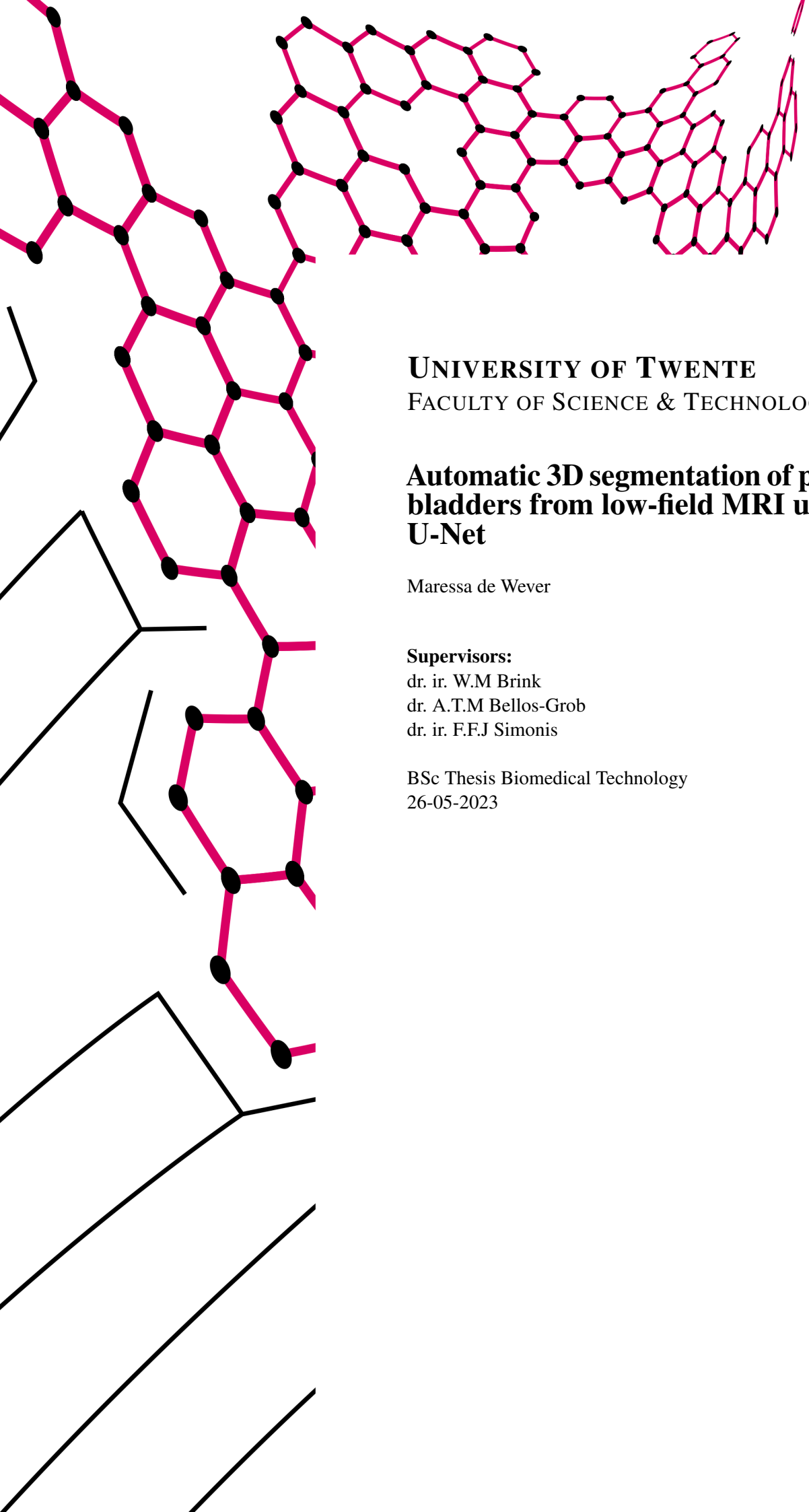


UNIVERSITY OF TWENTE
FACULTY OF SCIENCE & TECHNOLOGY

Automatic 3D segmentation of prolapsed bladders from low-field MRI using 3D U-Net

Maressa de Wever

Supervisors:
dr. ir. W.M Brink
dr. A.T.M Bellos-Grob
dr. ir. F.F.J Simonis

BSc Thesis Biomedical Technology
26-05-2023

Abstract

Anterior vaginal wall prolapse affects about 41% to 50% of women. Anterior Colporrhaphy is the most performed pelvic organ prolapse surgery. It has a 41% chance of operative failure and a 10% to 30% re-operation rate. There is currently no standardized technique among surgeons for colporrhaphy. A 3D model of the bladder could be useful for diagnosis, surgical planning and surgical evaluation. A deep learning model using a convolutional neural network in the form of a U-Net is trained to segment 3D models of all bladders (prolapsed and non-prolapsed) from MRI images.

The data used is gathered mainly from the TORBO study at the University of Twente. After a selection of the data, aided by classification of the bladders in degree of prolapse, contrast and artefacts, there are 4 test scans and 29 training scans (24 training and 5 validation). A selection is also made for the dataset sizes of 15, 20, 25. All sizes are used for training to evaluate the effect of dataset size on the test mean DSC and HD. The best model is used to obtain predictions on the test scans. The network prediction is then used in Matlab to evaluate the bladders pre- and post-op in the PICS coordinate system to evaluate the effect of the surgery.

Increasing the dataset size leads to an increase in test mean DSC and a decrease in the test mean HD. The model chosen to further evaluate the test scans has a DSC of 0.79 and a HD of 21.4. This model shows that the prolapsed bladders can be almost successfully segmented. Small extrusions in the prediction are still present and bladders strongly affected by bowel movement are not segmented correctly. Also the bladder wall is sometimes included in the prediction. With an increase of training data the model is expected to reach a mean DSC of around 0.85-0.90 which will further improve the results. The evaluation of the surgery can be done by evaluating the lowest point of the bladder from the 3D PICS coordinate system. It is concluded that the current model is a good step towards automated prolapsed bladder segmentation.

Samenvatting

Verzakking van de blaas komt voor bij ongeveer 41% tot 50% van de vrouwen. De voorwandplastiek is de meest voorkomende POP operatie en heeft een kans van 41% dat de operatie niet succesvol is en de verzakking terugkeert. De kans op een tweede operatie is zo'n 10% tot 30%. Momenteel is er geen techniek die als standaard wordt gebruikt onder chirurgen. Een 3D model van de blaas kan problemen verhelpen bij de diagnose van de verzakking maar ook bij het plannen van de operatie en het evalueren van de operatie. Een 'deep learning' model in de vorm van een U-Net kan worden getraind met een brede dataset om automatisch een 3D model van de blaas (verzakt en niet verzakt) te segmenteren uit MRI beelden.

De gebruikte data is voornamelijk afkomstig uit de TORBO studie aan de University of Twente. Na een selectie van de data, geholpen door een classificatie van alle blazen op gebied van verzakking, contrast en artefacten, zijn er 4 test scans en 29 training scans waarvan er 24 gebruikt worden voor training en 5 voor validatie. Een selectie wordt ook gemaakt voor 15, 20 en 25 scans. Alle grootten aan datasets worden getraind zodat het effect van de dataset grootte in kaart kan worden gebracht aan de hand van de gemiddelde DSC en HD. De netwerk predictions worden ingeladen in Matlab om de blazen voor en na de operatie te evalueren.

Het vergroten van de dataset leidt tot een verhoging van de gemiddelde DSC en een verlaging van de gemiddelde HD. Het model dat is gekozen om de test scans te analyseren heeft een DSC van 0.79 en een HD van 21.4. Dit model laat zien dat verzakte blazen kunnen worden gesegmenteerd door het netwerk. Er zijn wel nog kleine uitstulpingen die niet in de segmentatie horen en blazen die sterk beïnvloed zijn door de beweging van de darmen worden nog niet goed gesegmenteerd. Ook word de blaaswand af en toe meegenomen in de segmentatie. Een vergroting van de dataset zou in de toekomst kunnen leiden tot verbetering op deze punten en uitkomen op een DSC van ongeveer 0.85-0.90. Ook is het laten zien dat met behulp van het laagste punt van de blaas uit het 3D PICS coördinaat systeem het effect van operatie kan worden bestudeerd. Er kan worden geconcludeerd dat het model verkregen in deze studie een goede stap is richting de automatische segmentatie van verzakte blazen.

List of abbreviations

AI -Artificial Intelligence
AVP - Anterior Vaginal wall Prolapse
bSSFP - Balanced Steady State Free Precession
CNN - Convolutional Neural Network
DSC - Dice Similarity Coefficient
POP - Pelvic Organ Prolapse
HD - Hausdorff distance
MRI - Magnetic Resonance Imaging
PICS - Pelvic Inclination Correction System
Post-op- post-operative
Pre-op - pre-operative
SCIPP - Sacrococcygeal-inferior Pubic Point
SGD - Stochastic gradient descent
UT - University of Twente

Contents

1 Introduction

1.1 Pelvic Organ Prolaps

Anterior vaginal wall prolapse (AVP), also known as cystocele or prolapse of the bladder is a form of pelvic organ prolapse (POP) in women. AVP occurs when the muscles of the pelvic floor become too loose or if they sustain damage. [1] The most common causes of POP are vaginal childbirth and aging. During vaginal childbirth the chance is high that muscles and connective tissue of the pelvic floor sustain injury. [1] With aging, the muscles of the pelvic floor weaken because of declining estrogen levels during menopause. [2] Some form of prolapse is present in 41% to 50% of women. However only 3% have reported symptoms.[1]

AVP can be treated with pelvic floor exercises and/or a pessary. When conservative treatments are no longer effective, the colporrhaphy surgery can be necessary. This reconstructive surgery is performed vaginally. The vaginal wall is reinforced using dissolvable sutures to support the bladder. The goal of the surgery is to minimize the prolapse and reduce symptoms. [2] The risk of POP surgery being necessary for a woman at some point in her life is 6.3% to 11%. [1, 3] 81% of all surgical POP repairs involve the anterior wall and it is also the most frequent site of operative failure (41%). [4]. The re-operation rate for recurrence of the prolapse is estimated at 10% to 30%. [3] It was concluded from a study by Lensen et al. [5] that in the Netherlands, there is a large variety of techniques used for the anterior colporrhaphy surgery. There is currently no standard because there is not sufficient evidence of the best method to use. [5]. The result from surgery varies per patient. Recurrence of the prolapse is common and there is not a good understanding why. It probably depends on the used surgical techniques and surgical planning, however, surgeons have not been able to fix this problem.

1.2 3D segmentation

A 3D model reconstruction of the bladder can be used for diagnosis, surgical planning and surgical evaluation. It can aid in finding the best surgical method by evaluating the 3D bladder pre- and post-op. Eventually it may aid in the decrease of recurrence and re-operation. A 3D model can be made by manual segmentation of the bladder from MRI images. This is a time consuming process (with knowledge of the segmentation program, manual segmentation can take from half an our to an hour and a half) and is sensitive to intra- and inter-observer variability. Certain computer algorithms can aid manual segmentation. This is for example thresholding or region growing. [6] These techniques have limited success for organs with blurry boundaries, like the bladder. [7]

Bladders can vary widely between patients. The shape is strongly dependent on the degree of prolapse and the volume. The contrast is affected by the movement of the urine inside the bladder and can be high, low or varying inside the bladder. The bladder is also compromised by the movement of the bowel and other standard artefacts that occur with MRI. Because the bladder varies widely between patients it is challenging to find a segmentation method that can adapt to this inter-subject variability.[7] A solution can be automated segmentation techniques. These use artificial intelligence (AI), and more specifically deep learning with a convolutional neural network (CNN). CNN uses several convolutional layers to extract features and provide segmentations. The studies from Feng et al. [7] and L.Straetemans [8] show some success in using the deep learning method for bladder segmentation.

The study by L. Straetemans [8] was performed at the University of Twente (UT). A U-Net was trained with low-field MRI data of non-prolapsed bladders. The study showed success in automatically segmenting the non-prolapsed bladders. However, in the case of prolapsed bladders, the prolapse was not included in the prediction, see Figure 1. The results also showed some issues with bladders with inhomogeneous intensities. Bladders affected by artefacts were not included in this study.

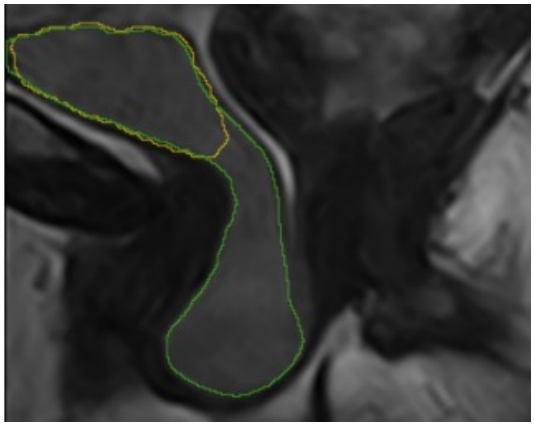


Figure 1: Bladder prediction (yellow) and ground truth bladder (green) by 3D U-Net [8]

1.3 Goal

The goal of this study is to re-train the 3D U-Net from L. Straetemans [8] to segment all bladders. That includes prolapsed bladders, non prolapsed bladders, bladders with high or low contrast, bladders that are affected by artefacts and as much combinations of these conditions as possible. Another goal is to gain knowledge in the effect of the size of the dataset on the resulting model to know how much data is needed in order to obtain this 3D U-Net that can segment all bladders. The last goal is to make a step towards the clinical evaluation of the bladders by comparing the network predictions before and after surgery. This can be done by comparing the badder volume below the PICS line and lowest point of the bladder.

2 Background

In this chapter, some concepts that are used for analysing the results are briefly explained. Including the PICS line for the clinical evaluation, the architecture and training of a neural network and the metrics used to evaluate the model.

2.1 PICS

In order to determine the dislocation of the bladder prolapse, certain landmarks (inferior pubic point of the symphysis, sacrococcygeal joint, left and right ischial spines) are selected. The landmarks form the basis for a 3D coordinate system that follows the movement and rotation of the bony pelvis. In this 3D coordinate system, the position of any point in the pelvis along the body axis can be determined in a normalized reference space. The SCIPP line is a reference line in this coordinate system that is traced form the inferior pubic point of the pubic symphysis to the sacrococcygeal joint. Another reference line can be drawn 34 (supine) of 29 (upright) degrees below the SCIPP line from the inferior pubic point. [9] This line is the Pelvic Inclination Correction System (PICS) line, see Figure 2. [10] The PICS line corrects for movement of the pelvis in the mid-sagittal plane. A non-prolapsed bladder is expected to be above the PICS line. [10]

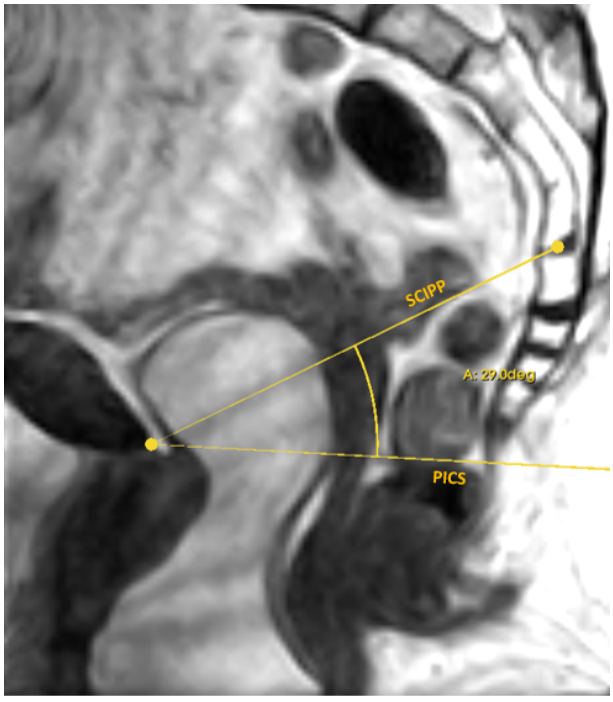


Figure 2: PICS and SCIPP line for an upright MRI

The PICS coordinate system can give information about the volume that lies below the PICS line and the lowest point of the bladder. The lowest point gives information about the extend of the prolapse and can aid in the quantification and diagnosis as seen in a study from Najjari et al. [11].

2.2 Deep learning

Deep learning can be used for organ segmentation, as discussed in the introduction. [7] Deep learning in this study uses CNN and is trained with MRI data from various patients which results in a good generalization performance. [7]

CNN
A convolutional neural network consists of different convolutional layers of neurons. These neural network neurons consists of a certain number of inputs. A weighted sum is performed on the inputs and an activation function is applied, which results in an output. The CNN is trained using a loss function. This function will aim to minimize errors in the network output, compared to the ground truth by adjusting the weights. [12]

Hyperparameters are an important subject in deep-learning. These parameters control the learning process of the network. These are for example weight decay, learning rate and number of epochs. An epoch is a single review of the data where the training data set is run through the network. The performance of a network increases with every epoch and will eventually plateau. [8]

U-Net
A U-Net is a deep learning model that uses CNN in the shape of the letter 'U' and is designed for image segmentation. [13] It has gained popularity for biomedical image processing and is commonly used for segmenting MRI images.[14] In the case of biomedical image segmentation, there is often limited training data available. A U-net can be used cases where the data sets are smaller because the architecture of the U-Net which is deeper then that of most networks. The deeper architecture implements more downsampling into the network, more downsampling allows for extended extraction of features of the image in lower resolutions and thus a more powerful and robust performance of the network, improving the training results. U-Net is also fast to train due to its context-based learning. [13]

The architecture of a U-Net consist of an encoder path which is followed by an decoder path. The encoder path is usually a classification network with convolution blocks followed by maxpool downsampling. The decoder path consists of up-sampling and concatenation of the features from the encoder path and projects the features to the pixel space. [13][15]

2.3 Neural network training

The neural network that was made by L. Straetemans [8] will be trained with new data. Training of a neural network is done for better performance and recognition of patterns. Training can be supervised or unsupervised. Biomedical image segmentation is often based on supervised training. During supervised training, the input data and corresponding output labels (in this case the manual segmentation labels) are given. The network is shown input images and produces an output. Initially, the network output will likely not resemble the ground truth. The loss function can measure what error the output has to the input, and the network can modify the internal adjustable parameters to reduce the error. The parameters are called the weights and define the input-output function of the network. [16] At the start of the training the initial weights are random. During the training process, the weights are adjusted to minimize the loss function and thus the error between the ground truth and the prediction. Backpropagation is used to minimize the loss by adjusting the weights with a step-size called the learning rate. An optimization method that is used for updating the weights is the stochastic gradient descent (SGD). [8]

2.4 Dice Similarity Coefficient and Hausdorff Distance

To indicate how well a network is trained, certain metrics are used. In this paper this will be the Dice Similarity Coefficient (DSC) and the Hausdorff Distance (HD). The DSC is a spatial overlap index. The value ranges from 0 to 1, 0 meaning that there is not spatial overlap between the ground truth and the prediction and 1 meaning complete overlap. [17] The HD is a measure for the distance between two segments. All voxels in the predicted segmentation and all voxels corresponding to that voxel from the ground truth segmentation are determined. The distance between all pairs of voxels are then determined. The largest distance between two sets of points is the HD. When the segmentations align nicely, The HD will be small. [8]

3 Method

3.1 Data

3.1.1 Data acquisition

The data in this study is gathered from an existing database of the TORBO study at the UT. The patients in this study are scanned before and six weeks after surgery in supine and upright position (81 degrees). A total of four scans per patient are available. At the time of this project there are 29 patients included in the study that have been scanned pre-op and six weeks post-op. Because the dataset is desired to be as large as possible for the most diverse training set, 4 scans from another study (EPPA) with the same field of view as the TORBO scans are added. Thus, data from 33 patients is used in total. The scans are made with the Esoate G-scan Brio 0.25 T MRI system. A balanced steady state free precession (bSSFP) sequence is used. The resolution of the images was 2.02 x 2.02 x 2.5 mm^3. After interpolation the voxel size is 1.5 x 1.5 x 1.5 mm^3.

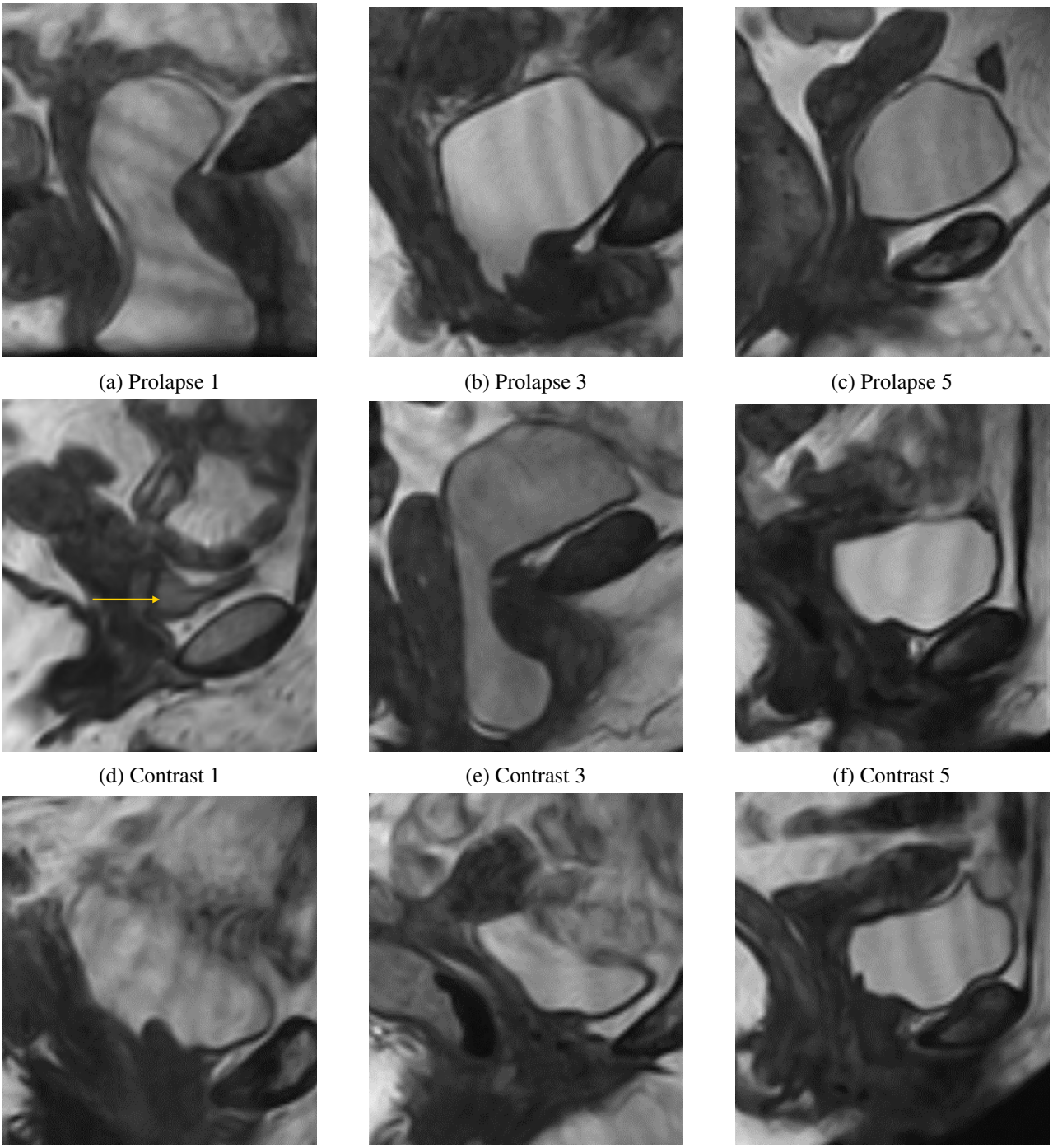
3.1.2 Data selection

As discussed in the introduction, bladders can vary greatly in shapes, sizes, contrast and artefacts. If a network that can segment all bladders is desired, regardless of the shape and quality, the training dataset has to be as diverse as possible. The network will then learn to recognize the different patterns. Using multiple scans from one patient for the training of a network should be avoided [8] because to obtain the best trained network, the training should have great anatomical diversity. Therefore, in this study, only one scan out of the available four per patient is used. This means a selection of the data has to be made.

A classification can aid the selection of the bladders and is done by assigning all bladders with a number from 1 through 5 for the degree of prolapse, contrast and artefacts, see Figure 3. For the degree of prolapse, 1 is maximal prolapse and 5 is no prolapse. For the contrast, 1 is low contrast or differential contrast within the bladder and 5 is high. For the artefacts 1 means a lot of artefact is present and 5 means very little.

When all scans are given a number for the three criteria, different possible combinations are proposed. For example: low contrast, low artefacts and no prolapse or high contrast, low artefacts and prolapse. If a scan is true to a certain combination it is assigned with a color. At least one scan of every color is included in the dataset. In total there are twelve combinations. In total there are 33 scans of which four will be used for testing, 24 for training and 5 for validation. Meaning 29 scans are the total training and validation dataset. The classification of these 29 chosen scans are seen in Figure 4

Because 29 training scans is the maximum training set size at the time of this study, it is unclear if more training scans would lead to a better model. To assess the networks dependency on the amount of training data, datasets with less then 29 scans are tested. This will be with 15, 20 and 25 scans. For these sizes, a classification and selection of the data is performed. All dataset sizes will be trained multiple times to check if the resulting model is consistent when trained under the same conditions.



(a) Prolapse 1 (b) Prolapse 3 (c) Prolapse 5

(d) Contrast 1 (e) Contrast 3 (f) Contrast 5

(g) Artefacts 1 (h) Artefacts 3 (i) Artefacts 5

Figure 3: Bladders and their assigned numbers (1, 3 and 5) for the degree of prolapse, contrast and artefacts

	Contrast	Artefact	Prolaps
p001	5	4	1
p002	2	5	5
p003	4	4	1
p004	2	4	2
p005	1	3	5
p006	5	3	5
p007	3	2	1
p008	1	3	3
p009	5	4	3
p010	2	2	2
p011	3	4	1
p012	3	3	4
p013	1	1	5
p014	3	3	1
p015	5	3	2
p020	4	3	1
p021	4	2	5
p022	3	1	2
p024	4	5	1
p025	5	5	3
p026	2	4	5
p027	5	4	1
p029	4	3	5
p035	2	4	1
p036	4	3	1
p312	4	4	5
p313	5	4	5
p315	4	5	5
p314	5	4	5

Figure 4: Selection result of the training dataset with the classification values of the contrast, artefacts and prolapse and the combination of the three

3.2 Manual segmentation

The selected bladders will form the training dataset for the U-net. To train the U-Net the output labels need to be generated by manually segmenting the bladders. The manual segmentation will serve as the ground truth. The segments are made in 3D Slicer using a threshold, cutting away large sections with the scissors and keep largest island function, the eraser and paintbrush are used for small details and to conclude surface smoothing is applied. Two examples of manual segmentations can be found in Figure 5. The bladder wall is not included in the segments. The four landmarks discussed in the background also marked for further evaluation of the volume and the PICS line. The segments are saved as NifTi files to serve as the ground truths during training, validation and testing.

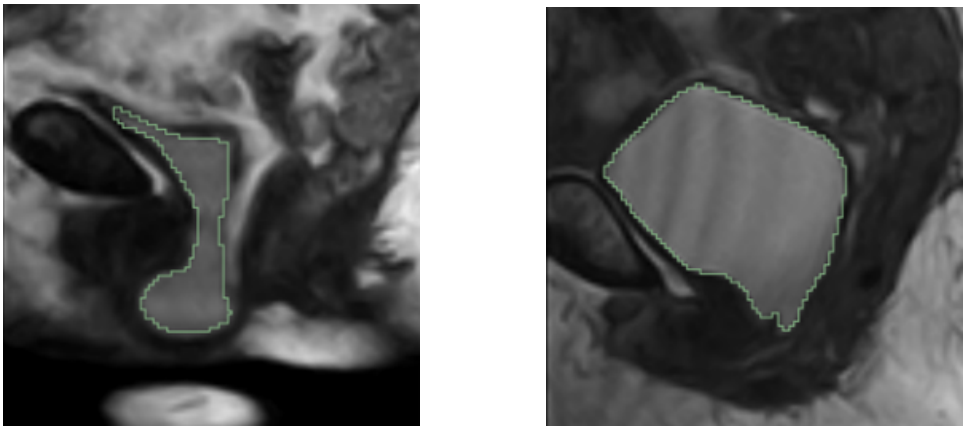


Figure 5: Examples of manual segmentations

3.3 U-Net

The U-net code from the study by L. Straetemans [8] is used. The hyperparameter optimization was performed by this previous study. The same hyperparameters are used in this study and are listed in Table 1

Hyperparameter	Value
Pixel dimensions	1.5 x 1.5 x 1.5 mm^3
Region of interest	[128, 128, 64] pixels
Number of Epochs	700
Batch size	1
Learning rate	0.0001
Weight decay	0.001

Table 1: Hyperparameters U-Net

For all parameters that are not in this list, the default settings are used. The pixel dimensions and batch size are chosen to be this value for memory reasons. The number of epochs is set to 700 which was chosen after a few trials in the study by L. Straetemans [8] because for this number, the metrics plateaued. However, because the size of the training set and the quality and shape of the bladders in the dataset from this study are different, a 1000 epochs will be tested as well to see if it leads to a better model.

To summarize, different sizes of training sets are put through the network to observe the effect of enlarging the dataset. Then the network is trained with the largest dataset of 29 scans and the resulting model is tested on the four test scans to get an objective reflection on data that was not used for training. All of the rounds of training will give a mean DSC and mean HD on the test scans and a visual prediction that is used to evaluate how good a model can predict the bladder.

3.4 Surgery evaluation

The effect of surgery is evaluated in more detail for two patients. This is will serve as an example of how the evaluation can be performed on the predictions of the network in the future. Because of time limitations, this is not done for every patient. The bladders from the upright pre- and post-op scans are predicted by the network and the predictions are analysed using Matlab. The four landmarks from the patient scan are loaded in Matlab and put into it's own 3D coordinate system, where the pubic symphysis landmark is at the origin, and the PICS line is on the x-axis. The predicted bladder volume is loaded into this coordinate system. The volume below the PICS line is calculated together with the lowest point of the bladder. These values will give information about the effect of the surgery. The main goal here is to give an example of how this research can be used for evaluating the surgery effects.

4 Results

First, the results from the network training with different dataset sizes and number of epochs will be discussed. The bladders from the test data are predicted by the best model and the results from this are shown in 2D and 3D, with a network prediction and a ground truth. The last result is the evaluation of pre- and post-operative bladders.

4.1 Network training

4.1.1 Dataset size and number of epochs

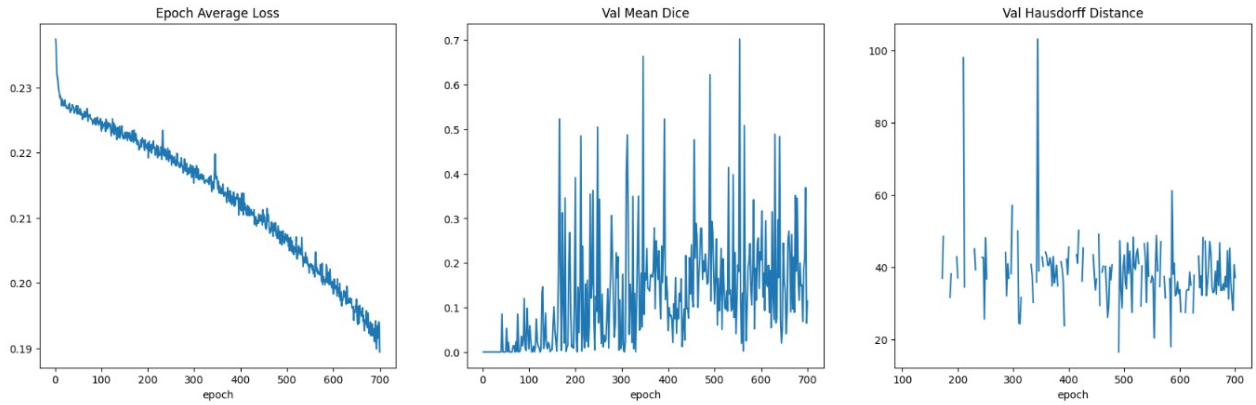


Figure 6: Metrics of 29 training scans, 700 epochs

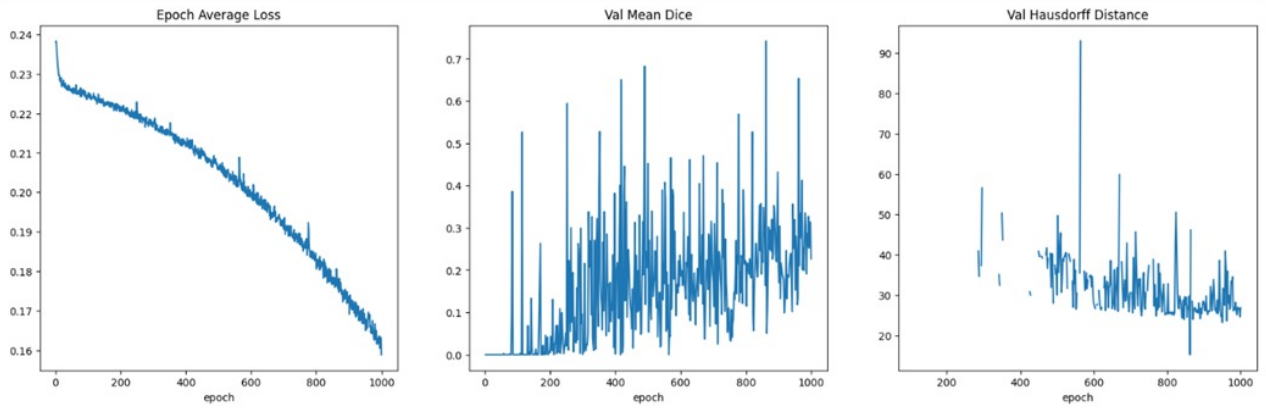
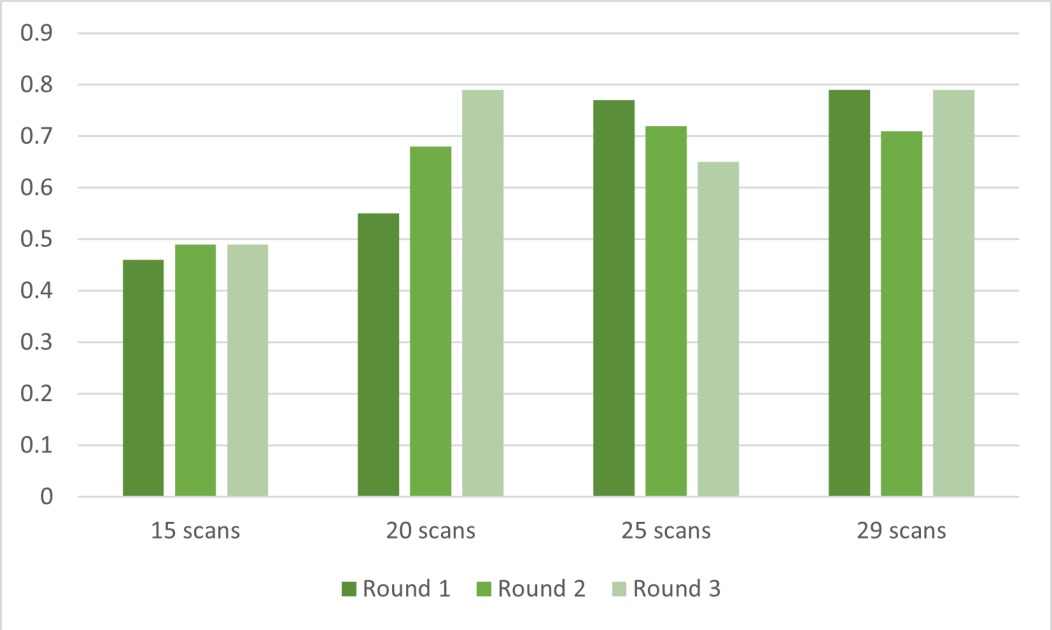
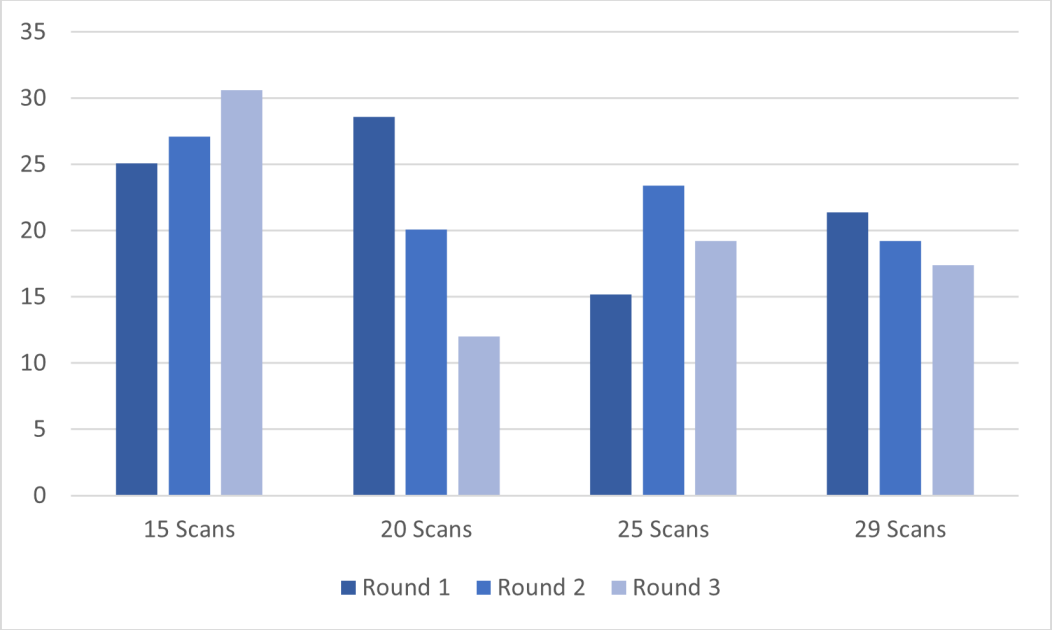


Figure 7: Metrics of 29 training scans, 1000 epochs

In Figure 6 the average loss, the validation mean DSC and the validation HD of the training of every epoch is shown. The DSC is increases to about an average of 0.2 with outliers from 0.1 to 0.7. The graph does not plateau. The loss function decreases from about 0.24 to 0.19. The HD is not declining. From the metric results of a training with 1000 epochs (Figure 7) it is observed that with a larger number of epochs, the loss and DSC increases and the HD decreases further. However, that did not mean the mean DSC and HD values of the test scans improved. Also there was no significant difference in the visual predictions from 700 and 1000 epochs.



(a) Test mean DSC



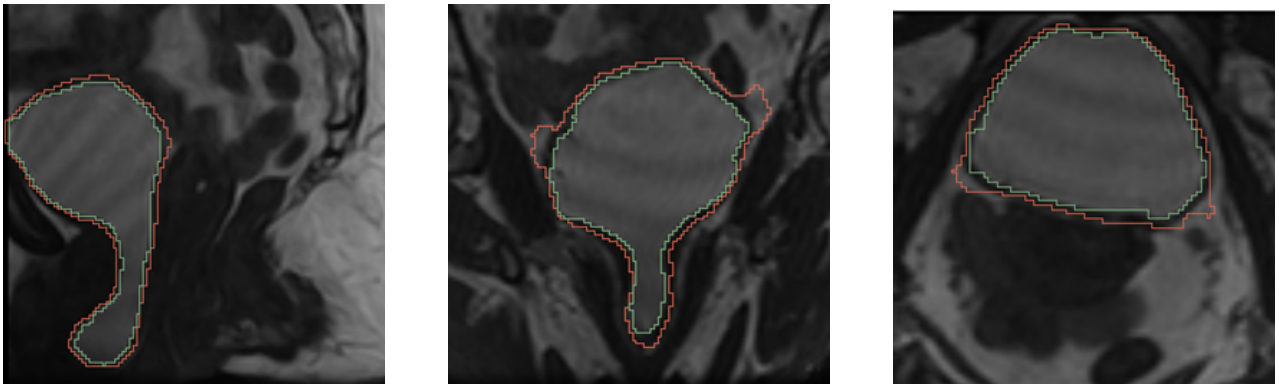
(b) Test mean HD

Figure 8: Test mean DSC (a) and test mean HD (b) for different sizes of training sets, three rounds of training per set size

In Figure 8 the metric results from four different sizes of training sets (15, 20, 25 and 29) on the test data of three different training rounds are displayed. All the training is done with the number of epochs at 700. The mean DSC increases strongly when comparing 15 training scans to 29 training scans. With 20 training scans, the mean DSC differs with about 0.35 between training rounds. The HD also has a difference of about 25. This difference between rounds decreases when looking at the DSC and HD of 25 and 29 training scans. 29 training scans has the highest DSC's and smallest variability in metrics between rounds, other training rounds are also taken into consideration. The 29 dataset size produced a mean DSC of 0.79 for four out of six training rounds. This is the only size that is trained more then three rounds.

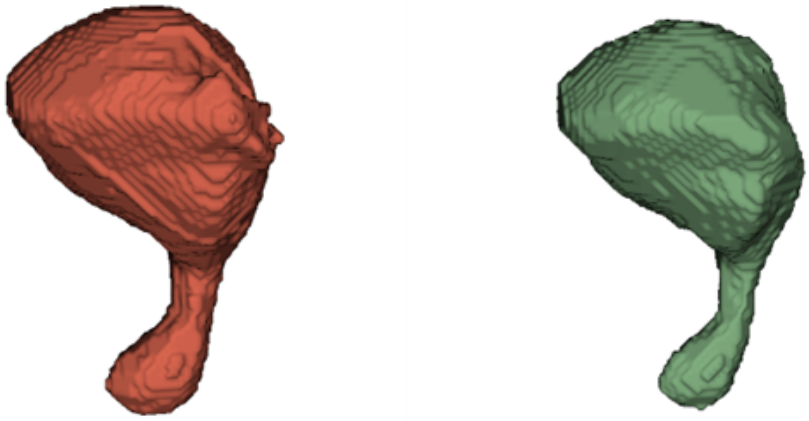
4.1.2 Test results

A training round with 29 scans resulted in the model that is used to evaluate the test scans. This model has a test mean DSC of 0.79 and a test mean HD of 21.4. In the Figures 9 to 16 the network prediction and the ground truth of the four test scans and the 3D bladder model are visualized.



(a) Sagittal (b) Frontal (c) Transverse

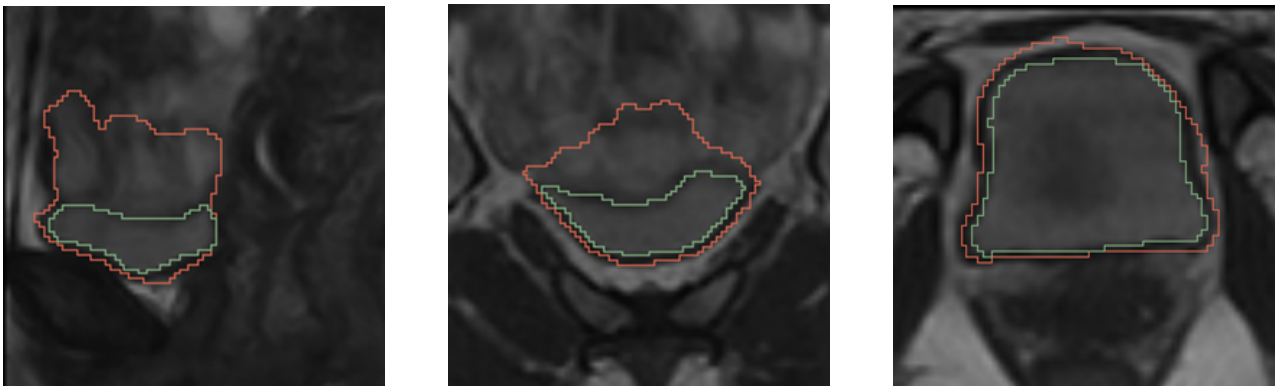
Figure 9: Image slice in all planes. Ground truth (green) and prediction (orange) (p019)



(a) (b)

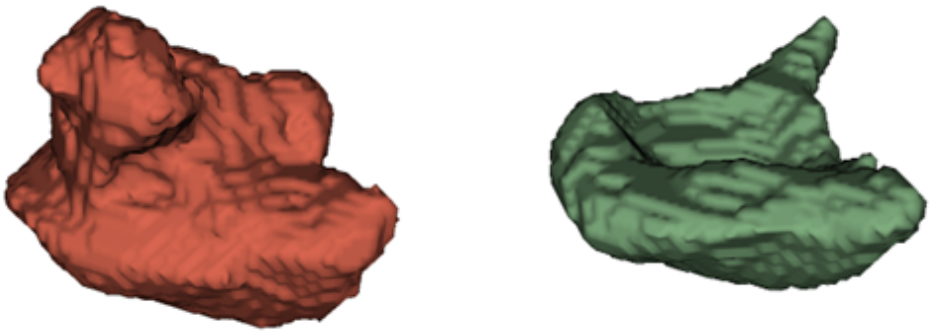
Figure 10: 3D prediction (a) and 3D ground truth (b) (p019)

In Figure 9 the sagittal, frontal and transverse planes are shown. The network prediction (orange) follows the ground truth (green) quite accurately. The prediction takes more of the bladder wall into the segmentation and there are some extrusions in the prediction that are not supposed to be there. These occurs mostly on the upper part of the bladder. When looking at the 3D models in Figure 10 it is also clear that the prediction takes more of the bladder wall into the segment because the entire bladder is thicker. The prolapse segmentation of the prediction is similar to the ground truth in shape, the upper part of the bladder has some small differences.



(a) Sagittal (b) Frontal (c) Transverse

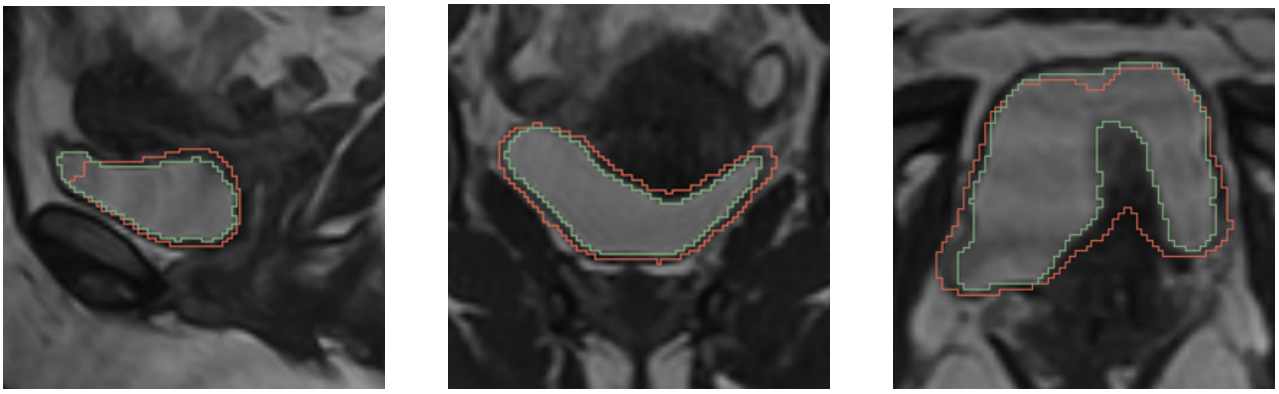
Figure 11: Image slice in all planes. Ground truth (green) and prediction (orange) (p028)



(a) (b)

Figure 12: 3D prediction (a) and 3D ground truth (b) (p028)

In Figure 11 the sagittal, frontal and transverse planes are shown. The network prediction (orange) is similar to the ground truth (green) at the bottom of the bladder. However, the upper border of the bladder is not correctly segmented by the prediction. It follows a random path. This is caused because the bladder wall is very influenced by artefacts from bowel movement which is not completely visible because of the line form the ground truth. Also, the bladder wall is added in the prediction. When observing the 3D models in Figure 12 the same is seen. The prediction is larger because of the inclusion of the bladder wall and the top part of the prediction includes a big portion that is not in the ground truth.



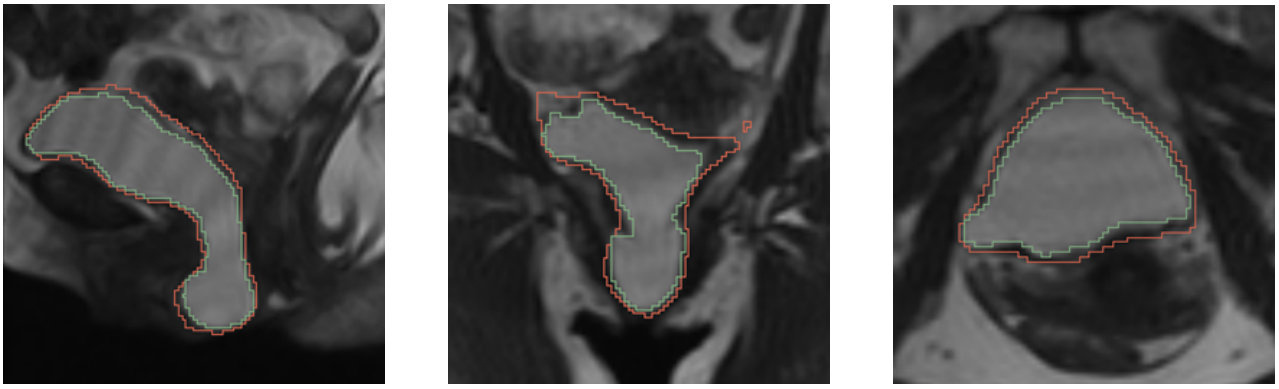
(a) Sagittal (b) Frontal (c) Transverse

Figure 13: Image slice in all planes. Ground truth (green) and prediction (orange) (p030)

(a) (b)

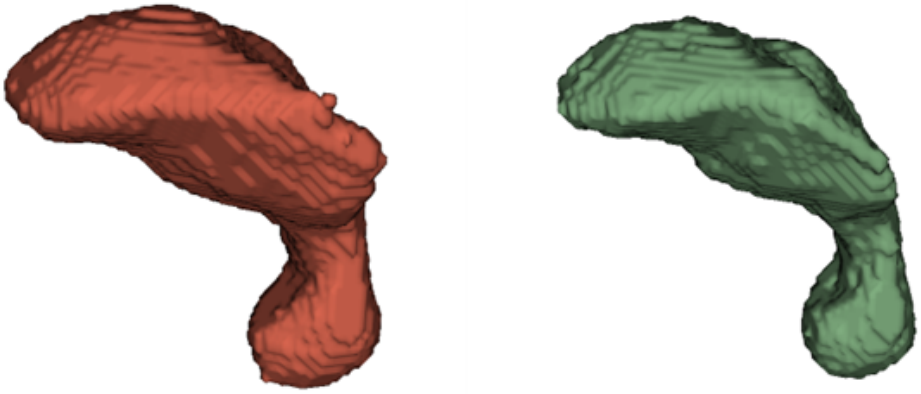
Figure 14: 3D prediction (a) and 3D ground truth (b) (p030)

In Figure 13 the sagittal, frontal and transverse planes are shown. The network prediction (orange) follows the ground truth (green) quite accurately is most places. A small part towards the front of the bladder, which can be seen in the sagittal plane, of the ground truth is not included in the prediction. In the transverse plane a there is a portion that the prediction mistakenly includes. The prediction also includes the bladder wall. The bottom part of the prediction is similar to the ground truth. When looking at the 3D model in Figure 14 the same observations are made.



(a) Sagittal (b) Frontal (c) Transverse

Figure 15: Image slice in all planes. Ground truth (green) and prediction (orange) (p031)



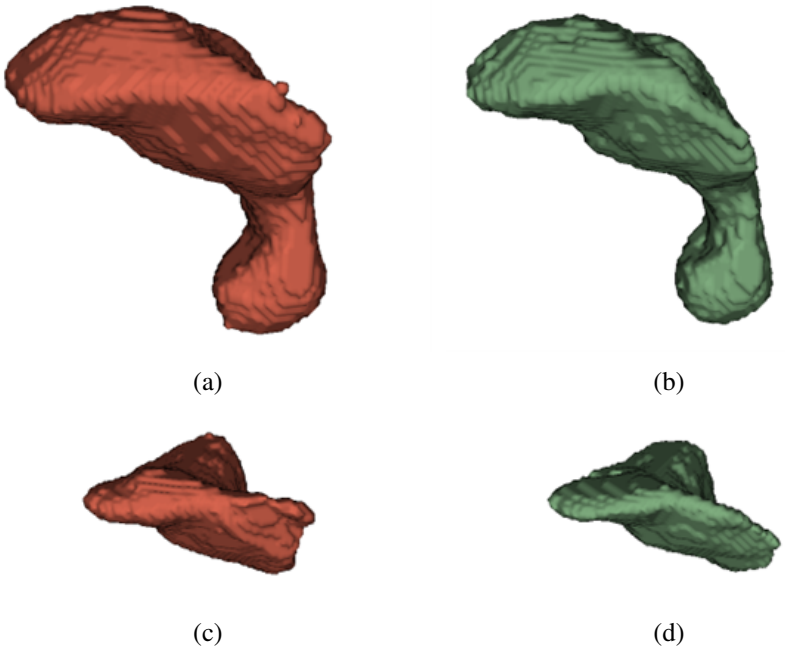
(a) (b)

Figure 16: 3D prediction (a) and 3D ground truth (b) (p031)

In Figure 15 the sagittal, frontal and transverse planes are shown. The network prediction (orange) follows the ground truth (green) quite accurately. At the upper part of the bladder are some extrusions in the prediction, best seen in the frontal plane. The prediction includes some of the bladder wall into the segmentation. In the 3D model in Figure 16 the prolapse part of prediction resembles the ground truth well, it is however a little larger because of the inclusion of the bladder wall. In the 3D model we can also see the extrusions that the prediction includes.

4.2 Surgery evaluation

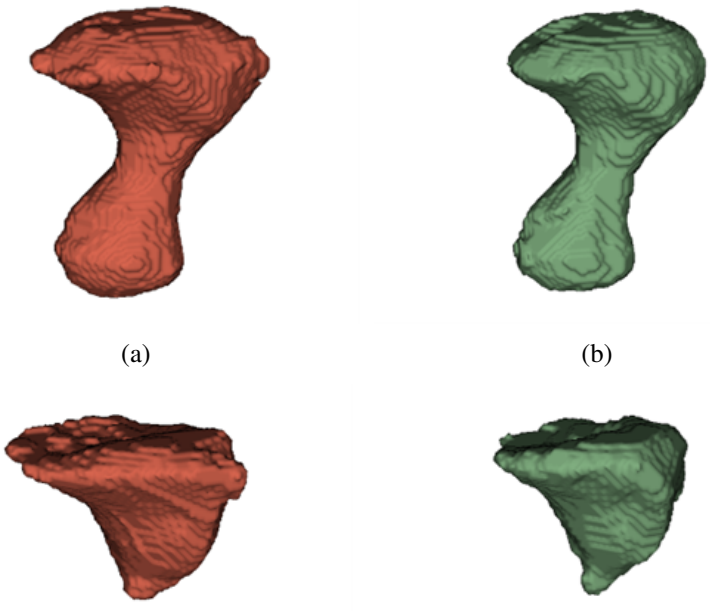
In the TORBO study by the UT, patients are scanned pre-op and six weeks, one year and two years post-op. In this study only the six weeks post-op scans are evaluated. With the U-Net network from this study the bladder predictions can be used to evaluate the surgery results. In the next section the network prediction and ground truth from two patients pre- and post-op are shown in Figures 17 and 18. In Table 2 the volume of the prolapse and the lowest points of the bladders are listed.



(a) (b)

(c) (d)

Figure 17: 3D model of p031 pre- (a,b) and post-op (c,d). Prediction (a,c) and ground truth (b,d)



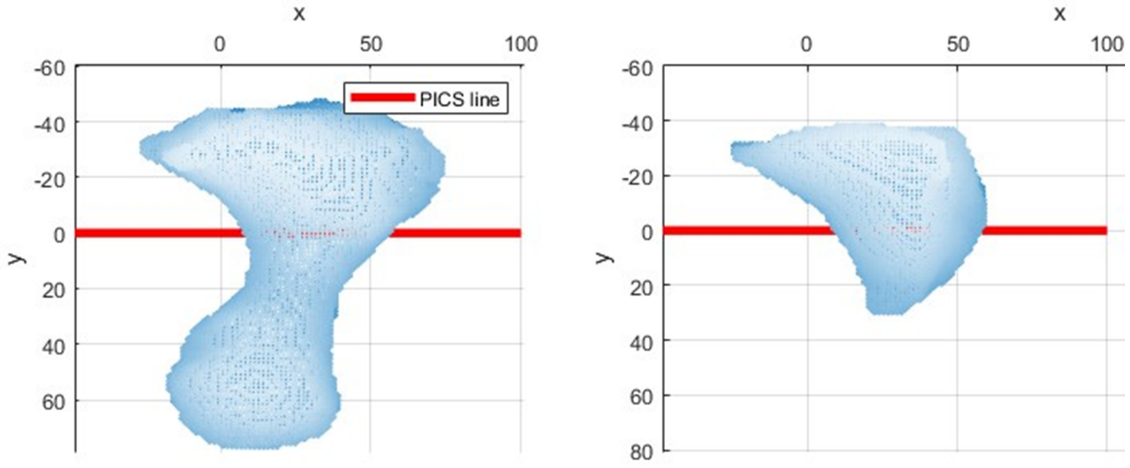
(a) (b)

(c) (d)

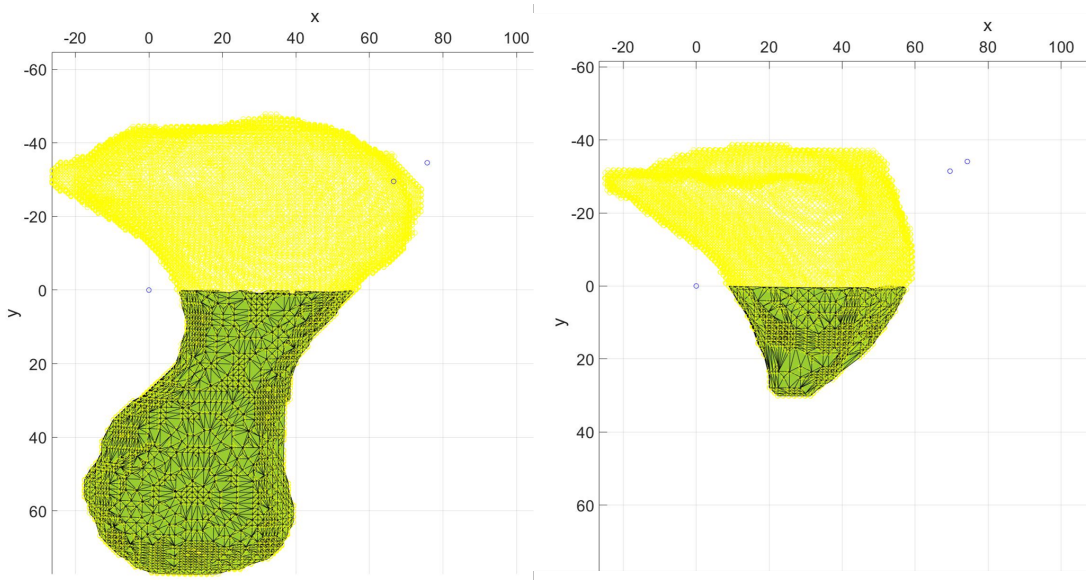
Figure 18: 3D model of p036 pre- (a,b) and post-op (c,d). Prediction (a,c) and ground truth (b,d)

In Figure 17 both post-op bladders do not show the extreme prolapse that is visible in the pre-op bladders. In Figure 18 the result is similar, however there is still some visible prolapse in the post-op bladder. In both cases there is improvement of the prolapse.

In Figure 19 the bladder prediction of p036 is loaded into the 3D PICS coordinate system. The bladder pre- and post-op lies partly beneath the PICS line but does decline from pre-op to post-op. The volume beneath the PICS line can be calculated, see Figure19b. The lowest point is read from this coordinate system as the lowest point straight down from the PICS line.



(a)



(b)

Figure 19: Evaluation of the network predicted bladder in the 3D PICS coordinate system. Pre- (left) and post-op (right)

The values of the PICS evaluation are shown in Table 2. The volume below the PICS line of the prediction is not close to that of the ground truth because the prediction includes most of the bladder wall in the segmentation, as seen in the previous results. The larger the volume below the PICS line, the larger the difference in volume between the prediction and ground truth. The lowest point of the bladder is the largest distance beneath the PICS line. This value can give an indication of the success of the surgery and is almost the same for the prediction and the ground truth, varying only millimeters. The lowest point of p031 post-op is almost above the PICS line, with a decline in distance of about 48 mm. For p036 this is not the case because the bladder is still about 30 mm below the PICS line, which indicates prolapse. There is however a large decline in prolapse, the lowest point rises about 45 mm. The distance decrease is similar to that of p031.

Segment	Volume PICS (mm^3)	Lowest point (mm)
p031_pre GT	18354	48.6
p031_pre PRED	28544	51.5
p031_post GT	111	2.8
p031_post PRED	187	3.1
p036_pre GT	8372	74.5
p036_pre PRED	10203	77.2
p036_post GT	3041	31.6
p036_post PRED	4109	30.2

Table 2: Volume below the PICS line and lowest point of the bladders from p031 and p036. The network prediction indicated by PRED and the ground truth indicated by GT

5 Discussion

5.1 Network training

The validation mean DSC, HD and loss function (Figure 6,7) were expected to plateau during training.[8, 18] The DSC should increase and HD decrease. This was not the case. With the increase of training scans the graphs came closer to the expectation. The same holds for the increase to a 1000 epochs. With more training scans the graphs can be brought up to the expectation. If the epoch size then still needs to be increased, it is recommended to perform the entire hyperparameter optimization again. Because the bladders used in this study are so different then the ones in the study by L.Straetemans, there is a chance that a more optimal hyperparameter combination exists. This would lead to a higher DSC and lower HD.

When the training set size is increased, an overall increase in test mean DSC and decrease in test mean HD is observed (Figure 8), as expected. The mean DSC of a model trained with fewer scans were in some cases similar to the mean DSC of models trained with more scans, for example in the second round with 20 scans. For the smaller data sets, the difference in mean DSC between training rounds was larger, especially with 20 and 25 scans. It was expected that training under the same conditions would lead to a similar result. When searching online, the same issue was brought up on platforms like 'stackoverflow' by several researchers. Some possible explanations given are the large differences in bladders present in the data, the use of stochastic learning algorithms or the use of stochastic evaluation procedures.

Stochastic algorithms incorporate elements of randomness into their behaviour. Meaning that small specific decisions during the training process can vary randomly. Neural networks are a stochastic machine learning algorithm. The initial weights are random which allows the model to learn from a different targeting point and allows the break of symmetry. This ensures that each weight update and gradient estimate are slightly different, leading to different result for several rounds of training. [19] With large differences in the shape and quality of the bladders, more training data would lead to a more steady outcome with each training round.

5.2 Bladder segmentations

The model that was chosen to evaluate the test data had a test mean DSC of 0.79 and a test mean HD of 21.4. This model was able to segment most of the test scans accurately. The bottom of the bladders was segmented correctly for the test scans, prolapsed and non-prolapsed. This is an improvement from the model made by L. Straetemans [8] where segmenting the prolapse was one of the main limitations. Some small extrusions in the prediction are still seen at the upper part of the bladders. For the test scan in Figure 12 a large part at the top of the bladder is taken into the prediction because this bladder is greatly affected by artefacts from bowel movement. To successfully segment bladders that are as affected by the movement of the bowel, more data of bladders impacted by bowel movement have to be added to the training data. One or two of these bladders will probably be sufficient.

The prediction includes some parts of the bladder wall into the segmentation. This is only a few millimeters which visually does not cause the prediction to look very different from the ground truth. The model would score better in DSC and HD if the prediction followed the ground truth more closely. It is possible that the bladder wall in the ground truth was not always consistently left out of the manual segmentations. The goal was to not include the bladder wall but with low contrast bladders and present artefacts, the chance for some mistakes is high, causing the network to have difficulty in segmenting the edges. However, it was seen in a training round with a 1000 epochs and 29 training scans that it is possible to get a prediction that does not include the bladder wall. This model was not chosen because it also included more extrusions in the prediction. This model was also better at segmenting the upper boundary the bladder from Figure 11.

It was discovered that the predicted segmentations were not binary. The background had an overall pixel value of about 0.2. There were some small segments that were not connected to the bladder segment included in the segmentation, but were not removed by the function of the Monai transforms 'KeepLargestComponent'.

This function only gets rid of binary segments. This problem occurs because of a bug in the code where the threshold of 0.5 does not work properly. The thresholding and removing of the extra segments were performed manually in 3D slicer. A new segment was made by simply applying a threshold from 0.5, which was supposed to be done by the code. Then the extra components were removed with the keep largest component button in slicer. In the future, it is desired to have these steps successfully performed by the code. It is unclear what effect the non-binary segmentations have on the DSC and HD. It is expected that the HD is larger because of the extra segments and will improve when the code can successfully remove the extra segments.

5.3 Clinical evaluation

For the evaluation of the surgery, the volume below the PICS line and the lowest point of the bladder were analysed pre- and post-op for two patients. The volume below the PICS line is not an accurate variable to compare a bladder pre- and post-op. This is because the bladder volume depends strongly on how much urine is present, which will differ between two different scanning moments. It does however give a good insight in the success of the network predicted bladder by comparing the volume to that of the ground truth. From this comparison it can be concluded that the network prediction is not yet accurate. The volume of the prediction is significantly larger (20-40 %) because of the inclusion of the bladder wall. The lowest point of the bladder does give important information that can be used in the evaluation and quantification of the bladder, as was shown in the study from Najjari et al. [11]. The lowest point between pre- and post-op for these two patients decreases a few centimeters. But this did not mean that the prolapse was no longer present.

There are currently not more patient scans available in the TORBO study that were not already used in the training of the model. This means that the evaluation can be performed on the four test scans or the scans used for training which the network has already seen before. The scans from the training will not give a non-biased evaluation of the model and network training but can be used for clinical evaluation.

5.4 Recommendations and further research

It is clear from the results that the mean DSC increases and HD decreases when the dataset is enlarged. The current performance of the network already ensures a good segmentation of the bladder (mean DSC of 0.79). To further increase the DSC to around 0.85-0.90, the data set has to be enlarged. It is not clear how much extra data is needed but it expected that about 5-20 more training scans is sufficient. This increase can improve the inclusion of the bladder wall and the small excess extrusions as well as the badly segmented boundaries affected by bowel movement. The data that is added should include bladders that are as diverse as the data used in this study.

When a network is trained that has an even higher DSC and lower HD and thus a better segmentation performance, the PICS evaluation of the bladder can be performed for all patients pre- and post-op.

During this research, 'no new U-Net' (nnU-Net) was considered instead of the U-Net code from the study by L. Straetemans [8]. nnU-Net is a U-Net that automatically configures itself, which includes network architecture, training and the optimization of the various associated hyperparameters. This means that the nnU-Net can be used for all segmentation problems and does not require expert knowledge of deep learning and network training. It is a simple package that can be downloaded into Python or another platform. This method was briefly looked into but it was not pursued because of time limitations. For future studies, this can be a innovative new concept for automated organ segmentation.

In future research, it could be considered to combine the automatic landmarks detection to the automatic segmentation code. The automated landmark detection has shown success in the study by G. Hurkemans [18]. This combination can be a big step towards automated clinical evaluation and can also aid diagnosis and evaluation of the bladder.

6 Conclusion

Increase in dataset size leads to an increasing test mean DSC and decreasing HD. The model used to predict the bladders of the test data had a mean DSC of 0.79 and a HD of 21.4. This model showed success in segmenting non prolapsed as well as prolapsed bladders which was a large limitation of previous studies. There are still problems with segmenting bladders with bowel movement artefacts and the inclusion of the bladder wall in the prediction. It is estimated that about 5-20 more training scans will result in a model that will overcome the problems with the inclusion of the bladder wall and the small extrusions and lead to a DSC of 0.85-0.90. It is also concluded that the network predictions can be used to evaluate the bladder in pre- and post-operative state using PICS evaluation which is a first step toward automated prolapse analysis.

References

[1] Iglesia CB, Smithling KR. Pelvic Organ Prolapse. Am Fam Physician. 2017 Aug;96(3):179-85. Available from: https://www.aafp.org/pubs/afp/issues/2017/0801/p179.html.

[2] Pelvic Organ Prolapse: Types, Causes, Symptoms & Treatment; 2023. [Online; accessed 28. Feb. 2023]. Available from: https://my.clevelandclinic.org/health/diseases/24046-pelvic-organ-prolapse.

[3] Zoua EP, Boulvain M, Dällenbach P. The distribution of pelvic organ support defects in women undergoing pelvic organ prolapse surgery and compartment specific risk factors. International Urogynecology Journal. 2022;33(2):405. doi:10.1007/s00192-021-04826-7.

[4] Gamal Mostafa Ghoniem AM. Cystocele Repair. 2021 Apr. Available from: https://emedicine.medscape.com/article/1848220-overview.

[5] Lensen EJM, Stoutjesdijk JA, Withagen MIJ, Kluivers KB, Vierhout ME. Technique of anterior colporrhaphy: a Dutch evaluation. International Urogynecology Journal. 2011;22(5):557. doi:10.1007/s00192-010-1353-4.

[6] Ma Z, Tavares JMRS, Jorge RN, Mascarenhas T. A review of algorithms for medical image segmentation and their applications to the female pelvic cavity. Comput Methods Biomech Biomed Eng. 2010 Apr;13(2):235-46. doi:10.1080/10255840903131878.

[7] Feng F, Ashton-Miller JA, DeLancey JOL, Luo J. Convolutional neural network-based pelvic floor structure segmentation using magnetic resonance imaging in pelvic organ prolapse. Med Phys. 2020 Sep;47(9):4281-93. doi:10.1002/mp.14377.

[8] Straetemans L. Automatic 3D bladder segmentation from low-field MR images using 3D U-Net. 2022 Mar.

[9] Morsinkhof LM, Schulten MK, DeLancey JOL, Simonis FFJ, Grob ATM. Pelvic inclination correction system for magnetic resonance imaging analysis of pelvic organ prolapse in upright position. International Urogynecology Journal. 2022 Oct;33:2801-7. doi:10.1007/s00192-022-05289-0.

[10] Reiner CS, Williamson T, Winklehner T, Lisse S, Fink D, DeLancey JOL, et al. The 3D Pelvic Inclination Correction System (PICS): A universally applicable coordinate system for isovolumetric imaging measurements, tested in women with pelvic organ prolapse (POP). Computerized medical imaging and graphics : the official journal of the Computerized Medical Imaging Society. 2017 Jul;59:28. doi:10.1016/j.compmedimag.2017.05.005.

[11] Najjari L, Hennemann J, Larscheid P, Papathemelis T, Maass N. Perineal Ultrasound as a Complement to POP-Q in the Assessment of Cystoceles. BioMed research international. 2014 11;2014:740925. doi:10.1155/2014/740925.

[12] Mittal A. Introduction to U-Net and Res-Net for Image Segmentation. Medium. 2022 Aug. Available from: https://aditi-mittal.medium.com/introduction-to-u-net-and-res-net-for-image-segmentation-9afcb432ee2f.

[13] Siddique N, Paheding S, Elkin CP, Devabhaktuni V. U-Net and Its Variants for Medical Image Segmentation: A Review of Theory and Applications. IEEE Access. 2021 Jun;9:82031-57. doi:10.1109/ACCESS.2021.3086020.

[14] Siddique N, Sidike P, Elkin C, Devabhaktuni V. U-Net and its variants for medical image segmentation: theory and applications. arXiv. 2020 Nov. doi:10.1109/ACCESS.2021.3086020.

[15] How U-net works? | ArcGIS API for Python; 2023. [Online; accessed 6. Mar. 2023]. Available from: https://developers.arcgis.com/python/guide/how-unet-works.

[16] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015 May;521:436-44. doi:10.1038/nature14539.

[17] Zou KH, Warfield SK, Bharatha A, Tempany CMC, Kaus MR, Haker SJ, et al. Statistical Validation of Image Segmentation Quality Based on a Spatial Overlap Index: Scientific Reports. Acad Radiol. 2004 Feb;11(2):178. doi:10.1016/S1076-6332(03)00671-8.

[18] Hurkmans G. Automatic 3D pelvic floor landmark detection from low-field MR images using 3D U-Net.

[19] Brownlee J. Why Do I Get Different Results Each Time in Machine Learning? - MachineLearningMastery.com. MachineLearningMastery. 2020 Aug. Available from: https://machinelearningmastery.com/different-results-each-time-in-machine-learning.