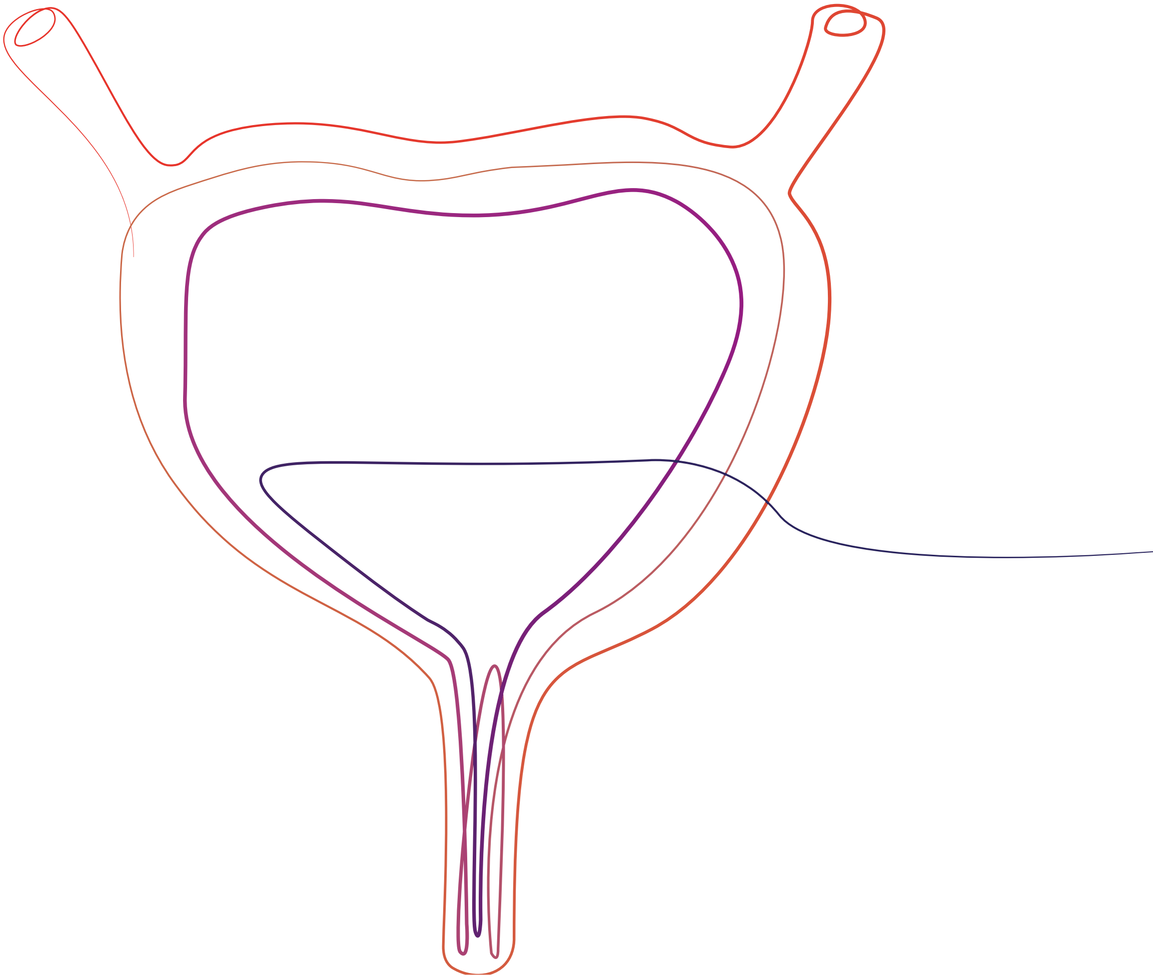
Artificial Intelligence based therapeutic response assessment of neoadjuvant immunotherapy for patients with bladder cancer

Joyce Greidanus



Artificial Intelligence in oncological imaging research

Therapeutic response assessment of neoadjuvant
immunotherapy for bladder cancer

J. GREIDANUS, BSc

Technical Medicine,
Track Medical Imaging & Intervention

M3 Internship, Department of Radiology at the Netherlands
Cancer Institute - Antoni van Leeuwenhoek (NKI-AVL)

October 7th, 2022

SUPERVISORS

Prof.dr. C. Brune	Chairman
Dr. S. Trebeschi	Technical supervisor, NKI-AVL
Z. Bodalal, MD MSc	Medical supervisor, NKI-AVL
Dr. T.N. Boellaard, MD PhD	Medical supervisor, NKI-AVL
Dr. C.O. Tan	Technical supervisor, University of Twente
Drs. A.G. Lovink	Process supervisor, University of Twente
Dr. E. Groot Jebbink	External member, University of Twente

ANTONI VAN LEEUWENHOEK
NEDERLANDS KANKER INSTITUUT

UNIVERSITY OF TWENTE.

Abstract

Introduction Clinical studies showed promising results for neoadjuvant immunotherapy for patients with muscle invasive bladder cancer (MIBC). Unfortunately, only a subset of patients responds, urging the quest for predictive image features. We hypothesise that Artificial Intelligence (AI) can automatically quantify predictive image features for immunotherapy response.

Patients and Method A retrospective study is performed in which n=79 patients from the PURE-01 study and n=19 patients from the NABUCCO study were included. These patients were diagnosed with MIBC (staged $\geq$ cT2N0M0), and received neoadjuvant immunotherapy, followed by a radical cystectomy. For each patient, we analysed two Magnetic Resonance Imaging (MRI) scans; the first one was acquired before neoadjuvant immunotherapy (baseline) and the second one after therapy (on-treatment). Pathological treatment response was divided between pathological Complete Response (pCR) (ypT0N0Mx) and non-pCR. A nnU-Net model was trained to, automatically detect tumour. Volumetric analysis was performed to determine its predictive value for therapeutic response. Radiological tumour characteristics were extracted and a Random Forest Classifier was trained to identify predictive image features. Finally, a Convolutional Neural Network (CNN) was trained to classify pathological outcomes.

Results Segmentation volume analysis showed a higher predictive performance for the pathological outcome for the on-treatment volumes, compared to the baseline volumes (respectively 0.91 AUC and 0.81 AUC). The predictive value for the radiomic features of baseline, on-treatment and difference over time was 0.49 AUC, 0.70 AUC and 0.74 AUC, respectively. The CNN models overfit on the training data, the highest AUC score (0.62) showed no significant predictive performance.

Conclusion The results show that the predictive performance for image features, obtained after treatment, is promising for pathological response classification. These features might be used to indicate organ-sparring treatment after neoadjuvant therapy. Predictions based on tumour features, derived from the radiological images before neoadjuvant immunotherapy, remain challenging.

Keywords: Artificial Intelligence, Neoadjuvant Immunotherapy, Radiological Imaging, Response Prediction

Acknowledgements

My gratitude goes to Prof. Dr. Regina Beets-Tan and Prof. Dr. Henk van der Poel for opening up their departments for my graduation internship. It was an honor to work with and learn from all the talented people in your departments.

Many thanks to Prof. Dr. Christoph Brune, for being the chairman of my committee. You always showed much interest in the project and gave me a different perspective on the research, which encouraged me to see the bigger picture. I'm extremely grateful for the daily supervision and guidance of Dr. Stefano Trebeschi. You helped me gain a better understanding in the broad field of Artificial Intelligence and scientific research. Thank you for your endless support, your patience and good ideas and suggestions. Even when I was really disappointed with some of the results, you kept me motivated and focused on the research goals. A special thanks to Zuhir Bodalal MD MSc and Dr. Thierry Boellaard for their medical supervision. Your profession has helped me gain new insights into immunotherapy pathways and the radiographic imaging analysis, which have been invaluable to the clinical relevance of this research. I would also like to thank Dr. Can Tan, for his technical supervision. You kept me focused on the statistical analysis and interpretation of my results, further enhancing the quality of my research. And a special thanks to Drs. Annelies Lovink, for your guidance and advice on my personal development. You always challenged me to ask myself deeper questions to stimulate my personal growth.

I would like like to acknowledge Dr. B. van Rhijn, Dr. L. Mertens and Dr. M. van de Kamp, of the Department of Urology, for their clinical guidance and especially their profession in bladder cancer. You helped me to develop myself in the clinical environment and to gain a better understanding of the clinical relevance of this research.

I had the pleasure of working with my amazing colleagues at the, always inspiring, O-building. You had to listen to all of my stories, whether I was happy and celebrating results or when I was mad and complaining, you were always supportive. In addition, your great knowledge of computer science has helped me further in my technical development.

For the great design of this thesis I would like to recognize Laura de Ridder, from Studio LAU. You helped me to worry a little less at the end of my graduation project by designing and finalising a great thesis lay-out.

Last but not least, I would like to thank my friends and family, especially my boyfriend Rob for always being an immense support over the past years. You have always believed in me and kept my motivation high.

Joyce Greidanus

Amersfoort
September 14, 2022

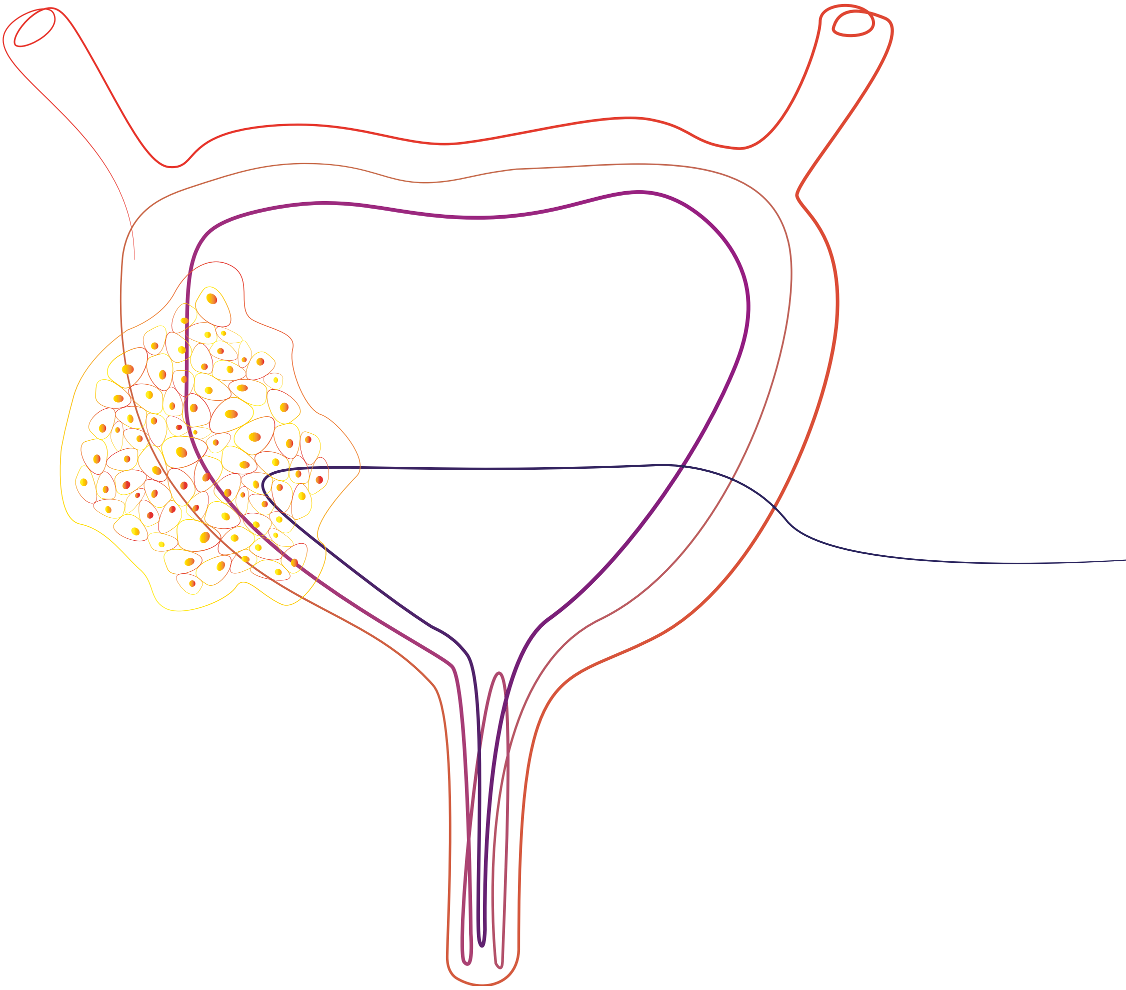
Contents

List of Abbreviations

ADC	**A**pparent **D**iffusion **C**oefficient
AI	**A**rtificial **I**ntelligence
AJCC	**A**merican **J**oint **C**ommittee on **C**ancer
ANN	**A**rtificial **N**eural **N**etwork
APC	**A**ntigen **P**resenting **C**ell
AUC	**A**rea **U**nder the **Curve**
CI	**C**onfidence **I**nterval
CNN	**C**onvolutional **N**eural **N**etwork
CT	**C**omputed **T**omography
CTLA-4	**C**ytotoxic **T-L**ymphocyte **A**ssociated protein 4
DCE	**D**ynamic **C**ontrast **E**nhanced
DL	**D**eep **L**earning
DWI	**D**iffusion **W**eighted **I**maging
GLCM	**G**ray **L**evel **C**o-occurence **M**atrix
GLDM	**G**ray **L**evel **D**ependence **M**atrix
GLRLM	**G**ray **L**evel **R**un **L**ength **M**atrix
GLSZM	**G**ray **L**evel **S**ize **Z**one **M**atrix
ICI	**I**mmune **C**heckpoint **I**nhibitors
MCCV	**M**onte **C**arlo **C**ross **V**alidation
MHC	**M**ajor **H**istocompatibility **C**omplex
MIBC	**M**uscle **I**nvasibe **B**ladder **C**ancer
ML	**M**achine **L**earning
mpMRI	**m**ulti **p**arametric **M**agnetic **R**esonance **I**maging
NGTDM	**N**eighbouring **G**ray **T**one **D**ifference **M**atrix
NMIBC	**N**on **M**uscle **I**nvasibe **B**ladder **C**ancer
pCR	**p**athological **C**omplete **R**esponse
PD-1	**P**rogrammed **D**eath receptor-1
PD-L1	**P**rogrammed **D**eath-**L**igand 1
ReLU	**R**ectified **L**inear **U**nit
RFC	**R**andom **F**orest **C**lassifier
ROC	**R**eceiver **O**perating **C**haracteristic
ROI	**R**egion **O**f **I**nterest
TCR	**T-C**ell **R**eceptor
TURBT	**T**rans **U**rothelial **R**esection of the **B**ladder **T**umour

Chapter 1

General Introduction



1.1 Introduction

The incidence of bladder cancer has increased in many European countries and morbidity and mortality are high.[1] The current, multimodal treatment for muscle invasive bladder cancer (MIBC), consists of transurethral resection of the bladder tumour (TURBT) and neoadjuvant chemotherapy, followed by surgical resection of the bladder (radical cystectomy).[2] Recent studies show that immunotherapy is a promising treatment for different types of cancer.[3–5] Its effectiveness has been confirmed in terms of long-term survival and reduction of toxicity, compared to traditional chemotherapy. Unfortunately, only a subset of the patients has a response to immunotherapy. Some patients may suffer from adverse effects of the treatment, without achieving any benefit. The NABUCCO study showed that 46% of the patients with MIBC had a pathological Complete Response (pCR) and 58% of the patients had no remaining invasive disease (<ypT2) after neoadjuvant treatment with immunotherapy.[6] In the PURE-01 study, 42% of the patients with MIBC had a pCR and downstaging of the tumour stage was achieved in 54% of the patients.[7] Neoadjuvant immunotherapy shows promising results, but to achieve optimal treatment, it should be decided, for each patient individually, whether immunotherapy is the most suitable option.

By predicting the patients' individual response to neoadjuvant immunotherapy before or during treatment, physicians can determine the most suitable course of action. Therefore, improving patients' quality of life and reducing unnecessary toxicity and costs.[8] Currently, histological- and molecular biomarkers are not sufficient to predict whether the patient will achieve any benefit from neoadjuvant immunotherapy. In addition, these biomarkers are obtained invasively, via tissue biopsies, and the prediction is only based on the pre-treatment status, while the inclusion of dynamic changes during therapy could increase the prediction efficiency.[7,9]

Medical imaging provides non-invasive, image features to determine tumour characteristics. With the use of Artificial Intelligence (AI), large amounts of data can be processed to detect and select image features that can predict therapeutic response.[8,10] AI has shown promising results for automatically quantifying predictive biomarkers for immunotherapy response.[11–13]

1.2 Clinical Background

Bladder cancer is the sixth most common malignant cancer among men, worldwide. The worldwide incidence is 9.6 per 100.000 persons per year for men and 2.4 for women. The overall European incidence is higher: 20.2 per 100.000 persons per year for men and 4.3 for women.[1] The morbidity and mortality are high.[1,14]

Bladder cancer is staged according to the TNM staging system, where Tumour (T) describes the invasion of the primary tumour into the bladder wall and its surrounding tissues, Node (N) describes if the tumour has spread to lymph nodes, and depends on number and location of the nodes. Lastly, Metastasis (M) describes if the tumour has spread to other organs. An overview of tumour staging in bladder cancer, according to the American Joint Committee on Cancer (AJCC) Staging System (8th edition, 2017), can be found in figure 1.1.[15] According to staging, bladder cancer can be divided into two groups; non-muscle invasive bladder cancer (NMIBC), which is staged as Ta/T1 and carcinoma *in situ* (CIS), requires only minimally invasive local treatment, and MIBC, which is staged as T2/T3/T4 and generally requires multimodal treatment[1,16,17]

1.2.1 Diagnosis

The diagnostic workflow for bladder cancer includes a cystoscopy, TURBT and radiological image assessment. During the cystoscopy, the bladder is viewed from the inside. Possible abnormalities can be seen and assessed based on the shape, texture, size and location. A urinary cytology sample can be drawn, to determine if there are tumour cells present within the urine. If abnormalities are observed during cystoscopy, a surgical procedure called TURBT is indicated. During the TURBT, tissue is obtained for histological assessment and tumour cells are removed to assess the invasiveness into the bladder wall.[13,18]

Radiological imaging is performed to assess whether cancer has spread outside the bladder wall, lymph nodes and other organs. In standard clinical practice, Computed Tomography (CT) scans are acquired for local staging and the assessment of metastasis. Magnetic Resonance Imaging (MRI) scans offers superior soft tissue contrast and provide more functional information, compared to CT. Nevertheless, a CT scan is still standard clinical

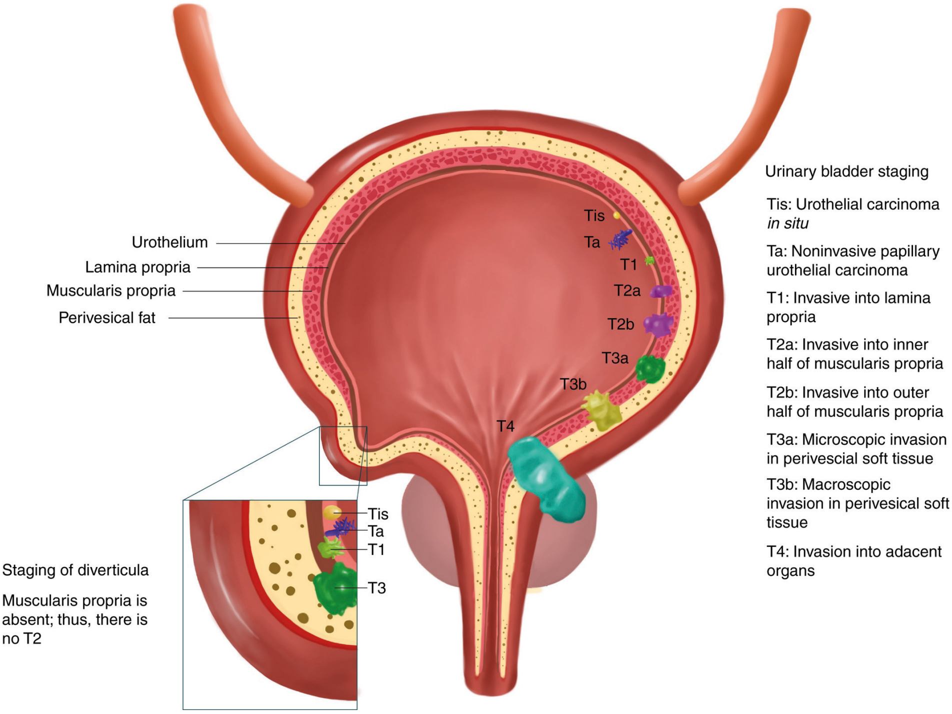

Urothelium
Lamina propria
Muscularis propria
Perivesical fat
Tis
Ta
T1
T2a
T2b
T3a
T3b
T4
Urinary bladder staging
Tis: Urothelial carcinoma *in situ*
Ta: Noninvasive papillary urothelial carcinoma
T1: Invasive into lamina propria
T2a: Invasive into inner half of muscularis propria
T2b: Invasive into outer half of muscularis propria
T3a: Microscopic invasion in perivescial soft tissue
T3b: Macroscopic invasion in perivesical soft tissue
T4: Invasion into adacent organs
Staging of diverticula
Muscularis propria is absent; thus, there is no T2
Tis
Ta
T1
T3

FIGURE 1.1: Overview of the staging of bladder cancer, according to the AJCC Staging System[15]

practice, because it is faster and more cost-effective than MRI.[19] For the scope of this study, pelvic multi-parametric MRI (mpMRI) scans were acquired. The tumour is assessed on T2-weighted images, diffusion-weighted images (DWI) and the dynamic contrast-enhanced (DCE) sequence (see figure 1.2). T2-weighted images are used for the morphological characteristics of the tumour, which is shown slightly hyperintense on this sequence. The DWI sequence is used to evaluate the cellular physiology: restriction of movement of water molecules, caused by hypercellularity of the tumour, will appear hyperintense on the DWI. The DWI should always be assessed together with the apparent diffusion coefficient (ADC) map, to determine whether there is true diffusion restriction, or that the hyperintensity is due to a high T2- signal also called T2-shine-through. True restricted diffusion is characterised by an increased signal on the DWI and a low signal on the ADC map. For the

DCE sequence, a contrast agent is administered intravenously, which will cause tissue enhancement on the MRI scan. Tumour lesions can be distinguished from other tissue because neoangiogenesis and tumour vascularization cause early enhancement. Primary bladder cancer starts to enhance after 6.5 ± 3.5 seconds, after the start of arterial enhancement.[20–23]

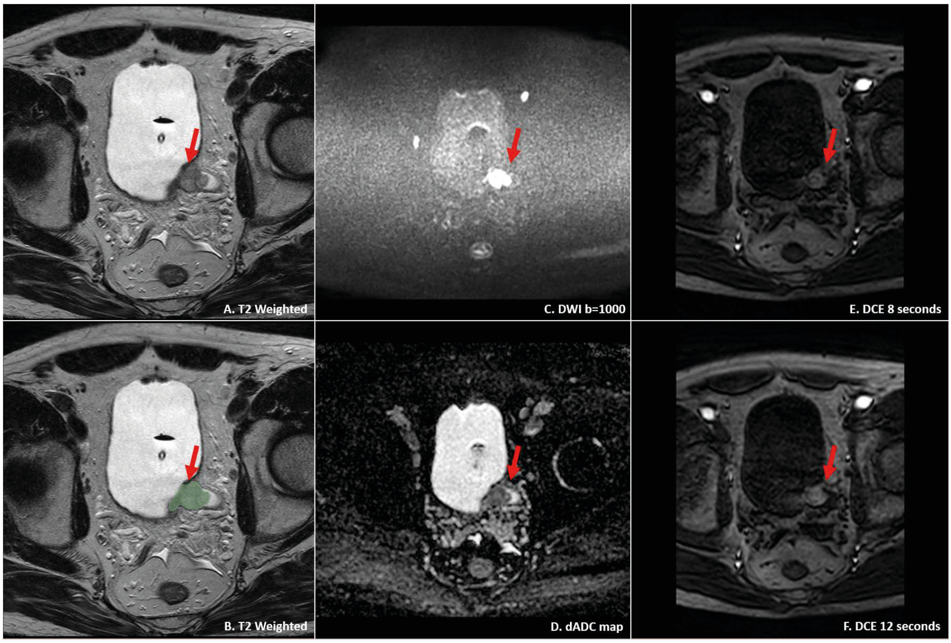


FIGURE 1.2: mpMRI acquisition of bladder cancer (stage T2) in a 75-year-old man. (A) The tumour is visible as a slightly hyperintense lesion in the right posterior bladder wall (red arrow) on an axial T2-weighted image. (B) An overlay of the tumour volume segmentation can be seen in green. True restricted diffusion is visible as the hyperintense lesion on the DWI (C) and the low signal on the corresponding ADC map (D). The DCE scans (E, 8 seconds; F, 12 seconds) show early enhancement of the tumour lesion.

1.2.2 Treatment

Treatment options depend on the type, stage and grade of the bladder cancer. For patients with NMIBC, intravesical chemo- or immunotherapy is given. With intravesical therapy, drugs are administered directly into the bladder. Therefore, this therapy only affects the cells lining the inside of the bladder and does not affect cells elsewhere. Bacillus Calmette-Guering (BCG) is the most common intravesical immunotherapy for non-invasive bladder cancer. Chemotherapy solutions (e.g. mitomycin, gemcitabine) are most often indicated when intravesical immunotherapy does not work.[24] For patients with high-grade NMIBC, the tumour can be fast-growing and the recurrence rate is high. Therefore BCG therapy might not be enough and chemo-radiation can be considered. Chemo-radiation therapy is not indicated if the tumour is located close to the ostia or the bladder neck and if CIS has been proven.

For patients with MIBC, radical cystectomy with bilateral pelvic lymph node dissection is currently the recommended standard. Neoadjuvant therapy might be indicated for systemic treatment, to reduce tumour volume and destroy microscopic cancer cells that might have spread beyond the bladder (which is not visible yet on medical imaging).[7] The current golden standard is neoadjuvant cisplatin-based chemotherapy, which shows a response rate of 22-40% pCR, but the overall survival benefit is only 5% at 5 years.[6] Neoadjuvant chemotherapy followed by a radical cystectomy has a high recurrence rate, metastases develop within two years after radical cystectomy in 50% of the patients.[13] Beside that, approximately 50% of the patients are ineligible to receive cisplatin-based chemotherapy, because of pre-existing contraindications; mainly impaired renal function.[7]

1.2.3 Immunotherapy for bladder cancer

Studies with neoadjuvant immunotherapy show promising results, regarding treatment response. The NABUCCO study[6] showed that 46% of the patients with MIBC had a pCR and 58% of the patients had no remaining invasive disease after neoadjuvant treatment with immunotherapy. Compared to neoadjuvant chemotherapy, neoadjuvant immunotherapy appears to give better and long-lasting responses, which is shown by the survival rate over time of patients treated with neoadjuvant immunotherapy, and a significant reduction of toxicity.[25]

The most widely used immunotherapy is immune checkpoint inhibitors (ICIs). It blocks specific immune interactions responsible for the suppression of an anti-tumour immune response. Tumours employ several strategies to mitigate the risk of detection and destruction by the immune system. Tumour cells suppress the host's immune system by expressing molecules, such as programmed death ligand 1 (PD-L1), which binds to programmed cell death protein 1 (PD-1), expressed on the surface of an (activated) T-cell. PD-1 binding to PD-L1 on the tumour surface blocks the activation of T-cells.[26] Besides, tumour cells also cause downregulation of specific molecules, such as the major histocompatibility complex (MHC). By downregulating MHC, recognition by and cytotoxicity of CD8+ T-cells is reduced.[27] Activation of T-cells starts the immune response cascade, which will eventually result in cell death. With the expression of PD-L1 and downregulation of MHC, the tumour suppresses this immune response cascade preventing cell death.

PD-1/PD-L1 and cytotoxic T-lymphocyte-associated protein 4 (CTLA-4) are inhibitory molecular checkpoints that are promising targets for the treatment of cancer.[28] PD-1/PD-L1 inhibitors are antibodies that block these molecule bindings, and therefore activate T-cell activation. CTLA-4 is expressed on the surface of an activated T-cell and has an inhibitory function. By introducing CTLA-4 antibodies, CTLA-4 is inhibited and therefore the regulation of the immune response will be increased. Activated T-cells will identify tumour cells and initiate a cytotoxic mechanism that will eventually result in tumour cell death (see Figure 1.3).[1]

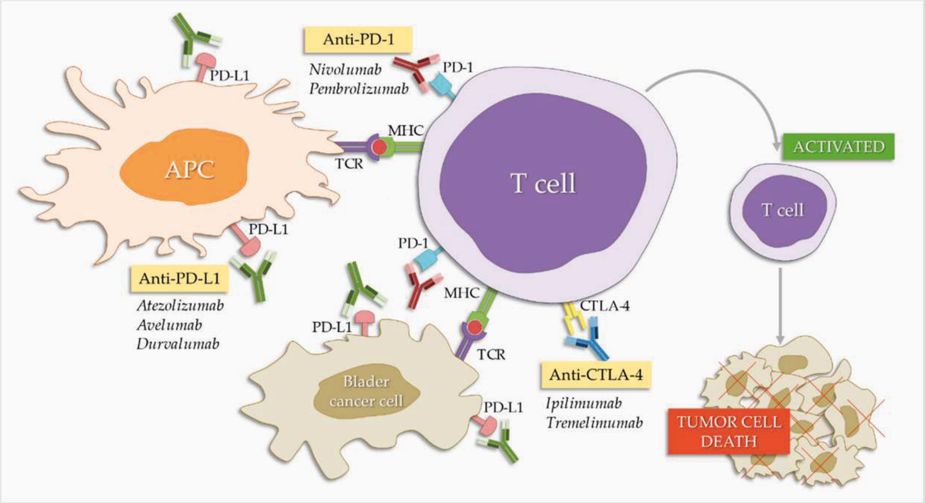


FIGURE 1.3: Effect of immune checkpoint inhibitors (ICIs). PD-1/PD-L1 and CTLA-4 block the inhibitory checkpoint that is responsible for immune suppression, leading to T-cell activation and therefore tumour cell death. APC= antigen-presenting cell; TCR = T-cell receptor.[1]

1.3 Technical Background

1.3.1 Artificial Neural Network

Machine Learning (ML) is used to create an algorithm that automatically extracts patterns out of data, to make predictions, e.g. classifications, segmentations, etc. There are several ML techniques, among which are Artificial Neural Networks (ANN). An ANN is an algorithm where the relationship of the input data is modelled by the interaction of multiple, stacked, non-linear layers of data processing:

$$x_l = f(x_{l-1}W_l + b_l)$$

where x is the data, W and b are the weights and biases, and l is the index of the layer.

Each input vector is multiplied with individual weights and bias and subsequently passed through an activation function, which will calculate the probability of the outcome (e.g. classification of diseases). The model is optimised using gradient descent:

$$W_{i+1} = W_i - \alpha \nabla L(W_i)$$

where W_i and W_{i+1} are the current and updated weights, respectively; L is the loss function, that estimates the error of the network prediction; and α is the learning rate, representing the size of the step to take in the opposite direction of the error.

A schematic representation of an ANN can be seen in Figure 1.4. The process of learning data relationships on these architectures with multiple layers is called Deep Learning (DL).

1.3.2 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are a class of DL algorithms, which is currently the state-of-the-art method for automatic image processing.[29] The model takes an image as input and through convolutions, the image is down-sampled to a highly informative feature representation. Features are extracted from the input image by the kernel. The kernel is a matrix (for example [3x3] or [5x5]) that slides over the image to detect features at a given spatial location, such as edges, corners, shape and size. Each kernel will have a corresponding

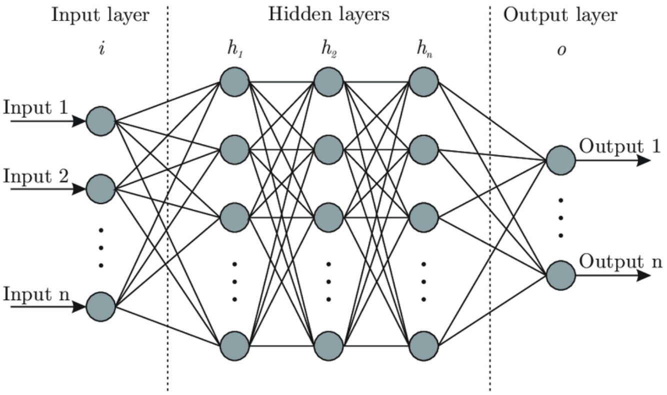


FIGURE 1.4: Artificial Neural Network, with the input data connected to all the nodes in the next hidden layers, resulting in an output

activation map. By adding multiple convolution layers and therefore increasing the number of kernels, more features can be identified - with the risk of overfitting: when a model performs too well on training data but is lacking performance on new, unseen data. After the convolutional layers, an activation function is applied to add non-linearity to the network, enabling CNN to learn complex relationships. The most often used activation function is the Rectified Linear Unit (ReLU), which converts all negative values from the filtered image to zero.

Pooling layers can be implemented to reduce the image spatial dimensionality of the activation map. After convolution and activation, the pooling layer takes the highest (max pooling) or average (average pooling) activation value in a small region of typically 2x2 voxels to be passed forward to the next layer. Pooling layers are used to reduce the spatial dimension of the input. Therefore the number of parameters and computation network will be reduced, which will also help in controlling overfitting.

Combinations of convolution-, activation- and pooling layers, can be arranged and repeated multiple times to create a deeper network and therefore model more complex relationships.

Finally, the output from the final convolutional layer (3D-matrix) is flattened to a 1D vector. This flattened vector is connected to a fully connected layer. The fully connected layer consists of nodes, where each node is connected to all nodes from the previous layer.

In the case of a classification or a segmentation problem, a softmax layer converts the results of the activation layer to a probability, which sums to 1. The sigmoid or softmax function can be used to calculate the probabilities out of network signals.

Just as described in the previous section, during training, the convolutional kernels (w) are learned through gradient descent. After each training epoch, the loss is calculated. The loss is defined as the difference between the predictions and their respective label. By using the derivative of the loss function, the loss can be minimised to optimise the network. During backpropagation, the gradient is calculated by taking the derivative of the loss function. The gradient is calculated through the chain rule, with respect to each weight of the network. With backpropagation, the weights and biases will be updated. The goal is to minimise the training loss, in order to optimise the model performance. The loss function should be set according to the application of the model.

1.4 Datasets

1.4.1 PURE-01

The PURE-01 study[7,30] was performed in two centres in Milan, Italy (IRCCS Istituto Nazionale dei Tumori and IRCCS San Raffaele Hospital). This study aimed to determine treatment response of neoadjuvant immunotherapy for patients with MIBC. Inclusion criteria for this study were predominant (ie, at least 50%) urothelial carcinoma histology and clinical T≥3bN0 stage tumour. The administered immunotherapy was pembrolizumab, given in three cycles of 200 mg every three weeks. After neoadjuvant therapy, a radical cystectomy was performed. The primary outcome of this study was pCR (pT0). The PURE-01 dataset included n=86 patients. For this thesis, n=14 patients were excluded due to the absence of MRI scans and/or sequences. This resulted in a total of n=72 patients from the PURE-01 dataset, of which n=28 patients (39%) had a pCR.

Images were acquired on a 1.5 Tesla MRI scanner (Philips MRI systems Ingenia). All mpMRI examinations were performed with bladder catheterization to allow for consistent bladder wall distension. The imaging protocol for mpMRI consisted of triplanar T2-weighted fast spin-echo sequences, DWI in transverse planes at four different b-values (0, 100, 500 and 1000 s/mm^2) and DCE sequences after injection of contrast agent (0.1 ml/Kg Gadovist) with a temporal resolution of 4/5 seconds. The general slice thickness was 3 mm.[7]

1.4.2 NABUCCO

The NABUCCO study[6] was performed at the NKI-AVL, with a multi-centre extension (Radboud UMC and UMC Utrecht). The primary outcome of this study was the number of patients that could have surgical resection 12 weeks after the study started. Secondary outcome measurement is the response rate after cystectomy according to pathological response criteria. Inclusion criteria according to the TNM-stage was high-risk resectable urothelial cancer, defined as stage III urothelial cancer: cT3-4aN0M0 or cT1-4aN1-3M0. The patients received three courses (similar to PURE-01) of ipilimumab and nivolumab combination as neoadjuvant immunotherapy, followed by a radical cystectomy. The NABUCCO study included 24 patients in cohort 1 and 30 patients in cohort 2. For this thesis, patients were excluded due to the absence of MRI scans/sequences. This resulted in a total of n=19 patients from the NABUCCO study, of which n=7 patients (41%) had a pathological complete response.

For mpMRI acquisition, two 3T MRI systems were used (Philips MR Systems Achieva dStream and Philips MRI Systems Ingenia). Before the acquisition, patients emptied their bladder and drank ± 0.5 litres of water, half an hour before the acquisition. The imaging protocol consisted of triplanar T2-weighted turbo spin echo sequences, DWI in transverse planes at four different b-values (0, 200, 800, 1000 s/mm^2) and DCE sequences after injection of contrast agent (Dotarem) with a temporal resolution of 6 seconds. The general slice thickness was 3 mm.[6]

For each patient, two MRI scans were acquired. The first scan was acquired before neoadjuvant immunotherapy, further referred to as the baseline scan. The second scan was acquired after neoadjuvant immunotherapy (and before radical cystectomy), further referred to as the on-treatment scan. See figure 1.5 for a schematic overview of the data collection timeline.

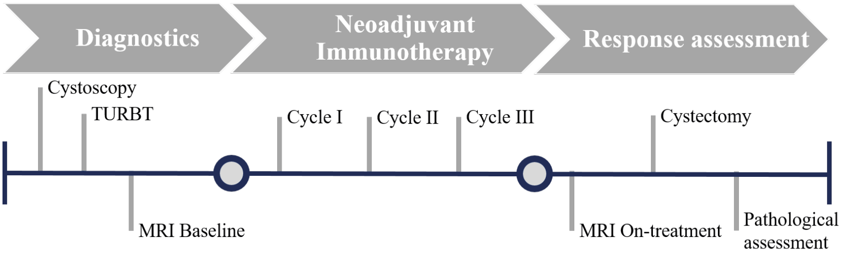


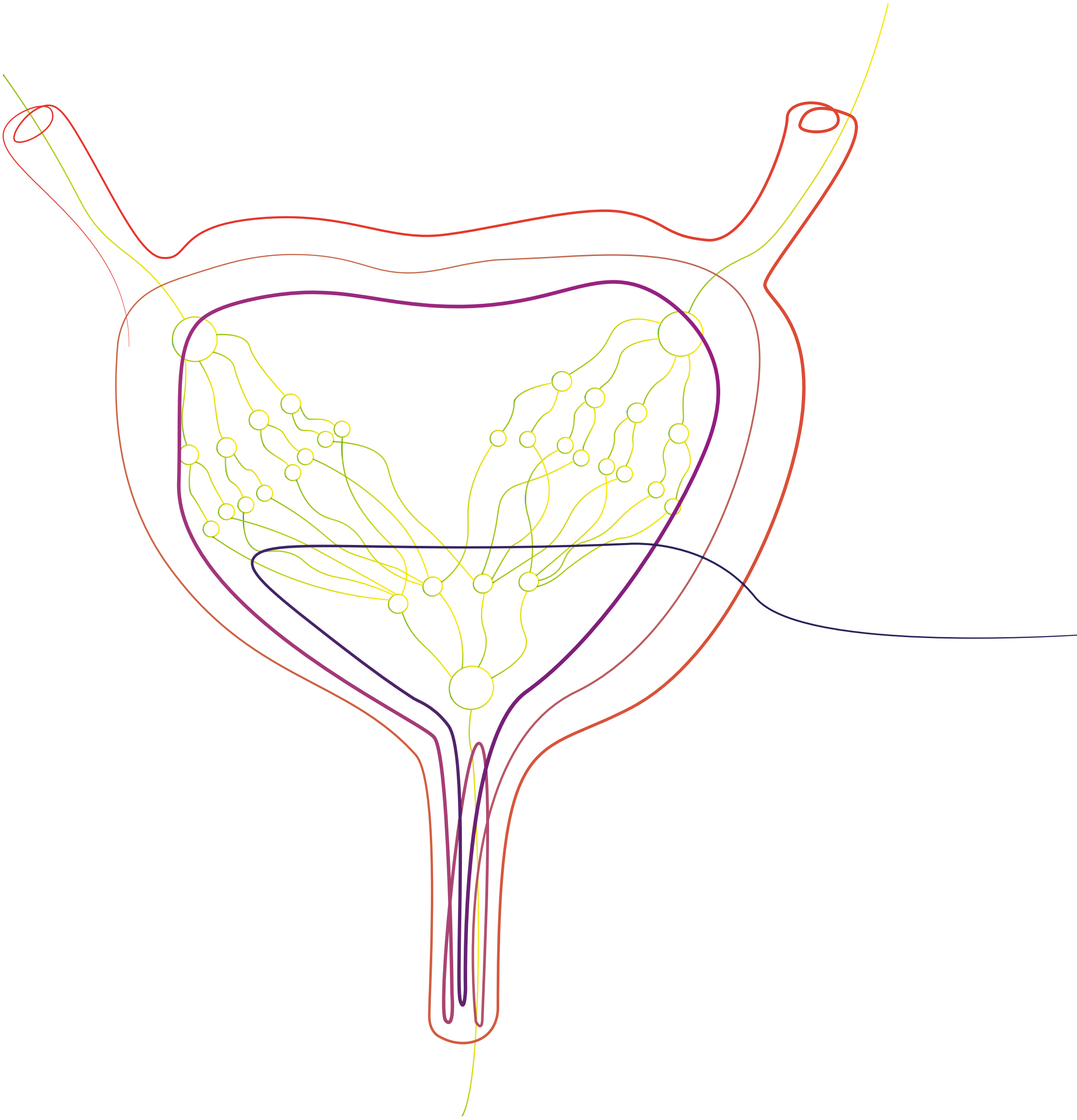
FIGURE 1.5: Timeline of data collection during the PURE-01 and NABUCCO study; starting with pre-treatment diagnostics, followed by three cycles of immunotherapy, a response assessment MRI and finally radical cystectomy with pathological assessment. TURBT= Transurethral Resection of Bladder Tumour; MRI = Magnetic Resonance Imaging

1.5 Research Aim and Thesis Outline

This thesis aimed to perform therapeutic response assessment and prediction of neoadjuvant immunotherapy for patients with MIBC. Clinical studies showed that neoadjuvant immunotherapy is a promising therapy, but until now there are no biomarkers that sufficiently predict whether a patient will benefit from neoadjuvant immunotherapy. The long-term clinical impact of this research is twofold: to prevent unnecessary toxicity in patients that will not benefit from neoadjuvant immunotherapy and, for organ preservation in patients that undergo complete response after receiving neoadjuvant treatment. To achieve this, image features from baseline and on-treatment mpMRI scans were identified and determined if these features had a predictive value for pathological response assessment. This will be presented threefold: in Chapter 2, we developed an automatic segmentation model for bladder cancer using a DL algorithm. Volumetric measures were determined to see if tumour volume and tumour volume changes might be predictive of therapeutic response. In Chapter 3, tumour-specific image features were extracted from mpMRI, based on the segmentation. Tumour characteristics were identified and, its predictive value for pathological treatment response was analysed by using an ML algorithm. Finally, in Chapter 4, an end-to-end CNN was implemented to automatically classify pathological treatment responses. In Chapters 2 and 3 the analysis was performed for tumour-specific characteristics that might be predictive of pathological response. While in Chapter 4, a classification model for the pathological outcome is presented.

Chapter 2

Automated detection of tumour



2.1 Introduction

Neoadjuvant immunotherapy showed promising treatment results for patients with muscle invasive bladder cancer (MIBC), compared to traditional neoadjuvant chemotherapy, but yet there is no reliable method to predict which patient is going to benefit from treatment.[6,7] As a result, a fraction of patients will receive ineffective therapy, causing unnecessary adverse effects and costs. Another fraction of patients will instead achieve pathological complete response (pCR) after neoadjuvant immunotherapy, meaning that no tumour cells are visible on pathological assessment of the resected bladder. Those patients could potentially forgo a surgical resection, which will increase their quality of life.

Tumour volume, estimated at the start or during treatment could be a reliable predictive biomarker of response.[16,31] Tumour volume delineation, when performed manually, by a radiologist, is time-consuming. Therefore, there is a need for an automatic segmentation algorithm of the bladder tumour on the radiological scans. This can be achieved with Deep Learning (DL).

The state-of-the-art segmentation model in Machine Learning (ML) is a nnU-Net.[32] nnU-Net is a deep-learning-based segmentation method, that automatises and optimises most of the image analytic pipeline, including pre-processing, network architecture, training and post-processing. By training an automatic segmentation model, segmentations can be obtained efficiently. Volume can be computed directly from these segmentations.

This chapter aims to investigate the predictive value of automatic DL bladder cancer segmentations on multi-parametric MR images of the pelvis.

2.2 Technical Background

The state-of-the-art model for medical image segmentation is nnU-Net. nnU-Net is a DL-based segmentation method, that automatizes the data pre- and post-processing pipeline. It provides an end-to-end automated pipeline, which can be used to train models with new data.[32] The automatic configuration is based on three parameter groups; fixed, rule-based and empirical parameters. The fixed parameters do not change between different datasets and describe for example the model's architecture, loss function, optimizer and data generation. Based on the data fingerprint, data-specific parameters are identified (e.g. image modality, shape, intensity distribution and distribution of spaces) as well as rule-based parameters (e.g. resampling, normalisation, batch and patch size). The fixed- and rule-based parameters together generate the pipeline fingerprint. Method configuration can be complex, since pipeline settings that are identified as optimal for one dataset, might not be suitable for other datasets. The Dice coefficient is used to monitor model performance during training.[32,33] The complete nnU-Net workflow can be found in figure 2.1.

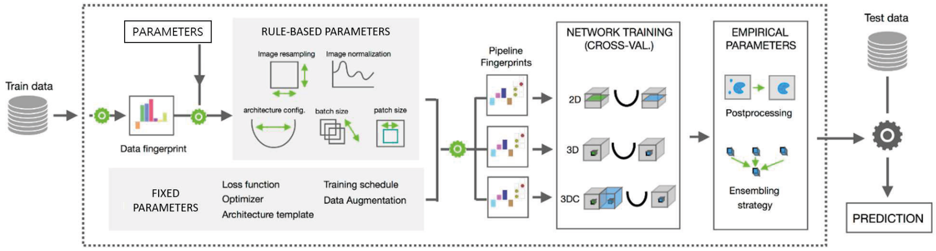


FIGURE 2.1: The workflow of nnU-Net; an automated configuration for DL-based biomedical image segmentation.[33]

A nnU-Net is based on the U-Net architecture. U-Nets are symmetrical network architectures, composed of an encoder and a decoder part. The output feature map of a convolutional layer in the encoder is copied over, and concatenated with the input feature map of the corresponding convolutional layer in the decoder.[34] Both encoder and decoder consist of a set of convolutional blocks. Where each block contains two 3x3 convolutions, followed by a ReLU and a max pooling operation. The pooling operation will reduce the spatial dimension. The opposite architecture is implemented in the decoder. At the lowest dimension space, the encoder produces a latent representation of the input, which is passed to the decoder. Because of the skip layers, spatial local information is added to the global abstract information and is therefore important for image segmentation. The U-Net architecture can be found in figure 2.2.

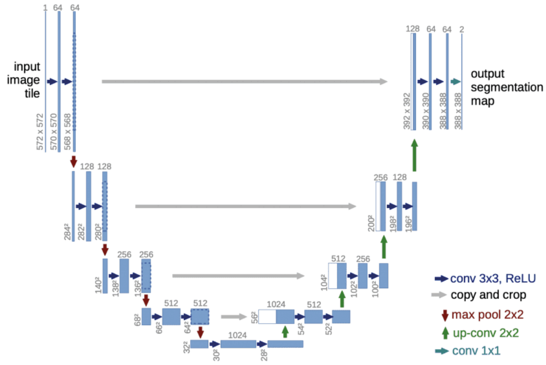


FIGURE 2.2: U-Net architecture, where each blue box corresponds to a multi-channel feature map. The number of channels is denoted on top of the box.

2.3 Method and Materials

2.3.1 Model Training

To study the applicability and limitations of DL-based segmentation in mpMRI scans of bladder cancer, multiple instances of the nnU-Net segmentation model were trained, each with different subsets of training data, from the PURE-01 and NABUCCO datasets (see section 1.4). Training input consisted of five sequences from each 3D scan; T2-weighted sequence, DWI with b-value 1000s/mm^2, ADC map, DCE at 6/8 seconds and DCE at 12 seconds after arterial contrast enhancement.

The selection of these sequences was based on radiological bladder cancer assessment on pelvic MRI, which is described in section 1.2.1. All models were trained with the nnU-Net 3D full-resolution configuration.

For this study, the segmentation labels were manually delineated by a radiologist, experienced in bladder imaging. This delineation was challenging due to the TURBT, which was performed before image acquisition. In some cases, a large part of the tumour volume was resected, with possibly only a few tumour cells left, hard to detect on the radiological images. Moreover, subtle evident tumour was hard to distinguish from the post-operative inflammatory response due to the TURBT.[30,35]

The radiologist was able to delineate a tumour-specific segmentation if any morphological anomaly was visible on the T2-weighted sequence, if any diffusion restriction was visible on the DWI sequence (which was confirmed by the ADC map) and if early contrast enhancement was shown on DCE sequence. In addition to these segmentations, the radiologist segmented a region of interest (ROI), corresponding to the area of the tumour bed, if there was no evident tumour visible. For model training, the ROI segmentations were included, but whenever there was a tumour-specific segmentation available, this one was included. The radiological segmentations are further referred to as ’ground truth’.

Four nnU-Net models were trained with different subsets of training data (see table 2.1).

1. The first model was trained on PURE-01 baseline scans (n=57) with the ground truth segmentations used as labels. The test set consisted of n=15 patients with PURE-01 baseline scans and n=72 patients with PURE-01 on-treatment scans. External validation was performed with the NABUCCO dataset (n=35 scans).
2. Second, a model was trained with a combination of PURE-01 baseline and on-treatment scans (n=57 patients for both baseline and on-treatment scans, so a total of 114 scans), again with the ground truth segmentations used as label. The test set consisted of the remaining n=15 patients from the PURE-01 dataset (both baseline and on-treatment scans, a total of 30 scans). External validation was performed with the NABUCCO dataset (n=35 scans). A training set with a combination of both baseline and on-treatment scans was investigated, to extend the training data with also scans that do not contain tumour.
3. The third model was trained on scans from both the PURE-01 and NABUCCO datasets, on both baseline and on-treatment scans. From the PURE-01 dataset, n=57 patients were included for training with both baseline and on-treatment scans. From NABUCCO, n=13 patients were included for training. A combination of the two datasets was investigated, to optimise the generalisation of the model.
4. The data subset of this model is comparable with the second model, but with different segmentation labels. For this model, the on-treatment segmentations for patients with a pCR were removed, causing the volume to be reduced to zero, instead of the original ROI. This is done to optimise the labelling according to the pathological outcome.

TABLE 2.1: The models were trained with the different subsets of training-, test- and external validation sets.

Model	Training		Test		External Validation		Labels
1	PURE-01 baseline	n=57	PURE-01 baseline	n=15	NABUCCO baseline	n=19	Ground truth
			PURE-01 on-treatment	n=72	NABUCCO n-treatment	n=19	
2	PURE-01 baseline	n=57	PURE-01 baseline	n=15	NABUCCO baseline	n=19	Ground truth
	PURE-01 on-treatment	n=57	PURE-01 on-treatment	n=15	NABUCCO on-treatment	n=19	
3	PURE-01 baseline	n=57	PURE-01 baseline	n=15	None		Ground truth
	PURE-01 on-treatment	n=57	PURE-01 on-treatment	n=15			
	NABUCCO baseline	n=13	NABUCCO baseline	n=4			
	NABUCCO on-treatment	n=13	NABUCCO on-treatment	n=4			
4	PURE-01 baseline	n=57	PURE-01 baseline	n=15	NABUCCO baseline	n=19	on-treatment volume
	PURE-01 on-treatment	n=57	PURE-01 on-treatment	n=15	NABUCCO on-treatment	n=19	to zero if pCR

Data Preprocessing

Preprocessing of the MR data is a key in many analyses, because MR images are acquired in arbitrary units, which cannot be compared between multiple images of a single patient nor across different patients.[36] Besides the automated nnU-Net preprocessing pipeline, intensity normalisation is applied to improve the generalisability of the model. For normalisation, a two-step method is applied.[37]

First, a standardisation scale was learned, for each sequence individually. From the PURE-01 and NABUCCO datasets, five scans of each were randomly selected. These scans were resampled to 1x1x1 mm spacing, a Bias Field Correction was applied, the scans were padded and cropped to the same geometry (256x256x256) and finally, the intensities were scaled between zero and one. The mean intensity histogram was determined from these scans, from which the standardised scaling factor was determined. The second step was to transform all images from the datasets with the trained standard scale. This resulted in MR images with consistent intensity histograms.[37] All four models, as described above, were trained again with the preprocessed data (further described as models 5-8).

2.3.2 Model Validation

Model performance was evaluated by the Dice score, Hausdorff distance and the volume correlation. The Dice score measures the overlap between the ground truth segmentation and the segmentation predicted by the trained model.[38–40] The higher the overlap between the two segmentations, the better the performance. The Hausdorff distance measures the maximum distance between two points of two different surfaces, one of the ground truth and one of the predicted segmentation.[40] The mean Hausdorff distance from all maximum distances was calculated. The correlation between the volume of the ground truth and the predicted segmentation was expressed with the Pearson Correlation Coefficient, measuring the linear relationship between the two variables. The correlation was also visualised in a scatter plot, where the regression line was plotted against the perfect prediction (x=y). Bland-Altman plots were used to determine the agreement (bias and variance) between the ground truth and predicted segmentation volumes.

2.3.3 Outlier Analysis

An observation of the outliers within the data was performed by calculating the Cook's distance. Cook's distance is an estimate of the influence of a specific data point. In other words, the distance describes how much the metrics will change if that data point is removed. Data points were suspect for outliers if the Cook's distance was greater than 3 times the mean value.[41,42] The suspect outliers were further analysed by visual inspection. We removed outiers if there were image artefacts or misalignment between the sequences.

2.3.4 Prediction of Pathological Response

Tumour volume was calculated from the predicted segmentation. Its predictive value for pathological treatment response was evaluated using the area under the curve (AUC) of the receiver operating characteristics (ROC) curve, sensitivity and specificity. An AUC value close to 1.0 indicates perfect discrimination between pCR and non-pCR, while an AUC value close to 0.5 indicates that the performance is as good as random guessing. The analysis was performed for both tumour volumes at baseline and on-treatment scans, as well as the difference over time, expressed in percentage of volume change, according to the following formula:

$$\frac{Vol_{on-treatment} - Vol_{Baseline}}{Vol_{Baseline}} \; x \; 100\% \tag{2.1}$$

2.4 Results

Eight instances of nnU-Net models were trained, each with a different subset of training data from the PURE-01 and NABUCCO datasets, see table 2.1. For 41% of the patients, a tumour-specific segmentation was delineated by the radiologist.

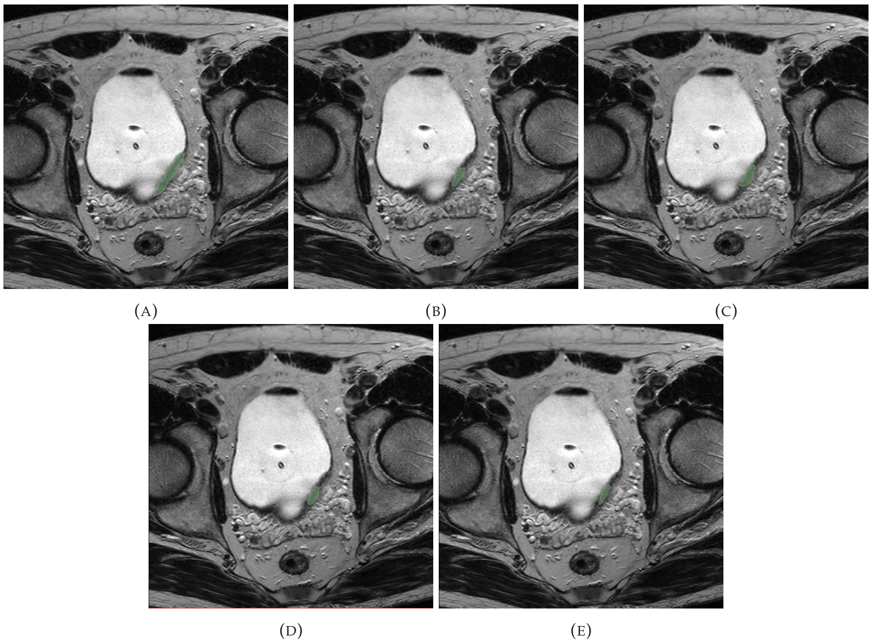
2.4.1 Model Performance

The results for model performance can be found in Appendix A. The segmentation volumes were calculated and it can be seen that the mean predicted segmentation volumes are smaller compared to the mean ground truth volumes.

Model 1, which was trained with the PURE-01 baseline scans, showed a Dice score of 0.39, the mean Hausdorff distance was 37.11 mm and the Pearson's correlation was 0.499 for the baseline test set. The PURE-01 on-treatment test set had a Dice score of 0.30, a Hausdorff distance of 31.35 mm and the Pearson's correlation was 0.693 (see Table A.1). Model 2 was trained with baseline as well as on-treatment scans. In Table A.2 it can be seen that the model performance for the PURE-01 on-treatment test set was getting better, while there was a slight decrease for the PURE-01 baseline test set, according to the Dice score. When, for model 4, the on-treatment labelling was changed according to the pathological outcome (instead of an ROI segmentation), the model performance for the PURE-01 on-treatment test set was even better; Dice score was 0.61, Hausdorff distance was 26.75 mm and the Pearson's correlation was 0.858 (see Table A.4). For model 3, where the training set was based on both the PURE-01 and the NABUCCO datasets, the performance for the PURE-01 test set was roughly equal to the performance of model 2. The NABUCCO performance could not be measured, since this was a very small subgroup (n=4). Instead, the overall model performance was calculated (see Table A.3).

Visual inspection of the predicted segmentation showed that the segmentation was within the ROI, but of a smaller volume. An example can be found in Figure 2.3.

The Pearson's correlation scatter plots can be seen in Appendix B(Figure B.1 for the baseline test set and Figure B.2 for the on-treatment test set). The regression line was plotted



(A) (B) (C) (D) (E)

FIGURE 2.3: Predicted segmentations on a T2-weighted on-treatment MRI scan from a 75-year-old man of the PURE-01 dataset. After cystectomy the pathological staging was ypT0N0Mx. The segmentations are visualised in green. (A) The radiological segmentation (ground truth) can be seen. The predicted segmentations can be seen for (B) model 1, (C) model 2, (D) model 3 and (E) model 4.

against the perfect prediction, to see how the predicted volumes correlate to the ground truth volumes. It can be seen that the regression line was bent due to the logarithmic scale.

Bland-Altman plots were used to determine the agreement (bias and variance) between the ground truth and predicted segmentation volumes (see Appendix C: Figure C.1 for the baseline test set and C.2 for the on-treatment test set). The mean difference was expressed in mm^3. The regression line was plotted with a 95% Confidence Interval (CI). For both baseline and on-treatment, it can be seen that the regression line goes up when the mean volume increases; the difference between the ground truth and predicted segmentation volume increased if the mean volume is higher.

For models 5 through 8, the same data subsets were used, but the data was first preprocessed based on Nyúl standard scaling, see Figure 2.4 for an example. The model performance, correlation plots and Bland-Altman plots can be found in respectively Appendix A

(Table A.5-A.8), Appendix B and Appendix C. The model performance for the PURE-01 test set was quite similar to the models trained without Nyúl standard scale preprocessing.

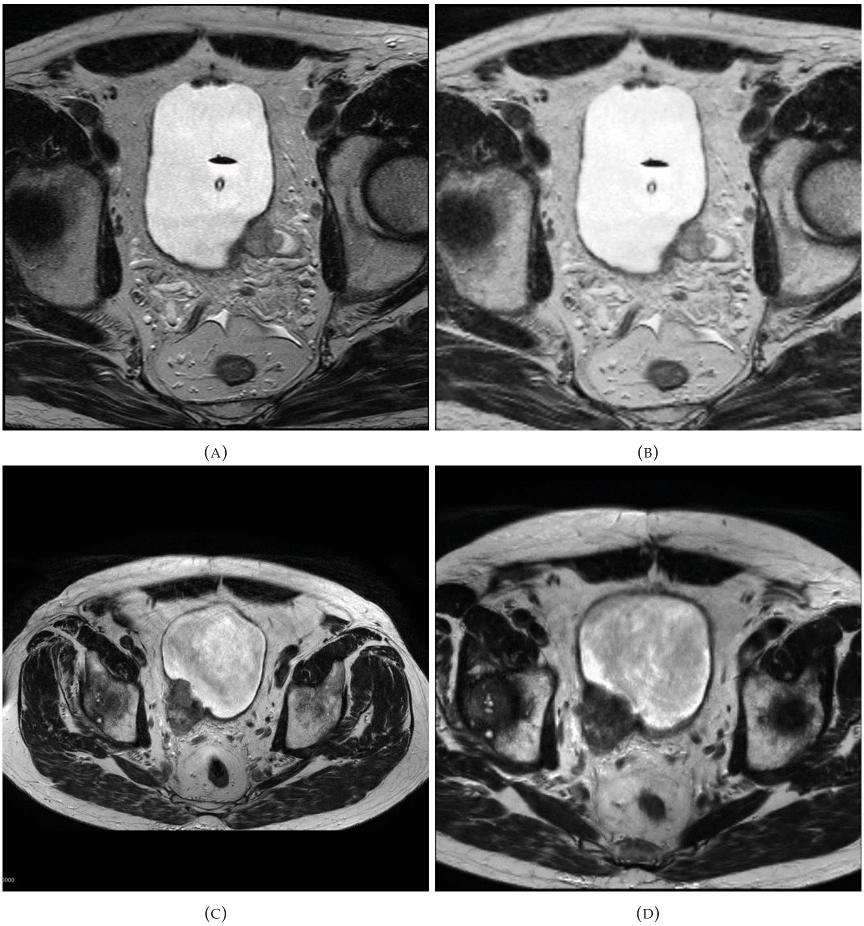
2.4.2 External Validation

On the external validation, which was performed with the NABUCCO dataset, models 1, 2 and 4 were performing poorly, which can be seen according to the low Dice score of 0.00 and high Hausdorff distances (see Appendix A). The correlation plots (Figure B.3 and B.4), showed that the overall correlation, especially for baseline segmentations, was poor. Visual inspection showed random predicted regions, sometimes within the actual tumour or ROI, as well as around the bladder in other anatomical structures (e.g. muscle, rectum and colon).

Model performance on the external validation set significantly improved for the models trained with Nyúl standard scale preprocessing. The Dice scores for the baseline segmentations were respectively 0.34, 0.36 and 0.28 for models 5, 6 and 8. This was comparable to the results of the PURE-01 test set. The Dice scores for the on-treatment segmentations were respectively 0.11, 0.27 and 0.09. It was noticeable that model 8 scores significantly lower on the NABUCCO on-treatment (external validation), compared to the PURE-01 on-treatment (test set); 0.09 vs 0.45. The baseline Bland-Altman plots (Figure C.3) for models 5, 6 and 8 showed a constant difference for different mean volumes. The on-treatment plots (Figure C.4) showed a higher difference in volume for higher mean volumes.

2.4.3 Outlier Analysis

For the PURE-01 test set, three outliers were identified by calculating Cook's distance. Visual inspection showed no signs of artefacts or misalignment's between the sequences, therefore the data points were not removed. The same goes for the NABUCCO external validation set. Three outliers were identified, but visual inspection showed no signs of artefacts or misalignment. It can be seen that the outliers, identified with Cook's distance, were the segmentations with the highest volumes.



(A)

(B)

(C)

(D)

FIGURE 2.4: T2-weighted pelvic MR images. (A) Original axial slices from a patient of the PURE-01 dataset with (B) the corresponding preprocessed slice. (C) Shows an original axial slice from a patient of the NABUCCO dataset, within (D) the corresponding preprocessed slice.

2.4.4 Prediction of Pathological Response

The predictive value of the volume was estimated using the ROC-AUC. An AUC value close to 1.0 indicates perfect discrimination between pCR and non-pCR, while an AUC value close to 0.5 indicates that the performance is as good as random guessing. In tables 2.2 and 2.3, the ROC-AUC values, sensitivity and specificity, with 95% CI, for respectively the PURE-01 test set and NABUCCO external validation set can be found. In Figure 2.5, the ROC curves for both PURE-01 and NABUCCO segmentation volumes can be found, with on the x-axis the False Positive Rate (1 - Specificity) and the y-axis the True Positive Rate (Sensitivity). Each figure shows, for each trained nnU-Net model, the ROC curves for baseline volume, on-treatment volume and the volume difference over time. It can be seen that the on-treatment volume predictions have a higher AUC value, compared to the baseline volumes and the volume difference over time.

2.5 Discussion

The goal of this study was to train an automatic segmentation model for bladder cancer. With this segmentation, tumour-specific features, like tumour volume, can be analysed for their predictive value of the therapeutic response.

2.5.1 Model Performance

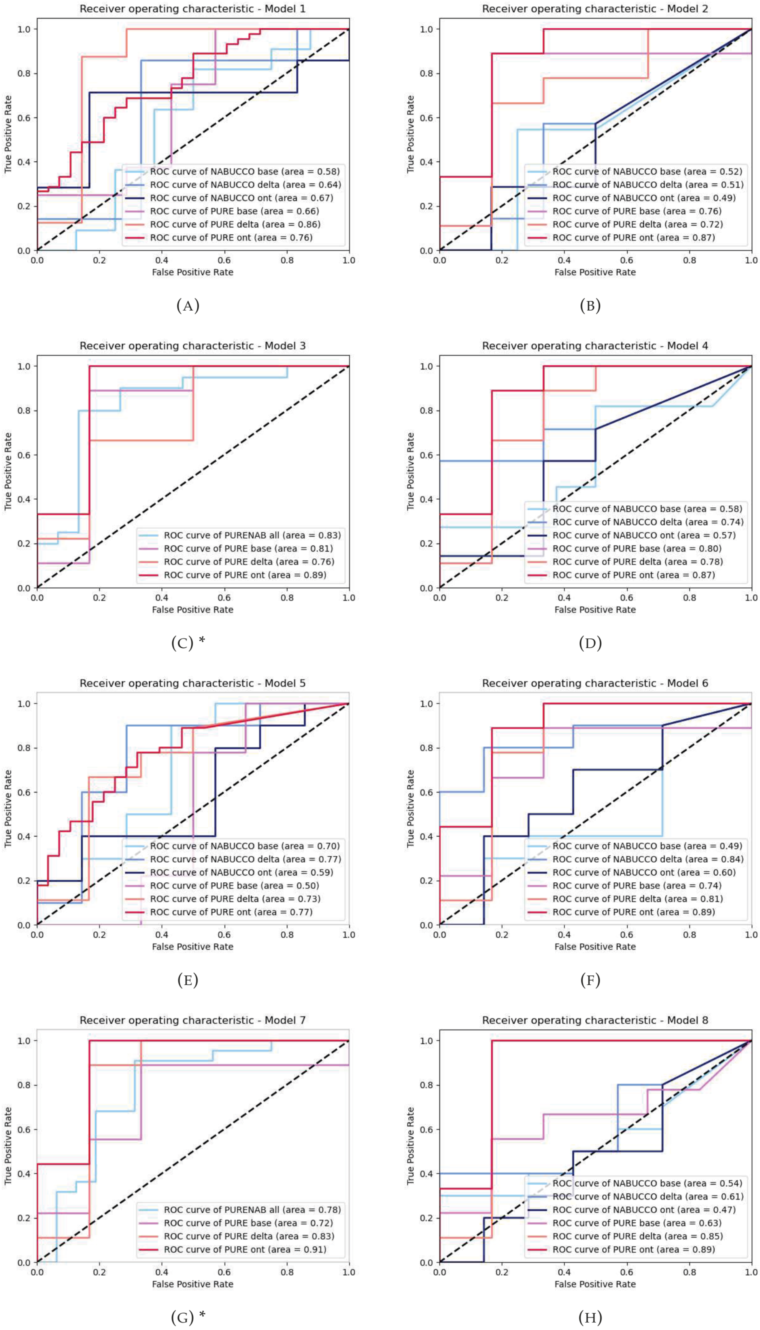
To train a model for tumour-specific segmentations, the labels should contain a radiological delineation of the tumour. Tumour delineation on baseline scans was complicated by the post-operative inflammatory response of the TURBT, which made it difficult to distinguish between tumour-specific tissue and the post-operative TURBT effect. If evident tumour could be delineated, the radiologist created a specific tumour segmentation. For model training, the tumour-specific segmentations were included if evident tumour could be delineated, else the ROI delineation was used. This resulted in a very weak linear correlation between the predicted and ground truth baseline segmentation volumes (Pearson's correlation coefficient between 0.051 - 0.523). In each trained model, the volume correlations for

TABLE 2.2: ROC-AUC values, sensitivity and specificity, expressed with a Confidence Interval (CI) of 95%, for the PURE-01 test set. Overall represents the average over all models.

		Volume AUC	**Sensitivity**	**Specificity**
Model 1	**Baseline**	0.66 (0.38 - 0.89)	0.50 (0.17 - 0.80)	0.57 (0.25 - 0.86)
	on-treatment	0.76 (0.66 - 0.85)	0.65 (0.52 - 0.76)	0.76 (0.61 - 0.88)
	Difference (%)	**0.86 (0.63 - 0.83)**	**0.75 (0.50 - 1.00)**	**0.86 (0.60 - 1.00)**
Model 2	**Baseline**	0.76 (0.50 - 1.00)	0.67 (0.40 - 0.90)	0.83 (0.56 - 1.00)
	on-treatment	0.87 (0.67 - 1.00)	0.67 (0.43 - 0.90)	0.83 (0.56 - 1.00)
	Difference (%)	0.72 (0.54 - 0.95)	0.56 (0.29 - 0.83)	0.67 (0.33 - 1.00)
Model 3	**Baseline**	**0.81 (0.56 - 1.00)**	**0.67 (0.40 - 0.89)**	**0.86 (0.57 - 1.00)**
	on-treatment	0.89 (0.68 - 1.00)	0.67 (0.40 - 0.91)	0.86 (0.57 - 1.00)
	Difference (%)	0.76 (0.50 - 0.96)	0.67 (0.40 - 0.90)	0.86 (0.57 - 1.00)
Model 4	**Baseline**	0.80 (0.55 - 1.00)	0.67 (0.40 - 0.90)	0.83 (0.56 - 1.00)
	on-treatment	0.87 (0.67 - 1.00)	0.67 (0.40 - 0.90)	0.83 (0.56 - 1.00)
	Difference (%)	0.78 (0.52 - 1.00)	0.67 (0.38 - 0.89)	0.83 (0.56 - 1.00)
Model 5	**Baseline**	0.50 (0.19 - 0.80)	0.44 (0.17 - 0.71)	0.50 (0.17 - 0.83)
	on-treatment	0.77 (0.68 - 0.86)	0.65 (0.52 - 0.76)	0.75 (0.62 - 0.88)
	Difference (%)	0.73 (0.52 - 0.97)	0.60 (0.33 - 0.86)	0.71 (0.40 - 1.00)
Model 6	**Baseline**	0.74 (0.50 - 0.97)	0.67 (0.38 - 0.89)	0.83 (0.56 - 1.00)
	on-treatment	0.89 (0.71 - 1.00)	0.67 (0.43 - 0.90)	0.83 (0.56 - 1.00)
	Difference (%)	0.81 (0.66 - 0.99)	0.70 (0.45 - 0.91)	0.88 (0.60 - 1.00)
Model 7	**Baseline**	0.73 (0.46 - 0.96)	0.56 (0.27 - 0.80)	0.67 (0.29 - 1.00)
	on-treatment	**0.91 (0.72 - 1.00)**	**0.67 (0.40 - 0.90)**	**0.86 (0.57 - 1.00)**
	Difference (%)	0.83 (0.64 - 1.00)	0.67 (0.40 - 0.90)	0.86 (0.57 - 1.00)
Model 8	**Baseline**	0.63 (0.38 - 0.88)	0.56 (0.29 - 0.83)	0.67 (0.33 - 1.00)
	on-treatment	0.89 (0.68 - 1.00)	0.67 (0.43 - 0.90)	0.86 (0.56 - 1.00)
	Difference (%)	0.85 (0.63 - 1.00)	0.67 (0.40 - 0.90)	0.86 (0.56 - 1.00)
Overall	**Baseline**	0.70	0.59	0.72
	on-treatment	0.86	0.67	0.82
	Difference (%)	0.79	0.66	0.81

TABLE 2.3: ROC-AUC values, sensitivity and specificity, expressed with a Confidence Interval (CI) of 95%, for the NABUCCO external validation set. Overall represents the average over all models.

		Volume AUC	**Sensitivity**	**Specificity**
Model 1	**Baseline**	0.60 (0.34 - 0.83)	0.55 (0.30 - 0.80)	0.64 (0.33 - 0.89)
	on-treatment	**0.67 (0.39 - 0.93)**	**0.71 (0.40 - 1.00)**	**0.83 (0.50 - 1.00)**
	Difference (%)	0.66 (0.33 - 0.91)	0.57 (0.22 - 0.88)	0.67 (0.33 - 1.00)
Model 2	**Baseline**	0.53 (0.31 - 0.77)	0.55 (0.27 - 0.82)	0.64 (0.33 - 0.90)
	on-treatment	0.50 (0.21 - 0.77)	0.43 (0.14 - 0.75)	0.50 (0.17 - 0.83)
	Difference (%)	0.51 (0.25 - 0.79)	0.57 (0.29 - 0.86)	0.67 (0.33 - 1.00)
Model 4	**Baseline**	0.59 (0.35 - 0.80)	0.45 (0.20 - 0.71)	0.50 (0.22 - 0.80)
	on-treatment	0.57 (0.30 - 0.85)	0.57 (0.25 - 0.88)	0.67 (0.33 - 1.00)
	Difference (%)	0.75 (0.48 - 0.95)	0.57 (0.25 - 0.88)	0.67 (0.33 - 1.00)
Model 5	**Baseline**	**0.70 (0.44 - 0.92)**	**0.50 (0.25 - 0.78)**	**0.57 (0.25 - 0.88)**
	on-treatment	0.59 (0.34 - 0.82)	0.40 (0.15 - 0.67)	0.43 (0.14 - 0.75)
	Difference (%)	0.77 (0.52 - 0.97)	0.60 (0.33 - 0.86)	0.71 (0.40 - 1.00)
Model 6	**Baseline**	0.49 (0.25 - 0.74)	0.40 (0.17 - 0.67)	0.43 (0.12 - 0.75)
	on-treatment	0.60 (0.36 - 0.85)	0.50 (0.25 - 0.75)	0.57 (0.25 - 0.88)
	Difference (%)	**0.84 (0.66 - 0.99)**	**0.70 (0.45 - 0.91)**	**0.88 (0.60 - 1.00)**
Model 8	**Baseline**	0.54 (0.31 - 0.77)	0.50 (0.22 - 0.78)	0.57 (0.25 - 0.88)
	on-treatment	0.47 (0.25 - 0.72)	0.50 (0.25 - 0.75)	0.57 (0.25 - 0.88)
	Difference (%)	0.61 (0.40 - 0.83)	0.50 (0.25 - 0.75)	0.57 (0.25 - 0.86)
Overall	**Baseline**	0.58	0.49	0.56
	on-treatment	0.57	0.52	0.60
	Difference (%)	0.67	0.59	0.70



(A)

(B)

(C) *

(D)

(E)

(F)

(G) *

(H)

FIGURE 2.5: ROC plots of the segmentation volume's predictive value for pathological response. For each trained nnU-Net model, the ROC curves for baseline volume, on-treatment volume and the volume difference over time. The x-axis represents the False Positive Rate (1 - Specificity) and the y-axis the True Positive Rate (Sensitivity). * For models 3 (C) and 7 (G), the NABUCCO data have too few data points for statistical analysis, therefore the ROC curve for the whole test set is calculated (*PURENAB all*). Abbreviations: base = baseline, ont = on-treatment, delta = difference between baseline and on-treatment.

the baseline segmentations were lower compared to the on-treatment segmentation volume correlations (0.693 - 0.942).

In the correlation scatter plots for the PURE-01 on-treatment set it can be seen that the predicted volumes underestimate the radiological ground truth (see Figure B.2). This was also seen during visual inspection (see Figure 2.3) and by the mean volume calculations (see Appendix A). This can be explained by the fact that the radiologist delineated an ROI for the on-treatment scans, based on the previously delineated ROI on the baseline scan. The segmentation was slightly adjusted to the anatomical structures, but no specific (residual) tumour segmentation was created. Since the on-treatment segmentations are not tumour-specific segmentations, the radiological ground truth will probably be an overestimation of the true pathological response, especially for those with a pCR.

When looking at performance evaluation only in terms of predicted segmentations, the Dice score, which determines the amount of overlap between both segmentations, and the Hausdorff distance are relatively poor. However, when we look at Pearson's correlation coefficient (Table Appendix A), the mean difference in the Bland-Altman plots (Figures C.1 and C.2) and the predictive value of segmentation volume for pathological response (Table 2.2), measured with the AUC, we see a significant performance. Since the model was trained on radiological segmentations, larger than the evident tumour, this could introduce a systemic error. A systemic error would throw off the performance metrics measuring overlap of volumes but would be accounted for in metrics that rely on rankings.

Further, it can be seen for the on-treatment segmentations that the difference between the ground truth and predicted segmentation volume was larger when the mean segmentation volume was higher (see Figure C.2). This proportional bias can be explained by the fact that for some patients the tumour specific segmentation was used as label and for others the large ROI.

Besides this mathematical approach, a visual inspection was performed as well during outliers detection and model failure. It can be seen that the segmentation was located correctly, i.e. in the bladder wall, but that the volume was systematically smaller than the ground truth (see Figure 2.3). This resulted in a low Dice score and the high Hausdorff distance as we have seen. However, the high correlation between the predicted and the ground truth volume, the agreement between AI- and radiologist-measured volume, and the good correlation between the on-treatment AI-measured volume and pathological response indicates the presence of a systematic error in volume measurement made by the AI algorithm - which can be also explained by the algorithm selecting the pixels providing the highest evidence of tumour.

2.5.2 External Validation

External validation was performed with the NABUCCO dataset (see 1.4.2). Model performance for external validation showed poor Dice scores, high Hausdorff distances and poor correlation coefficients (see Tables A.1, A.2 and A.4 in Appendix A). The Dice Score, determining the amount of overlap, was for all validation sets 0, except for model 1. This model was trained with PURE-01 baseline scans and the external validation with NABUCCO baseline showed a Dice score of 0.21. Visual inspection showed that the predicted segmentation was within the ROI, but there were also a lot of segmented areas outside of the ROI. The NABUCCO on-treatment predictions for model 1 showed a high mean predicted volume, but the Dice score was 0. Visual inspection showed that other (random) anatomical structures were segmented. Which corresponds to the low correlation coefficient. The predicted segmentations for models 2 and 4 had a very low mean volume, resulting in a Dice score of 0 and high Hausdorff distances.

Data preprocessing was applied to normalise the voxel intensities and therefore harmonise both datasets, in an attempt to make the models more generalisable. Comparing the models trained with the original data (models 1 to 4, see Table A.1 to A.4) with the models trained with preprocessed data (model 5 to 8, see Table A.5 to A.8), the models trained with preprocessed data showed significant improvement on the external validation set. The Dice scores for the models, both baseline and on-treatment, were comparable with the performance on the PURE-01 test sets. Only on-treatment segmentations for model 8, where the segmentation volume for pCR is set to zero, showed a really low Dice score for the external validation set. The significant improvement on the external validation set proves that intensity normalisation optimises the model performance on our external validation set.

Further, it can be seen that the mean ground truth volume for the NABUCCO baseline scans was significantly higher, compared to the NABUCCO on-treatment and the PURE-01 dataset. This is because both studies had different inclusion criteria according to the staging. The PURE-01 study included patients who had a clinical (c)T3bN0 stage tumour,[7] while the NABUCCO study included patients with a cT3-4aN0M0 or cT1-4aN1-3M0.[6] The ground truth volumes between the 'original' data and the preprocessed data slightly change, this is because the preprocessed data was resampled with nearest neighbour interpolation.

2.5.3 Outliers

Outliers were identified quantitatively, using the Cook's distance.[42,43] The Cook's distance identifies the influence of a data point on the model, so it measures how much the model will change if an individual data point is deleted. A high Cook's distance means a high influence on the model. If a data point is of high influence, it is suspect for an outlier, but it does not necessarily mean that it should be deleted from the dataset. The suspect outliers were analysed by visual inspection. Outliers were removed if there were image artefacts or misalignment between the sequences. For both the test- and external validation set, three outliers were identified with Cook's analysis, but we did not remove outliers. It turned out that the outliers were the ones with the highest segmentation volume. This is also reflected in the Bland-Altman plots, where we see that the difference between ground truth and predicted segmentation volume increased as the mean volume increased.

Outlier analysis relies on statistical assumptions that may not hold on small datasets. Using Cook's analysis on our datasets, the results are biased against larger volumes, therefore we decided to only remove the outliers if image artefacts or misalignment between the sequences was seen. Since those were not present, we did not remove outliers. For larger datasets in future studies, we should look into different methods for outlier removal.

2.5.4 Prediction of Pathological Response

The performance measurement for predicting pathological response can be found in Table 2.2 and the ROC curves can be found in Figure 2.5. The AUC values close to 1.0 indicate perfect discrimination between pCR and non-pCR, while the AUC values close to 0.5 indicate that the performance is as good as random guessing. Analysing the AUC scores showed that the on-treatment segmentation volumes have the highest predictive power. The on-treatment AUC scores are the highest for model 7, which was trained with both the PURE-01 and NABUCCO datasets (AUC score is 0.91).

Together with the AUC, sensitivity and specificity were calculated (95% CI). The sensitivity determines the ability to correctly identify patients with a disease (True Positive Rate), while the specificity determines the ability to correctly identify patients without the disease. If we look at the prediction of the pathological response, we want the sensitivity to be as high as possible, to minimise the number of False Negatives. In other words, we want to avoid that a patient with a non-pCR (labelled as 1) is classified as pCR (labelled as 0). But if we look at organ-sparring treatment options after neoadjuvant therapy, we would like to have a high specificity. For all models, it can be seen that the specificity is higher compared to the sensitivity. This means that the number of false positives is lower compared to false negatives. It is possible to choose a different threshold value to increase the sensitivity, however, this will result in a decrease in specificity. Since the current treatment protocol consists of a radical cystectomy, regardless of neoadjuvant therapy response, the high specificity might be an opening for possible organ-sparring protocols in the future.

2.5.5 Limitations and Outlook

The models were trained with an ROI segmentation, instead of a tumour-specific segmentation, for both the baseline and on-treatment scans. This is an overestimation of the actual tumour volume. More tumour-specific segmentations should be included in the training set, to create more accurate segmentations, which might probably increase the predictive value of segmentation volume for the pathological response. For tumour-specific segmentations we do not necessarily need immunotherapy studies, increasing the range of possible datasets that could be included.

For this study, the pathological response was defined as the pathological staging after radical cystectomy. Pathological staging was based on the location of remaining tumour cells within the bladder wall of the resected specimen. This means that pathological response was not based on the remaining tumour volume or tumour regression rate. For example, tumour regression can be close to 100%, but if only a few viable tumour cells are left within the perivesical fat, this will result in a pT3. For this study, therapeutic response was based on the pathological staging; pCR (no remaining tumour cells; ypT0N0) vs non-pCR (remaining tumour cells; $\geq$ ypTaNo). Since the results for both on-treatment volume (remaining tumour volume) and the volume difference over time (tumour regression) as a predictive value for pathological response is promising, volume and regression might be taken into account for pathological treatment response.

The AUC values for on-treatment tumour volume were between 0.76 and 0.91, and the AUC values for tumour difference over time were between 0.72 and 0.85. When looking at organ sparring treatment after neoadjuvant therapy, further prospective studies should determine the prognostic role of pathological staging and tumour regression in clinical practice. Follow-up treatment could, for example, be different for a complete response vs non-complete response, remaining non-muscle invasive tumour vs remaining muscle invasive tumour, the remaining tumour volume or the tumour regression rate.

For training the nnU-Net model, all five sequences were added as separate channels, therefore it is important that the geometry of the landmarks and structures represented on each sequence is the same, and that each voxel represents the same anatomical location. Especially for the bladder, the anatomical location can vary during the sequences since it is a very deformable organ (during the MRI acquisition bladder filling increases). For this study, each sequence was pre-processed to the same voxel size to match the same space. Unfortunately, this interpolation does not result in perfect anatomical alignment. Optimising sequence alignment ensures that the voxels in different channels will represent the same anatomical information, and will therefore result in more accurate segmentations. Optimising sequence alignment could be performed with the use of registration software that supports non-rigid image registration parameters. Non-rigid transformations can change in shape and size, therefore dilatation and shear transformations are possible. For bladder registration, these non-rigid transformations are essential because the correspondence between two acquired images cannot be described by only local translation or rotation.[44]

The sequences that were included in this study, are the same sequences as the radiologist uses for bladder cancer assessment. It is not known whether other sequences have a predictive value. The fact that the radiologist does not include them in the assessment, does not automatically mean that they do not contain predictive characteristics. Future studies could include different or multiple DWI sequences and/or different time moments for the DCE sequence and should point out the added value of each sequence.

Besides the DCE at 6/8 seconds, where early enhancement from tumour tissue is visible, the DCE at 12 seconds was included, because post-operative inflammatory tissue, due to the TURBT, starts to enhance at 13.6 seconds (± 4.2 seconds) after arterial enhancement.[20] Based on these two DCE scans, it might be possible to make a distinction between tumour and post-operative inflammatory tissue, but further research should be performed to determine the added value of each sequence. For this study, the DCE scans were manually selected, but it might be possible to create a new algorithm to automatically select the DCE scan with the most predictive information. Especially if more data is included, hand-picking the scans is time-consuming.

2.6 Conclusion

We developed an automatic segmentation model using a DL algorithm to segment a bladder tumour in mpMRI scans, from patients with MIBC, treated with neoadjuvant immunotherapy. Performance evaluation showed a systematic error, introduced by ground truth segmentations which is an overestimation of the real tumour volume. This systematic error throws off the performance metrics. Nevertheless, the predictive value of segmentation volumes for pathological response looks promising. Future clinical studies should prove the role of pathological staging, in terms of remaining tumour volume and tumour regression rate, for organ sparring treatment after neoadjuvant therapy.

Chapter 3

Identification of predictive features

3.1 Introduction

Radiomics is an emerging field in oncology and radiology, that aims to use computer algorithms to quantify tumour-specific characteristics from medical imaging, that can be used for clinical outcome predictions. In the current clinical practice, radiological images are qualitatively assessed by the radiologist. Research is increasing in the field of extracting quantitative tumour morphological characteristics from radiological images, using computer algorithms, that can be used for image feature identification and response prediction, allowing more objective and quantitative analysis of tumours on radiological images.[45] Radiomics uses a collection of methods that extracts features from radiological images with a data-characterisation algorithm. By quantitatively analysing the tumour characteristics on radiological images, radiomic features could provide a better understanding of bladder cancer pathology.[45–47] In our study, where we make use of multiple MRI sequences, morphological and functional features are extracted and quantified by the image intensity, shape and texture.

More specifically, in this chapter, radiomic features were extracted from MRI sequences, based on the DL-segmentations from the previous chapter, and used to predict immunotherapy response in patients with MIBC. We studied the predictive performance of radiomic features to select patients for immunotherapy, in case of baseline scans, or for potential organ-sparring treatment following complete response after neoadjuvant immunotherapy, in case of on-treatment scans.

3.2 Technical Background

3.2.1 Radiomic features

Radiomics uses data-characterisation algorithms to convert radiological imaging data into a high-dimensional feature space. This is done through mathematical extraction of the spatial distribution of signal intensities and their interrelationships.[48,49] There are multiple software libraries available to extract radiomic features. We used the open-source PyRadiomics package. Different types of features can be extracted from a radiological image. These are divided into three classes: shape, intensity and texture.[49] Shape based features are described in two- and three-dimensional size and shape of the segmentation mask, which is defined by the segmentation mesh grid. The intensity-based features are calculated with first-order statistics, these calculations describe the distribution of the voxel intensities. The textural, second-order statistics-based, features are derived from the textural matrices and the distribution of the grey level values in an image.[49,50]

There are five matrices to describe textural features: 1) Grey Level Co-occurrence Matrix (GLCM). This matrix shows different combinations of grey levels within the image and therefore textural features can be extracted. The GLCM are second-order derived features, which are computed considering the spatial relationship between two intensity levels of the ROI. 2) Grey Level Size Zone Matrix (GLSZM), which quantifies grey level zones, defined as the number of connected voxels with the same grey level intensity in an image. 3) Grey Level Run Length Matrix (GLRLM) features are defined by the length of consecutive voxels with the same grey level value. 4) Neighbouring Grey Tone Difference Matrix (NGTDM) features point out the difference between the grey value and the average grey value with its neighbours. 5) Grey Level Dependence Matrix (GLDM) features, which quantifies the grey level dependencies of voxels within an image.

3.3 Method and Materials

3.3.1 Feature Extraction

The features were derived from a segmentation mask on a specific MRI sequence. For each patient, radiomic features were extracted for both baseline and on-treatment scans, from five sequences (T2-weighted, DWI b=1000, ADC, DCE at 8 and 12 seconds after arterial enhancement), using the segmentations that were obtained with Model 1, as described in Chapter 2.3 (see also table 2.1). All available feature classes were enabled and extracted. See figure 3.1 for the schematic pipeline.

3.3.2 Feature Analysis

After extracting the features, a Random Forest classifier (RFC) was trained to predict pathological response to treatment. We investigated four different settings: baseline, on-treatment, the difference over time between baseline and on-treatment, or delta features (in percentages) and concatenation of both the baseline and on-treatment features.

The PURE-01 dataset was split into a train- (70%) and a test set (30%), via Monte Carlo Cross Validation (MCCV). With MCCV, random subsamples of train- and test sets are generated, where the data is tested arbitrary. This means that the data can be more than once in the test set, or not even at all. This differs from K-fold cross-validation, where each subset is only tested once. Due to the small dataset, MCCV wass introduced to avoid overfitting and reduce the variability of the model.

In each training fold, we selected the 20 most predictive features, based on the score function *f-classif*, which calculates the ANOVA F-value between the label and features. An RFC was trained to predict treatment response, with the pathological response used as the ground truth: pCR was labelled as 0 and non-pCR was labelled as 1. As a random forest fits several decision tree classifiers (n-trees=100), the maximum depth of the tree was set to two, to avoid overfitting.

Model performance was evaluated by fitting the model on the test fold. By using multiple iterations (number of iterations = 100), the average test error was calculated over all iterations. The accuracy of the classifier was expressed in the ROC-AUC with a 95% CI. A correlation analysis was performed to determine whether there is a linear relationship between the features of the sequences.

3.4 Results

The proposed pipeline was trained with n=55 patients and tested with n=24 patients, 36% of them were labelled with a pCR. Features were extracted from the baseline and on-treatment scans, for five different MRI sequences; T2-weighted, DWI b=1000, ADC, DCE=8s and DCE=12s. For both the PURE-01 baseline and on-treatment scans, 112 features were extracted.

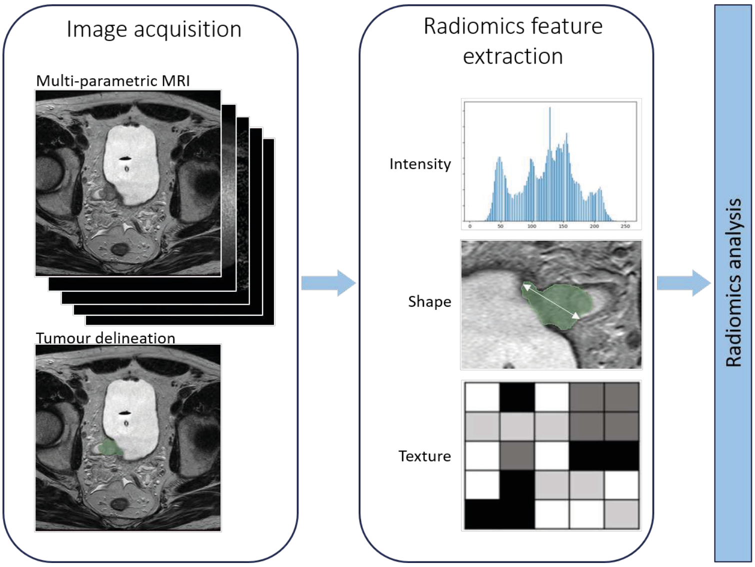


FIGURE 3.1: Workflow of radiomics: starting with the image acquisition of mpMRI scans. Tumour segmentation is automatically performed by a trained AI segmentation model, the mask can be seen in green. Second, radiomic features are extracted and finally, the predictive performance for the pathological treatment response is analysed.

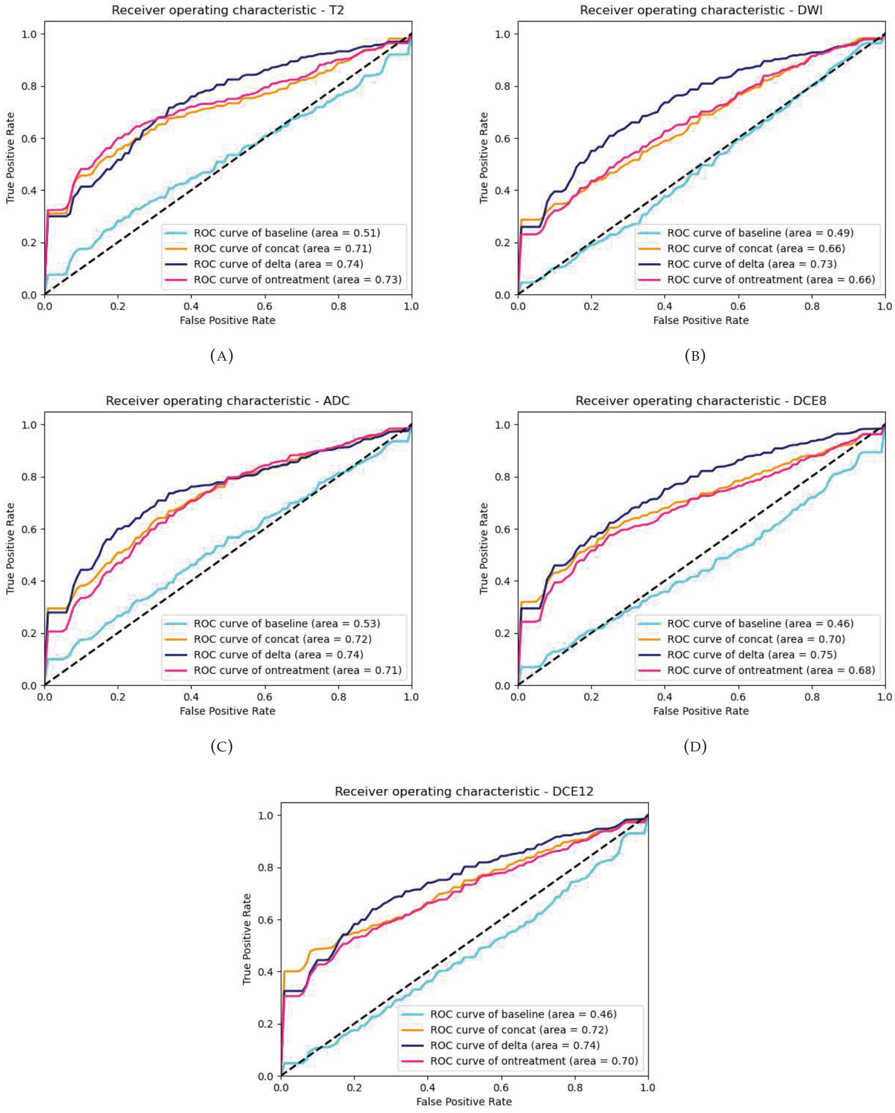
Classifier performance was expressed with the ROC-AUC. An AUC score close to 1.0 indicates perfect discrimination between pCR and non-pCR, while an AUC score close to 0.5 indicates that the performance is as good as random guessing. In Table 3.1, the AUC values can be found, expressed with a CI of 95%. The ROC curves can be found in Figure 3.2. Radiomic features determined on the on-treatment segmentation had a higher predictive value compared to the features extracted on the baseline segmentation, across all sequences (respectively 0.70 vs 0.49 average AUC). For on-treatment scans, the T2-weighted sequence had the highest AUC score (0.73). Baseline scans still provide valuable information, when combined with on-treatment, as the difference over time between baseline and on-treatment showed, overall, the highest AUC scores. Compared to the results reported in the previous chapter (Chapter 2, Table 2.2), radiomic features underperform compared to volumetric measurements (respectively; on-treatment 0.70 vs 0.77, and difference over time 0.74 vs 0.82)

In Table 3.2 the most predictive radiomic features were determined for each sequence, divided into feature classes; intensity, shape and texture. For the baseline scans the most predictive features were shape-based. While for the on-treatment scans, the most predictive features were intensity- and textural-based. For the delta features, it can be seen that surface area (shape-based) and texture features were the most predictive. It was noticeable that the most predictive radiomic features for the concatenated database were based on the on-treatment segmentations. This can be explained by the fact that the AUC scores for the on-treatment features were higher compared to the AUC scores of the baseline features (see Table 3.1).

In Figure 3.3, the correlation matrices of the most predictive radiomic features were presented. The correlation matrices describe the strength and direction of the linear relationship between two features. A correlation score close to 1 indicates a strong positive linear relationship, while scores close to -1 indicate a strong negative linear relationship. The correlation scores around 0 mean that there is a very weak linear relationship between the features. For example, among the baseline features, it can be seen that the kurtosis feature from the DWI had no linear correlation with any other feature. The on-treatment correlation matrix showed a low linear correlation between the GLCM-class features from the DCE scans and the shape-based features from the DWI, ADC and DCE at 8 seconds. The correlation matrix for the difference between the baseline and on-treatment features showed weak linear correlations. Since the most predictive features from the concatenated dataset were based on the on-treatment features, the concatenated correlation matrix is quite similar to the on-treatment correlation matrix.

TABLE 3.1: Results of the extracted image features. The results are expressed as the ROC-AUC with a 95% CI. The ROC-AUC value represents the predictive value of the extracted image features, for each segmentation, to predict the pathological therapeutic outcome. The predictive value was expressed as the ROC-AUC with a 95% CI, for the three different experiments.

	Radiomics Features				
	T2-weighted	**DWI**	**dADC map**	**DCE = 8s**	**DCE = 12s**
Baseline	0.51 (0.49 - 0.54)	0.49 (0.47 - 0.51)	0.53 (0.51 - 0.55)	0.46 (0.43 - 0.48)	0.46 (0.43 - 0.48)
On-treatment	0.73 (0.71 - 0.75)	0.66 (0.65 - 0.68)	0.71 (0.69 - 0.72)	0.68 (0.66 - 0.70)	0.70 (0.68 - 0.72)
Difference (%)	**0.74 (0.73 - 0.76)**	**0.73 (0.71 - 0.75)**	**0.74 (0.72 - 0.76)**	**0.75 (0.74 - 0.77)**	**0.74 (0.72 - 0.76)**
Concatenation	0.71 (0.69 - 0.73)	0.66 (0.64 - 0.68)	0.72 (0.70 - 0.73)	0.70 (0.68 - 0.73)	0.72 (0.70 - 0.74)



(A)

(B)

(C)

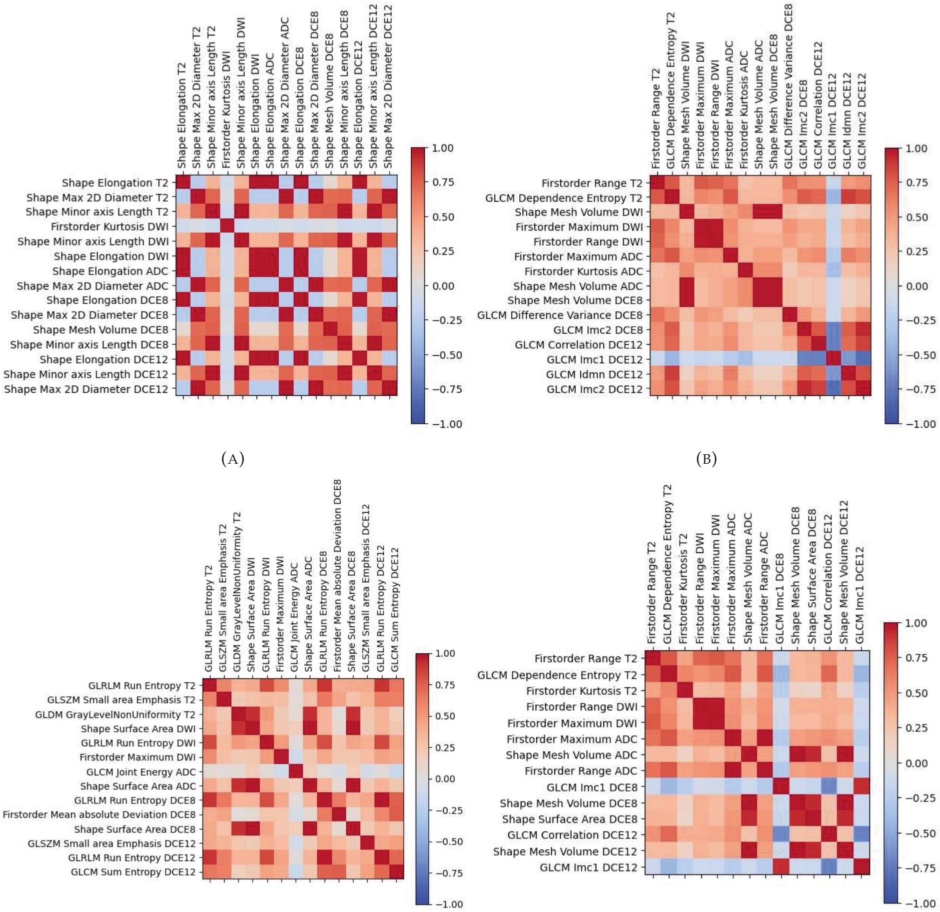
(D)

(E)

FIGURE 3.2: ROC curves of the predictive values for the pathological response, based on the radiomic features for the five sequences: (A) T2-weighted, (B) DWI, (C) ADC map, (D) DCE at 8 seconds after arterial enhancement and (E) DCE at 12 seconds after arterial enhancement. The 'area' within the legend corresponds with the AUC value.

TABLE 3.2: Most predictive features for treatment response. With Monte-Carlo cross-validation, model performance was determined based on 100 iterations. For each iteration, the most predictive feature was calculated, so the number behind each feature represents how many times this specific feature was the most predictive.

	T2	**DWI b1000**	**dADC**	**DCE=8s**	**DCE=12s**
Baseline	Shape_Elongation (27) Shape_Max 2D diameter (19) Shape_Minor axis length (16)	Firstorder_Kurtosis (19) Shape_Minor axis length (16) Shape_Elongation (14) Shape_Max 2D diameter (14)	Shape_Elongation (37) Shape_Max 2D diameter (13)	Shape_Elongation (25) Shape_Max 2D diameter (22) Shape_Mesh Volume (11) Shape_Minor axis length (11)	Shape_Elongation (26) Shape_Minor axis length (21) Shape_Max 2D diameter (10)
on-treatment	Firstorder_Range (26) GLDM_Dependence entropy (14)	Shape_Mesh volume (16) Firstorder_Maximum (11) Firstorder_Range (9)	Firstorder_Maximum (19) Firstorder_Kurtosis (14) Shape_Mesh volume(14)	Shape_Mesh volume (11) GLCM_Difference variance (10) GLCM_lmc2 (8)	GLCM_Correlation (16) GLCM_lmc1 (10) GLCM_ldmn / lmc2 (8 / 8)
Difference	GLRLM_Run entropy (20) GLSZM_SmallArea Emphasis (19) GLDM_Gray level non-uniformity (14)	Shape_Surface area (27) GLRLM_Run entropy (18) Firstorder_Maxima (17)	GLCM_joint energy (25) Shape_surface area (23)	GLRLM_Run entropy (40) Firstorder_Mean Absolute Deviation (17) Shape_SurfaceArea (16)	GLSZM_Small area emphasis (19) GLRLM_Run entropy (18) GLCM_Sum entropy (12)
Concatenate	Firstorder_Range.1 (18) GLDM_Dependence Entropy.1 (16) Firstorder_Kurtosis.1 (11)	Firstorder_Range.1 (14) Firstorder_maximum.1 (7)	Firstorder_maximum.1 (13) Shape_MeshVolume.1 (13) Firstorder_Range.1 (11)	GLCM_lmc1.1 (12) Shape_Meshvolume.1 (7) GLDM_Dependence entropy.1 (7)	GLCM_Correlation.1 (19) Shape_Meshvolume.1 (11) GLCM_lmc1.1 (9)



(A)

(B)

(C)

(D)

FIGURE 3.3: Correlation plot between the most predictive features of the sequences, for (A) baseline scans, (B) on-treatment scans, (C) difference over time and (D) the concatenated dataset.

3.5 Discussion

In this chapter, radiomic features were extracted from the baseline and on-treatment segmentations to predict pathological response. Delta features were calculated to analyse the difference in tumour morphology over time. As shown in Table 3.1, it can be seen that the AUC scores for the baseline segmentation were around 0.5 and therefore no distinction between the therapeutic response can be made, based on the baseline features. When looking at the features extracted from the on-treatment segmentations and the difference over time, the AUC scores showed significant predictive values and were better to distinguish cases of different pathological responses. For the concatenated dataset it can be seen that the predictive features were determined from the on-treatment segmentations, and was not considering baseline features at all. Considering the highest performance being reached by delta features, these results suggest that simple concatenation might not provide optimal utilisation of the feature set. However, it is difficult to formulate any sound hypothesis on this matter, due to the apparent lack of predictive value in the baseline features.

From Chapter 2 (see Table 2.2), the AUC scores for the segmented volume as predictive value for therapeutic response were calculated. The volume difference over time had the highest AUC score (0.82) and the on-treatment volume had an AUC score of 0.77. The radiomic features corresponded with these results because the Mesh Volume was one of the most predictive features for on-treatment scans and the Surface Area was one of the most predictive features for the difference over time (see Table 3.2). However, compared to these volumetric measurements, the radiomics features showed lower AUC scores and therefore underperformance. This might be the case if the RFC was overfitting on the training data. AUC performance for the test set will decrease. Therefore we implemented a Logistic Regression Classifier, to reduce the model complexity. Performance results showed lower AUC values, compared to the model trained with RFC, and therefore it was probably underfitting on the training data. Further study is needed to determine the optimal balance for variance-bias trade-off and minimise the total error.

The T2-weighted sequence was used by the radiologist to determine morphological characteristics of the tumour, which will appear slightly hyperintense.[22,23] The First-Order

Range defines the range of grey values within the ROI and was a predictive T2 feature for the therapeutic response. Then a second predictive T2 feature was the Dependence Entropy from the GLDM. This feature determines the heterogeneity between the voxels. The DWI sequence was used to evaluate diffusion restriction, caused by the tumour. True restricted diffusion will show an increased signal on the DWI and a low signal on the ADC.[22,23] This is why the maximal voxel intensity and intensity range (First-Order Maxima and -Range) was a predictive feature for the DWI sequence. For the ADC sequence, the maximal intensity was also predictive, together with the Kurtosis of voxel intensity. The Kurtosis represents the shape (peakedness) of the voxel intensity distribution. Tumour lesions can be distinguished from healthy tissue due to early enhancement visible on the DCE sequence. Early enhancement is recognised by hyperintensity on the early DCE time frames.[20,21,23] For the DCE sequences, the GLCM features were identified as predictive features. GLCM is a second-order statistical texture feature, which represents the spatial relationship, or distribution, of the grey values between voxels.[51] The distribution of enhancement might be predictive since it gives information about the extent of early enhancement, together with the shape of the mesh volume.

For the difference between baseline and on-treatment, it can be seen that the surface area was a predictive feature for multiple sequences. This can be explained by the fact that the surface area is a shape-based feature that is computed on the segmentation independently from the image itself. In combination with it, several textural features were predictive of therapeutic response. When the baseline and on-treatment features were concatenated, the predictive features were based on the on-treatment features, and therefore comparable with the on-treatment results.

Correlation matrices were generated to visualise the linear correlation between the most predictive features of each sequence. For baseline scans (Figure 3.3a) it can be seen that the kurtosis feature from the DWI had no linear correlation with any other feature from the other sequences. It should be noticed that the kurtosis was also the only feature from the First-Order class, all other features were shape-based. From the correlation matrix of the on-treatment features (Figure 3.3b), it can be seen that there was a very low linear correlation between the GLCM features from the DCE scans and the shape-based features from the

DWI, ADC and DCE at 8 seconds. The correlation matrix for the difference between the baseline and on-treatment features (Figure 3.3c) showed low linear correlations, especially if the features were from different classes. The features that showed weak linear correlation were independent of each other. Therefore, a combination of these features might improve the overall predictive value, increasing the AUC score. For this study, we only calculated the AUC scores for each sequence separately. Due to some weak correlations between the features, we might look further into the predictive value when all sequences are combined, especially for the on-treatment and difference over time features.

The most predictive features turned out to be all shape-based features. Therefore we assume that only (baseline) tumour volume will not have enough predictive performance for therapeutic response assessment.

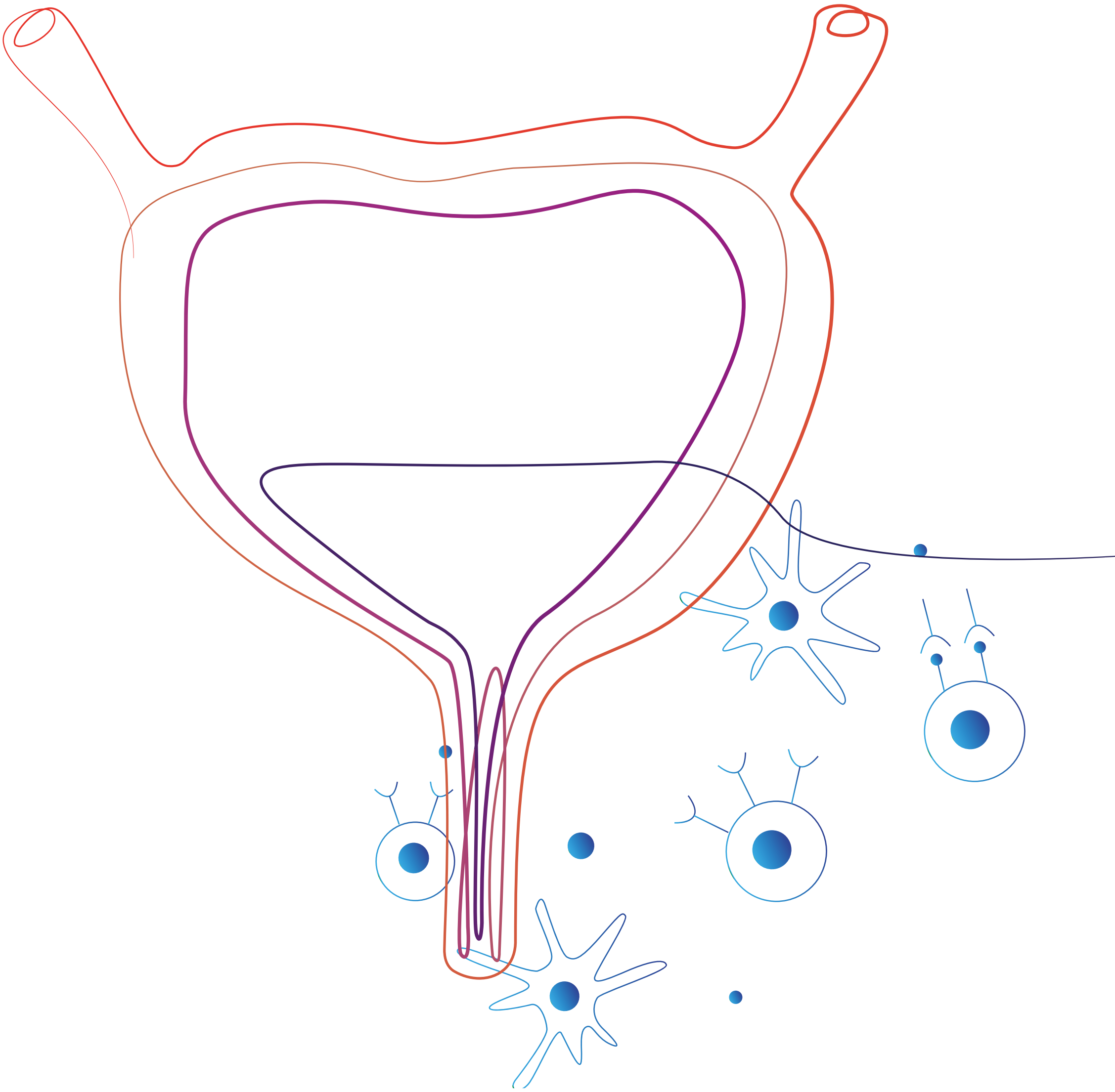
Some predicted segmentations for the on-treatment scans had a segmentation volume of zero, which means there is no segmentation available to extract radiomic features. Instead of removing these scans, the values were replaced by zeros for further analysis, but further research is needed.

3.6 Conclusion

In this study, we extracted tumour-specific image features based on DL segmentations from mpMRI in patients with MIBC, treated with neoadjuvant immunotherapy. Based on the baseline scans, almost no distinction can be made in the pathological treatment response, but looking at the difference over time between the features, a better distinction can be made. The features that are extracted after neoadjuvant immunotherapy are promising to predict the pathological response, while the features, derived before treatment are not.

Chapter 4

Automated classification of response



4.1 Introduction

Currently, there are no biomarkers sufficient enough to predict therapeutic responses for neoadjuvant immunotherapy for patients with MIBC. Neoadjuvant immunotherapy shows promising results,[6,7] but to achieve optimal treatment, it should be decided, for each patient individually, whether immunotherapy is the most suitable option.

Radiological images provide noninvasive, image features which might be useful to predict therapeutic responses. Especially, mpMRI plays an important role in therapeutic response assessment.[23] CNNs can be used to identify these image features and therefore a prediction model can be trained. In the previous chapters we have seen an automatic segmentation model and tumour-specific radiomic features are extracted from those segmentations. The features and predictions are based on tumour-specific characteristics. In this chapter, an end-to-end CNN is presented, where the prediction of therapeutic response is based on the full 3D mpMRI scan, rather than tumour-specific characteristics. The goal is to classify pathological staging and identify image features that might be predictive of the therapeutic response for neoadjuvant immunotherapy.

4.2 Method and Materials

For this study, T2-weighted, 3D MRI scans from the PURE-01 study were used. MRI scans are initially expressed in arbitrary units, therefore the voxel intensities cannot be compared between multiple images of a single patient nor across different patients.[36] Corresponding anatomical structures do not have similar voxel intensities, not even under the same scanner conditions, with the same acquisition protocol, and therefore image preprocessing is required. The T2-weighted images were normalised by a standardised scale, according to the method described by Nyúl et al.[37]

First, a standardisation scale was learned from 10 randomly selected on-treatment scans of the PURE-01 dataset. These scans were re-sampled to 1x1x1 mm spacing, a Bias Field Correction was applied, the scans were padded and cropped to the same geometry (265x256x256 and 192x192x192) and finally, the voxel intensities were scaled between zero and one. The mean intensity histogram was extracted from these scans, from which the standard scaling factor was determined. The second step was to transform all on-treatment scans with the trained standardised scale. This resulted in MR images with voxel intensities that had consistent tissue meaning.

Network Implementation

A 3D ResNet-18 architecture was implemented for model training. A ResNet, also called Residual Network, is a CNN-based network, where skip layers are implemented. The network architecture consists of residual blocks where skip connections are made between the activation layers (see figure 4.1). A ResNet was formed by stacking these residual blocks, whereas the ResNet-18 architecture consisted of 18 stacked residual blocks. ResNet has the ability to provide a deep neural network, which can be trained without increasing the training error percentage. Besides that, ResNet also overcomes the vanishing/exploding gradient problems, because of the skip layers.

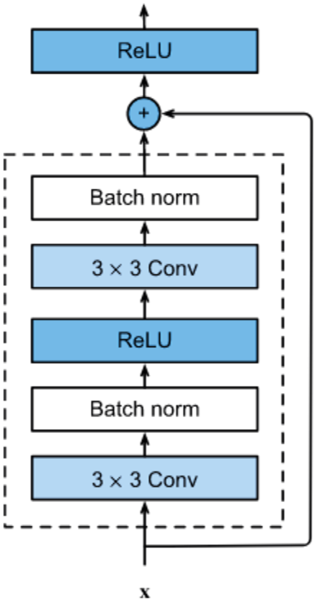


FIGURE 4.1: Residual block from a ResNet architecture

Multiple ResNet-18 models were trained, with n=82, T2-weighted, on-treatment scans from the PURE-01 dataset (see Table 4.1). A binary classification label was used, based on the pathological outcome: pCR was labelled as 0 and non-pCR as 1. Train- and validation sets were randomly divided (respectively 70% and 30%) concerning the pathological outcome. First, a model was trained with the entire 3D T2-weighted scans (256x256x256). A second model was trained with T2-weighted scans that were preprocessed to geometry of 192x192x192. The scans were cropped around the bladder, to decrease the complexity of the data and therefore overcome underfitting of the model. Lastly, a multichannel model was trained with the cropped T2-weighted scans and the radiological segmentations (as described in Chapter 2.3). The segmentations were added as a second input channel, to provide the model with a region of interest (ROI).

The model was compiled with the categorical cross-entropy loss function and Adam as an optimizer, with an initial learning rate of 1e-3. The batch size was set to 5, the model was trained for 50 epochs and after every epoch, the best model checkpoint was performed. To address overfitting, several methods were implemented to optimise the model towards a more generalisable model: First, we focused on the input data. We applied data augmentation with multiple rotation angles, to enlarge the dataset, but also to increase the heterogeneity of the training samples. Second, we modified the model architecture and parameters. Several drop-out layers (standard, spatial, alpha) were added and multiple normalisation methods have been applied (batch-, group-, layer normalisation). Lastly, we modified the model output by increasing regularisation.[52,53]

Model performance was determined by the validation accuracy of the trained model and overfitting was identified by analysing the training and validation loss. A model is overfitting when the validation loss starts increasing, while the training loss continues to decrease. The predictive value for pathological response was expressed with the AUC of the ROC curve, sensitivity and specificity, presented with a 95% CI.

Finally, gradient-based saliency maps were visualised to determine which regions of the MRI scan were relevant for the classification of pathological response. The saliency maps were rendered as a heatmap, where the ’hotness’ corresponds to the region that has a high impact on the models’ classification. It represents the features that lead to the final decision.

TABLE 4.1: Model specifications for the training of the 3D ResNet-18 models.

	Data input	Geometry	Additional Specifications
Model 1	3D T2-W MRI	256x256x256	None
Model 2	3D T2-W MRI	192x192x192	None
Model 3	3D T2-W MRI	192x192x192	Data Augmentation Standard drop-out 30% Spatial drop-out 30% Batch Normalisation
Model 4	3D T2-W MRI Segmentation ROI	192x192x192x2	Data Augmentation Spatial drop-out 30% Batch Normalisation
Model 5	3D T2-W MRI scan Segmentation ROI	92x114x93x2	Data Augmentation Spatial drop-out 30% Batch Normalisation

4.3 Results

A 3D ResNet-18 model was trained with T2-weighted, on-treatment MRI scans, from n=57 patients of the PURE-01 dataset. The validation set consisted of n=25 patients. Model performance results can be found in Table 4.2. First, a model was trained with the entire 3D, T2-weighted scans (256x256x256), resulting in a validation accuracy of 0.42, the AUC score for predicting pathological response was 0.34. Cropping of these scans to the geometry of 192x192x192 resulted in a validation accuracy of 0.63. Validation accuracy goes up, but so does the training accuracy (respectively 0.89 and 1.00), indicating that the model is overfitting. To prevent overfitting; data augmentation, drop-out methods, normalisation and L2-regularisation additions were performed. Model 3 showed a validation accuracy of 0.67 and an AUC value of 0.47. The validation accuracy for model 3 was the highest compared to the other trained models.

Models 4 and 5 were trained with the segmented ROI of the tumour added as a second channel. This resulted in a validation accuracy of 0.58 with a training loss of 0.53 and training accuracy was 0.88. The AUC score for model 4 was 0.62. For model 5, the MRI scans were cropped according to a general bounding box based on the ROI segmentation. The input layer had a dimension of 92x114x93x2, resulting in a lower validation accuracy (0.52 compared to 0.58) and lower AUC score (0.52 compared to 0.62). Training loss for model 5 was 0.12 and the training accuracy is 1.00, indicating overfitting during training. From the ROC curves in Figure 4.2, it can be seen that model 4 had the highest AUC score (0.62)

TABLE 4.2: Model performance of the trained 3D ResNet-18 models. The AUC, sensitivity and specificity are presented with a 95% CI. None of the models showed significant results.

	Validation Accuracy	Training Loss	Training Accuracy	AUC	Sensitivity	Specificity
Model 1	0.42	0.59	0.89	0.34 (0.14 - 0.56)	0.47 (0.28 - 0.68)	0.44 (0.14 - 0.73)
Model 2	0.63	0.26	1.00	0.42 (0.22 - 0.61)	0.42 (0.22 - 0.61)	0.33 (0.09 - 0.60)
Model 3	**0.67**	0.82	0.63	0.47 (0.24 - 0.72)	0.47 (0.27 - 0.67)	0.44 (0.17 - 0.75)
Model 4	0.58	0.53	0.88	**0.62 (0.42 - 0.82)**	0.57 (0.35 - 0.76)	0.67 (0.38 - 0.92)
Model 5	0.56	0.12	1.00	0.52 (0.30 - 0.73)	0.43 (0.22 - 0.67)	0.44 (0.17 - 0.73)

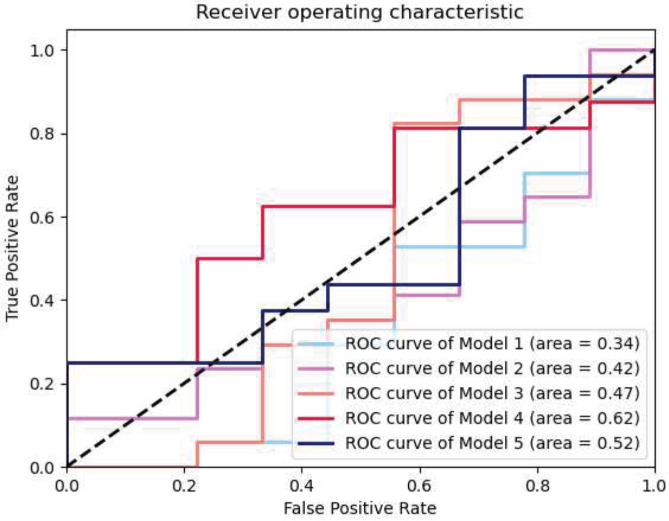
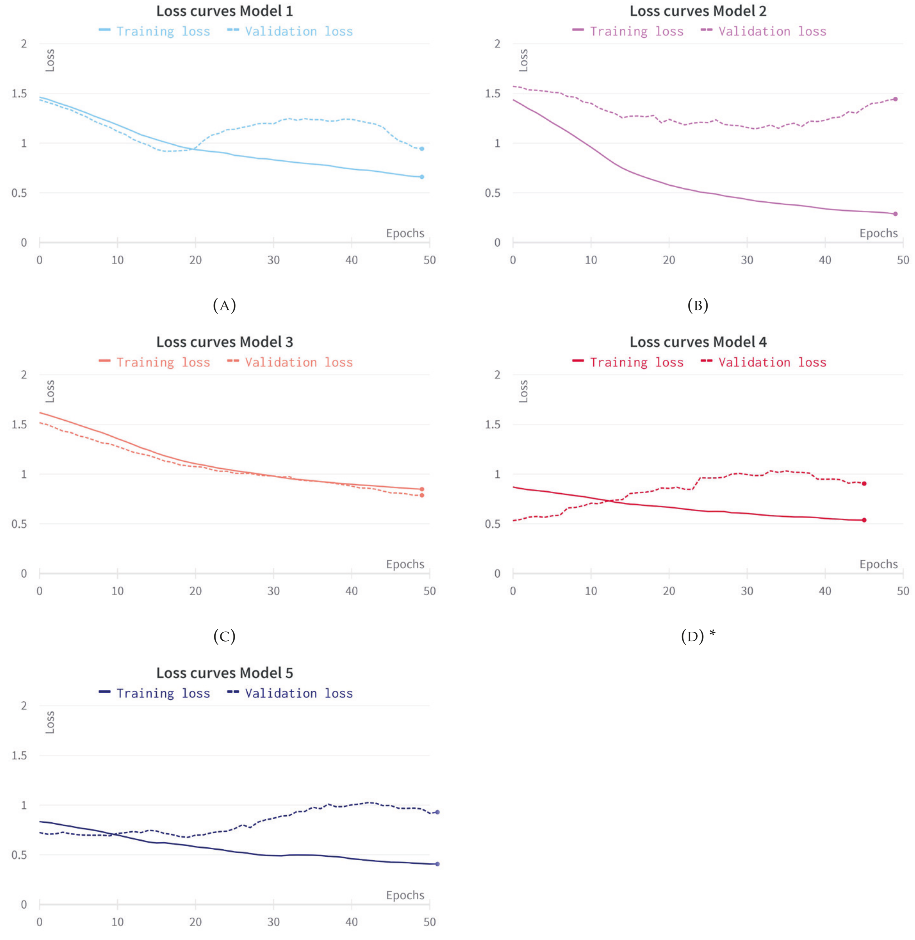


FIGURE 4.2: ROC curves of the 3D ResNet-18 models

compared to the other models.

In Figure 4.3, the loss curves of the training and validation set for each model were plotted. It can be seen that the validation loss increased at some point, while the training loss continued to decrease. This indicates overfitting of the model. Only model 3 did not show signs of overfitting.

In Figure 4.4, the gradient-based saliency maps were presented, where the tumour region is marked with a white circle. The highlighted areas on the image represent the data which is relevant for classification. It can be seen that, for models 1, 2, 4 and 5, there was a highlighted area around the bladder wall which contains tumour lesion. Model 1 showed also



(A)

(B)

(C)

(D) *

(E)

FIGURE 4.3: Loss curves of the training and validation for each model. * For model 4 it can be seen that the 50 epochs were not reached, this was because early-stopping had been applied for this model.

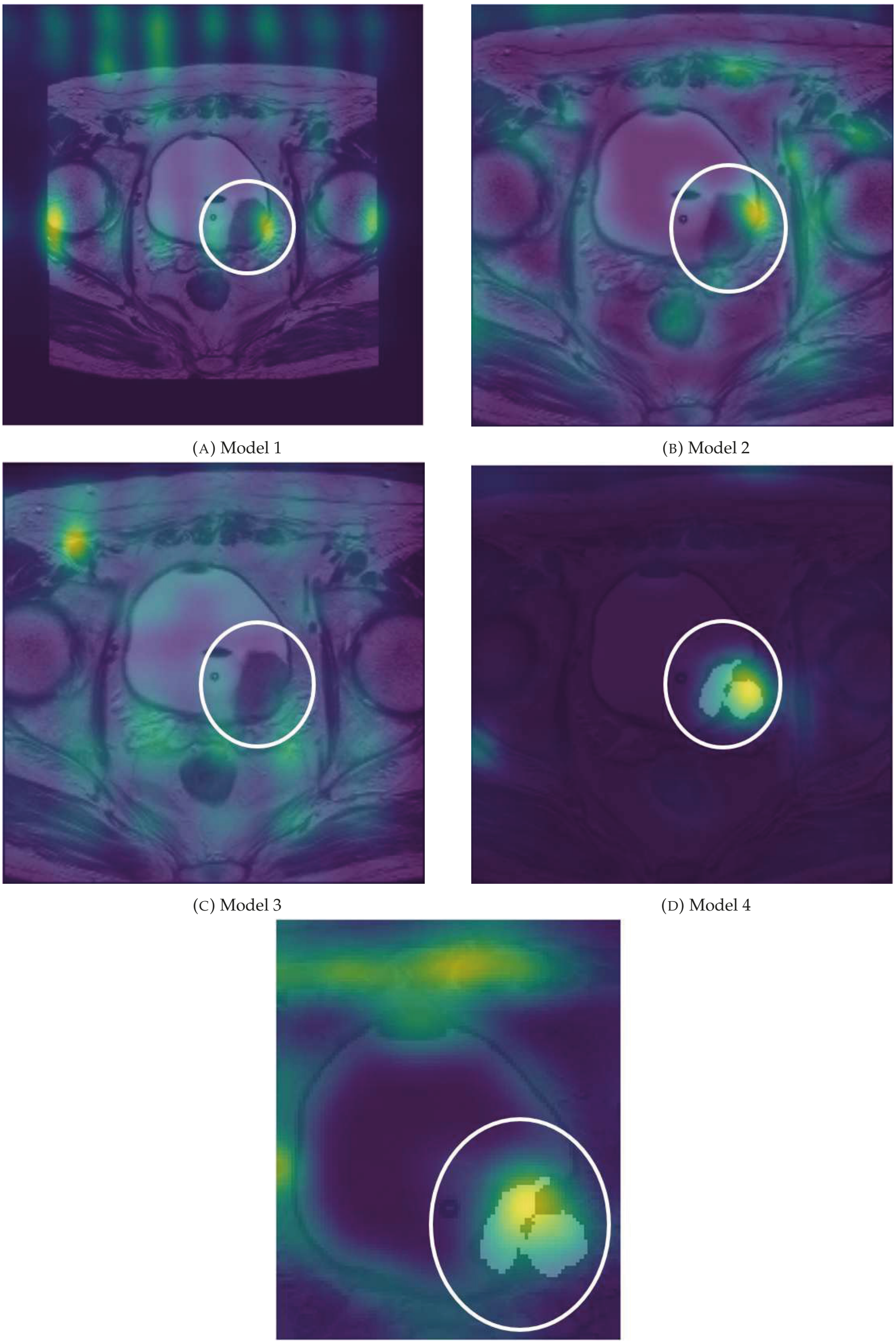
highlighted areas around the femur heads. Models 2 and 5 showed also highlighted areas around the pelvic musculature. Besides this, model 5 also showed a slightly highlighted area around the whole bladder wall. Model 4 showed solely a highlighted area within the ROI segmentation. At last, model 3 did not show any highlighted areas within the tumour region, only some in fatty tissue.

4.4 Discussion

A CNN, with ResNet-18 architecture, was implemented and trained on 3D, T2-weighted, on-treatment pelvic MRI scans. The goal was to classify the pathological outcome, pCR or non-pCR, for patients with bladder cancer, treated with neoadjuvant immunotherapy. A classification model was trained to determine the presence of remaining tumour cells after neoadjuvant immunotherapy, to classify the pathological outcome.

4.4.1 Model performance

Despite different setups, the models continued to overfit on the training data and therefore showed poor results on the validation set. To address overfitting, on the one hand, adjustment had been made regarding the input data; different input geometries, data augmentation and cropping of the scans to a bounding box based on the ROI segmentations. On the other hand, the ResNet-18 architecture was slightly adjusted by adding drop-out layers and network parameters, like normalisation and L2 regularisation, were modified.[52,53] The first three models were only trained on the T2-weighted MRI scans. According to the validation loss, models 1 and 2 were overfitting on the test data. Multiple network parameters were modified. Model 3, which was trained with data augmentation, standard drop-out (30%), spatial drop-out (30%) and batch normalisation, showed the best performance according to the validation loss and -accuracy (0.67). Compared to models 1 and 2, modifying network parameters prevented overfitting, but model performance remained poor. When looking at the AUC scores for classifying pathological response, model 3 had the highest score (0.47 compared to 0.34 and 0.42 for respectively model 1 and 2), however, these scores implicate



(A) Model 1

(B) Model 2

(C) Model 3

(D) Model 4

(E) Model 5

FIGURE 4.4: Gradient-based saliency maps show highlighted areas on the MRI scan which are relevant for the classification of pathological response. The tumour region is marked with a white circle.

that classification is still as good as random guessing. The heatmap plots for models 1 (Figure 4.4a) and 2 (Figure 4.4b) showed that there was at least a highlighted area within the tumour region, while for model 3 (Figure 4.4c) there was not. For model 3 we were able to prevent overfitting, but when visualising this network, the model did not recognize the tumour as a relevant area for classification.

Additionally, two models were trained with the radiological segmentations as a second input channel. According to the validation loss, both models were overfitting on the training data, which can also be seen for the validation accuracy (0.58 for model 4 and 0.56 for model 5). When looking at the AUC scores for the classification of pathological outcomes, model 4 showed a higher AUC score compared to model 5 (0.62 AUC vs 0.52 AUC). The AUC score of model 4 was slightly better than just random guessing, but still, the performance for making a distinction between pCR and non-pCR is weak. It should be noticed that none of the calculated AUC scores was significant. The heatmap for model 4 (Figure 4.4d) showed solely a highlighted area within the tumour region. The model recognised the tumour region as relevant for classification, yet it is not able to make the distinction between pCR and non-pCR. The heatmap for model 5 (Figure 4.4e) showed, besides the highlighted area within the tumour region, highlighted areas around the pelvic musculature as well.

When adding the segmentations as a second input channel, the predictive performance for pathological outcome slightly increased but is still too close to random guessing. The only model that was not overfitting, which was model 3, showed no highlighted areas on the heatmap within the tumour region. While the highlighted areas of model 4 were perfectly located within the tumour region, it had no predictive value for classifying pathological outcomes.

4.4.2 Limitations and Outlook

The trained CNN models showed poor performance according to the classification of pathological outcome. The main limitation of this study was the small dataset, only n=57 scans were included for model training. To address overfitting and increase model performance, we should add more training data. A way to increase the dataset might be by changing the inclusion criteria. This study focused on the response of immunotherapy, but for classification of the pathological outcome we can, for example, include neoadjuvant chemotherapy as well. Once we can classify pathological response, we can dive deeper into the image features which are specific for immunotherapy response.

Other options to address overfitting can be divided into multiple classes. The first class is about model inputs, such as data augmentation. With data augmentation, the training data is enlarged by cropping/rotating/mirroring the data. Second, model architecture and model parameters can be modified, such as adding drop-out layers and normalisation. Finally, we can also look into the models' output, for example, the regularisation of the learning rate can be modified.[52] In this study, we performed some first trial-and-error with the above-mentioned methods, but the lack of training data remained the Achilles heel of this study. By applying the different methods, we compared the validation loss with the validation loss of the model where nothing was done to prevent overfitting. It could be seen that, when applying methods to address overfitting, the validation loss increased at a later epoch or that the validation loss increased at a slower rate. Unfortunately, we were not able to fully prevent overfitting, which can be seen in the validation accuracy (Table 4.2). For further optimisation of this model, we should first include more training samples and then address overfitting again according to the above-mentioned methods.

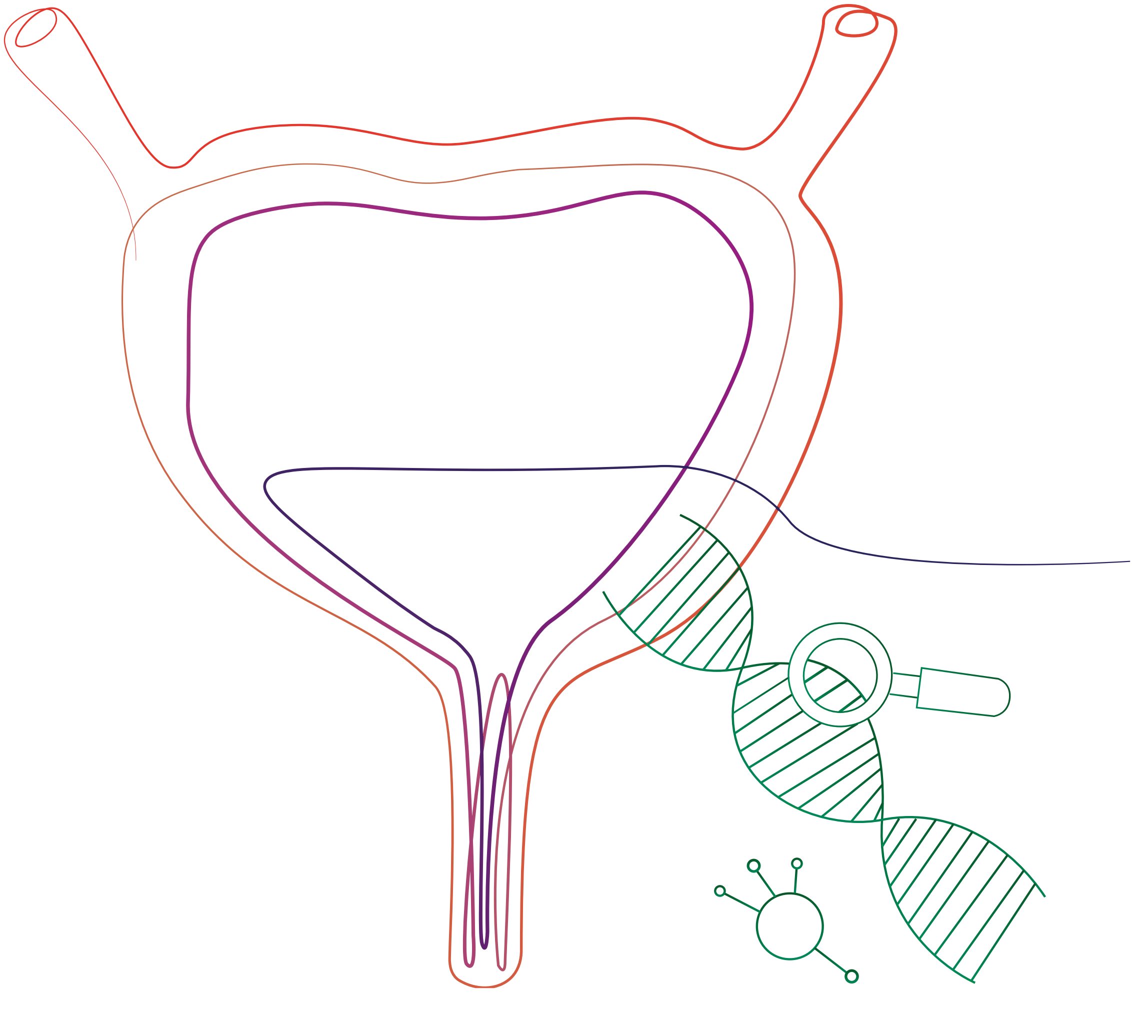
4.5 Conclusion

A 3D CNN, with Resnet-18 architecture, was trained on T2-weighted, on-treatment MRI scans, to classify the pathological outcome in patients with MIBC treated with neoadjuvant immunotherapy. Despite different network modifications, the models continued overfitting on the training data. Therefore the models do not generalise on new, unseen data. More training data should be included to address overfitting.

Further, it can be seen that adding segmentation as a second channel shows good results according to the saliency map. The model recognises the tumour region as relevant for classification, but cannot yet distinguish between pCR and non-pCR.

Chapter 5

General Discussion



This thesis aimed to perform therapeutic response assessment of neoadjuvant immunotherapy for patients with MIBC. Clinical studies showed that neoadjuvant immunotherapy is promising, but until now there are no biomarkers that sufficiently predict whether a patient will benefit from neoadjuvant immunotherapy. From the clinical perspective, there are two main research questions; can we predict whether a patient will benefit from neoadjuvant immunotherapy, based on the diagnostic MRI scan, and can we classify the pathological outcome after receiving treatment and therefore give an indication if organ-sparring treatment (e.g. chemo-radiation) is a possibility instead of radical cystectomy after neoadjuvant immunotherapy.

5.1 Limitations

The work in this thesis contains limitations. First, the dataset had a low sample size. For nnU-Net training we could only include n=72 patients, so training was performed on 114 scans. This resulted in model overfitting and therefore poor performance for new, unseen data. Statistical analysis could only be performed on n=25 patients from the PURE-01 test set and n=19 for the NABUCCO external validation set. Due to the small dataset, we introduced Monte Carlo Cross Validation to reduce the variability of the model and for the CNN we addressed overfitting with multiple techniques. Nevertheless, model performance will increase and the results will be more reliable with a bigger dataset.

Second, the labels used for model training were an overestimation of the actual tumour volume. Tumour delineation on baseline scans was complicated by the preceding TURBT, which made it difficult to distinguish specific tumour tissue and post-operative inflammatory response of the TURBT. Besides, for the segmentations on on-treatment scans, the radiologist delineated an ROI, based on the previously delineated ROI on the baseline scan. The segmentation was slightly adjusted to the anatomical structures, but no specific (residual) tumour segmentation was created.

5.2 Recommendations

For further studies, we would like to add more training data. For this study, we limited the included data to patients treated with neoadjuvant immunotherapy. To increase model performance for automatic tumour detection, we might consider adding more patients with MIBC, regardless of neoadjuvant treatment.

To overcome the second limitation, we advocate for standardisation of the segmentation process. With a standardised segmentation pipeline for new MRI scans, we can improve the accuracy and precision of the delineations. This also allows for several radiologists to segment cases, making the model more robust, since only one radiologist delineates the tumour ROI for this study. We could also add an iterative revision pipeline, where the radiologists adjust the models' predicted segmentations. The model can be retrained with these improved segmentations. By adding more data and applying a standardised segmentation pipeline, we aim for better model performance. Increasing the segmentation accuracy might also increase the predictive performance for the radiomic features.

We trained an automatic segmentation model on five mpMRI sequences. All five sequences were added as separate channels, therefore it is important that the geometry of anatomical structures, represented on each sequence, is similar and that each voxel represents the same anatomical information. Especially for the bladder, the anatomical location can vary during sequence acquisition, since it is a very deformable organ. Optimising sequence alignment ensures that the voxels in different channels will represent the same anatomical information, and will therefore result in more precise segmentations.

Predictive performance for the diagnostic MRI scan was lower compared to the on-treatment and difference over time. Results for volume and radiomic features showed low AUC scores, and are therefore poor in predicting pathological outcomes. It remains challenging to predict whether a patient will benefit from neoadjuvant immunotherapy. Future studies might consider a multi-modality approach, we could add, for example, genetic immunological biomarkers PD-L1 and CD8+ or look more into the predictive value of CT scans.[6]

The clinical goal would be to have a model, that can perform a prediction for the most suitable therapy. The outcome would be an indication for each possible neoadjuvant therapy, expressed in percentages, considering pathological outcome or overall survival.

Results for both on-treatment volume (remaining tumour volume) and the volume difference over time (tumour regression), as a predictive biomarker for pathological response assessment, are promising. Although, the predictive performance should be optimised, considering the low sensitivity. When looking at organ sparring treatment after immunotherapy, remaining tumour volume and tumour regression rate could be taken into account. Further clinical studies should determine the prognostic role of pathological staging and tumour regression in clinical practice. For the radiomic features, further studies might focus on the distinction between predictive features and prognostic features.[54]

Current clinical protocol, for patients with MIBC, consists of neoadjuvant chemotherapy followed by a radical cystectomy, regardless of the chemotherapy response. If the model would be able to classify pathological outcomes, we might have a good indication for organ-sparring treatment after neoadjuvant therapy. Clinical studies should define which treatment is suitable, according to the survival/recurrence rate.

Before introducing AI into clinical practice, we should compare the AI results against radiological assessment. A qualitative radiological assessment of therapeutic response and residual disease can be performed, which should be compared to the AI performance. Subsequently, the added value of AI within the current clinical protocol should be quantified.

Appendix A

Model performance tables

Model 1

TABLE A.1: Model performance for model 1. Model 1 is trained on PURE-01 baseline scans, performance is analysed over the PURE-01 test set (baseline and on-treatment) and the external validation set with NABUCCO. The volumes are expressed in cm^3, the Hausdorff distance in *mm* and the Dice Score with a confidence interval of 95%.

	Mean volume Ground truth (cm3)	Mean volume Prediction (cm3)	Dice score	Hausdorff distance (mm)	Pearson's r
PURE-01 baseline	5.906	3.359	0.39 (0.09 - 0.46)	37.11	0.499
PURE-01 on-treatment	10.790	4.940	0.30 (0.19 - 0.40)	31.35	0.693
NABUCCO baseline	34.977	39.067	0.21 (0.03 - 0.37)	136.76	0.787
NABUCCO on-treatment	10.367	17.418	0.00 (0.00 - 0.17)	124.34	0.090

Model 2

TABLE A.2: Model performance for model 2. Model 1 is trained on both PURE-01 baseline and on-treatment scans, performance is analysed over the PURE-01 test set (baseline and on-treatment) and the external validation set with NABUCCO. The volumes are expressed in cm^3, the Hausdorff distance in *mm* and the Dice Score with a confidence interval of 95%.

	Mean volume Ground truth (cm3)	Mean volume Prediction (cm3)	Dice score	Hausdorff distance (mm)	Pearson's r
PURE-01 baseline	7.992	7.192	0.33 (0.20 - 0.41)	33.46	0.414
PURE-01 on-treatment	12.909	6.901	0.52 (0.11 - 0.71)	30.12	0.896
NABUCCO baseline	34.977	0.812	0.00 (0.00 - 0.00)	83.21	0.736
NABUCCO on-treatment	10.367	0.377	0.00 (0.00 - 0.00)	106.59	-0.406

Model 3

TABLE A.3: Model performance for model 3. Model 3 is trained on both the PURE-01 and NABUCCO dataset. Performance is analysed over the PURE-01 test set and the overall model performance is presented, since the NABUCCO test set has too few datapoints for statistical analysis. No external validation set is used.The volumes are expressed in cm^3, the Hausdorff distance in *mm* and the Dice Score with a confidence interval of 95%.

	Mean volume Ground truth (cm3)	Mean volume Prediction (cm3)	Dice score	Hausdorff distance (mm)	Pearson's r
PURE-01 baseline	7.992	6.556	0.33 (0.22 - 0.46)	28.55	0.364
PURE-01 on-treatment	12.909	7.090	0.50 (0.11 - 0.72)	32.05	0.888
Overall Test set	10.605	6.086	0.34 (0.23 - 0.49)	34.32	0.517

Model 4

TABLE A.4: Model performance for model 4. Model 4 is trained on both PURE-01 baseline and on-treatment scans, where the labelling is adjusted to the pathological staging (if pCR, the segmentation volume is set to zero). Performance is analysed over the PURE-01 test set (baseline and on-treatment) and the external validation set with NABUCCO. The volumes are expressed in cm^3, the Hausdorff distance in *mm* and the Dice Score with a confidence interval of 95%.

	Mean volume Ground truth (cm3)	Mean volume Prediction (cm3)	Dice score	Hausdorff distance (mm)	Pearson's r
PURE-01 baseline	7.992	5.604	0.33 (0.20 - 0.42)	24.64	0.299
PURE-01 on-treatment	10.022	6.924	0.61 (0.32 - 0.71)	26.75	0.858
NABUCCO baseline	34.977	2.729	0.00 (0.00 - 0.00)	105.24	0.186
NABUCCO on-treatment	10.367	2.187	0.00 (0.00 - 0.00)	114.70	0.594

Model 5

TABLE A.5: Model performance for model 5. Model 5 is trained on preprocessed PURE-01 baseline scans, performance is analysed over the preprocessed PURE-01 test set (baseline and on-treatment) and the external validation set with NABUCCO. The volumes are expressed in cm^3, the Hausdorff distance in mm and the Dice Score with a confidence interval of 95%.

	Mean volume Ground truth (cm3)	Mean volume Prediction (cm3)	Dice score	Hausdorff distance (mm)	Pearson's r
PURE-01 baseline	6.621	2.147	0.31 (0.13 - 0.43)	22.77	0.051
PURE-01 on-treatment	10.817	2.749	0.18 (0.06 - 0.27)	29.98	0.847
NABUCCO baseline	29.588	7.614	0.34 (0.00 - 0.46)	51.05	0.037
NABUCCO on-treatment	7.344	6.562	0.11 (0.02 - 0.43)	52.68	0.561

Model 6

TABLE A.6: Model performance for model 6. Model 6 is trained on both preprocessed PURE-01 baseline and on-treatment scans, performance is analysed over the preprocessed PURE-01 test set (baseline and on-treatment) and the external validation set with NABUCCO. The volumes are expressed in cm^3, the Hausdorff distance in mm and the Dice Score with a confidence interval of 95%.

	Mean volume Ground truth (cm3)	Mean volume Prediction (cm3)	Dice score	Hausdorff distance (mm)	Pearson's r
PURE-01 baseline	7.852	4.346	0.33 (0.23 - 0.54)	24.30	0.511
PURE-01 on-treatment	12.839	4.168	0.35 (0.19 - 0.49)	29.13	0.845
NABUCCO baseline	29.588	9.734	0.36 (0.00 - 0.61)	31.99	0.156
NABUCCO on-treatment	7.344	7.290	0.27 (0.00 - 0.40)	32.44	0.524

Model 7

TABLE A.7: Model performance for model 7. Model 7 is trained on both the preprocessed PURE-01 and NABUCCO dataset. Performance is analysed over the preprocessed PURE-01 test set and the overall model performance is presented, since the NABUCCO test set has too few datapoints for statistical analysis. No external validation set is used.The volumes are expressed in cm^3, the Hausdorff distance in mm and the Dice Score with a confidence interval of 95%.

	Mean volume Ground truth (cm3)	Mean volume Prediction (cm3)	Dice score	Hausdorff distance (mm)	Pearson's r
PURE-01 baseline	7.852	4.323	0.32 (0.28 - 0.49)	24.14	0.523
PURE-01 on-treatment	12.839	4.448	0.45 (0.14 - 0.56)	29.81	0.942
Overall Test set	14.535	5.769	0.36 (0.28 - 0.48)	28.17	0.801

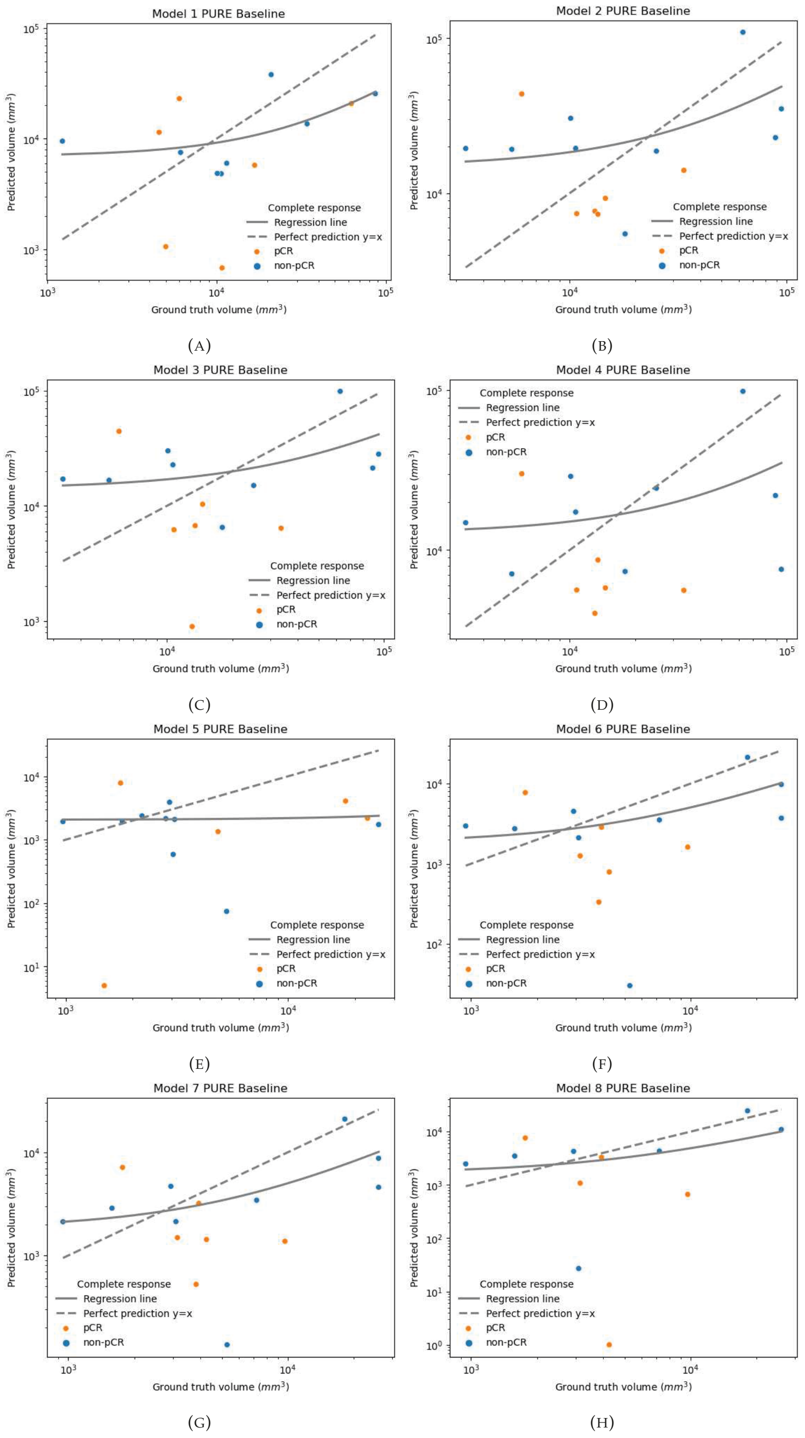
Model 8

TABLE A.8: Model performance for model 8. Model 8 is trained on both preprocessed PURE-01 baseline and on-treatment scans, where the labelling is adjusted to the pathological staging (if pCR, the segmentation volume is set to zero). Performance is analysed over the preprocessed PURE-01 test set (baseline and on-treatment) and the external validation set with NABUCCO. The volumes are expressed in cm^3, the Hausdorff distance in mm and the Dice Score with a confidence interval of 95%.

	Mean volume Ground truth (cm3)	Mean volume Prediction (cm3)	Dice score	Hausdorff distance (mm)	Pearson's r
PURE-01 baseline	7.852	4.191	0.30 (0.01 - 0.35)	27.28	0.424
PURE-01 on-treatment	9.693	3.689	0.45 (0.17 - 0.59)	17.09	0.788
NABUCCO baseline	29.589	7.763	0.28 (0.00 - 0.47)	29.37	0.205
NABUCCO on-treatment	7.344	6.959	0.09 (0.00 - 0.31)	60.94	0.545

Appendix B

Scatterplots



(A)

(B)

(C)

(D)

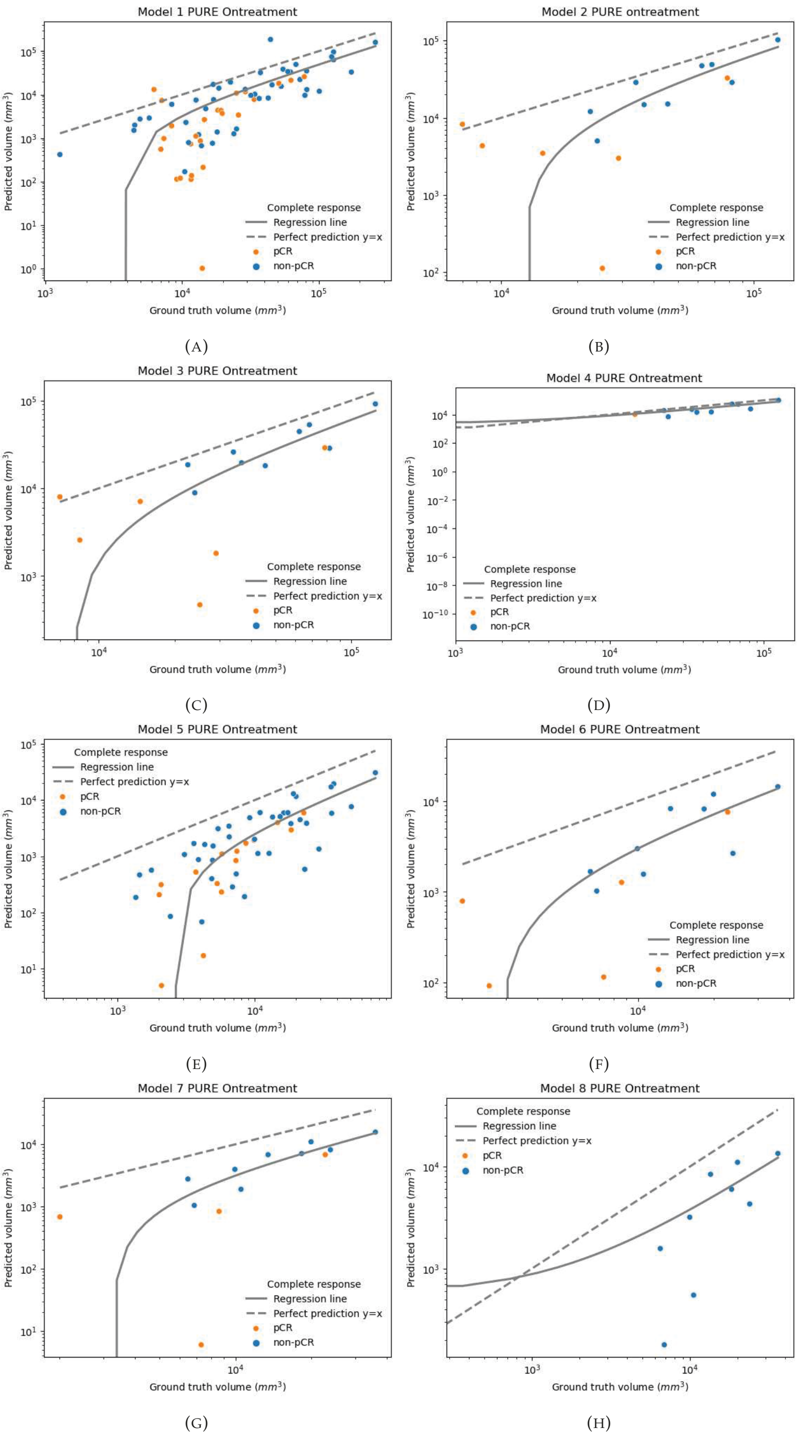
(E)

(F)

(G)

(H)

FIGURE B.1: Scatterplots of the PURE-01 baseline ground truth and predicted segmentation volumes (mm^3)



(A)

(B)

(C)

(D)

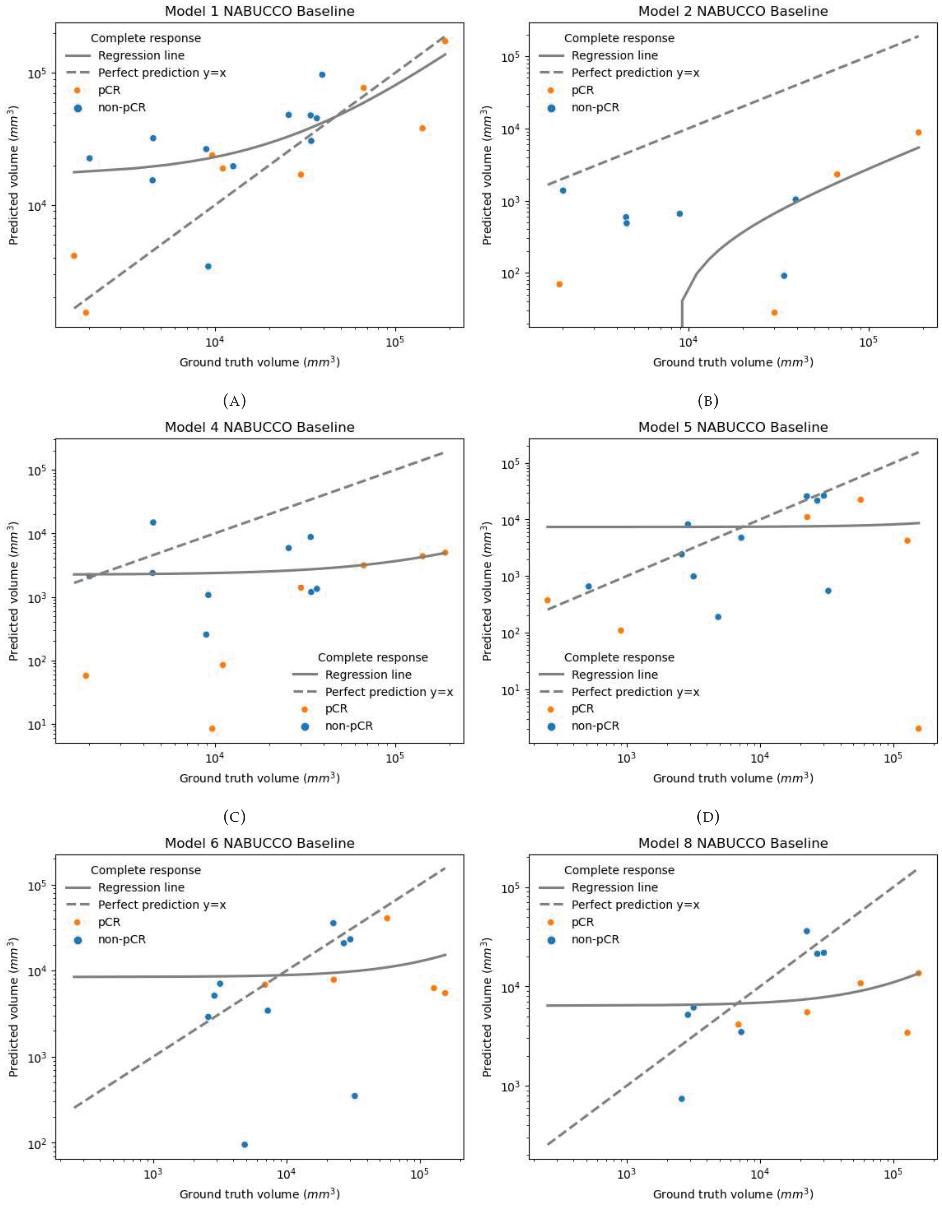
(E)

(F)

(G)

(H)

FIGURE B.2: Scatterplots of the PURE-01 ontreatment ground truth and predicted segmentation volumes (mm^3)



(A)

(B)

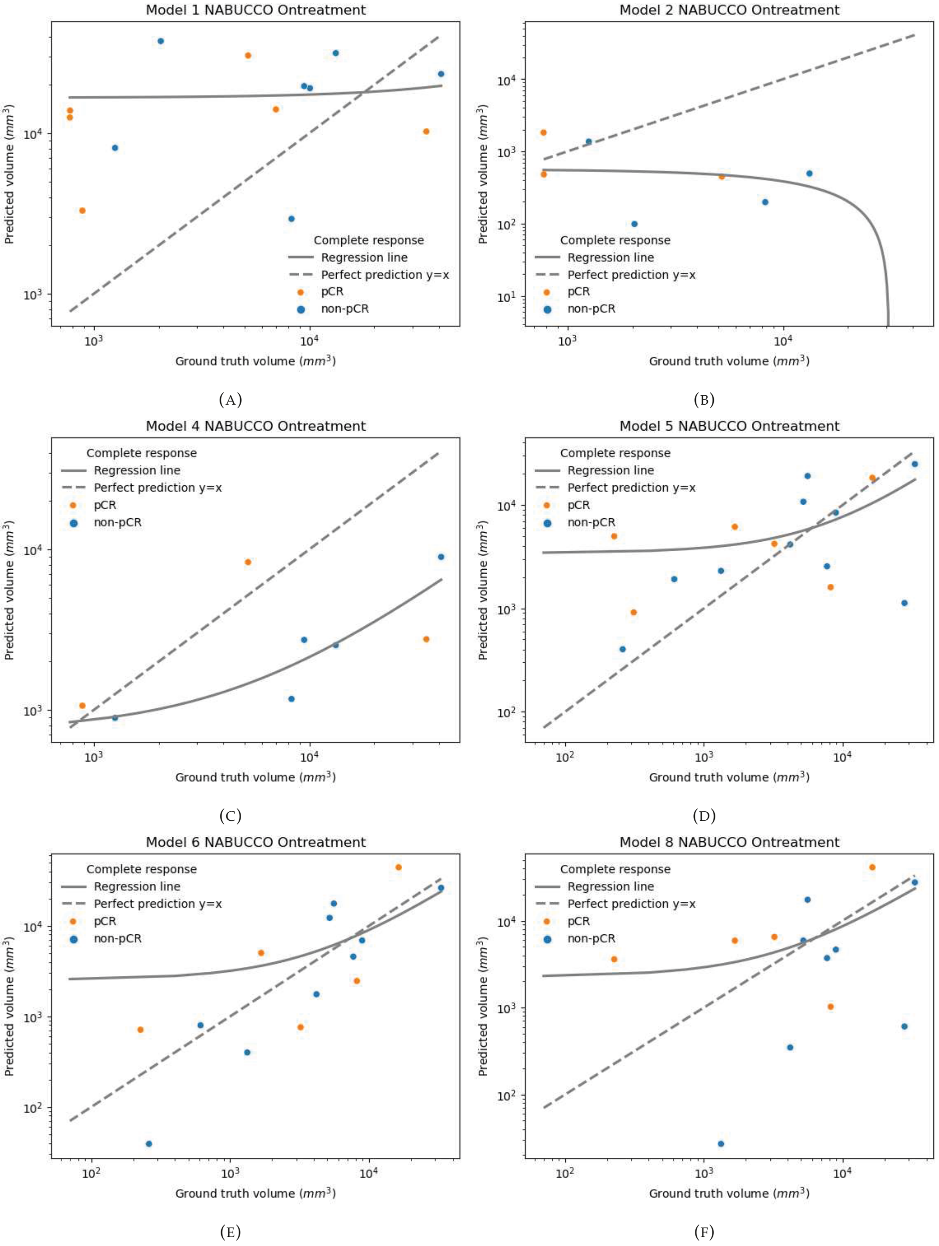
(C)

(D)

(E)

(F)

FIGURE B.3: Scatterplots of the NABUCCO baseline ground truth and predicted segmentation volumes (mm^3)



(A)

(B)

(C)

(D)

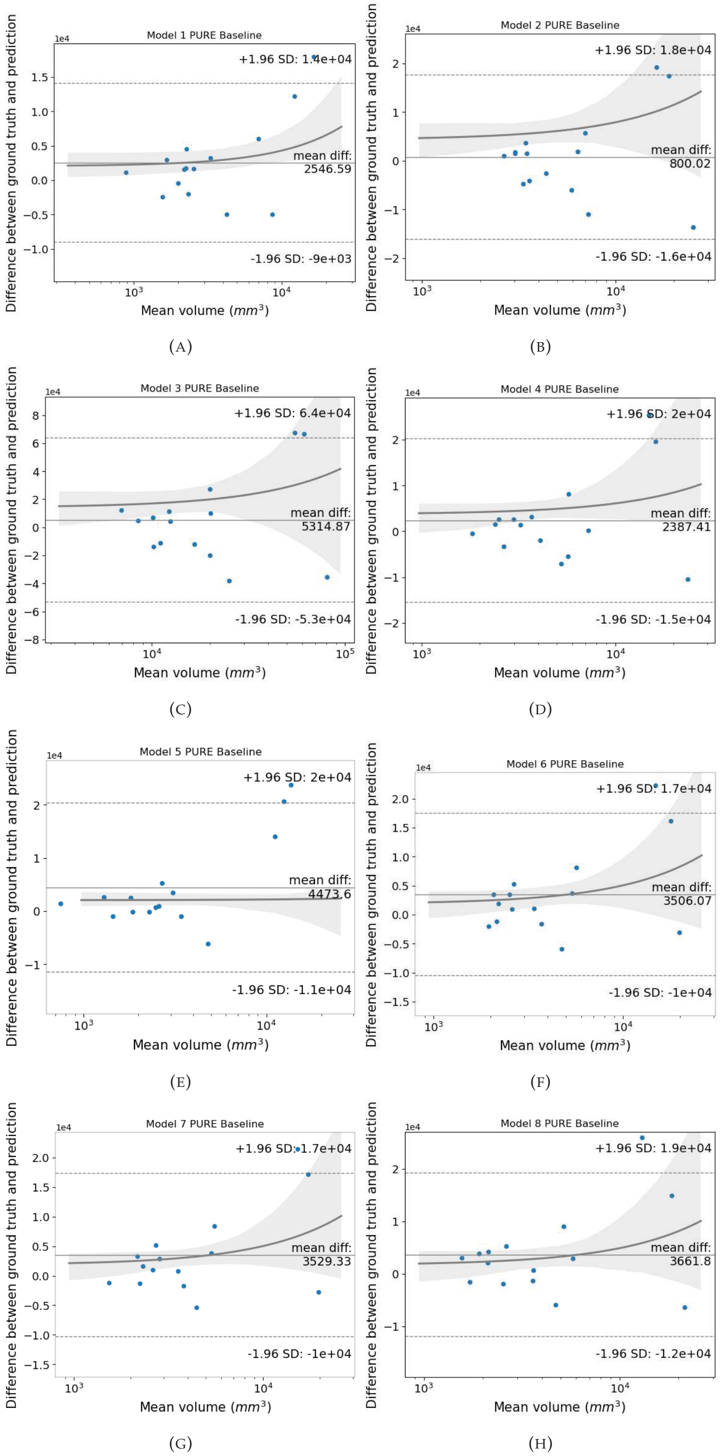
(E)

(F)

FIGURE B.4: Scatterplots of the NABUCCO ontreatment ground truth and predicted segmentation volumes (mm^3)

Appendix C

Bland-Altman plots



(A)

(B)

(C)

(D)

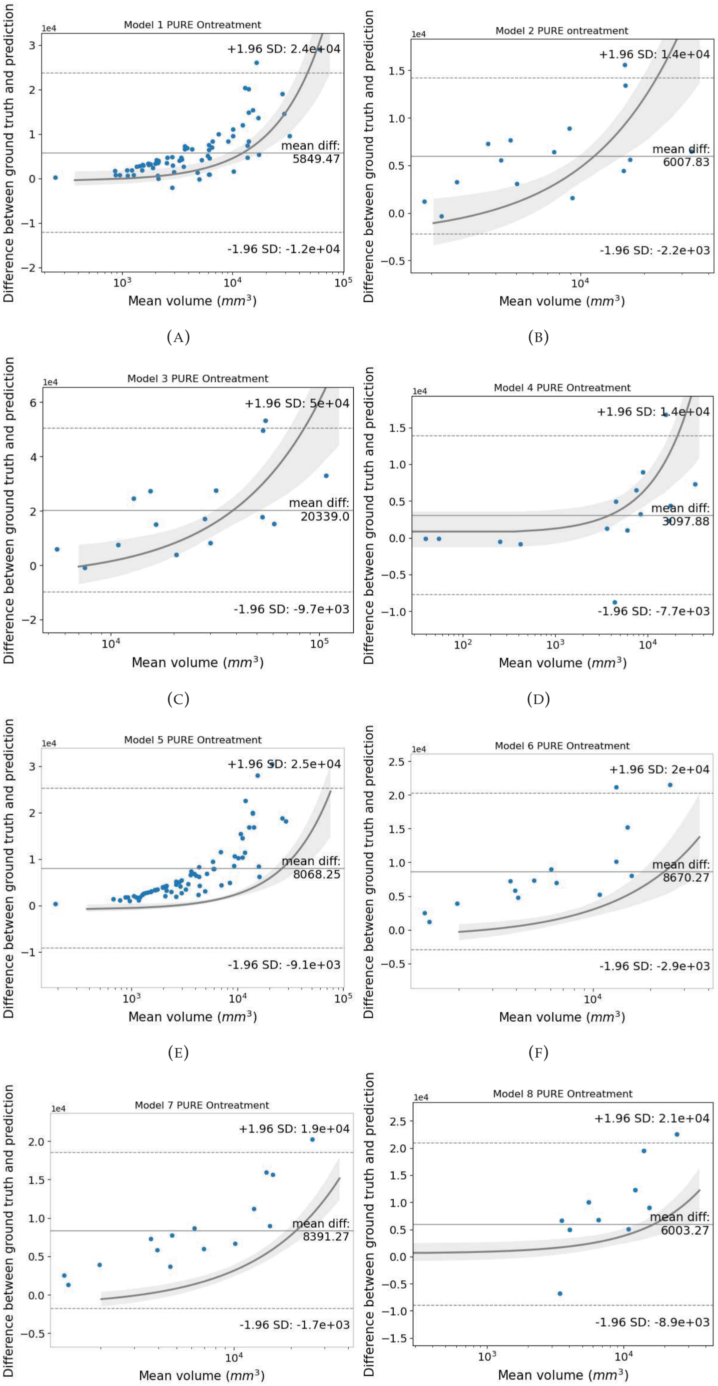
(E)

(F)

(G)

(H)

FIGURE C.1: Bland-Altman plots of the PURE-01 baseline (test set) ground truth and predicted segmentation volumes (mm^3)



(A)

(B)

(C)

(D)

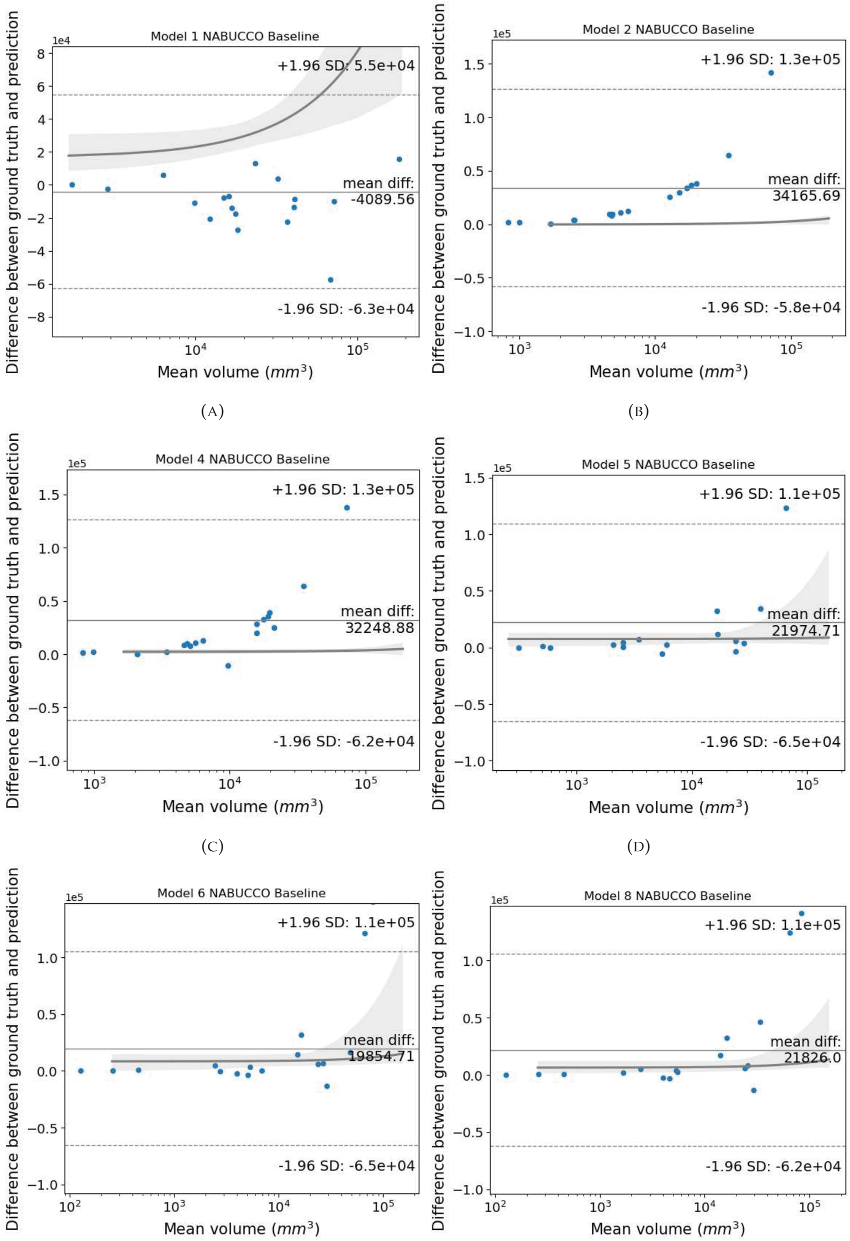
(E)

(F)

(G)

(H)

FIGURE C.2: Bland-Altman plots of the PURE-01 ontreatment (test set) ground truth and predicted segmentation volumes (mm^3)



(A)

(B)

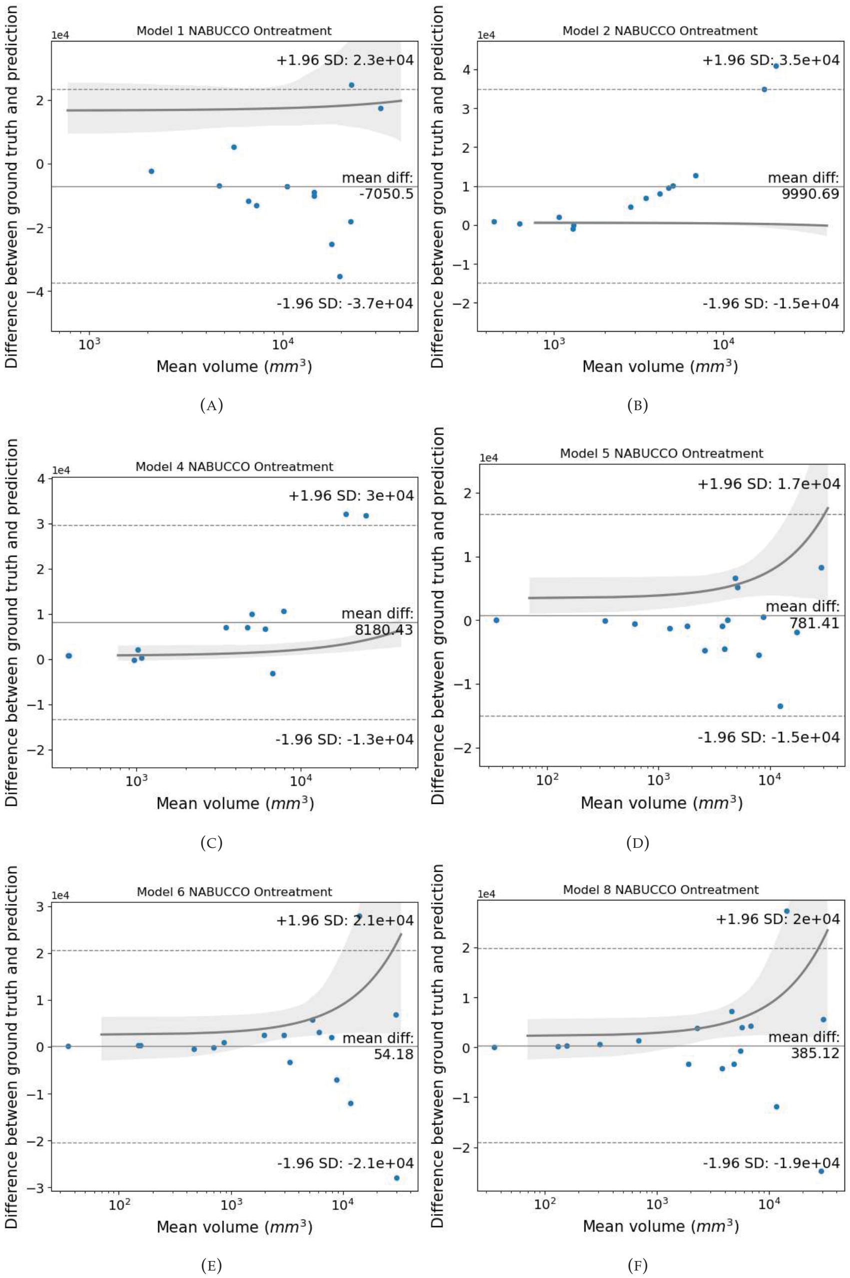
(C)

(D)

(E)

(F)

FIGURE C.3: Bland-Altman plots of the NABUCCO baseline (external validation) ground truth and predicted segmentation volumes (mm^3)



(A)

(B)

(C)

(D)

(E)

(F)

FIGURE C.4: Bland-Altman plots of the NABUCCO ontreatment (external validation) ground truth and predicted segmentation volumes (mm^3)

Bibliography

[1] M. Wołącewicz, R. Hrynkiewicz, E. Grywalska, T. Suchojad, T. Leksowski, J. Roliński, and P. Niedźwiedzka-Rystwej, “Immunotherapy in Bladder Cancer: Current Methods and Future Perspectives,” *Cancers 2020, Vol. 12, Page 1181*, vol. 12, p. 1181, 5 2020.

[2] “Cystectomy for Bladder Cancer — HealthLink BC.”

[3] F. Rundo, G. L. Banna, F. Trenta, C. Spampinato, L. Bidaut, X. Ye, S. Kollias, and S. Battiato, “Advanced Non-linear Generative Model with a Deep Classifier for Immunotherapy Outcome Prediction: A Bladder Cancer Case Study,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12661 LNCS, pp. 227–242, 1 2021.

[4] F. Rundo, C. Spampinato, G. L. Banna, and S. Conoci, “Advanced Deep Learning Embedded Motion Radiomics Pipeline for Predicting Anti-PD-1/PD-L1 Immunotherapy Response in the Treatment of Bladder Cancer: Preliminary Results,” *Electronics 2019, Vol. 8, Page 1134*, vol. 8, p. 1134, 10 2019.

[5] Q. Song, J. D. Seigne, A. R. Schned, K. T. Kelsey, M. R. Karagas, and S. Hassanpour, “A Machine Learning Approach for Long-Term Prognosis of Bladder Cancer based on Clinical and Molecular Features,” *AMIA Summits on Translational Science Proceedings*, vol. 2020, p. 607, 2020.

[6] N. van Dijk, A. Gil-Jimenez, K. Silina, K. Hendricksen, L. A. Smit, J. M. de Feijter, M. L. van Montfoort, C. van Rooijen, D. Peters, A. Broeks, H. G. van der Poel, A. Bruining, Y. Lubeck, K. Sikorska, T. N. Boellaard, P. Kvistborg, D. J. Vis, E. Hooijberg, T. N. Schumacher, M. van den Broek, L. F. Wessels, C. U. Blank, B. W. van Rhijn, and M. S. van der Heijden, “Preoperative ipilimumab plus nivolumab in locoregionally advanced urothelial cancer: the NABUCCO trial,” *Nature Medicine 2020 26:12*, vol. 26, pp. 1839–1844, 10 2020.

[7] A. Necchi, A. Anichini, D. Raggi, A. Briganti, S. Massa, R. Lucianò, M. Colecchia, P. Giannatempo, R. Mortarini, M. Bianchi, E. Farè, F. Monopoli, R. Colombo, A. Gallina, A. Salonia, A. Messina, S. M. Ali, R. Madison, J. S. Ross, J. H. Chung, R. Salvioni, L. Mariani, and F. Montorsi, “Pembrolizumab as Neoadjuvant Therapy Before Radical Cystectomy in Patients With Muscle-Invasive Urothelial Bladder Carcinoma (PURE-01): An Open-Label, Single-Arm, Phase II Study,” *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, vol. 36, pp. 3353–3360, 12 2018.

[8] K. H. Cha, L. Hadjiiski, H. P. Chan, A. Z. Weizer, A. Alva, R. H. Cohan, E. M. Caoili, C. Paramagul, and R. K. Samala, “Bladder Cancer Treatment Response Assessment in CT using Radiomics with Deep-Learning,” *Scientific Reports 2017 7:1*, vol. 7, pp. 1–12, 8 2017.

[9] D. Song, T. Powles, L. Shi, L. Zhang, M. A. Ingersoll, and Y. J. Lu, “Bladder cancer, a unique model to understand cancer immunity and develop immunotherapy approaches,” *Journal of Pathology*, vol. 249, pp. 151–165, 10 2019.

[10] U. Pinar, B. Pradere, and M. Roupret, “Artificial intelligence in bladder cancer prognosis: a pathway for personalized medicine,” *Current opinion in urology*, vol. 31, pp. 404–408, 7 2021.

[11] S. Trebeschi, S. G. Drago, N. J. Birkbak, I. Kurilova, A. M. Călin, A. Delli Pizzi, F. Lalezari, D. M. Lambregts, M. W. Rohaan, C. Parmar, E. A. Rozeman, K. J. Hartemink, C. Swanton, J. B. Haanen, C. U. Blank, E. F. Smit, R. G. Beets-Tan, and H. J. Aerts, "Predicting response to cancer immunotherapy using noninvasive radiomic biomarkers," *Annals of oncology : official journal of the European Society for Medical Oncology*, vol. 30, pp. 998–1004, 6 2019.

[12] K. Bera, N. Braman, A. Gupta, V. Velcheti, and A. Madabhushi, "Predicting cancer outcomes with radiomics and artificial intelligence in radiology," *Nature Reviews Clinical Oncology 2021*, pp. 1–15, 10 2021.

[13] K. Hammouda, F. Khalifa, A. Soliman, M. Ghazal, M. A. El-Ghar, M. A. Badawy, H. E. Darwish, A. Khelifi, and A. El-Baz, "A multiparametric MRI-based CAD system for accurate diagnosis of bladder cancer staging," *Computerized Medical Imaging and Graphics*, vol. 90, p. 101911, 6 2021.

[14] Z. Kirkali, T. Chan, M. Manoharan, F. Algaba, C. Busch, L. Cheng, L. Kiemeney, M. Kriegmair, R. Montironi, W. M. Murphy, I. A. Sesterhenn, M. Tachibana, and J. Weider, "Bladder cancer: Epidemiology, staging and grading, and diagnosis," *Urology*, vol. 66, pp. 4–34, 12 2005.

[15] M. J. Magers, A. Lopez-Beltran, R. Montironi, S. R. Williamson, H. Z. Kaimakliotis, and L. Cheng, "Staging of bladder cancer," *Histopathology*, vol. 74, pp. 112–134, 1 2019.

[16] C. TM, C. TC, H. SK, Y. BW, L. WM, H. CN, L. CC, W. WJ, and L. CF, "Role of Microtubule-Associated Protein 1b in Urothelial Carcinoma: Overexpression Predicts Poor Prognosis," *Cancers*, vol. 12, 3 2020.

[17] N. Mottet, J. Bellmunt, E. Briers, R. van den Bergh, M. Bolla, N. van Casteren, P. Cornford, S. Culine, S. Joniau, T. Lam, M. Mason, V. Matveev, H. van Der Poel, T. van der Kwast, . Rouviere, T. Wiegel, and Members of the European Association of Urology (Eau) Guidelines, "Diretrizes EAU. Edn. apresentado no EAU Annual Congress Milan 2021.," *European Association of Urology*, 2021.

[18] "Bladder Cancer: Introduction — Cancer.Net."

[19] E. Salmanoglu, E. Halpern, E. J. Trabulsi, S. Kim, and M. L. Thakur, "A glance at imaging bladder cancer," *Clinical and translational imaging*, vol. 6, p. 257, 8 2018.

[20] J. O. Barentsz, G. J. Jager, P. B. Van Vierzen, J. A. Witjes, S. P. Strijk, H. Peters, N. Karssemeijer, and S. H. Ruijs, "Staging urinary bladder cancer after transurethral biopsy: value of fast dynamic contrast-enhanced MR imaging.," *https://doi.org/10.1148/radiology.201.1.8816542*, vol. 201, pp. 185–193, 10 1996.

[21] J. O. Barentsz, M. Engelbrecht, G. J. Jager, J. A. Witjes, J. De Larosette, B. P. J. Van Der Sanden, H.-J. Huisman, and A. Heerschap, "Fast Dynamic Gadolinium-Enhanced MR Imaging of Urinary Bladder and Prostate Cancer," *J. Magn. Reson. Imaging*, vol. 10, pp. 295–304, 1999.

[22] V. Panebianco, Y. Narumi, E. Altun, B. H. Bochner, J. A. Efstathiou, S. Hafeez, R. Huddart, S. Kennish, S. Lerner, R. Montironi, V. F. Muglia, G. Salomon, S. Thomas, H. A. Vargas, J. A. Witjes, M. Takeuchi, J. Barentsz, and J. W. Catto, "Multiparametric Magnetic Resonance Imaging for Bladder Cancer: Development of VI-RADS," *European urology*, vol. 74, p. 294, 9 2018.

[23] K. C. Sim and D. J. Sung, "Role of magnetic resonance imaging in tumor staging and follow-up for bladder cancer," *Translational Andrology and Urology*, vol. 9, p. 2890, 12 2020.

[24] "Intravesical Therapy for Bladder Cancer."

[25] S. M. Einerhand, N. van Dijk, J. van Dorp, J. M. de Feijter, M. L. van Montfoort, M. W. van de Kamp, E. E. Schaake, T. N. Boellaard, K. Hendricksen, M. S. van der Heijden, and B. W. van Rhijn, "Survival after neoadjuvant/induction combination immunotherapy vs combination platinum-based chemotherapy for locally advanced (Stage III) urothelial cancer," *International journal of cancer*, 2022.

[26] N. Kuol, L. Stojanovska, K. Nurgali, and V. Apostolopoulos, “The mechanisms tumor cells utilize to evade the host’s immune system,” *Maturitas*, vol. 105, pp. 8–15, 11 2017.

[27] A. M. Cornel, I. L. Mimpen, and S. Nierkens, “MHC Class I Downregulation in Cancer: Underlying Mechanisms and Potential Targets for Cancer Immunotherapy,” *Cancers*, vol. 12, pp. 1–33, 7 2020.

[28] O. Kooshkaki, A. Derakhshani, N. Hosseinkhani, M. Torabi, S. Safaei, O. Brunetti, V. Racanelli, N. Silvestris, and B. Baradaran, “Combination of Ipilimumab and Nivolumab in Cancers: From Clinical Practice to Ongoing Clinical Trials,” *International Journal of Molecular Sciences 2020, Vol. 21, Page 4427*, vol. 21, p. 4427, 6 2020.

[29] H. Yu, L. T. Yang, Q. Zhang, D. Armstrong, and M. J. Deen, “Convolutional neural networks for medical image analysis: State-of-the-art, comparisons, improvement and perspectives,” *Neurocomputing*, vol. 444, pp. 92–110, 7 2021.

[30] A. Necchi, M. Bandini, G. Calareso, D. Raggi, F. Pederzoli, E. Farè, M. Colecchia, L. Marandino, M. Bianchi, A. Gallina, R. Colombo, N. Fossati, G. Gandaglia, U. Capitanio, F. Dehò, P. Giannatempo, R. Lucianò, A. Salonia, R. Madison, S. M. Ali, J. H. Chung, J. S. Ross, A. Briganti, F. Montorsi, F. De Cobelli, and A. Messina, “Multiparametric Magnetic Resonance Imaging as a Noninvasive Assessment of Tumor Response to Neoadjuvant Pembrolizumab in Muscle-invasive Bladder Cancer: Preliminary Findings from the PURE-01 Study,” *European Urology*, vol. 77, pp. 636–643, 5 2020.

[31] K. H. Cha, L. M. Hadjiiski, R. K. Samala, H.-P. Chan, R. H. Cohan, E. M. Caoili, C. Paramagul, A. Alva, and A. Z. Weizer, “Bladder Cancer Segmentation in CT for Treatment Response Assessment: Application of Deep-Learning Convolution Neural Network—A Pilot Study,” *Tomography*, vol. 2, p. 421, 12 2016.

[32] 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 2020 18:2*, vol. 18, pp. 203–211, 12 2020.

[33] “nnU-Net : The no-new-UNet for automatic segmentation — by Prateek Gupta — MICCAI Educational Initiative — Medium.”

[34] W. Weng and X. Zhu, “U-Net: Convolutional Networks for Biomedical Image Segmentation,” *IEEE Access*, vol. 9, pp. 16591–16603, 5 2015.

[35] H. Yan, X. Zhou, X. Wang, R. Li, Y. Shi, Q. Xia, L. Wan, G. Huang, and J. Liu, “Delayed 18 F FDG PET/CT Imaging in the Assessment of Residual Tumors after Transurethral Resection of Bladder Cancer,” *Radiology*, vol. 293, no. 1, pp. 144–150, 2019.

[36] R. T. Shinohara, E. M. Sweeney, J. Goldsmith, N. Shiee, F. J. Mateen, P. A. Calabresi, S. Jarso, D. L. Pham, D. S. Reich, and C. M. Crainiceanu, “Statistical normalization techniques for magnetic resonance imaging,” *NeuroImage: Clinical*, vol. 6, pp. 9–19, 2014.

[37] L. G. Nyúl, J. K. Udupa, and X. Zhang, “New variants of a method of MRI scale standardization,” *IEEE transactions on medical imaging*, vol. 19, no. 2, pp. 143–150, 2000.

[38] “MIC-DKFZ/nnUNet.”

[39] R. W. Pettit, B. B. Marlatt, S. J. Corr, J. Havelka, and A. Rana, “nnU-Net Deep Learning Method for Segmenting Parenchyma and Determining Liver Volume From Computed Tomography Images,” *Annals of Surgery Open*, vol. 3, p. e155, 6 2022.

[40] A. A. Taha and A. Hanbury, “Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool,” *BMC Medical Imaging*, vol. 15, p. 29, 8 2015.

[41] Q. Gao, M. Ahn, and H. Zhu, “Cook’s Distance Measures for Varying Coefficient Models with Functional Responses,” *Technometrics : a journal of statistics for the physical, chemical, and engineering sciences*, vol. 57, p. 268, 4 2015.

[42] "Identifying Outliers in Linear Regression — Cook's Distance — by Christian Thieme — Towards Data Science."

[43] R. Dennis Cook, "American Society for Quality Detection of Influential Observation in Linear Regression Detection o f Influential Observation i n Linear Regression," *Source: Technometrics*, vol. 19, no. 1, pp. 15–18, 1977.

[44] W. R. Crum, T. Hartkens, and D. L. Hill, "Non-rigid image registration: Theory and practice," *British Journal of Radiology*, vol. 77, no. SPEC. ISS. 2, 2004.

[45] H. J. Aerts, E. R. Velazquez, R. T. Leijenaar, C. Parmar, P. Grossmann, S. Cavalho, J. Bussink, R. Monshouwer, B. Haibe-Kains, D. Rietveld, F. Hoebers, M. M. Rietbergen, C. R. Leemans, A. Dekker, J. Quackenbush, R. J. Gillies, and P. Lambin, "Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach," *Nature Communications 2014 5:1*, vol. 5, pp. 1–9, 6 2014.

[46] G. E. Cacciamani, N. Nassiri, B. Varghese, M. Maas, K. G. King, D. Hwang, A. Abreu, I. Gill, and V. Duddalwar, "Radiomics and Bladder Cancer: Current Status," *Bladder Cancer*, vol. 6, pp. 343–362, 1 2020.

[47] Y. Wang, J. Lang, J. Z. Zuo, Y. Dong, Z. Hu, X. Xu, Y. Zhang, Q. Wang, L. Yang, S. T. C. Wong, H. Wang, H. Li, and W. A. Cn, "The radiomic-clinical model using the SHAP method for assessing the treatment response of whole-brain radiotherapy: a multicentric study," *European Radiology 2022*, pp. 1–11, 6 2022.

[48] J. J. Van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R. G. Beets-Tan, J. C. Fillion-Robin, S. Pieper, and H. J. Aerts, "Computational radiomics system to decode the radiographic phenotype," *Cancer Research*, vol. 77, pp. e104–e107, 11 2017.

[49] "Radiomic Features — pyradiomics v3.0.1.post15+g2791e23 documentation."

[50] J. Virmani, G. P. Singh, Y. Singh, and Kriti, "PNN-based classification of retinal diseases using fundus images," *Sensors for Health Monitoring*, pp. 215–242, 1 2019.

[51] P. Bhagat, P. Choudhary, and K. M. Singh, "A comparative study for brain tumor detection in MRI images using texture features," *Sensors for Health Monitoring*, pp. 259–287, 1 2019.

[52] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadhel, M. Al-Amidie, and L. Farhan, "Review of deep learning: concepts, CNN architectures, challenges, applications, future directions," *Journal of Big Data 2021 8:1*, vol. 8, pp. 1–74, 3 2021.

[53] M. Decuyper, M. Stockhoff, S. Vandenberghe, a. , and X. Ying, "An Overview of Overfitting and its Solutions," *Journal of Physics: Conference Series*, vol. 1168, p. 022022, 2 2019.

[54] F.-N. B. W. Group, "Understanding Prognostic versus Predictive Biomarkers," 12 2016.