DOI:
10.1016/j.bspc.2022.103724

Document Version
Peer reviewed version

Title page

Title

Deep Learning-based Quality-controlled Spleen Assessment from Ultrasound Images

Authors

Zhen Yuan[1], Esther Puyol-Antón[1], Haran Jogeesvaran[2], Nicola Smith[2], Baba Inusa[2], Andrew P. King[1]

Author addresses

[1]School of Biomedical Engineering and Imaging Sciences, King's College London, London, UK

[2]Evelina London Children's Hospital, Guy's and St Thomas NHS Foundation Trust, London, UK

Correspondence information

Corresponding author: Zhen Yuan

Email address: zhen.1.yuan@kcl.ac.uk

Telephone number: +44(0)7707631673

Address: 5th floor, Becket House, 1 Lambeth Palace Rd, London, UK, SE1 7EU

Abstract

Objective

Splenomegaly (abnormal splenic enlargement) is a potentially life-threatening condition that occurs in a range of clinical scenarios, including in patients suffering from Sickle cell disease (SCD). Therefore, spleen size assessments from ultrasound imaging are commonly performed in SCD clinics, and typically involve measuring the length of the spleen. However, the current workflow is prone to intra- and inter-observer variability and is dependent on the experience of the sonographer. Our objective was to automate the spleen length measurement process.

Methods

Two deep learning-based approaches were investigated to achieve automated spleen length measurement from ultrasound images. One is a segmentation-based approach, where we trained a modified U-Net to obtain a spleen segmentation and then applied post-processing to measure the spleen length from the segmentation. The second approach is based on direct regression of spleen length. We also incorporated a quality control (QC) model to help less experienced sonographers ensure the quality of ultrasound images before measurement.

Results

Our best model (segmentation-based approach) reached a mean percentage length error (MPLE) of 4.58% on good quality images, which is within the range of human expert

inter-observer variability (5.78%). After including bad quality images, the incorporation of the QC step resulted in a significant reduction in MPLE (from 5.76% to 4.88%).

Conclusion

Automated, quality-controlled spleen length measurement from ultrasound has been achieved with human-level accuracy.

Significance

This proposed framework has the potential to assist in making robust and accurate assessments of the spleen, especially in settings where there is a lack of experienced sonographers.

Keywords

Deep Learning, Spleen Ultrasound Image, Segmentation, Splenomegaly, Sickle Cell Disease

1. Introduction

1.1. Background

Splenomegaly refers to abnormal enlargement of the spleen, and is associated with a range of clinical conditions, such as infection [1], liver disease [2] and haematological diseases [3]. Sickle cell disease (SCD) is a genetic disorder of haemoglobin which causes the red blood cells to become abnormally sickle-shaped [4-6]. The spleen is acknowledged as the first organ to become damaged by sickle cell anaemia, and splenomegaly is frequent among children with SCD [7]. Sudden enlargement of the spleen (acute sequestration) can be an intractable and life-threatening condition without suitable treatment. Therefore, the size of the spleen is typically measured at routine clinical appointments for SCD patients.

Manual palpation, ultrasound imaging and computed tomography (CT) imaging can all be used to detect splenomegaly [3]. However, manual palpation is coarse and non-quantitative, and it only allows clinicians to form a preliminary judgement as to whether there is splenic enlargement and if a further examination is required. Compared to CT, ultrasound is portable, less expensive and does not involve ionizing radiation, therefore it is more commonly used in clinics for making quantitative measurements of spleen size. During ultrasound examinations, the length of the spleen has been the most frequently used surrogate for spleen size since it correlates well with splenic volume [8]. However, ultrasound imaging is subject to artefacts and poor signal-to-noise ratio. Measuring the length of the spleen from ultrasound images requires sonographer

expertise and is prone to inter-observer and intra-observer variability [9]. Furthermore, in Sub-Saharan Africa and the Middle East, where there is a high prevalence of splenomegaly caused by SCD [10,11], there can be a shortage of experienced sonographers to conduct this task.

The clinical benefits of using an automatic tool for spleen length measurement from ultrasound include speeding up the workflow, alleviating inter- and intra-observer variability and reducing the need for skilled sonographers. To the best of our knowledge, the only prior work to have attempted automatic spleen assessment from ultrasound images was our preliminary work in [12].

1.2. Related Works

Deep learning is a widely used technique in computer vision and image analysis that uses artificial neural networks to learn hidden features from raw data and performs analysis tasks automatically. Convolutional Neural Networks (CNNs) are one type of artificial neural network that allow for deeper networks due to the reduced number of parameters in each layer of neurons. In the medical field, CNNs have been broadly studied and have achieved good results in tasks such as classification, detection and segmentation for images of the breast [13], cardiovascular system [14], fetus [15], liver [16] and other organs. Specific examples from the field of ultrasound image analysis include [17], which applied the GoogLeNet CNN to breast ultrasound images to classify benign and malignant lesions. In [18], Li et al. added a spatial constrained layer to a Fast R-CNN [19] and gained encouraging results on the automatic detection of thyroid

papillary cancer regions from ultrasound images. Zhang et al. [20] proposed a regression CNN to predict fetal head circumference directly from ultrasound images. In [21], the EchoNet model was proposed to produce heart segmentations and quantify cardiac function from echocardiography. In [22], a regression CNN was used to estimate brain maturation from 3D fetal neurosonography images. [23] developed a model based on ResNet to estimate kidney function and evaluate the state of chronic kidney disease from ultrasound images.

A range of CNN architectures have been proposed for medical image analysis. Ronneberger et al. [24] proposed the U-Net CNN architecture, which is an encoder-decoder network with skip connections, achieving state-of-the-art results for the ISBI challenge in 2015. [25] and [26] proposed solutions based on the U-Net for segmentation in breast and fetal ultrasound images, respectively. Though there were a range of later adaptations, such as introducing residual blocks [27] and dense blocks [28] to the classic U-Net architecture, [29] suggested that with careful pre-processing and hyperparameter tuning, U-Net remained a very competitive network for medical image segmentation. Although automatic spleen segmentation has been attempted from CT images and magnetic resonance images (MRI) [30], automatic interpretation of splenic ultrasound images remains challenging due to the relatively poor image quality.

1.3. Contributions

In this paper, we propose a framework for the automatic assessment of spleen ultrasound images, expanding upon our previous work [12] in a number of ways. First,

we extend the analysis to a larger dataset. Second, we conduct more comprehensive experiments for the proposed CNN-based frameworks. Finally, we consider how our method could be incorporated into clinical workflows and propose an automated quality control (QC) step to ensure image quality before analysis. The main contributions of this paper are as follows:

(1) We investigate a range of models tailored to address a specific clinical problem – measuring spleen length from ultrasound images.

(2) We propose an automated method based on the U-Net to fully segment the spleen from ultrasound images. For measuring the length from the obtained segmentations, we investigate three post-processing methods.

(3) Two regression CNNs are investigated to estimate spleen length from ultrasound images directly, which allow end-to-end learning without the need for manual ground truth segmentations. In addition, transfer learning is studied by transferring the U-Net contracting path weights from the segmentation network.

(4) A QC network is developed to ensure the quality of the ultrasound images for the subsequent automated length measurement.

(5) Using our framework, we achieve human expert level on the task of spleen length measurement. To the best of our knowledge, this is the first work to automate the segmentation and the quantification of spleen from ultrasound images.

2. Materials and Methods

Section 2.1 describes the dataset we used to develop and evaluate our methods, including details of pre-processing steps that were applied to the acquired data before their use in developing the QC system and automated spleen length estimation techniques. In section 2.2, we describe the QC model for classifying good and bad quality images with the aim of improving the robustness of the length estimation. Two types of approach, including three different models, were investigated for estimating the spleen length, and these are described in section 2.3. A conceptual overview of the full pipeline, including the QC system and the automated length measurement, is shown in Figure 1.

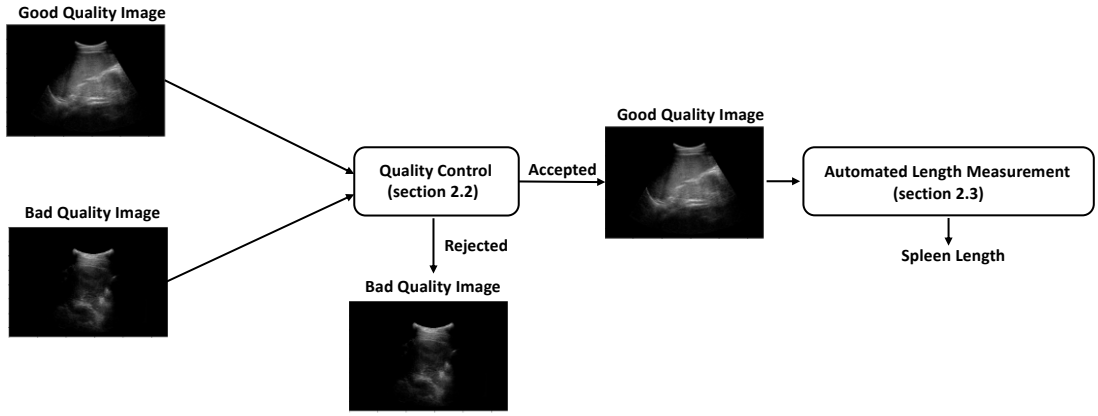


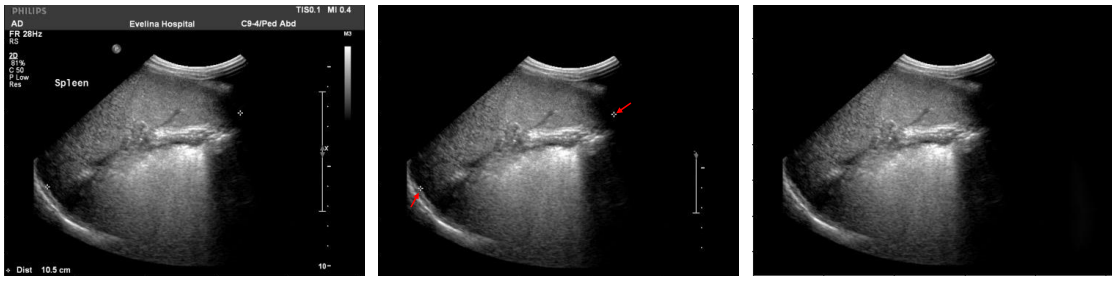
Figure 1: A conceptual overview of the full pipeline of our proposed approach. The quality control system screens out bad quality images in which length measurements would be unreliable, and the automated length measurement model estimates spleen length from good quality images. Refer to the indicated section numbers for further details of the individual steps.

2.1. Dataset and Pre-processing

The study was conducted on a total of 475 2D ultrasound images from a cohort of 200 patients (aged 0 to 18). All patients were children with SCD, who were under professional disease management at Evelina London Children's Hospital of Guy's and St Thomas' NHS Foundation Trust. They underwent multiple ultrasound examinations under the professional consultancy, and for each examination, 1 ultrasound image was recorded for spleen length measurement. Up to 4 images were stored for each patient, each acquired at an examination that occurred at a different time. Note that the spleen shape and length varied significantly between visits as patients grew and clinical conditions changed. Ultrasound imaging was carried out on Philips IU22, Philips EPIQ 7 (Philips Healthcare, Eindhoven, Netherlands), GE Healthcare Logiq E9 (GE Healthcare, Chicago, Illinois, United States) and Hitachi HI VISION Preirus (Hitachi, Tokyo, Japan) ultrasound machines. During the imaging process, an experienced sonographer placed the probe on the left intercostal of the patient and rotated the transducer in the coronal plane until the longitudinal view of the spleen was obtained. The length of the spleen was then estimated by measuring the distance between the superior and inferior points of the spleen, which were manually marked by the sonographer and appeared as crosses on the stored images (see Figure 2).

After obtaining the original images, they went through a series of pre-processing steps to enable them to be used for training our CNN models. First, the ultrasound images were converted from DICOM to NIFTI files, and then masks were manually defined to remove the unnecessary information outside of the ultrasound field of view.

To remove the manually annotated crosses on the superior and inferior points of the spleen, an image inpainting algorithm proposed by Telea [31] was employed for each image. Finally, the spleen was manually segmented by a trained observer. Figure 2 illustrates the pre-processing steps. All masks and segmentations were obtained using ITK-SNAP [32].



(a) (b) (c)

Figure 2: Visualisations of ultrasound images. (a) sample original ultrasound image; (b) same image after removal of the information outside the ultrasound field of view; (c) same image after inpainting to remove the annotations inside the field of view. The red arrows in (b) point to the superior and inferior points of the spleen annotated as crosses by the sonographer.

The processed images were then reviewed by two experienced sonographers. 55 images were labelled as bad quality images due to artefacts obscuring the spleen shape and boundary. The length measurements obtained from these images were mainly based on the prior experience of the sonographer. Therefore, these measurements were likely to be inaccurate compared to the measurements made from the other images. The remaining 420 images were marked as good quality images and used for developing and evaluating our automated length measurement models. Figure 3 shows examples of good quality and bad quality images. All 475 images were used for training and evaluating our QC system. In addition to the length measurement by the original

sonographer, two additional sonographers from Evelina London Children's Hospital were asked to make retrospective length measurements from 108 good quality images to allow quantification of inter-observer variability.

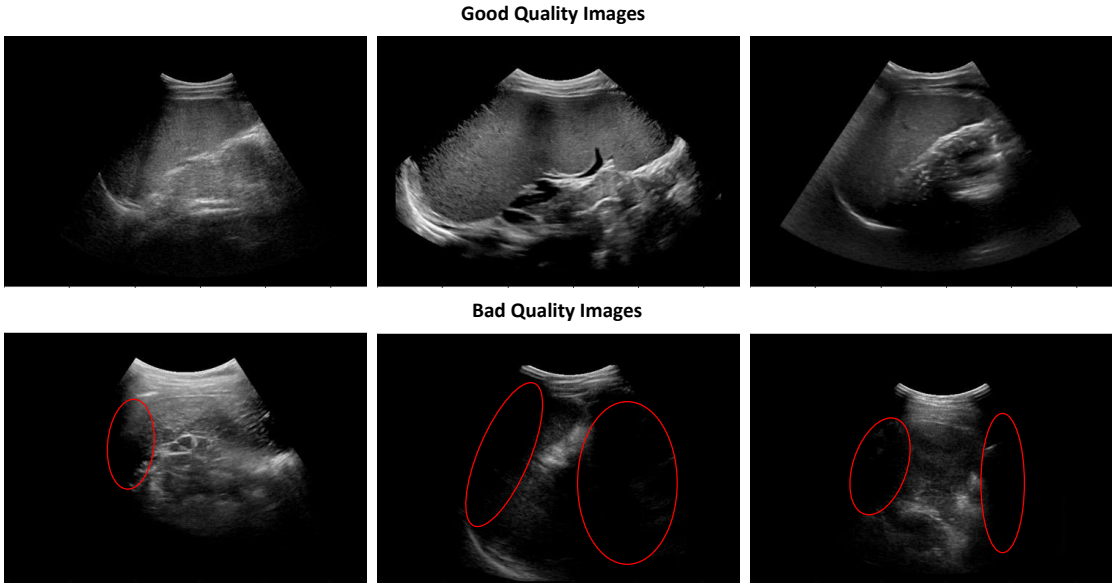


Figure 3: Visualisations of good quality and bad quality ultrasound images. The red circles in the bad quality images indicate the lack of visibility of the spleen and its boundary.

2.2. Quality Control System

An automated QC system was developed to enable a complete clinically usable pipeline for spleen length measurement. A ResNet18 [33] was adopted for the QC model to classify good quality and bad quality images. We added a fully connected layer with size 512 as an output layer after the global average pooling layer, followed by a Sigmoid activation function. We used the binary cross-entropy objective function when training the model. The good quality images were labelled as positive samples, while the bad quality images were labelled as negative samples. Receiver Operating Characteristic

(ROC) curve analysis was employed to identify the optimal threshold for classification. The implementation details are given in section 3.

2.3. Length Measurement

We investigated two different types of approach for spleen length measurement from ultrasound images in this study. The first approach was based on the use of a segmentation model (U-Net) followed by automated length measurement from the resulting segmentations (see section 2.3.1). The second approach was based on regression models to perform direct estimation of spleen length from the images (see section 2.3.2). All measurements were made in millimetres using the pixel size stored in the DICOM headers. The implementation details are given in section 3.

2.3.1. Segmentation-based Approach

We used a U-Net [24] as the backbone of our segmentation model due to its powerful capabilities in the medical image segmentation domain. To allow for learning of deeper features, we modified the original U-Net by adding an extra down-sampling block and up-sampling block to further compress the feature maps in the latent space (see Figure 4). We used the ReLU activation function after each convolutional layer and the Sigmoid activation function after the last convolutional layer to classify each pixel as either background or foreground. Each convolutional layer was followed by batch normalisation. The Dice objective function was used to penalise the difference between the predicted labels and the ground truth labels.

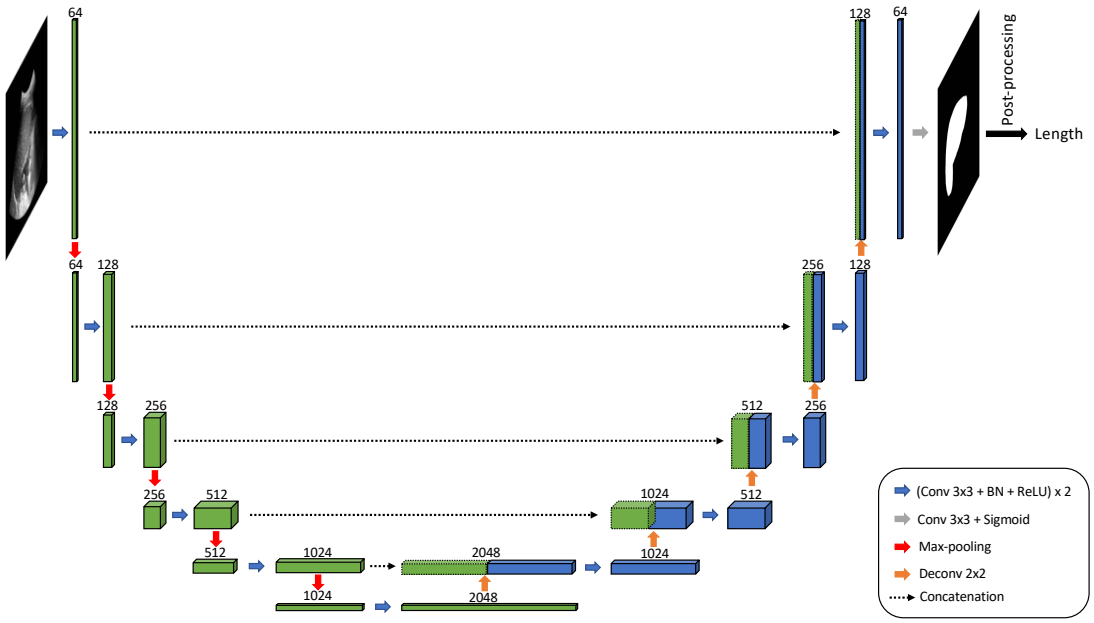


Figure 4: Diagram showing the architecture of the modified U-Net. After the segmentation is obtained, post-processing is applied to estimate the spleen length (see section 2.3.1).

Based on the predicted segmentation from the modified U-Net, we performed connected components analysis (CCA) to retain the largest connected region of the segmentation. Then we investigated three different methods for length measurement from the predicted segmentation. The first method was to calculate the longest distance between pairs of points on the segmentation boundary (LDP). The second one was based on principal components analysis (PCA). We performed PCA on the coordinates of the identified spleen pixels to identify the longest axis and then projected all spleen pixels onto this axis. The length was computed as the range of the projections. Figure 5 illustrates how PCA and LDP measure the length from the predicted segmentation. The last method used a VGG-19 [34] regression network to directly estimate length from the segmentation (Post VGG-19).

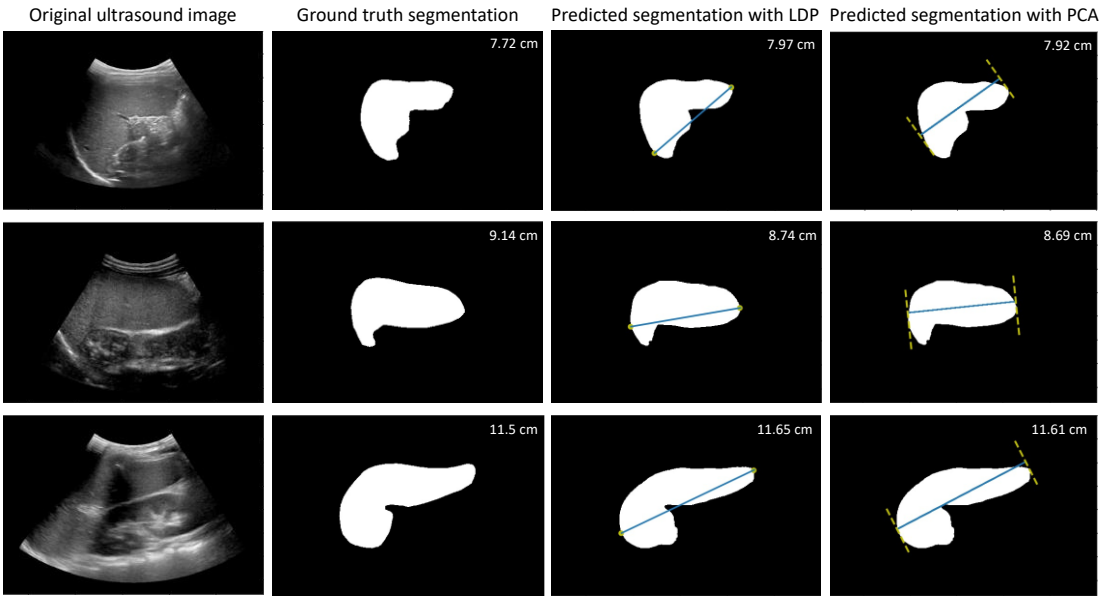


Figure 5: Visualisations for different post-processing length measurement techniques. The pictures from the left to the right column are the original ultrasound images, ground truth segmentations for the spleen, predicted segmentations with LDP length measurements and predicted segmentations with PCA length measurements. The ground truth lengths and the predicted lengths are in the top right corner of the ground truth segmentations and the predicted segmentations.

2.3.2. Regression-based Approach

We investigated two different models for direct regression of spleen length from ultrasound images (see Figure 6). In the first model, we used the standard regression model VGG-19 to perform the task. After the last max-pooling layer, we flattened the feature maps in the latent space and added two fully connected layers. We empirically set the number of nodes in the fully connected layers to 256. Each fully connected layer was followed by batch normalisation. The output layer followed the last fully connected layer to predict the length from the features.

The second model was based on the encoding path of the U-Net from section 2.3.1. To enable a fair comparison with the VGG-19, we added the same fully connected layers and output layer as VGG-19 after the final convolutional layer of the U-Net encoder, followed by batch normalisation. We trained the network in two settings: with and without weight transfer from the U-Net trained for the segmentation task. Note that the decoding path of the U-Net was not included in this model.

The mean squared error (MSE) was adopted as the loss function for both regression models.

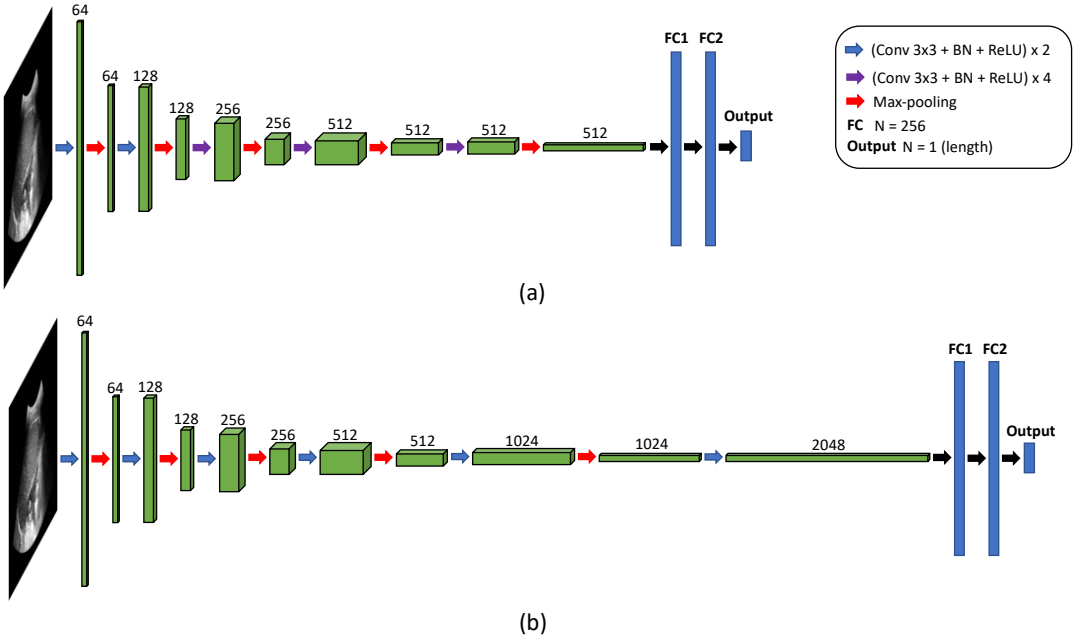


Figure 6: The direct regression CNNs. (a). VGG-19; (b). U-Net contracting path.

3. Experimental Details

A total of 420 good quality 2D ultrasound images were used in training and evaluating the length measurement models. We randomly split the data into sets of 252, 84 and 84 for training, validation and test, respectively. Because the different images acquired

from each subject all had widely varying appearances (i.e., probe angle, field of view, spleen size and shape), we did not enforce subject-level stratification in the training/validation/test split. For the segmentation-based approach, all images were resampled to ensure a consistent pixel spacing. We then calculated the bounding box for each ground truth segmentation and cropped all images and segmentations based on the largest bounding box across the dataset. For the regression-based approach, a centre-cropping method was applied to remove the unnecessary information. All images were cropped to 764×1112 pixels (for both segmentation- and regression-based approaches). The cropped images were subsampled by a factor of 2 and zero-padded to keep the consistency of the size of the feature maps in the down-sampling and up-sampling paths. The size of the pre-processed images used as input to the models was therefore 384×576 pixels. Data augmentation was applied to all images, consisting of spatial transformation (0 to 20-degree random rotations) and intensity transformations, including adaptive histogram equalisation and gamma correction. We also applied Z-Score normalisation to each image after augmentation. For the direct regression method, the length annotations were expressed in pixels and scaled (divided by 50) before using them for training/validation. Hyperparameter optimisation was performed for learning rate (values 10^{-3}, 10^{-4} and 10^{-5}) and weight decay (values 10^{-4}, 10^{-5}, 10^{-6} and 10^{-7}) using the validation dataset. Based on this, we set the learning rate to 10^{-3} and weight decay to 10^{-5}, and the Adam optimiser was adopted for training all models. The minibatch size was set to 4. We trained all models for up to 1000 epochs.

420 good images and 55 bad images selected by experienced sonographers were used to develop and evaluate the QC system. We combined the test dataset from the length measurement experiments (84 good quality images) with 11 bad quality images to form the held-out test set for the QC experiments. 4-fold nested cross-validation was applied to the remaining 336 good quality images and 44 bad quality images for hyperparameter optimisation and model performance evaluation. Based on this, we set the learning rate to 10^{-6} with learning rate decay and a minibatch size of 8 (7 good quality images and 1 bad quality image in each minibatch) to train the final model using all 336 good quality images and 44 bad quality images. The same data augmentation methods as were used for the length measurement models were applied to train the classifier. We trained all models for up to 600 epochs.

The proposed models were implemented in PyTorch and trained using an NVIDIA TITAN RTX (24GB).

4. Results

4.1. Evaluation Metrics

In this paper, to evaluate the performance of all proposed models (segmentation-based and regression-based) in estimating spleen length, we used the mean percentage length error (MPLE) and the Pearson's correlation coefficient (R) as performance metrics. These are defined as:

$$MPLE = 100\% \times \frac{1}{n}\sum_{i}^{n}\frac{|l_i' - l_i|}{l_i} \quad (1)$$

$$\rho_{L,L'} = \frac{cov(L,L')}{\sigma_L \sigma_{L'}} \quad (2)$$

where $L' = \{l'_1, l'_2, l'_3, \dots, l'_n\}$ stands for the estimated length, $L = \{l_1, l_2, l_3, \dots, l_i\}$ is the ground truth lengths and n denotes the number of test images.

We adopted different quantitative metrics to evaluate the performance of our models for the segmentation and regression tasks, respectively. To measure the performance of the segmentation-based model, we first computed Dice similarity coefficient and intersection over union (IoU), which both evaluate the overlap between the predicted segmentation A and the ground truth segmentation B. These are defined as:

$$D(A,B) = \frac{2|A \cap B|}{|A| + |B|} \quad (3)$$

$$IoU(A,B) = \frac{|A \cap B|}{|A \cup B|} \quad (4)$$

Both measures have values from 0 to 1, where 0 denotes no overlap between the two segmentations and 1 denotes perfect overlap.

We also computed the Hausdorff distance (HD) to evaluate the maximum distance between the predicted segmentation contour C_A and the ground truth segmentation contour C_B, defined as:

$$HD(C_A, C_B) = max\left\{\max_{a \in C_A} \min_{b \in C_B} \|a - b\|, \max_{b \in C_B} \min_{a \in C_A} \|b - a\|\right\} \quad (5)$$

For the regression-based model, we computed the mean square error (MSE) between the predicted and the ground truth lengths. Note that as mentioned in section

3, the ground truth length was expressed in pixels and scaled by a factor of 50 while training the models. As a result, the equation of the MSE evaluation metric is defined as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{l_i - l_i'}{\lambda \times s_i}\right)^2 \quad (6)$$

where λ denotes the scaling factor and s_i denotes the pixel spacing for test image i.

For evaluating the QC model, we computed sensitivity (SEN) and specificity (SPE). SEN describes the proportion of identified positives out of all ground truth positives, and SPE describes the proportion of identified negatives out of all ground truth negatives. In addition, the area under the ROC curve (AUC) was also calculated for the overall evaluation of the QC model. These metrics are defined as:

$$SEN = \frac{TP}{TP+FN} \quad (7)$$

$$SPE = \frac{TN}{FP+TN} \quad (8)$$

$$AUC = \int_{-\infty}^{\infty} TPR(t)FPR(t)dt \quad (9)$$

where TP represents true positives, FN is false negatives, FP is false positives and TN is true negatives. TPR represents the true positive rate while FPR represents the false positive rate for a given threshold t.

4.2. Spleen Length Measurement Results

We first conducted experiments to compare the two types of length measurement approach – one segmentation-based model and two regression-based models as follows:

(1) Segmentation-based: Modified U-Net with post-processing on predicted segmentation (SB)；
(2) Direct regression: U-Net contracting path direct regression without transferring weights (UC)；
(3) Direct regression: U-Net contracting path direct regression with transferring weights (UCW)；
(4) Direct regression: VGG-19 (VGG).

As mentioned in section 2.1, two additional sonographers performed retrospective length measurements on 108 ultrasound images in addition to the original length measurements. These were used to compute inter-observer variability in human expert length measurement to give context to the performance of our automated models. We refer to this variability as human error (HE) in the results below.

In Figure 7, we show examples of segmentation results from the U-Net in the SB approach. Table 1 shows the results from SB using the three different post-processing techniques described in section 2.3.1 to estimate the length from the predicted segmentation (LDP, PCA and Post VGG-19).

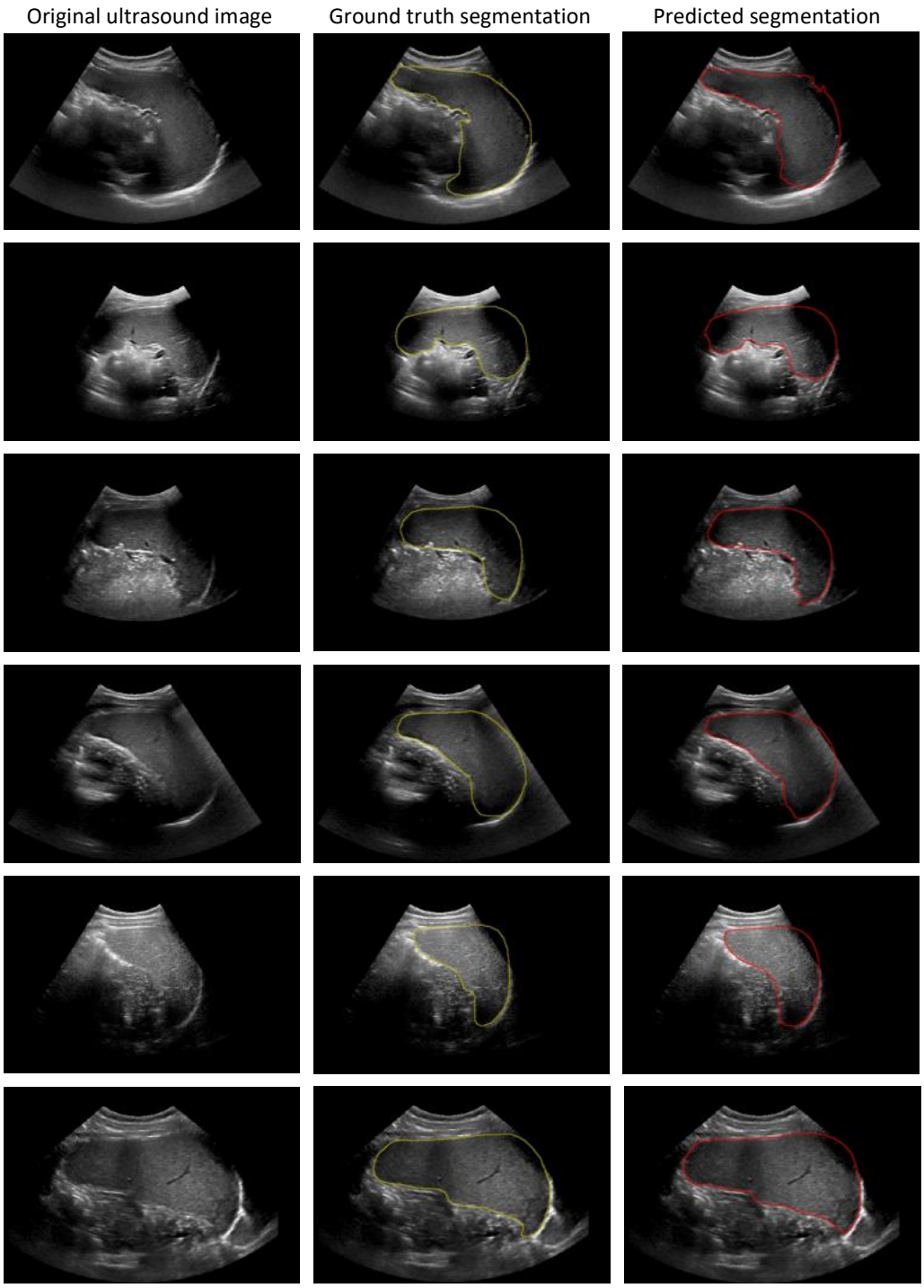


Figure 7: Visualisations of the segmentation results from U-Net. The pictures from the left to the right column are original ultrasound images, ground truth segmentations and predicted segmentations.

	MPLE	**R**
LDP	4.86%	**0.98**

PCA	**4.58%**	**0.98**
Post VGG-19	7.56%	0.95

Table 1: Comparison between results for three post-processing methods for length measurement from the predicted segmentation: longest distance between the points on the segmentation contour (LDP); length measurement based on principal components analysis (PCA); and VGG-19 (Post VGG-19). Two metrics are used here to evaluate the proposed methods: mean percentage length error (MPLE) and the Pearson's correlation coefficient (R). Best results shown in bold.

As can be seen from Table 1, PCA-based length measurement had the best performance with an MPLE of 4.58% and an R of 0.98. Therefore, we used this method with the SB model to compare its results with the other two models (direct regression). Table 2 presents the results from all proposed models.

	MPLE	**R**	**Dice**	**IoU**	**HD**	**MSE**
SB	4.58%	0.98	0.93	0.87	8.18mm	-
UC	8.76%	0.94	-	-	-	0.322
UCW	6.40%	0.97	-	-	-	0.165
VGG-19	8.53%	0.94	-	-	-	0.305
HE	5.78%	0.95	-	-	-	-

Table 2: Comparison of the results between four proposed models: segmentation-based (SB); U-Net contracting path regression without weight transfer (UC); U-Net contracting path regression with weight transfer (UCW); and VGG-19 direct regression (VGG-19). HE refers to human error in length measurement, i.e. inter-observer variability. Six metrics are adopted to compare the performance of different models: mean percentage length error (MPLE);

the Pearson's correlation coefficient (R); Dice; Intersection over union (IoU); Hausdorff distance (HD); and mean squared error (MSE). Best results in bold.

The results from Table 2 show that the U-Net with PCA-based post-processing (SB) outperformed the other three methods, and its performance reached the level of variability in human expert performance (MPLE of 5.78%) on the length measurement task. The regression-based methods performed less well (MPLE of 8.76%, 6.40%, 8.53% for UC, UCW and VGG-19, respectively), but the weight transfer from the segmentation task helped to improve the length estimation performance from an MPLE of 8.76% to 6.40% for the U-Net contracting path method (UCW).

4.3. Quality Control Results

Table 3 shows the average results for the 4-fold cross-validation of the QC model. It can be seen that SEN was higher than the SPE (0.958 compared to 0.773).

	SEN	**SPE**	**AUC**
QC CV	0.958	0.773	0.883

Table 3: Results for 4-fold cross-validation of the quality control system (QC CV). SEN: sensitivity, SPE: specificity, AUC: area under the ROC curve.

Table 4 illustrates the results of the complete pipeline, featuring automated QC and segmentation-based length measurement on the held-out test dataset (84 good quality images and 11 bad quality images). In addition, we also applied the segmentation-based length measurement method to the held-out test dataset without QC for comparison.

	SEN	SPE	AUC	MPLE	MPLE W/O QC
QC final	0.964	0.636	0.841	4.88%	5.76%

Table 4: The results for the final quality-controlled length measurement pipeline on the held-out test dataset. We performed length measurement on the ultrasound images selected by our QC model using the SB model with PCA (our best model). The MPLE was the final pipeline result with QC. We also conducted length measurement without the QC system as a comparison (MPLE W/O QC). SEN: sensitivity, SPE: specificity, AUC: area under the ROC curve, MPLE: mean percentage length error.

The MPLE was 5.76% without any QC. After applying our QC model, this improved to 4.88%. Both MPLEs were within the range of human inter-observer variability.

5. Discussion

We have proposed three models to enable automatic length measurement of the spleen from ultrasound images. In the segmentation-based approach, we investigated three different methods for predicting length from the segmentation. From Table 1, it can be seen that PCA was superior to LDP and Post VGG-19. We believe this is because the PCA method first finds the axis where the coordinates have the largest variance. Therefore, this measurement takes the overall shape of the spleen into account, which is similar to the way in which sonographers make these measurements. The LDP measurement does not consider the geometry of the spleen, and errors in the segmentation may worsen the performance of this method and the Post VGG-19

method. Comparing HE and SB from Table 2 shows that our best segmentation-based model with PCA has reached the human expert level. The MPLE and R for UC and VGG-19 were similar, but VGG-19 had fewer parameters and slightly better accuracy than UC. Comparing the results from UC and UCW, the weights transferred from the segmentation U-Net helped improve the model's performance to directly predict the length of the spleen. This suggests that the representations learnt for the segmentation and regression tasks have similarities. UCW's MPLE (6.40%) was close to the level of human expert variability, which demonstrates the ability to use an encoding CNN to perform direct length estimation. It is possible that with more images for training, UCW could reach a similar level of performance to SB. However, note that we used a segmentation-based cropping method when conducting the SB experiments. This showed promising results, but in order to deploy the SB-based pipeline in a clinical setting, an object instance detection method would need to be applied to crop the images (e.g., [35] and [36]) and we will investigate this in future work. Note that the regression-based approaches used centre-cropping so would not require such an additional step.

We also conducted experiments on the QC model. It can be seen from the cross-validation results in Table 3 that SEN was 0.958 and SPE was 0.773, meaning that the model can correctly identify 95.8% of good images and 77.3% of bad images. Table 4 shows similar results for the final QC model (0.964 SEN compared to 0.636 SPE). This indicates that the model is more sensitive to positive examples (good quality images) than negative examples (bad quality images). Such a result is mainly because of the data imbalance in our dataset (420 vs. 55), and it is possible that better SPE could be

achieved with a greater number of bad quality images for training. The results from the final pipeline suggest that with the QC, the MPLE on the selected images is superior to the MPLE without QC. However, both MPLEs were within the human expert variability, which shows the robustness of our length measurement system.

In this work, we have chosen to focus on estimating spleen size in a cohort of SCD patients, which is an application in which there is a clear clinical need for this technology. However, splenomegaly also occurs in patients with a range of conditions as mentioned in section 1, and we believe that the techniques we have proposed will have wider application. Because spleen ultrasound data from SCD patients will likely come from a different data distribution to similar data from other patient groups, training data from these groups would be needed to adapt our methods to other applications. This could be done by either training from scratch using the new data or fine-tuning the weights of our SCD patient-based model.

Note also that we have focused on the image analysis in this work, i.e., automated spleen length measurement from ultrasound images. However, in reality, acquiring images with good quality also requires significant expertise. Our intention is that the proposed quality-controlled framework would help ensure good image quality, thus making the length measurement more robust and enabling the process to be more easily applied in areas where there is often a shortage of experienced sonographers. In these places, QC becomes a key component of a clinically applicable technique for operators with little training and less experience. We envisage that an automated measurement

system with a QC step such as we have proposed could help to automate the selection and interpretation of images acquired by less experienced sonographers. In the future, we will consider incorporating a "quality score" system to mark each ultrasound image to ensure that the length is measured from the best quality image from each examination. Besides, it is also important to consider how machine learning tools can be integrated into imaging systems for practical use. For example, a software framework has recently been proposed [37] that enables machine learning models to be supplied as plug-ins to support real-time image acquisition and analysis. Such approaches are essential to enable clinical translation of methods such as the ones we have proposed in this paper.

The current clinical workflow uses spleen length as a surrogate for spleen volume, which is actually the measure that would be most clinically useful when assessing spleen size. More sophisticated 3D modalities such as MRI or CT could be used to determine the splenic volume [38]. Recently, a deep learning-based framework has been proposed to automatically estimate the spleen volume from CT images [39]. However, these modalities are expensive and hence are not widely available in parts of the global south, making ultrasound imaging a better choice. Apart from our preliminary work [12], no prior work has investigated the use of deep learning to assess spleen ultrasound images. 3D ultrasound could, in principle, be used to estimate spleen volume, but 3D ultrasound probes are not widely available. Machine learning techniques such as recurrent neural networks or transformers could be used to automatically process time series of 2D ultrasound images and make direct estimates of volume. We will investigate this possibility in future work.

6. Conclusions

We have proposed the first automated pipeline for spleen length measurement from ultrasound images. Building on our preliminary work [12], the pipeline consists of automated QC followed by deep learning-based length estimation, and we demonstrate that our best-performing model has reached human expert level. We believe that this approach could greatly improve the management of SCD, particularly in parts of the global south, and mitigate the difficulties caused by the lack of experienced sonographers.

7. Abbreviations and Acronyms

Abbreviation	Meaning
General Abbreviations	
SCD	Sickle cell disease
QC	Quality control
CT	Computed tomography
MRI	Magnetic resonance image
CNN	Convolutional neural network
ROC	Receiver operating characteristic
Method and Model Abbreviations	
CCA	Connected components analysis
LDP	Length measurement based on the longest distance between pairs of points on the segmentation boundary
PCA	Length measurement based on principal components analysis
Post VGG-19	Length measurement using VGG-19 to directly estimate the length from the segmentation
SB	Segmentation-based model: modified U-Net with post-processing on predicted segmentation

UC	Regression-based model: U-Net contracting path direct regression without transferring weights
UCW	Regression-based model: U-Net contracting path direct regression with transferring weights
VGG	Regression-based model: VGG-19
QC CV	Cross-validation of the quality control system
Evaluation Metric Abbreviations	
HE	Human error
MPLE	Mean percentage length error
MSE	Mean squared error
R	Pearson's correlation coefficient
IoU	Intersection over union
HD	Hausdorff distance
SEN	Sensitivity
SPE	Specificity
AUC	Area under the receiver operating characteristic curve

8. Acknowledgements

This work was supported by the Wellcome/EPSRC Centre for Medical Engineering [WT 203148/Z/16/Z]. The support provided by China Scholarship Council during PhD programme of Zhen Yuan in King's College London is acknowledged.

9. References

[1] A.W. Woodruff, Mechanisms involved in anaemia associated with infection and splenomegaly in the tropics, Trans. R. Soc. Trop. Med. Hyg. 67 (1973). https://doi.org/10.1016/0035-9203(73)90107-7.

[2] P.A. McCormick, K.M. Murphy, Splenomegaly, hypersplenism and coagulation abnormalities in liver disease, Best Pract. Res. Clin. Gastroenterol. 14 (2000). https://doi.org/10.1053/bega.2000.0144.

[3] A.L. Pozo, E.M. Godfrey, K.M. Bowles, Splenomegaly: Investigation, diagnosis and management, Blood Rev. 23 (2009). https://doi.org/10.1016/j.blre.2008.10.001.

[4] S. Chakravorty, T.N. Williams, Sickle cell disease: A neglected chronic disease of increasing global health importance, Arch. Dis. Child. 100 (2015). https://doi.org/10.1136/archdischild-2013-303773.

[5] G.J. Kato, F.B. Piel, C.D. Reid, M.H. Gaston, K. Ohene-Frempong, L. Krishnamurti, W.R. Smith, J.A. Panepinto, D.J. Weatherall, F.F. Costa, E.P. Vichinsky, Sickle cell disease, Nat. Rev. Dis. Prim. 4 (2018). https://doi.org/10.1038/nrdp.2018.10.

[6] B. Inusa, M. Casale, N. Ward, Introductory Chapter: Introduction to the History, Pathology and Clinical Management of Sickle Cell Disease, in: Sick. Cell Dis. - Pain Common Chronic Complicat., 2016. https://doi.org/10.5772/65648.

[7] V. Brousse, P. Buffet, D. Rees, The spleen and sickle cell disease: The sick(led) spleen, Br. J. Haematol. 166 (2014). https://doi.org/10.1111/bjh.12950.

[8] P.M. Lamb, A. Lund, R.R. Kanagasabay, A. Martin, J.A.W. Webb, R.H. Reznek, Spleen size: How well do linear ultrasound measurements correlate with three-dimensional CT volume assessments?, Br. J. Radiol. 75 (2002). https://doi.org/10.1259/bjr.75.895.750573.

[9] H.K. Rosenberg, R.I. Markowitz, H. Kolberg, C. Park, A. Hubbard, R.D. Bellah, Normal splenic size in infants and children: Sonographic measurements, Am. J. Roentgenol. 157 (1991). https://doi.org/10.2214/ajr.157.1.2048509.

[10] F.B. Piel, A.P. Patil, R.E. Howes, O.A. Nyangiri, P.W. Gething, M. Dewi, W.H. Temperley, T.N. Williams, D.J. Weatherall, S.I. Hay, Global epidemiology of Sickle haemoglobin in neonates: A contemporary geostatistical model-based map and population estimates, Lancet. 381 (2013). https://doi.org/10.1016/S0140-6736(12)61229-X.

[11] S.D. Grosse, I. Odame, H.K. Atrash, D.D. Amendah, F.B. Piel, T.N. Williams, Sickle cell disease in Africa: A neglected cause of early childhood mortality, Am. J. Prev. Med. 41 (2011). https://doi.org/10.1016/j.amepre.2011.09.013.

[12] Z. Yuan, E. Puyol-Antón, H. Jogeesvaran, C. Reid, B. Inusa, A.P. King, Deep Learning for Automatic Spleen Length Measurement in Sickle Cell Disease Patients, in: Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), 2020. https://doi.org/10.1007/978-3-030-60334-2_4.

[13] T.G. Debelee, F. Schwenker, A. Ibenthal, D. Yohannes, Survey of deep learning in breast cancer image analysis, Evol. Syst. 11 (2020). https://doi.org/10.1007/s12530-019-09297-2.

[14] G. Litjens, F. Ciompi, J.M. Wolterink, B.D. de Vos, T. Leiner, J. Teuwen, I. Išgum, State-of-the-Art Deep Learning in Cardiovascular Image Analysis, JACC Cardiovasc. Imaging. 12 (2019). https://doi.org/10.1016/j.jcmg.2019.06.009.

[15] J. Torrents-Barrena, G. Piella, N. Masoller, E. Gratacós, E. Eixarch, M. Ceresa, M.Á.G. Ballester, Segmentation and classification in MRI and US fetal imaging: Recent trends and future prospects, Med. Image Anal. 51 (2019). https://doi.org/10.1016/j.media.2018.10.003.

[16] L.Q. Zhou, J.Y. Wang, S.Y. Yu, G.G. Wu, Q. Wei, Y. Bin Deng, X.L. Wu, X.W. Cui, C.F. Dietrich, Artificial intelligence in medical imaging of the liver, World J. Gastroenterol. 25 (2019). https://doi.org/10.3748/wjg.v25.i6.672.

[17] S. Han, H.K. Kang, J.Y. Jeong, M.H. Park, W. Kim, W.C. Bang, Y.K. Seong, A deep learning framework for supporting the classification of breast lesions in ultrasound images, Phys. Med. Biol. 62 (2017). https://doi.org/10.1088/1361-6560/aa82ec.

[18] H. Li, J. Weng, Y. Shi, W. Gu, Y. Mao, Y. Wang, W. Liu, J. Zhang, An improved deep learning approach for detection of thyroid papillary cancer in ultrasound images, Sci. Rep. 8 (2018). https://doi.org/10.1038/s41598-018-25005-7.

[19] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, IEEE Trans. Pattern Anal. Mach. Intell. 39 (2017). https://doi.org/10.1109/TPAMI.2016.2577031.

[20] J. Zhang, C. Petitjean, P. Lopez, S. Ainouz, Direct estimation of fetal head circumference from ultrasound images based on regression CNN, 2020.

[21] A. Ghorbani, D. Ouyang, A. Abid, B. He, J.H. Chen, R.A. Harrington, D.H. Liang, E.A. Ashley, J.Y. Zou, Deep learning interpretation of echocardiograms, Npj Digit. Med. 3 (2020). https://doi.org/10.1038/s41746-019-0216-8.

[22] A.I.L. Namburete, W. Xie, J.A. Noble, Robust regression of brain maturation from 3D fetal neurosonography using CRNs, in: Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), 2017. https://doi.org/10.1007/978-3-319-67561-9_8.

[23] C.-C. Kuo, C.-M. Chang, K.-T. Liu, W.-K. Lin, H.-Y. Chiang, C.-W. Chung, M.-R. Ho, P.-R. Sun, R.-L. Yang, K.-T. Chen, Automation of the kidney function prediction and classification through ultrasound-based kidney imaging using deep learning, Npj Digit. Med. 2 (2019). https://doi.org/10.1038/s41746-019-0104-2.

[24] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), 2015. https://doi.org/10.1007/978-3-319-24574-4_28.

[25] Z. Zhuang, A.N.J. Raj, A. Jain, N. Ruban, S. Chaurasia, N. Li, M. Lakshmanan, M. Murugappan, Nipple Segmentation and Localization Using Modified U-Net on Breast Ultrasound Images, J. Med. Imaging Heal. Informatics. 9 (2019). https://doi.org/10.1166/jmihi.2019.2828.

[26] D. Qiao, F. Zulkernine, Dilated Squeeze-and-Excitation U-Net for Fetal Ultrasound Image Segmentation, in: 2020 IEEE Conf. Comput. Intell. Bioinforma. Comput. Biol. CIBCB 2020, 2020. https://doi.org/10.1109/CIBCB48159.2020.9277667.

[27] S.H. Gao, M.M. Cheng, K. Zhao, X.Y. Zhang, M.H. Yang, P. Torr, Res2Net: A New Multi-Scale Backbone Architecture, IEEE Trans. Pattern Anal. Mach. Intell. 43 (2021). https://doi.org/10.1109/TPAMI.2019.2938758.

[28] S. Jegou, M. Drozdzal, D. Vazquez, A. Romero, Y. Bengio, The One Hundred Layers Tiramisu: Fully Convolutional DenseNets for Semantic Segmentation, in: IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work., 2017. https://doi.org/10.1109/CVPRW.2017.156.

[29] F. Isensee, P. Kickingereder, W. Wick, M. Bendszus, K.H. Maier-Hein, No new-net, in: Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), 2019. https://doi.org/10.1007/978-3-030-11726-9_21.

[30] A. Mihaylova, V. Georgieva, A Brief Survey of Spleen Segmentation in MRI and CT Images, Int. J. Adv. Comput. Sci. Technol. 5 (2016) 72–77.

[31] A. Telea, An Image Inpainting Technique Based on the Fast Marching Method, J. Graph. Tools. 9 (2004). https://doi.org/10.1080/10867651.2004.10487596.

[32] P.A. Yushkevich, J. Piven, H.C. Hazlett, R.G. Smith, S. Ho, J.C. Gee, G. Gerig, User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability, Neuroimage. 31 (2006). https://doi.org/10.1016/j.neuroimage.2006.01.015.

[33] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., 2016. https://doi.org/10.1109/CVPR.2016.90.

[34] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: 3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc., 2015.

[35] U. Masud, T. Saeed, H.M. Malaikah, F.U. Islam, G. Abbas, Smart Assistive System for Visually Impaired People Obstruction Avoidance Through Object Detection and Classification, IEEE Access. 10 (2022) 13428–13441. https://doi.org/10.1109/ACCESS.2022.3146320.

[36] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask R-CNN, IEEE Trans. Pattern Anal. Mach. Intell. 42 (2020). https://doi.org/10.1109/TPAMI.2018.2844175.

[37] A. Gomez, V.A. Zimmer, G. Wheeler, N. Toussaint, S. Deng, R. Wright, E. Skelton, J. Matthew, B. Kainz, J. Hajnal, J. Schnabel, PRETUS: A plug-in based platform for real-time ultrasound imaging research, SoftwareX. 17 (2022) 100959. https://doi.org/10.1016/j.softx.2021.100959.

[38] E.M. Yetter, K.B. Acosta, M.C. Olson, K. Blundell, Estimating Splenic Volume: Sonographic Measurements Correlated with Helical CT Determination, Am. J. Roentgenol. 181 (2003). https://doi.org/10.2214/ajr.181.6.1811615.

[39] G.E. Humpire-Mamani, J. Bukala, E.T. Scholten, M. Prokop, B. van Ginneken, C. Jacobs, Fully Automatic Volume Measurement of the Spleen at CT Using Deep Learning, Radiol. Artif. Intell. 2 (2020). https://doi.org/10.1148/ryai.2020190102.