Technical note: automatic segmentation method of pelvic floor levator hiatus in ultrasound using a self-normalising neural network

Ester Bonmati, Yipeng Hu, Nikhil Sindhwani, Hans Peter Dietz, Jan D'hooge, et al.

Ester Bonmati, Yipeng Hu, Nikhil Sindhwani, Hans Peter Dietz, Jan D'hooge, Dean Barratt, Jan Deprest, Tom Vercauteren, "Technical note: automatic segmentation method of pelvic floor levator hiatus in ultrasound using a self-normalising neural network," Proc. SPIE 10576, Medical Imaging 2018: Image-Guided Procedures, Robotic Interventions, and Modeling, 105760K (13 March 2018); doi: 10.1117/12.2322403

SPIE.

Event: SPIE Medical Imaging, 2018, Houston, Texas, United States

Technical Note: Automatic segmentation method of pelvic floor levator hiatus in ultrasound using a self-normalising neural network

Ester Bonmati[*a,b,c], Yipeng Hu[a,b,c], Nikhil Sindhwani[d], Hans Peter Dietz[e], Jan D'hooge[d], Dean Barratt[a,b,c], Jan Deprest[b,d] , Tom Vercauteren[a,b,c,d]

[a]UCL Centre for Medical Image Computing, University College London, London, UK; [b]Wellcome/EPSRC Centre for Interventional and Surgical Sciences, University College London, London, UK; [c]Department of Medical Physics and Biomedical Engineering, University College London, London, UK; [d]Department of Development and Regeneration, Cluster Urogenital Surgery and Clinical Department of Obstetrics and Gynaecology, University Hospitals Leuven, KU Leuven, Leuven, Belgium; [e]Sydney Medical School Nepean, Nepean Hospital, Penrith, Australia

ABSTRACT

Segmentation of the levator hiatus in ultrasound allows to extract biometrics which are of importance for pelvic floor disorder assessment. In this work, we present a fully automatic method using a convolutional neural network (CNN) to outline the levator hiatus in a 2D image extracted from a 3D ultrasound volume. In particular, our method uses a recently developed scaled exponential linear unit (SELU) as a nonlinear self-normalising activation function. SELU has important advantages such as being parameter-free and mini-batch independent. A dataset with 91 images from 35 patients all labelled by three operators, is used for training and evaluation in a leave-one-patient-out cross-validation. Results show a median Dice similarity coefficient of 0.90 with an interquartile range of 0.08, with equivalent performance to the three operators (with a Williams' index of 1.03), and outperforming a U-Net architecture without the need for batch normalisation. We conclude that the proposed fully automatic method achieved equivalent accuracy in segmenting the pelvic floor levator hiatus compared to a previous semi-automatic approach.

Keywords: levator hiatus, automatic segmentation, SNN, ultrasound, CNN

1. INTRODUCTION

Pelvic Organ Prolapse (POP) is the abnormal downward descent of pelvic organs including, i.e., the bladder, uterus and/or the rectum or small bowel, through the genital hiatus, resulting in a protrusion through the vagina. In a previous study, 27,342 women between the age of 50-79 years were examined and found that about 41% showed some degree of prolapse[1]. At present, ultrasound is the most widely used imaging modality to assess the anatomical integrity and function of pelvic floor, because of availability and non-invasiveness. During a transperineal ultrasound examination, 3D volumes are acquired during Valsalva manoeuvre (act of expiration while closing the airways after a full inspiration), at pelvic floor muscle contraction and during rest. The hiatal dimensions and its area are then recorded by manually outlining the levator hiatus in the oblique axial 2D plane at the level of minimal anterioposterior hiatal dimensions (referred to as the C-plane hereinafter)[2] as these biometrics have been correlated with severity of prolapse, levator muscle avulsion and even prolapse recurrence after surgery[2–4]. The main limitation of this technique is the high variability between operators in assessing the images and the operator time required. Sindhwani et al.[5] earlier proposed a semi-automatic method to segment the levator hiatus in a predefined C-plane which was shown to reduce interoperator variability in comparison to manual segmentation. Overall, despite interesting results, the procedure still lacks of automation, limiting its reproducibility, and requires operator inputs and, consequently, time.

Recently, convolutional neural networks (CNN) have been shown to be able to successfully perform several tasks such as classify, detect or segment objects in the context of medical image analysis[6]. Litjens et al.[7] provide a good review on deep learning in medical image analysis. In this work, we propose a fully automatic method to segment, in manually defined 2D C-planes, the levator hiatus from ultrasound volumes thereby further automating the process of outlining the pelvic floor. In particular, we employ a self-normalising neural network (SNN) using a recently developed scaled exponential

Medical Imaging 2018: Image-Guided Procedures, Robotic Interventions, and Modeling, edited by Baowei Fei, Robert J. Webster III, Proc. of SPIE Vol. 10576, 105760K · © 2018 SPIE
CCC code: 1605-7422/18/$18 · doi: 10.1117/12.2322403

linear unit (SELU) as a nonlinear activation function, with and without SELU-dropout[8], showing competitive results compared to the equivalent network not using SELU. To the best of our knowledge, our work is the first attempt to combine SELU with CNN. SNNs have clear benefits in many medical imaging applications. These include the parameter-free and mini-batch independence nature of SNNs. In deep learning for medical imaging applications, memory constraints are frequently reached during training. Having opportunities to reduce the complexity of the network and being able to use a smaller mini-batch size (in contrast to batch normalisation), without sacrificing the generalisation performance, are both crucial for many applications. Full details of this work can be found in Bonmati et al.[9]

2. METHOD

2.1 Self-normalising neural networks for ultrasound segmentation

In this work, segmenting anatomical regions of interest in medical images is posed as a joint classification problem for all image pixels using a convolutional neural network. Ultrasound images, which contain relatively sparse features that are depth- and orientation-dependent representation of the anatomy, pose a challenging task for traditional CNNs. Therefore, the appropriate regularisation and robustness of the training may be important to successfully segment ultrasound images. In recent years, rectified linear units, has become the *de facto* standard nonlinear activation function for many CNN architectures due to its simplicity and provides partially constant, non-saturating gradient, while batch normalisation, retains a similar importance by effectively reducing the internal variate shift and therefore regularises and accelerates the network training[10]. However, the stochastic gradient descent with relatively small data and mini-batch sizes (commonly found in medical image analysis applications) may significantly perturb the training so that the variance of the training error becomes large. This work explores an alternative construction of the nonlinear activation function used in a self-normalising neural network, a recent development suggesting to use a scaled exponential linear unit (SELU) function[8]. The proposed SELU constructs a particular form of parameter-free scaled exponential linear unit so that the mapped variance can be effectively normalised. As a result, batch-dependent normalisation may not be needed. The SELU activation function is defined as:

$$SELU(x) = \lambda \begin{cases} x & if\ x > 0 \\ \alpha e^x - \alpha & if\ x \leq 0 \end{cases}, \quad (1)$$

where scale $\lambda = 1.0507$ and $\alpha = 1.6733$ (see Klambauer et al.[8] for details on the derivation of these two parameters). This specific form in Equation 1 ensures the mapped variance by the SELU activation is effectively bounded[8] thereby leading to a self-normalising property.

2.2 Network architecture

We adapt a U-Net architecture[11,12] as a baseline CNN to assess the segmentation algorithms. We refer to the proposed self-normalising U-Net-based network as SU-Net hereinafter. The detailed network architecture is illustrated in Figure 1. Each block consists of two convolutions, with a kernel size of 2x2, each followed by a SELU activation. Down-sampling is achieved with a max-pooling with a kernel size of 2x2 and stride 2x2 which halves the sizes of the feature maps preserving the number of channels, while up-sampling doubles the feature map sizes, also preserving the number of channels. Up-sampling is performed by a transposed convolution with a 2x2 stride. After each up-sampling, the feature maps are concatenated with the last feature maps of the same size (before pooling). The last block contains an extra convolution and the corresponding SELU activation. As shown in Figure 2, all the batch normalisation with rectified linear units (ReLU) blocks are replaced by a single SELU activation (described in Section 2.1). For the case of SU-Net with SELU-dropout, the dropout was applied after each convolution. SELU-dropout works with scaled exponential linear units by randomly setting activations to the negative saturation value (in contrast to zero variance in ReLU), in order to keep the mean and variance. The weighted sum of a L2 regularisation loss with of the probabilistic Dice score using label smoothing is used as a loss function[13,14].

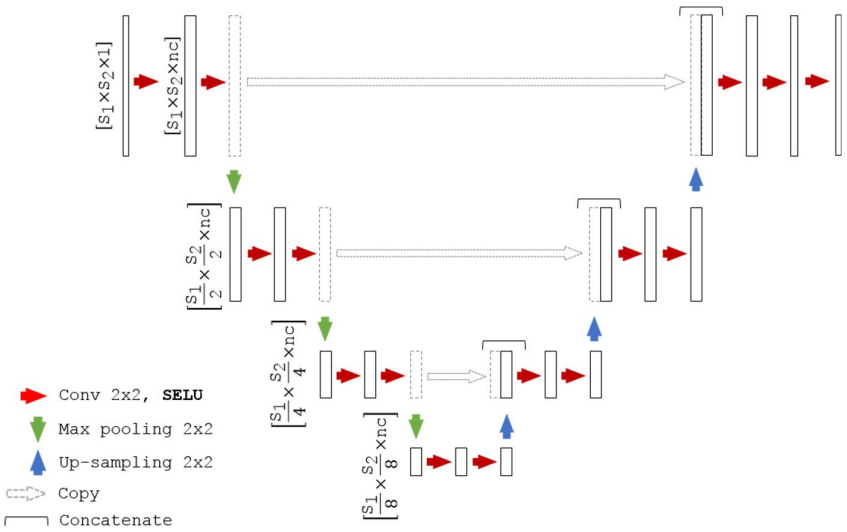


Figure 1. Network architecture, where S1 and S2 correspond to the spatial dimension and nc to the number of channels. For the U-Net, the SELU unit is replaced by batch normalisation and ReLU, and for the U-Net with dilated convolution (U-Net+DC), the last layer is also replaced by a dilated convolution.

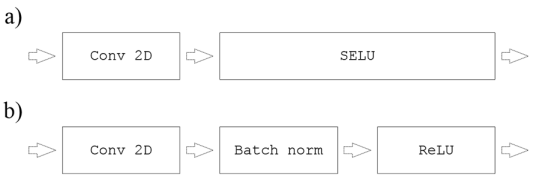


Figure 2. SU-Net architecture (a) versus a U-Net architecture (b).

2.3 Networks evaluation and metrics

Our benchmark include the proposed SU-Net using SELU (SU-Net), the SU-Net also using SELU-dropout (SU-Net+dropout), and a baseline U-Net using batch normalisation and ReLU (U-Net) sharing the same architecture as the SU-Net (Figure 1.). Other hyper-parameters are kept fixed for all these architectures. Additionally, similar to Vigneault et al.[12], we also compare the results with a U-Net in which the last layer convolutions are replaced by dilated convolutions (U-Net+DC), and with a ResNet architecture[15]. Hyper-parameters used in the implementation of the U-Net+DC and ResNet networks are described in Section 3.2. Evaluation is performed in a leave-one-patient-out cross-validation, in which the networks are trained 35 times using data from 34 patients while the contours from the different images of the left-out patient are used in testing. As a result, 91 automatic segmentations are obtained from the 35-fold validation, corresponding to the size of the original dataset.

Results are evaluated using two region-based measures: Dice similarity coefficient[16] and Jaccard coefficient[17], and two distance-based measures: symmetric Hausdorff distance and mean absolute distance (*MAD*). The choice of this comprehensive set of metrics aims to allow direct comparison with the results from a previous study using the same dataset[5]. Additionally, we include two more region-based measures, the false positive Dice (*FPD*) and the false negative Dice (*FND*)[18] and one distance-based measure, the symmetric mean absolute distance (*SMAD*) which is the symmetric version of *MAD*.

2.4 Statistical comparative analysis

Performance is quantified and compared by evaluating the computer-to-observer differences (COD) to determine the agreement between the automatic segmentation and the manual segmentations. A pairwise comparison approach between

each label obtained with the automatic method and the three labels available for each image is performed by considering all the metrics described in Section 2.3. Performance quantification is presented for all network architectures described. Furthermore, statistical analysis employing a paired two-sample Student's t-test is used to test whether the differences in performance between SU-Net and U-Net, U-Net+DC, ResNet and SU-Net+dropout are statistically significant different. Using a similar pairwise approach, interobserver differences (IOD) are quantified to determine the agreement between manual segmentations from the three operators and to allow a further comparison with the automatic methods.

The level of agreement between the automatic and manual segmentations is tested using the extended Williams' index (WI) and the metrics defined in Section 2.3. The WI is a statistical test for numeric multivariate data to test the null hypothesis that the automatic method agrees with the three operators, and that the three operators agree with each other[19,20]. This index quantifies the ratio of agreement by calculating the number of times that the automatic boundaries are within the observer boundaries. If the 95% confidence interval (CI) of the WI contains the value 1.0, it implies that the test fails in rejecting the null hypothesis that the agreement between the automatic method and the three operators is not significantly different.

2.5 Clinical impact

The dimensions of the levator hiatus on ultrasound is a biometric measurement used to assess the status of the levator hiatus, and is associated both with symptoms and signs of prolapse as well as with recurrence after surgical treatment[2]. Therefore, we include the area measurement from the manual and automatic segmentations in our analysis to provide further clinical relevance in assessing the segmentation algorithms. Evaluation is performed by grouping the images in the three different stages: during rest, Valsalva and contraction. WI is used to test the level of agreement between the automatic and manual labels.

3. EXPERIMENTS

3.1 Imaging

A dataset containing 91 ultrasound images, corresponding to the oblique axial plane at the level of minimal anteroposterior hiatal (C-plane), from 35 patients was used for validation[5]. All C-planes were selected by the same operator. The dataset had 35 images acquired during Valsalva, 20 images during contraction and 36 images at rest to cover all the stages during a standard diagnosis with some extreme cases and large anatomical variability. Images had a mean pixel size and standard deviation (SD) of 0.54±0.07 mm, with variable image resolutions ([199-286]×[176-223] pixels, for width and length, respectively). All 91 images were manually segmented by 3 different operators with at least 6 months of experience in evaluating pelvic floor 3D ultrasound images. More details on the dataset can be found in the work of Sindhwani et al.[5].

3.2 Implementation details

All original US images were automatically cropped or padded to 214x262 pixels primarily for normalisation and removing unnecessary background. In training, for the SU-Net and U-Net, we used a mini-batch size of 32 images, we linearly resized the data to 107x131 pixels and used a data augmentation strategy by applying an affine transformation with 6 degrees-of-freedom. The number of channels was fixed to 64. For the SU-Net with SELU-dropout, a dropout rate of 0.5 was used. During training, the images and labels from the three operators were both shuffled before feeding into respective mini-batches. The networks were implemented in TensorFlow[21] and trained with an Adam optimiser[22] with a learning rate of 0.0001, on a desktop with a 24GB NVIDIA Quadro P6000. For each automatic segmentation obtained, post-processing morphological operators to fill holes (i.e., flood fill of pixels that cannot be reached from the boundary of the image) and remove unconnected regions by selecting the region with the largest area, were also applied. For the U-Net+DC and ResNet we used a mini-batch size of 10, 128 initial channels and a learning rate of 0.001 (all the rest of hyper-parameters, pre-processing and post-processing were kept the same).

4. RESULTS

First, using the three manual labels available for each image as a ground truth, we evaluated the performance of the proposed network using the pairwise comparison strategy defined in Section 2.4 with the metrics described in Section 2.3. For comparison purposes, we also report the results obtained with the baseline U-Net architecture, and the U-Net+DC and

ResNet architectures. Median values and interquartile ranges for each metric are shown in Table 1. Statistical analysis comparing the mean values for each image (average of the operators) obtained with the U-Net and the SU-Net showed a statistically significant difference for the Dice, Jaccard, Hausdorff, *SMAD* and *FPD* metrics (*p-values=0.030, 0.022, 0.004, 0.027, 0.031,* respectively), and no significant difference for *MAD* an *FND* metrics (*p-values=0.064, 0.183,* respectively). However, when comparing the values of all metrics using SELU-dropout and without SELU-dropout, no statistically significant difference was found (all *p-values>0.37*). Furthermore, no statistically significant difference was found when comparing the SU-Net and U-Net+DC (all *p-values*>0.30), or when comparing the SU-Net with ResNet (all *p-values*>0.08). Differences between the three operators (i.e., interoperator differences), not considering the automatic segmentations, are reported using the same metrics and shown in Table 2**.** WIs are reported in Table 3 to compare the agreement between automatic and manual segmentations with the agreement among manual segmentations using the metrics described in Section 2.3. Table 4 shows the mean differences in area of the segmented regions in terms of computer-to-operator differences and interoperator differences during the three different stages, and with the corresponding WIs testing the performances.

Table 1. Performance of the SU-Net, SU-Net+dropout, U-Net, U-Net+DC and ResNet networks by employing a pairwise comparison with the three manual labels available for each ultrasound image. This table also contains results from a previous study (Sindhwani et al.[5]). Results are reported using median [interquartile range].

Method	Dice	Jaccard	Hausdorff (in mm)	*MAD* (in mm)	*SMAD* (in mm)	*FPD*	*FND*
SU-Net	0.90 [0.08]	0.82 [0.12]	4.21 [3.92]	1.19 [1.15]	1.16 [1.02]	0.07 [0.13]	0.09 [0.16]
SU-Net+dropout	0.90 [0.08]	0.81 [0.13]	3.90 [3.83]	1.21 [1.16]	1.23 [1.09]	0.07 [0.13]	0.09 [0.16]
U-Net	0.89 [0.11]	0.80 [0.18]	4.49 [5.67]	1.31 [1.42]	1.34 [1.41]	0.07 [0.16]	0.08 [0.16]
U-Net+DC	0.90 [0.08]	0.82 [0.13]	3.97 [3.87]	1.18 [3.86]	1.17 [1.23]	0.05 [0.13]	0.11 [0.15]
ResNet	0.91 [0.08]	0.83 [0.14]	3.59 [4.22]	1.13 [1.14]	1.10 [1.07]	0.06 [0.14]	0.07 [0.13]
Sindhwani et al.[5]	0.92 [0.05]	0.85 [0.09]	5.73 [3.90]	2.10 [1.54]	-	-	-

Table 2. IOD results reported using median [interquartile range].

Dice	Jaccard	Hausdorff (in mm)	*MAD* (in mm)	*SMAD* (in mm)	*FPD*	*FND*
0.92 [0.06]	0.85 [0.10]	3.05 [2.33]	1.01 [0.85]	1.01 [0.81]	0.03 [0.08]	0.08 [0.15]

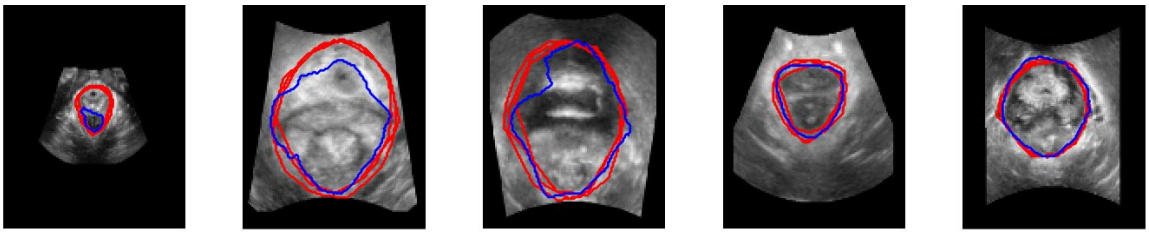
Table 3. WIs [95% CI] for the SU-Net, SU-Net+dropout, U-Net, U-Net+DC and ResNet architectures for each metric.

Method	WI Dice	WI Jaccard	WI Hausdorff (in mm)	*WI MAD* (in mm)	*WI SMAD* (in mm)	*WI FPD*	*WI FND*
SU-Net	1.032 [1.03,1.03]	1.052 [1.05,1.06]	0.677 [0.67,0.69]	0.738 [0.73,0.75]	0.776 [0.77,0.79]	0.425 [0.40,0.45]	0.588 [0.57,0.61]
SU-Net+dropout	1.032 [1.03,1.03]	1.051 [1.05,1.05]	0.701 [0.69,0.71]	0.751 [0.74,0.76]	0.784 [0.77,0.80]	0.420 [0.40,0.44]	0.591 [0.57,0.62]
U-Net	1.085 [1.08,1.09]	1.111 [1.10,1.12]	0.530 [0.52,0.54]	0.577 [0.56,0.59]	0.538 [0.52,0.56]	0.281 [0.26,0.30]	0.439 [0.42,0.46]
U-Net+DC	1.033 [1.03,1.04]	1.053 [1.05,1.06]	0.712 [0.70,0.72]	0.723 [0.71,0.74]	0.756 [0.74,0.77]	0.395 [0.37,0.42]	0.706 [0.69,0.72]
ResNet	1.037 [1.03,1.04]	1.061 [1.06,1.07]	0.717 [0.71,0.73]	0.726 [0.71,0.74]	0.731 [0.72,0.74]	0.533 [0.50,0.57]	0.52 [0.5,0.54]

Table 4. COD and IOD using SU-Net with the corresponding WIs and the 95% CI. Results are reported using mean (±SD).

Stage	Contraction	Valsalva	Rest
COD	0.62±0.91	0.86±1.89	0.60±1.22
IOD	0.52±0.70	0.62±1.03	0.61±0.92
WI [95% CI]	0.80 [0.72,0.89]	0.72 [0.68,0.76]	0.85 [0.80,0.90]

Figure 3 shows examples of original images with the corresponding segmentation results obtained with the automatic method together with the three manual labels used as a ground truth, and Figure 4 shows examples at the three different stages: rest, Valsalva and during contraction.



(a) (b) (c) (d) (e)

Figure 3. Segmentation of the levator hiatus using with the SU-Net architecture (blue) compared with the three manual labels (red) for the following percentiles of the Dice coefficient: 0th (a), 25th (b), 50th (c), 75th (d) and 100th (e).

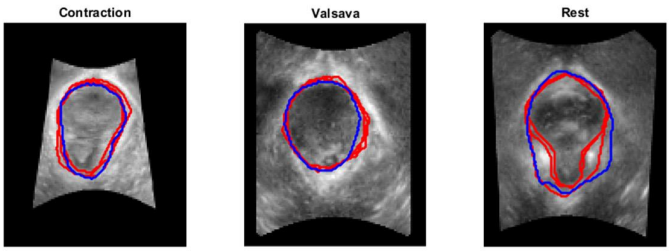


Figure 4. Segmentation examples of the levator hiatus at the three different stages (contraction, Valsalva and rest) using the proposed method (blue) compared to the outlines provided by the operators (red). Cases were chosen at the 75th percentile of the mean Dice coefficient considering the three operators.

5. DISCUSSION AND CONCLUSION

In this work, we have presented a fully automatic method, using a convolutional neural network, to segment the pelvic floor levator hiatus on a 2D image plane extracted from a 3D ultrasound volume. A large number of female patients may potentially benefit globally from this approach. We have adopted a recently proposed self-normalising neural network that replaces the batch normalisation with a SELU unit, which for the first time, has been applied in medical imaging to tackle a clinically important application, obtaining either superior or equivalent segmentation results compared to a number of state-of the-art network architectures with clear additional benefits in terms of complexity and memory requirements. We show that the method presented outperformed the U-Net-based architecture by considering region-based and contour-based metrics and confirmed by statistical tests. On the other hand, no statistical significant difference was found when SELU-dropout, U-Net+DC or ResNet were used. Therefore, SELU can potentially provide equivalent or improved results without the mini-batch size limitation.

Comparing the COD (Table 1) with IOD (Table 2), WIs CIs show that the automatic method strongly agrees with the observers in terms of Dice and Jaccard coefficients with a value very close to 1, but it is not the case for the distance metrics. This result may be due to a disagreement on local parts of the boundaries as shown in Figure 3 (c), which gives a higher Hausdorff distance value, or due to a larger part of the boundary in disagreement with the operators as shown in Figure 3 (b) which results in a higher SMAD value.

As a clinically relevant metric, we evaluated the differences in area at three different stages (contraction, Valsalva and rest). In this case, WIs were smaller than 1, showing some level of disagreement with the operators (Table 4). We believe that the results can be further improved by increasing the number of images during training, as the current dataset size is limited and contains some extreme cases with a high variability. Compared to a previous study[5], in which at least three anatomical points have to be manually identified on the C-plane, we proposed a fully-automatic segmentation algorithm that is able to segment the pelvic floor on the C-plane without operator input of any form, achieving comparable accuracy. Note that, the previous study already achieved competitive results obtaining a good agreement with the three operators and demonstrated to be clinically useful. Furthermore, compared to a solution that requires human interaction (i.e. manual definition of several anatomical landmarks), fully-automatic methods, such as the one proposed in this work, have significant advantages including minimising subjective factors due to intra- and inter observer variations, simpler clinical workflow with minimal uncertainty and quantifiable, repeatable procedure outcome.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Friyan Tuyrel, Dr. Ixora Atan for providing the data and the manual ground truth labels used in this study. This work was supported by the Wellcome/EPSRC [203145Z/16/Z, WT101957, NS/A000027/1]; and the Royal Society [RG160569].

REFERENCES

[1] Hendrix, S. L., Clark, A., Nygaard, I., Aragaki, A., Barnabei, V. and McTiernan, A., "Pelvic organ prolapse in the Women's Health Initiative: gravity and gravidity.," American journal of obstetrics and gynecology **186**(6), 1160–1166 (2002).

[2] Dietz, H. P., Shek, C. and Clarke, B., "Biometry of the pubovisceral muscle and levator hiatus by three-dimensional pelvic floor ultrasound," Ultrasound in Obstetrics and Gynecology **25**(6), 580–585 (2005).

[3] Abdool, Z., Shek, K. L. and Dietz, H. P., "The effect of levator avulsion on hiatal dimension and function," American Journal of Obstetrics and Gynecology **201**(1), 89.e1-89.e5 (2009).

[4] Dietz, H. P., Chantarasorn, V. and Shek, K. L., "Levator avulsion is a risk factor for cystocele recurrence," Ultrasound in Obstetrics and Gynecology **36**(1), 76–80 (2010).

[5] Sindhwani, N., Barbosa, D., Alessandrini, M., Heyde, B., Dietz, H. P., D'Hooge, J. and Deprest, J., "Semi-automatic outlining of levator hiatus," Ultrasound in Obstetrics & Gynecology **48**(1), 98–105 (2016).

[6] Gibson, E., Li, W., Sudre, C., Fidon, L., Shakir, D. I., Wang, G., Eaton-Rosen, Z., Gray, R., Doel, T., Hu, Y., Whyntie, T., Nachev, P., Modat, M., Barratt, D. C., Ourselin, S., Cardoso, M. J. and Vercauteren, T., "NiftyNet: a deep-learning platform for medical imaging" (2017).

[7] Litjens, G., Kooi, T., I.Sánchez, B., Bejnordi, E., Setio, A. A. A., Clara, G., Ciompi, F., Ghafoorian, M., Laak, J. A. W. M. va. der, Ginneken, B. van and Sánchez, C. I., "A survey on deep learning in medical image analysis," Medical Image Analysis **42**, 60–88 (2017).

[8] Klambauer, G., Unterthiner, T., Mayr, A. and Hochreiter, S., "Self-Normalizing Neural Networks," Advances in Neural Information Processing Systems (2017).

[9] Bonmati, E., Hu, Y., Sindhwani, N., Dietz, H. P., D'hooge, J., Barratt, D., Deprest, J. and Vercauteren, T., "Automatic segmentation method of pelvic floor levator hiatus in ultrasound using a self-normalizing neural network," Journal of Medical Imaging **5**(2), 1 (2018).

[10] Ioffe, S. and Szegedy, C., "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift" (2015).

[11] Ronneberger, O., Fischer, P. and Brox, T., "U-Net: Convolutional Networks for Biomedical Image Segmentation" (2015).

[12] Vigneault, D. M., Xie, W., Bluemke, D. A. and Noble, J. A., "Feature Tracking Cardiac Magnetic Resonance via Deep Learning and Spline Optimization" (2017).

[13] Pereyra, G., Tucker, G., Chorowski, J., Kaiser, Ł. and Hinton, G., "Regularizing Neural Networks by Penalizing Confident Output Distributions" (2017).

[14] Milletari, F., Navab, N. and Ahmadi, S.-A., "V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation" (2016).

[15] Li, W., Wang, G., Fidon, L., Ourselin, S., Cardoso, M. J. and Vercauteren, T., "On the Compactness, Efficiency, and Representation of 3D Convolutional Networks: Brain Parcellation as a Pretext Task," Springer, Cham, 348–360 (2017).

[16] Dice, L. R., "Measures of the Amount of Ecologic Association Between Species," Ecology **26**(3), 297–302 (1945).

[17] Jaccard, P., "The distribution of the flora in the alpine zone," New Phytologist **11**(2), 37–50 (1912).

[18] Babalola, K. O., Patenaude, B., Aljabar, P., Schnabel, J., Kennedy, D., Crum, W., Smith, S., Cootes, T., Jenkinson, M. and Rueckert, D., "An evaluation of four automatic methods of segmenting the subcortical structures in the brain," NeuroImage **47**(4), 1435–1447 (2009).

[19] Chalana, V. and Kim, Y., "A methodology for evaluation of boundary detection algorithms on medical images," IEEE Transactions on Medical Imaging **16**(5), 642–652 (1997).

[20] Williams, G. W., "Comparing the joint agreement of several raters with another rater," Biometrics **32**(3), 619–627 (1976).

[21] Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M.,

et al., “TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems” (2016).
[22] Kingma, D. P. and Ba, J., “Adam: A Method for Stochastic Optimization” (2014).