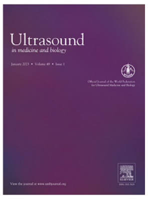
Development of a Deep Learning–Based System for Optic Nerve Characterization in Transorbital Ultrasound Images on a Multicenter Data Set

Ultrasound in Medicine & Biology 49 (2023) 2060−2071

Contents lists available at ScienceDirect

Ultrasound in Medicine & Biology

journal homepage: www.elsevier.com/locate/ultrasmedbio



Original Contribution

Development of a Deep Learning−Based System for Optic Nerve Characterization in Transorbital Ultrasound Images on a Multicenter Data Set

Francesco Marzola [a,*], Piergiorgio Lochner [b], Andrea Naldi [c], Robert Lemor [d], Jakob Stögbauer [b], Kristen M. Meiburger [a]

[a] Biolab, Department of Electronics and Communications, Politecnico di Torino, Torino, Italy
[b] Saarland University Medical Center, Homburg, Germany
[c] Neurology Unit, San Giovanni Bosco Hospital, Turin, Italy
[d] Department of Biomedical Engineering, Saarland University of Applied Sciences, Saarbrücken, Germany

ARTICLE INFO

Keywords:
Deep learning
Segmentation
Optic nerve
Optic nerve sheaths
Ultrasound
Transorbital ultrasound

Objective: Characterization of the optic nerve through measurement of optic nerve diameter (OND) and optic nerve sheath diameter (ONSD) using transorbital sonography (TOS) has proven to be a useful tool for the evaluation of intracranial pressure (ICP) and multiple neurological conditions. We describe a deep learning−based system for automatic characterization of the optic nerve from B-mode TOS images by automatic measurement of the OND and ONSD. In addition, we determine how the signal-to-noise ratio in two different areas of the image influences system performance.
Methods: A UNet was trained as the segmentation model. The training was performed on a multidevice, multicenter data set of 464 TOS images from 110 subjects. Fivefold cross-validation was performed, and the training process was repeated eight times. The final prediction was made as an ensemble of the predictions of the eight single models. Automatic OND and ONSD measurements were compared with the manual measurements taken by an expert with a graphical user interface that mimics a clinical setting.
Results: A Dice score of 0.719 ± 0.139 was obtained on the whole data set merging the test folds. Pearson's correlation was 0.69 for both OND and ONSD parameters. The signal-to-noise ratio was found to influence segmentation performance, but no clear correlation with diameter measurement performance was determined.
Conclusion: The developed system has a good correlation with manual measurements, proving that it is feasible to create a model capable of automatically analyzing TOS images from multiple devices. The promising results encourage further definition of a standard protocol for the automatization of the OND and ONSD measurement process using deep learning−based methods. The image data and the manual measurements used in this work will be available at 10.17632/kw8gvp8m8x.1.

Introduction

Measurement of the optic nerve diameter (OND) and optic nerve sheath diameter (ONSD) using transorbital sonography (TOS) is an effective non-invasive technique for monitoring and identifying several neurologic disorders through the estimation of intracranial pressure (ICP) [1]. The reference tool for ICP monitoring is the invasive ICP evaluation, but this is not a viable technique in emergency settings or for centers without access to neurosurgical tools.

The optic nerve (ON) is enveloped by a meningeal sheath, including the subarachnoid space containing cerebrospinal fluid (CSF). It has been proven that the optic nerve sheath (ONS), because of its elastic properties, can expand in response to progressive increases in ICP. In fact, ONSD has been found to increase simultaneously with increases in ICP [2,3]. Several studies have determined the validity of ONSD assessment in detecting intracranial hypertension (ICH) as compared with invasive ICP measurement [4−6], and recent meta-analyses have confirmed the findings [7]. However, despite the numerous studies conducted in this field, a unique cut-off value for the determination of ICH has not yet been identified. On the other hand, TOS presents several advantages, including the possibility of performing repeated and non-invasive bedside examinations in both regular and intensive care settings [1,8]. Potential clinical applications are numerous and include evaluation of the conditions able to modify the ICP for the ONSD measurement (*i.e.,* traumatic brain injury, cerebral vascular diseases, brain infections, idiopathic intracranial hypertension) [9−11] and the evaluation of multiple sclerosis, optic neuritis and other demyelinating diseases for the OND measurement [8,10,12,13]. Moreover, ONSD and OND can also be measured using pocket-sized devices, expanding the potential field of application of this technique to point-of-care solutions [14].

[*] Corresponding author. Department of Electronics and Communications, Politecnico di Torino, Torino, Italy.
E-mail address: francesco.marzola@polito.it (F. Marzola).

https://doi.org/10.1016/j.ultrasmedbio.2023.05.011
Received 10 December 2022; Revised 16 May 2023; Accepted 21 May 2023

Manual measurements of OND and ONSD achieve relatively good scan−rescan reproducibility and inter- and intra-observer reliability [15]. Still, the acquisition and interpretation of the ultrasound image need to be carried out by an expert operator [16,17]. The standardization of the acquisition and the analysis of TOS images could provide benefits in multiple aspects of the examination. Fixing a range for machine settings (gain, dynamic range, enhancement filters) and probe positioning would cut acquisition time. The acquired images could then be processed automatically to reduce processing time, operator dependency and error. This would enable a more reliable cutoff threshold definition for ONSD to identify increased intracranial pressure. Moreover, more robust studies could be performed to assess interactions between the optic bulb diameter and ONSD and to evaluate the effect of measuring ONSD at different depths.

In recent literature, several methods for the automatic or semi-automatic segmentation of TOS images for OND and ONSD measurement have been presented. Gerber et al. [18] reported a two-step approach to automatically measure the ONS on a 3-D-printed phantom using a portable US probe. First, they located the optic bulb fitting an ellipse in the largest dark area in the image. Then they constructed two boundaries on the walls of the optic nerve acoustic shadow. They measured a correlation >0.8 with novice and expert operators on the measurements from 23 images. In 2019, Soroushmehr et al. [19] proposed a superpixel-based segmentation method to measure the ONSD. They developed their method on 50 videos of 25 patients, with an intra-class correlation coefficient (ICC) of 0.70 with an average of two experts. In 2020, Meiburger et al. [20] developed AUTONoMA on a data set of 75 images, with a correlation >0.6 for ONSD measurement compared with two expert operators. Their method was based on the coarse localization of the ocular bulb through a column-wise search on the output of a Gaussian derivative filter. Then, a similar method was used to locate the ON centerline. Finally, the precise segmentation of the ONS was obtained using a double snake model. In 2021, Stevens et al. [21] described a method based on active contours. They detected the borders of the retrobulbar fat using a signed asymmetry map and used the detected profile as the initialization for the active contour method. They tested this method on 42 images, and comparing the automatic ONSD measurements with two operators, they achieved an R^2 of 0.31 with the first operator and 0.46 with the second operator (first operator vs. second operator $R^2 = 0.63$).

Our group developed the first deep learning (DL)−based approach for this task in 2021 when we proved the advantages of a DL approach with respect to the previously published methods, obtaining a mean absolute error of 0.48 ± 0.48 mm on ONSD measurements [22]. Although the ONSD measurement error was relatively high because of the high variability in the images, it was the first attempt to use a single model to segment images from a multicentric data set. This work also underlined how rule-based methods can tend to fail when applied to heterogenous data sets.

The choice of using a DL-based approach arises from the limitations of rule-based methods that lack generalization capabilities. The TOS acquisition is dependent on the operator, who might select different acquisition parameters to better visualize the ON. The image interpretation system must be robust over these variations; thus, the DL systems have a clear advantage over the rule-based systems for this application. At the same time, to achieve good generalization performance, it is important to train the DL system on a diverse and representative data set similar to the one used in this study that is openly available for further studies.

In this work, we aim to expand our earlier analysis proving the robustness of a DL approach across a challenging data set that includes images from four different machines. We validate our results across multiple runs and hold-out test sets from external machines. In our previous study [22], the reference OND and ONSD were extracted solely from the manual segmentation masks. In this study, the automatic measurements were compared against those from manual segmentation masks and against manual diameter measurements using an *ad hoc* graphical user interface (GUI), representing a more realistic clinical setting. Lastly, we investigate how the signal-to-noise ratio (SNR) influences segmentation and diameter measurement performance. The aim is to guide the definition of a standard acquisition protocol for the creation of a data set for the training of a baseline model for the segmentation of ON structures.

Methods

Database description

A private data set of 464 TOS ultrasound images in .bmp format acquired with four different machines from 110 different subjects was used for this study. A smaller version of this data set was used in our previous study [22] where the DL approach was compared with traditional heuristic-based methods. Two hundred and seven images were acquired using Esaote MyLab Gold 30 and MyLab Seven in Homburg, 113 using Esaote MyLab in Turin, 93 using Toshiba Medical Systems Aplio 300 and AplioXG and 51 using the GE-Vivid7 sonography system. The images were acquired by expert sonographers using linear transducers with frequency ranges of 3−11 and 7.2−14 MHz. The lateral resolution was 0.47 and 0.40 mm at 10 and 15 MHz, respectively, for the Esaote Turin and Homburg devices. This information was not available for the Toshiba Aplio 300, AplioXG and GE-Vivid7 systems. The median conversion factor was 0.1061 mm/pixel, ranging from 0.0527 to 0.1210 mm/pixel. Examples of images from the different data sets are available in Figure 1, and device characteristics are summarized in Table 1. This is a retrospective study; all the image data were acquired during regular clinical practice and de-identified. No further approval was requested by the local institutional review board of the four centers.

Manual and automated OND and ONSD measurement

Data pre-processing and manual segmentation and diameter measurement

Before the manual and automated measurement of OND and ONSD, all the images were automatically cropped. A window of 256 × 256 pixels was automatically centered on the Optic Nerve. The window centering operation was briefly described in [22]. The image was cropped exploiting the intensity values of the pixels and the calibration factor (CF). The row indexes for the first crop were defined at fixed positions corresponding to 15% and 90% of the height of the image to eliminate the auxiliary information on the US acquisitions. Then, the row having the maximum mean intensity value was searched inside this range. The column indexes for the first crop were found as the indexes of the columns delimiting the largest sequence of non-zero pixels in the previously defined row. Results of this step are illustrated in Figure 2a (*green box*). After the first crop, the image was thresholded at a pixel value of 20 (range: 0−255), and the circumference of the optic bulb was detected from the binary map searching for circumferences with a radius in the range of 80 to 120 pixels using the circular Hough transform (CHT) [23]. The lowest point of the detected circumference is the center of the cropping window. The optic bulb detection and square window crop are depicted in Figure 2a.

To obtain the targets of the segmentation network, all the cropped images were manually segmented using the MATLAB ImageLabeler tool. Three labels were assigned: 0 for the background, 1 for the optic nerve sheaths and 2 for the optic nerve. The manual segmentation process included the tracing of the contours of the ONS. Contours were traced only where the sheaths were distinguishable (*i.e.,* there was a good visual SNR between ON and ONS). The ON label was defined as the area between the two ONS labels. It was obtained by creating a mask using the convex hull of the ONS labels. The ONS labels were subtracted from this mask, and the remaining object was labeled as the optic nerve. A screenshot from the segmentation GUI is available in Figure 3. The step is also depicted in the pipeline visualization in Figure 2.

As a means to have a ground truth (GT) for diameter measurement, a simple GUI was developed to measure the OND and ONSD in the 3 mm zone, which resembles customary practice in a clinical setting [24]. To

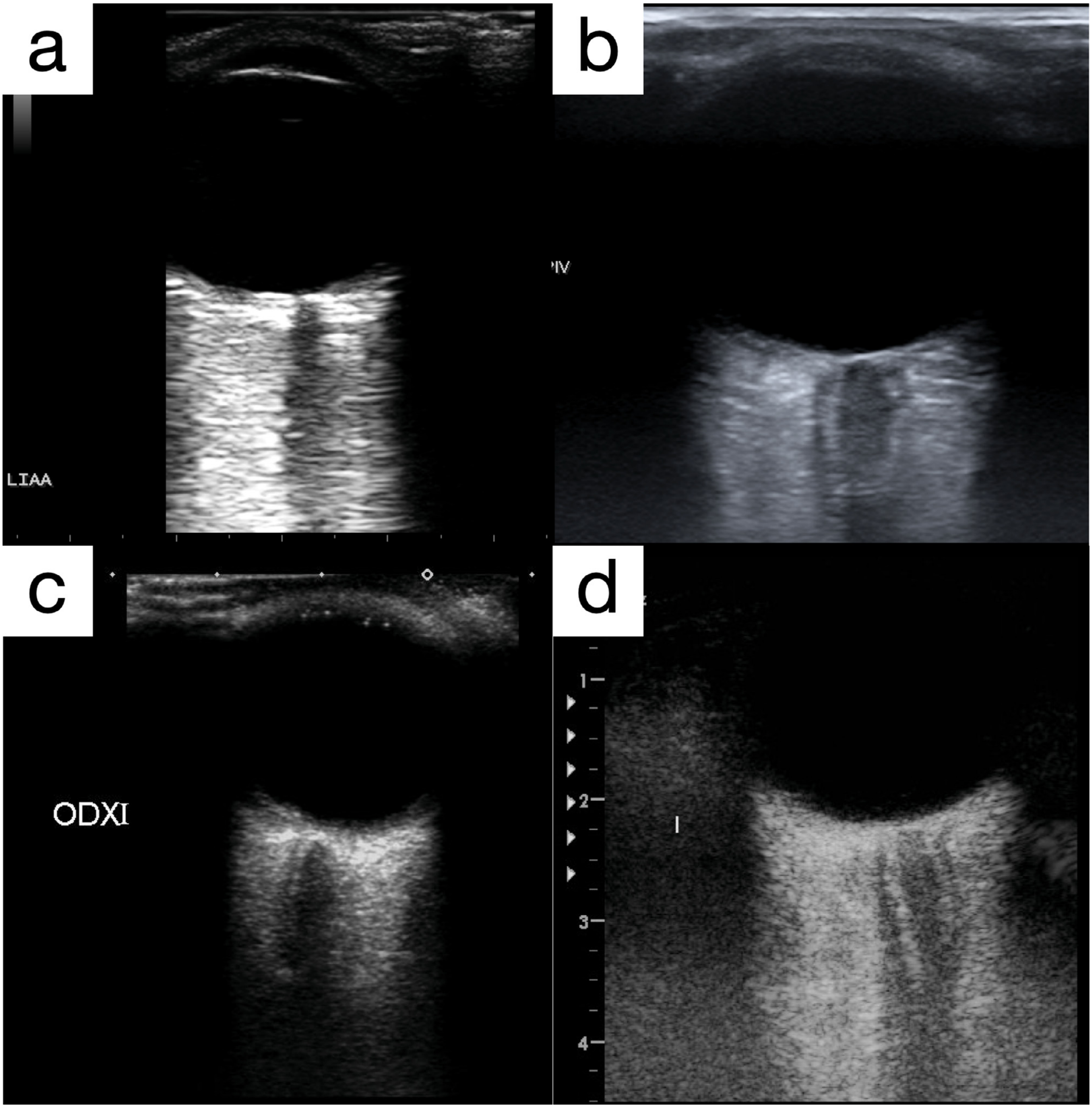


Figure 1. Examples of images available in our data set from different centers and machines. (a) Homburg, Germany (MyLab Gold 30). (b) Turin, Italy (Esaote MyLab). (c) Novara, Italy (Toshiba Medical System Aplio 300). (d) GE-Vivid7.

Table 1
Characteristics of devices

Device	Linear probe frequency (MHz)	Lateral resolution	Conversion factor (mm/pixel)	Numerosity
MyLab Esaote, Homburg	10	0.6 mm (6.6 MHz) 0.47 mm (10 MHz)	0.1029 ± 0.0124	207
MyLab Esaote, Turin	10	0.6 mm (6.6 MHz) 0.47 mm (10 MHz)	0.0848 ± 0.0120	113
Toshiba Aplio300 and AplioXG, Novara	7.2−14	N/A	0.1020 ± 0.0108	93
GE-Vivid7	7−11	N/A	0.1124 ± 0.0097	51

N/A, not available.

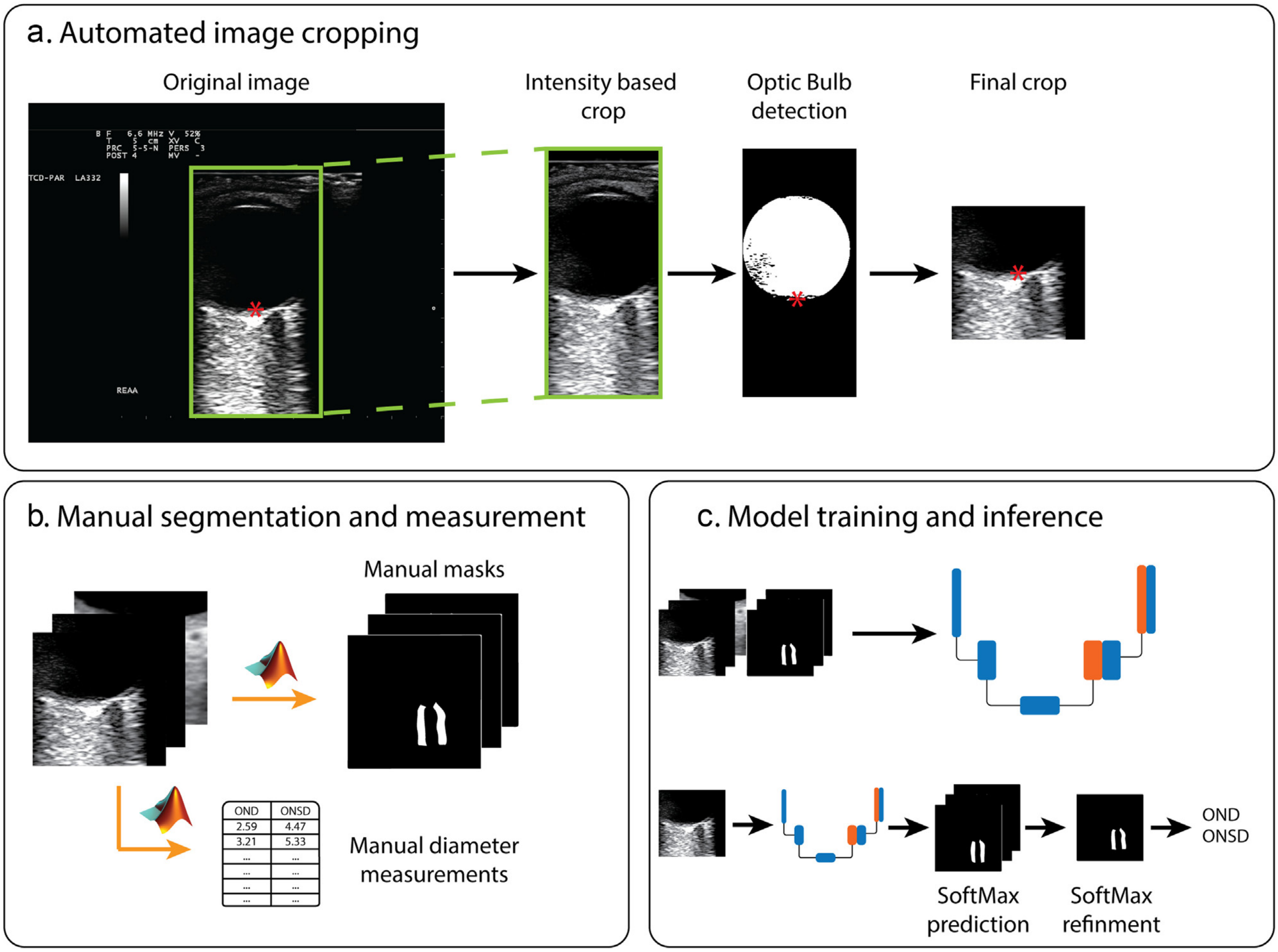


Figure 2. (a) Image cropping process. The first crop is based on pixel intensity, then the optic bulb is detected using the Hough transform and the 256 × 256 is centered on the lowest point of the bulb. (b) Use of two MATLAB-based GUIs to extract manual masks and manual OND and ONSD measurements. (c) Image−mask pairs used to train the UNet and the inference process followed by the post-processing of the eight predictions and the automated diameter measurements. GUI, graphical user interface; OND, optic nerve diameter; ONSD, optic nerve sheath diameter.

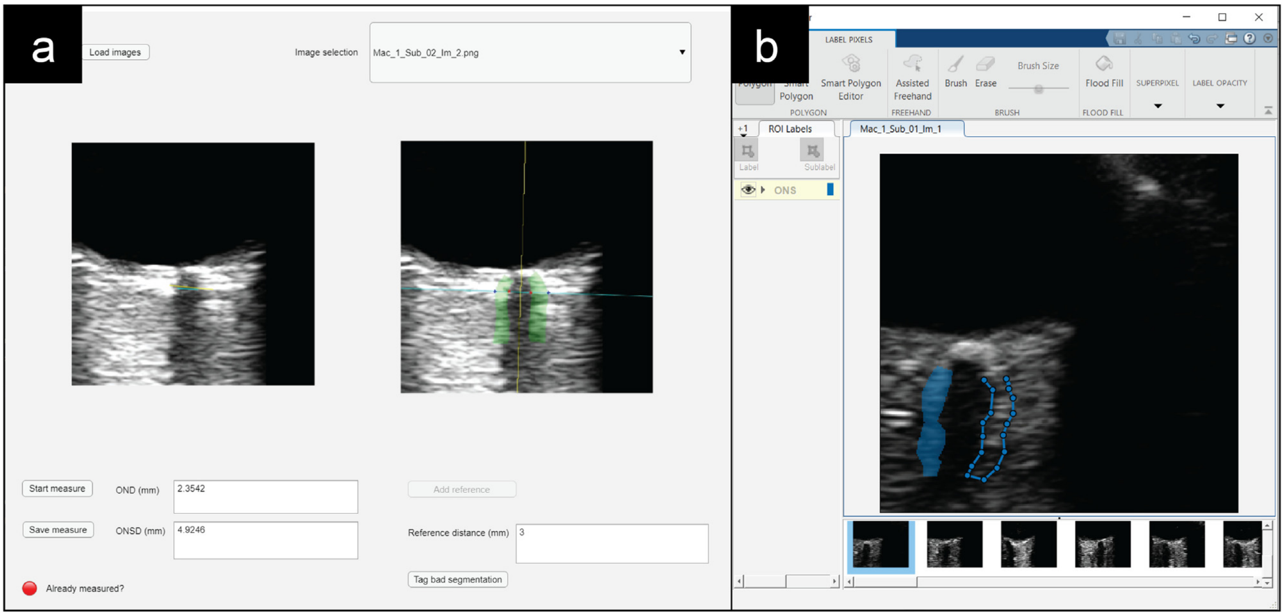


Image selection: Mac_1_Sub_02_Im_2.png

OND (mm): 2.3542

ONSD (mm): 4.9246

Reference distance (mm): 3

Figure 3. (a) Custom user interface for diameter measurement. Left: The operator can measure OND and ONSD by tracking two lines. Right: There is a guide that highlights the 3 mm zone. (b) ImageLabeler User Interface for ONS segmentation. OND, optic nerve diameter; ONSD, optic nerve sheath diameter; ROI, region of interest.

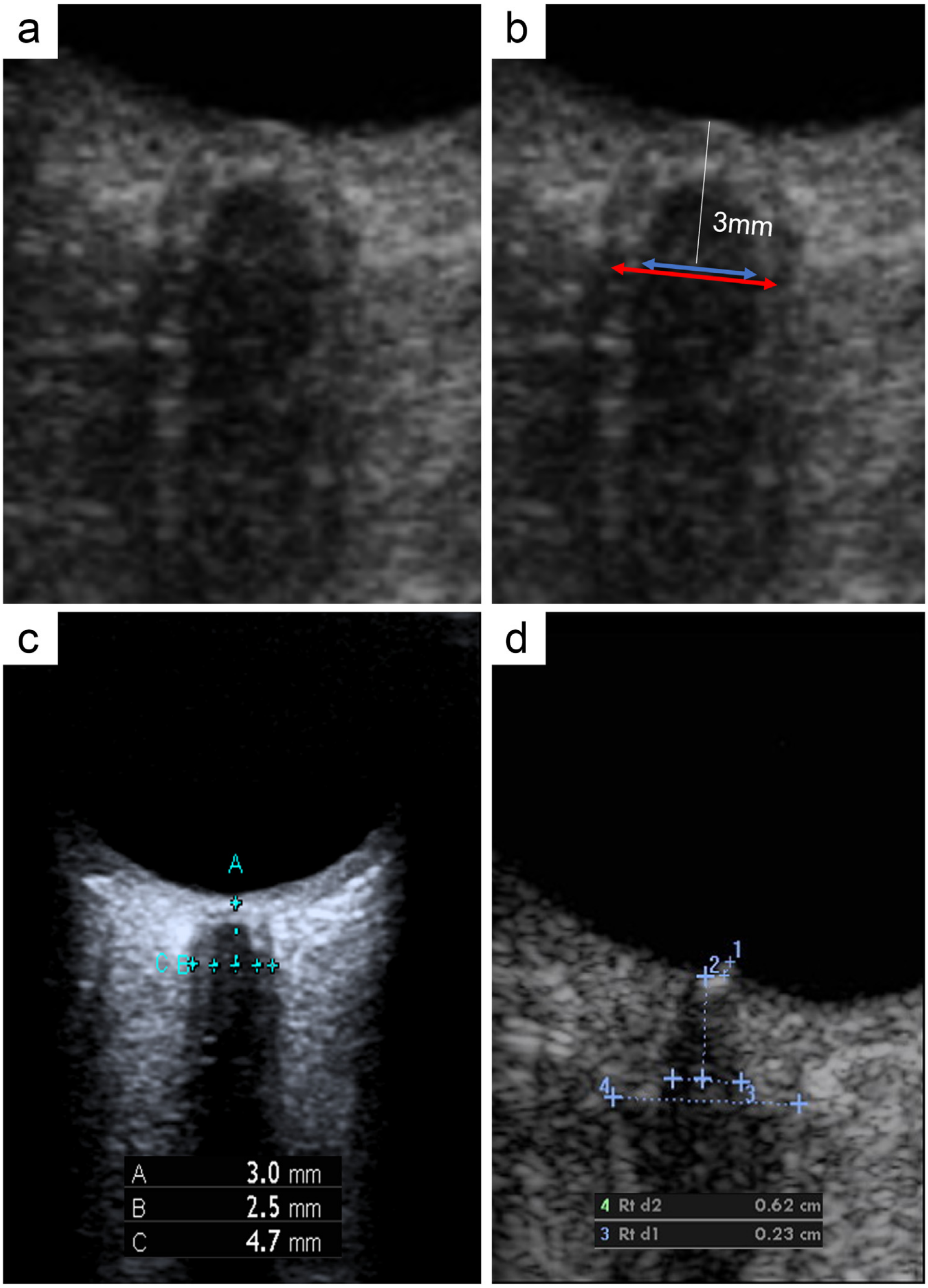


Figure 4. (a) Depiction of the optic nerve and its meningeal sheath as visualized by transorbital sonography. (b) OND (*blue arrow*) and ONSD (*red arrow*) measured 3 mm behind the papilla. (c) Measurements of the OND and ONSD in a healthy participant. (d) Measurements of the OND and ONSD in a patient with idiopathic intracranial hypertension. OND, optic nerve diameter; ONSD, optic nerve sheath diameter.

define the 3 mm zone, a line standing for the axis of the optic nerve was extracted from the ON label. To calculate the ON axis, the ON label was morphologically thinned up to a line 1 pixel wide (*i.e.,* the "skeleton" of the ON label). The ON axis was then calculated as the linear interpolation of the skeleton of the ON label. Using the conversion factor of each image, 3 mm was measured from the point at which the ON axis intersects the circumference of the optic bulb. When using the GUI, the operator is prompted with an image and the annotation of the 3 mm zone; then the operator tracks two straight lines perpendicularly to the ON axis, one to measure the OND and one to measure the ONSD. ONSD was defined as the outer borders of the hyperechogenic area surrounding the ON, corresponding to the subarachnoid space; OND was defined as the distance between the inner borders of the hyperechogenic area surrounding the ON. An example of the standard used for ON and ONSD measurement for healthy and pathological images is available in Figure 4. Measurements were then saved in a text file. The manual measurement of the two diameters was performed by A1 (A.N. with 10+ y of experience in clinical practice of TOS), considered as the ground truth, and by A2 (F.M. with 5 y of experience in ultrasound image analysis, who trained on 25 independent examinations and performed the ONSD measurement in a standardized view of the ONSD) considered as the human control. Each operator was blinded to the other operator's measurements. Two manual diameter measurements were obtained to have an estimate of inter-rater variability.

Manual segmentation of the ON sheaths to create the ground truth for the segmentation network is a time-consuming operation and, thus, was performed only by A2. To assess that manual segmentations by A2 were performed correctly, the OND and ONSD values were extracted from the manual masks and compared with the diameters measured by A1. No outliers were detected (ONSD max absolute error <1.5 mm), so the manual segmentations were considered usable as ground truth for the segmentation network.

The same geometric approach applied to define the 3 mm zone was used for the automatic measurement of OND and ONSD. At the 3 mm mark, a line was traced orthogonal to the ON axis, and the intersections with the ONS mask were marked on the ONS contours, for a total of four points. The two medial marks defined the OND, while the two outer marks defined the ONSD, thus including in the measured diameter also the subarachnoid space. Examples of user interfaces for segmentation and diameter measurement are displayed in Figure 3.

The image data and the manual measurements used in this work will be available at 10.17632/kw8gvp8m8x.1.

Model training and inference

A UNet with five down sampling steps and a Resnet50 encoder pre-trained on ImageNet was used as the segmentation model. The Dice loss was used as the objective function, with Adam optimization (initial learning rate = 0.0002) and a batch size of 32 images. During training, the Dice loss was tracked on the validation set to guide the selection of the best model. The learning rate was halved after 20 and 40 epochs. The training was stopped after 10 epochs without an improvement of at least 1% in the Validation loss with respect to the best epoch. To regularize model training, data augmentation was applied to the input images and masks. Each transformation had a probability (p) of being applied to each batch that was collected from the data set. The transforms were applied as a random permutation of horizontal flipping ($p = 0.5$), scaling and rotation (scale = ±10%, rotate = ±10°, $p = 1$), Gaussian noise addition ($p = 0.3$), contrast-limited adaptive histogram equalization (CLAHE) ($p = 0.3$), multiplicative noise addition ($p = 0.2$), optical distortion ($p = 0.2$), sharpening ($p = 0.25$) and blurring ($p = 0.5$). More details can be found in the shared code. Our approach was developed using Pytorch and with the libraries Albumentations and Segmentation Models [25−27].

The low numerosity of our data set is a limitation when applying a DL model. Hence, a fivefold cross-validation (CV) was applied to have a robust estimate of network performance and generalization. The comparisons to the ground truth were performed merging the five folds selected for testing, thus standing for the full data set. At each run, three subsets were used for training and one each for validation and testing. To improve the segmentation of the ON and ONSD, the fivefold CV was repeated eight times and the SoftMax outputs of the models were averaged to obtain a more precise final Ensemble prediction. This was possible given the fast training time of the models: each run was completed in under 20 min using an RTX 3060. The total training process took 12 h.

To obtain the Ensemble prediction, each SoftMax output was binarized with a threshold set at 0.5 creating segmentation masks. Morphological operations were applied to each mask to fill holes, smooth boundaries (*i.e.,* opening with a disk structural element of radius 3 pixels, followed by a closing with a disk structural element of radius 5 pixels) and retain only the two largest connected objects in the mask.

The eight post-processed masks were averaged, resulting in a probability map. The probability map was binarized with the threshold set at 0.6 (0−1 range), and the same morphological operations applied on the single segmentation masks were used to smooth the predicted area. Sizes of thresholds and structural elements were chosen empirically. All details of the post-processing operations can be found in the shared repository. A detailed illustration of the morphological operations on the SoftMax and their average is provided in Figure 5.

As an added check of the generalization performance of our approach, the training of the UNet model was also performed on the images from three machines using the images from the fourth machine as an external test set. A fivefold CV strategy was used as in the earlier example, but with no repetition; four folds were used for training and one for validation. The test set was defined by the images of the left-out machine. Hence, for each test set, five different predictions were available. The SoftMax predictions from the 5 folds were averaged as previously described.

Model validation

To validate the performance of our model, the Dice score and the Hausdorff distance were calculated between the manual masks and the automatic ONS masks. These comparisons were done only in the image section where the manual masks were defined. This was done because the automatic network segmented the ONS also in the areas with a low SNR while the manual segmentation was limited to areas with a high SNR. Moreover, a paired Wilcoxon test was performed to verify that the distributions of these two metrics in each single run were significantly different from those of the ensemble prediction.

The main goal of this approach was to obtain correct measurements of the OND and ONSD. The automatically obtained diameter measurements were compared with our ground truth (*i.e.,* manual measurements by A1). The comparisons included three different methods: the first compared the diameters measured by another operator (A2) that used the same GUI as the ground truth, the second extracted the diameters from the manual masks (A2$_{\text{mask}}$) and the third measured the diameters from the masks obtained using the network (UNet). The three methods were compared by computing the absolute error defined as

$$\text{Absolute error}_{\text{Diameter}} = \left|\text{Diameter}_{\text{Method}} - \text{Diameter}_{\text{AN}}\right| \quad (1)$$

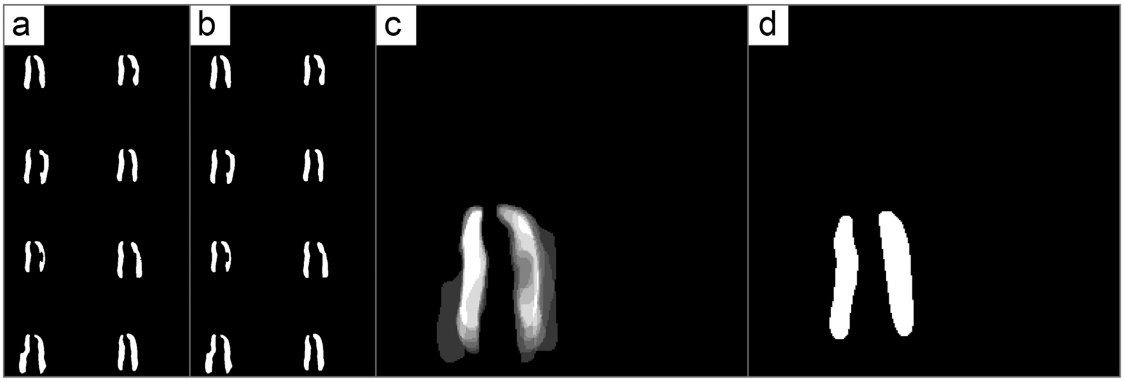


Figure 5. (a) Binarized SoftMax outputs from eight separately trained models. (b) Post-processed SoftMax. (c) Averaging of the 8 SoftMax results in a probability map. (d) Post-processing of the probability map (thresholding at 0.6 and morphological operations).

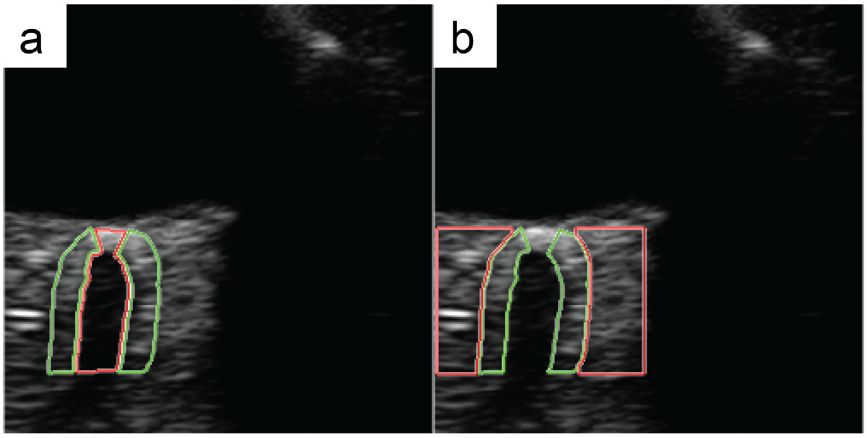


Figure 6. Areas considered during signal and noise power calculation for signal-to-noise ratio evaluation. (a) Signal power (*green label*) is calculated in the optic nerve sheath area, while noise power (*red label*) in the optic nerve area. (b) Signal power is calculated in the optic nerve sheath area, and noise power in the retrobulbar fat area.

the intra-class correlation (ICC2,1), Bland−Altman plots and correlation plots. One of the aims of this work was to give some guidance on the acquisition of TOS images to reach human-level performance by an automated system. To do so, segmentation and diameter measurement performance was compared at different SNR levels. SNR was defined as SNR = $20 \times \log(P_s/P_n)$, where P_s is signal power and P_n is noise power. The SNR was computed in two distinct positions using the labels defined manually by A2. In the first, the signal power was computed inside the ONS labels and the noise power inside the ON label. In the second, the signal power was computed inside the ONS labels, and the noise power was computed in the retrobulbar fat area surrounding the ONS labels. The two different measurements are described in Figure 6. Three categories were established:

- Low SNR: $\text{SNR}_i < \overline{\text{SNR}} - \sigma_{\text{SNR}}$
- Average SNR: $\overline{\text{SNR}} - \sigma_{\text{SNR}} > \text{SNR}_i < \overline{\text{SNR}} + \sigma_{\text{SNR}}$
- High SNR: $\text{SNR}_i > \overline{\text{SNR}} + \sigma_{\text{SNR}}$

Where SNR_i is the SNR measured in the i-th image. $\overline{\text{SNR}}$ and σ_{SNR} are respectively the average SNR and the standard deviation of the SNR in

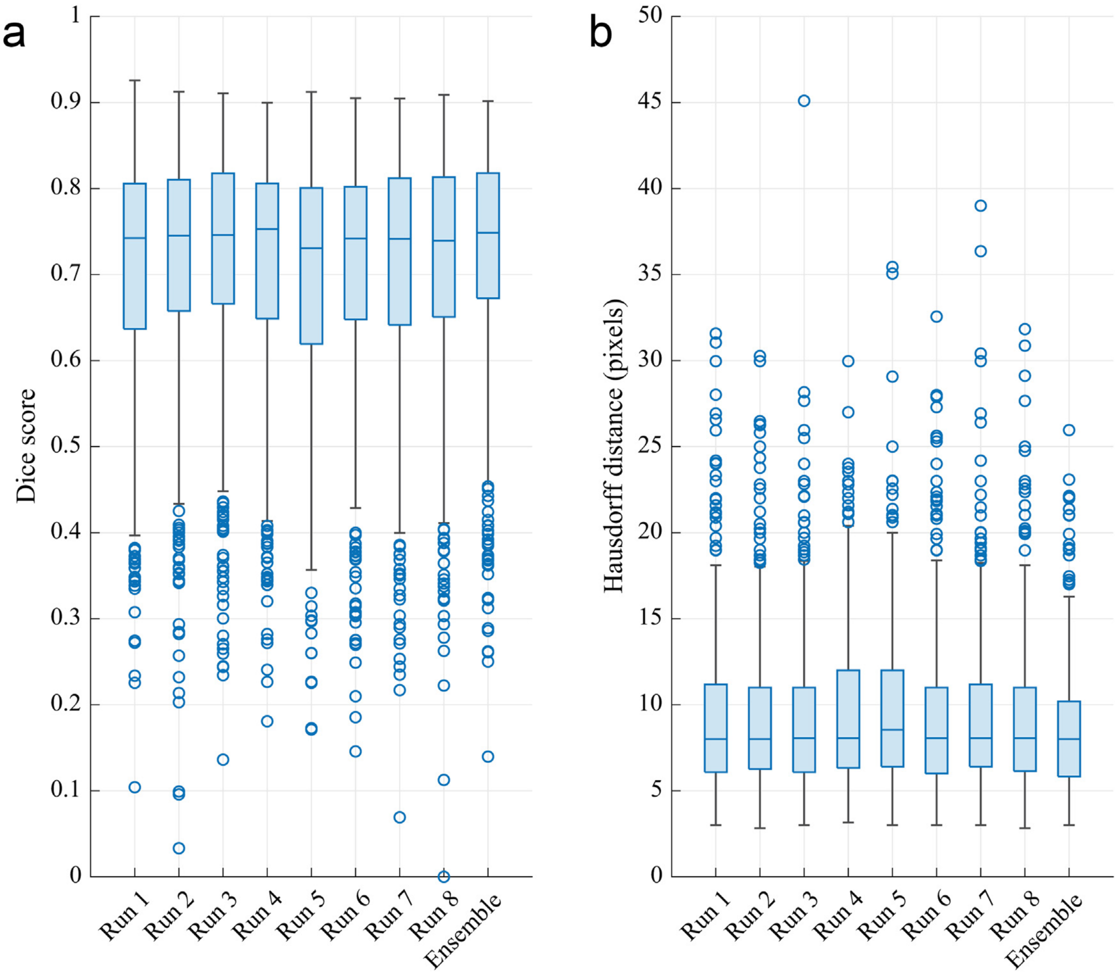


Figure 7. (a) Boxplots of the Dice scores for the eight runs and those resulting from the ensemble of the SoftMax. (b) Boxplots of the Hausdorff distances in pixels between the optic nerve sheath contours of the ground truth and the automated masks for the eight runs and the ensemble mask.

the dataset. SNR impact on segmentation performances was studied by comparing the Dice scores at different SNR levels. To study SNR impact on diameter measurement, the distribution of SNR values was investigated depending on the correctness of the OND and ONSD measurements. Measures within 0.2 mm from GT were labeled as "correct," whereas outside this range measures were labeled as "over" or "under" estimated. The distribution on SNR values for each category was then evaluated.

Results

The average Dice score for each repetition ranged from 0.695 to 0.713, and the standard deviation ranged from 0.141 to 0.147, while the average Dice score for the ensemble of the prediction was 0.719 ± 0.139. Each Wilcoxon test resulted in a significant difference ($p < 0.01$) between the Dice scores from the single run and those of the ensemble predictions. Similarly, the average HD ranged from 8.9 to 9.3 pixels with a standard deviation from 4.0 to 4.4 pixels for each run, while the ensemble prediction produced an HD of 8.2 ± 0.5 pixels, being significantly lower than all the single-run distributions ($p < 0.01$). Boxplots for the Dice score and HD for each run and the ensemble prediction are provided in Figure 7.

Regarding the generalization performance of the network on a different device considered as the test set, the Dice score was 0.658 ± 0.156 for machine 1 (MyLab Esaote Homburg, N = 207), 0.726 ± 0.121 for machine 2 (MyLab Esaote Turin, N = 113), 0.612 ± 0.157 for machine 3 (Toshiba Aplio 300, N = 93) and 0.689 ± 0.112 for machine 4 (GE-Vivid7, N = 51). Boxplots for the Dice score and HD for this analysis are provided in Figure 8.

During the manual diameter measurement process, 45 images were defined by A1 as not usable for measuring the OND and ONSD because of their poor SNR. The numerosity of the excluded images for each device is given in Table 1. All analyses concerning the OND and ONSD measurements are made excluding those images, while they were not excluded in the segmentation performance assessment, as A2 manually annotated all the images. The results for the comparisons of diameter measurements with the ground truth are summarized in Table 2.

OND is slightly underestimated by the automatic measure with a mean error of −0.253 mm (mean absolute error [MAE] = 0.323 ± 0.246 mm, mean square error [MSE] = 0.165 ± 0.273 mm) and a Pearson's correlation of 0.69. The correlation between A2 and A1 was equal to 0.85, whereas the correlation between the A1 measures and the diameters extracted from the manual masks ($A2_{mask}$) was equal to 0.84. The ONSD measure has a much lower mean bias with a mean error of 0.027 mm (MAE = 0.394 ± 0.320 mm, MSE = 0.257 ± 0.426 mm), and its correlation is at the same level as OND with correlations of 0.69 between A1 and UNet, 0.86 between A1 and $A2_{mask}$ and 0.89 between A1 and A2. The Bland−Altman plots do not reveal any clear trend in

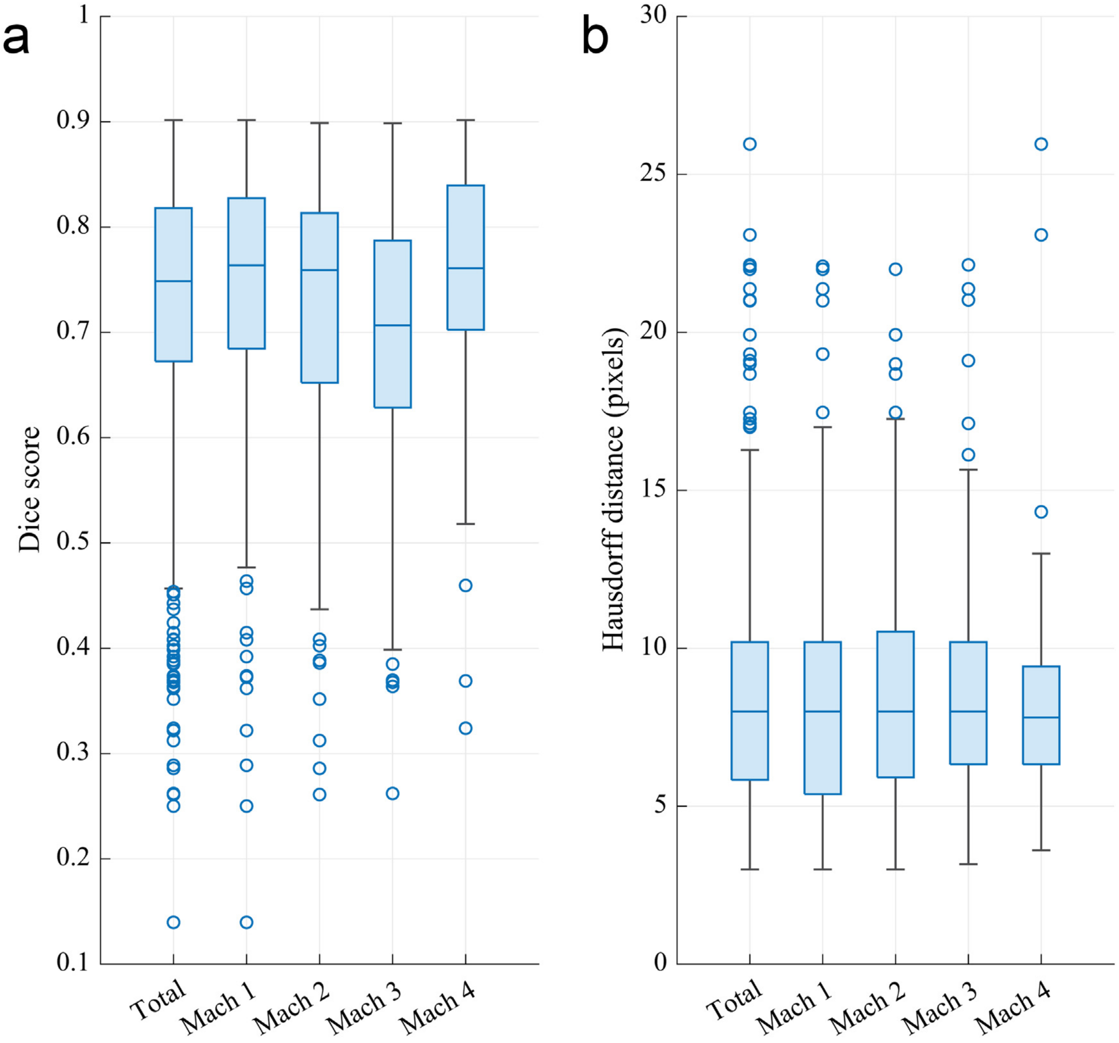


Figure 8. (a) Boxplots of the Dice scores for the subsets when the images from that machine were used as the external test set. The first column is the union of the four runs. (b) Same plot revealing the Hausdorff distances. Mach 1 (MyLab Esaote, Homburg), Mach 2 (MyLab Esaote, Turin), Mach 3 (Toshiba Aplio300 and AplioXG, Novara), Mach 4 (GE-Vivid7). Mach, machine.

Table 2
Evaluation of measurement errors with respect to the ground truth of the compared methods

Pool	Name	OND error (mm)	ONSD error (mm)	OND absolute error (mm)	ONSD absolute error (mm)	iccOND	iccONSD	Low quality
Full	A2$_{\text{mask}}$	−0.193 ± 0.244	0.068 ± 0.368	0.244 ± 0.194	0.288 ± 0.238	0.752	0.861	45
	UNet	−0.253 ± 0.318	0.027 ± 0.507	0.323 ± 0.246	0.394 ± 0.319	0.518	0.661	
	A2	0.051 ± 0.226	0.166 ± 0.337	0.180 ± 0.146	0.293 ± 0.234	0.845	0.860	
Machine 1	A2$_{\text{mask}}$	−0.211 ± 0.249	0.091 ± 0.411	0.260 ± 0.197	0.321 ± 0.272	0.631	0.810	20
	UNet	−0.261 ± 0.291	0.026 ± 0.510	0.324 ± 0.218	0.410 ± 0.303	0.402	0.610	
	A2	0.046 ± 0.226	0.226 ± 0.297	0.178 ± 0.147	0.295 ± 0.229	0.770	0.844	
Machine 2	A2$_{\text{mask}}$	−0.158 ± 0.196	0.117 ± 0.315	0.196 ± 0.157	0.268 ± 0.201	0.840	0.873	8
	UNet	−0.229 ± 0.267	0.083 ± 0.440	0.281 ± 0.212	0.332 ± 0.298	0.618	0.710	
	A2	0.030 ± 0.219	0.161 ± 0.411	0.167 ± 0.144	0.345 ± 0.273	0.877	0.794	
Machine 3	A2$_{\text{mask}}$	−0.196 ± 0.281	−0.023 ± 0.359	0.249 ± 0.234	0.281 ± 0.222	0.752	0.885	15
	UNet	−0.257 ± 0.467	−0.022 ± 0.644	0.413 ± 0.335	0.479 ± 0.428	0.334	0.559	
	A2	0.079 ± 0.275	0.117 ± 0.363	0.225 ± 0.176	0.303 ± 0.229	0.818	0.876	
Machine 4	A2$_{\text{mask}}$	−0.195 ± 0.256	0.033 ± 0.290	0.275 ± 0.166	0.220 ± 0.189	0.653	0.913	2
	UNet	−0.260 ± 0.229	0.026 ± 0.436	0.274 ± 0.211	0.350 ± 0.257	0.503	0.714	
	A2	0.087 ± 0.170	0.047 ± 0.218	0.156 ± 0.110	0.178 ± 0.132	0.856	0.947	

A2, manual diameter measurement from the second operator; A2$_{\text{mask}}$, measurements made on the manual masks. UNet, measurements made on the predicted masks. OND, optic nerve diameter; ONSD, optic nerve sheath diameter; icc, intraclass correlation coefficient; machine 1, MyLab Esaote, Homburg; Machine 2, MyLab Esaote, Turin; Machine 3, Toshiba Aplio300 and AplioXG, Novara; Machine 4, GE-Vivid7.

error distribution for OND measurements. The ONSD measures exhibit a slight overestimation by the automatic method when the manual ONSD increases. Also, the correlation plots do not give rise to doubts about the homoscedasticity of the error deviation for OND and ONSD. These plots are available in Figure 9 for OND and ONSD.

In Figure 10a are boxplots visualizing the Dice score for the left and right ON sheaths and for the total image. Measurements are grouped by SNR category, and the Dice score is shown on the y-axis. In these graphs, we can appreciate how the Dice distribution is influenced by the SNR level, but outliers and errors are still possible in images with high SNR.

Figure 10b compares SNR levels with the correctness of OND measurements. Measurements are grouped by measurement correctness category, and the SNR level is shown on the y-axis. We can see that SNR is distributed similarly for each of the three categories. We can appreciate a slightly higher distribution of SNR values only for correct OND values. Similar considerations arise from Figure 10c, in which the SNR level is compared against the ONSD measurement.

Discussion

The DL-based segmentation method described in this article achieves promising results in the identification of the optic nerve and the optic nerve sheaths in TOS images. From a technical perspective, we introduced a refining operation based on the averaging of the predictions from multiple models and light morphological operations. This operation improves the ability of the system to avoid coarse segmentation errors that are typical when segmenting ultrasound images given their low SNR. This behavior can be visualized as seen in Figure 5, where the outputs from the single networks (Fig. 5a, 5b) can portray non-physiological shapes, blobs and vacancies in the segmented sheaths (*i.e.,* coarse segmentation errors). The averaging step and the final prediction are displayed in panels C and D where the physiological shape of the sheaths is restored, and the segmentation quality is hence improved. This step, along with a fivefold CV, introduces a multiplicative factor in training time but provides beneficial robustness to the segmentation and validation process. Both the CV and the prediction refinement through averaging of multiple outputs could in theory be avoided if the considered data set is large enough, but they were considered necessary given the sample size for this specific study. The OND and ONSD measurements exhibited robust inter-observer agreement in previous literature, and this result was confirmed by our analysis comparing diameters from manual masks with the ground truth measurements. The DL system needs to be improved to reach an accuracy that would bring its performance inside the inter-rater variability range. This achievement has already been described in literature in other clinical applications, where DL-based segmentation networks applied to ultrasound images reach Dice scores >0.8 and have an agreement with the ground truth comparable to intra-operator variability [28−31]. The lower performance compared with literature is explained by the small number of training images and the high difficulty of our data set. It includes data from multiple centers and devices and presents images with ONSD diameters ranging widely from 3.62 to 8.40 mm. So, performance is in line with expectations and can scale accordingly with an augmentation of training data or a decrease in data set variability.

The generalization ability of the network was tested both in cross-validation folds and by leaving one machine out of the training process. The first gave stable results on all folds across multiple repetitions, so we can say that the network can robustly segment images from the same distribution of training data. The second task of generalization across ultrasound devices was the most challenging, producing reduced performance especially for machine 3. This result illustrates how this method can generalize over different machines, but it needs to be tuned on a single machine to reach a robust agreement with an expert operator. This is owing to the different settings and image enhancement algorithms that are present in machines from different vendors that affect the general appearance and texture of the US image.

The correlation and Bland−Altman plots reveal a good correlation between the automatic OND and ONSD measurements and the ground truth. Still, the correlation coefficients are lower than those between ground truth and manual measurements from another operator and the ones obtained from the manual masks. This result is in line with the segmentation performance analysis and highlights the limitations of this approach. This limitation is confirmed also by the ICC values in Table 2 that indicate how the automated measurements still have a lower consistency with ground truth with respect to another manual operator.

Reaching consistency in diameter measurement is especially challenging because of the millimeter/pixel conversion factor and lateral resolution in US images. In our data set, the median conversion factor is 0.1061 mm/pixel, and the absolute error standard deviation for ONSD for UNet measurements is 0.507 mm. Hence, an error of a few pixels can hinder the possibility of obtaining an accurate measurement. The lateral resolution of the ultrasound acquisitions ranged between 0.47 and 0.60 mm; thus, the ONSD absolute error standard deviation lies in this range. The uncertainty level of this approach is thus approaching the physical limitations of US. To reach better accuracy levels, improvements in both the ultrasound signal and image processing are required. For the signal processing and image reconstruction step, new DL-based methods are being explored [32], as are super-resolution ultrasound techniques [33].

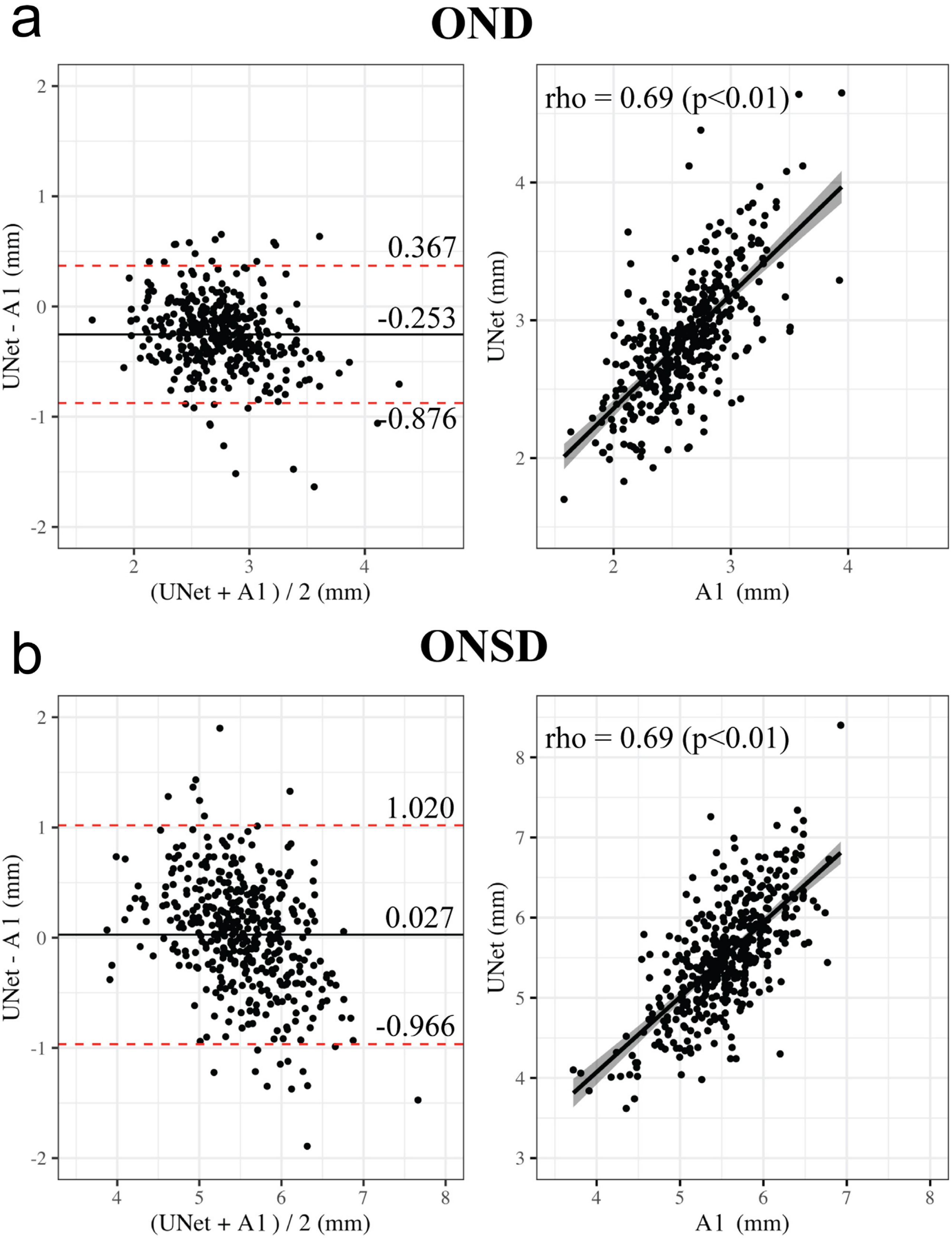


Figure 9. (a) Bland−Altman and correlation plots comparing OND measurements between ground truth and those made using the automated method. (b) Same graphs for ONSD. On the Bland−Altmann plot, the *red dashed lines* are at 1.96 SD from the mean value. In the correlation plots, rho is calculated using Pearson's correlation. A1, manual measurements obtained by the most experienced operator; OND, optic nerve diameter; ONSD, optic nerve sheath diameter.

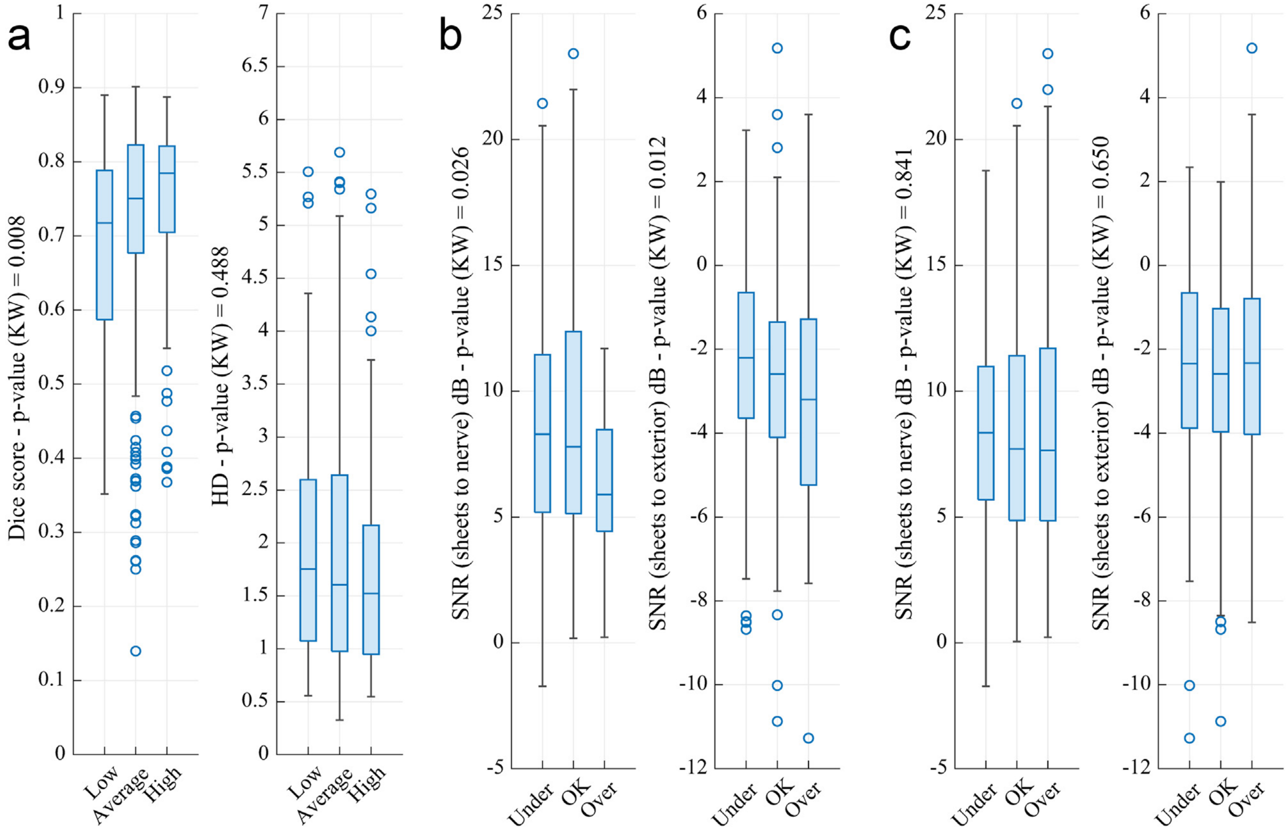


Figure 10. (a) Boxplots visualizing Dice score and Hausdorff distance distributions for distinct categories of SNR. (b) These boxplots visualize SNR distribution depending on the correctness of the OND measurement (GT ± 0.2 mm). SNR is calculated first from the ON sheaths to the ON (left), then from the ON sheaths to their peripheral area (right). (c) SNR distribution depending on the correctness of ONSD measurement. ON, optic nerve; OND, optic nerve diameter; ONSD, optic nerve sheath diameter; SNR, signal-to-noise ratio.

From the image processing standpoint, a possible solution for this problem and future development could include artificially augmenting the resolution of the images using a super-resolution algorithm based on GANs or diffusion models.

The obtained Dice score and OND/ONSD correlation analysis gave rise to similar conclusions. It is important to highlight how both these aspects need to be separately evaluated when assessing an automated system for ON and ONS analysis. The Dice score is better suited when comparing different segmentation approaches on the same data set. Conversely, to assess the effectiveness of the system in performing the clinical task, the correlation analysis with one or more manual operators is to be preferred. The SNR analysis reveals a correlation between SNR level and segmentation performances, but not with diameter measurement accuracy. This is explainable by considering that segmentation accounts for a wider area and ON sheaths are less and less distinguishable at higher acquisition depths. So, a high SNR helps in the detection of the sheaths, especially in the most challenging areas. Instead, the 3 mm zone lies in an area with relatively high SNR, so its effect is less prominent for automatic diameter accuracy. Moreover, it is important to have a high SNR between the ONS and the ON, but this is not true when studying the SNR between the ONS and its peripheral structures (*i.e.,* retro-orbital fat). This is so because of the appearance of these structures in a TOS image where ON and ONS have a higher contrast compared with ONS and retro-orbital fat, which have a similar echogenicity. Thus, to maximize the performance of an automated algorithm, the operators that perform the acquisition should focus on the ON being less echogenic than the ONS, and the ONS having a clear boundary with its peripheral structures. Developing a system that accurately segments the optic nerve sheaths in all their length might also make it possible to study the variability of the diameter of the ON and ONS width depending on the measurement depth.

As a final remark, having a larger or less variable data set should improve both segmentation and diameter measurement accuracy. A larger number of samples would improve the generalization capabilities of this method, enhancing the development of an accurate baseline model that could be transferable across multiple devices and acquisition protocols. Then, tuning the model to a less variable data set, obtained with a standardization of the acquisition process, should additionally improve the system's accuracy. The combination of these two improvements would help to create a system that reliably performs the segmentation of the ONS and the measurement of the OND and ONSD.

Conclusions

We have described a DL approach to the automatic segmentation of ON structures in TOS images with a novel refinement step that involves multiple predictions to avoid coarse segmentation errors. Our approach was validated in two highly challenging scenarios using hold-out test sets on external machines, proving the robustness of our approach. Moreover, the measurements extracted using our method correlate both with those measured by an expert operator and with the measurements extracted from the manual masks.

Focusing on the relationship between SNR and segmentation performances, a first step was made in the direction of setting up a protocol for the acquisition of TOS images that can be easily interpreted by a DL system. With these guidelines, an operator could focus on the correct acquisition of the US image without having to perform an error-prone task such as diameter measurement.

Future work will be in the direction of increasing data set size to develop a robust baseline model that can be fine-tuned on a single machine to achieve human-level precision in ON segmentation. With a larger data set, a statistical analysis to see how each machine setting influences the segmentation capabilities of the DL system would be possible, thus improving the explainability of the DL model.

Overall, this work is a firm statement of the possibility of investigating TOS images with a DL system and should encourage further research toward the creation of a model that can be exploited by practitioners to improve and standardize ON analysis.

Conflict of interest

The authors declare no competing interests.

Data availability statement

The image data and the manual measurements used in this work will be available at 10.17632/kw8gvp8m8x.1.

References

[1] Lochner P, Czosnyka M, Naldi A, Lyros E, Pelosi P, Mathur S, et al. Optic nerve sheath diameter: present and future perspectives for neurologists and critical care physicians. Neurol Sci 2019;40:2447–57.

[2] Liu D, Kahn M. Measurement and relationship of subarachnoid pressure of the optic nerve to intracranial pressures in fresh cadavers. Am J Ophthalmol 1993;116:548–56.

[3] Helmke K, Hansen HC. Fundamentals of transorbital sonographic evaluation of optic nerve sheath expansion under intracranial hypertension: II. Patient study. Pediatr Radiol 1996;26:706–10.

[4] Robba C, Cardim D, Tajsic T, Pietersen J, Bulman M, Donnelly J, et al. Ultrasound non-invasive measurement of intracranial pressure in neurointensive care: a prospective observational study. PLoS Med 2017;14:e1002356.

[5] Moretti R, Pizzi B, Cassini F, Vivaldi N. Reliability of optic nerve ultrasound for the evaluation of patients with spontaneous intracranial hemorrhage. Neurocrit Care 2009;11:406–10.

[6] Geeraerts T, Launey Y, Martin L, Pottecher J, Vigué B, Duranteau J, et al. Ultrasonography of the optic nerve sheath may be useful for detecting raised intracranial pressure after severe brain injury. Intensive Care Med 2007;33:1704–11.

[7] Sallam A, Alkhatip AAAM, Kamel MG, Hamza MK, Yassin HM, Hosny H, et al. The diagnostic accuracy of noninvasive methods to measure the intracranial pressure: a systematic review and meta-analysis. Anesth Analg 2021;132:686–95.

[8] Soliman I, Johnson RJ, Gillman LM, Zeiler FA, Faqihi F, Aletreby WT. New optic nerve sonography quality criteria in the diagnostic evaluation of traumatic brain injury. Crit Care Res Pract 2018;2018:3589762.

[9] Lochner P, Fassbender K, Knodel S, Andrejewski A, Lesmeister M, Wagenpfeil G, et al. B-Mode transorbital ultrasonography for the diagnosis of idiopathic intracranial hypertension: a systematic review and meta-analysis. Ultraschall Med 2019; 40:247–52.

[10] Lochner P, Fassbender K, Andrejewski A, Behnke S, Wagenpfeil G, Fousse M, et al. Sonography of optic nerve sheath diameter identifies patients with middle cerebral artery infarction at risk of a malignant course: a pilot prospective observational study. J Neurol 2020;267:2713–20.

[11] Stead GA, Cresswell FV, Jjunju S, Oanh PKN, Thwaites GE, Donovan J. The role of optic nerve sheath diameter ultrasound in brain infection. eNeurologicalSci 2021; 23:100330.

[12] Carraro N, Servillo G, Maria Sarra V, Bignamini A, Pizzolato G, Zorzon M. Optic nerve and its arterial−venous vascularization: an ultrasonologic study in multiple sclerosis patients and healthy controls. J Neuroimaging 2014;24:273–7.

[13] Schroeder C, Katsanos AH, Ayzenberg I, Schwake C, Gahlen A, Tsivgoulis G, et al. Atrophy of optic nerve detected by transorbital sonography in patients with demyelinating diseases of the central nervous system. Eur J Neurol 2020;27: 626–32.

[14] Johnson GG, Jelic T, Derksen A, Unger B, Zeiler FA, Ziesmann MT, et al. Accuracy of optic nerve sheath diameter measurements in pocket-sized ultrasound devices in a simulation model. Front Med 2022;9:831778.

[15] Bäuerle J, Schuchardt F, Schroeder L, Egger K, Weigel M, Harloff A. Reproducibility and accuracy of optic nerve sheath diameter assessment using ultrasound compared to magnetic resonance imaging. BMC Neurol 2013;13:187.

[16] Zeiler FA, Ziesmann MT, Goeres P, Unger B, Park J, Karakitsos D, et al. A unique method for estimating the reliability learning curve of optic nerve sheath diameter ultrasound measurement. Crit Ultrasound J 2016;8:9.

[17] Copetti R, Cattarossi L. Optic nerve ultrasound: artifacts and real images. Intensive Care Med 2009;35:1488–9.

[18] Gerber S, Jallais M, Greer H, McCormick M, Montgomery S, Freeman B, et al. Automatic estimation of the optic nerve sheath diameter from ultrasound images. Imaging Patient Cust Simul Syst Point Care Ultrasound 2017;10549:113–20.

[19] Soroushmehr R, Rajajee K, Williamson C, Gryak J, Najarian K, Ward K, et al. Automated optic nerve sheath diameter measurement using super-pixel analysis. Annu Int Conf Eng Med Biol Soc 2019;2019:2793–6.

[20] Meiburger KM, Naldi A, Michielli N, Coppo L, Fassbender K, Molinari F, et al. Automatic optic nerve measurement: a new tool to standardize optic nerve assessment in ultrasound B-mode images. Ultrasound Med Biol 2020;46:1533–44.

[21] Stevens RRF, Huberts W, Gommer ED, Ertl M, Aries M, Mess WH, et al. An automated algorithm for optic nerve sheath diameter assessment from B-mode ultrasound images. J Neuroimaging 2021;31:724–32.

[22] Meiburger KM, Naldi A, Marzola F, Lochner P. Automatic segmentation of the optic nerve in transorbital ultrasound images using a deep learning approach. Proc 2021 IEEE Int Ultrason Symp 2021:1–4.

[23] Rizon M, Yazid H, Saad P, Shakaff AYM, Saad AR, Sugisaka M, et al. Object detection using circular Hough transform. Am J Appl Sci 2005;2:1606–9.

[24] Geeraerts T, Merceron S, Benhamou D, Vigué B, Duranteau J. Non-invasive assessment of intracranial pressure using ocular sonography in neurocritical care patients. Intensive Care Med 2008;34:2062–7.

[25] Collobert R, Kavukcuoglu K, Farabet C. Torch7: A Matlab-like Environment for Machine Learning. In: Proceedings, BigLearn, NIPS Workshop. 25th Annual Conference on Neural Information Processing Systems; December 12−17, 2011.

[26] Buslaev A, Iglovikov VI, Khvedchenya E, Parinov A, Druzhinin M, Kalinin AA. Albumentations: fast and flexible image augmentations. Information 2020;11:125.

[27] Iakubovskii P. Segmentation Models Pytorch. Github Repository. Github, 2019. segmentation-models-pytorch · PyPI.

[28] Ilesanmi AE, Chaumrattanakul U, Makhanov SS. A method for segmentation of tumors in breast ultrasound images using the variant enhanced deep learning. Biocybernet Biomed Eng 2021;41:802–18.

[29] Poudel P, Illanes A, Sheet D, Friebe M. Evaluation of commonly used algorithms for thyroid ultrasound images segmentation and improvement using machine learning approaches. J Healthc Eng 2018;2018:8087624.

[30] Marzola F, van Alfen N, Doorduin J, Meiburger KM. Deep learning segmentation of transverse musculoskeletal ultrasound images for neuromuscular disease assessment. Comput Biol Med 2021;135:104623.

[31] Meiburger KM, Marzola F, Zahnd G, Faita F, Loizou CP, Lainé N, et al. Carotid Ultrasound Boundary Study (CUBS): technical considerations on an open multi-center analysis of computerized measurement systems for intima−media thickness measurement on common carotid artery longitudinal B-mode ultrasound scans. Comput Biol Med 2022;144:105333.

[32] Luijten B, Chennakeshava N, Eldar YC, Mischi M, van Sloun RJG. Ultrasound signal processing: from models to deep learning. Ultrasound Med Biol Volume 2023; 49:677–98.

[33] Christensen-Jeffries K, Couture O, Dayton PA, Eldar YC, Hynynen K, Kiessling F, et al. Super-resolution ultrasound imaging. Ultrasound Med Biol 2020;46: 865–91.