Uncertainty Quantification for Amniotic Fluid Segmentation and Volume Prediction

Daniel Csillag [1] **Lucas Monteiro** [1] **Thiago Ramos** [1] **João Vitor Romano** [1] **Rodrigo Schuller** [1] **Roberto B. Seixas** [1] **Roberto I. Oliveira** [1] **Paulo Orenstein** [1]

Abstract

Recent advances in neural networks have led to significant improvements in medical image segmentation. However, in many cases, it is crucial to quantify the confidence of such predictions. We present a novel dataset for fetal MRI and show that segmenting amniotic fluid can be done accurately with recent architectures. We then propose and evaluate different ways of creating confidence intervals for segmentation and volume forecasts via Conformal Prediction. We find that methods that yield variable-length intervals are better-suited to medical segmentation tasks in general, and illustrate this in the amniotic fluid case.

1. Introduction

The use of AI techniques in medical imaging and diagnostics requires not only accurate algorithms, but also adequate quantification of the uncertainty of these methods. This paper develops and evaluates ways of creating confidence intervals for such tasks. In particular, we focus on the emblematic case of segmenting amniotic fluid (AF) in fetal MRI exams and present theoretically sound, high-confidence prediction methods for the volume and shape of AF.

Amniotic fluid (AF) is a liquid that surrounds the fetus during gestation. Abnormal AF volumes are linked to pregnancy complications and negative outcomes (Chamberlain et al., 1984; Moore, 2011; Moore & Cayle, 1990). Polyhydramnios, or excessive AF volume, occurs in 1 to 2% of pregnancies, and corresponds to a 2 to 5-fold increase in perinatal morbidity and mortality. By contrast, reduced AF volume, or oligohydramnios, occurs in up to 12% of pregnancies, and corresponds to a 15 to 50-fold increase in perinatal morbidity and mortality. Thus, accurate prediction of AF volume is crucial to ensure a healthy pregnancy.

Ultrasound exams (US) are the primary method for AF volume assessment (Magann et al., 2001; Werner et al., 2015). The two most important techniques are Amniotic Fluid Index and the Deepest Vertical Pocket (Moore & Cayle, 1990; Phelan et al., 1987; Phelan & Smith, 1987). These techniques are semiquantitative, indirect, and not very precise. Moreover, the cutoff points used to define oligohydramnios and polyhydramnios are still the object of controversy.

Magnetic Resonance Imaging (MRI) exams are used to obtain additional information about fetal abnormalities and conditions in situations where US cannot provide high-quality images (Hellinger & Epelman, 2010; Kubik-Huch et al., 2001; Moschos et al., 2017; Prayer et al., 2004). However, the evaluation of AF volume from MRI is subjective.

This paper evaluates different methodologies for uncertainty quantification in medical segmentation tasks, with a focus on AF volume estimation. We devise algorithms that, when presented with a fetal MRI exam, output an AF segmentation, a confidence interval for the AF volume in the exam, and a predictive set for the shape of the AF in the exam. To achieve these goals, we combine U-Net neural networks with techniques from the growing literature of Conformal Prediction (CP) (Shafer & Vovk, 2008; Lei et al., 2018; Bates et al., 2020). All CP methods are backed by theoretical results that guarantee their probabilistic coverage. Empirical results using a novel fetal MRI dataset show that our prediction methodology as well as the confidence intervals produced are accurate in properly segmenting AF and estimating fluid volume. Many of the results and algorithms presented extend to other medical segmentation tasks.

2. Dataset, pipeline, and segmentation models

We retrospectively evaluate 652 fetal MRI performed by the same fetal radiologist from January 2015 to April 2021. The gestational age was between 19 to 38 weeks of gestation. Over 80% of the subjects present some degree of pathology, which can sometimes be reflected in the resulting exam.

MRI images were produced using a 1.5-T scanner. The 3D reconstruction protocol was T2-weighted true fast imaging with a steady-state precession (TrueFisp) sequence in the sagittal plane, FOV $380\,\mathrm{mm}$, voxel size: $1 \times 1 \times 1\,\mathrm{mm}$; acquisition time, $0.24\,\mathrm{s}$. Maternal sedation was not used.

[1]Instituto de Matemática Pura e Aplicada, Rio de Janeiro, Brazil. Correspondence to: Paulo Orenstein <pauloo@impa.br>.

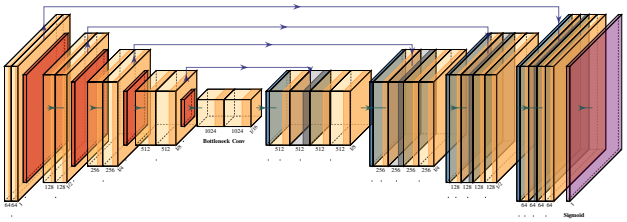

Figure 1. U-Net architecture (via PlotNeuralNet (Iqbal, 2018))

The uterine contents (fetus, placenta, AF and umbilical cord) of the MRI scans were segmented by humans, under supervision of the radiologist who performed the exams, and the AF was highlighted (see Figure 2). The end result is a set of pairs $(X_i, Y_i)_{i=1}^{\ell}$ of $\ell = 652$ segmented exams, where X_i is the 3D exam image and Y_i is the highlighted AF. The dataset was divided into three disjoint sets: $(X_i, Y_i)_{i=1}^{n}$ for training ($n = 420$ exams), $(X_i, Y_i)_{i=n+1}^{m}$ for validation ($m - n = 120$ exams) and $(X_i, Y_i)_{i=m+1}^{l}$ for testing ($l - m = 112$ exams). Models were trained on the training set, hyperparameters and early stopping were decided based on the validation set and the best model was selected according to the Dice coefficient on the test set.

As hyperparameters, we used the Adam optimizer (Kingma & Ba, 2014) with learning rate of 0.001; batch size of 4; and maximum of 100 epochs, with early stopping after 7 epochs. The networks were not trained on the full 3D exams, but on "augmented 2D data." That is, each exam was decomposed into 2D grayscale slices (about 160 per exam), scaled to dimensions 256×256. Moreover, we also passed the values of corresponding pixels in the previous and following slices of the same exam. This overcame the limitations of using purely 2D data. Neither data augmentation nor additional information about more distant slices were helpful.

The neural network architectures tested were: a standard U-Net convolutional network (Ronneberger et al., 2015) with 17 million parameters (see Figure 1); Fast-SCNN (Poudel et al., 2019), a convolutional model with 1.1 million parameters; and Small U-Net, a version of U-Net slimmed down to 1.9 million parameters. Soft Dice was used as the loss function; binary cross entropy was also considered. We observed that mis-segmentation happens mostly close to the border of AF, so Active Contour loss (Chen et al., 2019) was used in conjunction with BCE. The results are in Section 4.

3. Methods for uncertainty quantification

Conformal prediction (CP) (Shafer & Vovk, 2008) is a general family of uncertainty quantification methods. They provide confidence regions with valid coverage from any arbitrary predictive method. In this paper, we build *volume-predictive intervals* and *shape-predictive regions* via CP.

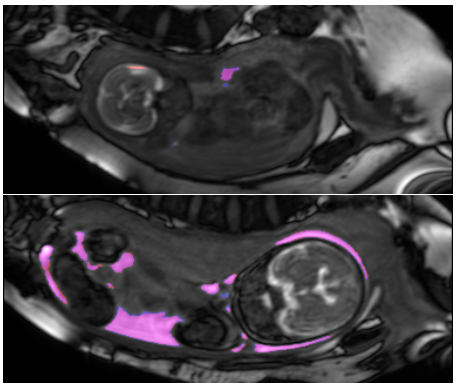

Figure 2. Slice from hard exam (Dice: 0.5354) on top, and slice from a typical exam (Dice: 0.9352) on the bottom. Red is $\mathcal{M}(X_i) \setminus Y_i$, blue is $Y_i \setminus \mathcal{M}(X_i)$ and magenta is $Y_i \cap \mathcal{M}(X_i)$.

Consider a trained model $\mathcal{M}$ mapping 3D exams X to vectors $\mathcal{M}(X) \in [0,1]^v$ where v is the number of voxels. Given $0 \leq t \leq 1$, $\mathcal{M}(X)_{\geq t}$ is the thresholded output where values above t are replaced by 1s and other values are replaced to 0s. Our standard value for predicting AF is $t = 0.5$. Let (X_j, Y_j) with $m + 1 \leq j \leq l$ be a point in the test sample, and let $\mathrm{Vol}(Y_j)$ denote the volume of AF. A *volume-predictive interval at confidence level* $1 - \alpha$ is a mapping from X to an interval $\mathcal{I}_\alpha(X) \subset \mathbb{R}$, also depending on training and validation data, such that:

$$\mathbb{P}[\mathrm{Vol}(Y_j) \in \mathcal{I}_\alpha(X_j)] \geq 1 - \alpha.$$

A *shape-predictive region at confidence level* $1-\alpha$ is another map $\mathcal{C}_\alpha : X_j \mapsto \mathcal{C}_\alpha(X_j)$, which also depends on training and validation data, taking X_j to a subset of the space of segmentations, in a way that

$$\mathbb{P}[Y_j \in \mathcal{C}_\alpha(X_j)] \geq 1 - \alpha.$$

Shape-predictive regions control the geometry of the amniotic fluid prediction, and not just its volume. Our methods for volume prediction are tighter than those for shapes. However, the latter can be advantageous because they are easier to be interpreted by doctors and checked against the exam.

3.1. Volume-predictive intervals

We build on two CP methods for uncertainty quantification of volume prediction. The first one is Algorithm 1, relying on a normalizing function g. This method gives confidence intervals whose lengths on each test point are of the form $2g(\mathcal{M}(X_j)) \cdot \mathrm{radius}$ where X_j is a test point and radius is a X_j-independent value chosen from validation data. The case $g \equiv 1$ amounts to (Lei et al., 2018) and is a standard method for building constant-length predictive intervals for scalar responses. In our tests, we also consider a volume normalization $g(\mathcal{M}(X_j)) = \mathrm{Vol}\,(\mathcal{M}(X_j))$; this has the

Algorithm 1 Normalized Standard Volume Prediction

Input: model $\mathcal{M}$, validation set $\{(X_i, Y_i)\}_{i=n+1}^{m}$, test set $\{(X_i, Y_i)\}_{i=m+1}^{l}$, confidence $1-\alpha \in (0,1)$, and normalizing function $g : \Im(\mathcal{M}) \to \mathbb{R}_+$.

radii $\leftarrow [\,]$
for $i \in \{n+1, ..., m\}$ **do**
  append $\left(\frac{\left|\mathrm{Vol}\left(\mathcal{M}(X)_{\geq .5}\right) - \mathrm{Vol}(Y_i)\right|}{g(\mathcal{M}(x_i)_{\geq .5})}\right)$ to radii
end
radius $\leftarrow (1-\alpha)$-quantile of radii
for $j \in \{m+1, ..., l\}$ **do**
  $dv \leftarrow g(\mathcal{M}(X_j)_{\geq .5}) \cdot$ radius
  lower volume $\leftarrow \mathrm{Vol}\left(\mathcal{M}(X)_{\geq .5}\right) - dv$
  upper volume $\leftarrow \mathrm{Vol}\left(\mathcal{M}(X)_{\geq .5}\right) + dv$
  $\mathcal{I}_\alpha(X_j) \leftarrow$ [lower volume, upper volume]
end

advantage of making interval lengths adaptive to the wide range of volumes in the data. Importantly, the results of (Lei et al., 2018) can be adapted to show that Algorithm 1 has adequate coverage for any g.

A different method is described in Algorithm 2 (Thresholded Volume Prediction), which uses the model output $\mathcal{M}(X)$ thresholded at different values of t. Intuitively, the magnitudes of the values of $\mathcal{M}(X)$ at each voxel give additional information about how likely each voxel is to correspond to AF. Mathematically, the sets $\mathcal{M}(X)_{\geq t}$ are nested, and one can use the general theory of (Bates et al., 2020) to prove that Thresholded Volume Prediction has correct coverage.

3.2. Shape-predictive regions

Algorithm 3 (Segmentation Prediction) gives shape-predictive regions for the segmentation. Given α and an additional, user specified leniency parameter λ, the confidence region $\mathcal{C}_\alpha(X_j)$ takes the following form: $Y_j \in \mathcal{C}_\alpha(X_j)$ if and only if $\max\{\mathrm{Vol}\left(Y_j \backslash \mathcal{U}_{\alpha,\lambda}(X_j)\right), \mathrm{Vol}\left(\mathcal{L}_{\alpha,\lambda}(X_j) \backslash Y_j\right)\}$ is less than or equal to $\lambda \mathrm{Vol}\left(Y_j\right)$, where $\mathcal{U}_{\alpha,\lambda}(X_j) \supset \mathcal{L}_{\alpha,\lambda}(X_j)$ are obtained from $\mathcal{M}(X_j)$ by thresholding at values learned from validation data. The choice of $\lambda = 0$ guarantees that $\mathcal{L}_{\alpha,\lambda}(X_j) \subset Y_j \subset \mathcal{U}_{\alpha,\lambda}(X_j)$ with probability $1-\alpha$, but in practice leads to pessimistic choices of $\mathcal{U}_{\alpha,\lambda}(X_j)$ and $\mathcal{L}_{\alpha,\lambda}(X_j)$. Therefore, we allow for a positive leniency $\lambda > 0$ in the results we report. The theoretical coverage property of Algorithm 3 follows from (Bates et al., 2020).

4. Results

4.1. Image segmentation

The best-performing model was U-Net with binary cross-entropy as loss function. Its Dice coefficient on the test dataset was of approximately 0.92, showing that it can capably segment previously unseen exams on average. The performance of this model is further illustrated by Figure 2. Table 1 below summarizes the results of our neural network architectures trained with different losses.

It is natural to assume that segmentations are noisy at the boundaries of AF: these are the hardest regions for humans perform segmentation, and the decision of where exactly to place the boundary is somewhat arbitrary. Accordingly, most of the errors of the best performing neural network are concentrated on the borders. This is visible in Figure 2 and can be made more evident via dilations and erosions. Dilating the prediction masks by a box kernel with size $3 \times 3 \times 3$ covers 83% of $Y_i \setminus \mathcal{M}(X_i)$ on average. Eroding the prediction mask by a box kernel with size $3 \times 3 \times 3$ excludes 87% of $\mathcal{M}(X_i) \setminus Y_i$ on average.

4.2. Performance of volume-predictive regions

Figure 3 (up) presents the average on test data of the lengths of intervals $\mathcal{I}_\alpha(X_j)$ normalized by $\mathrm{Vol}\left(Y_j\right)$. for different values of confidence $1-\alpha$. The base NN $\mathcal{M}$ is the top performing model from the previous section. Unnormalized Standard Volume Prediction has the widest intervals (as expected); and normalized Standard Volume Prediction and Thresholded Volume Prediction have similar average lengths up to confidence $\sim .96$. The lengths of the intervals in the last two methods strongly correlated with the

Algorithm 2 Thresholded Volume Prediction

Input: model $\mathcal{M}$, validation set $\{(X_i, Y_i)\}_{i=n+1}^{m}$, test set $\{(X_i, Y_i)\}_{i=m+1}^{l}$ and confidence $1-\alpha \in (0,1)$

thresholds $\leftarrow [\,]$
for $i \in \{n+1, ..., m\}$ **do**
  $p \leftarrow$ proportion of ones in Y_i
  best threshold $\leftarrow p$-quantile$(\mathcal{M}(X_i))$
  append best threshold to thresholds
end
upper bound$_t \leftarrow -(1-\alpha/2)$-quantile of list $-$thresholds
lower bound$_t \leftarrow (1-\alpha/2)$-quantile of list thresholds
for $j \in \{m+1, ..., l\}$ **do**
  lower volume $\leftarrow \mathrm{Vol}\left(\mathcal{M}(X_j)_{\geq \text{lower bound}_t}\right)$
  upper volume $\leftarrow \mathrm{Vol}\left(\mathcal{M}(X_j)_{\geq \text{upper bound}_t}\right)$
  $\mathcal{I}_\alpha(X_j) \leftarrow$ [lower volume, upper volume]
end

Table 1. Average Dice coefficient and standard deviation across 112 exams (test set) of networks trained with diverse loss functions.

Model	Soft Dice	BCE	AC+BCE
U-Net	0.908 ± 0.10	$\mathbf{0.924 \pm 0.06}$	0.923 ± 0.07
Fast-SCNN	0.871 ± 0.11	0.870 ± 0.08	0.872 ± 0.09
Small U-Net	0.903 ± 0.09	0.911 ± 0.08	0.921 ± 0.08

Algorithm 3 Segmentation Prediction

Input: model $\mathcal{M}$, validation set $\{(X_i, Y_i)\}_{i=n+1}^{m}$, test set $\{(X_i, Y_i)\}_{i=m+1}^{l}$, leniency $\lambda \in (0, 1)$ and confidence $1 - \alpha \in (0, 1)$

upper thresholds $\leftarrow [\,]$; lower thresholds $\leftarrow [\,]$

for $i \in \{n+1, ..., m\}$ **do**

  $\lambda_{\text{upper}} \leftarrow \lambda$

  upper threshold $\leftarrow \lambda_{\text{upper}}$-quantile$(\mathcal{M}(X_i)|Y_{i,v} > 0.5)$

  append $\min$(upper threshold, 0.5) to upper thresholds

  $\lambda_{\text{lower}} = 1 - \lambda \cdot \text{Vol}\,(Y_i)\,/\text{Vol}\,(1 - Y_i)$

  lower threshold = λ_{lower}-quantile$(\mathcal{M}(X_i)|Y_{i,v} < 0.5)$

  append $\max$(lower threshold, 0.5) to lower thresholds

end

upper bound$_t \leftarrow -(1-\alpha/2)$-quantile of $-$upper thresholds

lower bound$_t \leftarrow (1-\alpha/2)$-quantile of lower thresholds

for $j \in \{m+1, ..., l\}$ **do**

  $\mathcal{U}_{\alpha,\lambda}(X_j) \leftarrow \mathcal{M}(X_j)_{\geq \text{upper bound}_t}$

  $\mathcal{L}_{\alpha,\lambda} \leftarrow \mathcal{M}(X_j)_{\geq \text{lower bound}_t}$

end

volume prediction error $\text{Vol}\,(\mathcal{M}(X_i)_{\geq .5}) - \text{Vol}\,(Y_i)$, which suggests that adaptive interval lengths are indeed important for strong results. Figure 3 (low) shows that, in our test data, the empirical coverage of Standard Volume Prediction and its normalized version are closer to the nominal coverage $1 - \alpha = .9$ than for Thresholded Volume Prediction.

4.3. Performance of shape-predictive regions

Now, consider Algorithm 3 with different values of the leniency parameter λ. Figure 5 shows the average value of the difference of volumes $\text{Vol}\,(\mathcal{U}_{\alpha,\lambda}(X_j)) - \text{Vol}\,(\mathcal{L}_{\alpha,\lambda}(X_j))$ for different confidence levels and leniencies. This is not

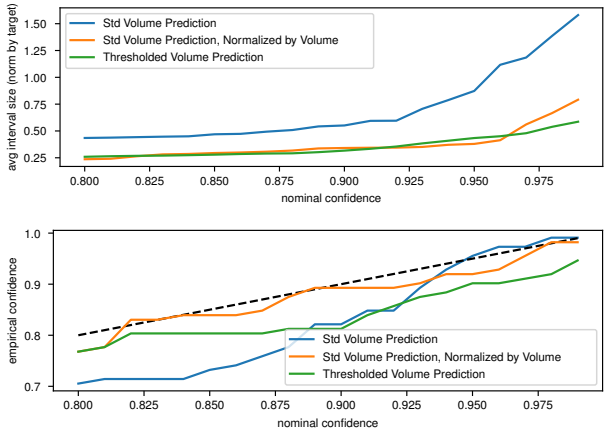

Figure 3. Average interval sizes (normalized by target volume) for different nominal confidences (top) and empirical vs. nominal coverage (bottom).

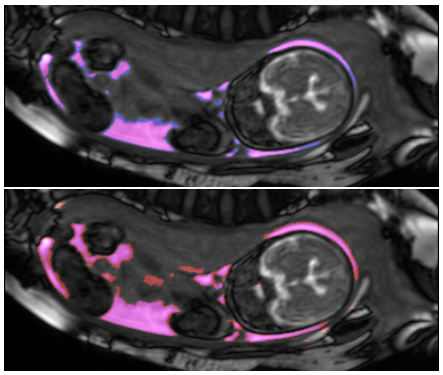

Figure 4. Segmentation Predictionwith $\alpha = 0.1$ and $\lambda = 0.05$. Up shows $Y_j \setminus \mathcal{L}_{\alpha,\lambda}(X_j)$ in blue and $Y_j \cap \mathcal{L}_{\alpha,\lambda}(X_j)$ in magenta. Down shows $\mathcal{U}_{\alpha,\lambda}(X_i) \setminus Y_j$ in red and $Y_i \cap \mathcal{U}_{\alpha,\lambda}(X_i)$ in magenta.

the same as the length of a volume-predictive interval, since part of Y_j might not lie between $\mathcal{U}_{\alpha,\lambda}(X_j)$ and $\mathcal{L}_{\alpha,\lambda}(X_j)$. Larger leniency leads to narrower intervals.

Figure 4 illustrates the typical behavior of Algorithm 3. We find that $\mathcal{L}_{\alpha,\lambda}(X_j)$ only misses boundary areas and narrow regions of the AF, whereas $\mathcal{U}_{\alpha,\lambda}(X_j)$ moderately enlarges the segmentation and adds small disconnected areas.

5. Discussion

The segmentation of AF in MRI is prototypical of a class of medical imaging problems that can be effectively done using neural networks. Still, a key challenge is guaranteeing high-confidence, informative bounds on the performance of different methods. Our work is a first step in this direction.

We found that volume-predictive intervals with adaptive sizes can be substantially narrower than the original proposal by (Lei et al., 2018), while maintaining good coverage. The results for shape-predictive regions are promising, and suggest that methods such as Algorithm 3 might visually aid radiologists in MRI segmentation. Crucially, both methods come with theoretical guarantees.

There are two important directions for further study. One is to design novel models that may aid the downstream

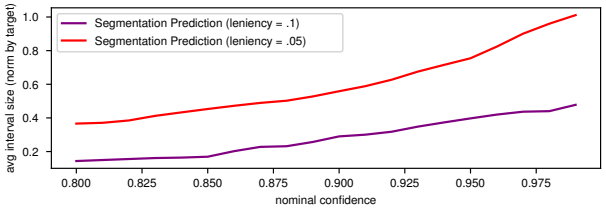

Figure 5. Average interval sizes for different leniency values.

Conformal Prediction method. Although CP is general, its practical performance may depend heavily on the method used. Different architectures, training regimes or losses might lead to improved results.

A second direction is obtaining stronger theoretical guarantees. Recent work in classification (Cauchois et al., 2021) has noted that CP methods may achieve coverage by ignoring hard examples. Similarly improving our algorithms would be of great theoretical and practical interest.

References

Bates, S., Angelopoulos, A. N., Lei, L., Malik, J., and Jordan, M. I. Distribution-free, risk-controlling prediction sets. *arXiv preprint arXiv:2101.02703*, 2020.

Cauchois, M., Gupta, S., and Duchi, J. C. Knowing what you know: valid and validated confidence sets in multiclass and multilabel prediction. *Journal of Machine Learning Research*, 22(81):1–42, 2021. URL http://jmlr.org/papers/v22/20-753.html.

Chamberlain, P., Manning, F., Morrison, I., and et al. Ultrasound evaluation of amniotic fluid volume i. the relationship of marginal and decreased amniotic fluid volumes to perinatal outcome. *Am J Obstet Gynecol*, 1(150(3)): 245–249, 1984.

Chen, X., Williams, B. M., Vallabhaneni, S. R., Czanner, G., Williams, R., and Zheng, Y. Learning active contour models for medical image segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 11632–11640, 2019.

Hellinger, J. and Epelman, M. Fetal mri in the third dimension. *Applied Radiology*, 39:8–22, 2010.

Iqbal, H. PlotNeuralNet. https://github.com/HarisIqbal88/PlotNeuralNet, 2018.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kubik-Huch, R., Wildermuth, S., Cettuzzi, L., Rake, A., Seifert, B., and et al., R. C. Fetus and uteroplacental unit: fast mr imaging with three-dimensional reconstruction and volumetry-feasibility study. *Radiology*, 219:567–573, 2001.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018. doi: 10.1080/01621459.2017.1307116. URL https://doi.org/10.1080/01621459.2017.1307116.

Magann, E., Chauhan, S., Whitworth, N., and et al. Subjective versus objective evaluation of amniotic fluid volume of pregnancies of less than 24 week's gestation: how can we be accurate? *J Ultrasound Med.*, 20:191–195, 2001.

Moore, T. The role of amniotic fluid assessment in evaluating fetal well-being. *Clinics in perinatology*, 38:33–46, 2011.

Moore, T. and Cayle, J. The amniotic fluid index in normal human pregnancy. *Am J Obstet Gynecol.*, 162:1168–1173, 1990.

Moschos, E., Gullmar, D., Fiedler, A., and et al. Comparison of amniotic fluid volumetry between fetal sonography and mri – correlation to mr diffusion parameters of the fetal kidney. *Birth Defect*, 1(1):1–7, 2017.

Phelan, J. and Smith, C. Amniotic fluid volume assessment with the four-quadrant technique at 36-42 week's gestation. *J Reprod Med*, 32:540–542, 1987.

Phelan, J., Ahn, M., Smith, C., and et al. Amniotic fluid index measurements during pregnancy. *J Reprod Med*, 32:601–604, 1987.

Poudel, R., Liwicki, S., and Cipolla, R. Fast-scnn: Fast semantic segmentation network. In *30th British Machine Vision Conference 2019, BMVC 2019*, 2019.

Prayer, D., Brugger, P., and Prayer, L. Fetal mri: techniques and protocols. *Pediatr Radiol*, 34:685–693, 2004.

Ronneberger, O., P.Fischer, and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 9351 of *LNCS*, pp. 234–241. Springer, 2015. URL http://lmb.informatik.uni-freiburg.de/Publications/2015/RFB15a. (available on arXiv:1505.04597 [cs.CV]).

Shafer, G. and Vovk, V. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(12):371–421, 2008. URL http://jmlr.org/papers/v9/shafer08a.html.

Werner, H., Lopes, J., and Junior, E. A. Physical model from 3d ultrasound and magnetic resonance imaging scan data reconstruction of lumbosacral myelomeningocele in a fetus with chiari ii malformation. *Childs Nerv Syst*, 31(4):511–513, 2015.