], and Added Path Length (APL) [76]. These approaches differ fundamentally from DSC_τ. Firstly, surface DSC and APL evaluate the overlap of surfaces rather than volumetric regions as is done by the conventional Dice Score Coefficient and DSC_τ. Meanwhile, WDC incorporates the distance between predicted and ground truth voxels when computing volumetric overlap. Although WDC reduces the impact of minor boundary discrepancies compared to the conventional DSC, it still penalizes them more heavily than DSC_τ, which explicitly accounts for clinically acceptable spatial tolerance.

Furthermore, DSC and DSC_τ cannot be calculated for cases where the ground truth does not contain any cancerous regions, as this would result in an undefined outcome. Therefore, false positive patients are calculated for each ensemble as well as the baseline model. The false positive patients metric represents the percentage of patients whose predictions include a cancerous region, while their ground truth does not. Finally, we report performance at the lesion level. Specifically, the distributions of lesion counts generated by each ensemble and the baseline model, along with the distributions of undetected lesions, corresponding to lesion level false negatives. Furthermore, our study includes the distributions of spurious foci produced by each ensemble and by the baseline model, representing lesion level false positives.

Lastly, we perform statistical tests to determine statistically significant difference between the compared approaches. We use Friedman tests, introduced by Friedman [27], to assess the significance of the difference between the results of:

- Compared prostate zone segmentation workflows.
- Compared deep neural network architectures used in the experiments on the prostate zone segmentation task.
- Compared ensembles used for the prostate cancer segmentation task.

The Friedman test is a non-parametric statistical test used for related groups, so it is appropriate to assess whether the mean ranks

of the group population differ for small and non-normally distributed samples.

Furthermore, the Conover post hoc test for unreplicated block design, introduced by Conover [20], are conducted whenever the statistically significant difference is determined. Conover post hoc test allows us to determine which groups are responsible for the observed difference. The p-values for these tests are adjusted by employing the Benjamini/Hochberg (non-negative) method.

Conover post hoc test is used whenever there are three or more groups. However, it is more appropriate to use post hoc Wilcoxon signed-rank test whenever there are precisely three groups. This test was introduced by Wilcoxon [82]. Considering that three prostate zone segmentation workflows are compared, a post hoc Wilcoxon signed-rank test is applied to determine which workflow pairs exhibit significant differences. Just as with Conover post hoc test, the p-values of post hoc Wilcoxon signed-rank tests are adjusted by using the Benjamini/Hochberg (non-negative) method.

Finally, a two-sided Wilcoxon signed-rank test is performed to evaluate the significant difference between collected DSC and DSC_τ. Just as the Friedman test, a two-sided Wilcoxon signed test is a non-parametric statistical test. This test is used for two sample tests and it ranks the differences between paired observations and evaluates whether the ranks of positive and negative differences differ significantly.

2.6. Summary of the Chapter

This chapter provides a detailed overview of the proposed approaches for prostate cancer and prostate zone segmentation tasks as well as methods employed in these approaches. Proposed PCa segmentation approaches consist of four unique ensembles. Each ensemble utilizes prostate zonal information to segment prostate cancer. Furthermore, the prostate zone segmentation is used to acquire prostate zones. Namely, transitional and peripheral zones. These zones are later employed to improve PCa segmentation solutions. The proposed approaches for prostate zone segmentation consist of workflow and DNN architecture, which is used to substitute the original U-Net used in the nnU-Net method. The first workflow is Separate Models for the Transitional Zone and Peripheral Zone Segmentation. The other workflow

is Prostate Segmentation, Transitional Zone Segmentation, and Peripheral Zone Calculation. Meanwhile, DNN architectures are U-Net modifications with an attention mechanism. These architectures consist of grid attention U-Net, single attention U-Net, multidimensional attention U-Net and dual-attention U-Net. All of the approaches employ the nnU-Net method for training and inference of deep neural network models.

Next, a novel metric, DSC_τ is defined. This metric addresses a high inter-observer variance which arises due to the nature of manual prostate cancer segmentation. The other collected metrics, DSC and false positives per patient, are described in this chapter together with DSC_τ.

Finally, performed statistical tests on collected metrics are described in this chapter. These tests evaluate the presence of the statistically significant difference between compared approaches. These statistical tests consist of Friedman tests, two-sided Wilcoxon signed-rank tests, Conover post hoc tests for unreplicated block design and post hoc Wilcoxon signed-rank tests.

This chapter presents the empirical experiments on the methods described in the previous chapter, the data used in the experiments and the data sources. Firstly, the dataset, used for the experiment on the prostate cancer segmentation task, comes from a different source than the dataset, used for the experiment on the prostate zone segmentation task. Moreover, the data format is not suitable for the experiments. Therefore, this chapter outlines the required format for each researched task and the preprocessing steps for the datasets from each source. Lastly, this chapter details setups of the experiments. This information includes used validation and testing techniques employed for each experiment, collected metrics and performed statistical tests.

Moreover, each training run is executed using a different random seed. All training procedures and inference experiments are carried out on an NVIDIA DGX-1 GPU server equipped with an architecture comprising eight V100 GPUs. These experiments were presented in the main papers by Vaitulevičius et al. [A.1], [A.2] and [B.4].

3.1. Dataset of the Prostate Cancerous Region Segmentation Task

The dataset used for experiment on the prostate cancerous region segmentation task consists of T2W, ADC and DWI modalities and ground truth masks. The data is provided by the National Cancer Institute (NVI). The dataset is not publicly available and it was granted under the terms of the bioethical agreement (number 2020/5-1229-714). It consists of all the necessary modalities of the mpMRI for the experiment on the prostate cancer segmentation task. All of the modalities are volumetric images. The dataset was acquired across 11 medical centres in Lithuania. The initial cohort comprised 146 patients; however, 26 cases were excluded due to imaging artefacts, such as the presence of hip implants, or missing MRI sequences. This resulted in a dataset of 120 patients. The centres employed a variety of MRI scanners, each operating with distinct acquisition parameters. Seven centres utilised 1.5T scanners, whereas the remaining centres used scanners with a magnetic field strength of 3T. In addition, the mpMRI acquisitions varied in both echo time and repetition time across institutions. Summary statistics

for these parameters are presented in Table 3.1, where the Modality column denotes the mpMRI sequence and the Parameter column specifies the parameter. For each parameter, the minimum, maximum, mean, and median values are reported. All scanners produced volumetric datasets with an inter-slice spacing of 2.5 mm, while the in-plane voxel dimensions ranged from 0.39 mm to 0.78 mm.

Table 3.1. Statistical summary of the echo time and repetition time parameters in the acquired T2-Weighted Imaging and Diffusion-Weighted Imaging modalities of the NVI dataset (in ms).

Modality	Parameter	min	mean	median	max
T2W	Repetition time	3000	5334.74	4310.31	9847
	Echo time	85	132.81	120	162
DWI	Repetition time	1294.29	4114.89	3918.35	10267
	Echo time	66.19	79.57	78.25	93.8

NVI also provided manually delineated prostate and cancerous region masks as well as the results and the masks of the biopsies. These masks are volumetric images just as the modality images. Manual delineation of cancerous regions was conducted by a radiologist with a decade of professional experience. All segmentations were completed before any modelling or evaluation took place, ensuring that the annotator had no exposure to the outputs of either the ensemble methods or the baseline model. Biopsy sampling was performed independently by an interventional radiologist with 15 years of experience. One patient exhibited manually annotated cancerous region that was not substantiated by biopsy analysis. Owing to this inconsistency, the patient was removed from the study cohort, yielding a final dataset of 119 cases. The full patient selection workflow is illustrated in Figure 3.1.

In this final cohort, the ages of the patients span from 46 to 82 years, with both the mean and median age being 64. The biopsy burden varied across individuals: patients received an average of 15 biopsies, with a median of 16, and the total number ranged between 3 and 25 per patient. Table 3.2 summarises the distribution of biopsy outcomes by the Gleason score. The GS column enumerates the assigned Gleason grades, while the Category column specifies the associated diagnostic classification. The Number of patients column reports how many pa-

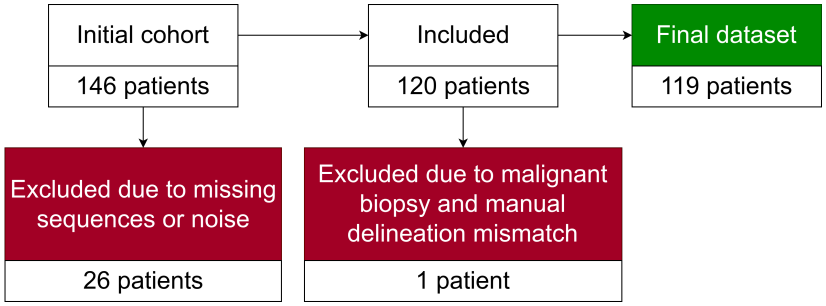


Figure 3.1. Flow diagram illustrating the patient selection and exclusion process.

tients had at least one biopsy exhibiting a given GS, and the Number of biopsies column counts the total number of samples with that score. The average biopsies per patient describes the mean number of biopsies with that GS among patients with at least one such sample.

Table 3.2. Gleason score distribution.

Category	GS	Number of patients	Number of biopsies	Average biopsies per patient
Malignant	3+4	34	97	3
Malignant	4+3	16	31	2
Malignant	4+4	5	11	2
Undetermined	3+3	72	293	4
Benign	0+0	112	1316	12

Moreover, the dataset is extended with prostate zonal masks, and all imaging sequences and segmentations are provided as volumetric datasets. Furthermore, the ground truth is calculated from masks provided by NVI, namely, the prostate, cancerous region and biopsy results masks. Firstly, connected components within manually delineated cancerous regions are calculated. Each component represents an individual manually delineated lesion within the segmentation mask. The ground truth is defined as the subset of manually delineated lesions that are corroborated by malignant biopsy findings. Prostatic biopsies are acquired by employing a standard grid, which contains evenly spaced holes as described by Bhat et al. [10], and Sahni et al. [68]. The distance between holes is 5 mm, which introduces a biopsy

localization error of 5 mm radius. Therefore, if the manually delineated lesion is not within 5 mm distance from malignant biopsy mask or the malignant biopsy does not intersect with the lesion, then manually delineated lesion is excluded from the mask. In other words, the manually delineated lesion C is removed if:

$$\neg\exists x \in (dilate(biopsies_{malignant}) \wedge C) \tag{3.1}$$

where:

- x is a pixel of the $dilate(biopsies_{malignant}) \wedge C$ result with a binary value.
- $dilate$ is a dilation operation that uses a ball-shaped kernel. The radius of this kernel is calculated so that it represents a physical size of 5 mm, based on the pixel spacing information provided in the image metadata.
- $biopsies_{malignant}$ is a biopsy result mask that includes only the biopsies labelled as malignant.

The resulting distribution of lesion counts and lesion size statistics across prostate zones within the final dataset is reported in Table 3.3. In this table Zone column corresponds to the prostate zone. Each column within the Number of lesions group contains the number of patients exhibiting the corresponding number of lesions within the specified prostate zone. Meanwhile, the lesion size statistics group contains the lesion size statistics within the given prostate zone. These statistics consist of minimum, maximum, mean and median lesion sizes.

Table 3.3. Distribution of lesion counts and lesion volume statistics across prostate zones, with volumes reported in mm^3.

Zone	Number of lesions		Lesion size statistics			
	1	2	min	mean	median	max
prostate	36	6	71.38	1557.71	905.53	7694.17
TZ	32	2	0.67	1037.24	440.89	7283.42
PZ	35	6	12.98	670.9	412.5	3356.48

As mentioned before, our research inspects the use of prostate zone information in prostate cancer segmentation task. The prostate zones

are acquired by employing a SM4TZPZS workflow, which is described in Chapter 2. However, further inspection indicated faults within the resulting zones. These faults include gaps between prostate zones and overlaps. These inaccuracies stem from inherent constraints of deep neural network models, including prediction ambiguity near zonal interfaces and susceptibility to partial volume artefacts. Consequently, such errors have the potential to adversely influence the performance of both the ensembles and the baseline model. Moreover, the acquired prostate zones very often contain areas that do not intersect with NVI provided prostate masks. Therefore, the prostate zones are further processed. Firstly, the PZ is adjusted by employing the following function:

$$f_{PZ}(PZ, TZ) = dilate(PZ) \wedge closing(PZ \vee TZ) \wedge \neg TZ \tag{3.2}$$

where:

- $dilate$ is a dilation operation with a ball kernel of a 5-pixel radius.
- $closing$ is a binary morphological closing operation, which is defined as $closing(x) = erosion(dilation(x))$. Just as $dilation$, $erosion$ operation was introduced by Vincent [79]. This operation is performed with a ball shaped kernel of a 10-pixel radius.

The $dilate(PZ) \wedge closing(PZ \vee TZ)$ part reduces the gaps between zones, while conjunction with a negated TZ removes the overlaps. Afterwards, the adjusted PZ is used for processing the TZ. The adjustment is the following function:

$$f_{TZ}(PZ, TZ) = dilate(TZ) \wedge closing(PZ \vee TZ) \wedge \neg f_{PZ}(PZ, TZ) \tag{3.3}$$

Once again, the $dilate(TZ) \wedge closing(PZ \vee TZ)$ part further reduces the gaps between zones, while conjunction with a negated processed PZ ensures that new overlaps would not appear. The kernel dimensions used for the dilation and morphological closing operations are determined empirically, with the aim of bridging minor discontinuities while maintaining the anatomical integrity and overall geometric characteristics of the prostate zones.

Lastly, prostate zone areas that do not intersect with prostate mask provided by NVI are removed. This last step is only performed on training set.

The resulting dataset statistics is provided in Table 3.4, with the following columns:

- Zone indicates the zone of the prostate (TZ, PZ or the whole prostate).
- has cancerous region - does the patient have a cancerous region or not.
- number of cases - number of patients, who either have or does not have a lesion in the ground truth (indicated by the column has cancerous region) within a Zone.

Table 3.4. Processed dataset size statistics.

Zone	has cancerous region	number of cases	percentage
Prostate	Yes	42	$\sim 35.29\%$
	No	77	$\sim 64.71\%$
TZ	Yes	34	$\sim 28.57\%$
	No	85	$\sim 71.43\%$
PZ	Yes	41	$\sim 34.45\%$
	No	78	$\sim 65.55\%$
Total		119	

3.2. Experiment Setup of the Prostate Cancer Segmentation Task

The cases with a cancerous region in the ground truth are under-represented as indicted by Table 3.4. Moreover, due to the limited size of the available dataset, a conventional split into separate training and test sets is not employed, as this would result in an insufficient number of samples for robust model training and evaluation. Instead, the proposed ensembles are assessed using stratified five-fold cross-validation, which enables a more reliable estimation of model performance by iteratively training and validating on different data partitions. Therefore, a dedicated evaluation on an independent held-out test set should be undertaken for future research. Moreover, the use of five-fold cross-validation mitigates the influence of random seed initialization on the

training outcomes. The stratification is performed by evenly distributing patients with cancerous regions in the ground truth across the folds. Accordingly, all deep neural network models are trained five times, with each training run corresponding to a distinct fold. As a consequence, both the ensembles and the baseline model are likewise evaluated five times, each using a separate test fold. Four of the five folds contain 24 patients, resulting in training sets of 95 cases for those folds and test sets comprising 24 cases. The remaining fold includes 23 patients; therefore, in this instance each model contributing to the ensembles is trained on 96 cases, and the resulting ensembles are evaluated on a test set of 23 cases. The results of all five folds are merged into a single test result set.

A total of four ensembles are compared. These ensembles are described in Chapter 2 together with the baseline model, $prostate(All, Zones)$. The scheme of this experiment is visualized in Figure 3.2.

All deep neural network models are trained with the no-new U-Net method using a batch size of 4 for 1,000 epochs. The test results consist of prostate cancer segmented for each patient by employing compared approaches separately. Thus, each patient has 5 segmentations as we compare 4 ensembles together with one baseline model. Each segmentation of each patient is evaluated by calculating DSC and a novel metric, DSC_τ. Lastly, the metrics are aggregated by calculating mean, standard deviation (std), median and Median Absolute Deviation (MAD) for each ensemble together with the baseline model. However, these metrics cannot be calculated for cases where the ground truth does not contain any cancerous regions, as this would result in an undefined outcome. Therefore, false positive patients are calculated for each ensemble as well as the baseline model. A total of 42 cases have cancerous region in their ground truth. Hence, DSC and DSC_τ are calculated for these 42 patients only. Meanwhile, false positive patients are calculated for the remaining 77 patients.

Additionally, we report performance at the lesion level: the distributions of lesion counts generated by each ensemble and the baseline model, the distributions of undetected lesions, and the distributions of spurious foci produced by each ensemble and by the baseline model.

Moreover, Friedman tests are performed in order to determine the presence of statistically significant difference between the results of the

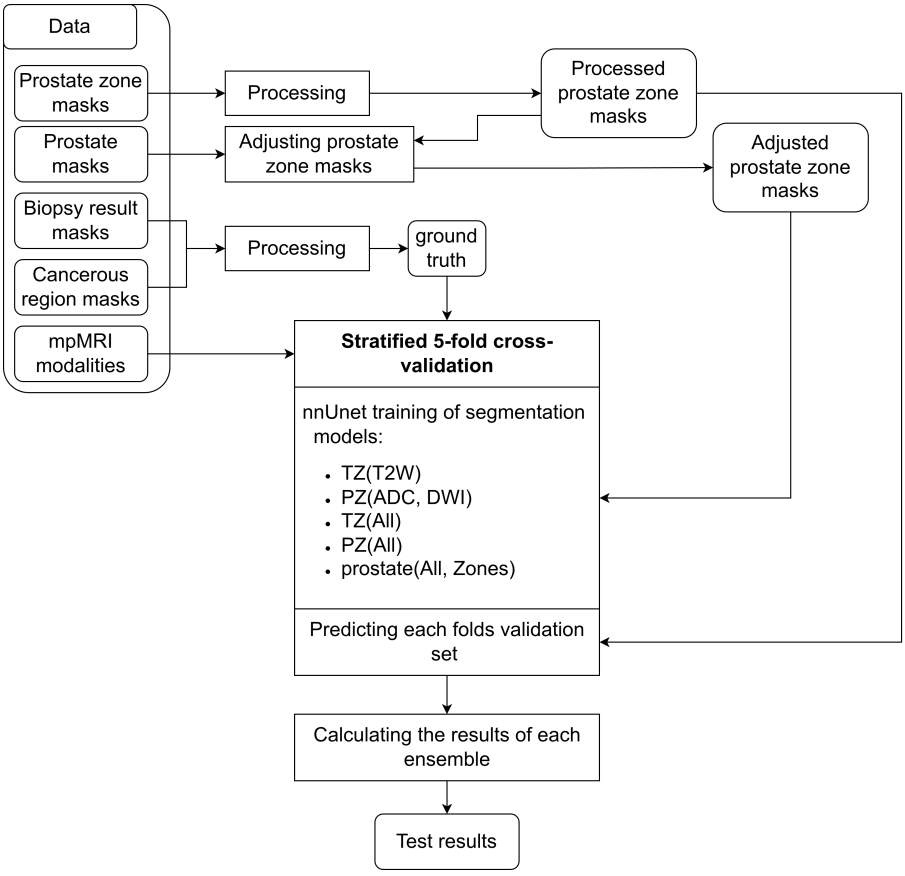


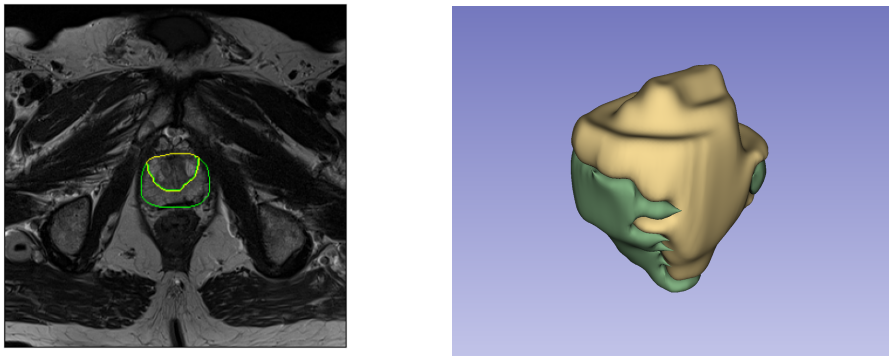
Figure 3.2. Experiment schema for the prostate cancer segmentation task.

ensembles and the baseline model. Friedman tests are performed twice - once with the collected sample of DSC and with the collected sample of DSC_τ. The chosen confidence level is 95% (significance level is 0.05). If statistically significant difference is determined, then Conover post hoc test for unreplicated block design are performed.

Lastly, we perform Wilcoxon signed-rank tests to determine whether the observed difference between DSC and DSC_τ are statistically significant. We use a standard confidence level of 95% (significance level 0.05). These tests are performed for each compared ensemble as well as the baseline model independently from one another. The sample size is 42 as DSC and DSC_τ are only calculated for the cases that possess cancerous regions in their ground truth.

3.3. Dataset of Prostate Zone Segmentation

The dataset used for experiment on the prostate zone segmentation task consists of T2W modality images and prostate zone masks. T2W images are volumetric and the example of a single slice is visualized in Figure 3.3 (a) with contours marking the prostate zones. Meanwhile, prostate zone masks are used as ground truth in the experiment on the prostate zone segmentation task. These masks consists of volumetric categorical images with two categories: peripheral and transitional zones. The example of prostate zone mask is provided in Figure 3.3 (b).



(a) (b)

Figure 3.3. (**a**) A single slice example of T2W modality. The yellow contour marks the transitional zone while the green one is the peripheral zone. (**b**) A volumetric mask of transitional (yellow) and peripheral (green) zones.

The dataset used in the prostate zone segmentation task comes from two different data sources. Each source requires different preprocessing steps. This section describes data of each source and the preprocessing steps taken to prepare the data for the experiments.

The first source is the dataset from the Decathlon challenge, which is described by Antonelli et al. [3]. This dataset is publicly available and it was provided by Radboud University Nijmegen Medical Center. It consists of 32 cases with prostate zone labels and T2W modality volumetric images. However, a single patient is excluded from the cohort as he does not contain the PZ. Thus, the resulting cohort of the Decathlon challenge dataset contains 31 cases. The magnetic field strength, repe-

tition time and echo time used for the acquisition of this dataset are not publicly available. Meanwhile, the volumetric images have inter-slice spacing ranging from 3 mm to 4 mm, while the in-plane voxel dimensions range from 0.6 mm to 0.75 mm. The majority of the cases (23) have an inter-slice spacing of 3.6 mm and an in-plane voxel dimension of 0.625 mm. No publicly accessible information is available concerning either the methodology employed for prostate zone annotation or the demographic details, including patient age. The images are upside down on the second axis. Therefore, direct discrete Fourier transformation is performed on the images and masks in order to flip them. Lastly, the prostate mask is acquired by calculating a union of TZ and PZ regions.

The other dataset is taken from the ProstateX challenge, which is described by Meyer et al. [49]. This dataset is publicly available as well. It consists of 98 cases with volumetric T2W modality images and ground truth masks. All 98 patients are included in the experiment of our study. The youngest patient in the cohort is 47 years old, while the oldest is 76 years old. The mean and median age of the patients is 64. All of the mpMRI images are acquired with the magnetic field strength of 3T. Moreover, the summary statistics of employed echo time and repetition time parameters in T2W image acquisition is provided in Table 3.5. Furthermore, images with the dataset exhibit an inter-slice spacing varying between 3 mm and 5 mm, whereas the in-plane voxel resolution ranges from 0.3 mm to 0.6 mm. Most cases (76) are characterized by an inter-slice spacing of 3 mm and an in-plane voxel size of 0.5 mm. The masks consist of the following categories: transitional, peripheral zones, distal prostatic urethra and fibromuscular stroma. The ground truth masks are labelled by a medical student experienced in prostate segmentation. Moreover, an experienced urologist provided segmentation instructions to the aforementioned student and conducted a review of the segmentation results.

The masks are processed in order to reduce the categories to TZ and PZ regions as our experiment only uses those prostate zones. A new TZ mask is calculated by using this equation:

Table 3.5. Statistical summary of the echo time and repetition time parameters in the acquired T2W modalities of the ProstateX dataset in ms.

Parameter	min	mean	median	max
Repetition time	4480	5974	5910	7435
Echo time	101	104	104	104

$$TZ_{processed} = TZ \vee stroma \vee (dilate(TZ) \wedge urethra_{TZ,PZ}) \vee (urethra_{TZ} \wedge \neg urethra_{TZ,PZ}) \quad (3.4)$$

where:

- $TZ_{processed}$ is the acquired TZ mask.
- TZ is an unprocessed transitional zone.
- $stroma$ is the fibromuscular stroma.
- $dilate$ is a dilation operation with a ball kernel of 4-pixel radius.
- $urethra$ is the prostatic urethra.
- $urethra_{TZ,PZ}$ is the prostatic urethra with slices that do not contain either transitional and peripheral zones set to 0.
- $urethra_{TZ}$ is the prostatic urethra with slices that do not contain TZ region set to 0.

The $dilate(TZ) \wedge urethra_{TZ,PZ}$ assigns part of the urethra to TZ if the slice contains both TZ and PZ regions. Meanwhile, a new PZ, $PZ_{processed}$, mask is calculated by employing a newly acquired TZ mask, $TZ_{processed}$, and this equation:

$$PZ_{processed} = PZ \vee (urethra \wedge \neg TZ_{processed}) \quad (3.5)$$

Lastly, the prostate mask, $prostate$, is calculated with this equation:

$$prostate = TZ \vee PZ \vee stroma \vee urethra \quad (3.6)$$

3.4. Experiment Setup of the Prostate Zone Segmentation Task

The preprocessed datasets are used to train and evaluate compared deep neural network architectures and workflows on prostate zone segmentation task. The dataset provided by NVI does not contain prostate zone masks. Therefore, training and evaluation of prostate zone segmentation approaches are conducted on publicly available datasets. Firstly, 88 patients from the processed ProstateX dataset are randomly selected for training models. The remaining 10 cases are left for testing. Meanwhile, all 31 cases of the Decathlon dataset are used for test set.

In total 15 approaches are compared. Each approach consists of a DNN and a workflow pair. Compared DNNs are U-Net (baseline), grid attention U-Net, single attention U-Net, multidimensional attention U-Net and dual-attention U-Net. Each of the incorporated DNN is described in Chapter 2. Meanwhile, compared workflows are Single Peripheral Zone and Transitional Zone Segmentation Model (SPZTZSM), SM4TZPZS and PSTZSPZC. Each proposed workflow is described in the chapter 2. Furthermore, the baseline workflow, SPZTZSM is just a single model trained with the nnU-Net method. The input of this model is a T2W modality just as of the proposed workflows and the output is prostate zone segmentation binary maps. The approaches are evaluated on the ProstateX and Decathlon source test sets separately. The results, acquired on the ProstateX source test set, are used to indicate the overall accuracies of the compared approaches. Meanwhile, the results calculated from the Decathlon source dataset also indicate the generalizability between different datasets of the compared approaches. This experiment scheme is visualized in Figure 3.4.

All deep neural network models are trained with the no-new U-Net method using a batch size of 2 for 1,000 epochs. The sufficiently long training procedure reduce the sensitivity of the model to random seed initialization, thereby mitigating its influence on the training outcomes. The test results consist of prostate zone segmentations predicted for each patient individually using all compared approaches. Consequently, the number of segmentations per patient corresponds to the total number of evaluated methods (15), with each method providing a separate segmentation for the same patient. DSC is calculated for each segmentation within each test result on prostate zone basis.

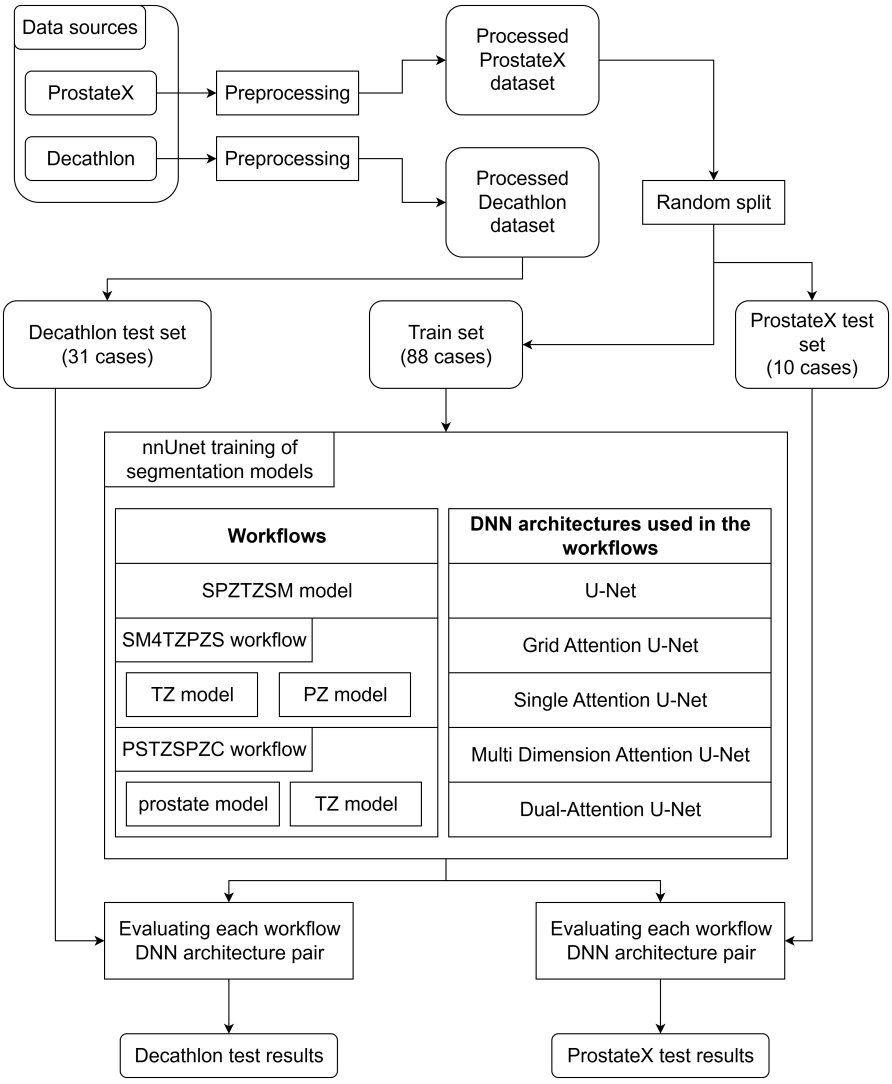


Figure 3.4. Experimental schema for the prostate zone segmentation task.

Therefore, two metrics are calculated for each segmentation, one for the transitional zone and the other for the peripheral zone. These metrics are further examined separately.

Friedman tests are performed on acquired metrics. We use these tests to indicate a presence of statistically significant difference between the compared approaches. Friedman tests are performed for each compared DNN architecture separately. The results of these tests deter-

mine the presence of statistically significant difference between compared workflows. In addition, Friedman tests are also performed for each compared workflow separately as well. The latter results determine the presence of statistically significant difference between compared DNN architectures. The confidence level chosen for these tests is 95% (significance level is 0.05). Furthermore, the Conover post hoc test for unreplicated block design are conducted whenever the statistically significant difference between the compared DNN architectures is determined. This allows us to determine which deep neural network architecture are responsible for the observed difference. Meanwhile, post hoc Wilcoxon signed-rank tests are conducted whenever the statistically significant difference between compared workflows is determined. This allows determining which workflows are responsible for the observed difference.

Lastly, the metrics are aggregated for each DNN architecture and workflow pair. The aggregation is performed by calculating mean, std, median and MAD.

3.5. Summary of the Chapter

This chapter outlines the dataset preparation for the empirical experiments and the overall schema of those experiments. The experiments are designed to compare proposed approaches together with baseline approaches. Comparison is performed by training models necessary for these approaches and evaluating the approaches themselves.

Different approaches are compared for PCa and prostate zone segmention tasks. Furthermore, different datasets are collected for training and evaluation of approaches of each task. PCa segmentation task dataset source is NVI. Meanwhile prostate zone segmentation task dataset sources are ProstateX and Decathlon challenge datasets. Each source has its own preprocessing workflows provided in this chapter.

The result of preprocessing are three datasets: ProstateX dataset, Decathlon dataset and prostate cancer segmentation task dataset. Stratified five-fold cross validation is performed on prostate cancer segmentation task dataset. Meanwhile, the prostate zone segmentation task training set consists of randomly selected 88 patients from the ProstateX dataset. Furthermore, the prostate zone segmentation task approaches are evaluated on two test sets separately:

- ProstateX test set which consist of remaining 10 cases of the ProstateX dataset.
- Decathlon test set which consists of all 31 cases of the Decathlon dataset.

The PCa segmentation task approaches are evaluated on the PCa segmentation task validation set of each fold by calculating DSC and a novel metric, DSC_τ. Meanwhile, the approaches for the prostate zone segmentation task are evaluated on both prostate zone segmentation task test sets by calculating DSC.

The metrics are calculated on a single patient basis. Thus, the comparison of the approaches is performed by aggregating the collected metrics and performing statistical tests. The aggregation is performed by calculating mean, std, median and MAD for both tasks.

The results of the prostate cancer segmentation task approaches are compared on an ensemble and metric basis. Moreover, the Friedman test is performed to evaluate the statistical significance of difference between compared ensembles as well as the baseline model.

The results of prostate zone segmentation task approaches are compared on each prostate zone separately. Therefore, the aggregation of prostate zone segmentation task results are performed on the test set, prostate zone, deep neural network architecture and workflow basis. Moreover, the presence of statistically significant difference between the compared prostate zone segmentation task approaches are evaluated by conducting Friedman statistical tests. Firstly, the Friedman test is performed to evaluate the statistical significance of difference between the compared DNN architectures. Secondly, the Friedman test is performed to evaluate the statistical significance of difference between the compared workflows.

Lastly, for any result where the Friedman test indicates a statistically significant difference, the Conover post hoc test is performed. Except for the results on which Friedman tests indicated statistically significant difference between the compared workflows. On these results, post hoc Wilcoxon signed-rank tests are performed.

4. RESULTS

This chapter presents the results of the empirical experiments on prostate zone segmentation and prostate cancer segmentation tasks. Namely, it provides the aggregated collected metrics on all of the proposed approaches and baselines. Moreover, it contains the statistical experiments that evaluate the significance of the difference between the compared approaches.

Firstly, the results on the prostate zone segmentation task are provided. The test set on which these results are acquired is divided into two test sets – ProstateX and the Decathlon test sets. The results obtained on those test sets are reported separately. The results, acquired on the ProstateX source test set, are used to indicate the overall accuracies of the compared approaches. Meanwhile, the results calculated from the Decathlon source data set also indicate the generalizability between different datasets of the compared approaches. Furthermore, the results on both prostate zones, peripheral and transitional, are reported separately.

The results of the experiments on the prostate zone segmentation task include Friedman test results. These tests are performed twice – once to evaluate the significance of the difference between the results of the compared workflows, second time – the significance of the difference between the results of the deep neural network architectures, which are used to replace the conventional U-Net architecture in the no-new U-Net method. Lastly, the aggregated collected metrics of the experiment on the prostate zone segmentation task are provided.

Secondly the results on the PCa segmentation task are provided. There is only one test set on which these results are acquired. Once again, this chapter contains the results of Friedman tests performed on the collected metrics. These tests are performed twice – once on the DSC, second time on the DSC_τ. Furthermore, the aggregated collected metrics of the experiment on the PCa segmentation task are provided. Lastly, Wilcoxon signed-rank tests are performed to determine whether the observed difference between DSC and DSC_τ are statistically significant. These tests are performed for each compared ensemble as well as the baseline model independently from one another. These results were presented in the main papers by Vaitulevičius et al. [A.1], [A.2] [B.4].

4.1. Results of Experiment on the Prostate Zone Segmentation Task

This section provides the results of the experiments on the prostate zone segmentation task. Firstly, the results on the ProstateX test set is reported. Afterwards, the results on the Decathlon test set is presented.

4.1.1. Results on the ProstateX Test Set

The compared approaches for the prostate zone segmentation task consists of two components: workflow and deep neural network architecture, which is used as a substitute of the conventional U-Net architecture in the no-new U-Net method. Therefore, two sets of Friedman tests are performed. The first set of of tests is performed for each compared workflow separately. These tests determine the significance of difference between the results of compared DNN architectures. The degrees of freedom of each test is four as there are five compared DNN architectures. Table 4.1 contains the results of Friedman tests performed for each compared workflow separately where column:

Table 4.1. Friedman test results. These tests determine the presence of significant difference between compared deep neural network architectures. The test results are calculated on the ProstateX test set.

Zone	Workflow	W	Q	p-value
TZ	SM4TZPZS	0.080	3.20	0.524931
	PSTZSPZC	0.120	4.80	0.308441
	SPZTZSM	0.066	2.64	0.619754
PZ	SM4TZPZS	0.080	3.20	0.205203
	PSTZSPZC	0.120	4.80	0.564541
	SPZTZSM	0.066	2.64	0.864760

- Zone indicates on which zone was the test performed. It can either be transitional or peripheral zones.
- Workflow represents on which workflow was the test performed. It can either be SM4TZPZS, PSTZSPZC or SPZTZSM.
- W contains Kendall's coefficient of concordance corrected for ties.

- Q represents the Friedman chi-square statistic corrected for ties.
- p-value indicates the p-value obtained by performing a test.

Since none of the statistical tests produced a p-value below the selected significance level, the null hypothesis cannot be rejected. Consequently, no statistically significant differences were found between the deep neural network architectures being compared.

The second set of Friedman tests are performed for each compared DNN architecture independently from one another. These tests determine the significance of difference between the results of compared workflows. The degrees of freedom of each test is two as there are three compared DNN architectures. Table 4.2 contains the results of Friedman tests performed for each compared DNN architecture separately. The columns of this table are analogous to the columns of Table 4.1. The only difference is that instead of column workflow it contains column DNN architecture, which represents the DNN architecture which is used to substitute the original U-Net architecture in the nnU-Net method. The possible values of this column are the grid attention U-Net, single attention U-Net, multidimensional attention U-Net, dual-attention U-Net and U-Net.

Table 4.2. Friedman test results. These tests determine the presence of significant difference between compared workflows. The test results are calculated on the ProstateX test set.

Zone	DNN architecture	W	Q	p-value
TZ	Grid Attention U-Net	0.010	0.20	0.904837
	Attention U-Net	0.130	2.60	0.272532
	Multidimensional Attention U-Net	0.090	1.80	0.406570
	Dual-Attention U-Net	0.010	0.20	0.904837
	U-Net	0.130	2.60	0.272532
PZ	Grid Attention U-Net	0.250	5.00	0.082085
	Attention U-Net	0.090	1.80	0.406570
	Multidimensional Attention U-Net	0.120	2.40	0.301194
	Dual-Attention U-Net	0.030	0.60	0.740818
	U-Net	0.010	0.20	0.904837

Once again as none of those tests produced a p-value lower than the significance level, the null hypothesis is not rejected. Hence, no statistically significant difference is observed between compared workflows. However, the p-value obtained from the statistical test conducted on PZ segmentation results using the grid attention U-Net is an outlier, as it is substantially smaller than the remaining p-values. The post hoc Wilcoxon signed-rank test on these results yield a p-value of 0.0644531 for the SPZTZSM and SM4TZPZS pair, a p-value of 0.1601562 for the SPZTZSM and PSTZSPZC pair, and a p-value of 0.625 for the SM4TZPZS and PSTZSPZC pair. As expected, none of the comparisons indicate a statistically significant difference. However, the results suggest that the largest discrepancy is observed between SPZTZSM and PSTZSPZC.

Finally, aggregated test results on the ProstateX test set are displayed in Table 4.3 and Table 4.4. Table 4.3 contains the results obtained on the transitional zone, while Table 4.4 - on the peripheral zone. The columns of these tables are defined as follows:

- DNN architecture presents the deep neural network architecture which was used to substitute the original U-Net in the nnU-Net method.
- Workflow indicates the compared workflow.
- Mean +/- std contains the average DSC and std.
- Median +/- MAD contains the median DSC and median absolute deviation.

Just as statistical tests indicated no significant difference between compared approaches can be observed.

Figure 4.1 provides representative single slice examples of the prostate zone segmentation results. These segmentations were obtained using the nnU-Net method with a U-Net architecture and SM4TZPZS on the ProstateX test set. The left column shows PZ segmentation results (see Fig. 4.1a, Fig. 4.1c and Fig. 4.1e), and the right column shows TZ segmentation results (see Fig. 4.1b, Fig. 4.1d and Fig. 4.1f). Each row corresponds to the same slice from the same patient. The first case (see Fig. 4.1a and Fig. 4.1b) illustrates one of the most accurately segmented examples from the ProstateX test set, whereas the second (see

Table 4.3. Mean, standard deviation, median and MAD of transitional zone segmentation results calculated on the ProstateX test set.

DNN architecture	workflow	Mean +/- std	Median +/- MAD
Grid Attention U-Net	SM4TZPZS	0.9170+/-0.0379	0.9232+/-0.0435
	PSTZSPZC	0.9180+/-0.0323	0.9244+/-0.0320
	SPZTZSM	0.9175+/-0.0335	0.9201+/-0.0218
Attention U-Net	SM4TZPZS	0.9189+/-0.0372	0.928+/-0.04140
	PSTZSPZC	0.9177+/-0.0315	0.9217+/-0.0239
	SPZTZSM	0.9143+/-0.0394	0.9225+/-0.0410
Multidimensional Attention U-Net	SM4TZPZS	0.9175+/-0.0391	0.9277+/-0.0590
	PSTZSPZC	0.9184+/-0.0323	0.923+/-0.03010
	SPZTZSM	0.9171+/-0.0356	0.9258+/-0.0394
Dual-Attention U-Net	SM4TZPZS	0.9189+/-0.0335	0.9255+/-0.0438
	PSTZSPZC	0.9192+/-0.0308	0.9237+/-0.0270
	SPZTZSM	0.9186+/-0.0327	0.9248+/-0.0459
U-Net	SM4TZPZS	0.9181+/-0.0371	0.9256+/-0.0347
	PSTZSPZC	0.9192+/-0.0319	0.9248+/-0.0294
	SPZTZSM	0.9179+/-0.0361	0.9269+/-0.0364

Fig. 4.1c and Fig. 4.1d) and third (see Fig. 4.1e and Fig. 4.1f) cases illustrate typical segmentation errors produced by the workflow. Due to the large number of approaches considered in this study (15 in total) as well as a relatively large dataset (129 patients in total), conducting a detailed qualitative analysis of each approach is beyond the scope of this thesis. Furthermore, such an analysis would require expert evaluation of the segmentation results by medical professionals, making it a resource-intensive undertaking. Nevertheless, such an analysis represents a valuable avenue for future research.

4.1.2. Results on the Decathlon Test Set

The results on the Decathlon test set are reported in the same structure. The first set of Friedman tests is performed separately for each workflow being compared. Table 4.5 presents the results of the Friedman tests for each compared workflow independently from one another.

Comparative analyses of the DNN architectures performed using

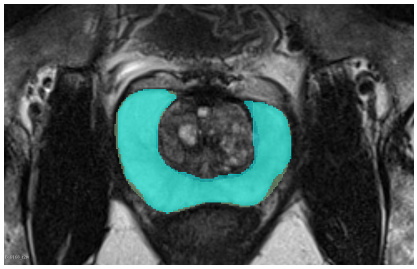
Table 4.4. Mean, standard deviation, median and MAD of peripheral zone segmentation results calculated on the ProstateX test set.

DNN architecture	workflow	Mean +/- std	Median +/- MAD
Grid Attention U-Net	SM4TZPZS	0.8199+/-0.0561	0.8188+/-0.0187
	PSTZSPZC	0.8173+/-0.0595	0.8143+/-0.0218
	SPZTZSM	0.8248+/-0.0529	0.8181+/-0.0217
Attention U-Net	SM4TZPZS	0.8194+/-0.0602	0.8225+/-0.0215
	PSTZSPZC	0.819+/-0.05500	0.8149+/-0.0340
	SPZTZSM	0.8245+/-0.0518	0.8147+/-0.0251
Multidimensional Attention U-Net	SM4TZPZS	0.8161+/-0.0587	0.8188+/-0.0053
	PSTZSPZC	0.8202+/-0.0592	0.8169+/-0.0200
	SPZTZSM	0.8218+/-0.0539	0.815+/-0.01790
Dual-Attention U-Net	SM4TZPZS	0.8156+/-0.0582	0.8177+/-0.0142
	PSTZSPZC	0.8185+/-0.0618	0.8215+/-0.0232
	SPZTZSM	0.8191+/-0.0573	0.8155+/-0.0104
U-Net	SM4TZPZS	0.8207+/-0.0601	0.8217+/-0.0262
	PSTZSPZC	0.821+/-0.05760	0.8238+/-0.0119
	SPZTZSM	0.8272+/-0.0500	0.8134+/-0.0264

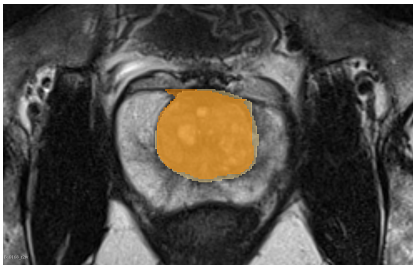
Table 4.5. Friedman test results. These tests determine the presence of significant difference between the compared deep neural network architectures. The test results are calculated on the Decathlon test set.

Zone	Workflow	W	Q	p-value
TZ	SM4TZPZS	0.08	9.95	0.041336
	PSTZSPZC	0.059	7.31	0.120543
	SPZTZSM	0.14	17.36	0.001645
PZ	SM4TZPZS	0.076	9.47	0.050347
	PSTZSPZC	0.083	10.25	0.036494
	SPZTZSM	0.195	24.15	0.000074

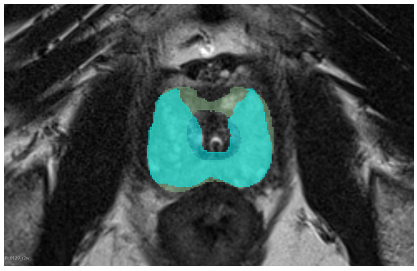
PSTZSPZC workflow for the TZ segmentation task and SM4TZPZS workflow for the PZ segmentation demonstrated no statistically significant differences, as the Friedman tests yielded p-values exceeding the significance level. However, statistically significant differences between the compared DNN architectures are observed when analysing



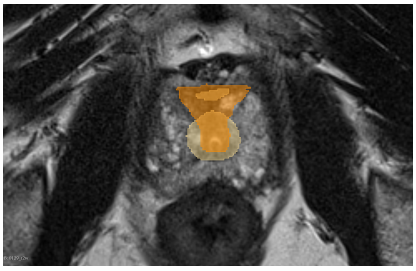
(a) Peripheral zone segmentation.



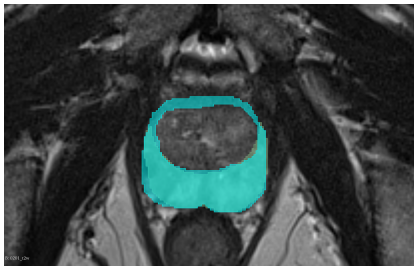
(b) Transitional zone segmentation.



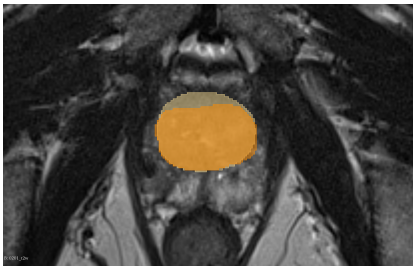
(c) Peripheral zone segmentation.



(d) Transitional zone segmentation.



(e) Peripheral zone segmentation.



(f) Transitional zone segmentation.

Figure 4.1. Single slice examples of prostate zone segmentations on the ProstateX test set. The segmentations are performed with Separate Models for the Transitional Zone and Peripheral Zone Segmentation. The deep neural network architecture used for these segmentations is U-Net.

TZ segmentation performance obtained using SM4TZPZS and SPZTZSM workflows, as well as PZ segmentation performance derived from PSTZSPZC and SPZTZSM workflows, with Friedman tests yielding p-values below the significance level. Thus, the Conover post hoc test for unreplicated block design is performed in this thesis on the De-

cathlon test set TZ segmentation results obtained using SM4TZPZS and SPZTZSM workflows. Moreover, this test is also performed in this thesis on the Decathlon test set PZ segmentation results obtained using PSTZSPZC and SPZTZSM workflows. The outcomes of both tests are reported in Table 4.6 and Table 4.7.

Table 4.6. Conover post hoc test for unreplicated block design results. These tests indicate which deep neural network architectures are responsible for the observed statistical difference in the transitional zone segmentaion results. The segmentations are performed on the Decathlon test set.

Comparison	SPZTZSM	SM4TZPZS
U-Net - Multidimensional Attention U-Net	0.6657436	0.1623296
U-Net - Grid Attention U-Net	0.2845869	0.8020478
U-Net - Dual-Attention U-Net	0.1707531	0.7512701
U-Net - Attention U-Net	**0.0351308**	0.1046590
Multidimensional Attention U-Net - Grid Attention U-Net	0.4857866	0.2238420
Multidimensional Attention U-Net - Dual-Attention U-Net	0.0795479	0.1046590
Multidimensional Attention U-Net - Attention U-Net	0.0795479	0.7512701
Grid Attention U-Net - Dual-Attention U-Net	**0.0251200**	0.7201986
Grid Attention U-Net - Attention U-Net	0.2405340	0.1410870
Dual-Attention U-Net - Attention U-Net	**0.0004348**	0.1046590

The post hoc analysis of the TZ segmentation results obtained using SM4TZPZS does not reveal any statistically significant differences between the compared DNN architectures, as the Friedman test produced p-values exceeding the adopted significance level. This result is not unexpected, given that the corresponding Friedman test yielded a p-value close to the significance threshold. Furthermore, the post hoc analysis of the TZ segmentation results obtained using SPZTZSM in-

dicates statistically significant differences between the following pairs of these DNN architectures: U-Net and attention U-Net, grid attention U-Net and dual-attention U-Net, as well as attention U-Net and dual-attention U-Net. Although the results of these tests do not indicate that any of the DNN architectures clearly stand out or form distinct clusters, the results corresponding to the identified architecture pairs are subjected to further analysis.

Table 4.7. Conover post hoc test for unreplicated block design results. These tests indicate which deep neural network architectures are responsible for the observed statistical difference in the peripheral zone segmentaion results. The segmentations are performed on the Decathlon test set.

Comparison	SPZTZSM	PSTZSPZC
U-Net - Multidimensional Attention U-Net	**0.0182317**	**0.0216176**
U-Net - Grid Attention U-Net	0.4225707	0.1130058
U-Net - Dual-Attention U-Net	**0.0000557**	0.2392860
U-Net - Attention U-Net	0.4225707	0.1377493
Multidimensional Attention U-Net - Grid Attention U-Net	0.0955599	0.5740625
Multidimensional Attention U-Net - Dual-Attention U-Net	0.0901108	0.2392860
Multidimensional Attention U-Net - Attention U-Net	0.0955599	0.4761054
Grid Attention U-Net - Dual-Attention U-Net	**0.0005821**	0.5740625
Grid Attention U-Net - Attention U-Net	1.0000000	0.8049505
Dual-Attention U-Net - Attention U-Net	**0.0005821**	0.6905740

The post hoc analysis of the PZ segmentation results obtained using PSTZSPZC revealed a statistically significant difference only between U-Net and multidimensional attention U-Net. Moreover, the post hoc analysis of the PZ segmentation results obtained using SPZTZSM indicates statistically significant differences between the following pairs

of DNN architectures: U-Net and multidimensional attention U-Net; U-Net and dual-attention U-Net; grid attention U-Net and dual-attention U-Net; as well as attention U-Net and dual-attention U-Net. The results of these tests indicate that the performance of dual-attention U-Net on the PZ segmentation task differs significantly from that of the other DNN architectures when using SPZTZSM. However, the results of the post hoc analysis do not indicate that any of the remaining DNN architectures clearly stand out or form distinct clusters. Nevertheless, the results corresponding to the identified architecture pairs are also subjected to further analysis.

The second set of Friedman tests was conducted for each compared DNN architecture independently. Table 4.8 presents the results of the Friedman tests performed for each DNN architecture. None of the p-values are lower than significance level when segmenting the TZ, except when testing with the results obtained with DNN architecture, attention U-Net. Considering, that no statistically significant difference is observed between compared DNN architectures, a single positive result among five tests can be ignored. Thus, the results of these tests indicate no statistically significant difference between compared workflows when segmenting the transitional zone. However, there are a lot of p-values lower than selected significance level when segmenting the peripheral zone. Hence, there is a statistically significant difference between compared workflows when segmenting the PZ.

Thus, we perform post hoc Wilcoxon signed-rank tests on the Decathlon test set PZ segmentation results. These tests determine which workflows cause the aforementioned significance. Table 4.9 displays the results of post hoc Wilcoxon signed-rank tests where column:

- DNN architecture describes the deep neural network architecture used to replace the original U-Net in the nnU-Net method.
- Comparison indicates which workflows are being compared in the test.
- p-value indicates the p-value obtained by performing the post hoc Wilcoxon signed-rank tests.

The results of these tests indicate that there is a statistically significant difference between PSTZSPZC and the other workflows as all cor-

Table 4.8. Friedman test results. These tests determine the presence of significant difference between the compared workflows. The test results are calculated on the Decathlon test set.

Zone	DNN architecture	W	Q	p-value
TZ	Grid Attention U-Net	0.087	5.40	0.0672055
	Attention U-Net	0.137	8.47	0.0145040
	Multidimensional Attention U-Net	0.046	2.87	0.2385126
	Dual-Attention U-Net	0.003	0.20	0.9048374
	U-Net	0.042	2.60	0.2725318
PZ	Grid Attention U-Net	0.257	15.94	0.0003465
	Attention U-Net	0.426	26.39	0.0000019
	Multidimensional Attention U-Net	0.482	29.87	0.0000003
	Dual-Attention U-Net	0.316	19.61	0.0000551
	U-Net	0.326	20.19	0.0000412

responding p-values are below the significance level. Thus, the null hypothesis is rejected. Meanwhile, the post hoc Wilcoxon signed-rank test results indicate conflicting conclusions when comparing SPZTZSM and SM4TZPZS. On one hand, the results indicate a significant difference between compared workflows when conducting tests obtained with these DNN architectures – dual-attention U-Net and U-Net – as all of the corresponding p-values are below the significance level. Hence, the null hypothesis is rejected. On the other hand, the post hoc Wilcoxon signed-rank test results, obtained with DNN architectures – grid attention U-Net, attention U-Net and multidimensional attention U-Net – bear no p-values below significance level. Thus, the statistically significant difference between workflows SM4TZPZS and baseline workflow, SPZTZSM is undetermined.

Furthermore, the aggregated test results on the Decathlon test set are provided in Table 4.10 and Table 4.11. Table 4.10 contains the results obtained on the TZ, while the PZ is in Table 4.11. The tables are analogous to Table 4.3 and Table 4.4.

The results presented in Table 4.10 indicate that attention U-Net achieved higher mean and median DSC values than multidimensional

Table 4.9. Post hoc Wilcoxon signed-rank test results. These tests indicate which workflows are responsible for the observed statistical difference in the peripheral zone segmentation results.

DNN architecture	Comparison	p-value
Grid Attention U-Net	SPZTZSM - SM4TZPZS	0.3883800
	SPZTZSM - PSTZSPZC	0.0007822
	SM4TZPZS - PSTZSPZC	0.0030005
Attention U-Net	SPZTZSM - SM4TZPZS	0.0694814
	SPZTZSM - PSTZSPZC	0.0001776
	SM4TZPZS - PSTZSPZC	0.0000012
Multidimensional Attention U-Net	SPZTZSM - SM4TZPZS	0.0062173
	SPZTZSM - PSTZSPZC	0.0025741
	SM4TZPZS - PSTZSPZC	0.0000018
Dual-Attention U-Net	SPZTZSM - SM4TZPZS	0.0005177
	SPZTZSM - PSTZSPZC	0.0374298
	SM4TZPZS - PSTZSPZC	0.0004353
U-Net	SPZTZSM - SM4TZPZS	0.0304844
	SPZTZSM - PSTZSPZC	0.0000940
	SM4TZPZS - PSTZSPZC	0.0000590

attention U-Net when using SPZTZSM. Considering that the Conover post hoc test revealed a statistically significant difference between these DNN architectures, this result can be regarded as promising. Meanwhile, dual-attention U-Net achieved lower mean and median DSC values than both attention U-Net and grid attention U-Net. The Conover post hoc test indicates that the differences between dual-attention U-Net and these two DNN architectures are statistically significant. These results suggest the need for a larger training dataset for dual-attention U-Net, as it contains a higher number of parameters compared to the other DNN architectures. The results displayed in Table 4.10 indicate no other significant difference between compared approaches. Thus, these results confirm the observation made from Friedman test results obtained on the Decathlon test set transitional zone segmentation results.

However, the results provided in Table 4.10 indicate a trend of SM4TZPZS outperforming other workflows. SM4TZPZS achieves the

Table 4.10. Mean, standard deviation, median and MAD of transitional zone segmentation results calculated on the Decathlon test set.

DNN architecture	workflow	Mean +/- std	Median +/- MAD
Grid Attention U-Net	SM4TZPZS	**0.8142+/-0.2171**	0.8822+/-0.0070
	PSTZSPZC	0.7794+/-0.2493	0.8802+/-0.0027
	SPZTZSM	0.7784+/-0.2563	**0.8825+/-0.0024**
Attention U-Net	SM4TZPZS	**0.8061+/-0.2329**	**0.8874+/-0.0000**
	PSTZSPZC	0.7709+/-0.2684	0.8832+/-0.0030
	SPZTZSM	0.8032+/-0.2149	0.8825+/-0.0000
Multidimensional Attention U-Net	SM4TZPZS	**0.8117+/-0.2206**	0.8796+/-0.0103
	PSTZSPZC	0.7943+/-0.2473	**0.8834+/-0.0072**
	SPZTZSM	0.7593+/-0.2698	0.8786+/-0.0121
Dual-Attention U-Net	SM4TZPZS	0.7173+/-0.3025	**0.8838+/-0.0006**
	PSTZSPZC	**0.7442+/-0.2781**	0.8804+/-0.0000
	SPZTZSM	0.7307+/-0.2760	0.8576+/-0.0000
U-Net	SM4TZPZS	**0.7832+/-0.2502**	**0.8858+/-0.0012**
	PSTZSPZC	0.7817+/-0.2513	0.8818+/-0.0012
	SPZTZSM	0.7728+/-0.2580	0.8792+/-0.0000

highest average DSC on the results obtained with all compared deep neural network architectures except for dual-attention U-Net. The highest mean DSC is achieved by PSTZSPZC workflow on the results obtained with dual-attention U-Net. Moreover, SM4TZPZS achieves the highest median DSC on the results obtained with all compared DNN architectures except for grid attention U-Net and multidimensional attention U-Net. The highest median DSC is achieved by SPZTZSM on the results obtained with grid attention U-Net. Meanwhile, the highest median DSC is achieved by PSTZSPZC workflow on the results obtained with multidimensional attention U-Net. Thus, for TZ segmentation, the preferred workflow is SM4TZPZS. Lastly, the results provided in Table 4.10 show that the highest mean DSC is achieved by employing grid attention U-Net and SM4TZPZS. Meanwhile, these results indicate that the highest median DSC is achieved by employing U-Net and SM4TZPZS.

The results presented in Table 4.11 indicate that multidimensional attention U-Net achieved higher mean and median DSC values than U-

Table 4.11. Mean, standard deviation, median and MAD of peripheral zone segmentation results calculated on the Decathlon test set.

DNN architecture	workflow	Mean +/- std	Median +/- MAD
Grid Attention U-Net	SM4TZPZS	**0.6794+/-0.1050**	**0.6975+/-0.1101**
	PSTZSPZC	0.5962+/-0.2169	0.6413+/-0.1010
	SPZTZSM	0.6735+/-0.1113	0.6778+/-0.1317
Attention U-Net	SM4TZPZS	**0.6909+/-0.1129**	**0.7088+/-0.1428**
	PSTZSPZC	0.5775+/-0.2438	0.6443+/-0.1181
	SPZTZSM	0.6751+/-0.1056	0.6898+/-0.1480
Multidimensional Attention U-Net	SM4TZPZS	**0.6920+/-0.1117**	**0.7098+/-0.1277**
	PSTZSPZC	0.5725+/-0.2452	0.6366+/-0.1023
	SPZTZSM	0.6281+/-0.1510	0.6503+/-0.1104
Dual-Attention U-Net	SM4TZPZS	**0.6806+/-0.1265**	**0.7205+/-0.1261**
	PSTZSPZC	0.5221+/-0.2635	0.6322+/-0.0973
	SPZTZSM	0.6395+/-0.1323	0.6732+/-0.1134
U-Net	SM4TZPZS	**0.6978+/-0.1075**	**0.7249+/-0.1766**
	PSTZSPZC	0.5376+/-0.2643	0.6333+/-0.0994
	SPZTZSM	0.6684+/-0.1165	0.6692+/-0.1208

Net when using PSTZSPZC. However, when using SPZTZSM, U-Net achieved higher mean and median DSC values than multidimensional attention U-Net. Therefore, although post hoc statistical tests indicated a statistically significant difference between these two DNN architectures, the results are inconsistent across workflows, and no DNN architecture can be considered consistently more accurate than the other. Meanwhile, dual-attention U-Net achieved lower mean and median DSC values than U-Net, Attention U-Net, and grid attention U-Net. The Conover post hoc test indicates that the differences between dual-attention U-Net and these DNN architectures are statistically significant. This observation further supports the need for a larger training dataset for dual-attention U-Net.

The results displayed in Table 4.11 indicate that the highest DSC on peripheral zone segmentation task is obtained by employing SM4TZPZS. Thus, for PZ segmentation, the preferred workflow is SM4TZPZS. Therefore, the results obtained with SM4TZPZS are further inspected.

Table 4.12 contains the Decathlon test set PZ segmentation results

obtained strictly with SM4TZPZS. The columns of this table are analogous to the columns of Table 4.3, Table 4.4, Table 4.10 and Table 4.11. These results indicate no significant difference between compared DNN architectures, thereby confirming the observations made from the post hoc statistical tests. However, the highest DSC is achieved by U-Net. Thus, for PZ segmentation, the preferred deep neural network architecture is U-Net.

Table 4.12. Mean, standard deviation, median and MAD of the Decathlon test set peripheral zone segmentation results obtained with Separate Models for the Transitional Zone and Peripheral Zone Segmentation.

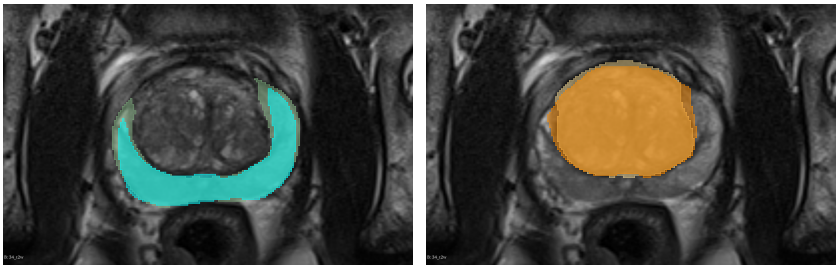
DNN architecture	Mean +/- std	Median +/- MAD
Grid Attention U-Net	0.6794+/-0.1050	0.6975+/-0.1101
Attention U-Net	0.6909+/-0.1129	0.7088+/-0.1428
Multidimensional Attention U-Net	0.692+/-0.11170	0.7098+/-0.1277
Dual-Attention U-Net	0.6806+/-0.1265	0.7205+/-0.1261
U-Net	**0.6978+/-0.1075**	**0.7249+/-0.1766**

Figure 4.2 contains representative single slice examples of prostate zone segmentation results. These segmentations were obtained using the nnU-Net method with a U-Net architecture and SM4TZPZS on the Decathlon test set. The structure of this figure is analogous to Figure 4.1. The first case illustrates one of the most accurately segmented examples from the Decathlon test set, whereas the second and third cases illustrate typical segmentation errors produced by the workflow.

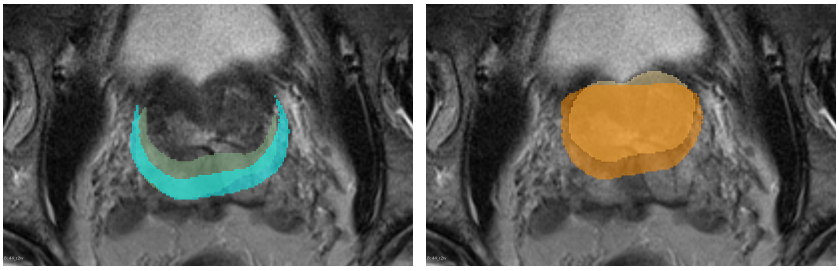
4.2. Results of Experiment on the Prostate Cancer Segmentation Task

This section contains the results of the experiments on the prostate cancer segmentation task. Firstly, aggregated collected DSC and DSC_τ are reported in Table 4.13 and Table 4.14. Table 4.13 contains mean and std, while Table 4.14 provides median and MAD. The columns of these tables are defined as follows:

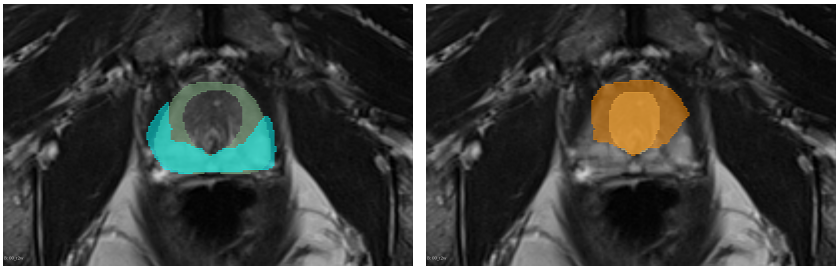
- Ensemble/Model represents the proposed ensemble and the baseline model.



(a) Peripheral zone segmentation. (b) Transitional zone segmentation.



(c) Peripheral zone segmentation. (d) Transitional zone segmentation.



(e) Peripheral zone segmentation. (f) Transitional zone segmentation.

Figure 4.2. Single slice examples of prostate zone segmentations on the Decathlon dataset. The segmentations are performed with Separate Models for the Transitional Zone and Peripheral Zone Segmentation. The deep neural network architecture used for these segmentations is U-Net.

- DSC represents the aggregated DSC results.
- DSC_τ indicates the aggregated DSC_τ results.

These results indicate that Ensemble 3 achieves the highest segmentation accuracy, followed by Ensemble 4. Moreover, the aggregated

Table 4.13. Mean and std of prostate cancer segmentation results. These values are calculated based on the collected metrics DSC_τ and DSC.

Ensemble/Model	DSC	DSC_τ
Ensemble 1	0.1613 +/- 0.2271	0.2116 +/- 0.3015
Ensemble 2	0.1828 +/- 0.2381	0.2512 +/- 0.3269
Ensemble 3	**0.2264 +/- 0.2611**	**0.2977 +/- 0.3443**
Ensemble 4	0.2155 +/- 0.2580	0.2829 +/- 0.3395
Baseline	0.1787 +/- 0.2387	0.2360 +/- 0.3152

Table 4.14. Median and MAD of prostate cancer segmentation results. These values are calculated based on the collected metrics DSC_τ and DSC.

Ensemble/Model	DSC	DSC_τ
Ensemble 1	0.0000 +/ 0.0000	0.0000 +/ 0.0000
Ensemble 2	0.0165 +/ 0.0165	0.0174 +/ 0.0174
Ensemble 3	**0.0608 +/ 0.0608**	**0.0683 +/ 0.0683**
Ensemble 4	0.0521 +/ 0.0521	0.0599 +/ 0.0599
Baseline	0.0000 +/ 0.0000	0.0000 +/ 0.0000

results indicate that the novel metric, DSC_τ, yields higher scores than the conventional DSC, indicating that the proposed metric accounts for less severe penalties on boundary differences within the allowable tolerance.

Furthermore, the low MAD values presented in Table 4.14 indicate that the evaluated ensembles produce highly consistent results for the majority of cases. In contrast, the relatively high std values reported in Table 4.13 suggest the presence of a small number of atypical cases that substantially increase the overall variability of the results. Given that the mean values of the evaluation metrics exceed their corresponding median values, it can be inferred that these outlying cases are associated with considerably higher segmentation accuracy. Consequently, although the ensembles achieve favourable performance for a limited subset of cases, the median-based evaluation indicates that segmentation performance remains modest for the majority of the dataset. Considering the limited size of the available dataset, such behaviour is

not unexpected. While the conducted experiment enables the identification of the most promising prostate cancer segmentation approach among the evaluated alternatives, a more comprehensive assessment is required to establish the robustness and generalisability of the obtained results Therefore, future research should focus on training and evaluating the proposed solutions using substantially larger and more diverse datasets.

Furthermore, both Friedman tests produced p-values lower than the significance level. The test on DSC_τ resulted in the p-value equal to 0.003107. The Kendall's coefficient of concordance for this test, adjusted for ties, is W = 0.083. The corresponding Friedman chi-square statistic, also adjusted for ties, is Q = 15.94. Meanwhile, the Friedman test on DSC resulted in the p-value equal to 0.00067. The W of this test is 0.101, while Q is 19.35. Thus, the null hypothesis is rejected, which indicates the statistically significant difference between the results of the ensembles and the baseline model.

Therefore, the Conover post hoc tests for unreplicated block design is performed. As both metrics indicated a statistically significant difference between ensembles, these tests are performed twice. Once on the collected DSC metrics, and the other time on the collected DSC_τ. The results of Conover post hoc tests for unreplicated block design are displayed in Table 4.15. The columns of this table are defined as follows:

- Comparison indicates the proposed ensemble/baseline model being compared.
- p-value(DSC) represents the p-value of the test on collected DSC metrics.
- p-value(DSC_τ) represents the p-value of the test on collected DSC_τ metrics.

Since the p-values are below the significance level, the null hypothesis is rejected, indicating a statistically significant difference between the results of the baseline model and Ensembles 3 and 4. In contrast, no statistically significant difference is observed between the baseline model and Ensembles 1 and 2, as the corresponding p-values exceed the significance level Therefore, the null hypothesis is not rejected. The remaining results from the Conover post hoc test show no statistically

Table 4.15. Conover post hoc test results on the prostate cancer segmentation task. The tests were conducted on the collected metrics DSC_τ and DSC. The metric corresponding to each reported p-value is indicated in parentheses.

Comparison	p-value (DSC)	p-value (DSC_τ)
Ensemble 1 - Ensemble 2	0.7059	0.5352
Ensemble 1 - Ensemble 3	**0.0079**	**0.0189**
Ensemble 1 - Ensemble 4	**0.0179**	0.0656
Ensemble 1 - Baseline	0.7059	0.5758
Ensemble 2 - Ensemble 3	**0.0148**	0.0741
Ensemble 2 - Ensemble 4	**0.0462**	0.2490
Ensemble 2 - Baseline	0.6042	0.2533
Ensemble 3 - Ensemble 4	0.7059	0.5710
Ensemble 3 - Baseline	**0.0043**	**0.0069**
Ensemble 4 - Baseline	**0.0096**	**0.0189**

significant difference between Ensembles 3 and 4, nor between Ensembles 1 and 2. However, the collected DSC metrics for Ensembles 1 and 2 differ statistically significantly from those of Ensembles 3 and 4, while the DSC_τ metric does not.

Moreover, Table 4.16 presents the false positive patients. The False positive patients column is the percentage of patients for whom the model's prediction includes a cancerous region while the ground truth does not. As in Table 4.13, the column Ensemble/Model lists the compared ensembles and the baseline model. The column False positive patients shows the calculated value of this metric. This metric indicates that the baseline model produces the fewest erroneous prostate cancer segmentations.

Wilcoxon signed-rank test results are provided in Table 4.17. These tests inspect the presence of the statistical significance of the difference between metrics DSC and DSC_τ. Just as in Table 4.13, the Ensemble/Model column represents the compared ensembles together with the baseline model. Meanwhile the column statistic and p-value are the results of the Wilcoxon signed-rank test. The column statistic contains the sum of signed ranks of paired differences. The results indicate that all DSC_τ are greater than DSC as reflected by statistics presented in Ta-

Table 4.16. False positive patients calculated on the results of the experiment on the prostate cancer segmentation task.

Ensemble/Model	False positive patients
Ensemble 1	24.68
Ensemble 2	28.57
Ensemble 3	35.06
Ensemble 4	27.27
Baseline	**19.48**

bles 4.13 and 4.14 as well as Table 4.17, where all corresponding statistic values are equal to 0. Considering that all of the acquired p-values are lower than the significance level, the null hypothesis is rejected. Thus, there is a significant difference between the compared metrics.

Table 4.17. Results of Wilcoxon signed-rank test performed on the prostate cancer segmentation results. These tests are conducted to determine the statistical significance of difference between the DSC metric and the novel metric, DSC_τ.

Ensemble / Model	statistic	p-value
Ensemble 1	0.0	0.000132
Ensemble 2	0.0	0.000027
Ensemble 3	0.0	0.000012
Ensemble 4	0.0	0.000027
Baseline	0.0	0.000132

Table 4.18 summarizes the distribution of predicted lesion counts per patient for each proposed ensemble and for the baseline model. The Ensemble/Model column enumerates all compared ensembles together with the baseline. Within the Number of predicted lesions group, each column specifies the number of patients for whom a given ensemble or the baseline model predicted the corresponding lesion count.

Overall, the predicted distributions approximate the ground truth lesion distribution. The dataset includes 77 patients with no lesions, 36 with a single lesion, and 6 with two lesions. Ensemble 4 most closely reproduces this distribution: it predicts zero lesions for exactly 77 patients, underestimates the number of patients with one lesion by four,

Table 4.18. Distribution of predicted lesion counts across patients. Each distribution represents the number of patients for whom a given ensemble or the baseline model predicted the corresponding lesion count.

Ensemble / Model	Number of predicted lesions					
	0	1	2	3	4	6
Ensemble 1	78	29	9	1	1	1
Ensemble 2	71	31	15	2	0	0
Ensemble 3	67	32	17	3	0	0
Ensemble 4	77	32	8	1	1	0
Baseline	90	24	4	0	1	0

and overestimates the number of patients with two lesions by two.

The lesion count distribution produced by Ensemble 1 shows slightly greater deviation from the ground truth. It predicts zero lesions for one more patient, one lesion for seven fewer cases, and two lesions for three additional patients compared with the true distribution. In contrast, Ensembles 2 and 3 demonstrate more pronounced discrepancies. Ensemble 2 predicts zero lesions for six fewer patients and one lesion for five fewer patients, while overestimating the two lesion category by nine patients. Similarly, Ensemble 3 predicts zero lesions for ten fewer patients and one lesion for four fewer patients, while overestimating the two lesion category by eleven patients.

The baseline model exhibits the largest deviation from the true lesion distribution. Specifically, it predicts zero lesions for 13 additional patients, underestimates the one lesion category by 12 patients, and underestimates the two lesion category by 2 patients. These findings suggest that the baseline model has a tendency toward lesion under detection, whereas the ensemble approaches generally yield predictions that more closely align with the true lesion counts.

Furthermore, although the training dataset contains at most two lesions per patient, the ensemble models occasionally predict more than two lesions. This overestimation is limited in scope, affecting only two to three patients per ensemble. The baseline model demonstrates even fewer such instances, predicting four lesions in only a single patient.

Table 4.19 presents the distribution of missed (false negative) lesions per patient for each proposed ensemble and for the baseline model.

The Ensemble/Model column lists all compared ensembles together with the baseline. Within the Number of missed lesions group, the first column for each lesion count category indicates the number of patients for whom the corresponding ensemble or the baseline model failed to detect lesions overlapping with that specific number of true lesions. The second column reports the corresponding percentage of such patients. These percentages are based on the 42 patients in the dataset who have at least one lesion.

Table 4.19. Distribution of missed lesions across the patients (42 cases with at least one lesion). Each distribution shows the number and percentage of patients for whom the ensemble or the baseline model failed to detect the corresponding number of lesions.

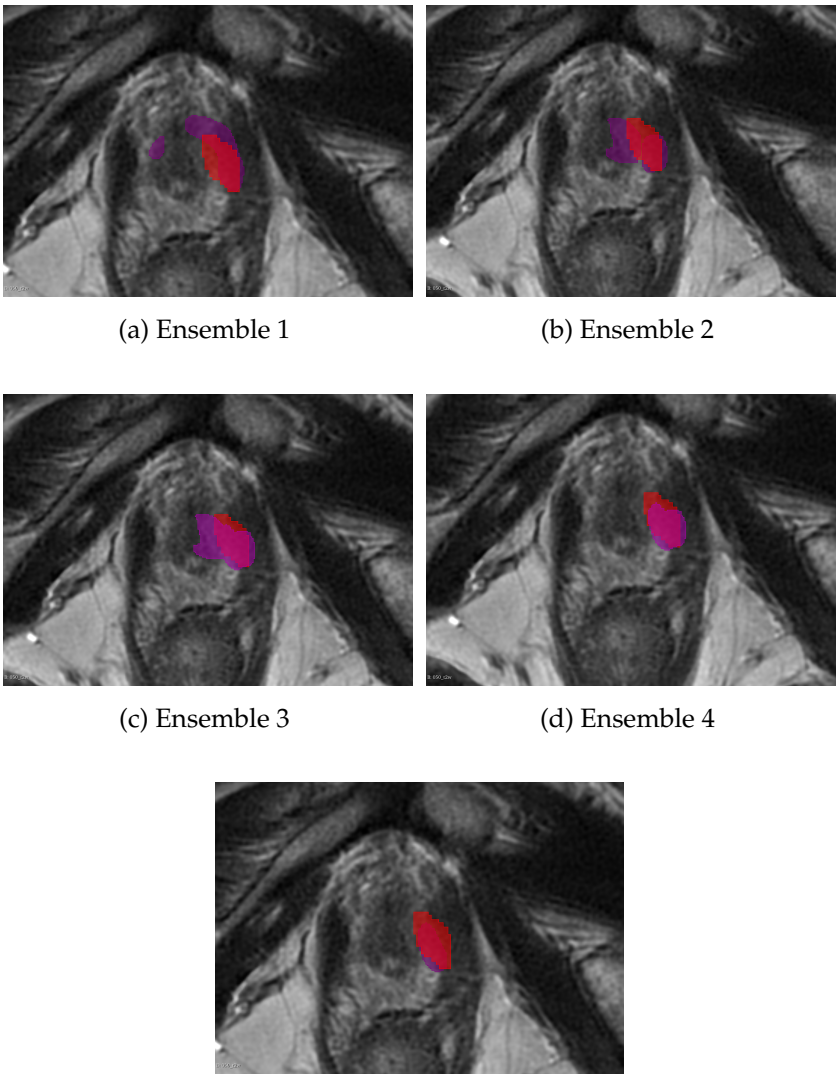
Ensemble / Model	Number of missed lesions					
	0		1		2	
Ensemble 1	16	38.10%	24	57.14%	2	4.76%
Ensemble 2	19	45.24%	21	50.00%	2	4.76%
Ensemble 3	**22**	**52.38%**	**19**	**45.24%**	**1**	**2.38%**
Ensemble 4	20	47.62%	21	50.00%	**1**	**2.38%**
Baseline	16	38.10%	25	59.52%	**1**	**2.38%**

The results summarized in Table 4.19 demonstrate that Ensemble 3 achieves the highest performance in terms of lesion localization, as reflected by the lowest proportion of missed lesions. Ensemble 2 and Ensemble 4 exhibit slightly lower, yet comparable, localization accuracy. In contrast, Ensemble 1 and the baseline model show the weakest performance in lesion localization. This observation is consistent with the comparative results reported in Table 4.13.

Figure 4.3 illustrates representative single slice segmentation results for each proposed ensemble alongside the baseline model, highlighting areas where the predicted lesion masks overlap with the ground truth. All examples are taken from the same patient and correspond to the same anatomical slice. Predicted lesions from the ensembles or the baseline model are shown in purple, while the ground truth is shown in red. The segmentation results and ground truth are overlaid on the T2W image.

Additionally, Figure 4.4 presents single slice examples of cases in

which all proposed ensembles and the baseline model failed to detect the lesion. In these examples, the ground truth is shown in red and overlaid on the T2W image.



(a) Ensemble 1

(b) Ensemble 2

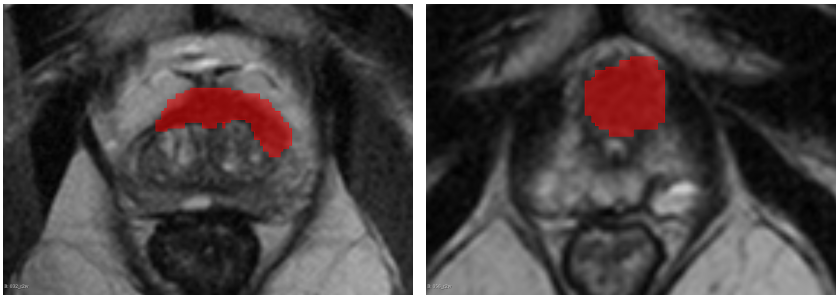
(c) Ensemble 3

(d) Ensemble 4

(e) Baseline model

Figure 4.3. Single slice segmentation examples from a single patient for each ensemble and the baseline model, showing regions where predicted lesions overlap with the ground truth. Segmentations are overlaid on the T2W image.

Two instances were observed in which the deep neural network



(a) (b)

Figure 4.4. Single slice examples of patients in whom none of the ensembles or the baseline model successfully localized the lesions. The ground truth is overlaid on the T2W image.

models produced fragmented lesion predictions. In the first case, both the $PCa_{seg}(All \times PZ)$ model and the baseline $Pca_{seg}((All, Zones) \times prostate)$ model segmented a single lesion into two separate components. Consequently, Ensemble 2 and the baseline model also generated a fragmented representation of this lesion. In contrast, Ensemble 3 and Ensemble 4 produced a single, continuous lesion, as the combination of their constituent models prevented fragmentation.

The second instance of fragmentation was caused solely by the $Pca_{seg}((All, Zones) \times TZ)$ model, resulting in a fragmented lesion in the outputs of Ensemble 2 and Ensemble 3. Notably, this lesion was not detected by Ensemble 1, Ensemble 4, and the baseline model. No further cases of lesion fragmentation were observed, indicating that these two examples represent outliers and that the proposed ensembles generally maintain lesion integrity without fragmentation.

Table 4.20 and Table 4.21 summarize the distributions of spurious foci (false positive) generated across patients by the compared ensembles and the baseline model. The Ensemble/Model column lists all proposed ensembles along with the baseline model. In Table 4.20, each column within the Number of spurious foci group reports the number of patients for whom a given ensemble or the baseline model produced that specific number of spurious lesions, i.e., lesions that do not overlap with any ground truth lesion. Meanwhile, in Table 4.21, each corresponding column reports the percentage of patients with that number

of spurious foci. These percentages are calculated relative to the full dataset of 119 patients. The results indicate that the baseline model generates the fewest spurious foci, consistent with the findings reported in Table 4.16.

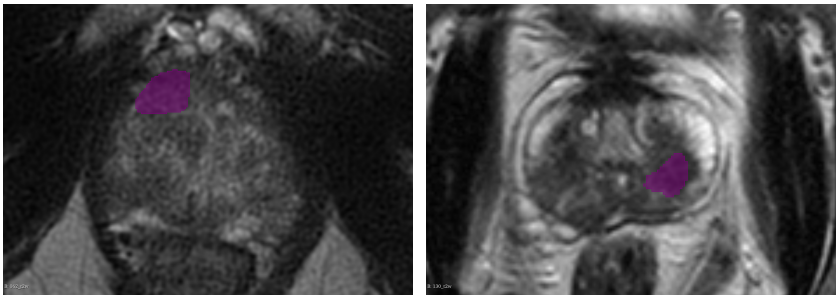
Table 4.20. Distribution of spurious foci across patients. Each distribution shows the number of patients for whom a given ensemble or the baseline model produced the corresponding number of spurious lesions.

Ensemble / Model	Number of spurious foci					
	0	1	2	3	4	6
Ensemble 1	94	17	5	1	1	1
Ensemble 2	86	24	9	**0**	**0**	**0**
Ensemble 3	83	26	9	1	**0**	**0**
Ensemble 4	95	20	3	**0**	1	**0**
Baseline	**108**	**10**	**0**	**0**	1	**0**

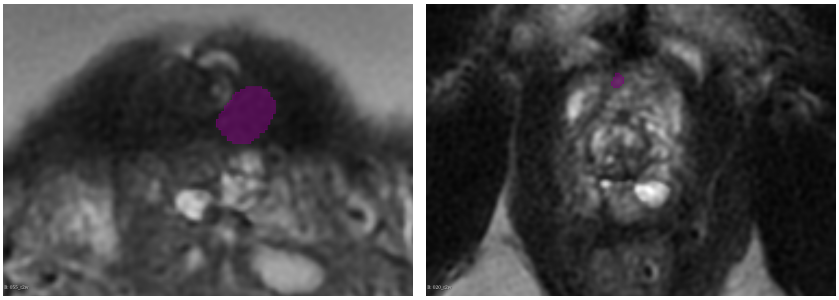
Table 4.21. Distribution of spurious foci across all 119 patients. Each distribution shows the number and percentage of patients for whom a given ensemble or the baseline model produced the corresponding number of spurious lesions.

Ensemble / Model	Number of spurious foci					
	0	1	2	3	4	6
Ensemble 1	78.99	14.29	4.20	0.84	0.84	0.84
Ensemble 2	72.27	20.17	7.56	**0.00**	**0.00**	**0.00**
Ensemble 3	69.75	21.85	7.56	0.84	**0.00**	**0.00**
Ensemble 4	79.83	16.81	2.52	**0.00**	0.84	**0.00**
Baseline	**90.76**	**08.40**	**0.00**	**0.00**	0.84	**0.00**

Figure 4.5 presents single slice examples of spurious foci generated by each proposed ensemble. Each example corresponds to a different patient in which the baseline model did not produce any spurious foci. The predicted spurious foci are shown in purple and overlaid on the T2W images.



(a) Ensemble 1 (b) Ensemble 2



(c) Ensemble 3 (d) Ensemble 4

Figure 4.5. Single slice examples of spurious foci generated by each ensemble. The predicted spurious lesions are overlaid on the T2W image.

4.3. Summary of the Chapter

This chapter provides the results of the experiments conducted on prostate zone segmentation and prostate cancer segmentation tasks. The research consists of developing the solution to these tasks, the no-new U-Net method.

The compared approaches to the first task, the prostate zone segmentation task, are evaluated on two test sets – ProstateX and Decathlon test sets. The results acquired on the ProstateX are used to determine the overall accuracy of the compared approaches. Meanwhile, the results acquired on the Decathlon test set are employed to determine the generalization of the approaches between different datasets.

This study found that there is no significant difference between the prostate zone segmentation results of the compared deep neural network architectures, which are used to substitute the original U-Net

architecture within the nnU-Net method. These architectures are U-Net variants, that employ attention mechanisms. The compared DNN architectures consists of grid attention U-Net, single attention U-Net, multidimensional U-Net and dual-attention U-Net and baseline, U-Net.

Furthermore, in this study several prostate zone segmentation workflows, which employs the nnU-Net method, are compared. These workflows include SM4TZPZS, PSTZSPZC and SPZTZSM. The experimental results show that SM4TZPZS generalizes slightly better across different datasets on peripheral zone segmentation. It achieves the highest DSC on the test set of the source which is different than the source of the training set. However, no significant results are acquired on transitional zone segmentation. Moreover, the result of the comparison on the test set of the source, which is the same as the training set's, yields no significant difference between the compared workflows. Overall, this study shows that the most preferred approach to the prostate zone segmentation task is SM4TZPZS workflow which employs the nnU-Net with the conventional U-Net architecture.

The resulting solution of the prostate zone segmentation task is used in the second task, PCa segmentation. The compared prostate cancer segmentation ensembles employ segmented prostate zones to enhance the accuracy of the PCa segmentation. A total of four ensembles are compared together with the baseline model. Moreover, a novel metric, DSC_τ is used together with the convention DSC metric to evaluate those ensembles and the baseline model. The results indicate that DSC_τ is higher than the DSC. This observation indicates that the proposed metric takes into account the error of the manual ground truth segmentation.

Ensemble 3 achieves the highest DSC and DSC_τ among all compared ensembles and the baseline model. Furthermore, Ensemble 4 is only marginally worse in terms of both DSC and DSC_τ and there is no statistically significant difference between the results of those two ensembles. The other ensembles and the baseline model are significantly worse.

Lastly, the statistical significance of the difference between collected metrics DSC and DSC_τ on the prostate cancer segmentation task is evaluated. The result indicates that the difference between those metrics is significant.

In this thesis, novel deep learning approaches are proposed for prostate zone and prostate cancer segmentation. Specifically, the thesis (a) reviewed the current state-of-the-art prostate zone and PCa segmentation methods, highlighting the robustness of the no-new U-Net method for medical image segmentation tasks; (b) proposed a method to incorporate attention mechanisms into the nnU-Net method by replacing conventional U-Net deep neural network architecture with architectures that employ attention mechanisms; (c) introduced several prostate zone segmentation workflows which employ the nnU-Net method; (d) evaluated and compared the segmentation accuracy of the proposed workflows and the nnU-Net modifications together with the current state-of-the-art solution; (e) proposed several PCa segmentation ensembles which employ the nnU-Net method and prostate zone information; (f) evaluated and compared the segmentation accuracy of the proposed ensembles together with the current state-of-the-art solution. (g) introduced a DSC modification, DSC_τ, a metric which accounts for the high inter-observer variability in the ground truth; (h) inspected the significance of the difference between the evaluations of PCa segmentation results with the conventional DSC and DSC_τ. The key conclusions are:

1. The results presented in this thesis demonstrate that replacing the conventional U-Net in the nnU-Net method with attention-based U-Net variants does not lead to statistically significant improvements in either overall segmentation accuracy or generalizability across datasets from different sources. Considering that attention mechanisms introduce additional computational overhead, the conventional U-Net remains the preferred architecture within the nnU-Net method.

2. The proposed prostate zone segmentation workflow, SM4TZPZS, demonstrated superior generalizability across datasets from different sources compared to other workflows, achieving statistically significantly higher segmentation accuracy than PSTZSPZC. While no statistically significant difference was observed compared to SPZTZSM, SM4TZPZS consistently matched or outperformed it in nearly all generalizability evaluations, with mean DSC improvements of 0.3% to 5.2% for TZ and 0.6% to 6.4% for

PZ across all DNN architectures. Median DSC improvements ranged from 0.1% to 2.6% (TZ) and 1.9% to 5.9% (PZ), with only two exceptions showing marginal decreases of 0.02% (median DSC) and 1.3% (mean DSC).

3. Ensemble 4 is identified as the most suitable solution for the PCa segmentation task, statistically significantly outperforming the baseline model, Ensemble 1, and Ensemble 2 by 3.17% to 7.13% in aggregated DSC and DSC_τ. Although Ensemble 3 achieved 1% higher aggregated metrics, the difference was not statistically significant, while Ensemble 4 produced significantly fewer false positive patients. As segmentation accuracy is top priority, false positive patients of other approaches were not considered. Lastly, Ensemble 4 requires only two models while Ensemble 3 consists of three models.

4. The novel DSC_τ metric accounts for inter-observer variability in manual delineation of prostate cancer lesions. The results indicate that DSC_τ systematically yields statistically significantly higher values than DSC, suggesting partial compensation for ground truth uncertainty: mean DSC_τ exceeds mean DSC by 5.03% to 7.12% across all evaluated ensembles as well as the baseline model, while median improvements range from 0.09% to 0.79%, with equality observed for the baseline model and Ensemble 1.

[1] T. Akinci D'Antonoli, L. K. Berger, A. K. Indrakanti, N. Vishwanathan, J. Weiss, M. Jung, Z. Berkarda, A. Rau, M. Reisert, T. Küstner, et al. Totalsegmentator mri: Robust sequence-independent segmentation of multiple anatomic structures in mri. *Radiology*, 314(2):e241613, 2025.

[2] N. Aldoj, F. Biavati, F. Michallek, S. Stober, and M. Dewey. Automatic prostate and prostate zones segmentation of magnetic resonance images using densenet-like u-net. *Scientific reports*, 10(1): 14315, 2020.

[3] M. Antonelli, A. Reinke, S. Bakas, K. Farahani, A. Kopp-Schneider, B. A. Landman, G. Litjens, B. Menze, O. Ronneberger, R. M. Summers, et al. The medical segmentation decathlon. *Nature communications*, 13(1):4128, 2022.

[4] D. Bahdanau. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[5] V. Baliyan, C. J. Das, R. Sharma, and A. K. Gupta. Diffusion weighted imaging: technique and applications. *World journal of radiology*, 8(9):785, 2016.

[6] J. O. Barentsz, J. Richenberg, R. Clements, P. Choyke, S. Verma, G. Villeirs, O. Rouviere, V. Logager, and J. J. Fütterer. Esur prostate mr guidelines 2012, 2012.

[7] E. Bass, A. Pantovic, M. Connor, R. Gabe, A. Padhani, A. Rockall, H. Sokhi, H. Tam, M. Winkler, and H. Ahmed. A systematic review and meta-analysis of the diagnostic accuracy of biparametric prostate mri for prostate cancer in men at risk. *Prostate Cancer and Prostatic Diseases*, 24(3):596–611, 2021.

[8] M. Bhalerao and S. Thakur. Brain tumor segmentation based on 3d residual u-net. In *International MICCAI brainlesion workshop*, pages 218–225. Springer, 2019.

[9] S. Bhandary, D. Kuhn, Z. Babaiee, T. Fechter, M. Benndorf, C. Zamboglou, A.-L. Grosu, and R. Grosu. Investigation and benchmarking of u-nets on prostate segmentation tasks. *Computerized Medical Imaging and Graphics*, 107:102241, 2023.

[10] K. R. S. Bhat, S. Samavedi, M. C. Moschovas, F. F. Onol, S. Roof, T. Rogers, V. R. Patel, and A. Sivaraman. Magnetic resonance imaging-guided prostate biopsy—a review of literature. *Asian*

Journal of Urology, 8(1):105–116, 2021.

[11] F. Bray, M. Laversanne, H. Sung, J. Ferlay, R. L. Siegel, I. Soerjomataram, and A. Jemal. Global cancer statistics 2022: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 74(3):229–263, 2024.

[12] G. B. Chavhan, P. S. Babyn, B. Thomas, M. M. Shroff, and E. M. Haacke. Principles, techniques, and applications of t2*-based mr imaging and its special applications. *Radiographics*, 29(5):1433–1449, 2009.

[13] C. Chen, Y. Wang, J. Niu, X. Liu, Q. Li, and X. Gong. Domain knowledge powered deep learning for breast cancer diagnosis based on contrast-enhanced ultrasound videos. *IEEE Transactions on Medical Imaging*, 40(9):2439–2451, 2021.

[14] H. Chen, Q. Dou, L. Yu, J. Qin, and P.-A. Heng. Voxresnet: Deep voxelwise residual networks for brain segmentation from 3d mr images. *NeuroImage*, 170:446–455, 2018.

[15] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

[16] M. Y. Chen, M. A. Woodruff, P. Dasgupta, and N. J. Rukin. Variability in accuracy of prostate cancer segmentation among radiologists, urologists, and scientists. *Cancer Medicine*, 9(19):7172–7182, 2020.

[17] B. S. Chhikara, K. Parang, et al. Global cancer statistics 2022: the trends projection analysis. *Chemical Biology Letters*, 10(1):451–451, 2023.

[18] A. Colonna and F. Scarpa. Improving corneal nerve segmentation using tolerance dice loss function. *Signal, Image and Video Processing*, 18(2):1069–1077, 2024.

[19] A. Comelli, N. Dahiya, A. Stefano, F. Vernuccio, M. Portoghese, G. Cutaia, A. Bruno, G. Salvaggio, and A. Yezzi. Deep learning-based methods for prostate segmentation in magnetic resonance imaging. *Applied Sciences*, 11(2):782, 2021.

[20] W. J. Conover. *Practical nonparametric statistics*. John Wiley & sons, 1999.

[21] R. Cuocolo, M. B. Cipullo, A. Stanzione, L. Ugga, V. Romeo,

L. Radice, A. Brunetti, and M. Imbriaco. Machine learning applications in prostate cancer magnetic resonance imaging. *European radiology experimental*, 3(1):35, 2019.

[22] A. V. Dalca, J. Guttag, and M. R. Sabuncu. Anatomical priors in convolutional networks for unsupervised biomedical segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9290–9299, 2018.

[23] C. de Vente, P. Vos, J. Pluim, and M. Veta. Simultaneous detection and grading of prostate cancer in multi-parametric mri. 2019.

[24] C. De Vente, P. Vos, M. Hosseinzadeh, J. Pluim, and M. Veta. Deep learning regression for prostate cancer detection and grading in bi-parametric mri. *IEEE Transactions on Biomedical Engineering*, 68 (2):374–383, 2020.

[25] A. Duran, P.-M. Jodoin, and C. Lartizien. Prostate cancer semantic segmentation by gleason score group in bi-parametric mri with self attention model on the peripheral zone. In *Medical Imaging with Deep Learning*, pages 193–204. PMLR, 2020.

[26] A. Duran, G. Dussert, O. Rouvière, T. Jaouen, P.-M. Jodoin, and C. Lartizien. Prostattention-net: A deep attention model for prostate cancer segmentation by aggressiveness in mri scans. *Medical Image Analysis*, 77:102347, 2022.

[27] M. Friedman. A comparison of alternative tests of significance for the problem of m rankings. *The annals of mathematical statistics*, 11 (1):86–92, 1940.

[28] E. Gibson, F. Giganti, Y. Hu, E. Bonmati, S. Bandula, K. Gurusamy, B. Davidson, S. P. Pereira, M. J. Clarkson, and D. C. Barratt. Automatic multi-organ segmentation on abdominal ct with dense v-networks. *IEEE transactions on medical imaging*, 37(8):1822–1834, 2018.

[29] D. Gillespie, C. Kendrick, I. Boon, C. Boon, T. Rattay, and M. H. Yap. Deep learning in magnetic resonance prostate segmentation: A review and a new perspective. *arXiv preprint arXiv:2011.07795*, 2020.

[30] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, and D. Xu. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 574–584, 2022.

[31] K. He, X. Zhang, S. Ren, and J. Sun. Identity mappings in deep residual networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam , The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 630–645. Springer, 2016.

[32] M. Hosseinzadeh, P. Brand, and H. Huisman. Effect of adding probabilistic zonal prior in deep learning-based prostate cancer detection. *arXiv preprint arXiv:1907.12382*, 2019.

[33] M. Hosseinzadeh, A. Saha, P. Brand, I. Slootweg, M. de Rooij, and H. Huisman. Deep learning–assisted prostate cancer detection on bi-parametric mri: minimum training data size requirements and effect of prior knowledge. *European Radiology*, pages 1–11, 2022.

[34] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[35] 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*, 18(2):203–211, 2021.

[36] S. Jetley, N. A. Lord, N. Lee, and P. H. Torr. Learn to pay attention. *arXiv preprint arXiv:1804.02391*, 2018.

[37] Z. Jiang, C. Ding, M. Liu, and D. Tao. Two-stage cascaded u-net: 1st place solution to brats challenge 2019 segmentation task. In *International MICCAI brainlesion workshop*, pages 231–241. Springer, 2019.

[38] L. Joskowicz, D. Cohen, N. Caplan, and J. Sosna. Inter-observer variability of manual contour delineation of structures in ct. *European radiology*, 29:1391–1399, 2019.

[39] J. Jucevičius, P. Treigys, J. Bernatavičienė, M. Trakymas, I. Naruševiči ūtė, and R. Briedienė. Segmentation mask resampling for mri prostate localization improvement. In *2022 International Conference on Electrical, Computer and Energy Technologies (ICECET)*, pages 1–6. IEEE, 2022.

[40] J. Jucevičius, P. Treigys, J. Bernatavičienė, M. Trakymas, and I. Naruševiči ūtė. Usage of isotropic mri images improves prostate cancer localization results. *Baltic journal of modern computing.*, 11 (4):703–725, 2023.

[41] S. Klein, U. A. Van Der Heide, I. M. Lips, M. Van Vulpen, M. Star-

ing, and J. P. Pluim. Automatic segmentation of the prostate in 3d mr images by atlas matching using localized mutual information. *Medical physics*, 35(4):1407–1417, 2008.

[42] G. Litjens, O. Debats, J. Barentsz, N. Karssemeijer, and H. Huisman. Computer-aided detection of prostate cancer in mri. *IEEE transactions on medical imaging*, 33(5):1083–1092, 2014.

[43] G. Litjens, R. Toth, W. Van De Ven, C. Hoeks, S. Kerkstra, B. Van Ginneken, G. Vincent, G. Guillard, N. Birbeck, J. Zhang, et al. Evaluation of prostate segmentation algorithms for mri: the promise12 challenge. *Medical image analysis*, 18(2):359–373, 2014.

[44] G. Litjens, O. Debats, J. Barentsz, N. Karssemeijer, and H. Huisman. Prostatex challenge data. *The cancer imaging archive*, 10: K9TCIA, 2017.

[45] J. Ma, Y. He, F. Li, L. Han, C. You, and B. Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024.

[46] S. M. McKinney, M. Sieniek, V. Godbole, J. Godwin, N. Antropova, H. Ashrafian, T. Back, M. Chesus, G. S. Corrado, A. Darzi, et al. International evaluation of an ai system for breast cancer screening. *Nature*, 577(7788):89–94, 2020.

[47] S. Mehralivand, D. Yang, S. A. Harmon, D. Xu, Z. Xu, H. Roth, S. Masoudi, D. Kesani, N. Lay, M. J. Merino, et al. Deep learning-based artificial intelligence for prostate cancer detection at biparametric mri. *Abdominal Radiology*, 47(4):1425–1434, 2022.

[48] P. Mehta, M. Antonelli, S. Singh, N. Grondecka, E. W. Johnston, H. U. Ahmed, M. Emberton, S. Punwani, and S. Ourselin. Autoprostate: towards automated reporting of prostate mri for prostate cancer assessment using deep learning. *Cancers*, 13(23):6138, 2021.

[49] A. Meyer, D. Schindele, D. von Reibnitz, M. Rak, M.and Schostak, and C. Hansen. Prostatex zone segmentations [data set]. the cancer imaging archive, 2020. figshare `https://doi.org/10.7937/TCIA.NBB4-4655`.

[50] F. Milletari, N. Navab, and S.-A. Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. Ieee, 2016.

[51] S. Montagne, D. Hamzaoui, A. Allera, M. Ezziane, A. Luzurier, R. Quint, M. Kalai, N. Ayache, H. Delingette, and R. Renard-

Penna. Challenge of prostate mri segmentation on t2-weighted images: inter-observer variability and impact of prostate morphology. *Insights into imaging*, 12(1):1–12, 2021.

[52] G. Mooij, I. Bagulho, and H. Huisman. Automatic segmentation of prostate zones. *arXiv preprint arXiv:1806.07146*, 2018.

[53] A. Myronenko. 3d mri brain tumor segmentation using autoencoder regularization. In *International MICCAI brainlesion workshop*, pages 311–320. Springer, 2018.

[54] Nacionalinis vėžio centras. Vėžio registro informacinė sistema, 2026. Unpublished clinical dataset.

[55] S. Nikolov, S. Blackwell, A. Zverovitch, R. Mendes, M. Livne, J. De Fauw, Y. Patel, C. Meyer, H. Askham, B. Romera-Paredes, et al. Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy. *arXiv preprint arXiv:1809.04430*, 2018.

[56] O. Oktay, J. Schlemper, L. L. Folgoc, M. C. H. Lee, M. P. Heinrich, K. Misawa, K. Mori, S. G. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert. Attention u-net: Learning where to look for the pancreas. *CoRR*, abs/1804.03999, 2018. URL `http://arxiv.org/abs/1804.03999`.

[57] A. Oto, A. Kayhan, Y. Jiang, M. Tretiakova, C. Yang, T. Antic, F. Dahi, A. L. Shalhav, G. Karczmar, and W. M. Stadler. Prostate cancer: differentiation of central gland cancer from benign prostatic hyperplasia by using diffusion-weighted and dynamic contrast-enhanced mr imaging. *Radiology*, 257(3):715–723, 2010.

[58] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello. Enet: A deep neural network architecture for real-time semantic segmentation. *arXiv preprint arXiv:1606.02147*, 2016.

[59] M. Piert, P. R. Shankar, J. Montgomery, L. P. Kunju, V. Rogers, J. Siddiqui, T. Rajendiran, J. Hearn, A. George, X. Shao, et al. Accuracy of tumor segmentation from multi-parametric prostate mri and 18 f-choline pet/ct for focal prostate cancer therapy applications. *EJNMMI research*, 8:1–14, 2018.

[60] O. Rainio and R. Klén. Modified dice coefficients for evaluation of tumor segmentation from pet images: A proof-of-concept study. *Journal of Imaging Informatics in Medicine*, pages 1–9, 2025.

[61] T. Riepe, M. Hosseinzadeh, P. Brand, and H. Huisman. Anisotropic deep learning multi-planar automatic prostate segmentation. In *Proceedings of the 28th International Society for Magnetic Resonance in Medicine Annual Meeting. http://indexsmart. mirasmart. com/ISMRM2020/PDFfiles/3518. html*, 2020.

[62] E. Romera, J. M. Alvarez, L. M. Bergasa, and R. Arroyo. Erfnet: Efficient residual factorized convnet for real-time semantic segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 19(1): 263–272, 2017.

[63] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich , Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[64] A. B. Rosenkrantz, S. Kim, N. Campbell, B. Gaing, F.-M. Deng, and S. S. Taneja. Transition zone prostate cancer: revisiting the role of multiparametric mri at 3 t. *American Journal of Roentgenology*, 204 (3):W266–W272, 2015.

[65] A. Saha, M. Hosseinzadeh, and H. Huisman. Encoding clinical priori in 3d convolutional neural networks for prostate cancer detection in bpmri. *arXiv preprint arXiv:2011.00263*, 2020.

[66] A. Saha, M. Hosseinzadeh, and H. Huisman. End-to-end prostate cancer detection in bpmri via 3d cnns: effects of attention mechanisms, clinical priori and decoupled false positive reduction. *Medical image analysis*, 73:102155, 2021.

[67] A. Saha, J. S. Bosma, J. J. Twilt, B. van Ginneken, A. Bjartell, A. R. Padhani, D. Bonekamp, G. Villeirs, G. Salomon, G. Giannarini, et al. Artificial intelligence and radiologists in prostate cancer detection on mri (pi-cai): an international, paired, non-inferiority , confirmatory study. *The Lancet Oncology*, 25(7):879–887, 2024.

[68] D. S. Sahni, J. Morrison, and H. Y. Leung. Grid-based cognitive diagnostic prostatic biopsy without transrectal ultrasonography. *BJU international*, 134(4):659–663, 2024.

[69] J. Schlemper, O. Oktay, M. Schaap, M. Heinrich, B. Kainz, B. Glocker, and D. Rueckert. Attention gated networks: Learning to leverage salient regions in medical images. *Medical image analysis*, 53:197–207, 2019.

[70] R. L. Siegel, K. D. Miller, N. S. Wagle, and A. Jemal. Cancer statistics, 2023. *CA: a cancer journal for clinicians*, 73(1), 2023.

[71] L. A. Simmons, H. U. Ahmed, C. M. Moore, S. Punwani, A. Freeman, Y. Hu, D. Barratt, S. C. Charman, J. Van der Meulen, and M. Emberton. The picture study—prostate imaging (multi-parametric mri and prostate histoscanning™) compared to transperineal ultrasound guided biopsy for significant prostate cancer risk evaluation. *Contemporary clinical trials*, 37(1):69–83, 2014.

[72] L. A. Simmons, A. Kanthabalan, M. Arya, T. Briggs, D. Barratt, S. C. Charman, A. Freeman, J. Gelister, D. Hawkes, Y. Hu, et al. The picture study: diagnostic accuracy of multiparametric mri in men requiring a repeat prostate biopsy. *British journal of cancer*, 116 (9):1159–1165, 2017.

[73] A. L. Simpson, M. Antonelli, S. Bakas, M. Bilello, K. Farahani, B. Van Ginneken, A. Kopp-Schneider, B. A. Landman, G. Litjens, B. Menze, et al. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv preprint arXiv:1902.09063*, 2019.

[74] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 71(3):209–249, 2021.

[75] B. Turkbey, A. B. Rosenkrantz, M. A. Haider, A. R. Padhani, G. Villeirs, K. J. Macura, C. M. Tempany, P. L. Choyke, F. Cornud, D. J. Margolis, et al. Prostate imaging reporting and data system version 2.1: 2019 update of prostate imaging reporting and data system version 2. *European urology*, 76(3):340–351, 2019.

[76] F. Vaassen, C. Hazelaar, A. Vaniqui, M. Gooding, B. Van der Heyden, R. Canters, and W. Van Elmpt. Evaluation of measures for assessing time-saving of automatic organ-at-risk segmentation in radiotherapy. *Physics and Imaging in Radiation Oncology*, 13:1–6, 2020.

[77] S. Vaccarella, M. Li, F. Bray, R. Kvale, D. Serraino, V. Lorenzoni, A. Auvinen, and L. Dal Maso. Prostate cancer incidence and mortality in europe and implications for screening activities: population based study. *bmj*, 386, 2024.

[78] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[79] L. Vincent. Morphological transformations of binary images with arbitrary structuring elements. *Signal processing*, 22(1):3–23, 1991.

[80] C. Wachinger, M. Reuter, and T. Klein. Deepnat: Deep convolutional neural network for segmenting neuroanatomy. *NeuroImage*, 170:434–445, 2018.

[81] J. Wasserthal, H.-C. Breit, M. T. Meyer, M. Pradella, D. Hinck, A. W. Sauter, T. Heye, D. T. Boll, J. Cyriac, S. Yang, et al. Totalsegmentator: robust segmentation of 104 anatomic structures in ct images. *Radiology: Artificial Intelligence*, 5(5):e230024, 2023.

[82] F. Wilcoxon. Probability tables for individual comparisons by ranking methods. *Biometrics*, 3(3):119–122, 1947.

[83] T. E. Yankeelov and J. C. Gore. Dynamic contrast enhanced magnetic resonance imaging in oncology: theory, data acquisition, analysis, and examples. *Current Medical Imaging*, 3(2):91–107, 2007.

[84] S. Yoo, I. Gujrathi, M. A. Haider, and F. Khalvati. Prostate cancer detection using deep convolutional neural networks. *Scientific reports*, 9(1):19518, 2019.

[85] J. Yu, A. S. Fulcher, S. G. Winks, M. A. Turner, R. D. Clayton, M. Brooks, and S. Li. Diagnosis of typical and atypical transition zone prostate cancer and its mimics at multiparametric prostate mri. *The British journal of radiology*, 90(1073):20160693, 2017.

[86] X. Yu, R. Liu, L. Song, W. Gao, X. Wang, and Y. Zhang. Differences in the pathogenetic characteristics of prostate cancer in the transitional and peripheral zones and the possible molecular biological mechanisms. *Frontiers in Oncology*, 13:1165732, 2023.

[87] G. Zhang, Z. Yang, B. Huo, S. Chai, and S. Jiang. Multiorgan segmentation from partially labeled datasets with conditional nnu-net. *Computers in Biology and Medicine*, 136:104658, 2021.

[88] H. Zheng, A. L. Y. Hung, Q. Miao, W. Song, F. Scalzo, S. S. Raman, K. Zhao, and K. Sung. Atpca-net: anatomical-aware prostate cancer detection network on multi-parametric mri. *Scientific Reports*, 14(1):5740, 2024.

[89] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang. Unet++: Redesigning skip connections to exploit multiscale features in im-

age segmentation. *IEEE transactions on medical imaging*, 39(6):1856–1867, 2019.

Articles in international research journals with a citation index in the Clarivate Analytics Web of Science (CA WoS) database.

Articles in CA WoS journals:

[A.1] **Aleksas Vaitulevičius** & Jolita Bernatavičienė & Jurgita Markevičiutė & Ieva Naruševičiūtė & Mantas Trakymas & Povilas Treigys. Advancements in prostate zone segmentation: integrating attention mechanisms into the nnU-Net framework. *Machine Learning: Science and Technology*. IOP Publishing. vol. 5, no. 4, art. no. 045003, 2024. p. 2632-2153. ISSN 2632-2153. DOI: 10.1088/2632-2153/ad7f24.

[A.2] **Aleksas Vaitulevičius** & Jolita Bernatavičienė & Ieva Naruševičiūtė & Jurgita Markevičiutė & Mantas Trakymas & Povilas Treigys. Advancements in Prostate Cancer Segmentation: Integrating Prostate Zonal Information. *Digital Health*. SAGE Publications. vol. 12, 2026. DOI: 10.1177/20552076261439004.

Other publications:

[B.1] **Aleksas Vaitulevičius** & Roman Surkant & Povilas Treigys & Jurgita Markevičiutė & Mantas Trakymas & Jolita Bernatavičienė & Ieva Naruševičiūtė. DCE MRI modality investigation for cancerous prostate region detection: case analysis. *Journal of WSCG*, Union Agency Science Press. vol. 30, no. 1-2, 2022. p. 48-55. ISSN 1213-6972. eISSN 1213-6964. DOI: 10.24132/CSRN.3201.7.

[B.2] Povilas Treigys & **Aleksas Vaitulevičius** & Jolita Bernatavičienė & Jurgita Markevičiutė & Ieva Naruševičiūtė & Mantas Trakymas. Detecting Cancerous Regions in DCE MRI using Functional Data, XGboost and Neural Networks. *Communication Papers of the 17th Conference on Computer Science and Intelligence Systems*, Annals of Computer Science and Information Systems. vol. 32, 2022. p. 23-30. ISSN 2300-5963. DOI: 10.15439/2022F128.

[B.3] **Vaitulevičius, Aleksas** & Bernatavičienė, Jolita & Markevičiutė, Jurgita & Naruševičiūtė, Ieva & Trakymas, Mantas & Treigys, Povilas. Impact of timestamp and segmentation map selection for

cancerous prostate regions in DCE MRI classification. *DAMSS: 13th Conference on Data Analysis Methods for Software Systems*, Vilnius University Press. ISSN 2669-0233. 2022, vol. 31, no. 1-2, p. 97. DOI: 10.15388/DAMSS.13.2022.

[B.4] **Aleksas Vaitulevičius** & Jolita Bernatavičienė & Povilas Treigys. nnU-Net: investigation of attention mechanism for prostate zone segmentation task. *2024 E-Health and Bioengineering Conference (EHB)*, IEEE. 2024. p. 1-4. ISSN 2575-5137. eISSN 2575-5145. DOI: 10.1109/EHB64556.2024.10805724.

Presentations at scientific conferences:

[C.1] **Vaitulevičius, Aleksas** & Surkant, Roman & Treigys, Povilas & Markevičiutė, Jurgita & Trakymas, Mantas & Bernatavičienė, Jolita & Naruševičiūtė, Ieva. DCE MRI modality investigation for cancerous prostate region detection: case analysis. *WSCG 2022*. Pilzen, Czech, May 17-20, 2022.

[C.2] **Vaitulevičius, Aleksas** & Bernatavičienė, Jolita & Treigys, Povilas. nnU-Net: investigation of attention mechanism for prostate zone segmentation task. *2024 E-Health and Bioengineering Conference (EHB)*. IASI, Romania, November 14-15, 2024.

[C.3] **Vaitulevičius, Aleksas** & Bernatavičienė, Jolita & Markevičiutė, Jurgita & Naruševičiūtė, Ieva & Trakymas, Mantas & Treigys, Povilas. Prostate Cancer Segmentation Evaluation: Tolerant Dice Score Coefficient and Dice Score Coefficient Comparison. *International Vilnius Conference on Statistics and Its Applications 2025*. Vilnius, Lithuania, August 27-29, 2025.

Presentations at national scientific conferences:

[D.1] **Vaitulevičius, Aleksas**. Detecting Cancerous Regions in DCE MRI using Functional Data. *Dirbtinio intelekto technologijų taikymai vaizdų analizėje*. Vilnius, Lithuania, October 5, 2022.

[D.2] **Vaitulevičius, Aleksas** & Bernatavičienė, Jolita & Markevičiutė, Jurgita & Naruševičiūtė, Ieva & Trakymas, Mantas & Treigys, Povilas. Impact of timestamp and segmentation map selection for cancerous prostate regions in DCE MRI classification. *DAMSS:*

13th Conference on Data Analysis Methods for Software Systems. Druskininkai, Lithuania, December 1-3, 2022.

CURRICULUM VITAE

Aleksas Vaitulevičius graduated from the Faculty of Mathematics and Informatics at Vilnius University in 2018, obtaining a Bachelor's degree in Informatics. In 2021, he completed a Master's degree in Informatics at the same institution. From 2018 to 2019, he worked as a Java Programmer at VĮ Registrų Centras. Subsequently, he was employed as a Data Analyst at UAB Euromonitor International -Eastern Europe until 2022. Since 2022, he has been working as a Machine Learning Specialist at UAB Novian PRO, where he develops and implements solutions based on deep learning methods.

Aleksas Vaitulevičius 2018 metais baigė Vilniaus universiteto Matematikos ir informatikos fakultetą ir įgijo informatikos bakalauro laipsnį. 2021 metais tame pačiame universitete jam suteiktas informatikos magistro laipsnis. 2018–2019 m. A. Vaitulevičius dirbo „Java“ programuotoju VĮ Registrų centre. Vėliau, iki 2022 metų, jis dirbo duomenų analitiku UAB „Euromonitor International – Eastern Europe“. Nuo 2022 metų A. Vaitulevičius dirba mašininio mokymosi specialistu UAB „Novian PRO“, kur kuria ir diegia giliojo mokymosi metodais paremtus sprendimus.

Remiantis naujausiais pasauliniais onkologinių susirgimų statistiniais duomenimis, prostatos vėžys – viena iš dažniausiai diagnozuojamų piktybinių ligų pasaulyje. Svarbus diagnostikos etapas yra multiparametrinė magnetinio rezonanso tomografijos (mpMRT) vaizdų analizė. Pastaraisiais metais vis plačiau taikomi giliojo mokymosi metodai, siekiant pagerinti prostatos vėžio aptikimą ir segmentavimą naudojant mpMRT duomenis. Nepaisant šios pažangos, prostatos vėžio segmentavimas vis dar išlieka sudėtinga užduotis. Šiame darbe nagrinėjami prostatos vėžio segmentavimo ansambliai, siekiant padidinti segmentavimo tikslumą ir patikimumą. Ypatingas dėmesys skiriamas pažangiausiam medicininių vaizdų segmentavimo metodui no-new U-Net (nnU-Net).

Daugybė tyrimų rodo, kad prostatos zonų informacijos įtraukimas į prostatos vėžio segmentavimo procesą didina aptikimo ir segmentavimo tikslumą. Todėl šiame darbe taip pat analizuojamos prostatos zonų segmentavimo veiksmų sekos. Šioje disertacijoje pasiūlytos dvi naujos prostatos zonų segmentavimo veiksmų sekos, abi paremtos giliaisiais neuroniniais tinklais (angl. *Deep Neural Networks*, DNN), apmokytais naudojant nnU-Net metodą. Eksperimentiniai rezultatai parodė, kad tiksliausi rezultatai gaunami naudojant veiksmų seką atskirų modelių tranzitorinei ir periferinei zonoms segmentuoti. Šiame darbe taip pat tiriama, kaip prostatos zonų segmentavimo rezultatus veikia nnU-Net naudojamos U-Net architektūros keitimas į jos variacijas su dėmesio mechanizmais. Nagrinėjamos kelios architektūros: viengubo dėmesio U-Net (angl. *single attention U-Net*), tinklinio dėmesio U-Net (angl. *grid attention U-Net*), daugiamačio dėmesio U-Net (angl. *multidimensional attention U-Net*) bei dvigubo dėmesio U-Net (angl. *dual-attention U-Net*). Eksperimentai neatskleidė statistiškai reikšmingų skirtumų tarp šių architektūrų rezultatų. Prostatos zonų segmentavimo eksperimentams naudoti du viešai prieinami duomenų rinkiniai: ProstateX ir Decathlon iššūkio duomenys.

Disertacijoje taip pat pasiūlyti keturi nauji prostatos vėžio segmentavimo ansambliai, sudaryti iš nnU-Net metodu apmokytų DNN modelių. Visi pasiūlyti ansambliai naudoja prostatos zonų kaukes. Eksperimentai parodė, kad tiksliausi segmentavimo rezultatai pasiekiami sujungiant modelio, skirto periferinės zonos pažeidimams segmentuoti,

rezultatus su modelio, segmentuojančio prostatos vėžinius regionus visoje prostatoje ir naudojančio zonų kaukes kaip papildomą įvesties kanalą, rezultatus. Be to, šiame darbe siūlomas modifikuotas Dice įverčio koeficiento (angl. *Dice Score Coefficient*, DSC) matas, tolerantiškas Dice įverčio koeficientas DSC_τ. Ši metrika sukurta atsižvelgiant į didelę tarpstebėtojinę variaciją, būdingą rankiniams prostatos vėžio segmentavimo duomenims. Matų lyginimas parodė, kad DSC_τ nuosekliai suteikia aukštesnius įverčius už standartinį DSC, o tai rodo, kad DSC_τ atsižvelgia į toleruojamą variaciją tarp piktybinių navikų apibrėžimų. Prostatos vėžio segmentavimo eksperimentams naudoti Nacionalinio vėžio instituto pateikti duomenys, kurie nėra viešai prieinami.

Prostatos vėžys – viena iš labiausiai paplitusių onkologinių ligų pasaulyje. Naujausioje pasaulinėje onkologinių ligų statistikoje, kurią pateikė Bray et al. [11], analizuojami 2022 metų duomenys, apimantys 36 vėžio tipus 185 šalyse. Pagal šį tyrimą, prostatos vėžys yra antras pagal dažnumą vyrams diagnozuojamas vėžys pasaulyje, sudarantis 14,2 % visų naujai vyrams nustatytų onkologinių ligų 2022 metais. Jį lenkia tik plaučių vėžys, kuris sudaro 15,3 % naujų atvejų. Be didelio sergamumo, prostatos vėžys taip pat reikšmingai prisideda prie su onkologinėmis ligomis susijusio mirtingumo. 2022 metais jis sudarė 7,3 % visų vyrų mirčių nuo onkologinių ligų ir užėmė penktą vietą tarp pagrindinių vyrų mirtingumo nuo onkologinių ligų pasaulyje. Atsižvelgiant į didelį paplitimą ir reikšmingą mirtingumą, tikslus, greitas ir didelio masto prostatos vėžio regionų nustatymas yra itin svarbus.

Multiparametrinė magnetinio rezonanso tomografija (mpMRT) tapo vienu iš pagrindinių modalumų prostatos vėžiui nustatyti ir diagnozuoti. Tarp struktūrizuotų ataskaitų sistemų mpMRT vertinti viena iš plačiausiai naudojamų yra prostatos skenavimo ataskaitų ir duomenų sistema (angl. *Prostate Imaging Reporting and Data System*, PI-RADS), kurią pristatė Barentsz et al. [6]. Autoriai standartizavo mpMRT radinių interpretavimą ir pateikimą klinikinėje prostatos vėžio diagnostikoje, taip gerindami nuoseklumą, atkuriamumą ir komunikaciją tarp radiologų bei gydymo institucijų. Nuo jos įvedimo PI-RADS išsivystė į visame pasaulyje taikomą sistemą, suteikiančią struktūrizuotą ir atkuriamą prostatos mpMRT radinių interpretavimo ir pateikimo būdą. Naujausia versija, PI-RADS v2.1, kurią pristatė Turkbey et al. [75], tikslina ankstesnes rekomendacijas ir aiškiau apibrėžia sekų specifinius vertinimo kriterijus. Svarbu tai, kad PI-RADS v2.1 nustato skirtingus vertinimo kriterijus skirtingoms zonoms [75], atspindinčius skirtingą magnetinio rezonanso tomografijos (MRT) sekų diagnostinį tikslumą įvairiose prostatos anatominėse srityse. Vertinimas daugiausia atliekamas tranzitorinėje (TZ) ir periferinėje zonose (PZ), nes čia aptinkama didžioji dalis prostatos vėžio atvejų.

Nepaisant plataus mpMRT ir struktūrizuotų ataskaitų sistemų taikymo, tikslus prostatos vėžio regionų lokalizavimas ir segmentavimas išlieka techniškai sudėtingas uždavinys net ir patyrusiems radiologams. Segmentavimo uždavinys – skaidyti vaizdą į prasmingas sritis. At-

liekant šį uždavinį, kiekvienam vaizdo taškeliui priskiriama tam tikra klasė. Užduoties tikslas – aptikti ir tiksliai apibrėžti vaizde esančius objektus ar anatomines struktūras. Kelete tyrimų pabrėžta, kad ypač sunku nustatyti naviko sritis tranzitorinėje zonoje. Oto et al. [57], Rosenkrantz et al. [64] ir Yu et al. [85] nurodė, kad tranzitorinėje zonoje atsirandantys navikai dažnai sunkiai atskiriami nuo gerybinės prostatos hiperplazijos mazgų, nes jų vaizdinės savybės gali sutapti. Rosenkrantz et al. [64] taip pat pabrėžė, kad tranzitorinei zonai būdingas difuzinis foninis heterogeniškumas ir daugiamazgė architektūra reikšmingai apsunkina piktybiškų navikų atpažinimą šioje srityje. Panašiai Yu et al. [85] teigė, kad stromos asimetrija tranzitorinėje zonoje gali imituoti vėžinį naviką, o lėtinis prostatitas periferinėje zonoje sukelia vaizdinių požymių, kurie primena vėžinius navikus ir taip didina klaidingai teigiamų interpretacijų riziką.

Taigi matyti aiškus naujų, patikimų ir lengvai pritaikomų metodų, galinčių tiksliai segmentuoti ir lokalizuoti vėžinius regionus prostatoje, siekiant pagerinti diagnostinį tikslumą, poreikis. Didėjantis kruopščiai atrinktų istorinių duomenų prieinamumas dar labiau skatina duomenimis, ypač giliuoju mokymusi (angl. *Deep Learning*, DL), grįstų metodų kūrimą. Šie metodai plačiai taikomi įvairiose medicininių vaizdų segmentavimo užduotyse ir demonstruoja pažangiausius rezultatus skirtinguose vaizdiniuose modalumuose. Prostatos mpMRT kontekste gilių jų neuroninių tinklų (angl. *Deep Neural Networks*, DNN) architektūros geba įsisavinti sudėtingas hierarchinių požymių išraiškas tiesiogiai iš tūrinės duomenų informacijos. Šie modeliai gali veiksmingai atpažinti subtilius erdvinius ir intensyvumo požymius mpMRT vaizduose, leidžiančius tiksliai išskirti ir lokalizuoti prostatos vėžinius regionus.

Tyrimo problema

mpMRT susideda iš kelių vaizdinių sekų, iš kurių kiekviena suteikia anatominės ir funkcinės informacijos, reikalingos prostatos vėžiui diagnozuoti skirtingose prostatos zonose. PI-RADS v2.1 sistemoje pagrindinės sekos yra T2-svertiniai (angl. *T2-Weighted*, T2W) vaizdai, regimieji difuzijos koeficientai (angl. *Apparent Diffusion Coefficient*, ADC), difuzijos svertiniai vaizdai (angl. *Diffusion-Weighted Imaging*, DWI) ir dinaminio kontrastinio sustiprinimo (angl. *Dynamic Contrast-Enhanced*, DCE) vaizdai. DWI ir ADC sekas išsamiai aprašė Baliyan et al. [5]. DWI

sekos tiesiogiai gaunamos naudojant magnetinio rezonanso tomografiją, o ADC žemėlapiai apskaičiuojami iš kelių DWI vaizdų, įgytų taikant skirtingus difuzijos svorius. Abi šios sekos yra pagrindinės aptinkant vėžinius regionus periferinėje prostatos zonoje. T2-svertiniai vaizdai, kuriuos aprašė Chavhan et al. [12], pasižymi aukšta erdvine raiška ir suteikia detalią anatominę informaciją. Ši seka yra pagrindinė aptinkant vėžinius regionus tranzitorinėje zonoje. O štai DCE vaizdai apima pakartotinius tūrinio vaizdo įgijimus suleidus intraveninės kontrastinės medžiagos kontrasto kaupimosi dinamikai laike įvertinti. Vis dėlto Bass et al. [7], Turkbey et al. [75] nustatė, kad DCE indėlis į prostatos vėžio diagnostiką, ypač periferinėje zonoje, yra antrinis – jo diagnostinė vertė ribota, palyginti su DWI ir ADC, todėl ši seka dažniausiai atlieka pagalbinį vaidmenį. Visos minėtos sekos yra tūrinės ir paprastai pasižymi anizotropine erdvine raiška.

Ankstesniuose tyrimuose, pavyzdžiui, [25, 33, 66], buvo analizuojamas prostatos vėžio segmentavimas, pasitelkiant prostatos zonų informaciją. Šie tyrimai pademonstravo, kad prostatos zonų anatomijos įtraukimas gali pagerinti vėžio segmentavimo tikslumą. Tačiau esamuose darbuose trūksta metodų, kurie būtų aiškiai suderinti su PI-RADS sistema, pateikiančia kliniškai pagrįstas mpMRT interpretavimo rekomendacijas. Ypač pastebima spraga giliojo mokymosi metodų srityje, kai modeliai neintegruoja zonai specifinių mpMRT sekų pagal PI-RADS rekomendacijas. Šioje disertacijoje ši spraga sprendžiama sistemingai tiriant prostatos vėžio segmentavimo metodus, kurie remiasi prostatos zonų informacija, ir derinant sprendinių architektūrą bei naudojamas sekas su PI-RADS sistema, siekiant pagerinti tiek klinikinį pritaikomumą, tiek segmentavimo tikslumą.

Prostatos vėžio segmentavimas, naudojant prostatos zonų informaciją, neišvengiamai reikalauja tikslaus pačių prostatos zonų segmentavimo. Šiam uždaviniui skirta nemažai tyrimų, kuriuose siūlomi įvairūs giliojo mokymosi metodai automatiniam prostatos zonų segmentavimui. Šioje disertacijoje ši tyrimų kryptis toliau plėtojama, o gauti rezultatai tiesiogiai integruojami į pagrindinę tyrimo sritį – prostatos vėžio segmentavimą. Integracija leidžia atlikti prostatos zonų informacija pagrįstą prostatos vėžio segmentavimą, derinant giliojo mokymosi modelių rezultatus su anatominiu ir klinikiniu kontekstu, apibrėžtu PI-RADS sistemoje, ir taip siekiant didesnio segmentavimo tikslumo.

Galiausiai, eksperimentiniai tyrimai, kuriuos aprašė Chen et al. [16], rodo, kad prostatos vėžio segmentavimas išlieka sudėtinga užduotis. Net ir tarp patyrusių radiologų pastebima didelė rezultatų variacija, kuri atspindi subjektyvų prostatos vėžio ribų apibrėžimo sudėtingumą mpMRT vaizduose. Ši tarpstebėtojinė variacija (angl. *inter-observer variability*) daro tiesioginę įtaką anotacijų patikimumui, todėl apsunkina tiek automatinių metodų mokymą, tiek objektyvų jų vertinimą. Atsižvelgiant į tai, disertacijoje nagrinėjamos alternatyvios ir patikimesnės prostatos vėžio segmentavimo vertinimo strategijos.

Apibendrinant disertacijoje plėtojama vientisa automatizuoto prostatos vėžio segmentavimo veiksmų seka. Joje integruojamas atskiras prostatos zonų segmentavimo komponentas, kurio rezultatai panaudojami vėžio segmentavimo etape, siekiant pagerinti navikinių pakitimų išskyrimo tikslumą ir anatominį nuoseklumą. Pagrindinis sistemos komponentas – prostatos vėžio segmentavimas – vertinamas taikant specialiai šiai užduočiai pritaikytą metodiką, atsižvelgiančią į prostatos vėžio heterogeniškumą ir tarpstebėtojinę variaciją. Tokia integruota prieiga užtikrina metodologinį nuoseklumą ir aukštą klinikinį aktualumą.

Tyrimo objektas

Tyrimo objektas – tūriniai mpMRT duomenys, naudojami prostatos vėžiui segmentuoti, įtraukiant prostatos anatominių zonų informaciją. Šiam objektui tirti taikomi giliaisiais neuroniniais tinklais pagrįsti modeliai.

Tyrimo tikslas ir uždaviniai

Tyrimo tikslas – sukurti ir kritiškai įvertinti giliaisiais neuroniniais tinklais pagrįstų veiksmų sekų ir ansamblių, skirtų prostatos anatominėms zonoms ir prostatos vėžiui segmentuoti tūriniuose mpMRT duomenyse, segmentavimo tikslumą. Šiam tikslui pasiekti keliami šie uždaviniai:

1. Atlikti išsamią šiuolaikinių giliojo mokymosi metodų, taikomų prostatos zonoms ir prostatos vėžiui segmentuoti, literatūros analizę, ypatingą dėmesį skiriant neuroninių tinklų architektūrų įvairovei, duomenų paruošimo strategijoms, vertinimo metodikoms bei jų taikymui medicininių vaizdų segmentavimo srityje.

2. Suprojektuoti giliuoju mokymusi paremtus ansamblius, kurie prostatos vėžio segmentavimo užduočiai įtraukia informaciją apie prostatos zonas, ir empiriniais eksperimentais įvertinti jų segmentavimo tikslumą, siekiant nustatyti tiksliausią ansamblį.

3. Pasiūlyti ir ištirti DL pagrindu sukurtas prostatos zonų segmentavimo veiksmų sekas, dėmesį skiriant skirtingų giliųjų neuroninių tinklų architektūrų ir veiksmų sekų dizainų empiriniam lyginimui, kad būtų nustatytas tiksliausias sprendimas prostatos zonoms atskirti.

4. Pasiūlyti prostatos vėžio segmentavimo vertinimo metodą, atsižvelgiantį į anotacijų tarpstebėtojinę variaciją, ir empiriniais eksperimentais įvertinti jos pranašumus, lyginant su standartiniais vertinimo metodais.

Mokslinis darbo naujumas ir praktinė vertė

Šiame tyrime pristatomi ir sistemiškai įvertinami keli nauji metodologiniai sprendimai, prisidedantys prie prostatos vėžio bei prostatos anatominių zonų segmentavimo uždavinių tobulinimo. Pagrindiniai darbo rezultatai ir indėliai yra šie:

- Nenustatyta statistiškai reikšmingų skirtumų tarp dėmesio mechanizmu paremtų U-Net architektūrų variantų (viengubo dėmesio U-Net (angl. *single attention U-Net*), tinklinio dėmesio U-Net (angl. *grid attention U-Net*), daugiamačio dėmesio U-Net (angl. *multidimensional attention U-Net*), dvigubo dėmesio U-Net (angl. *dual-attention U-Net*)) bei standartinio U-Net, atliekant prostatos zoninę segmentaciją.

- Nustatyta, kad naujai pasiūlyta dviejų modelių veiksmų seka, kurioje kiekvienas modelis yra specializuotas atskirai prostatos zonai, yra veiksmingiausias iš visų vertintų sprendimų, nes užtikrina geriausią generalizuojamumą tarp skirtingų duomenų šaltinių.

- Identifikuota veiksmingiausia prostatos vėžio segmentacijos ansamblio konfigūracija, sudaryta iš bazinio modelio ir modelio, apmokyto naudojant visas mpMRT sekas vėžio segmentacijai PZ zonoje. Šis ansamblis pasiekia panašų segmentavimo tikslumą

kaip ir tiksliausiai veikiantis ansamblis, tačiau labai sumažina klaidingai teigiamų pacientų skaičių.

- Pasiūlytas tolerancija paremtas Dice įverčio koeficiento matas prostatos vėžio segmentacijai vertinti, kuris atsižvelgia į anotacijų tarpstebėtojinę variaciją. Šis matas suteikia aukštesnius įverčius už klasikinį Dice įverčio koeficiento matą, o tai rodo, kad ši metrika ignoruoja nedidelius ribų neatitikimus apibrėžtoje tolerancijos srityje.

Ginamieji teiginiai

Šioje disertacijoje ginami šie tyrimu pagrįsti teiginiai:

1. U-Net architektūros išplėtimas dėmesio mechanizmais nepadidina prostatos zonų segmentavimo tikslumo.
2. Naudojant du atskirus giliųjų neuroninių tinklų modelius, iš kurių kiekvienas yra skirtas konkrečiai prostatos zonai segmentuoti, segmentavimo tikslumas didėja.
3. Prostatos vėžio segmentavimo ansambliai, sudaryti iš prostatos zonai specifinių giliųjų neuroninių tinklų modelių, pasiekia didesnį segmentavimo tikslumą už vieną prostatos vėžio segmentavimo modelį.
4. Įtraukus toleranciją į Dice įverčio koeficientą, kompensuojama aukšta tarpstebėtojinė variacija, esanti rankiniu būdu apibrėžtuose vėžinio regiono ribose.

Įvado skyriuje pristatoma bendra prostatos vėžio segmentavimo situacija, temos aktualumas, tyrimo tikslas, tyrimo uždaviniai ir ginamieji teiginiai. S.1 skyrius apima literatūros apžvalgą, sutelkiant dėmesį į šiuolaikinius DL metodus prostatos zonoms ir prostatos vėžiui segmentuoti. S.2 skyriuje pateikiama tyrime taikyta metodika, tarp jų ir siūlomos prostatos zonų segmentavimo veiksmų sekos, dėmesio mechanizmus integruojančios DNN architektūros, siūlomi prostatos vėžio segmentavimo ansambliai ir segmentavimo tikslumo vertinimo matai. S.3 skyriuje aprašomi tyrimo duomenys, jų paruošimas ir empirinio eksperimento nustatymai. S.4 skyriuje pateikiami tyrimo rezultatai, o galutinėje disertacijos dalyje – bendros išvados.

Disertaciją sudaro 158 puslapių; santrauka lietuvių kalba prasideda nuo 136 puslapio; pateiktos 27 lentelės, 29 paveikslai ir 89 literatūros šaltiniai.

S.1. Literatūros analizė

Vienas iš pagrindinių iššūkių automatizuojant prostatos vėžio segmentavimą – didelė tarpstebėtojinė variacija [16]. Šioje disertacijoje ši problema sprendžiama siūlant naują metriką (žr. S.2.3).

Daugelyje tyrimų prostatos vėžiui segmentuoti taikomi DNN modeliai, naudojant prostatos zonų informaciją. Parodyta, kad tikimybinės zonų segmentacijos didina prostatos vėžio segmentavimo tikslumą [32, 33], o kituose darbuose siūlomos DNN architektūros, naudojančios zonų kaukes, pvz., AtPCa-Net [88]. Apibendrinant zoninė informacija yra svarbi šio uždavinio dalis. Šiame darbe ši kryptis plėtojama toliau, siūlant naujus ansamblius tikslesniam prostatos vėžio segmentavimui.

Nuo savo pristatymo [4] dėmesio mechanizmai plačiai taikomi medicininės kompiuterinės regos uždaviniuose. Prostatos vėžio segmentavimo srityje pasiūlytos įvairios architektūros, tokios kaip ProstAttention-Net [25, 26], paremta U-Net su tinklinio dėmesio (angl. *grid attention*) mechanizmu [56, 69], bei dviejų DNN modelių sistemos su sprendimų suliejimu (angl. *decision fusion*) [66], kuriose pirminio DNN modelio architektūra yra dvigubo dėmesio U-Net, tai yra UNet++ [89], išplėsta tinklinio dėmesio mechanizmu. Atsižvelgiant į ankstesnius tyrimus, šiame darbe dėmesio mechanizmų taikymas prostatos vėžiui segmentuoti nėra toliau nagrinėjamas, tačiau vertinamas jų poveikis prostatos zonoms segmentuoti.

Tyrimai rodo, kad prostatos zonų segmentacijoje tarpstebėtojinė variacija yra ypač didelė kraštinėse srityse, ypač mažo tūrio prostatose [51]. Tai pabrėžia reprezentatyvių ir morfologiškai įvairių duomenų svarbą. Šiame darbe ši problema sprendžiama naudojant du viešai prieinamus duomenų rinkinius (žr. S.3).

Prostatos zonoms segmentuoti pasiūlyta įvairių metodų. Tyrimai rodo, kad lengvieji modeliai, tokie kaip efektyvus neuroninis tinklas (angl. *Efficient Neural network*) [58] ir efektyvus faktorizuotas likutinis konvoliucinis tinklas (angl. *Efficient Residual Factorized convolution network*) tinkami ribotų resursų sąlygomis, tačiau U-Net išlieka veiksmingiausias turint pakankamai skaičiavimo galios [19, 62]. Taip pat siūlomos U-Net variacijos, pvz., dense-2 U-Net [2]. Medicininės segmentacijos iššūkių rezultatai parodė, kad no-new U-Net (nnU-Net) metodas pasiekia didžiausią prostatos zonų segmentavimo tikslumą [3,

35] ir yra plačiai taikomas įvairiuose medicininių vaizdų uždaviniuose [1, 40, 45, 81, 87]. Todėl šiame darbe nnU-Net pasirinktas kaip pagrindas tolesniems metodams kurti.

S.2. Metodų taikymas

S.2.1. Lyginti prostatos vėžio segmentavimo ansambliai

Visuose ansambliuose naudojami nnU-Net metodu apmokyti DNN modeliai. Prostatos vėžiui segmentuoti įvestis modifikuojama už prostatos zonos esančius vokselius nustatant į nulį. mpMRT modalumai pateikiami kaip atskiri įvesties kanalai. Siūlomi ansambliai yra:

- 1 $ansamblis = PCa_{seg}(T2W \times TZ) \vee PCa_{seg}((ADC, DWI) \times PZ)$;
- 2 $ansamblis = PCa_{seg}(Visos \times TZ) \vee PCa_{seg}(Visos \times PZ)$;
- 3 $ansamblis = PCa_{seg}(Visos \times TZ) \vee PCa_{seg}(Visos \times PZ) \vee PCa_{seg}((Visos, Zonos) \times prostata)$;
- 4 $ansamblis = PCa_{seg}(Visos \times PZ) \vee PCa_{seg}((Visos, Zonos) \times prostata)$.

Ansamblių apibrėžimuose $PCa_{seg}(x)$ žymi prostatos vėžio segmentavimo modelį su įvestimi x. $T2W$, ADC ir DWI nurodo mpMRT modalumus, o žymėjimas $Visos$ – visus 3 mpMRT modalumus. PZ, TZ žymi prostatos zonų binarines kaukes, $Zonos$ – abiejų zonų daugiaklases kaukes, o $prostata$ žymi binarinę visos prostatos kaukę. $PCa_{seg}((Visos, Zonos) \times prostata)$ yra laikomas baziniu šio tyrimo DNN modeliu.

S.2.2. Lyginti prostatos zonų segmentavimo sprendimai

Abiejose pasiūlytose prostatos zonų segmentavimo veiksmų sekose naudojami DNN modeliai, apmokyti taikant nnU-Net metodą. DNN modelių įvestį sudaro T2W modalumas. Vieno periferinės ir tranzitorinės zonos segmentavimo modelio (angl. *Single Peripheral Zone and Transitional Zone Segmentation Model*, SPZTZSM) veiksmų seka šioje disertacijoje yra laikoma bazine. Siūlomos veiksmų sekos:

- Atskiri modeliai tranzitorinei ir periferinei zonai segmentuoti (angl. *Separate Models for the Transitional Zone and Peripheral Zone Segmentation*, SM4TZPZS) – šioje darbo eigoje sujungiamas modelio, segmentuojančio TZ, rezultatas su modelio, segmentuojančio PZ, rezultatu į daugiaklasį tūrinį žemėlapį.
- Prostatos segmentavimas, tranzitorinės zonos segmentavimas ir periferinės zonos apskaičiavimas (angl. *Prostate Segmentation, Transitional Zone Segmentation, and Peripheral Zone Calculation*, PSTZSPZC) – šioje veiksmų sekoje pirmiausia naudojamas DNN modelis, skirtas segmentuoti visą prostatą. Paskui taikomas DNN modelis tranzitorinei zonai segmentuoti tik gautoje prostatos segmentacijoje, o periferinė zona gaunama iš bendros prostatos srities atėmus TZ.

Originalus no-new U-Net, naudojantis U-Net standartinę architektūrą, šioje disertacijoje yra laikomas bazine architektūra. Taip pat šioje disertacijoje siūloma no-new U-Net metode naudojamą U-Net architektūrą pakeisti dėmesio mechanizmą turinčiomis DNN architektūromis prostatos zonoms segmentuoti:

- tinklinio dėmesio U-Net [56, 69];
- viengubo dėmesio U-Net [56, 69];
- daugiamačio dėmesio U-Net [56, 69];
- dvigubo dėmesio U-Net [66].

S.2.3. Segmentavimo tikslumo vertinimo matai

Pagrindinis naudojamas matas yra Dice įverčio koeficientas (angl. *Dice Score Coefficient*, DSC), pasiūlytas Klein et al. [41]. Be to, šioje disertacijoje siūlomas modifikuotas DSC matas, DSC_τ, leidžiantis įvertinti kliniškai priimtiną ribų neapibrėžtumą. Abu šie matai pateikti lygtyse:

$$DSC_\tau = \frac{2TP}{2TP + FN_\tau + FP_\tau} \qquad DSC = \frac{2TP}{2TP + FN + FP} \tag{S.1}$$

FN_τ (toleruojamos negatyvios klaidos) skaičiuojamos plečiant sprendimo segmentacijos tūrį, o FP_τ (toleruojamos pozityvios klaidos) – plečiant anotacijos tūrį. Plėtimo operacija [79] atliekama su 5 mm žiedo

formos branduoliu. Taip pat skaičiuojamas klaidingai teigiamų pacientų procentas, rodantis pacientų dalį, kuriems modelis ar ansamblis aptinka vėžį, nors anotacijoje jo nėra.

Rezultatų skirtumams tarp prostatos zonų segmentavimo veiksmų sekų, DNN architektūrų ir prostatos vėžio segmentavimo ansamblių vertinti taikomi Friedmano testai [27]. Nustačius reikšmingus skirtumus, atliekami post hoc testai: Conoverio testai DNN architektūroms ir prostatos vėžio segmentavimo ansambliams [20], o prostatos zonų segmentavimo veiksmų sekų rezultatams lyginti – Wilcoxono suderintų rangų testai [82]. P-reikšmės koreguojamos Benjamini-Hochbergo metodu. Skirtumai tarp DSC ir DSC_τ vertinami dvipusiu Wilcoxono testu. Visuose testuose taikomas 95 % pradinis pasikliovimo lygmuo (reikšmingumo lygmuo – 0,05).

S.3. Duomenų ir eksperimento eiga

Prostatos vėžio segmentavimo eksperimentui naudoti T2W, ADC ir DWI vaizdai su Nacionalinio vėžio centro pateiktomis prostatos, biopsijų ir vėžinių regionų kaukėmis. Duomenys papildyti prostatos zonų kaukėmis taikant SM4TZPZS veiksmų seką, o vėžiniai regionai atrinkti, atmetant regionus, esančius >5 mm nuo biopsijos arba su ja nesikertančius. Galutinę imtį sudaro 119 atvejų. Taikyta stratifikuota 5 dalių kryžminė validacija (angl. *5 fold cross-validation*), užtikrinant pacientų su vėžiniais regionais pasiskirstymą. Apmokymai vykdyti su paketu dydžiu, lygiu 4 ir 1000 epochų. DSC ir DSC_τ skaičiuoti 42 pacientams su vėžiniais regionais, o pozityviai klaidingų pacientų procentas – likusiems 77 pacientams.

Prostatos zonų segmentavimo eksperimentams naudoti T2W vaizdai iš dviejų šaltinių: Decathlono (31 atvejis) ir ProstateX (98 atvejai) [3, 49]. Iš ProstateX 88 atvejai skirti mokyti, 10 – testuoti, o Decathlono duomenys naudoti papildomai testuoti siekiant įvertinti generalizaciją tarp skirtingų duomenų šaltinių. Apmokymai vykdyti su paketu dydžiu, lygiu 2 ir 1000 epochų. Tikslumas vertinamas DSC matu kiekvienai prostatos zonai, PZ ir TZ, atskirai.

S.4. Rezultatai

S.4.1. Prostatos zonų segmentavimo rezultatai

Vertinant atskirai kiekvienos imties prostatos zonų segmentavimo rezultatų (DSC) skirtumų statistinį reikšmingumą, atlikti du Friedmano testų rinkiniai. Pirmasis rinkinys taikytas atskiroms veiksmų sekoms, siekiant nustatyti statistiškai reikšmingus skirtumus tarp DNN architektūrų. Rezultatai pateikiami S.1 lentelėje, kur nurodoma prostatos zona, analizuojama veiksmų seka ir atitinkamai gautos p-reikšmės kiekvienai testavimo imčiai. Kadangi visos p-reikšmės, gautos naudojant ProstateX testavimo imtį, nesiekė pasirinkto reikšmingumo lygio, nėra pagrindo atmesti nulinę hipotezę. Taip pat Conoverio post hoc testų, atliktų su Decathlon testavimo imtimi, rezultatų analizė rodo, kad nė viena DNN architektūra statistiškai reikšmingai neišsiskiria iš kitų. Todėl statistiškai reikšmingų skirtumų tarp nagrinėtų DNN architektūrų rezultatų nenustatyta.

S.1 lentelė. Friedmano testų rezultatų p-reikšmės. Šie testai vertina, ar tarp lyginamų DNN architektūrų yra statistiškai reikšmingų skirtumų.

Prostatos zona	Veiksmų seka	ProstateX	Decathlon
TZ	SM4TZPZS	0,524931	0,041336
	PSTZSPZC	0,308441	0,120543
	SPZTZSM	0,619754	0,001645
PZ	SM4TZPZS	0,205203	0,050347
	PSTZSPZC	0,564541	0,036494
	SPZTZSM	0,864760	0,000074

Antrasis Friedmano testų rinkinys taikytas kiekvienai DNN architektūrai atskirai, siekiant įvertinti skirtumus tarp prostatos zonų segmentavimo veiksmų sekų. Rezultatai pateikiami S.2 lentelėje, kurios struktūra tokia pati kaip S.1 lentelėje, tačiau vietoje veiksmų sekos nurodoma nnU-Net metode naudojama DNN architektūra.

Kadangi visų testų su ProstateX imtimi p-reikšmės viršijo reikšmingumo lygį, nulinė hipotezė neatmetama, tad statistiškai reikšmingų skirtumų tarp veiksmų sekų rezultatų nenustatyta. Analogiškai keturių iš penkių TZ segmentavimo testų p-reikšmės taip pat viršija reikšmin-

S.2 lentelė. Friedmano testų rezultatų p-reikšmės. Šie testai naudojami įvertinti, ar tarp lyginamų prostatos zonų segmentavimo veiksmų sekų yra statistiškai reikšmingų skirtumų.

Prostatos zona	DNN architektūra	ProstateX	Decathlon
TZ	Tinklinio dėmesio U-Net	0,904837	0,0672055
	Viengubo dėmesio U-Net	0,272532	0,0145040
	Daugiamačio dėmesio U-Net	0,406570	0,2385126
	Dvigubo dėmesio U-Net	0,904837	0,9048374
	U-Net	0,272532	0,2725318
PZ	Tinklinio dėmesio U-Net	0,082085	0,0003465
	Viengubo dėmesio U-Net	0,406570	0,0000019
	Daugiamačio dėmesio U-Net	0,301194	0,0000003
	Dvigubo dėmesio U-Net	0,740818	0,0000551
	U-Net	0,904837	0,0000412

gumo lygį, todėl skirtumai laikomi nereikšmingais. O visų PZ segmentavimo testų p-reikšmės nesiekia reikšmingumo lygio, tad nustatyti statistiškai reikšmingi skirtumai. Todėl su Decathlono rinkinio PZ segmentavimo rezultatais atliekami Wilcoxono suderintų rangų post hoc testai, siekiant nustatyti, kurios veiksmų sekos lemia šiuos skirtumus. Šių testų rezultatai pateikti S.3 lentelėje, kur stulpelis „DNN architektūra“ nurodo DNN architektūrą, stulpelis „Lyginimas“ nurodo, kurios prostatos zonų segmentavimo veiksmų sekos yra lyginamos, o stulpelis „p-reikšmė“ nurodo p-reikšmę, gautą atliekant Wilcoxono suderintų rangų post hoc testą. Rezultatai rodo, kad PSTZ-SPZC rezultatai statistiškai reikšmingai skiriasi nuo kitų veiksmų sekų rezultatų, nes visos p-reikšmės nesiekia reikšmingumo lygio. O lyginant SPZTZSM ir SM4TZPZS rezultatus gauti prieštaringi rezultatai: su dvigubo dėmesio U-Net ir U-Net nustatyti reikšmingi skirtumai, o su kitomis DNN architektūromis – ne, todėl statistiškai reikšmingo skirtumo tarp SM4TZPZS ir bazinės SPZTZSM veiksmų sekos rezultatų nustatyti negalima.

S.4 lentelėje pateikiami agreguoti PZ segmentavimo Decathlono rinkinyje rezultatai. Nurodoma naudota DNN architektūra, prostatos zonų segmentavimo veiksmų seka bei pagrindinės statistikos: vi-

S.3 lentelė. Wilcoxono suderintų rangų post hoc testų rezultatai. Šie testai rodo, kurios prostatos zonų segmentavimo veiksmų sekos lemia statistiškai reikšmingus skirtumus segmentuojant periferinę zoną. Rezultatai apskaičiuoti naudojant Decathlono testavimo rinkinio rezultatus.

DNN architektūra	Lyginimas	p-reikšmė
Tinklinio dėmesio U-Net	SPZTZSM - SM4TZPZS	0,3883800
	SPZTZSM - PSTZSPZC	0,0007822
	SM4TZPZS - PSTZSPZC	0,0030005
Viengubo dėmesio U-Net	SPZTZSM - SM4TZPZS	0,0694814
	SPZTZSM - PSTZSPZC	0,0001776
	SM4TZPZS - PSTZSPZC	0,0000012
Daugiamačio dėmesio U-Net	SPZTZSM - SM4TZPZS	0,0062173
	SPZTZSM - PSTZSPZC	0,0025741
	SM4TZPZS - PSTZSPZC	0,0000018
Dvigubo dėmesio U-Net	SPZTZSM - SM4TZPZS	0,0005177
	SPZTZSM - PSTZSPZC	0,0374298
	SM4TZPZS - PSTZSPZC	0,0004353
U-Net	SPZTZSM - SM4TZPZS	0,0304844
	SPZTZSM - PSTZSPZC	0,000094
	SM4TZPZS - PSTZSPZC	0,000059

durkis, standartinis nuokrypis (angl. *standard deviation*, std), mediana ir medianinis absoliutus nuokrypis (angl. *Median Absolute Deviation*, MAD). Rezultatai rodo, kad didžiausias DSC pasiekiamas naudojant SM4TZPZS veiksmų seką.

S.4.2. Prostatos vėžio segmentavimo rezultatai

Šiame skyriuje dėstomi prostatos vėžio segmentavimo eksperimentų rezultatai. Pirmiausia S.5 lentelėje pateikiami surinkti DSC ir DSC_τ vidurkiai ir standartiniai nuokrypiai. Stulpelis „Ansamblis / modelis“ nurodo siūlomą ansamblį ir bazinį modelį, stulpelis „DSC“ – agreguotus DSC rezultatus, o stulpelis „DSC_τ“ – agreguotus DSC_τ rezultatus. Iš rezultatų aiškėja, kad didžiausią tikslumą pasiekia 3 ansamblis, po jo – 4 ansamblis. Be to, agreguoti duomenys rodo, kad DSC_τ verti-

S.4 lentelė. Aritmetinis vidurkis, standartinis nuokrypis, mediana ir MAD reikšmės apskaičiuotos iš periferinės zonos segmentavimo rezultatų, gautų naudojant Decathlono testavimo rinkinį.

DNN architektūra	Veiksmų seka	Vidurkis +/- std	Mediana +/- MAD
Tinklinio dėmesio U-Net	SM4TZPZS	**0,6794+/-0,1050**	**0,6975+/-0,1101**
	PSTZSPZC	0,5962+/-0,2169	0,6413+/-0,1010
	SPZTZSM	0,6735+/-0,1113	0,6778+/-0,1317
Viengubo dėmesio U-Net	SM4TZPZS	**0,6909+/-0,1129**	**0,7088+/-0,1428**
	PSTZSPZC	0,5775+/-0,2438	0,6443+/-0,1181
	SPZTZSM	0,6751+/-0,1056	0,6898+/-0,1480
Daugiamačio dėmesio U-Net	SM4TZPZS	**0,692+/-0,11170**	**0,7098+/-0,1277**
	PSTZSPZC	0,5725+/-0,2452	0,6366+/-0,1023
	SPZTZSM	0,6281+/-0,1510	0,6503+/-0,1104
Dvigubo dėmesio U-Net	SM4TZPZS	**0,6806+/-0,1265**	**0,7205+/-0,1261**
	PSTZSPZC	0,5221+/-0,2635	0,6322+/-0,0973
	SPZTZSM	0,6395+/-0,1323	0,6732+/-0,1134
U-Net	SM4TZPZS	**0,6978+/-0,1075**	**0,7249+/-0,1766**
	PSTZSPZC	0,5376+/-0,2643	0,6333+/-0,0994
	SPZTZSM	0,6684+/-0,1165	0,6692+/-0,1208

na palankiau už įprastą DSC. Tai patvirtina, kad DSC_τ metrika apima leistinus segmentavimo ribų nuokrypius ir yra atsparesnė nedideliems kontūrų neatitikimams.

S.5 lentelė. Prostatos vėžio segmentavimo rezultatų vidurkis ir standartinis nuokrypis (DSC_τ ir DSC).

Ansamblis / modelis	DSC	DSC_τ
1 ansamblis	0,1613 +/- 0,2271	0,2116 +/- 0,3015
2 ansamblis	0,1828 +/- 0,2381	0,2512 +/- 0,3269
3 ansamblis	**0,2264 +/- 0,2611**	**0,2977 +/- 0,3443**
4 ansamblis	0,2155 +/- 0,2580	0,2829 +/- 0,3395
Bazinis modelis	0,1787 +/- 0,2387	0,2360 +/- 0,3152

Friedmano testai ($p = 0,003107$ su DSC_τ ir $p = 0,00067$ su DSC)

leidžia atmesti nulinę hipotezę, nes parodo reikšmingus skirtumus tarp ansamblių ir bazinio modelio rezultatų. Todėl taikomi Conoverio post hoc testai, o jų rezultatai pateikti S.6 lentelėje, kur stulpelis „Lyginimas“ nurodo, kurie ansambliai ar bazinis modelis yra lyginami. Stulpeliai „p-reikšmė“ nurodo testo rezultato p-reikšmę, skliausteliuose nurodomas naudotas matas. Kadangi p-reikšmės nesiekia reikšmingumo lygio, nustatyti statistiškai reikšmingi skirtumai tarp bazinio modelio ir 3 ansamblio bei 4 ansamblio rezultatų, tačiau tarp bazinio modelio ir 1 ansamblio bei 2 ansamblio rezultatų skirtumų nenustatyta. Taip pat reikšmingų skirtumų nėra tarp 3 ansamblio ir 4 ansamblio rezultatų bei tarp 1 ansamblio ir 2 ansamblio rezultatų. Vis dėlto statistiniai testai su DSC įverčiais rodo, kad 1 ansamblis ir 2 ansamblis reikšmingai skiriasi nuo 3 ansamblio ir 4 ansamblio, o testai su DSC_τ įverčiais tokių skirtumų neatskleidžia.

S.6 lentelė. Conoverio post hoc testų rezultatai prostatos vėžio segmentavimo užduotyje. Kiekvienai pateiktai p-reikšmei naudotas matas nurodytas skliaustuose.

Lyginimas	p-reikšmė (DSC)	p-reikšmė (DSC_τ)
1 ansamblis - 2 ansamblis	0,7059	0,5352
1 ansamblis - 3 ansamblis	**0,0079**	**0,0189**
1 ansamblis - 4 ansamblis	**0,0179**	0,0656
1 ansamblis - Bazinis modelis	0,7059	0,5758
2 ansamblis - 3 ansamblis	**0,0148**	0,0741
2 ansamblis - 4 ansamblis	**0,0462**	0,2490
2 ansamblis - Bazinis modelis	0,6042	0,2533
3 ansamblis - 4 ansamblis	0,7059	0,5710
3 ansamblis - Bazinis modelis	**0,0043**	**0,0069**
4 ansamblis - Bazinis modelis	**0,0096**	**0,0189**

1 ansamblis, 2 ansamblis, 3 ansamblis, 4 ansamblis ir bazinis modelis atitinkamai pasiekia 24,68 %, 28,57 %, 35,06 %, 27,27 % ir 19,48 % pozityviai klaidingų pacientų. Šis matas rodo, kad bazinis modelis segmentuoja prostatos vėžį darydamas mažiausiai klaidų.

Wilcoxono suderintų rangų testų, nustatančių skirtumo tarp DSC ir DSC_τ statistinį reikšmingumą, rezultatų p-reikšmės yra 0,000132, 0,000027, 0,000012, 0,000027 ir 0,000132, gautos atitinkamai su 1 an-

sambliu, 2 ansambliu, 3 ansambliu, 4 ansambliu ir baziniu modeliu. Kadangi visos p-reikšmės nesiekia reikšmingumo lygio, nulinė hipotezė atmetama, o tai rodo statistiškai reikšmingą skirtumą tarp lyginamų matų.

Šioje disertacijoje pasiūlyti nauji giliojo mokymosi metodai prostatos zonoms ir prostatos vėžiui segmentuoti. Konkrečiai disertacijoje atlikta: (a) šiuolaikinių prostatos zonų ir prostatos vėžio segmentavimo metodų apžvalga, atskleidžiant no-new U-Net metodo patikimumą medicininių vaizdų segmentavimo uždaviniuose; (b) pasiūlytas metodas dėmesio mechanizmams integruoti į nnU-Net metodą, keičiant įprastą U-Net architektūrą į dėmesio mechanizmus naudojančias architektūras; (c) pasiūlytos kelios prostatos zonų segmentavimo veiksmų sekos, pagrįstos nnU-Net metodu; (d) palyginti pasiūlytų veiksmų sekų ir nnU-Net modifikacijų segmentavimo tikslumai kartu su šiuo metu naudojamo prostatos zonų segmentavimo sprendimo tikslumu; (e) pasiūlyti keli prostatos vėžio segmentavimo ansambliai, naudojantys no-new U-Net metodą ir prostatos zonų informaciją; (f) palyginti pasiūlytų ansamblių segmentavimo tikslumai kartu su šiuolaikiniais prostatos vėžio segmentavimo sprendimais; (g) pasiūlyta DSC modifikacija, DSC_τ, kuri atsižvelgia į didelę tarpstebėtojinę variaciją prostatos vėžio anotacijose; (h) palygintas pasiūlytas prostatos vėžio segmentavimo rezultatų vertinimo matas, DSC_τ, su standartiniu DSC matu. Pagrindinės išvados:

1. Šiame darbe pateikti rezultatai rodo, kad klasikinį U-Net pakeitus dėmesio mechanizmais paremtomis U-Net variacijomis nnU-Net metode, statistiškai reikšmingų segmentavimo rezultatų skirtumų nepastebima nei vertinant bendrą tikslumą, nei generalizuojamumą tarp skirtingų šaltinių duomenų rinkinių. Atsižvelgiant į tai, kad dėmesio mechanizmai padidina skaičiavimo sąnaudas, klasikinė U-Net architektūra išlieka praktiškiausias pasirinkimas nnU-Net metode.

2. Pasiūlyta prostatos zonų segmentavimo veiksmų seka, SM4TZPZS, pasižymėjo geresniu generalizuojamumu tarp skirtingų šaltinių duomenų rinkinių už kitas lygintas veiksmų sekas ir pasiekė statistiškai reikšmingai aukštesnį segmentavimo tikslumą už PSTZ-SPZC. Nors statistiškai reikšmingo skirtumo su SPZTZSM nenustatyta, SM4TZPZS daugumoje generalizuojamumo vertinimų nuosekliai prilygo bazinei veiksmų sekai arba ją pranoko. DSC vidurkio gerėjimai siekė nuo 0,3 % iki 5,2 % TZ srityje ir nuo 0,6 %

iki 6,4 % PZ srityje tarp visų palygintų DNN architektūrose. DSC medianos gerėjimai svyravo nuo 0,1 % iki 2,6 % (TZ) ir nuo 1,9 % iki 5,9 % (PZ), o tik dviem atvejais fiksuotas nežymus mažėjimas – 0,02 % (DSC mediana) ir 1,3 % (DSC vidurkis).

3. 4 ansamblis identifikuotas kaip tinkamiausias sprendimas prostatos vėžio segmentavimo uždaviniui, statistiškai reikšmingai pranokęs bazinį modelį, 1 ansamblį ir 2 ansamblį pagal agreguotus DSC ir DSC_τ matus nuo 3,17 % iki 7,13 %. Nors 3 ansamblis pasiekė 1 % aukštesnius agreguotus matus, skirtumas nebuvo statistiškai reikšmingas, o 4 ansamblis pasižymėjo reikšmingai mažesniu klaidingai teigiamų pacientų skaičiumi. Kadangi pagrindinis vertinimo kriterijus buvo segmentavimo tikslumas, kitų metodų klaidingai teigiamų pacientų skaičius nebuvo laikomas reikšmingu veiksniu. Be to, 4 ansamblis susideda iš dviejų modelių, o 3 ansamblis – iš trijų.

4. Pasiūlytas matas, DSC_τ, atsižvelgia į prostatos vėžio anotacijų tarpstebėtojinę variaciją. Gauti rezultatai rodo, kad DSC_τ nuosekliai suteikia statistiškai reikšmingai aukštesnes įverčių reikšmes už DSC, o tai leidžia daryti prielaidą, kad šis matas bent iš dalies kompensuoja anotacijų paklaidą. Visų vertintų ansamblių bei bazinio modelio atveju DSC_τ vidurkis viršijo DSC vidurkį nuo 5,03 % iki 7,12 %, o medianų skirtumas svyravo nuo 0,09 % iki 0,79 %. Bazinio modelio ir 1 ansamblio atveju DSC ir DSC_τ medianos sutapo.

NOTES

Aleksas Vaitulevičius
Research and Modeling of Digital Signals for the Detection of Cancerous Areas of Tissue in Different MRI Modalities
Doctoral Dissertation
Natural Sciences
Informatics (N 009)
Thesis Editor: Zuzana Šiušaitė

Aleksas Vaitulevičius
Skaitmeninių signalų tyrimas ir modeliavimas audinio vėžinėms zonoms aptikti skirtinguose MRT modalumuose
Daktaro disertacija
Gamtos mokslai
Informatika (N 009)
Santraukos redaktorė: Jorūnė Rimeisytė-Nekrašienė

Vilnius University Press
9 Saulėtekio Ave., Building III, LT-10222 Vilnius
Email: info@leidykla.vu.lt, www.leidykla.vu.lt
bookshop.vu.lt, journals.vu.lt
Print run of 20 copies