Automatic 3D pelvic landmark detection and 3D bladder segmentation from low-field MRI using 2.5D ForkNet

Huib Schulte

BSc thesis biomedical engineering

Supervisors:

Dr. Ir. W.M. Brink

Dr. Ir. F.F.J. Simonis

Dr. F. Van Den Noort PhD

University of Twente

Magnetic Detection & Imaging

Faculty of science and technology

University of Twente

2023

Summary

Pelvic organ prolapse is a component of pelvic floor dysfunction and is a big issue as it affects half of all women over the age of 50 years. With POP surgery, the risk of recurrence is about 10-30%. To evaluate the effect of surgery the patients are scanned with MRI before and after the operation. Manual delineation can be labour-intensive, and having a deep learning model that is able to do this properly, saves a lot of time and is more consistent than manual selection.

This study uses the MRI data from women who participated in the TORBO study at the University of Twente. The U-Net model uses all the slices of one orientation in one batch and trains the model on patterns in 2D images in three directions, this is called 2.5D. ForkNet is used to integrate the 2.5D landmark detection and bladder segmentation in one model.

The 2.5D U-Net model outperforms the 3D U-Net in bladder segmentation but falls short in landmark detection, while ForkNet presents an opportunity to integrate different anatomical features, albeit with the need for optimization in POP assessment.

Contents

1 Introduction

1.1 Pelvic organ prolapse

Pelvic organ prolapse (POP) is a component of pelvic floor dysfunction and is a big issue as it affects half of all women over the age of 50 years [1]. POP is the descent of the pelvic organs that result in a protrusion of the vagina, uterus or both [2]. Some factors increase the risk of POP, with the largest being vaginal delivery. Some of the other factors are nonmodifiable, for example, advancing age and connective tissue disorder. Some factors are modifiable, for example, obesity and smoking.

The assessment of POP is currently done by clinical examination, using the pelvic organ prolapse quantification (POP-Q) system. A system introduced by the International Continence Society. A medical expert defines six points around the vagina [3]. The position of these points is measured during coughing relative to the hymen. This relative position determines what stage of POP the patient has. This system however is based on a moving structure as reference, which may not be optimal for the purpose [4].

1.2 Clinical interference

The most common type of POP is anterior vaginal wall prolapse (AVWP), here the muscles above the vagina have weakened [5]. This can cause the bladder to slip out of place and bulge onto the vagina. The conservative nonsurgical treatments are pelvic floor exercises and a vaginal pessary. However, if these nonsurgical treatments do not improve the symptoms, surgery may be an option.

With POP surgery, the risk of recurrence is about 10-30% [6]. To evaluate the effect of surgery the patients are scanned before and after the operation. The scans will be both in a supine and standing position. This is important because in a supine position, the pelvic anatomy can be very different than when in a standing position, due to gravity pulling organs down.

Magnetic resonance imaging (MRI) is a promising complementary tool for the examination of POP. In a clinical POP-Q examination, it is hard to identify the underlying pathophysiology of the POP problem. With the use of MRI, there is more anatomical information on the pelvic region, which may give relevant information regarding POP, that could make for a better examination [7].

To evaluate the effect of surgery, several anatomical landmarks are needed, as well as a segmentation from the bladder. Manual delineation can be labour intensive, taking at least 20 minutes and up to several hours per subject, and is prone to inter-operator variations. This limits large-scale studies in adequately evaluating POP interventions and so, having a deep learning model that is able to do this properly, saves a lot of time and is more consistent than manual selection.

1.3 Previous research

Previous research projects at the University of Twente have used deep learning to identify landmarks in the pelvic region [8, 9]. They show potential for landmark detection. If this model is trained for anatomical landmarks relevant to the assessment of POP, these can be used for AI-based POP assessment.

Another previous research project at our institution used deep learning to segment the bladder in low-field MR images [10]. Here was shown the model could predict the lower part of the bladder correctly for patients without POP. The lower part is the most relevant part for the assessment of POP. However, the model was not trained with data from POP patients. Resulting in poorer results segmenting prolapsed bladders. Following this, another approach was taken. This approach took data from POP patients as part of the training data [11]. This study also showed good segmentation. Again, especially on the lower part, which is the most important for POP assessment.

1.4 Report outline

In this report, the three approaches mentioned above will be optimized and integrated into one AI tool for the automatic assessment of pelvic organ prolapse. First, the models will be adapted to be trained on landmarks and segmentation, factors used for AI-based POP assessment, and some of the discussion points of the reports will be taken into account.

This report is divided into different sections. Starting with the introduction in section 1. In section 2 relevant theory, concerning the evaluation of POP and the use of deep learning will be given for this report. Materials and methods, like the used data and applied AI models will be described in section 3, followed by the results in section 4. In the end, there will be a discussion and conclusion in sections 5 and 6 respectively.

2 Theory

In this chapter, some concepts used in this study will be briefly explained.

2.1 Anterior colporrhaphy

Colporrhaphy is a surgical treatment performed to treat AVWP. With this surgery, the vaginal walls are reinforced with dissolvable sutures to support the bladder and rectum [5]. A questionnaire evaluating the technique of anterior colporrhaphy was conducted among the members of the Dutch Urogynecologic Society [12]. The conclusion was that there was no widely accepted opinion on the best surgical approach.

To evaluate the results of the anterior colporrhaphy one can use MRI. In the TORBO study patients are scanned before and after the surgery in standing and supine position. The evaluation is performed by looking at the extent of bladder prolapse, before and after the surgery because the anterior colporrhaphy should support the bladder.

2.2 Magnetic resonance imaging

The MRI data used in this study was collected using the Esaote G-scan BRIO 0.25T. This MRI scanner has the possibility to scan in a supine and in a standing position. The POP-Q examination is done in a supine position. In patients with POP, there are significant anatomical differences when in a supine or standing position[13], as is visible in figure 1. These differences can be very important during the examination of POP.

2.3 PICS

For reliable pelvic measurements, a 3D coordinate system is often used, based on bony structures as reference points [15]. These landmark reference points provide a well-defined and rigid 3D coordinate system, which can follow movement. Also,

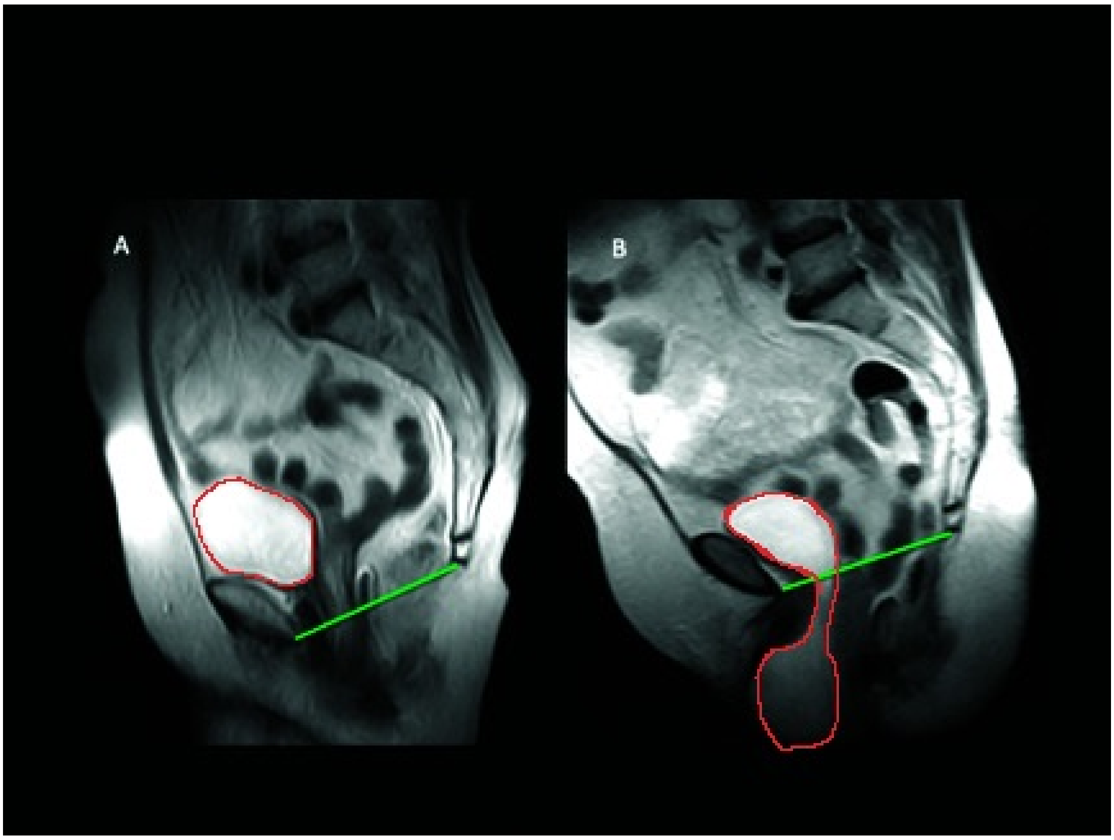


Figure 1: *In this figure the differences between supine and standing position are clearly visible. The left MR image is acquired in a supine position and the right image is acquired in a standing position. The red line is drawn around the bladder. In the standing position a prolapse is clearly visible, whereas in the supine position the bladder seems fairly normal. [14]*

bone structures are easy to find in an MRI scan.

Four landmarks are used to define this 3D coordinate system. Two are along the midline, those are: the inferior pubic point and the sacrococcygeal point (see figure 2). Laterally, the left and right ischial spine points are chosen. These are used to make a 3D cartesian coordinate system in which the PICS-plane is made. For POP assessment, the volume of the bladder underneath the PICS-plane can be looked at.

2.4 Deep learning

The interpretation of MR images can be a cumbersome process and very costly [16]. However, in recent years the impact of artificial intelligence (AI) technology in

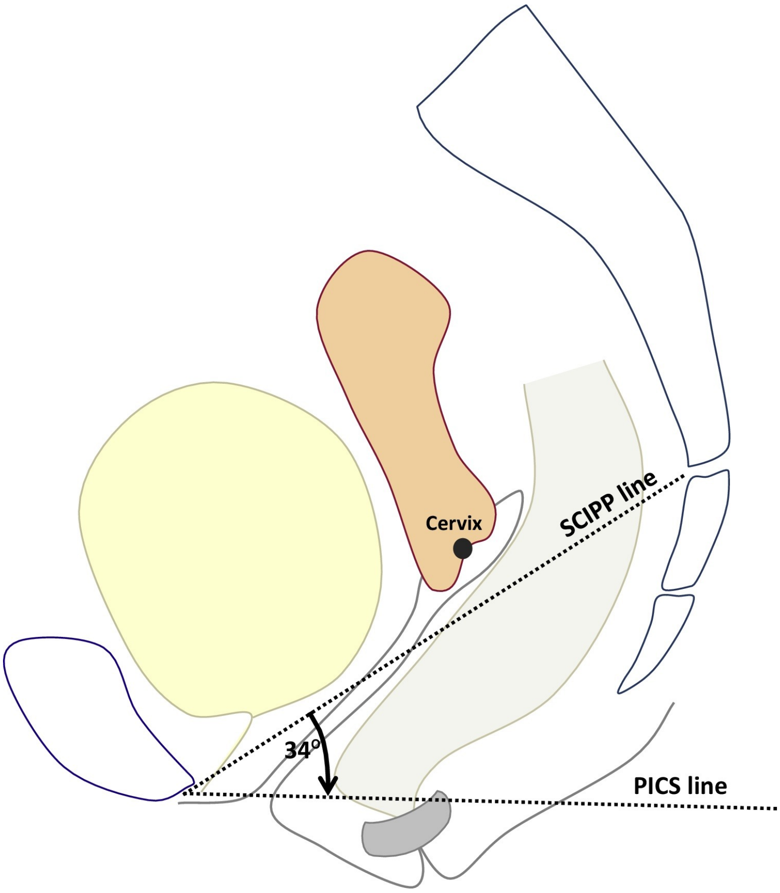


Figure 2: *In this figure both the landmarks on the midline are shown. On the left is the inferior pubic point. On the right is the sacrococcygeal point the PICS line is rotated 34°to align with the longitudinal body axis in the standing position. [15]*

healthcare has grown [17]. AI uses neural networks [18], which consist of artificial neurons, with considerable equivalence to the human brain. A neural network typically consists of input nodes, hidden layers and output nodes[19]. An artificial neuron consists of different components [20], as shown in figure 3. By using AI, image segmentation can be done automatically, this saves a lot of time.

2.4.1 Weights and biases

Weights are the values attached to each input. They convey the importance of that corresponding input in predicting the final output [20]. Inputs with weights closer to zero are less important for prediction compared to inputs with a higher weight value. Biases are used for shifting the activation function. Weights and biases are parameters trained by the model to have a predicted value as close as

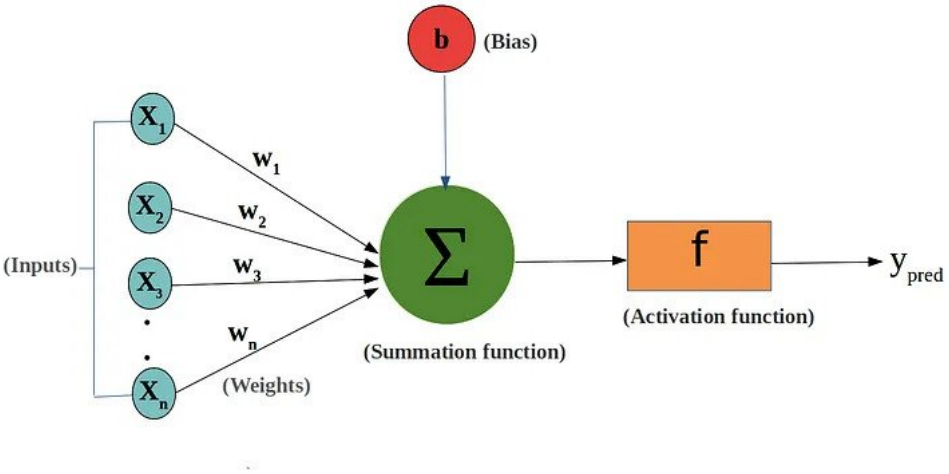


Figure 3: *The schematic representation of an artificial neuron. With different inputs, each with its own weight. All summated with a bias, after which it goes trough an activation function. [20]*

possible to the ground truth value.

2.4.2 Activation functions

The transfer function is a function that determines the output of a node. It normalizes the output. This output determines if the neuron is activated or not [21]. There are different activation functions, that all influence the input in another way. Below the two most common are briefly explained.

Sigmoid function The sigmoid function exists between 0 and 1, see figure 4. Therefore it is especially used for models where the probability is predicted, since probability ranges from 0 to 1.

ReLU function The Rectified Linear Unit (ReLU) function, is a function that exists between 0 and infinity, see figure 4. The function rectifies all values lower than zero, to zero. All the values above zero, keep their value.

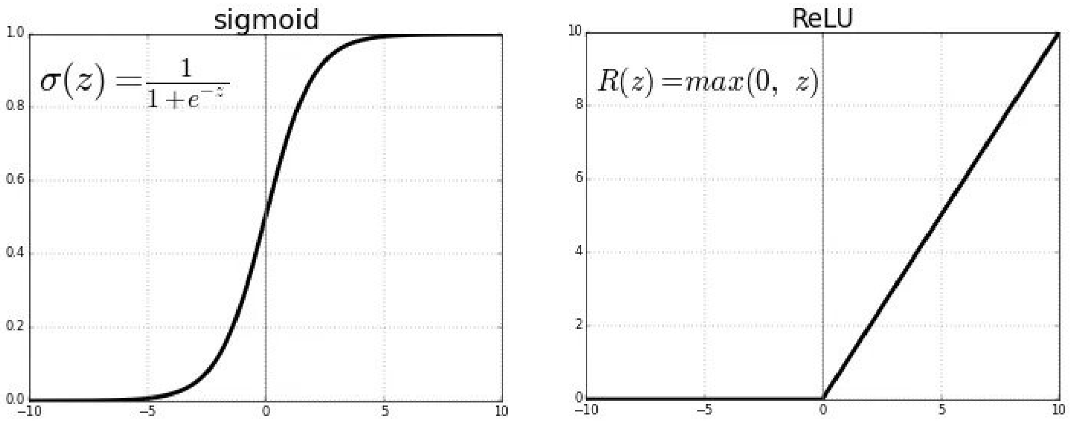


Figure 4: *The left graph is the visualization of the sigmoid function, with its formula. The right graph is the visualization of the rectified linear unit function, with its formula. [21]*

2.4.3 Error and loss function

The error in deep learning is the difference between the predicted output and the desired output. In a model, a loss function is used. A common example of a loss function is the mean squared error (MSE). A loss function measures the calculated error for a single training.

2.4.4 Optimization functions

Optimization functions are algorithms or methods used to change attributes of your neural network in order to reduce losses [22]. An example is ADAM, adaptive moment estimation. This optimizer is a bit slower than some of the other options, but this is to ensure a smooth convergence to the global minimum.

2.4.5 Hyperparameters

Hyperparameters are parameters that define the model's architecture [23]. Examples of hyperparameters are the learning rate (lr) and weight decay (wd).

2.5 Convolutional neural networks

Convolutional neural networks (CNN) are a building block for deep learning methods, primarily used for image-driven pattern recognition tasks. CNNs are computational processing systems that are based on the biological nervous systems[24]. The biggest benefit of using CNNs instead of other artificial neural networks is that CNNs are reduced in the number of parameters. This is because it looks at local regions instead of the whole image [25].

2.5.1 Components of convolutional neural networks

CNNs have three different kinds of layers [26]: convolutional layers, pooling layers and fully connected layers. In a convolutional layer a kernel slides over the input data with strides, performing an elementwise multiplication [27]. It sums up the results into a single output pixel. The kernel will perform the same operation for every location it slides over, transforming an input matrix into a feature matrix, see figure 5. A pooling layer reduces the number of parameters of the input tensor. It does this by moving a kernel over the matrix and taking one value. There are two types of pooling layers: max pooling and average pooling. With max pooling the maximum value of the matrix is put in the corresponding output matrix, see figure 6. With average pooling the average of the matrix is put in the corresponding output matrix.

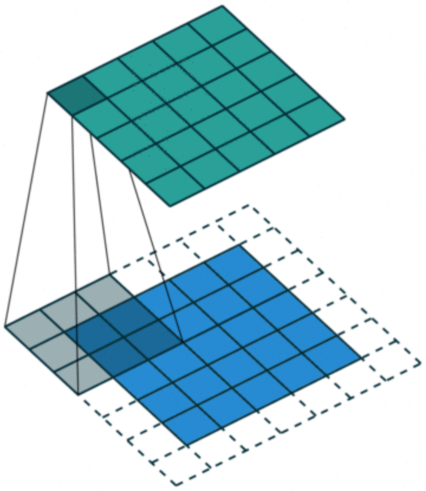
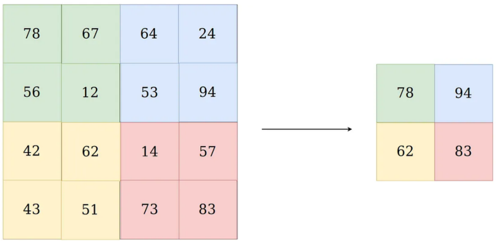


Figure 5: *Visual representation of convolution, where the kernel slides over the input matrix with strides of 1.*



78	67	64	24
56	12	53	94
42	62	14	57
43	51	73	83

78	94
62	83

Figure 6: *Visual representation of max pooling where the max value of each 2x2 block is put in the corresponding output matrix*

2.5.2 U-Net

U-Net is a deep-learning technique widely used in the medical imaging community [28]. The basic structure of a U-Net architecture consists of two paths: an encoding path and a decoding path. The encoding path is similar to a regular CNN. The U-Net however, distinguishes itself by using the decoding path. In each stage, it upsamples the feature map using up-convolution. Then the feature map from the corresponding layer in the encoding path is cropped and concatenated onto the upsampled feature map. Following this, there will be two successive convolutions and ReLu activations. At last, a 1x1 convolution is applied to reduce the feature map to the required number of channels and make the segmented image. The decoding and encoding path are more or less symmetric [29], this yields a U-shaped architecture as visible in figure 7

2.5.3 ForkNet

ForkNet is a CNN architecture originally proposed for the construction of human head models from MRI images[30]. The architecture is different from conventional U-Net structures, through the way it handles individual decoder paths for each individual anatomical structure. The basic architecture is visible in figure 8, here a ForkNet with 2 outputs is shown.

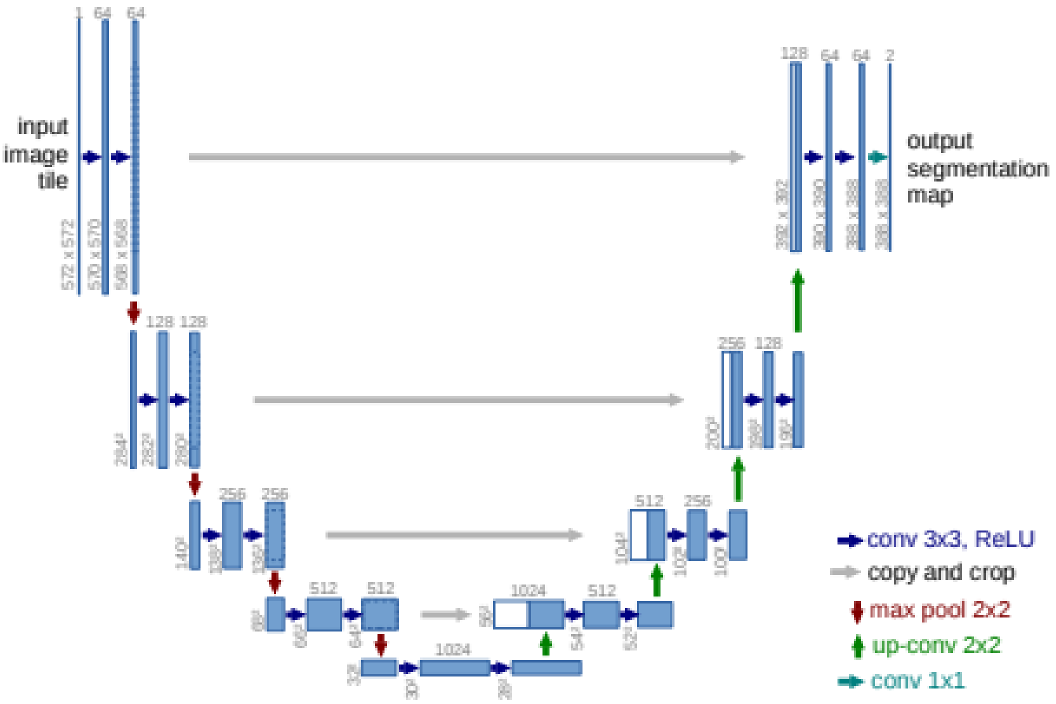


Figure 7: *Visualisation of U-Net architecture, with on the left the input. Different convolutional and pooling layers in the middle and on the right the output. [29]*

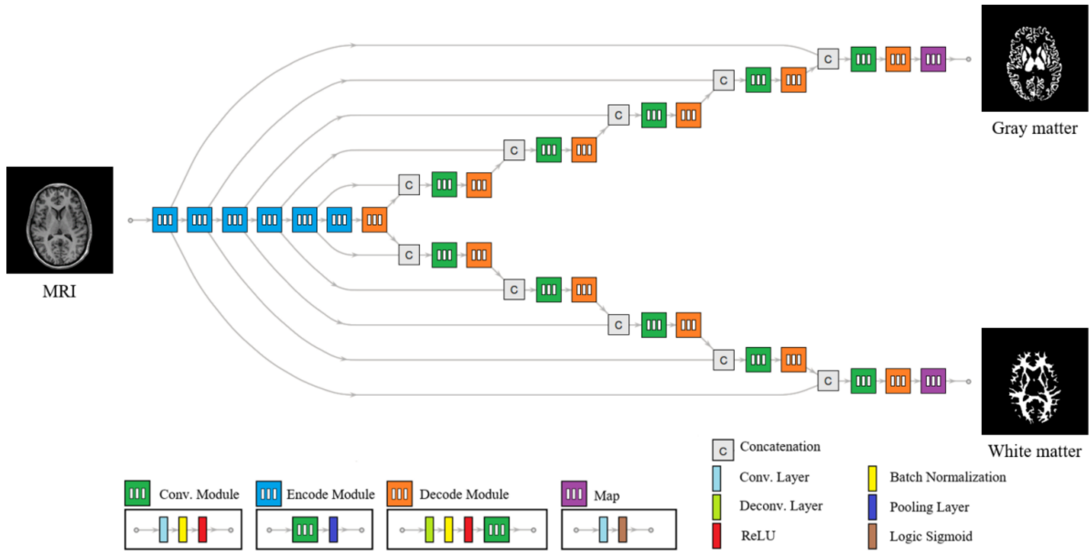


Figure 8: *Visualisation of basic ForkNet architecture, with on the left the input. Different convolutional and pooling layers in the middle and on the right the two outputs, this ForkNet has a degree of N=2. In the middle, the decoder track for each anatomical structure starts its own path. [30]*

2.6 2.5D Deep learning

Previous approaches [10, 11] for bladder segmentation used 3D deep learning networks. In other segmentation research, another approach is being used [31]. This approach uses a 2.5D approach, ensembling 3 orthogonal views to segment. With MRI scans this means it trains the model in three different directions: sagittal, coronal and transversal. This approach uses less computing power and therefore is quicker.

2.7 Evaluation

For landmarks and segmentation, different evaluation methods will be used.

2.7.1 Euclidian distance

For evaluating the performance of the model in landmark detection, Euclidian distance will be used. The Euclidian distance is the distance from the predicted point to the ground truth point. Euclidian distance= $\sqrt{(x_p - x_g)^2 + (y_p - y_g)^2 + (z_p - z_g)^2}$. With x$_p$ being the predicted x position of the landmark and x$_g$, the x position of the ground truth landmark.

2.7.2 Dice similarity coefficient

For evaluating the performance of the model in bladder segmentation, the dice similarity coefficient (DSC) will be used. The dice similarity coefficient is a statistical tool which measures the similarity between two sets of data [32]. The two sets in this study are the predicted bladder segmentation and its ground truth. The score can be between 0 and 1, 0 being no overlap and 1 being perfect overlap. The formula is as follows: DSC $= 2$ * $(|p_{bladder}\cap$ g$_{bladder}|)/(|p_{bladder}| + |g_{bladder}|)$, where $p_{bladder}$ are pixels that are considered part of the bladder by the prediction of the model and $g_{bladder}$ are pixels that are part of the bladder of the ground truth.

3 Materials and methods

In this chapter, the used methods will be explained

3.1 Data

This study uses the MRI data from women who participated in the TORBO study at the University of Twente, see figure 9 for an example of an MRI scan. For this study, researchers are looking for the effect of an anterior colporrhaphy operation. Patients are scanned before the operation and 6 weeks after. They are scanned in the supine position and in a standing position which is an angle of 81 degrees with respect to the supine position. So, there are four scans per patient. For the different sets of training for the model, 29 scans were used. For testing, 4 other scans were used. They were chosen from different patients at different times concerning the operation and in different positions. This is to try and train the model with as varied an input as possible. The 29 training and 4 test scans are the same as previous research [11], so a good comparison could be made between models regarding segmentation.

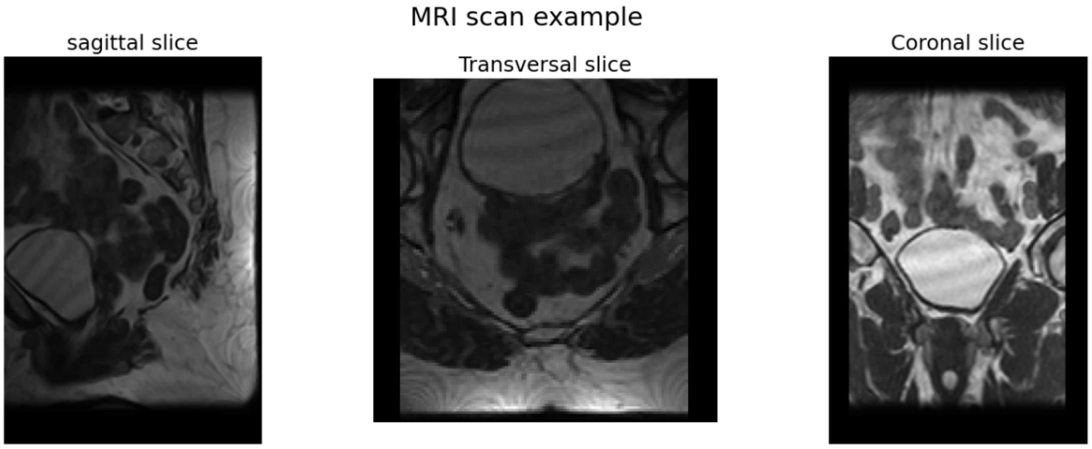


Figure 9: *An example of an MRI scan from the TORBO study at the UT. All three orientations are shown in this figure.*

3.1.1 Input

Different inputs are needed to train the model for segmentation and landmarks detection using MRI images. The MRI scan will be given as input as a NIfTI file.

Landmark detection uses heatmaps as ground truth. A heatmap is an image with values between 0 and 1. With the pixel on the place of the coordinate of the landmark having 1 as the value, from there, a Gaussian or Laplacian distribution is made, an example is visible in figure 10. This study will look at what is the best distribution to train the model. This will be done by trying Gaussian and Laplacian distributions with varying standard deviations of 2, 4, 8 and 16 pixels. **Bladder segmentation** uses manual bladder segmentation as input. This is an image with binary values. So, if a pixel is part of the bladder it has a value of 1. Otherwise, it has a value of 0, in figure 11 is an example.

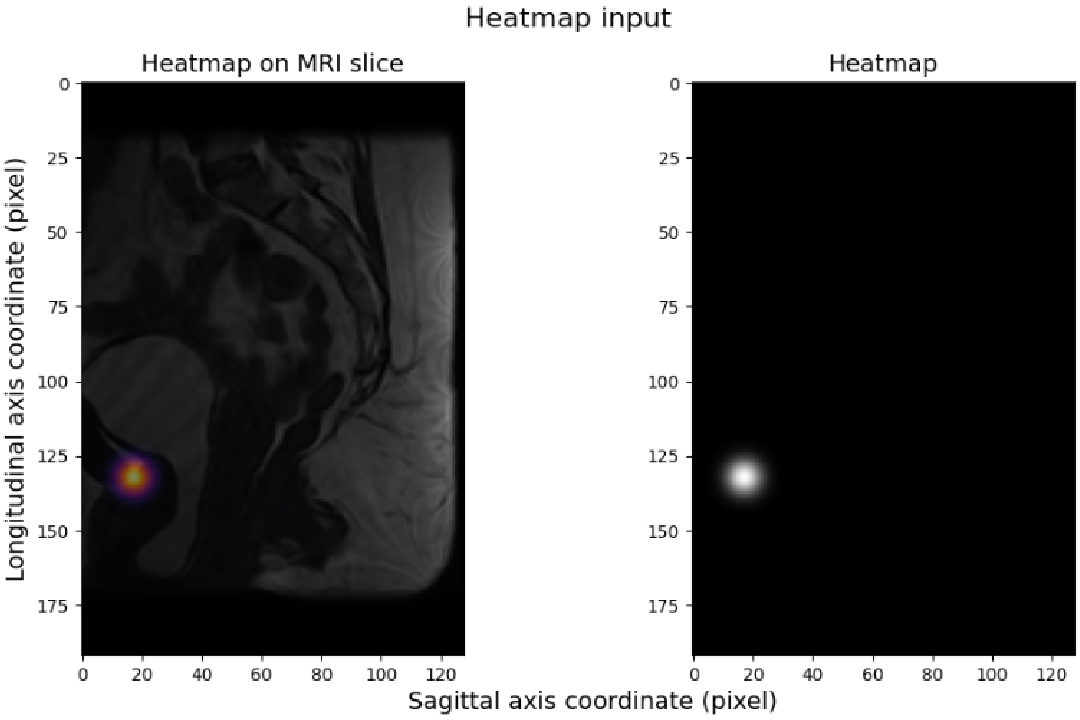


Figure 10: *On the left is a sagittal MRI image with the heatmap on the inferior point of the pubic bone. On the right, only the heatmap is visible, which will be an input to the model. The heatmap is Gaussian distributed with a standard deviation of 4 pixels.*

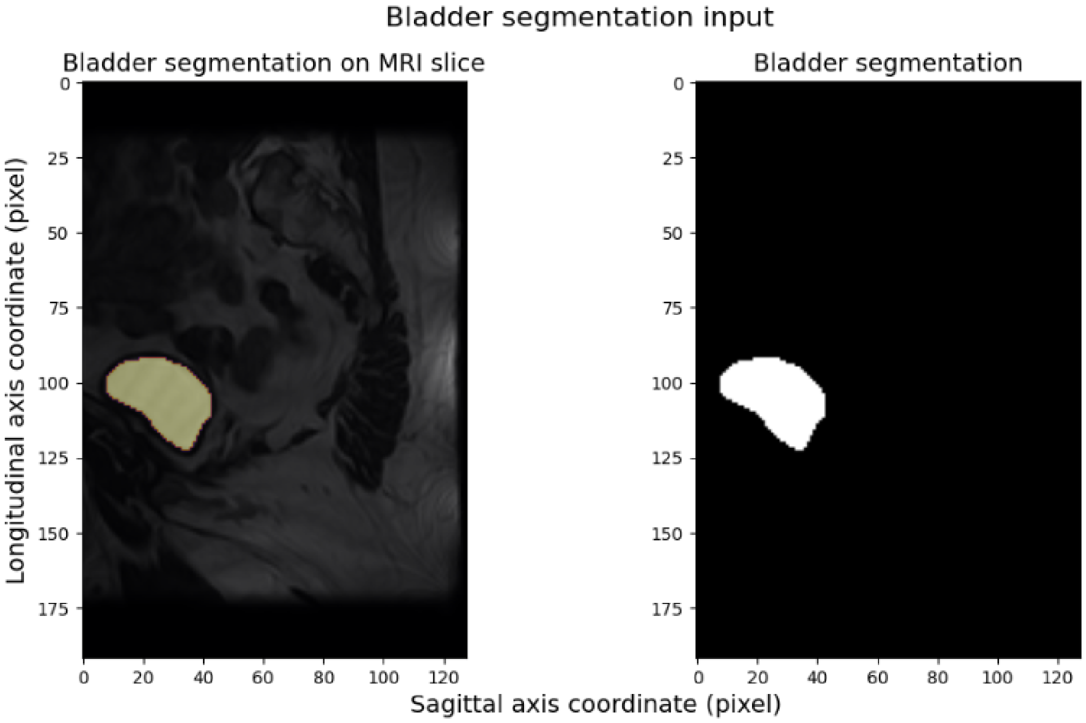

Bladder segmentation input
Bladder segmentation on MRI slice
Bladder segmentation
Longitudinal axis coordinate (pixel)
Sagittal axis coordinate (pixel)
0
25
50
75
100
125
150
175
0
20
40
60
80
100
120

Figure 11: *On the left, a sagittal MRI image with the segmentation of the bladder in yellow. On the right, only the segmentation is visible, which will be an input to the model. The map is binary, so it only contains values that are either 0 or 1.*

3.2 2.5D U-Net

This study uses U-Net, like the other approaches mentioned earlier. However, this study approaches the U-Net differently. The previous studies used 3D U-Net. This means that the model is trained on 3D information. It takes voxels as input and tries to recognize patterns in the volume of an MR scan.

In this study, a 2D U-Net in three directions is used. This divides the scans into slices in different orientations, namely sagittal, coronal and transversal. The model then combines all the slices of one orientation in one batch and trains the model on patterns in 2D images. After this, the model will be trained for the other orientations. As the model will be trained on 2D data in 3 directions, this is called 2.5D.

This means that three different models will be trained in 2D. When testing the model, each test scan will be tested in 2D. After which they will be summed up and divided by three.

3.3 Landmarks

Landmarks are used for the assessment of POP. For landmark detection in this study, scripts were used from previous studies [8, 9]. After this, landmark detection with a 2.5D U-Net was tried. As ground truth, different variations of heatmaps were tried. This is to test which ground truth is the best for the model to learn. Both Gaussian and Laplacian heatmaps were tried with the standard deviation varying between 2, 4, 8 and 16 pixels. In figure 12 is an example of different distributions in 2D.

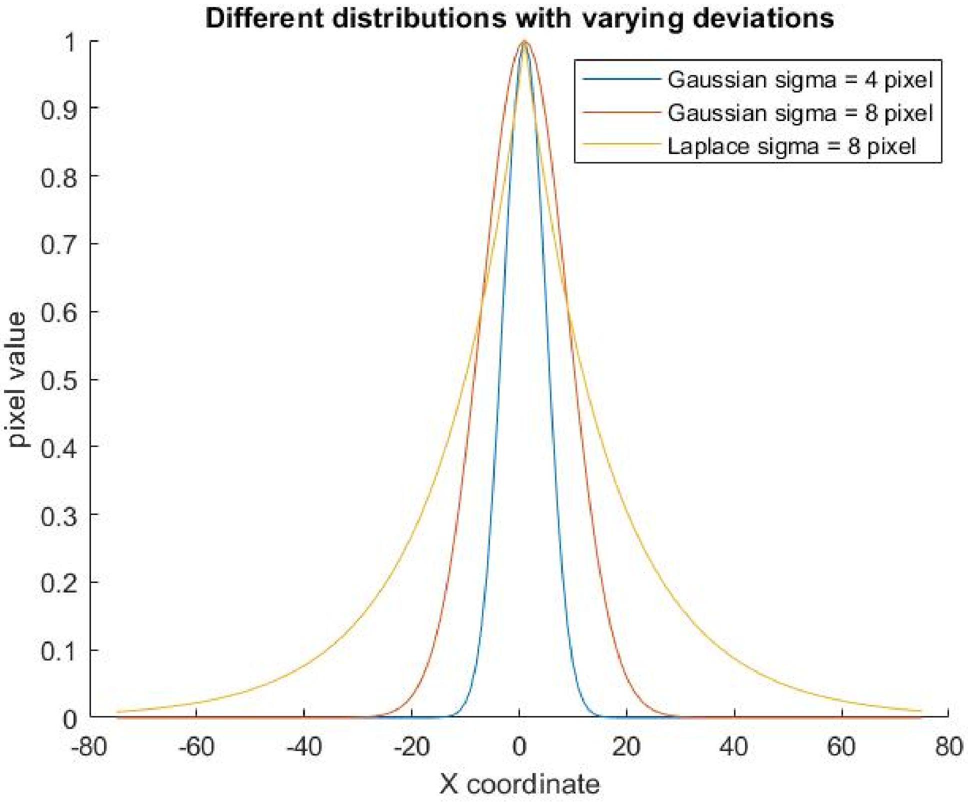


Figure 12: *In this figure different distributions with different deviations are visible. In yellow a Laplacian distribution is visible. In red and blue Gaussian distributions are visible.*

The output of the model was fitted to the heatmap distribution that was used for the input. So, if the model was trained with a heatmap with a Gaussian distribution with $\sigma = 4$ pixels, the output will be fitted to a Gaussian distribution with $\sigma = 4$ pixels. Fitting means that it will construct a Gaussian heatmap with $\sigma = 4$ pixels that fits best with the output data points. By fitting the output, a predicted landmark coordinate can be determined from a heatmap. This coordinate is needed for the assessment of POP.

3.4 Bladder segmentation

For bladder segmentation, the 2.5D U-Net was used. In previous studies, 3D U-Net was used. The goal of doing the 2.5D U-Net was to see if fewer scans would give a better result in segmenting the bladder than a 3D U-Net. To test this, a comparison with a different study was made [11].

The 2.5D U-Net will be compared to the 3D U-Net based on the DSC score and usability for POP assessment. It will be compared using different amounts of training data. The model will be trained with 15, 20, 25 and 29 training scans. The training scans are the same as the ones used in the previous study. The models will be tested on 4 scans, these also will be the same as in the previous study.

For bladder segmentation, the output of every orientation will be subjected to a sigmoid activation function. Which makes the value 0 if it predicts that the pixel is not part of the bladder and 1 if it is part of the bladder.

3.5 ForkNet architecture

In this study, ForkNet is used to integrate the 2.5D landmark detection and bladder segmentation in one model. The inputs of the model are the MRI scans, the heatmaps for the different landmarks and the manual segmentation of the bladder. The inputs will follow the same encoding and decoding path, with the same architecture as in figure 8. The convolutional and pooling layers also have the same architecture for segmentation, for landmark detection, in the last step the logic sigmoid will be replaced with a ReLu activation function. The ForkNet used in this study has a degree of N=5. That means that there will be 5 different decoding paths, whereas in figure 8, there are only 2. The 5 outputs will be the bladder segmentation, a heatmap for the pubic bone, a heatmap for the sacrococcygeal point, a heatmap for the left ischial spine and a heatmap for the right ischial spine respectively.

This model will be built in Python, using the JupyterLab environment of the University of Twente. The model will be built using the Keras interface in Python. The optimizer ADAM will be used with its standard hyperparameters. The parameters for the Batchnormalization function are momentum = 0.9 and epsilon = 0.001.

4 Results

In this chapter, the results of this study will be shown.

4.1 heatmap shape

A heatmap is an input for the training of the model for landmark detection. Different ground truths were compared as input. The outputs of the different training sessions were used to evaluate which heatmap variant leads to the most accurate landmark detection. The different variants of heatmaps were Gaussian and Laplace-distributed heatmaps. Standard deviations were 2, 4, 8 and 16 pixels. All variations were tested on 5 test scans. In table 1 the average distance is shown per heatmap variant. The results indicate that a Gaussian-distributed heatmap with a standard deviation of 4 pixels as ground truth leads to the best results and therefore will be applied in all analyses this study.

Euclidian distance								
variation	Gaussian				Laplace			
Standard deviation (pixel)	$\sigma = 16$	$\sigma = 8$	$\sigma = 4$	$\sigma = 2$	$\sigma = 16$	$\sigma = 8$	$\sigma = 4$	$\sigma = 2$
Average distance (mm)	7.3	6.1	4.3	105.5	13.7	9.0	5.5	5.7

Table 1: *Different heatmap variations and its Euclidian distance from the predicted landmark coordinate to the ground truth coordinate.*

4.2 Model output

4.2.1 Landmark detection

The output of the model also is in the form of a heatmap, shown in figure 13. The maximum value is not 1. Also, the distribution is not a perfect Gaussian distribution, which is represented in figure 14. The output was fitted to a Gaussian distribution with a standard deviation of 4 pixels. Fitting the output determined a point, this is what the model predicts as the coordinate of the landmark. Then the distance to the real coordinate of the landmark is calculated in mm.

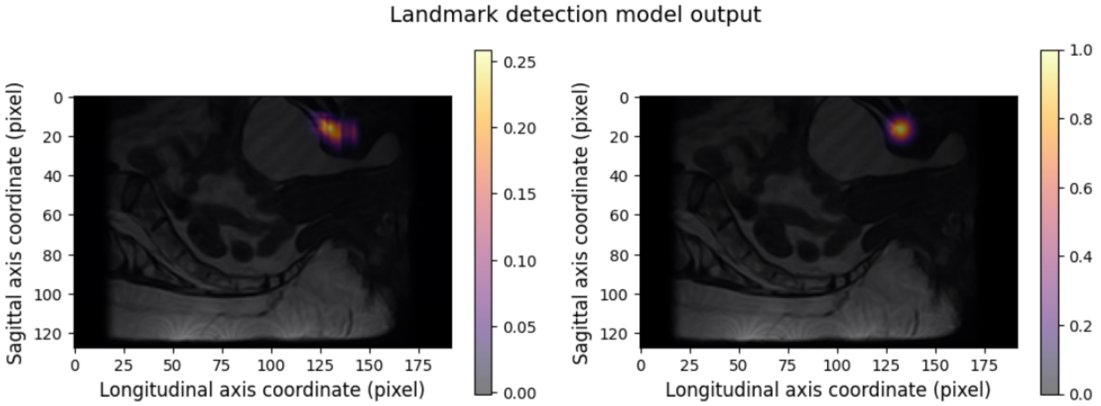

Landmark detection model output
Sagittal axis coordinate (pixel)
Longitudinal axis coordinate (pixel)
0
20
40
60
80
100
120
0
25
50
75
100
125
150
175
0.25
0.20
0.15
0.10
0.05
0.00
1.0
0.8
0.6
0.4
0.2
0.0

Figure 13: *The output of the model of landmark detection on the inferior pubic bone on the left. The maximum value is not 1, which is the case with the input on the right.*

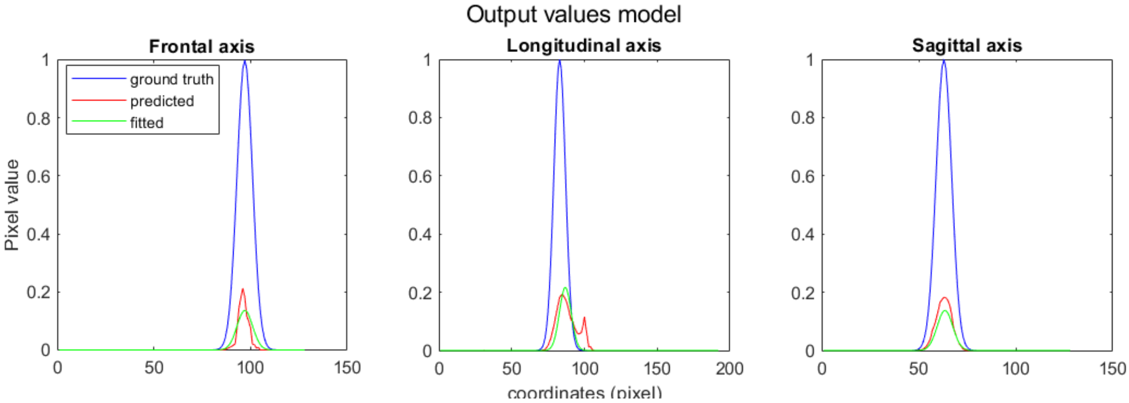


Figure 14: *The output values of the model are represented in red. These are taken from the slice of the ground truth maximum. The green is the fitted Gaussian distribution to these values. The blue is the ground truth value as a reference.*

4.2.2 Segmentation

A manually segmented bladder is an input for the model for the training of segmentation. This has binary values, if the value is 0, the pixel is not part of the bladder. If the value is 1, the pixel is part of the bladder. An output of the model is shown in figure 15. Here four different shades are visible: white, light grey, dark grey and black. These shades have values of 1, 0.67, 0.33 and 0 respectively. This is because the activation function used in the 2.5D model is a sigmoid. This leads to a value being 0 or 1. When the different orientations are summed and divided by three, the values mentioned above are established. After which the values will be made binary with a threshold of 0.5.

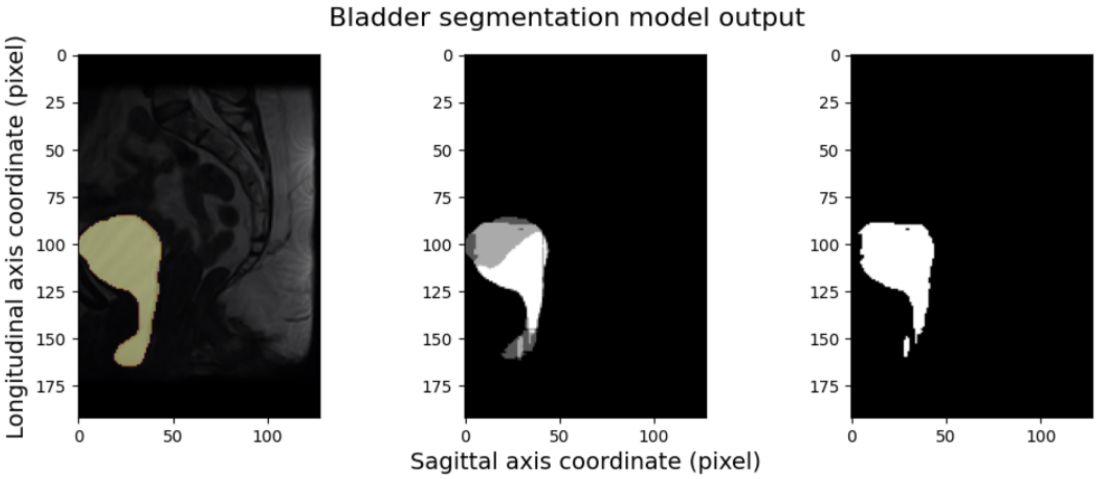

Bladder segmentation model output
Longitudinal axis coordinate (pixel)
Sagittal axis coordinate (pixel)
0
25
50
75
100
125
150
175
0
50
100

Figure 15: *On the left a sagittal view of a ground truth bladder segmentation over a MRI scan is visible. In the middle is an output of the model for bladder segmentation shown. On the right is the binary version of the model prediction, with a threshold of 0.5.*

4.3 2.5D and 3D U-Net

In this section, the results of 2.5D will be presented. To make a good comparison with 3D, results from earlier studies will also be included.

4.3.1 Landmark detection

The results of landmark detection with 2.5D ForkNet are shown in figure 16. Here, the average Euclidian distance is visible for each landmark with different amounts of training data. The average distance is the distance of the predicted landmark position to the ground truth landmark position over the four test scans.

The average Euclidian distance ranges from 3.00mm to 49.51mm. To put those distances into perspective, different predictions and ground truths are plotted onto an MRI scan in figure 17.

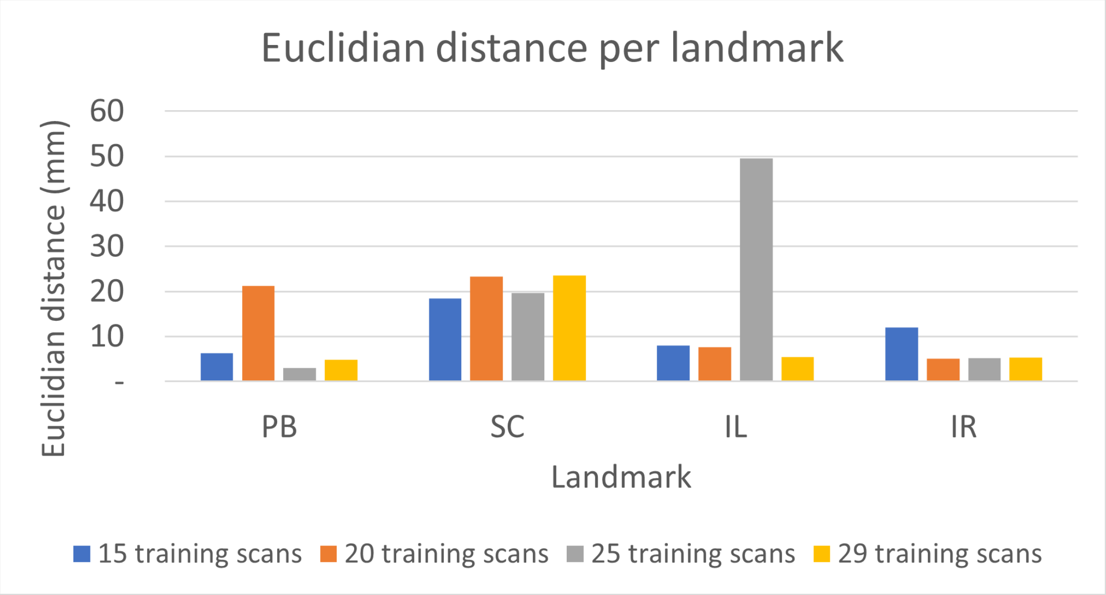


Figure 16: *In this figure the different average Euclidian distances are plotted. The average Euclidian distance is the average of the 4 test scans. The average Euclidian distances are plotted per landmark and per amount of training scans.*

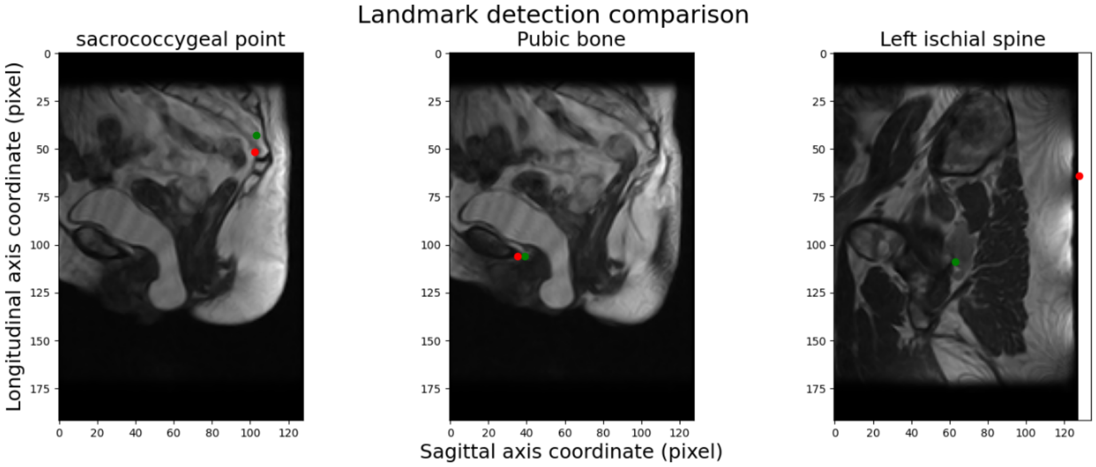

Landmark detection comparison
sacrococcygeal point
Pubic bone
Left ischial spine
Longitudinal axis coordinate (pixel)
Sagittal axis coordinate (pixel)
0
25
50
75
100
125
150
175
0
20
40
60
80
100
120

Figure 17: *Examples of landmark detection predictions by the model, for visualization of distances. In all images, a sagittal MRI image is shown. The ground truth is the green dot, and the predicted point is the red dot. In the left image, the landmark of interest is the sacrococcygeal point, here the Euclidian distance is 12.81 mm. In the middle image, the landmark of interest is the inferior pubic point, here the Euclidian distance is 5.49mm. In the right image, the landmark of interest is the left ischial spine, here the Euclidian distance is 118.55mm.*

In figure 18 the average Euclidian distances are plotted per landmark and per amount of training data. In this figure, the standard deviation is also plotted. In the graph for the different landmarks, there are large differences in the average Euclidian distance per landmark. Also, the standard deviation varies a lot and

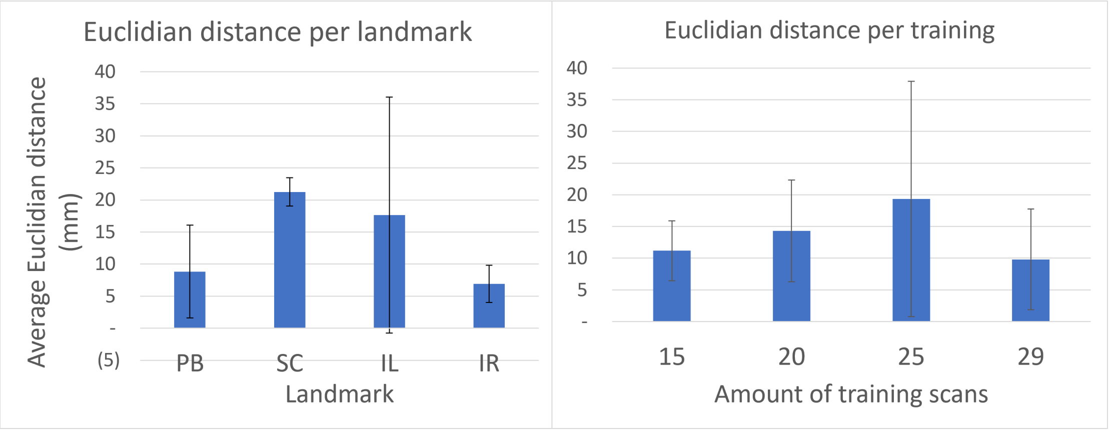


Figure 18: *Average Euclidian distance for different landmarks and different training sessions. In the left graph, the average Euclidian distance per landmark is visible. The standard deviation is included. In the right graph, the average Euclidian distance per amount of training data is visible. The standard deviation is included.*

is high, especially with the sacrococcygeal point. In the graph for the different amounts of training scans, the average Euclidian distance becomes larger from 15 to 20 and 25 training scans, which is against expectations. The batch with 29 training scans has the lowest Euclidian distance. Here, the standard deviation is also high.

The inferior pubic point is the only landmark in the 3D study, so the comparison is made based on this landmark. The 3D study used 36 training scans and 3 validation scans. The 3D U-Net training needs the validation data and training data assigned beforehand. The 2.5D splits the data into validation data and training data during the training, to make a comparison, the training and validation scans of the 3D study are counted as one batch of 39 training scans. In figure 19 the Euclidian distances are shown.

4.3.2 Segmentation

The dice similarity coefficients of the 2.5D model are shown in figure 20. The results of 3D are also included in this graph to make a comparison between the 2.5D and 3D U-Net model. The 2.5D model has DSC scores ranging from 0.80 to 0.92 with the different training scans.

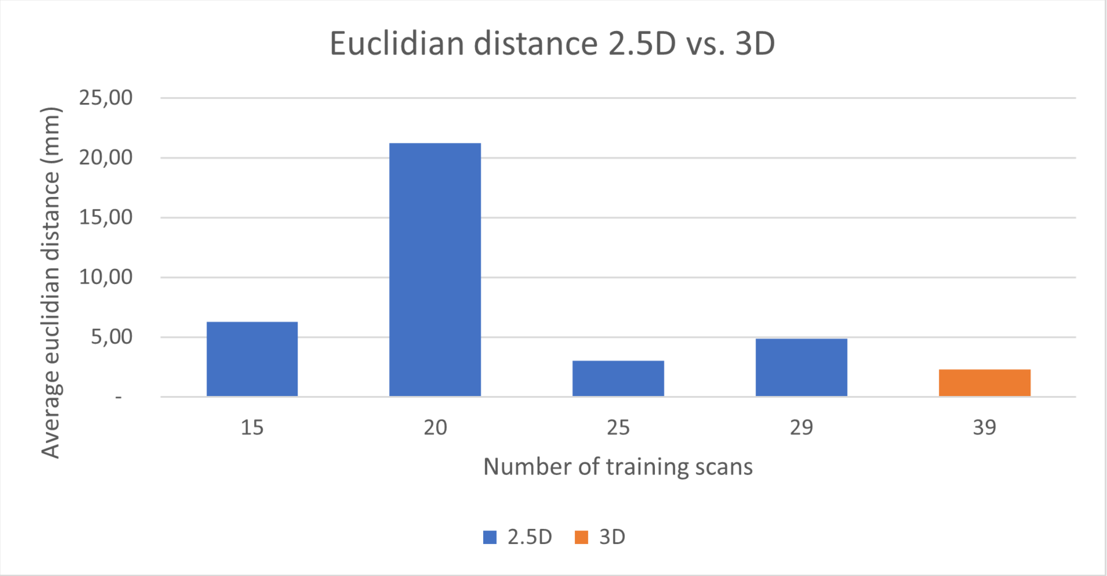


Figure 19: *The Euclidian distances of the different models, with different amounts of training data.*

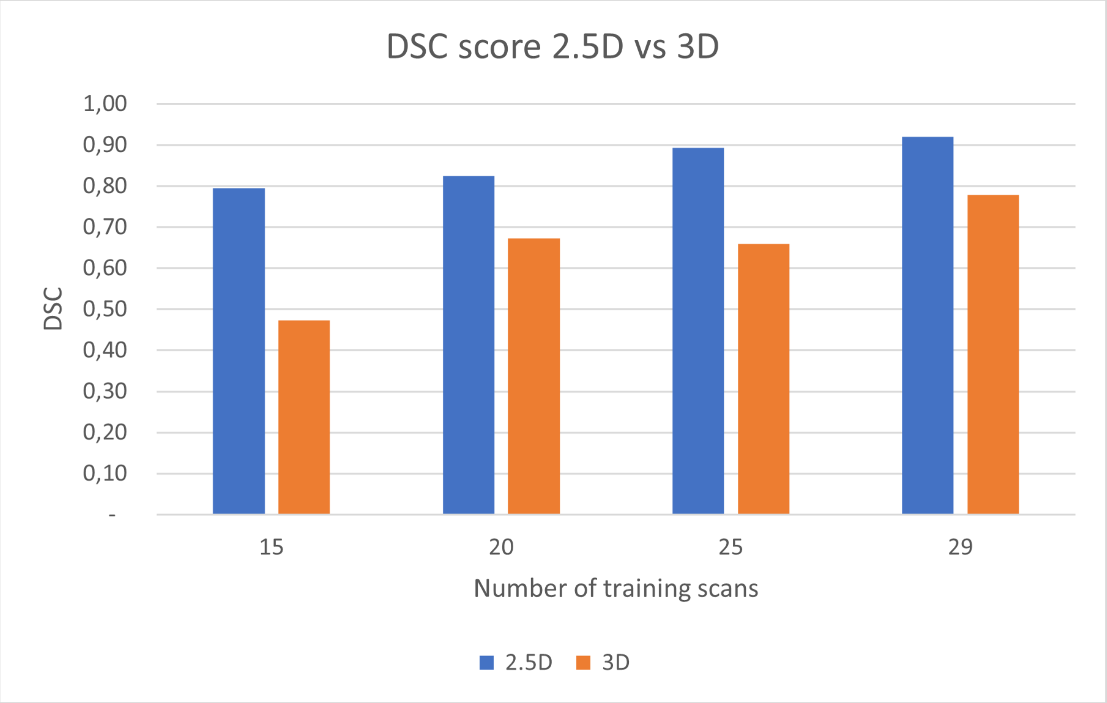


Figure 20: *The dice similarity coefficients of 2.5D and 3D U-Net models, with different amounts of training data.*

For the assessment of the anterior colporrhaphy surgery, the volume of the bladder underneath the PICS plane may be considered. So, especially the lower part of the

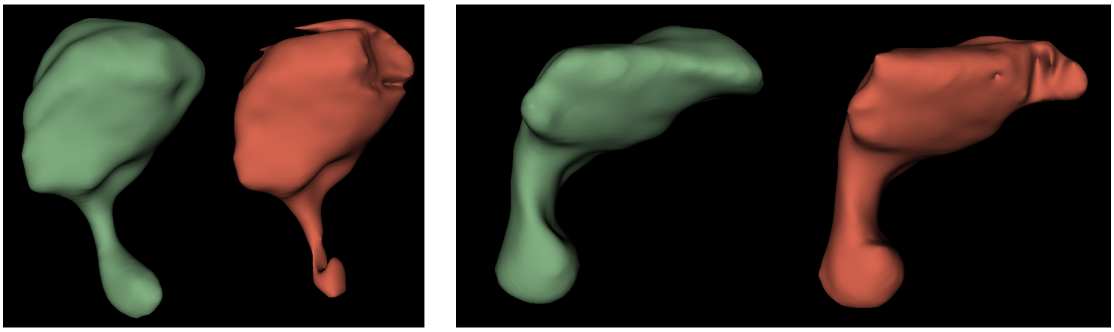


Figure 21: *In this figure model predictions of bladder segmentation are plotted next to their ground truth. The ground truths are in green, and the model predictions are in red. In the left image, a test scan (p019) is visible. On the right another test scan (p031) is visible.*

bladder is the region of interest for this study. The segmentations are compared to their ground truths in figure 21.

5 Discussion

This study shows mixed results. 2.5D U-Net seems to be better for the segmentation of the bladder than 3D U-Net. The DSC scores obtained with a 2.5D U-Net are consistently higher with the same amount of training data. Although there are limited test scans, it seems that the lower part of the bladder is not always segmented right. Which may be a problem in the context of POP assessment. The 2.5D U-Net seems to be worse for landmark detection than the 3D U-Net. The Euclidian distances are quite high and mostly inconsistent. Although we anticipated that with more training the Euclidian distance decreases, this was not always the case.

In this study the use ForkNet was successful in determining landmarks and segmentation of the bladder in one model. However, in ForkNet it is possible to only use one U-Net model. So either 2.5D or 3D has to be used.

The DSC for 2.5D was higher with the same amount of training scans, this means that when there is little data available, 2.5D would be better to perform bladder segmentation. Combining the landmarks and segmentation into one model using ForkNet in principle works, and could therefore be an option for research that is in its starting phase.

Landmark detection using a 2.5D U-Net with limited training data is very inconsistent. With more training data, it may be anticipated that the network performance improves, however, in this study, this was not the case. Also, the training results were very inconsistent. This means that the results vary a lot from test scan to test scan. It could possibly be due to the input being a 2D heatmap. In 3D there is more information from the surrounding anatomy. In 2.5D it only has the anatomical information of the slice as surrounding. Having less information in training can lead to a worse result. Also, the fitting of the output to a Gaussian heatmap model can lead to worse results. Because the prediction is based on 2D heatmap slices, there is no perfect 3D Gaussian distribution as output.

The approach using 3D landmark detection mentioned above had a 2.3 ± 0.8 mm error [8], which is a lower average Euclidian distance than this study, also much more consistent. When comparing to the literature, another 3D approach tried to localize 4 landmarks [33]. That study had errors ranging from 0.9 to 3.6 mm with 73 training images. Which is also a better result than this study. One of the previous approaches used 2.5D for landmark detection [9]. That study however used only relevant slices, whereas this study used all the slices from an MRI scan. Which can make comparing not really representative. The two landmarks that both studies have in common are the pubic bone and the sacrococcygeal point. The other study achieved Euclidian distances of 2.0 and 11.1mm respectively, which is better than the 4.84 and 23.55mm this study achieved. However, the other study used 156 training scans, and only relevant slices, which makes the training easier.

Future research could look into improving landmark detection in 2.5D, this can be done by trying other forms of input than a heatmap. Also, other loss functions can be tried. Now, the MSE is used which in combination with heatmaps as input can be the cause for poorer results. Also, the model can be tuned with parameter optimization, which could lead to better results in both landmark detection and segmentation. This could also lead to shorter training times. Finally, 2.5D ForkNet can be compared to 3D ForkNet with the four landmarks and bladder segmentation used as parameters. To compare the volume of the bladder underneath the PICS plane, also the lowest point of the bladder can be determined. Both can be compared to the ground truth, to see which could work better for POP assessment.

6 Conclusion

The 2.5D U-Net model outperforms the 3D U-Net in bladder segmentation but falls short in landmark detection, while ForkNet presents an opportunity to integrate different anatomical features, albeit with the need for optimization in POP assessment. The 2.5D U-Net model gets a higher dice similarity coefficient with bladder segmentation than the 3D U-Net, by using the same amount of input data. For landmark detection, the used 2.5D U-Net model is inconsistent and does not give better results than 3D U-Net. Using ForkNet, it is possible to combine different anatomical features into one AI model. However, it needs to be optimized to the goal of POP assessment.

References

[1] Denise Chow and Larissa V. Rodríguez. Epidemiology and prevalence of pelvic organ prolapse. *Current Opinion in Urology*, 23(4):293, July 2013.

[2] J. Eric Jelovsek, Christopher Maher, and Matthew D. Barber. Pelvic organ prolapse. *Lancet*, 369(9566):1027–1038, March 2007.

[3] Chendrimada Madhu, Steven Swift, Sophie Moloney-Geany, and Marcus J. Drake. How to use the Pelvic Organ Prolapse Quantification (POP-Q) system? *Neurourol. Urodyn.*, 37(S6):S39–S43, August 2018.

[4] K. L. Shek and H. P. Dietz. Assessment of pelvic organ prolapse: a review. *Ultrasound Obstet. Gynecol.*, 48(6):681–692, December 2016.

[5] Medical Professional. Pelvic Organ Prolapse: Types, Causes, Symptoms & Treatment - Cleveland Clinic, June 2023. [Online; accessed 22. Jun. 2023].

[6] Emmanuel Payebto Zoua, Michel Boulvain, and Patrick Dällenbach. The distribution of pelvic organ support defects in women undergoing pelvic organ prolapse surgery and compartment specific risk factors. *Int. Urogynecol. J.*, 33(2):405–409, February 2022.

[7] Suzan R. Broekhuis, Jurgen J. Fütterer, Jelle O. Barentsz, Mark E. Vierhout, and Kirsten B. Kluivers. A systematic review of clinical studies on dynamic magnetic resonance imaging of pelvic organ prolapse: the use of reference lines and anatomical landmarks. *Int. Urogynecol. J.*, 20(6):721–729, June 2009.

[8] Gijs Hurkmans. Automatic 3d pelvic floor landmark detection from low-field mr images using 3d u-net, 2022.

[9] Kyra de Bree. Development of a convolutional neural network for landmark detection of the levator plate, 2023.

[10] L Straetemans. Automatic 3d bladder segmentation from low-field mr images using 3d u-net, 2022.

[11] Maressa de Wever. Automatic 3d segmentation of prolapsed bladders from low-field mri using 3d u-net, 2023.

[12] Ellen J. M. Lensen, Jackie A. Stoutjesdijk, Mariella I. J. Withagen, Kirsten B. Kluivers, and Mark E. Vierhout. Technique of anterior colporrhaphy: a Dutch evaluation. *Int. Urogynecol. J.*, 22(5):557–561, May 2011.

[13] Boris Friedman, Lynn Stothers, Darren Lazare, and Andrew Macnab. Positional pelvic organ prolapse (POP) evaluation using open, weight-bearing magnetic resonance imaging (MRI). *Canadian Urological Association Journal*, 9(5-6):197, May 2015.

[14] Lynn Stothers, Jennifer A. Locke, Marwa Abdulaziz, Darren Lazare, Alex Kavanagh, and Andrew Macnab. Standing open magnetic resonance imaging improves detection and staging of pelvic organ prolapse. *Canadian Urological Association Journal*, 16(1):E20, January 2022.

[15] Caecilia S. Reiner, Tom Williamson, Thomas Winklehner, Sean Lisse, Daniel Fink, John O. L. DeLancey, and Cornelia Betschart. The 3D Pelvic Inclination Correction System (PICS): A universally applicable coordinate system for isovolumetric imaging measurements, tested in women with pelvic organ prolapse (POP). *Comput. Med. Imaging Graph.*, 59:28–37, July 2017.

[16] Mark G. Bandyk, Dheeraj R. Gopireddy, Chandana Lall, K. C. Balaji, and Jose Dolz. MRI and CT bladder segmentation from classical to deep learning based approaches: Current limitations and lessons. *Comput. Biol. Med.*, 134:104472, July 2021.

[17] Dinggang Shen, Guorong Wu, and Heung-Il Suk. Deep Learning in Medical Image Analysis. *Annu. Rev. Biomed. Eng.*, 19(1):221–248, June 2017.

[18] Muhammad Imran Razzak, Saeeda Naz, and Ahmad Zaib. Deep Learning for Medical Image Processing: Overview, Challenges and the Future. In *Classifi-*

cation in BioApps, pages 323–350. Springer, Cham, Switzerland, November 2017.

[19] What Is Deep Learning? | How It Works, Techniques & Applications, June 2023. [Online; accessed 22. Jun. 2023].

[20] Kuruva Satya Ganesh. Weights and Bias in a Neural Network | Towards Data Science. *Medium*, April 2023.

[21] Sagar Sharma. Activation Functions in Neural Networks - Towards Data Science. *Medium*, November 2022.

[22] Sanket Doshi. Various Optimization Algorithms For Training Neural Network. *Medium*, December 2021.

[23] Li Yang and Abdallah Shami. On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing*, 415:295–316, November 2020.

[24] Keiron O'Shea and Ryan Nash. An Introduction to Convolutional Neural Networks. *arXiv*, November 2015.

[25] Saad Albawi, Tareq Abed Mohammed, and Saad Al-Zawi. Understanding of a convolutional neural network. In *2017 International Conference on Engineering and Technology (ICET)*, pages 1–6. IEEE, August 2017.

[26] Arc. Convolutional Neural Network - Towards Data Science. *Medium*, December 2021.

[27] What is a Convolutional Layer?, September 2023. [Online; accessed 23. Jun. 2023].

[28] Nahian Siddique, Sidike Paheding, Colin P. Elkin, and Vijay Devabhaktuni. U-Net and Its Variants for Medical Image Segmentation: A Review of Theory and Applications. *IEEE Access*, 9:82031–82057, June 2021.

[29] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. *arXiv*, May 2015.

[30] Essam A. Rashed, Jose Gomez-Tames, and Akimasa Hirata. Development of accurate human head models for personalized electromagnetic dosimetry using deep learning. *arXiv*, February 2020.

[31] Saikat Roy, David Kügler, and Martin Reuter. Are 2.5D approaches superior to 3D deep networks in whole brain segmentation? In *International Conference on Medical Imaging with Deep Learning*, pages 988–1004. PMLR, December 2022.

[32] Daniel J. Bell. Dice similarity coefficient. *Radiopaedia*, August 2021.

[33] Fei Feng, James A. Ashton-Miller, John O. L. DeLancey, and Jiajia Luo. Feasibility of a deep learning-based method for automated localization of pelvic floor landmarks using stress MR images. *Int. Urogynecol. J.*, 32(11):3069–3075, November 2021.