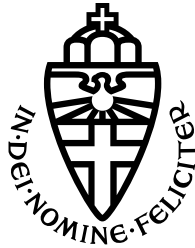


RADBOUD UNIVERSITY



Faculty of Social Sciences

Automatic Abdominal Aortic Aneurysm Detection from Ultrasound Imaging using Deep Learning

MASTER THESIS IN ARTIFICIAL INTELLIGENCE

Author:

Douwe van Erp
s4258126

Radboud University Nijmegen

Internal Supervisor:

Tom Heskes
Radboud University Nijmegen

Affiliated Supervisor:

Thomas van den Heuvel
Radboud University Medical Center

December 2021

Abstract

An abdominal aortic aneurysm (AAA) is an enlargement of the abdominal aorta. If an AAA ruptures, it leads to death in 48.5% to 81% of the cases. Detection and monitoring of AAAs is therefore vital and is currently performed by a trained sonographer at a hospital. A general practitioner is not able to perform an ultrasound, since it requires months of training. In this study, we present a method based on deep learning to automatically detect an aorta from ultrasound (US) imaging and to automatically measure the aortic diameter, so untrained people would be able to measure the aortic diameter without extensive training. The method consists of two steps. In the first step, a deep learning model with a U-Net architecture segments the aorta for each acquired US frame. In the second step, connected-component labeling (CCL) is used to find the segmented aorta, and a direct least-squares ellipse fit is performed to measure the aortic diameter. Data from 100 patients was acquired. A handheld US device was used to make an axial sweep of 7 seconds from the xiphoid process up until the umbilicus. Data of 80 patients was used to train the algorithms. 20 patients were used as a test set which showed a median Dice of 0.88 (IQR = 0.78 - 0.92). The segmentation model was included into a smartphone application, which was used to acquire data from 44 additional patients which also received a computed tomography (CT) as ground truth. The results show that the CT-US maximum diameter differences had a median of 6.0 mm (IQR = 4.0 - 9.6 mm) and 73.8% of the measurements fell within the clinically acceptable limits of agreement of ± 5 mm.

1 Introduction

An abdominal aortic aneurysm (AAA) is a localized enlargement of the abdominal aorta. An AAA is defined as having an abdominal aortic diameter larger than 3.0 cm [47]. In a patient with an AAA there is a 9.4% to 32.5% chance that the aorta may rupture, depending on its size [17]. The mortality rate from a ruptured AAA lies between 48.5% and 81% [28, 9]. Therefore early detection and monitoring is crucial. Most studies show that men at the age of 65 have an increased chance of having an AAA. The overall prevalence of AAAs in the general population is 6.0% for men and 1.6% in women [19].

Before the age of 55-65 years the prevalence is negligible, and afterwards the prevalence increases steadily with age [31]. When an AAA is detected it needs to be monitored. The monitoring interval is determined by the size of the AAA. Safe surveillance intervals have been established at every three years for small aneurysms (3-3.9 cm in diameter), annually for aneurysms 4.0-4.9 cm, and every 3-6 months for ≥ 5.0 cm [47]. Intervention policies such as AAA repair are recommended for men when the diameter is ≥ 5.5 cm, while for women they may be considered when the diameter is ≥ 5.0 cm [47].

Ultrasound imaging is a widely used method for detection and monitoring of AAA that is performed by trained sonographers. US is relatively safe and has a noninvasive nature. The sonographer can use US to measure the diameter of the aorta. US devices have recently become cheaper and more portable. Hand-held devices such as the MicrUs Pro C60S (Teleded, Lithuania) can be connected to the PC, tablet and smartphone. This makes them accessible to more users, including primary care physicians. The use of portable US devices to do an examination at a patient's bedside is referred to as point-of-care US (POCUS) [27].

Computer vision is the field that focuses on automatically extracting useful information from images. In this work, it is attempted to apply computer vision to US images, specifically by using deep learning. Deep learning is the approach of using a neural network that contains layers that are built on top of each other. If visual features are encoded within a layer, a hierarchy of features allows the model to represent complex features by building them out of simpler ones. This approach has shown great success for applications in image analysis and computer vision [16, 14, 35, 8]. Similarly, deep learning has also been applied to the medical imaging domain. The main tasks of segmentation (finding the outline of an object in an image) and classification from US imaging have been widely applied to different anatomical structures, such as the breast, prostate, liver, heart, brain and more [20, 21]. Van den Heuvel et al. [41] combined the use of low-cost, point-of-care ultrasound equipment and deep learning for the automated detection of fetal risk factors. This system was deployed on a smartphone and could be used by healthcare workers within a few hours of training [40].

To the author's knowledge there are no published works on automatic AAA detection from POCUS imaging by using deep learning in the literature. However, more traditional methods have been used in the past. In 1996, Vorp et al. [46] demonstrated the use of an automated border detection

algorithm to locate the AAA region with the largest diameter. However, this algorithm relies tissue back-scatter data and can not be used independent of the ultrasound machine. Van Essen et al. [44] used a minimum-cost algorithm to detect changes in the echo intensity corresponding to boundaries of arterial structures, but this method requires the use of a 3-dimensional (3D) intravascular ultrasound system. Rouet et al. [30], Lopata et al. [24] and Long et al. [22] all performed semi-automatic segmentation with image kernels, and automatic selection of the maximum diameter by performing least-squares fitting of an ellipse. The limitation of these methods is that they are semi-automatic. Lastly, van Disseldorp et al. [42, 43] used an active contour model for automatic segmentation, but this method relies on the use of a 3D ultrasound system. None of the above methods are both fully automatic and can be performed on 2-dimensional (2D) ultrasound images. Deep learning is fully automatic and can be performed on individual 2D US images, which makes it suitable for POCUS. Abdominal aortas can have varying features, such as shape, size, contrast and location. Additionally, the algorithm needs to be able to differentiate the aorta from other arteries and veins. To deal with the complexities of this task, deep learning was chosen as opposed to traditional computer vision method.

In this study we present a deep-learning method for automatic detection and measurement of the abdominal aorta from ultrasound imaging obtained by using a hand-held scanner. We optimized this method for efficiency, such that it could perform real-time inference on a smartphone. This enables users, e.g. a general practitioner, to use it at the point-of-care outside of the hospital in the future. The main contribution of this work is to show the feasibility of a method that automatically measures the aortic diameter of a patient.

2 Data

2.1 Data acquisition

Two datasets were acquired. In dataset 1, a total of 100 patients were included. In dataset 2, a total of 44 patients were included. The patients who were included were all scheduled for a CT scan at the Radboud university medical center, Nijmegen, the Netherlands. All patients signed a written informed consent prior to inclusion. The collection of the data used in this

study was approved by the local ethics committee. All included patients received a CT scan of the abdominal area. After the CT scan the patient directly received an ultrasound scan on the bed of the CT scanner using the MicrUs Pro C60S (Telemed, Lithuania). The MicrUs Pro C60S was connected to a smartphone.

For dataset 1, we set out to evaluate a standardized scanning protocol that could be used by non-specialists with no prior training in sonography. Scanning was performed by the author of this thesis, having no prior experience in sonography. The acquisition protocol consists of one predefined sweep over the abdomen of the patient from the xiphoid process up until the umbilicus. The sweep was performed in the axial plane. During a sweep, the ultrasound device recorded at 20 frames per second. The smartphone provided an indicator bar to show the connectivity between the transducer and the skin. US images were not shown, so there was no visual feedback during a scan. This sweep was performed within a duration of 7-9 seconds.

For dataset 2, two consecutive acquisitions were made. The first acquisition was exactly the same as the acquisition for dataset 1. The second acquisition included the display of the segmentation by the deep-learning model that was developed using dataset 1 and implemented on the smartphone (Figure 1). The aim of the second acquisition was to provide guidance during the sweep, in order to investigate whether this improved the image quality. The second acquisition had a maximum duration of 30 seconds, but recording could be stopped earlier when the user determined when the acquisition was completed. During this acquisition the user scanned along the same path as the first acquisition. During the second acquisition, it was attempted to use the segmentation and the ultrasound image as aids to follow the aorta accurately and to aim for the best possible scan quality.

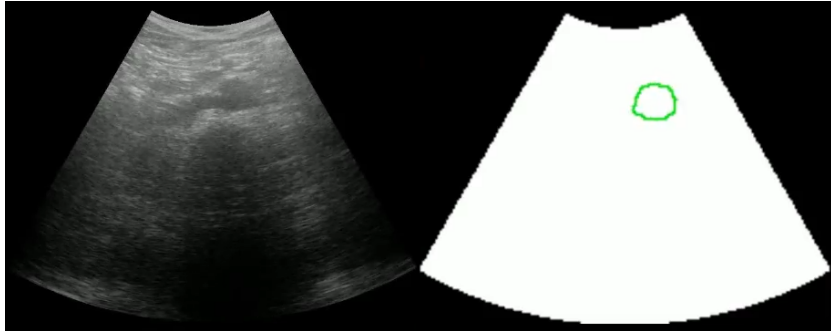


Figure 1: Left: US image for segmentation. Right: segmentation in real-time by the deep-learning model in the second acquisition. The aorta is identified by the green outline.

2.2 Beam-forming

To form a US image, sound waves are transmitted by the US transducer into the bodily tissue. The waves are then reflected back and recorded by the transducer. The waves weaken as they travel through tissue, a phenomenon known as attenuation. B-mode (brightness mode) is the most common ultrasound mode that displays a two-dimensional image where the brightness of each dot reflects the amplitude of the sound wave that is returned. The MicrUs Pro C60S transducer uses a 64-element array to send and receive sound waves. The process to convert the received data to the B-mode image is called beam-forming. Within this study we used the standard beam-forming of the MicrUs Pro C60S.

2.3 Annotation

The author of this thesis annotated the surfaces of the aortas in each frame. The aorta was annotated by including the outer aortic wall, which corresponds to the outer-to-outer method when measuring the AP diameter (Figure 2). Delineation of the lateral aortic wall is less precise in ultrasound, which is also evidenced by transverse diameter measurements being less accurate than AP diameter measurements [3]. Therefore, the annotations were made as wide as possible in the lateral direction.

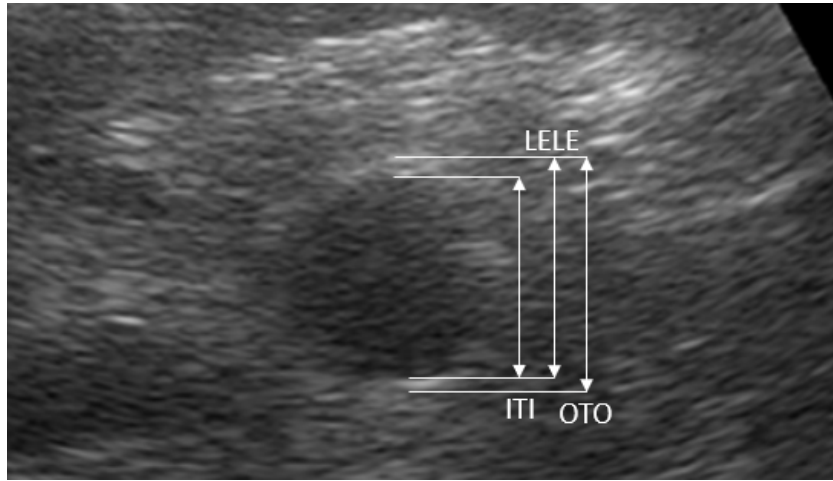


Figure 2: An ultrasound frame which indicates the outer-to-outer (OTO) measurement between the outer layers of the aortic wall, inner-to-inner (ITI) measurement between the inner layers of the aortic wall, and leading edge-to-leading edge (LELE) measurement between the outer layer of the anterior wall and the inner layer of the posterior wall.

The annotation started at the first frame in which the aorta was clearly visible. This means the complete aortic wall could be observed. After that, the aorta was annotated one in every five frames. The annotation was stopped at the last clear instance of the aorta. A margin of 10 frames was added before the first and after the last annotated frame. In the frames in the margins the aorta could still be distinguished by a human, but the aortic walls were not clearly visible. Because the deep learning algorithms should not be punished for detecting the aorta in these frames, they were not included in the training data. Outside the margins, the aorta is not distinguishable anymore. This range of annotations, including the margins, defines a positive range. The aorta could reappear further on in the scan depending on the scan quality. Therefore this process was repeated until all ranges with positive frames were annotated. A total of 16432 frames were acquired from 100 patients. A total of 4027 frames were identified as positive frames and 12405 were identified as negative frames. Out of the positive frames, 549 were annotated. Annotations were made using grand-challenge.org [45].

2.4 Pre-processing

The pixel values of input images and their corresponding segmentation maps were rescaled from the range $[0, 255]$ to $[0, 1]$. This was done with min-max scaling given by the formula:

$$\mathbf{X}_{new} = \frac{\mathbf{X} - x_{min}}{x_{max} - x_{min}} \quad (1)$$

where \mathbf{X} is a 2D matrix representing an input image, x_{min} is the minimum value in the matrix, and x_{max} is the maximum value in the matrix. Because the original values were in the range $[0, 255]$, scaling was simply done by dividing the pixel values by 255.

2.5 CT references

Each included patient received both a US scan and a CT scan. In order to compare the US with the CT, the aorta was segmented in the CT scans with an nnUNet [10]. The nnUNet was pre-trained on the abdomen segmentation task of the MICCAI2015 challenge “Multi-Atlas Labeling Beyond the Cranial Vault” [15].

For each CT scan, a starting point was selected at the umbilicus, and an end point was selected at the xiphoid process. Over this region the aortic diameter will be extracted, in order to be compared with the diameters from US scans.

It is known from the literature that aortic maximum diameter measurements from CT are significantly larger compared to US by 4.2 - 9.6 mm difference [25, 18, 38].

The quality of the segmentations was manually evaluated to ensure that they were correct. In dataset 2 there were two cases in which nnUNet failed segmentation. 42 cases were correctly segmented and were used for further analysis.

3 Methods

3.1 Segmentation model

Segmentation on the US scans was performed with a fully convolutional neural network based on the U-Net implementation of Ronneberger et al. [29].

This implementation was chosen because it remains a state-of-the-art architecture for segmentation problems, and it is often successful for US applications [34]. The U-Net architecture can be described as an encoder-decoder architecture. The decoder consists of a series of two convolutions, a rectified linear unit (ReLU) and a max pool operation (the contracting path). The encoder consists of a series of an up-convolution, two convolutions and a ReLU (the extending path). At every step the contracting and expanding paths are connected by skip connections. The final layer is a sigmoid activation layer, which returns for every pixel the probability that a pixel belongs to the foreground. In our implementation we made two adjustments compared to the U-net implementation of Ronneberger et al. We used padded convolutions instead of unpadded convolutions and we varied the number of model channels at the first step in the contracting path. Padded convolutions were used to retain the size of the input, and to preserve border information of the input image.

3.1.1 Model channels

The number of feature channels is doubled after each max pool layer in the network. The number of model channels in the first convolutional layer (referred to in this work as model channels) therefore determine the total number of parameters in the model. In the standard U-Net model there are 64 channels in the first layer. Decreasing the number of model channels lowers the model’s complexity and avoids overparameterization. In our optimization experiment we examined model combinations with 64, 32, 16 and 8 model channels respectively.

3.1.2 Downsampling

The input images were downsampled at different rates. Downsampling offers the advantage that it reduces the number of parameters in the network, and therefore also the computation time and memory cost. This is required to be able to run the model on a smartphone. Additionally, pooling produces invariance to small translations in the image [5]. Because each max pool operation in the network halves the size of the input image, the input image size needs to be divisible by 2^n , where n is the number of steps in the contracting path. In the standard implementation there are 4 steps, so the image sizes were chosen such that they were divisible by 16. Downsampling was done at

a factor of 1 (original, 752×576 pixels), 2 (384×288 pixels), 4 (192×144 pixels), 6 (128×96 pixels) and 8 (96×80 pixels).

Annotations were resized using nearest-neighbor interpolation such that they matched the size of the input.

3.1.3 3D input

An additional experiment was set up to train a model with three-dimensional (3D) input as opposed to using two-dimensional (2D) frames as input. The motivation behind this is that when humans evaluate a scan, they look at multiple successive frames to determine where the aorta is present. Therefore, this additional information might also help the network to generalize better.

The 3D input consisted of the current 2D frame and the preceding three frames. As a result, each input image has four channels. An alternative option would be to add surrounding frames to the current frame, but this option was not explored in this work due to time constraints. The 2D model on the downsampling factor and selected model channels that gave the highest performance on the validation set was retrained with this 3D input.

This method takes in account 3D information, but it is not the same as volumetric segmentation. V-Net [26] is an instance of an architecture for volumetric segmentation by using 3D convolutions. The advantage of the method we use is that it is not as computationally expensive, and it does not require modifications to the U-Net architecture. Additionally, U-Net can be used on the smartphone for real-time inference, whereas V-Net would require the whole US scan volume as input. An advantage of V-Net could be that it interpolates better between segmentations of the aorta, because it considers the whole volume at once. A V-Net model could then be used for off-line inference.

3.1.4 Training and evaluation

The data of dataset 1 was partitioned in five folds of 20 patients, such that the percentages of positive and negative frames and the number of annotated frames were approximately equal in each fold. Non-nested cross-validation was applied for all experiments. All networks were trained using a four-fold cross-validation with the train and validation set. This corresponds to a 60% train, 20% validation and 20% test split.

During training the Dice loss was used as a the loss function, defined by

$1 - \text{Dice}(X, Y)$. The Dice similarity coefficient [37, 2] (Dice) is a measure of spatial overlap accuracy, defined as

$$\text{Dice} = \frac{2|X \cap Y|}{|X| + |Y|} \quad (2)$$

where X is the matrix of predicted segmentations and Y is the matrix of the annotated segmentations. The Dice similarity coefficient has a range of $[0, 1]$, where 0 means the two segmentations have no overlapping pixels, and 1 means all pixels overlap.

Before calculating the Dice metric on the validation set, the probability values in the predicted segmentation maps were binarized with a threshold of 0.5:

$$x' = \begin{cases} 1 & \text{if } x > 0.5 \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

where x' is the binarized probability and x is the original probability value.

To fully utilize the memory of the video card, we chose the batch size s as

$$s = \min\left(256\frac{d^2}{m}, 128\right) \quad (4)$$

where d is the downsampling factor and m is the number of model channels. The batch size was limited to 128 since the number of training images (the annotated frames) was relatively small (always below 549).

Two metrics were used to evaluate the performance of the trained model: the Dice and the average number of false positive pixels per negative frame (referred to in this work as the false positive pixels metric). The false positive pixels metric is used to judge the performance on the negative frames. The Dice cannot be computed for these frames, since there are no annotations present. The average number of false positive pixels on a negative frame tells us how likely the model is to segment pixels when there is no aorta present.

The U-Net was trained on the training frames and their corresponding segmentation maps. Adam [13] was used to optimize the network with an initial learning rate of 0.001. He-normalization [7] was used to initialize the weight matrices. Early stopping was used to stop training when no improvement was found in the validation Dice for 50 epochs. The maximum number of epochs was set at 500. A fixed seed was used for all experiments to ensure reproducible results. Experiments were implemented using the Keras framework from Tensorflow 1.15.0 [1]. The models were trained and

evaluated on a single GeForce RTX 2080 Ti (Nvidia Corp., Santa Clara, CA, USA) graphics card.

3.2 Aortic diameter measurement from segmentation

The maximum aorta diameter was automatically measured from the segmentations made by the U-net. First connected-component labeling (CCL) [4] was used to determine the different components from the segmentation output. For the US scans, the relevant 3D components were selected manually. For the CT scans, the largest component was selected. Next, binary erosion is applied to each 2D segmentation. The eroded segmentation is subtracted from the segmentation to obtain the contour of the segmentation. Finally, a direct least-squares ellipse fitting [6] is computed on the contour. As a result, each frame has a fitted ellipse associated with it if a segmentation was made. The diameter d was calculated as the average of the major axis (the longest ellipse diameter) and minor axis (the shortest ellipse diameter).

3.3 Deployment on smartphone

The model with the highest Dice on the validation set that could also perform real-time inference on a smartphone was converted with TFLite (included with TensorFlow 1.15 [1]), and subsequently deployed on a OnePlus 7T (Android 11). This model was used for the acquisition of dataset 2.

3.4 Statistical analyses

The Shapiro-Wilk test [33] was performed to determine whether the Dice metric, false positive pixels metric, and aortic diameter measurements were normally distributed. If this is the case, a t-test [39] was performed and the mean and the standard deviation (SD) will be reported. Otherwise a Wilcoxon signed rank test [48] was performed, and the median and interquartile range (IQR) will be reported.

4 Results

4.1 Experiment 1

Figure 3 shows the results of the experiments in which the downsampling and model channel hyperparameters were varied. At each point, the mean metric over the four folds \pm one standard deviation (colored area) is shown. The models with 8 model channels appear most robust since they have the lowest standard deviation. The model with a downsampling factor of 2 is the only model to converge consistently over all model channel options. None of the models converged for 64 model channels. For the models with downsampling factors 4, 6 and 8, the false positive pixels metric goes above 3000 for ≥ 16 model channels. This shows that over-segmentation tends to take place in the models that do not converge. The model with downsampling factor 2 and 32 model channels shows the best performance. The model with downsampling 4 and model channels 8 is best performing model that can run in real-time (20 fps) on a smartphone. This model was therefore selected to be used on the smartphone for the collection of dataset 2.

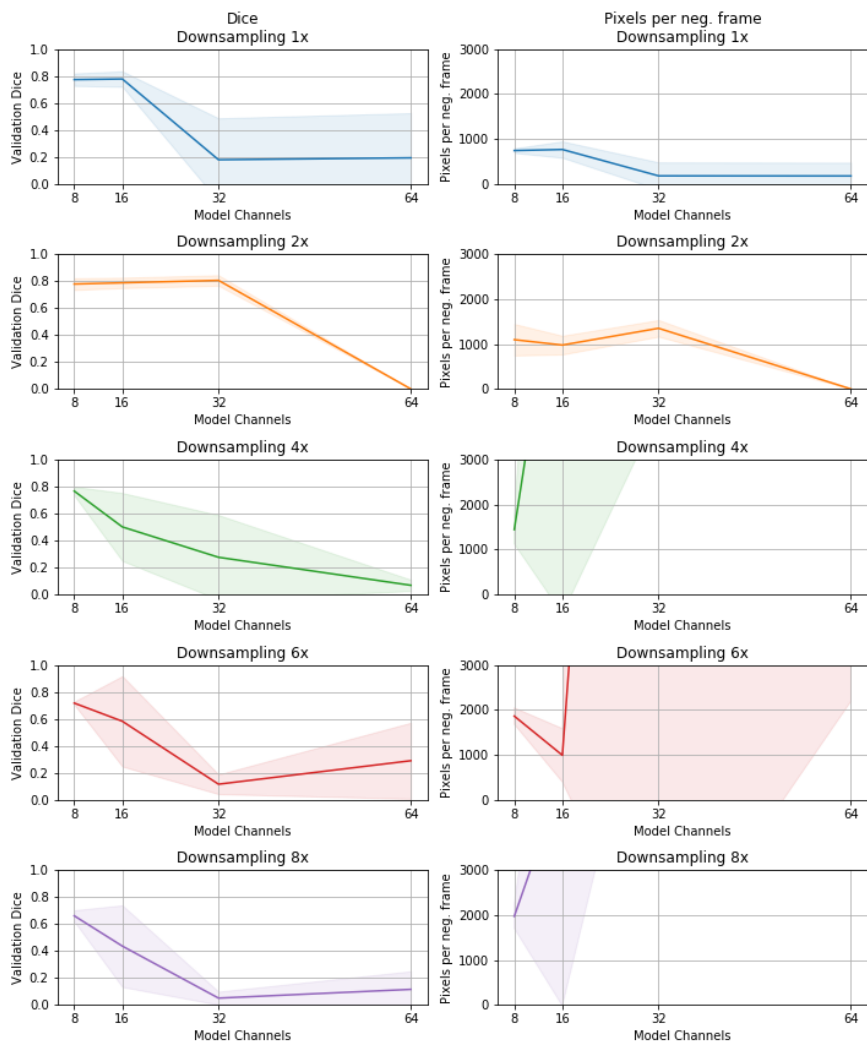


Figure 3: The validation Dice and false positive pixels metric performance for varying downsampling rates and model channels.

4.2 Experiment 2

A comparison is made between the 2D model for the smartphone “2D model^S” (downsampling 4 and 8 model channels), the best 2D model “2D model^B” (downsampling 2 and 32 model channels) and the 3D model with the same hyperparameters as 2D model^B. For each model, the Dice metric (Table 1)

is presented. Additionally the percentages of negative frames with no segmentation are reported (Table 2). For all models, the Dice scores are not normally distributed according to a Shapiro-Wilk test ($p < 0.05$), therefore the median and the IQR are reported.

Table 1: Median Dice metric

	Validation	Test
2D model ^S	0.81 (IQR = 0.73 - 0.87)	0.83 (IQR = 0.71 - 0.89)
2D model ^B	0.86 (IQR = 0.78 - 0.9)	0.86 (IQR = 0.78 - 0.9)
3D model	0.87 (IQR = 0.8 - 0.91)	0.88 (IQR = 0.78 - 0.92)

Table 2: Percentage of negative frames with no segmentation

	Validation	Test
2D model ^S	36.9%	59.2%
2D model ^B	53.7%	76.3%
3D model	61.5%	82.0%

A paired one-sample Wilcoxon Signed rank test shows that there is a significant difference in Dice scores and number of false positive pixels between all three models on the test set (Table 3). Compared to the smartphone 2D model, the best 2D model achieves a higher Dice score and a lower number of false positive pixels. The 3D model achieves both a better performance on the Dice metric and false positive pixels metric, compared to the 2D models.

Table 3: Model comparison

	p -value (Dices)	p -value (pixels)
2D model ^S vs 2D model ^B	< .001	< .001
2D model ^S vs 3D model	< .001	< .001
2D model ^B vs 3D model	.002	< .001

The segmented aorta of an example patient is shown in figure 4. This example was selected to show a fully successful segmentation of the aorta

over the annotated region. Note that in the 2D models, there are more false positive segmentations in comparison to the 3D model (around frames 100 for 2D model^B and frame 150 for 2D model^S).

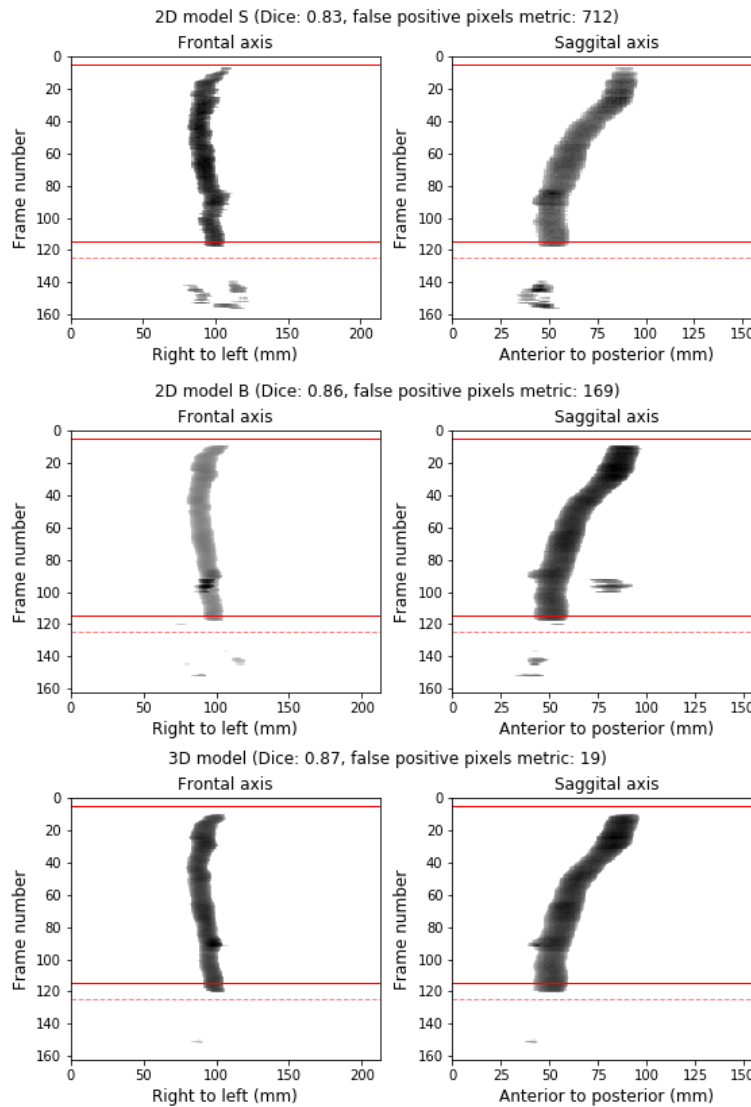


Figure 4: 2D plots of an example patient for different models. In the frontal plot, the pixels of each segmentation are summed over the y-axis of all frames. The x-axis of the plot corresponds to the right (0) and left of the patient. In the sagittal plot, the pixels of each segmentation are summed over the x-axis of the frame. The x-axis of the plot corresponds to the anterior (0) and posterior of the patient. The frame number is shown on the y-axis of the plot and progresses from superior (frame 0) to inferior. The solid red lines indicate the positive range and the dotted red lines indicate the margins (determined during annotation).

The aortic diameter measurement of the US is compared with the aortic diameter measurement of the CT, for the same example patient (Figure 5). The diameters from the US models closely follow the the CT model, although the CT diameters tend to be generally higher.

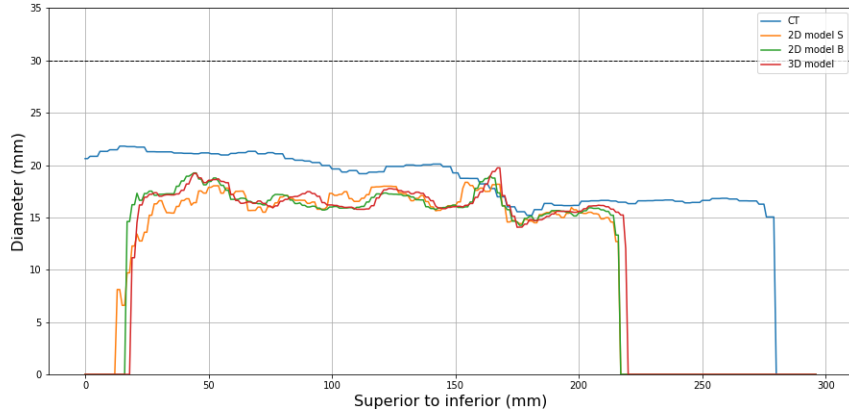


Figure 5: The course of the CT diameter for an example patient in comparison with the extracted diameters that were inferred with the segmentation models. The dotted line at 30 mm indicates the threshold for aneurysms.

4.3 Experiment 3

Table 4 shows the percentages of frames with a diameter measurement on the first acquisition in dataset 1 and dataset 2, respectively. For all models, the percentages from both dataset 1 and dataset 2 are not normally distributed according to a Shapiro-Wilk test ($p < 0.05$). For the percentage metric, the median and the IQR are reported. An unpaired two-samples Wilcoxon Signed rank test shows that there is no significant difference between the quality of the first acquisition in dataset 1 and dataset 2.

Table 4: Median percentages of frames with a diameter in acquisition 1.

	% (dataset 1)	% (dataset 2)	<i>p</i> -value
2D model ^S	17.3 (IQR = 0.0 - 52.1)	42.4 (IQR = 12.0 - 57.1)	0.08
2D model ^B	17.0 (IQR = 0.0 - 43.0)	42.6 (IQR = 12.1 - 54.7)	0.05
3D model	19.4 (IQR = 0.0 - 40.2)	37.2 (IQR = 10.0 - 53.0)	0.09

An example shows what the diameter measurement of the 3D model with 19% frames with a diameter and a model with 43.4% looks like (Figure 6).

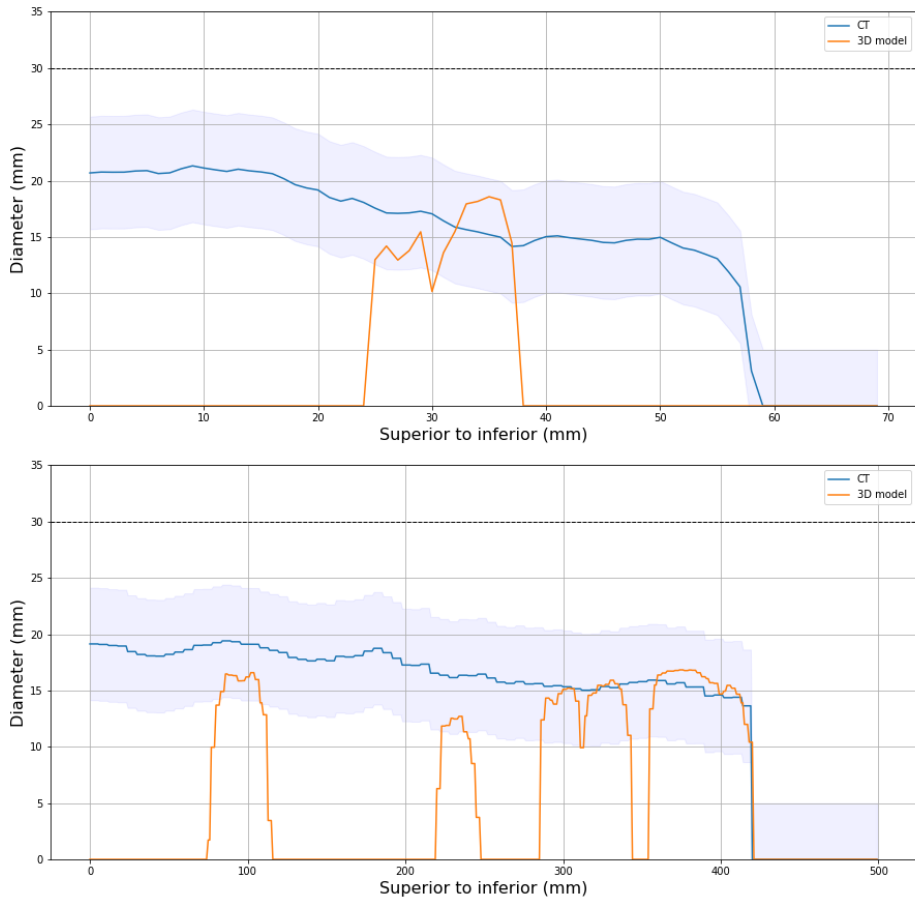


Figure 6: CT and US diameter measurements. Top: 19% frames with a diameter measurement. Bottom: 43.4% frames with a diameter measurement.

4.4 Experiment 4

A comparison was made between the first acquisition and second acquisition of data set 2, in order to investigate whether the guidance with the deep learning system resulted in better scan quality.

Dataset 2 consisted of 44 patients. The nnUNet did not perform a correct segmentation in two cases, so the results are shown for the remaining 42 cases. For differences between the automatically measured maximum aortic diameters from the CT and from the US segmentation, the median and IQR are reported (Table 5). The US diameters against the CT diameters are shown in a scatter plot (Figure 7). The diameter measurements obtained with the 3D model on acquisition 2 had significantly smaller CT-US differences than those obtained from 2D model S on acquisition 1 ($p = 0.025$) and acquisition 2 ($p = 0.042$), and the 3D model on acquisition 1 ($p = 0.005$).

Table 5: Median CT-US differences for 2D model^S, 3D model and their acquisitions. A positive median means that the CT diameter is larger than the US diameter.

Model	Acquisition 1 (mm)	Acquisition 2 (mm)
2D model ^S	7.1 (IQR = 4.6 - 10.7)	6.5 (IQR = 4.0 - 9.8)
3D model	7.4 (IQR = 4.7 - 11.8)	6.0 (IQR = 4.0 - 9.6)

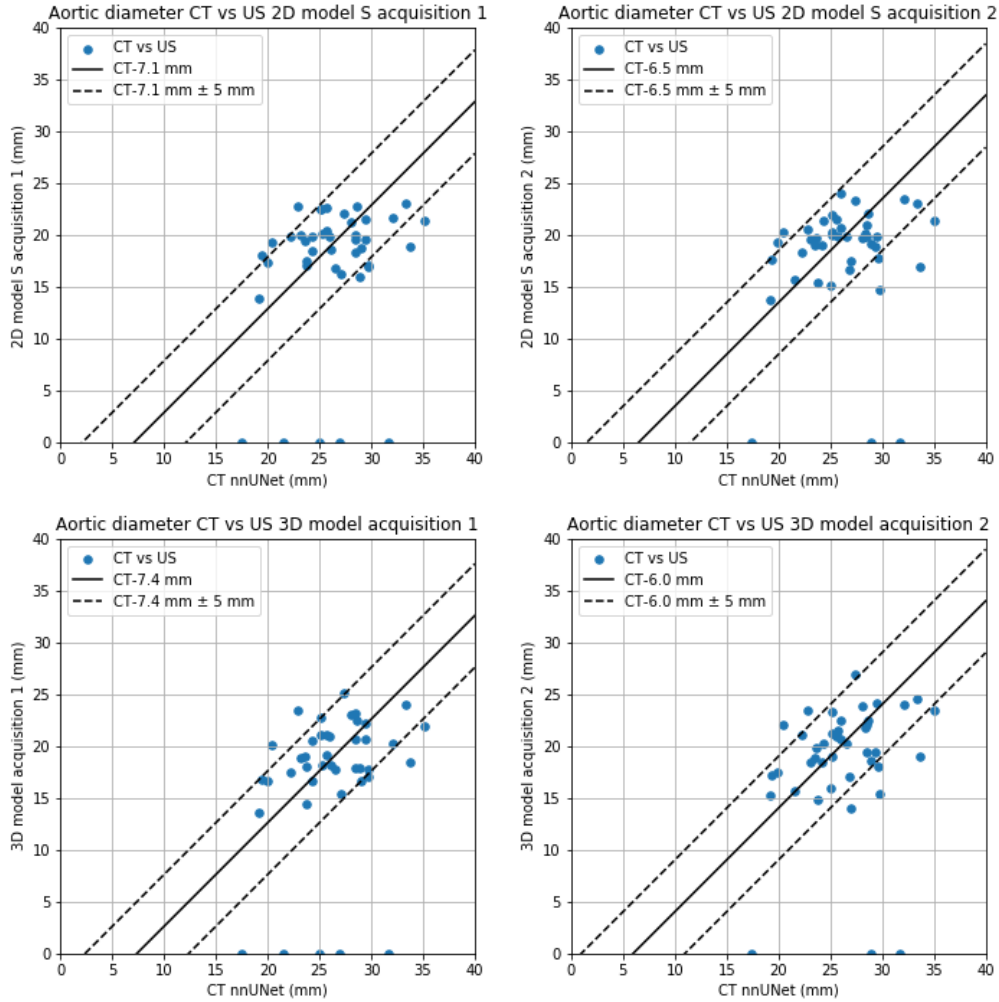


Figure 7: Scatter plots of the US diameter against the CT diameter for acquisition 1 and 2. The CT diameters were determined from the nnUNet segmentations and the US diameters were determined from the 2D model^S and the 3D model segmentations. The dashed lines shows the limits of agreement of ± 5 mm.

The percentages of the US-CT differences that are within the clinically accepted limits of agreement of ± 5 mm are reported (Table 6).

Table 6: The percentages of CT-US differences that are between the limits of agreement of ± 5 mm.

	Acquisition 1	Acquisition 2
2D model ^S	69%	78.6%
3D model	71.4%	73.8%

5 Discussion

In this study, we presented a deep-learning method to automatically detect and measure the abdominal aortic diameter from US imaging using the MicrUs Pro C60S. For dataset 1, data of 100 patients were collected. These data were used to train deep-learning models with varying downsampling factors and model channels. The model with the highest validation Dice was additionally trained on 3D input. This 3D model performed significantly better compared to all other models on the test set and showed a median Dice of 0.88 (IQR = 0.78 - 0.92). The 3D model could not run real-time on a smartphone. The best performing 2D model that could also run real-time on a smartphone (2D model^S) had a median Dice of 0.83 (IQR = 0.71 - 0.89) on the test set. 2D model^S was used to collect a second dataset of 44 patients. During the acquisition of this data the 2D model^S was included on a smartphone for real-time inference. The CT-US maximum diameter differences from the first acquisition and second acquisition were compared. The differences obtained with the 3D model on acquisition 2 are significantly smaller than all other differences. This model had a median difference of 6.0 mm (IQR = 4.0 - 9.6 mm), with 73.8% of cases falling within the clinically acceptable limits of agreement of ± 5 mm. compared to CT.

In the first experiment the performance of the Dice and false positive pixels metric investigated under different hyperparameter combinations of the downsampling factor and model channels. It was found that the models that were trained were most stable at 8 model channels. This means that they always trained and converged to a solution for each of the four validation folds. It is possible that using more than 8 model channels creates an overparameterized model in our case, because we had limited data to train on (60% of 549 annotated frames). The same was observed by a similar re-

search that performed placenta localization from US with limited amounts of data [32]. The only model that was stable across all settings for model channels, was the model trained on images with a downsampling factor of 2. We hypothesize that downsampling, at least up to a factor of 2, also acts as a regularizer. For the downsampling 2 model, the Dice scores were the highest and the number of false positive pixels were lower than all models with a higher downsampling factor. Specifically, we chose the downsampling 2 model with 32 model channels as our best model for inference. A downsampling factor of 2 achieved the highest Dice on the validation set and 32 model channels produced the most spherical segmentations. The model with downsampling 4 and model channels 8 was chosen to be used for real-time inference on the smartphone, since its Dice performance was highest for all models that could be run in real-time on a smartphone.

The best performing model on the test set was the 3D Model with a median dice of 0.88 (IQR = 0.78 - 0.92). To assess the performance of negative frames, we looked at the percentage of negative frames with no segmentation. The 3D model was again the best performing model with 82% on the test set. We also found that the 3D model is significantly better compared to both 2D models in terms of the Dice and false positive pixels metric. Additionally, the 2D model^B outperforms the model^S, which was to be expected because of the lower downsampling factor. This means that there is added benefit to giving 3D information to the U-Net as input. It corresponds to how humans interpret an US scan, and helps to model to interpolate between frames. An example of this can be seen in a case where the 3D model does not make false positive segmentations in regions where there is no aorta, as opposed to the 2D models (Figure 4). An additional possibility is to use volumetric segmentation such as V-Net [26] to take into account 3D information and also use 3D convolutions. This model could be computationally more expensive but it might be possible to run this on a smartphone in the future. If a V-Net model turns out to be more robust for this problem, it could be used on a computer for off-line inference.

The second part of this work investigated whether the scan quality would be improved when we were guided by the deep learning model, to provide real-time inference while scanning a patient. Our quantitative measure to assess the scan quality was the percentage of frames where there was no segmentation done and no diameter measured. It was found that there was no significant difference between the first acquisition of dataset 1 and dataset 2. For the other models there is a trend towards significance, so it could be

that the number of test cases (20) is too low. It is debatable whether the percentage of frames with a diameter is a good measure for determining scan quality. During scanning, it is possible that the transducer is held still at the start of the scan and also at the end. This can result in a lower percentage. It should be determined what the minimum percentage of frames with a diameter is for aneurysm cases with a sufficient quality.

The results show that the maximum aortic diameter CT-US differences from acquisition 2 obtained with the 3D model, are significantly smaller compared to the differences for acquisition 1. The differences on acquisition 2 from the 3D model have a median of 6.0 mm (IQR = 4.0 - 9.6 mm). The CT-US differences have a positive median for each model and acquisition, meaning that the CT diameters are larger than the US diameters. This is consistent with the observations from the literature. The results demonstrate that an acquisition with guidance from the deep-learning system does improve the scan quality. An example of this are three cases for which US segmentation failed in acquisition 1, but not in acquisition 2 (Figure 7). 73.8% of the CT-US differences obtained from the 3D model on acquisition 2 are within the clinically acceptable limits of agreement of ± 5 mm. Jaakkola et al. [11] compared the aortic diameter measurements made by a radiologist on CT and ultrasound. They showed that 83% of these measurements have a difference of ± 5 mm. This shows that the method we use to obtain an automatic diameter measurement from US is almost comparable to that of a radiologist. However, we have to consider that a large portion of the US scans were not of sufficient quality. This work therefore shows feasibility to automatically measure the aortic diameter, but future research on a larger population, which includes larger aneurysms, is required to validate our approach.

Future research directions could focus on improving the model’s robustness or on obtaining US scans of better quality. To improve the model’s robustness it is possible to use data augmentation. B-mode images could be flipped horizontally and rotated. Another possible augmentation is to try out changing the contrast of the B-mode images. From a theoretical perspective, it could be interesting to adapt the loss function to the specific problem of abdominal aorta detection, similar to how humans would detect it. The loss function could include the sphericity of the segmentation and the criterion that only a single component is segmented. This first criterion could help because the aorta always has the shape of a circle or an ellipse. The second criterion ensures that there can only be one segmented structure. The model

should for instance not segmented both the inferior vena cava (IVC) and the aorta, or various abdominal arteries. These two criteria could be included in a weighted loss function, along with the Dice loss term. Finally, since we have now obtained a working model, it can be easily applied to infer segmentations new on scans, which could then be used as annotations for new training data. This is a quick way to gather more training data and to make the model more robust.

A limitation that emerged from this work was the difficulty to obtain US scans of sufficient quality. In the future it is important to find a solution to this problem. A scan is of sufficient quality if the whole aorta is visible during the scan and scan be segmented after recording. If parts of the aorta are missing, it cannot be ruled out that an aneurysm was present but was not detected. Specifically shadowing in abdominal US is hard to avoid due to collection of bowel gas. The aim of the application is that untrained people can measure the aortic diameter without extensive training. The quality of scans recorded with the MicrUs Pro C60S by trained sonographers could be compared to the quality of scans by a layman. If the quality of scans is sufficient for trained sonographers, their input could be used to design a protocol that reduces the current quality problems. There is also a bias in our study population, because they all received an CT scan. A study cohort in AAA patients and the general population is required to investigate if the quality of scans in this population is the same.

A follow-up to this work is to include training data for aneurysm cases. In the current work we have trained a model on healthy aortas, but the ultimate goal would be to monitor the aortic diameter of patients with an aneurysm. The current model could be used as a pre-trained network to train aneurysm cases, or the annotations for aneurysms could be combined with the annotations for healthy aortas.

6 Conclusion

In this study, we show a deep-learning method to automatically detect and measure the aortic diameter from US imaging. Data was collected from 144 patients. Data from 100 of those patients was used to create a deep-learning model. The best performing model achieved a median Dice of 0.88 on the test set. This demonstrates the feasibility of the deep-learning approach for automated aortic diameter detection. We investigated whether a layman

could obtain US scans of better quality when guided with real-time inference from a deep learning system on a smartphone. The laymen acquired US data from 44 patients which also received a CT scan. The best performing 3D model showed a median difference of 6.0 mm (IQR = 4.0 - 9.6) between the US and CT measurement. 73.8% of all cases fell within the clinically acceptable limits of agreement of ± 5 mm. This approach shows promising results for automated aortic diameter measurement for laymen. A future study should investigate this approach on aneurysm cases.

References

- [1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016.
- [2] L. R. Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945.
- [3] M. Ellis, J. Powell, and R. Greenhalgh. Limitations of ultrasonography in surveillance of small abdominal aortic aneurysms. *Journal of British Surgery*, 78(5):614–616, 1991.
- [4] R. C. Gonzalez, R. E. Woods, et al. Digital image processing, 2002.
- [5] I. Goodfellow, Y. Bengio, and A. Courville. *Deep learning*. MIT press, 2016.
- [6] R. Halir and J. Flusser. Numerically stable direct least squares fitting of ellipses. In *Proc. 6th International Conference in Central Europe on Computer Graphics and Visualization. WSCG*, volume 98, pages 125–132. Citeseer, 1998.
- [7] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, 2015.
- [8] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

- [9] L. Hoornweg, M. Storm-Versloot, D. Ubbink, M. Koelemay, D. Legemate, and R. Balm. Meta analysis on mortality of ruptured abdominal aortic aneurysms. *European Journal of Vascular and Endovascular Surgery*, 35(5):558–570, 2008.
- [10] F. Isensee, J. Petersen, A. Klein, D. Zimmerer, P. F. Jaeger, S. Kohl, J. Wasserthal, G. Koehler, T. Norajitra, S. Wirkert, et al. nnu-net: Self-adapting framework for u-net-based medical image segmentation. *arXiv preprint arXiv:1809.10486*, 2018.
- [11] P. Jaakkola, M. Hippeläinen, P. Farin, H. Rytönen, S. Kainulainen, and K. Partanen. Interobserver variability in measuring the dimensions of the abdominal aorta: comparison of ultrasound and computed tomography. *European journal of vascular and endovascular surgery*, 12(2):230–237, 1996.
- [12] N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.
- [13] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [14] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.
- [15] B. Landman, Z. Xu, J. E. Igelsias, M. Styner, T. Langerak, and A. Klein. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In *Proc. MICCAI: Multi-Atlas Labeling Beyond Cranial Vault-Workshop Challenge*, 2015.
- [16] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- [17] F. A. Lederle, G. R. Johnson, S. E. Wilson, D. J. Ballard, W. D. Jordan Jr, J. Blebea, F. N. Littooy, J. A. Freischlag, D. Bandyk, J. H. Rapp, et al. Rupture rate of large abdominal aortic aneurysms in patients refusing or unfit for elective repair. *Jama*, 287(22):2968–2972, 2002.

- [18] F. A. Lederle, S. E. Wilson, G. R. Johnson, D. B. Reinke, F. N. Littooy, C. W. Acher, L. M. Messina, D. J. Ballard, H. J. Ansel, C. S. P. C. Center, et al. Variability in measurement of abdominal aortic aneurysms. *Journal of vascular surgery*, 21(6):945–952, 1995.
- [19] X. Li, G. Zhao, J. Zhang, Z. Duan, and S. Xin. Prevalence and trends of the abdominal aortic aneurysms epidemic in general population—a meta-analysis. *PloS one*, 8(12):e81260, 2013.
- [20] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.
- [21] S. Liu, Y. Wang, X. Yang, B. Lei, L. Liu, S. X. Li, D. Ni, and T. Wang. Deep learning in medical ultrasound analysis: a review. *Engineering*, 5(2):261–275, 2019.
- [22] A. Long, L. Rouet, A. Debreuve, R. Ardon, C. Barbe, J. Becquemin, and E. Allaire. Abdominal aortic aneurysm imaging with 3-d ultrasound: 3-d-based maximum diameter measurement and volume quantification. *Ultrasound in medicine & biology*, 39(8):1325–1336, 2013.
- [23] A. Long, L. Rouet, J. S. Lindholt, and E. Allaire. Measuring the maximum diameter of native abdominal aortic aneurysms: review and critical analysis. *European Journal of Vascular and Endovascular Surgery*, 43(5):515–524, 2012.
- [24] R. Lopata, S. Meesters, V. Nguyen, G. Schurink, and F. Van De Vosse. Automated 2d ultrasound fusion imaging of abdominal aortic aneurysms. In *2012 IEEE International Ultrasonics Symposium*, pages 354–357. IEEE, 2012.
- [25] B. J. Manning, T. Kristmundsson, B. Sonesson, and T. Resch. Abdominal aortic aneurysm diameter: a comparison of ultrasound measurements with those from standard and three-dimensional computed tomography reconstruction. *Journal of vascular surgery*, 50(2):263–268, 2009.
- [26] F. Milletari, N. Navab, and S.-A. Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016*

- fourth international conference on 3D vision (3DV)*, pages 565–571. IEEE, 2016.
- [27] C. L. Moore and J. A. Copel. Point-of-care ultrasonography. *New England Journal of Medicine*, 364(8):749–757, 2011.
- [28] J. J. Reimerink, M. J. van der Laan, M. J. Koelemay, R. Balm, and D. A. Legemate. Systematic review and meta-analysis of population-based mortality from ruptured abdominal aortic aneurysm. *Journal of British Surgery*, 100(11):1405–1413, 2013.
- [29] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.
- [30] L. Rouet, R. Ardon, J.-M. Rouet, B. Mory, C. Dufour, and A. Long. Semi-automatic abdominal aortic aneurysms geometry assessment based on 3d ultrasound. In *2010 IEEE International Ultrasonics Symposium*, pages 201–204. IEEE, 2010.
- [31] U. K. Sampson, P. E. Norman, F. G. R. Fowkes, V. Aboyans, Y. Song, F. E. Harrell Jr, M. H. Forouzanfar, M. Naghavi, J. O. Denenberg, M. M. McDermott, et al. Estimation of global and regional incidence and prevalence of abdominal aortic aneurysms 1990 to 2010. *Global heart*, 9(1):159–170, 2014.
- [32] M. Schipzand. Automatic placenta localization from ultrasound imaging in a resource-limited setting using a predefined ultrasound acquisition protocol and deep learning. 2020.
- [33] S. S. Shapiro and M. B. Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611, 1965.
- [34] N. Siddique, S. Paheding, C. P. Elkin, and V. Devabhaktuni. U-net and its variants for medical image segmentation: A review of theory and applications. *IEEE Access*, 2021.
- [35] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

- [36] K. Singh, K. Bønaa, S. Solberg, D. Sørli, and L. Bjørk. Intra-and interobserver variability in ultrasound measurements of abdominal aortic diameter. the tromsø study. *European journal of vascular and endovascular surgery*, 15(6):497–504, 1998.
- [37] T. A. Sorensen. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on danish commons. *Biol. Skar.*, 5:1–34, 1948.
- [38] L. R. Sprouse II, G. H. Meier III, C. J. LeSar, R. J. DeMasi, J. Sood, F. N. Parent, M. J. Marcinyck, and R. G. Gayle. Comparison of abdominal aortic aneurysm diameter measurements obtained with ultrasound and computed tomography: is there a difference? *Journal of vascular surgery*, 38(3):466–471, 2003.
- [39] Student. The probable error of a mean. *Biometrika*, pages 1–25, 1908.
- [40] T. van den Heuvel. *Automated low-cost ultrasound: improving antenatal care in resource-limited settings*. PhD thesis, [Sl: sn], 2019.
- [41] T. L. A. Van den Heuvel, H. Petros, S. Santini, C. L. de Korte, and B. van Ginneken. Automated fetal head detection and circumference estimation from free-hand ultrasound sweeps using deep learning in resource-limited countries. *Ultrasound Med Biol*, 45(3):773–785, 3 2019.
- [42] E. van Disseldorp, J. van Dronkelaar, J. Pluim, F. van de Vosse, M. van Sambeek, and R. Lopata. Automatic segmentation and registration of abdominal aortic aneurysms using 3d ultrasound. In *2016 IEEE International Ultrasonics Symposium (IUS)*, pages 1–4. IEEE, 2016.
- [43] E. M. van Disseldorp, J. J. van Dronkelaar, J. P. Pluim, F. N. van de Vosse, M. R. van Sambeek, and R. G. Lopata. Ultrasound based wall stress analysis of abdominal aortic aneurysms using multiperspective imaging. *European Journal of Vascular and Endovascular Surgery*, 59(1):81–91, 2020.
- [44] J. A. van Essen, E. J. Gussenhoven, J. D. Blankensteijn, J. Honkoop, L. C. van Dijk, M. R. van Sambeek, and A. van der Lugt. Three-dimensional intravascular ultrasound assessment of abdominal aortic aneurysm necks. *Journal of Endovascular Therapy*, 7(5):380–388, 2000.

- [45] K. G. van Leeuwen, S. Schalekamp, M. J. Rutten, B. van Ginneken, and M. de Rooij. Artificial intelligence in radiology: 100 commercially available products and their scientific evidence. *European radiology*, 31(6):3797–3804, 2021.
- [46] D. Vorp, W. Mandarino, M. Webster, and J. Gorcsan III. Potential influence of intraluminal thrombus on abdominal aortic aneurysm as assessed by a new non-invasive method. *Cardiovascular Surgery*, 4(6):732–739, 1996.
- [47] A. Wanhainen, F. Verzini, I. Van Herzele, E. Allaire, M. Bown, T. Cohnert, F. Dick, J. van Herwaarden, C. Karkos, M. Koelemay, et al. Editor’s choice–european society for vascular surgery (esvs) 2019 clinical practice guidelines on the management of abdominal aorto-iliac artery aneurysms. *European Journal of Vascular and Endovascular Surgery*, 57(1):8–93, 2019.
- [48] F. Wilcoxon. Individual comparisons by ranking methods. In *Break-throughs in statistics*, pages 196–202. Springer, 1992.

A Appendix

A.1 Inter-observer variability

Two observers measured the maximum anterior–posterior (AP) diameter of the aorta for each of the 100 scans. One observer was the author of this thesis, who was trained by a researcher with six years of ultrasound experience. The AP diameter is a standard metric for measuring the aorta diameter, and has been shown to have a better reproducibility over the transverse diameter [3]. To measure the aorta we used the outer-to-outer (OTO) method (Figure 2). These measurements were compared to examine the inter-observer variability. Measurements were made by using grand-challenge.org [45].

Ultrasound is highly dependent upon the operator. To assess the inter-observer variability in the scans, the diameter measurements from both observers were compared. The clinically acceptable limits of agreement between aortic diameter ultrasound measurements are ± 5 mm [23]. This means that the absolute mean difference between measurements is < 5 mm for 95% of

the measurements. Inter-observer differences of maximum AP diameter measurements by trained sonographers range from 2 mm or less in 75% of cases, to 4 mm or less in 96% of cases [36].

The differences are shown in a Bland-Altman plot (Figure 8). The observers were able to find and measure a diameter in 62 out of 100 cases. The absolute mean difference is 1.51. In total 75.8% (47 out of 62) measurements differ 2 mm or less, and 91.9% (57 out of 62) of measurements differ 4 mm or less. Only 2 cases were outside of the clinical limits of agreement of ± 5 mm. This is comparable to what trained sonographers achieve. In conclusion, the reproducibility of aortic measurements from ultrasound images in the axial plane, recorded with the MicrUs Pro C60S, seems to be clinically acceptable.

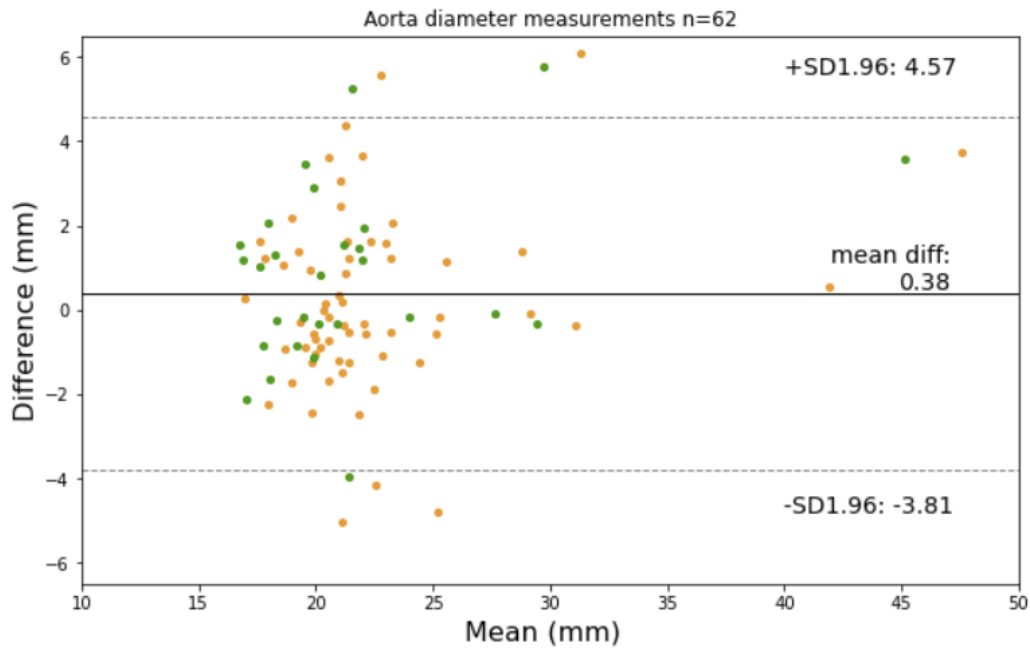


Figure 8: Bland-Altman plot of the sufficient cases (green points) and insufficient cases (orange points). The mean aortic diameter is shown on the x-axis. The inter-observer difference is shown on the y-axis. \pm SD1.96 shows the 1.96 times the standard deviation from the mean difference.

A.2 Manual quality evaluation

For dataset 1, it was manually assessed whether the quality of an US scan was sufficient or insufficient. Rater 1 was a researcher with six years of ultrasound experience, and rater 2 was the author of this report. A scan was labeled as insufficient when the course of the aorta was partly missing, and when we deemed it possible that an aorta with a larger diameter may have been present in the missing region. There are multiple factors that can cause the image of the aorta to be missing, such as obesity, excess bowel gas or air, cysts in the abdomen, or bad contact between the transducer and skin. The number of patients with scans that both raters assessed as sufficient was 28% (28 out of 100) (Table 7).

Table 7: Number of patients with sufficient and insufficient scan quality, assessed by rater 1 and rater 2.

		Rater 1	
		Sufficient	Insufficient
Rater 2	Sufficient	28	18
	Insufficient	2	52

A.3 Batch size experiment

In a similar research that performed automatic placenta localization from ultrasound imaging with deep learning [32], it was found that small batch sizes performs better for downsampling factors 4, 6 and 8. Therefore, for the smartphone model (downsampling 4, model channels 8), further optimization experiments with smaller batch sizes (4, 8, 16, 32 and 64) were carried out to see if this leads to a better performance. For smaller batch sizes there is indeed an increase in the Dice score (Figure 9). However, this also seems to have the trade-off that the average number of false positive pixels on negative frames becomes higher.

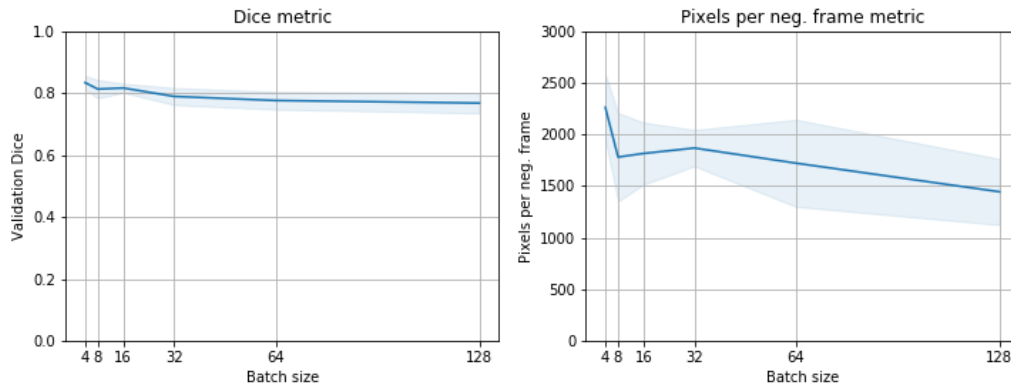


Figure 9: The mean Dice and false positive pixels metric for different batch sizes for the smartphone model. At each batch size the mean \pm SD of all cross-folds is shown.

A qualitative evaluation shows that a model with a batch size of 16 tends to segment slightly more and shows a better generalization ability in comparison to a model with a batch size of 128. This can be observed in an example patient where the top of the first component (around frame 25) and bottom of the second component (around frame 125) are more fully segmented with the batch size 16 model (Figure 10). Overall, this results in a higher Dice score.

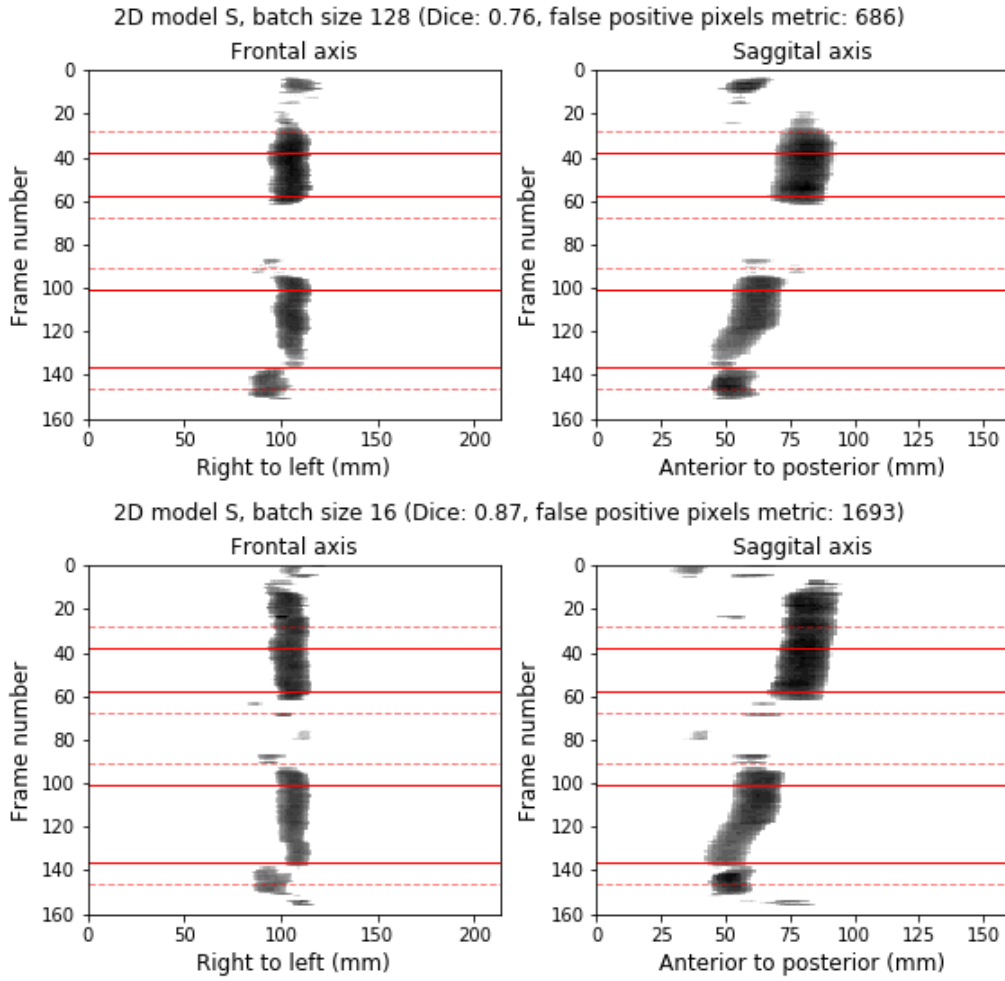


Figure 10: 2D plots showing a case with a more complete segmentation for the batch size 16 model. Top row: batch size 128 model. Bottom row: batch size 16 model.

In contrast, the model with batch size 16 performed worse on the false positive pixels metric. This means that it would more easily give false positive segmentations. For instance in a case with insufficient quality, the batch size 128 model segments an average of 793 pixels per negative, while the batch size 16 model segments 1927 (Figure 11). For this reason, the model was not chosen to be implemented on the smartphone, but the model with the lowest number of false positive pixels (batch size 128) was chosen instead.

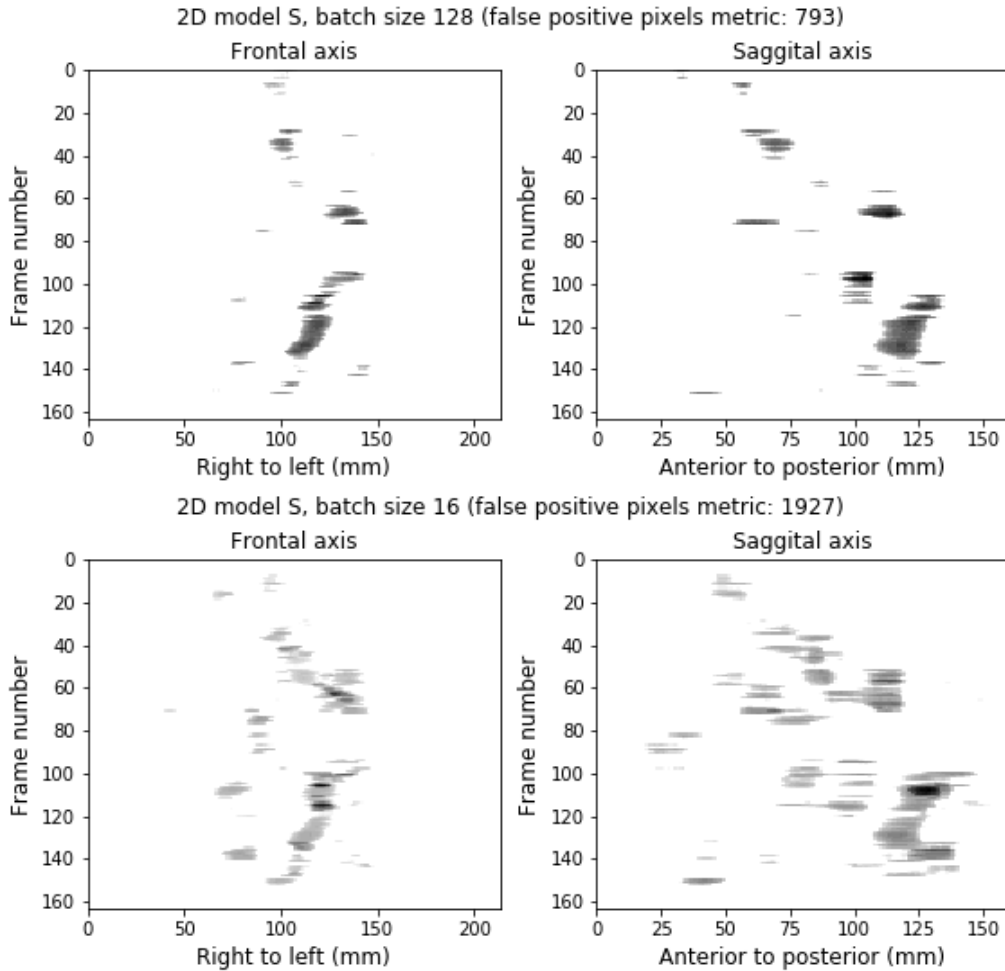


Figure 11: 2D plots for an insufficient case. Top row: batch size 128 model. Bottom row: batch size 16 model.

A possible explanation for the increased generalization ability for smaller batch size is because it leads convergence to *flat minimizers*, while using larger batch sizes leads converge to *sharp minimizers* [12]. Flat minimizers can be described with lower precision and have a better generalization performance. An experiment for future research could be compare the sharpness of minima for a small batch-size regime and a large batch-size regime, to validate if this phenomenon also occurs in our problem setting.