Ultrasound in Medicine & Biology

Automatic extraction of hiatal dimensions in 3D transperineal pelvic ultrasound recordings

--Manuscript Draft--

Manuscript Number:	UMB-D-21-00110R2
Article Type:	Original Contribution
Keywords:	Ultrasound; levator hiatus; transperineal ultrasound; segmentation; deep learning; automatic clinical workflow
Corresponding Author:	Helena williams leuven, Europe BELGIUM
First Author:	Helena williams
Order of Authors:	Helena williams
	Laura Cattani
	Dominique Van Schoubroeck
	Mohammad Yaqub
	Carole Sudre
	Tom Vercauteren
	Jan D’hooge
	Jan Deprest
Abstract:	The objective of this work was to create a robust automatic software tool for measurement of the levator hiatal area on Transperineal ultrasound (TPUS) volumes, and to measure the potential reduction in error and time taken for analysis in a clinical setting. The proposed tool automatically detects the C-plane (i.e. the plane of minimal hiatal dimensions) from a 3D transperineal UltraSound (US) volume and subsequently uses the extracted plane to automatically segment the levator hiatus, using a convolutional neural network (CNN). The automatic pipeline was tested using 73 representative TPUS volumes. Reference hiatal outlines were obtained manually by two experts and compared with the pipeline’s automated outlines. The Hausdorff distance, area, a clinical quality score, C-plane angle, and the C-plane Euclidean distance were used to evaluate C-plane detection and quantify levator hiatus segmentation accuracy. A visual Turing Test was created to compare the performance of the software to the expert, based on the visual assessment of C-plane and hiatal segmentation quality. The overall time taken to extract the hiatal area with both measurement methods (i.e. manual and automatic) was measured. Each metric was calculated both for computer-observer differences, and for inter-and intra-observer differences. The automatic method gave similar results to the expert when determining the hiatal outline from a TPUS volume. Indeed, the hiatal area measured by the algorithm and by an expert were within the intra-observer variability. Similarly, the method identified the C-plane with an accuracy of 5.76 ± 5.06 ° and 6.46 ± 5.18 mm in comparison to the inter-observer variability of 9.39 ± 6.21° and 8.48 ± 6.62 mm. The visual Turing Test suggested that the automatic method identified the C-plane position within the TPUS volume visually as well as the expert. The average time taken to identify the C-plane and segment the hiatal area manually was 2 minutes and 35 ± 17 seconds, compared to 35 ± 4 seconds for the automatic result. This study presents a method for automatically measuring the levator hiatal area using AI-based methodologies whereby the C-plane within a TPUS volume is detected and subsequently traced for the levator hiatal outline. The proposed solution was demonstrated to be accurate, relatively quick, robust and reliable, and – importantly - to reduce time and expertise required for pelvic floor disorder assessment.
Suggested Reviewers:	Hans Peter Dietz hans.dietz@sydney.edu.au Expert in pelvic floor disorder assessment

Opposed Reviewers:	

Cover letter

Helena Williams
KU Leuven, Belgium
Helena.williams@kuleuven.be

2nd February 2021

Dear Editor in Chief of Ultrasound in medicine and biology,

We wish to submit an original research article entitled "Automatic extraction of hiatal dimensions in 3D transperineal pelvic ultrasound recordings" for consideration of publication in your journal. We confirm that this work is original and has not been published elsewhere, nor is it currently under consideration for publication elsewhere.

In this paper, we present a novel, automatic software tool able to identify the plane of minimal hiatal dimensions and further delineate the levator hiatus from a 3D Transperineal Ultrasound volume. The assessment of the levator hiatus area is helpful for tailoring treatment of patients with pelvic organ prolapse. We feel this work is significant because currently, within clinic, it is a manual process requiring a high level of training and expertise. Furthermore, the current clinical workflow is time-consuming, labour-intensive and prone to inter-observer error. Experts may have different techniques, in locating the minimal hiatal dimensions and delineating the levator hiatus, therefore, automation may standardise the procedure.

In this work we perform extensive validation on a challenging clinical dataset (with a high proportion of pathology cases). We show that our proposed method is clinically acceptable. The proposed tool performs with a higher accuracy than the recorded inter-observer error, thus it will reduce error in the clinical setting. The proposed pipeline is roughly 2 minutes quicker than an expert, meaning it may save clinicians' time to spend on patient counselling and treatment planning. The proposed pipeline also lowers the expertise required to perform Transperineal ultrasound imaging.

We believe that this manuscript is appropriate for publication by Ultrasound in medicine and biology because it is clinically relevant research within the field of ultrasound in medicine and pelvic floor disorder assessment. Furthermore, this work is novel, intuitive and performs well with a small training dataset- rare for deep learning applications, thus it could be applied to other clinical applications where plane selection or landmark detection is required for medical imaging analysis.

Please address all correspondence concerning this manuscript to me at helena.williams@kuleuven.be.

Thank you for your consideration of this manuscript.

Sincerely,

Helena Williams

Reply to Editor and Reviewers

August 4, 2021

These parts are annotated with $[R_i C_j]$ referring to the comment j associated with reviewer i.

1 Manuscript Modifications

There were no main modifications which were asked of by the reviewers in the second reading.

We thank the reviewer's again for their careful reading and feedback, it has greatly improved the clarity of the paper.

2 Answer to Reviewer 1's second review

C$[R_1 C_1]$**:** *The authors have made all the changes I requested making the paper clearer.*

R: Thank you for your comments and feedback.

3 Answer to Reviewer 2's second review

C$[R_2 C_1]$**:** *Thank you for revising this manuscript. The additions and rephrasing of sections of the manuscript make it stronger, logical and coherent. The work adds to the evidence base for the use of deep learning for patients with suspected pelvic disease.*

Well done.

A minor change is to use third person rather than 'herself'

I look forward to seeing your work published and reading future work of yours on this topic.

R: Thank you for your kind comments and feedback. The paper has been changed to 'themself' instead of 'herself'. P15 L318.

Automatic extraction of hiatal dimensions in 3D transperineal pelvic ultrasound recordings

Helena Williams[1,2,4], Laura Cattani[1], Dominique Van Schoubroeck[1], Mohammad Yaqub[3], Carole Sudre[2], Tom Vercauteren[2], Jan D'hooge[4] and Jan Deprest[1]

1 Department of Obstetrics and Gynaecology, University Hospitals Leuven, Belgium
2 School of Biomedical Engineering & Imaging Sciences, King's College London, UK
3 Department of Computer Vision, Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, United Arab Emirates
4 Department of Cardiovascular Sciences, KU Leuven, Belgium

Corresponding Author: Helena Williams
Institute Department of Obstetrics and Gynaecology, University Hospitals Leuven, 49 Herestraat, 3000, Leuven, Belgium

Email: Helena.williams@kuleuven.be
Phone: +447949421813
Running Title: Automatic extraction of hiatal dimensions

Abstract

The aim of this work was to create a robust automatic software tool for measurement of the levator hiatal area on Transperineal ultrasound (TPUS) volumes, and to measure the potential reduction in variability and time taken for analysis in a clinical setting. The proposed tool automatically detects the C-plane (i.e. the plane of minimal hiatal dimensions) from a three-dimensional (3D) transperineal UltraSound (US) volume and subsequently uses the extracted plane to automatically segment the levator hiatus, using a convolutional neural network (CNN). The automatic pipeline was tested using 73 representative TPUS volumes. Reference hiatal outlines were obtained manually by two experts and compared with the pipeline's automated outlines. The Hausdorff distance, area, a clinical quality score, C-plane angle, and the C-plane Euclidean distance were used to evaluate C-plane detection and quantify levator hiatus segmentation accuracy. A visual Turing Test was created to compare the performance of the software to the expert, based on the visual assessment of C-plane and hiatal segmentation quality. The overall time taken to extract the hiatal area with both measurement methods (i.e. manual and automatic) was measured. Each metric was calculated both for computer-observer differences, and for inter-and intra-observer differences. The automatic method gave similar results to the expert when determining the hiatal outline from a TPUS volume. Indeed, the hiatal area measured by the algorithm and by an expert were within the intra-observer variability. Similarly, the method identified the C-plane with an accuracy of 5.76 ± 5.06 ° and 6.46 ± 5.18 mm in comparison to the inter-observer variability of 9.39 ± 6.21° and 8.48 ± 6.62 mm. The visual Turing Test suggested that the automatic method identified the C-plane position within the TPUS volume visually as well as the expert. The average time taken to identify the C-plane and segment the hiatal area manually was 2 minutes and 35 ± 17 seconds, compared to 35 ± 4 seconds for the automatic result. This study presents a method for automatically measuring the levator hiatal area using AI-based methodologies whereby the C-plane within a TPUS volume is detected and subsequently traced for the levator hiatal outline. The proposed solution was demonstrated to be accurate, relatively quick, robust and reliable, and – importantly - to reduce time and expertise required for pelvic floor disorder assessment.

Keywords: ultrasound, levator hiatus, transperineal ultrasound, segmentation, deep learning, automatic clinical workflow

Introduction

Pelvic Floor Ultrasound examination (PFUS) is increasingly being used in the assessment of the pelvic floor anatomy in women with pelvic floor dysfunction (IUGA, 2019). Typically, an abdominal 3D transducer is placed on the labia to assess the urogenital organs, the levator ani muscle (LAM), and where indicated, the anal sphincter. The LAM is a broad muscular sheet attached to the internal surface of the pelvis and supports the urogenital organs and ano-rectum(Schwertner-Tiepelmann, et al. 2012). Levator integrity and the hiatal area assessment for organ descent is helpful when counselling and tailoring treatment of patients with pelvic organ prolapse. Levator avulsion enlarges the genital hiatus (Abdool, et al. 2009) and is associated with anterior and middle compartment prolapse, as well as recurrence of prolapse after native tissue repair, hence can be considered a biomarker to assess pelvic floor dysfunction(Dietz and Simpson 2008, Ismail, et al. 2016). Additionally, delivery-induced sarcomeric hyperelongation may cause substantial, irreversible ultrastructural trauma in the LAM(Brooks, et al. 1995, Lien, et al. 2004). Irreversible over-distention of the levator hiatus ('microtrauma') has been described in postpartum women as a possible consequence of muscular atrophy, reduction in function and can alter pelvic floor distensibility after vaginal delivery (Shek and Dietz 2010).

Manual detection of the levator hiatus in a 3D transperineal ultrasound (TPUS) acquisition requires significant offline post-processing of the volumetric recordings by specifically trained sonographers. UltraSound (US) manufacturers have implemented and previous works (Li, et al. 2019, Sindhwani, et al. 2016, van den Noort, et al. 2019) have developed semi-automatic and automatic tools to aid PFUS. For instance, real-time visualisation of the desirable C-plane from a manually identified approximation of the C-plane was developed in Omniview-VCI (GE Healthcare, Austria). Clinicians assess the levator hiatus on the plane where the anteroposterior distance (between the dorsocaudal end of the symphysis pubis (SP), and the ventral end of the levator), is the smallest, and refer to this plane as the plane of minimal hiatal dimensions (MHD) or C-plane. However, a fully automatic levator hiatus detection from a TPUS volume should obtain a more accurate representation of the anatomical findings, would be less operator-dependent, and may save clinicians time to allow more focus on patient care and counselling. Automation would lower the minimal threshold of expertise for clinicians to be using TPUS. This study aimed to build a fully automatic workflow that consists of C-plane detection followed by hiatal segmentation. A solution to this clinical problem (IUGA, 2019) which ensures the trustworthiness and interpretability from experts while following the clinical guidelines is likely to have a strong clinical value.

Material and Methods

Abbreviations

In order to make tables and figures more readable, abbreviations have been introduced throughout this paper. A table of abbreviations and definitions can be seen in table S1, in the Supplementary material.

Manual C-plane detection

3D-TPUS acquisition is performed orienting the 3D abdominal probe as on conventional transvaginal ultrasound images (cranioventral aspects to the left, dorsocaudal to the right)

(Dietz 2010). The so-acquired 3D-image of the pelvic floor shows the midsagittal plane in the top left corner (A), the axial plane in the top right corner (B) and the coronal plane in the bottom left corner (C) (Figure 1). In order to visualise the C-plane on the coronal plane of the US image, clinicians manually align the SP and the LAM to a horizontal direction on the midsagittal plane. Eventually, the LAM lies on the axial plane, as shown in Figure 1. This makes the levator hiatus clearly visible on the coronal plane as the pubic bones ventrally, and the LAM dorsally is hyper-echogenic compared with the hypo-echogenic pelvic organs. By analysing the levator hiatus, one can diagnose levator avulsion and hiatal ballooning (IUGA,2019).

Proposed biomarker extraction pipeline

The proposed automatic data analysis pipeline is composed of two sequential parts: a C-plane extractor, and a levator hiatus outline extractor (Figure 2). The C-plane extractor is based on our previous work, see (Williams, et al. 2020) for more in-depth technical details. The proposed pipeline expands on this work to automatically outline the levator hiatus from the C-plane extractor's output. The proposed pipeline utilises advances in CNNs, landmark detection, semantic segmentation and follows the IUGA/AIUM (IUGA, 2019) clinical guidelines to ensure interpretability of the results. The solution requires no user input and is thus completely automatic. In brief, the pipeline starts by automatically detecting the SP and LAM extreme coordinates within a TPUS via CNN landmark regression. The extreme coordinates are defined as the voxel coordinates with the shortest Euclidean distance between the 3D segmentations of the SP and LAM within a Mid-Sagittal (MS) slice, as shown in Figure 3. Post-processing identifies the vector of MHD, and a transformation matrix can be formed to resample the TPUS volume as the desired 2D C-plane. The extracted C-plane is then used as input to a pre-trained 2D semantic segmentation CNN model that segments the levator hiatus, defining the hiatal area.

Description of the biomarker pipeline

i) 3D landmark regression of the SP and LAM extreme coordinates

The first step of the C-plane extractor accepts a TPUS volume as input and results in a heatmap of the SP and LAM extreme coordinates within the TPUS volume. The heatmap is a data visualisation technique which encodes the probability of a landmark being located at a certain voxel position within the TPUS volume. In this study, the heatmap voxels near the extreme coordinate have high values (with the highest at the extreme coordinate), and they smoothly and rapidly decrease with increasing distance from the extreme coordinate, as shown in Figure 3.

The rationale behind this approach was that regressing one coordinate from a large volume can be difficult to train, and a heatmap is more robust (Williams, et al. 2020). The CNN architecture used was an adaptation of U-Net (Çiçek, et al. 2016) and the heatmaps were regressed in training. A multi-task approach was used, to determine the distinct SP and LAM heatmaps simultaneously, utilising transfer learning between the two tasks. Finally, the SoftMax layer of U-Net was removed to generate a continuous output.

ii) Post-processing to identify minimal hiatal dimension

The second step identifies the extreme coordinates from the regression output. This was achieved with a computational post-processing step inspired by the IUGA clinical guidelines (IUGA,2019). While our landmark regression was performed in 3D, clinicians normally identify the plane defining ‘extreme coordinates’ within a single 2D MS plane (Williams, et al. 2020). Thus, to follow clinical guidelines, a 2D approach was also followed in our automatic pipeline, to create a workflow that was comparable to the clinical one. The combined voxel maxima of the SP and LAM heatmaps were determined within a small range of 2D MS planes to reduce computational load and running time. Thus, the SP and LAM combined overall voxel maxima, corresponding extreme coordinates and MS plane were identified.

iii) Extraction of the C-plane

The final step of task one was to slice and resample the 3D TPUS as the automatically defined 2D C-plane. The C-plane was defined as the plane orthogonal to the depth direction of the TPUS volume at acquisition, thus contains the orthogonal vector, $|001|$. The C-plane also contains the vector, $\overrightarrow{AB}$, that joins the extreme coordinates of the SP and LAM identified in the previous step. The cross product of these two orthogonal vectors defines the final orthogonal vector as $-AB_y\boldsymbol{i} + \mathrm{A}B_x\boldsymbol{j} + 0\boldsymbol{k}$. Clinical guidelines suggest the vector $\overrightarrow{AB}$ has a magnitude within the x and y directions only, as the extreme coordinates lie within the same MS plane (z slice) (Williams, et al. 2020), which was determined in the previous section. Therefore, the bases of the C-plane are defined as,

$$\|b_x\|\|b_y\|\|b_z\| = \left\|\begin{matrix} AB_x & -AB_y & 0 \\ AB_y & AB_x & 0 \\ 0 & 0 & 1 \end{matrix}\right\|. \tag{1}$$

Once the TPUS volume was rotated, the C-plane was extracted at the mid-point between the SP and LAM extreme coordinates.

iv) Levator hiatus segmentation

The second task of the proposed pipeline was to automatically define the hiatal area from the extracted 2D C-plane, elaborating on previous work (Bonmati, et al. 2018, Sindhwani, et al. 2016). In this study, a 2D CNN accepts the extracted 2D C-plane from the previous task and automatically classifies the voxels as levator hiatus (1) or background (0). The network architecture utilised was an implementation of 2D U-Net (Ronneberger, et al. 2015). Due to the nature of US, segmentation can be difficult due to noise, artefacts and blurring, thus advanced data augmentation was used including elastic deformation and our own adaptation of the original mix-up (Zhang, et al. 2018), where three images and their corresponding ground-truth labels were linearly combined instead of two. Post-processing morphological operators were applied to the CNN output, such as connected component analysis, fill-holes and Gaussian blur of sigma value 0.5. This post-processing was used to ensure that the segmentation was complete (i.e. no holes) and that the boundary was smooth, which ensures the hiatal output was more realistic.

Implementation details

The CNN models were implemented using NiftyNet (Gibson, et al. 2018) on a desktop with a 24GB NVIDIA Quadro P6000 (NVIDIA, California, United States)

3D landmark regression

The network architecture of 3D U-Net (Çiçek, et al. 2016) was adapted to have one input (i.e. TPUS volume) and two outputs (i.e. SP and LAM heatmaps) at testing, to ensure a multi-task approach to learning. The final SoftMax layer was removed to output a continuous value which ranges between zero and a maximum value. The loss function was a combined L2 loss of the SP and LAM heatmaps with an initial learning rate of 10^{-4}. A RMSprop optimiser, parametric ReLU activation function, weighted decay factor of 10^{-5} and batch size of six were used. Histogram based normalisation and whitening were used, thus the volume was set to have zero-mean and unit variance. A combined smooth version of the heatmaps was used for weighted sampling during training. The following data augmentation were used: random scaling (with a range of -10, +10%), random rotation of all axes (with a range of -10°, +10°) and our own adaptation of *mixup* (Zhang, et al. 2018). Methods were optimised until network convergence of a validation set (i.e. subset of TPUS volumes from the training dataset).

C-plane hiatal area segmentation

The network architecture used was an adaptation of 2D U-Net (Ronneberger, et al. 2015) as it has proven to perform well in other 2D US semantic medical imaging tasks (Bonmati, et al. 2018, Li, et al. 2019). An Adam optimiser, ReLU activation function, weighted decay factor of 10^{-5} and batch size of 32 were used. Whitening was applied to reduce the effects of noise; thus, the image was set to have zero-mean and unit variance, and histogram normalisation was further performed (Pal and Sudeep 2016). A loss function of combined cross entropy and Dice score was used, with an initial learning rate of 10^{-3}. Balanced window sampling was used during training (i.e. regions of label and background were equally sampled). During training the following data augmentation were used: random rotation (with a range of -5°,+5°), elastic deformation (deformation sigma = nine, number of control points= four and proportion to deform 0.5), random scaling (range of -20,+20%), vertical 'flipping' and our implementation of *mixup* (Zhang, et al. 2018).

Data collection

Analysis of anonymised, archived, ultrasound images was retrospective, therefore, no ethics committee approval was required by KU Leuven, Belgium.

Training data - C-plane detection

Regarding the 3D C-plane detection task, a training dataset of 25 3D TPUS volumes was used. This was the same dataset used in our previous study (Williams, et al. 2020). The training dataset comprised of 13 clinical cases with a range of pelvic floor dysfunctions, assessed at the pelvic floor clinic in UZ Leuven, Belgium. Multiple TPUS volumes were obtained from the 13 clinical cases (i.e. at rest, Valsalva and/or pelvic floor contraction). 3D segmentations of the SP and LAM were provided by an expert human annotator (referred to as expert 1) and used to generate the heatmaps via the process shown in Figure 3. Expert 1 was chosen for their experience in this domain and in annotating the 3D LAM and SP structures from a TPUS volume, expert 1 had 12 months of experience in annotating the SP and LAM in 3D TPUS volumes prior to data curation.

Training data- 2D levator hiatus segmentation

Regarding levator hiatus segmentation, a training dataset of 256 2D C-planes and corresponding ground truth labels of the levator hiatus were used to train the CNN segmentation model. The training dataset comprised two sets of archived clinical images with expert annotations, acquired by several operators, which allows the CNN to learn a variety of acquisition parameters and image qualities. Within the training dataset a subset of 91 2D C-planes with expert annotations were used in our previous studies (Bonmati, et al. 2018, Sindhwani, et al. 2016), in this dataset the expert had over four years of experience in acquiring and analysing pelvic floor TPUS volumes.

Test data

The test data included a randomised selection of 73 anonymised TPUS volumes from 37 other *symptomatic* women assessed at the pelvic floor clinic, between February and June 2019. There is no patient overlap across training and testing sets. The test data was evaluated in a previous study (Williams, et al. 2020) and was not used to train the CNN models; it was used purely for testing the proposed pipeline. Detailed patient information is included in Table S2 in the Supplementary material.

Hiatal measurements were delineated by expert 1, resulting in Gold Standard (GS) C-plane orientations and levator hiatus segmentations used for validation. The GS C-plane orientations were extracted using GE 4DView software (GE Healthcare, Zipf, Austria) and the corresponding GS hiatal segmentations were delineated using 3D Slicer software (Slicer 2020, Fedoroy, et al. 2012). Two operators (expert 1 and expert 2) participated in the inter-operator reliability studies. At the time of the analysis, both experts had over four years of experience in acquiring and analysing pelvic floor TPUS volumes. Both experts work as clinicians in the pelvic floor disorder clinic at UZ Leuven, Belgium, and were asked to identify the C-plane following the IUGA guidelines (IUGA, 2019). The experts identified the C-plane using the multi-planar technique (Williams et al., 2020) on GE 4D View software (GE Healthcare, Zipf, Austria). The experts performed manual hiatal outlining and C-plane detection on all 73 TPUS volumes.

Quantitative metrics for evaluation

Several metrics were used to describe the similarity of the manual C-plane detection and the levator hiatus segmentation to the computer-generated output. As this was a 'two-task' pipeline both 'tasks' were evaluated independently as well as jointly.

C-plane detection

Validation of the C-plane detection task was similar to the previous study (Williams, et al. 2020). To validate the accuracy of the plane detection task, the angular difference between the identified C-plane against the GS plane was measured as well as the Euclidean distance of the midpoints of the planes within the TPUS volume. The angular difference computed was the averaged x axis and y axis angular difference, as the z axis was fixed as per guidelines (IUGA, 2019). To evaluate clinical relevance, a visual *Turing Test* was proposed and evaluated on 10 TPUS volumes. Hereto, expert 1 was asked to blindly rate a randomised selection of (manually and automatically detected C and MS planes to give a Likert scale score from zero to five (5 being excellent, 4, above average, 3, average, 2, below average, 1, poor, and 0, of no clinical use). Test one was based on the placement of the C-plane within the TPUS volume. Test two was based on the C-plane quality for clinical diagnosis. A paired Wilcoxon test was performed to compare the performance of the proposed method against expert 1's GS recording, this generated an output score which was averaged per TPUS volume. The paired Wilcoxon test is calculated by deducting expert 1's GS score from the method's (i.e. algorithm, inter-observer or intra-observer) score. The score ranges from a negative value to a positive value, depending on the performance of the detected C-plane against the GS. If the score was positive, it suggests the detected C-plane method performed 'better' visually than the manual GS; if the overall score was negative it suggests the detected method performed 'worse' visually than the manual GS, and a score of zero means the methods performed the same.

Levator hiatus localisation and segmentation

The levator hiatus outline (i.e. hiatal area) identified in the C-plane is a biomarker used for the analysis of given pelvic floor disorders. In order to assess the extracted biomarker quality, the following metrics were computed: the Hausdorff Distance (HD) and the Robust 95th percentile HD of the levator hiatus segmentation that were evaluated against the GS manual hiatal segmentation from the GS C-planes. The hiatal area is an important biomarker; thus, the area of the GS hiatal outline was compared to the hiatal outline of the automatic extracted C-plane. Moreover, the hiatal area difference and absolute hiatal area difference were calculated. To evaluate clinical acceptability of the segmentations, another visual *Turing* Test (Turing Test 3) is proposed and evaluated on 10 extracted C planes and corresponding segmentations. The hiatal segmentations were rated a 'clinical score' by expert 1, from zero to five as above, and compared to the score of the GS in a paired Wilcoxon test. Expert 1 performed the test three months after they annotated the GS hiatal segmentations to limit the impact of pre-learning bias. The average result per TPUS volume was presented and the score will range between +5 and -5.

Computer-observer, intra-observer and inter-observer differences

The computer-observer differences (COD), intra-observer differences (IAOD) and inter-observer differences (IEOD) were evaluated. COD were evaluated by calculating all similarity metrics between automatic hiatal segmentations on automatic C-planes and expert 1 manual hiatal segmentations on GS C-planes. The IAOD was evaluated by calculating similarity metrics between identified C-planes and hiatal outlines generated by expert 1 GS and a second analysis from expert 1 taken a month after the GS was generated. The second analysis was undertaken two months before the Turing test analysis, in order to reduce bias and the risk of the expert recognising their analysis and thus rating it higher subconsciously. In addition as the experts are active members of the clinical team at UZ Leuven, they analyse new TPUS volumes daily and we assume bias is limited as this is a common and repetitive task. Finally, IEOD was evaluated by calculating similarity metrics between expert 2 and the first assessment from expert 1.

Statistical analysis

To evaluate the reliability of the automatic method, a paired f-test was used to test several null hypotheses. The first was that the automated method agreed with expert 1's GS at least as well as expert 1 agreed with themself (i.e. the variance of the differences between the automatic method and the GS was not larger than the variance in intra-observer differences). The second null hypothesis tested was that the automated method agreed with expert 2 at least as well as expert 2 agreed with expert 1's GS result (i.e. the variance of the differences between the automatic method and the GS was not larger than the variance in inter-observer differences). The final null hypothesis tested was that expert 2 agreed less with expert 1's GS results than expert 1 agreed with themself (i.e. the variance in inter-observer differences was statistically greater than the variance intra-observer differences). Type one statistical errors (i.e. multiple testing) were accounted for using a Bonferroni correction, hence the p-value obtained was reduced by a factor of three. Therefore, the p-value ≤ 0.017 was used as a cut-off to show statistical significance. To further evaluate the reliability of the automatic method, the Bland-Altman limits of agreement were calculated for COD, IAOD and IEOD.

To evaluate the possibility of bias between the methods (i.e. automatic, expert 1 and expert 2) to expert 1's GS, several paired t-tests were used to test several null hypotheses. The null hypotheses were the same as above, however, based on the mean difference, i.e. bias, rather than on the variance of the differences. As above, type one statistical errors (i.e. multiple testing) were accounted for using a Bonferroni correction, and a p-value ≤ 0.017 was used as a cut-off to show statistical significance.

Results

Figure 4 shows examples of the C-plane position within the TPUS volume and the corresponding extracted C-planes and hiatal segmentations. The images represent the 0^{th}, 25^{th}, 50^{th}, 75^{th} and 100^{th} percentiles, respectively, of the 95^{th} Hausdorff distance metric (the corresponding 95th HD distance are included along with the hiatal areas). The red lines and masks represent the automatic method, and the green lines and masks the GS. Qualitatively, the computer-generated C-planes and levator hiatus segmentations matched well with the GS C-planes and hiatal segmentations.

Table 1 shows the semi-qualitative average result of the COD, IEOD and IAOD from the *visual* Turing Tests. The results show that the pipeline performs better than IEOD and IAOD, in relation to the C-plane detection task within a TPUS volume, as COD scored 0.00±0.77 for Turing Test 1. The pipeline scored comparable to the GS with a score of 0.00±1.07 for Turing Test 2 (C-plane quality), which was a lower error than IEOD scoring -0.20±0.77. The proposed pipeline scored -1.50±1.01 for Turing Test 3 (hiatal segmentation quality) whereas IEOD scored -0.80±0.60.

The quantitative results from the C-plane detection task (Table 2) demonstrate that the COD's bias and variance are not significantly higher than the IAOD and are significantly smaller than the IEOD. Moreover, as expected, for all C-plane detection metrics the IEOD was statistically larger than the IAOD for both bias and variance.

The quantitative results from the second task (i.e. levator hiatus segmentation) are given in Table 3. The COD and IEOD both have a statistically higher bias and variance than the IAOD's bias and variance, for the 95th Robust HD, and the HD. However, the variance and bias of the COD were not statistically different from that of the IEOD for the 95th Robust HD and HD.

Table 3 shows that for the hiatal area difference, the COD's bias and variance were not statistically higher than the IEOD's bias and variance. The COD's variance was statistically higher than the measured IAOD's variance, and the IEOD's bias was statistically higher than the IAOD's bias.

Regarding the absolute hiatal area, the bias and variance of the COD was not statistically different from those of the IAOD or IEOD, and the IEOD was not statistically different from the IAOD for the bias and variance.

Table 4 shows the limits of agreement of COD, IAOD and IEOD for all metrics evaluated in this study.

The computer automated C-plane detection and hiatal segmentation pipeline took 35 ± 4 seconds, and the manual process took Expert 1 on average 2 minutes and 35 ± 17 seconds to identify the C-plane and segment the hiatal area on GE software.

Discussion

This study presents a fully automatic hiatal biomarker extraction pipeline from a TPUS volume. Previous studies showed promising results for automatic hiatal segmentation, but required manual determination of the 2D C-plane (Bonmati, et al. 2018, Li, et al. 2019) which is time-consuming and prone to error.

Qualitatively in Figure 4 there was minimal difference between the automatically extracted and GS C-planes. The 100th-25th percentile results show accurate SP and LAM positioning within the TPUS volume. However, the 0th percentile shows an inaccurate SP position for the automated task. This particular case was a patient with severe hiatal ballooning. Ballooning may be that severe that the SP is not fully present within the TPUS. In those circumstances, the operator will watch the SP move during Valsalva in real-time and *estimate* the position. Unfortunately, this was not exploited by the proposed method, thus it would not perform as good as an expert in these extreme cases.

Table 1 shows that in Test 1 (C-plane position quality) the automated method performed as good as the GS and better than IEOD and IAOD when visually assessed for the true C-plane position. The results from Turing Test 2 and 3 are based on C-plane and hiatal area segmentation quality respectively. Despite COD achieving a high-quality C-plane, on average the segmentation quality scored noticeably worse. The lower accuracy may be due to the variety of image qualities and pathologies within the testing dataset, hence including more pathological training data may improve results. Nevertheless, the average score for the proposed method was above three (average) hence still clinically acceptable.

Table 2 indicates that the pipeline performed with a lower bias and variability w.r.t. expert 1 than expert 2 did; and performed similar to expert 1 in the C-plane detection task. Table 4 indicates that the automated C-plane detection task performed within the limits of agreement of the measured inter-observer difference, highlighting that the first part of the pipeline may reduce the observed bias and variability, below the inter-observer variability measured in this study. This may be due to subtle difference of techniques used by the experts to identify the C-plane, although the experts were instructed to follow the IUGA clinical guidelines using GE 4D View software (GE Healthcare, Zipf, Austria) and the multi-planar technique (Williams et al., 2020).

To reduce bias experts were not managed during the testing phase, in order to measure the real-world inter-observer variability of experts working in the same pelvic floor clinic at the same institute, and both with at least four years' experience. The training data used for the C-plane detection task were generated by expert 1, who also identified the GS C-plane orientations. Thus, it may be assumed the network learnt to identify the extreme coordinates more similarly to expert 1 or as the pipeline was based on the extreme coordinate position, expert 1 followed the IUGA guidelines more closely than expert 2, and the C-plane was positioned closer to the extreme coordinates.

This is a common trait for a majority of supervised learning tasks that utilise CNNs, the network is trained on data from a specific observer and hence will learn to identify features similarly. This trait may be seen as an advantage or disadvantage based on the application. For example, it can learn the behaviour of a specific expert or in this case a clinical guideline, and can create a personalised automatic workflow that mirrors the expert with the lowest intra-observer variability and most experience, or it may mirror a standardised clinical guideline.

Nevertheless, for other applications (i.e. not guideline related) if desired it can be beneficial to expand the training dataset across several experts, to learn to identify features similarly to several experts rather than one in particular, which makes the CNN more generalisable. This approach was taken for the hiatal area segmentation task of this pipeline. However, a disadvantage of this approach is that the accuracy can reduce if experts disagree, or if one expert delineates with a large error. This approach could leave to no experts being satisfied with the algorithm's result. Therefore, quality control should be conducted to assess the training segmentation data prior to training, regarding testing this is less important and a variety of experts with adequate experience may be included, to gauge the current clinical world inter-observer variability of a specific task.

The pipeline was able to extract the hiatal area to a high level of accuracy. In Table 3 the bias and variance of the COD were not statistically higher than the IEOD regarding hiatal area error metrics (i.e. hiatal area difference and absolute hiatal area). This suggests that the proposed method extracts hiatal biomarkers as good as experts and thus is clinically acceptable. The IEOD's bias for hiatal area difference was statistically higher than the IAOD, indicating that the proposed method may reduce the bias below the measured IEOD. The COD's variance for the hiatal area difference was statistically higher than the IAOD's variance, however, as it was not statistically higher than the IEOD's variance, it is still clinically acceptable.

Literature records a hiatal area difference (bias) of $0.61cm^2$ (Bonmati, et al. 2018) and $0.23cm^2$, $1.1cm^2$ (for U-Net and Dense U-Net respectively) (Li, et al. 2019). This study recorded a bias of $0.91cm^2$. The bias will be higher in this study as the levator hiatus is a 3D structure and between C-plane positions the area will differ, thus there is an accumulation of error and is not directly comparable. Nevertheless, the bias may be higher due to the 2D levator hiatus segmentation training dataset used; it consists of contrasted post-processed C-planes, whereas the testing dataset is un-post-processed. In addition, unlike literature, the training dataset was from a different data centre to the testing dataset, hence the image qualities differ. To improve results annotated un-post-processed C-planes may be used in training.

The method in this study was tested on a clinical dataset of patients with a range of anatomical variability and pathological conditions, such as severe hiatal ballooning, levator avulsion, and bladder neck hypermobility, as well as patients without pathology. The dataset was even more challenging as up to 81.1% of the patients had pelvic organ prolapse, hence had a wide range of extreme coordinate movement.

The approach taken in this study utilises information extracted from data, the geometry of the patient and clinical guidelines, to drive a hybrid approach to extract hiatal dimensions. The proposed method accomplished relatively low errors with a small training dataset, typically rare in deep learning applications. The proposed method performs faster than an expert, however, not in real-time. For real-time clinical implementation, the pipeline would have to be optimised. The proposed method performs within inter-observer and intra-observer error (for most evaluated metrics); thus, a high level of pelvic floor disorder analysis training may no longer be required for experts to extract high-quality hiatal biomarkers. Furthermore, the output is interpretable to clinicians as the extreme coordinates are well known and recognisable, thus if the C-plane is incorrect it is easy to identify the problem (i.e. misplacement of the SP due to shadowing).

Clinically, experts commonly acquire a 4D TPUS volume, referred to as a Cine loop. Currently the volume of interest (i.e. volume of maximal contraction) is selected manually by the expert. In future work, one aims to expand this method to localise the volume of interest from the Cine loop. Finally, the proposed pipeline will be made interactive, to allow operators to adapt the C-plane position and/or the 2D hiatal segmentation.

Conclusion

In conclusion, our method was able to extract high-quality C-planes and hiatal area measurements from TPUS volumes without user input. The time taken for hiatal extraction decreased by 120 seconds, saving clinicians time. Furthermore, the automated pipeline reduces error below the inter-observer variability for evaluated metrics within this study.

Acknowledgments

We gratefully acknowledge the support of GE Healthcare Women's Health Ultrasound (Zipf, Austria) for their on-going research and data support and NVIDIA Corporation for the GPU grant (California, United States).

Supplementary information:

Table of abbreviations is shown in Table S1.

Demographics of the 37 patients used for evaluating the proposed pipeline is shown in Table S2.

References

AIUM/IUGA Practice Parameter for the Performance of Urogynecological Ultrasound Examinations: Developed in Collaboration with the ACR, the AUGS, the AUA, and the SRU. Journal of Ultrasound in Medicine 2019; 38:851-64.

Abdool Z, Shek KL, Dietz HP. The effect of levator avulsion on hiatal dimension and function. American Journal of Obstetrics and Gynecology 2009; 201:89.e1-89.e5.

Bonmati E, Hu Y, Sindhwani N, Dietz HP, D'hooge J, Barratt D, Deprest J, Vercauteren T. Automatic segmentation method of pelvic floor levator hiatus in ultrasound using a self-normalizing neural network. Journal of Medical Imaging 2018; 5:021206.

Brooks SV, Zerba E, Faulkner JA. Injury to muscle fibres after single stretches of passive and maximally stimulated muscles in mice. The Journal of Physiology 1995; 488:459-69.

Çiçek Ö, Abdulkadir A, Lienkamp S, Brox T, Ronneberger O. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. 2016.

Dietz H, Simpson J. Levator trauma is associated with pelvic organ prolapse. BJOG: An International Journal of Obstetrics & Gynaecology 2008; 115:979-84.

Dietz HP. Pelvic floor ultrasound: a review. American Journal of Obstetrics and Gynecology 2010; 202:321-34.

Fedorov A, Beichel R, Kalpathy-Cramer J, Finet J, Fillion-Robin J-C, Pujol S, Bauer C, Jennings D, Fennessy F, Sonka M, Buatti J, Aylward S, Miller JV, Pieper S, Kikinis R. 3D Slicer as an image computing platform for the Quantitative Imaging Network. Magn Reson Imaging 2012; 30:1323-41.

Gibson E, Li W, Sudre C, Fidon L, Shakir DI, Wang G, Eaton-Rosen Z, Gray R, Doel T, Hu Y, Whyntie T, Nachev P, Modat M, Barratt DC, Ourselin S, Cardoso MJ, Vercauteren T. NiftyNet: a deep-learning platform for medical imaging. Computer Methods and Programs in Biomedicine 2018; 158:113-22.

Ismail S, Duckett J, Rizk D, Sorinola O, Kammerer-Doak D, Contreras-Ortiz O, Al-Mandeel H, Svabik K, Parekh M, Phillips C. Recurrent pelvic organ prolapse: International Urogynecological Association Research and Development Committee opinion. Int Urogynecol J 2016; 27:1619-32.

Li X, Hong Y, Kong D, Zhang X. Automatic segmentation of levator hiatus from ultrasound images using U-net with dense connections. Physics in Medicine & Biology 2019; 64:075015.

Lien K-C, Mooney B, DeLancey JOL, Ashton-Miller JA. Levator ani muscle stretch induced by simulated vaginal birth. Obstet Gynecol 2004; 103:31-40.

Pal KK, Sudeep KS. 2016 Preprocessing for image classification by convolutional neural networks. *2016 IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)*, 1778-81.

Ronneberger O, Fischer P, Brox T. 2015 U-Net: Convolutional Networks for Biomedical Image Segmentation, In: Navab N, Hornegger J, Wells WM, and Frangi AF, eds. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Cham: Springer International Publishing, 234-41.

Schwertner-Tiepelmann N, Thakar R, Sultan AH, Tunn R. Obstetric levator ani muscle injuries: current status. Ultrasound in Obstetrics & Gynecology 2012; 39:372-83.

Shek K, Dietz H. Intrapartum risk factors for levator trauma. BJOG: An International Journal of Obstetrics & Gynaecology 2010; 117:1485-92.

Sindhwani N, Barbosa D, Alessandrini M, Heyde B, Dietz HP, D'Hooge J, Deprest J. Semi-automatic outlining of levator hiatus. Ultrasound in Obstetrics & Gynecology 2016; 48:98-105.

Slicer D. 2020, 3D Slicer web site.

van den Noort F, van der Vaart CH, Grob ATM, van de Waarsenburg MK, Slump CH, van Stralen M. Deep learning enables automatic quantitative assessment of puborectalis muscle and urogenital hiatus in plane of minimal hiatal dimensions. Ultrasound in Obstetrics & Gynecology 2019; 54:270-75.

Williams H, Cattani L, Yaqub M, Sudre C, Vercauteren T, Deprest J, D’hooge J. 2020 Automatic C-Plane Detection in Pelvic Floor Transperineal Volumetric Ultrasound, In: Hu Y, Licandro R, Noble JA, Hutter J, Aylward S, Melbourne A, Abaci Turk E, and Torrents Barrena J, eds. *Medical Ultrasound, and Preterm, Perinatal and Paediatric Image Analysis*. Cham: Springer International Publishing, 136-45.

Williams H, Cattani L, Yaqub M, Vercauteren T, Deprest J, D’Hooge J. 2020 Data augmentation to aid 3D convolutional neural network segmentation of landmarks in a small volumetric ultrasound dataset of the pelvic floor

. IUS IEEE: KU Leuven, KCL.

Zhang H, CissÈ M, Dauphin Y, Lopez-Paz D. mixup: Beyond Empirical Risk Minimization. ArXiv 2018; abs/1710.09412.

Legends of figures:

Figure 1: The typical acquisition and evaluation screen on Voluson systems shows three orthogonal planes: A- sagittal, B- coronal and C- axial; the bottom right image (3D) is the Axial plane rendered volume. This volume has been aligned as the desired MHD position and the extreme coordinates are marked by red dots. Abbreviations: A, anal canal; B, bladder; LAM, levator ani muscle; R, rectum; SP, symphysis pubis; U, urethra; V, vagina.

Figure 2: Overall levator hiatus analysis pipeline split into two tasks by colour, the first section (pink) is the automatic detection and extraction of the C-plane; the second task (orange) being the automatic segmentation of the hiatal area within this C-plane.

Figure 3: Visualisation of the steps to generate the ground-truth heatmaps used in this study. The desired heatmap of the extreme coordinates (red dots) of the SP (left) and LAM (right) are identified from 3D segmentations manually-delineated by experts. The first row shows the segmentation of the SP and LAM, the second row shows the distance heatmap (i.e., extreme coordinate = 0 and the voxel value radially increases with distance) and the third row shows the smooth inverse distance heatmap (i.e. extreme coordinate is the maximum value and the voxel value radially decreases with distance).

Figure 4: The GS C-plane position is shown by a green line and the computer automated C-plane position is shown by a red line for each corresponding TPUS. The corresponding GS manual segmentation of the hiatal area is the green mask and the automated segmentation of the hiatal area is the red mask under its corresponding TPUS image. TPUS images show an increasing computer-generated hiatal outline quality that represent the 0th, 25th, 50th, 75th and 100th percentiles, respectively, of the 95th Hausdorff distance.

Table 1: Turing Test score per TPUS volume, a negative result indicates the GS performed better than the other method in comparison. The scores can range from -5 to +5 (dependent on the GS score and evaluated method score). A score of 0 means that the GS performed equally to the other method evaluated. A positive score would mean the method outperformed the GS and a negative score implies that the GS performed better than the evaluated method.

	COD	IEOD	IAOD
Turing Test 1	0.00±0.77	-1.00±1.34	-0.20±0.98
Turing Test 2	0.00±1.07	-0.20±0.77	0.30±0.92
Turing Test 3	-1.50±1.01	-0.80±0.60	0.00±0.44

Table 2: COD, IAOD, IEOD differences and standard deviations of C-plane detection metrics: angular difference of the C-planes and Euclidean distances of the C-plane midpoints.

	COD	IAOD	IEOD
Angular difference (°)	5.76 ± 5.06†	4.94±4.24	9.39±6.21§*
Euclidean distance (mm)	6.46±5.18	5.80±4.15	8.48±6.62§*

† Mean statistically significantly different from IEOD

§ Mean statistically significantly different from IAOD

* Variance statistically significantly different from IAOD

Table 3: COD, IAOD, IEOD errors of hiatal segmentation metrics.

	COD	IAOD	IEOD
95^{th} Robust Hausdorff distance (mm)	7.30±4.99§*	5.10±3.45	8.48±6.13§*
Hausdorff distance (mm)	11.26±5.95§*	7.62±3.88	11.52±6.60§*
Hiatal area difference cm^2	0.98±3.74*	-0.52±2.74	2.05±2.86§
Absolute hiatal area cm^2	2.66±2.78	1.81±2.12	2.53±2.34

§ Mean statistically significantly different from IAOD

* Variance statistically significantly different from IAOD

Table 4: The COD, IAOD, IEOD limits of agreement of all pipeline metrics are shown. The limits of agreement are presented as {lower limit, upper limit}.

	COD	IAOD	IEOD
Angular difference (°)	{-4.15, 15.68}	{-3.37, 13.25}	{-2.78, 21.56}
Euclidean distance (mm)	{-3.69, 16.61}	{-2.33, 13.93}	{-4.50, 21.46}
95^{th} Robust Hausdorff distance (mm)	{-2.48, 17.08}	{-1.66, 11.86}	{-3.53, 20.49}
Hausdorff distance (mm)	{-0.40, 22.92}	{0.02, 15.22}	{-1.42, 24.46}
Hiatal area difference (cm^2)	{-6.35, 8.31}	{-5.89, 4.85}	{-2.63, 6.55}
Absolute hiatal area difference (cm^2)	{2.79, 8.11}	{-2.35, 5.97}	{-2.06, 7.12}

Table S1: Abbreviations within text and their corresponding definitions

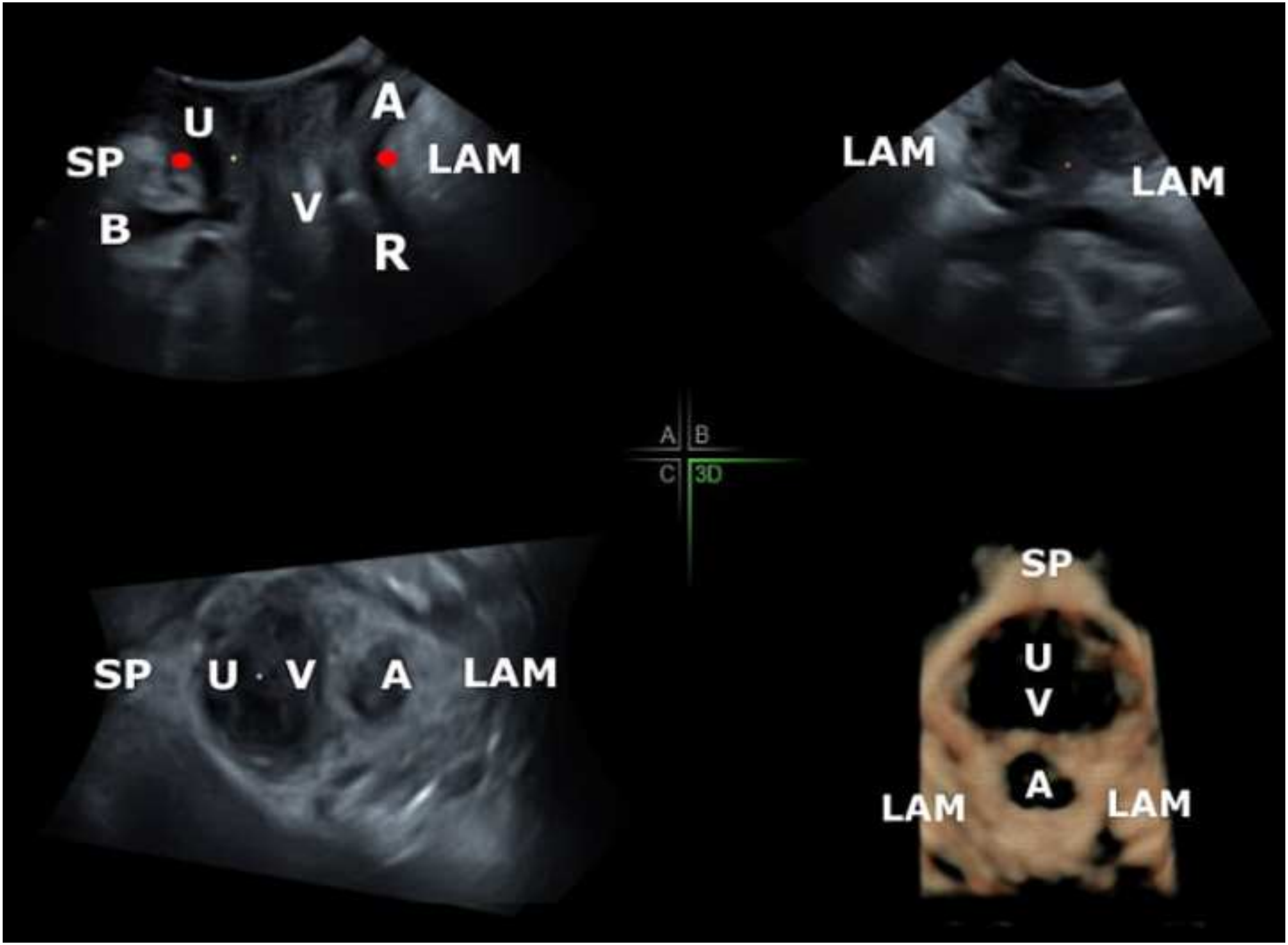
Abbreviation	Definition
TPUS	Transperineal ultrasound
CNN	Convolutional neural network
3D	Three-dimensional
US	Ultrasound
PFUS	Pelvic floor ultrasound examination
LAM	Levator ani muscle
SP	Symphysis pubis
MHD	Minimal hiatal dimensions
MS	Mid-sagittal
HD	Hausdorff distance
COD	Computer-observer differences
IAOD	Inter-observer differences
IEOD	Intra-observer differences
U	Urethra
V	Vagina
A	Anal canal
B	Bladder
R	Rectum
2D	Two-dimensional

Table S2. Characteristics of the study population. Data are presented as mean (standard deviation), as prevalence in % (ratio) or as median [IQR].

Demographic variables	**Values**
Age *years)	57.6 (14.3)
BMI (kg/m^2)	26.7 (3.8)
Obstetric variables	
Vaginally parous	75.7 % (28/37)
Only caesarian section	8.1 % (3/37)
Nulliparous	5.4 % (2/37)
Vaginal parity	2 [1.25]
Max birth weight in grams	3741 (439)
Symptoms of pelvic floor dysfunction	
Urinary incontinence	
- Stress urinary incontinence	48.7 % (18/37)
- Urge urinary incontinence	21.6 % (8/37)
Pelvic organ prolapse	81.1 % (30/37)
Anal incontinence	2.7 % (1/37)

Figure 1

Click here to access/download;Figure;figure1_journal_HW.png



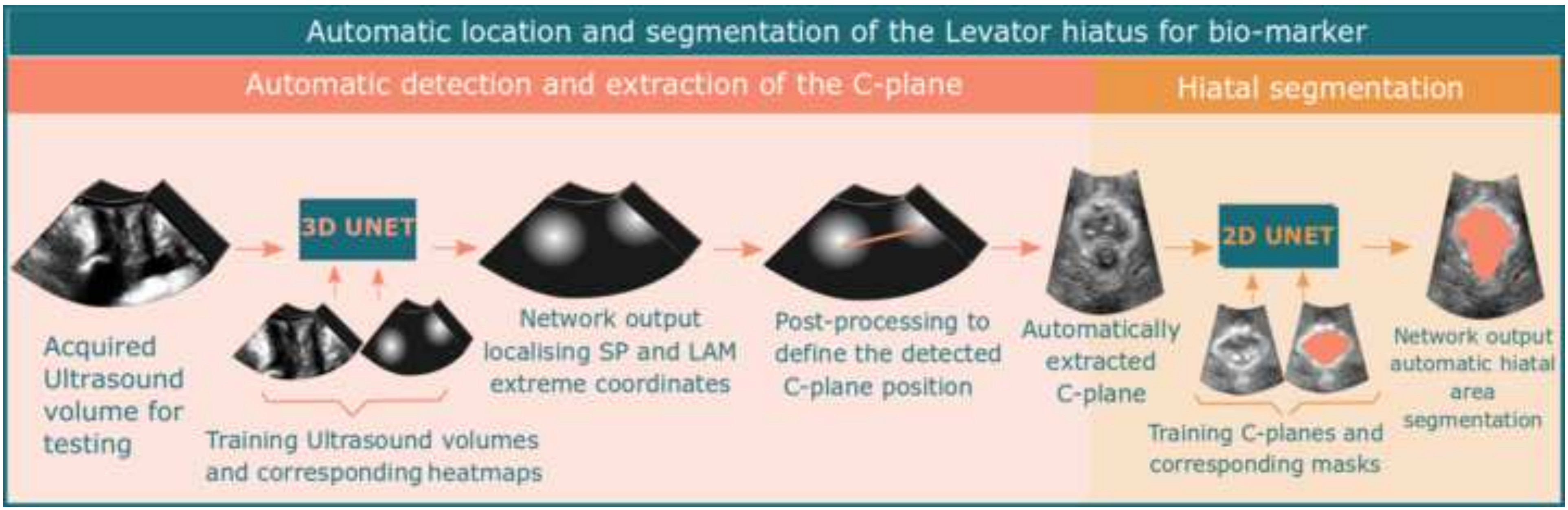
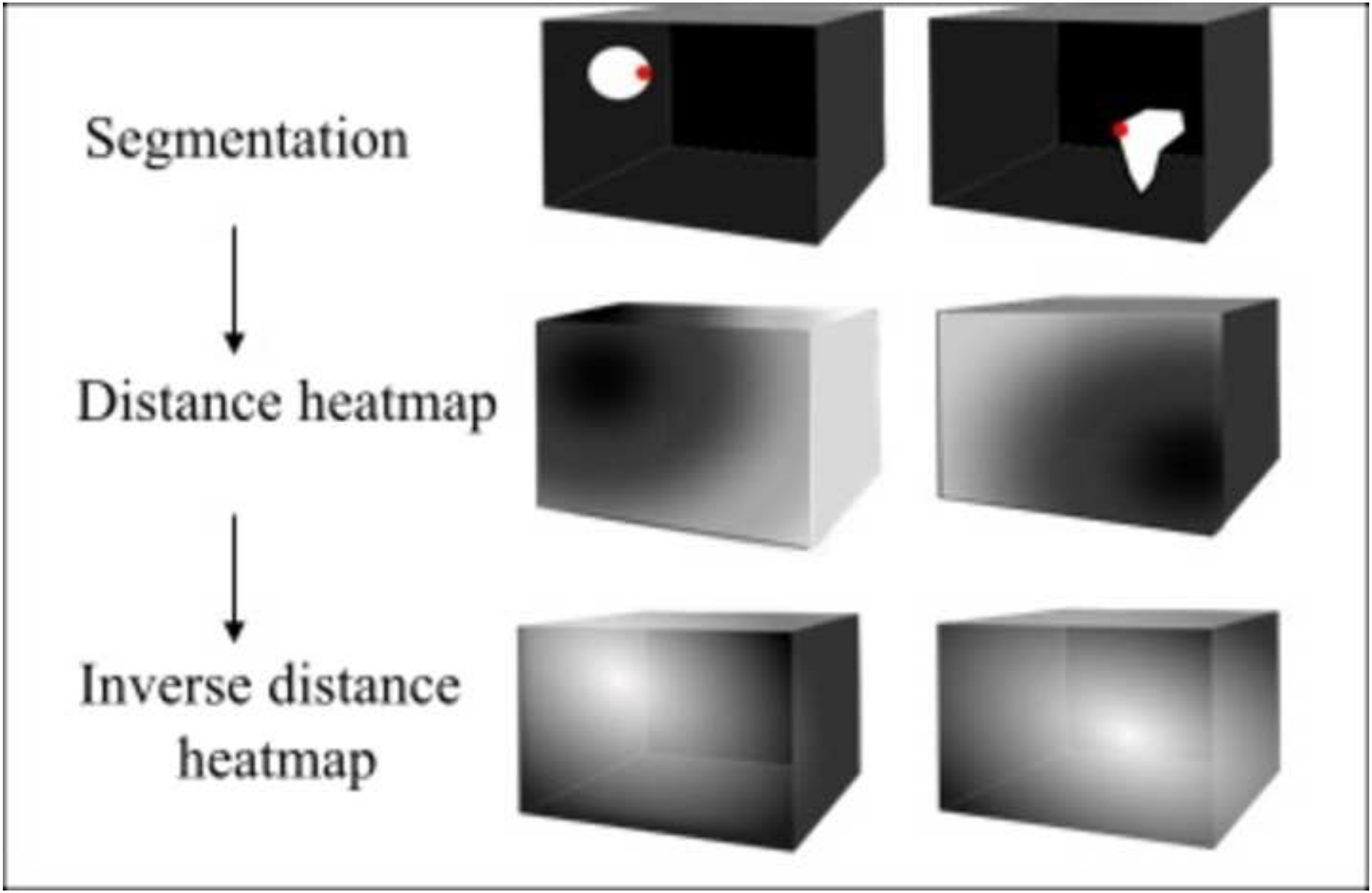
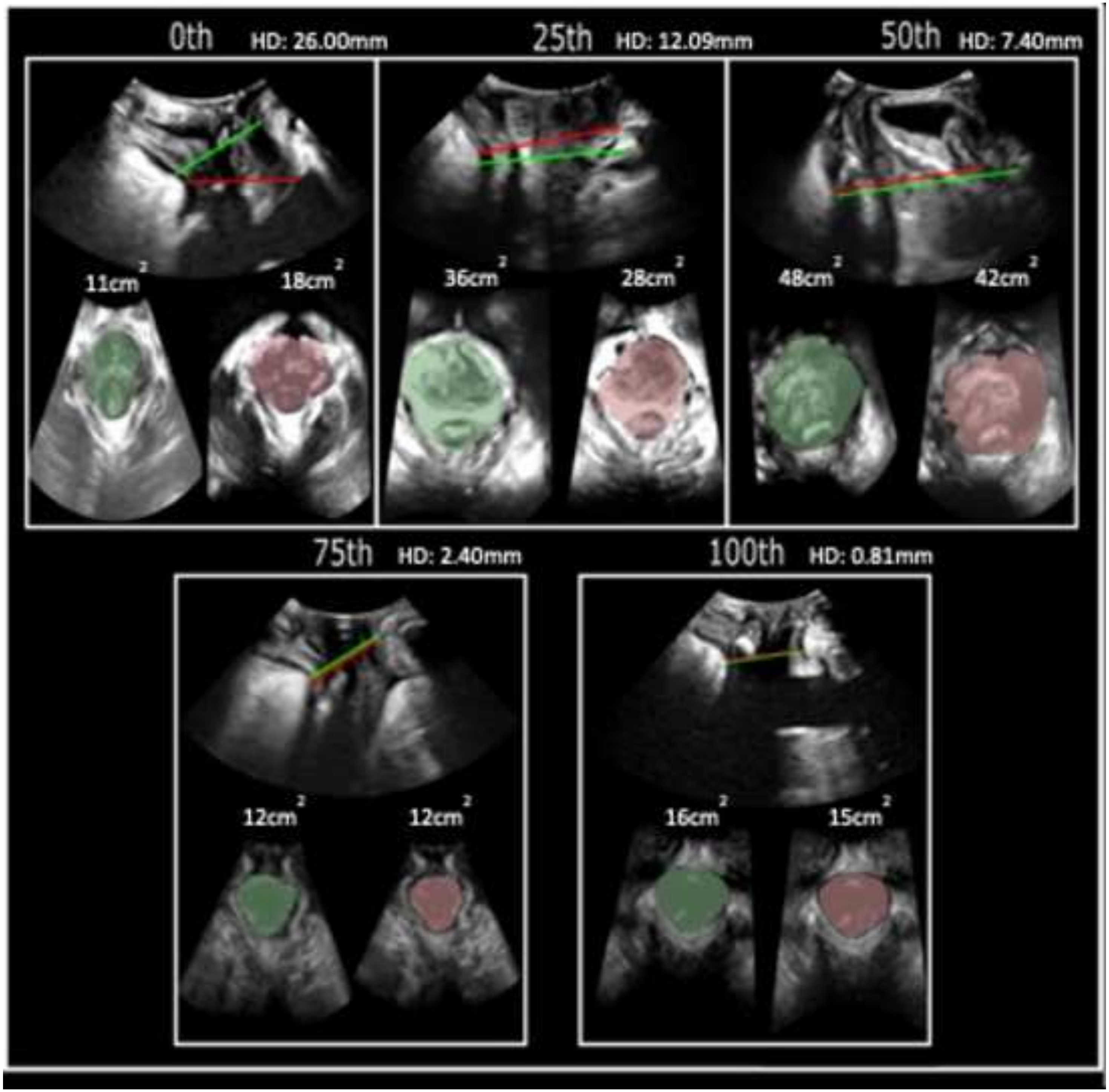

Automatic location and segmentation of the Levator hiatus for bio-marker
Automatic detection and extraction of the C-plane
Hiatal segmentation
3D UNET
2D UNET
Acquired Ultrasound volume for testing
Training Ultrasound volumes and corresponding heatmaps
Network output localising SP and LAM extreme coordinates
Post-processing to define the detected C-plane position
Automatically extracted C-plane
Training C-planes and corresponding masks
Network output automatic hiatal area segmentation

Figure 3

Click here to access/download;Figure;heatmap_workflow.png




0th HD: 26.00mm
25th HD: 12.09mm
50th HD: 7.40mm
11cm2
18cm2
36cm2
28cm2
48cm2
42cm2
75th HD: 2.40mm
100th HD: 0.81mm
12cm2
12cm2
16cm2
15cm2