AmnioML: Amniotic Fluid Segmentation and Volume Prediction with Uncertainty Quantification

Daniel Csillag[1], **Lucas Monteiro Paes**[2], **Thiago Ramos**[1], **João Vitor Romano**[1], **Rodrigo Schuller**[1], **Roberto B. Seixas**[1], **Roberto I. Oliveira**[1], **Paulo Orenstein**[1]

[1] IMPA, Rio de Janeiro, Brazil
[2] Harvard University, Cambridge, USA
daniel.csillag@impa.br, lucaspaes@g.harvard.edu, thiagorr@impa.br, joao.vitor@impa.br, rodrigo.loro@impa.br, rbs@impa.br, rimfo@impa.br, pauloo@impa.br

Abstract

Accurately predicting the volume of amniotic fluid is fundamental to assessing pregnancy risks, though the task usually requires many hours of laborious work by medical experts. In this paper, we present AmnioML, a machine learning solution that leverages deep learning and conformal prediction to output fast and accurate volume estimates and segmentation masks from fetal MRIs with Dice coefficient over 0.9. Also, we make available a novel, curated dataset for fetal MRIs with 853 exams and benchmark the performance of many recent deep learning architectures. In addition, we introduce a conformal prediction tool that yields narrow predictive intervals with theoretically guaranteed coverage, thus aiding doctors in detecting pregnancy risks and saving lives. A successful case study of AmnioML deployed in a medical setting is also reported. Real-world clinical benefits include up to 20x segmentation time reduction, with most segmentations deemed by doctors as not needing any further manual refinement. Furthermore, AmnioML's volume predictions were found to be highly accurate in practice, with mean absolute error below 56mL and tight predictive intervals, showcasing its impact in reducing pregnancy complications.

Introduction

Amniotic fluid (AF), the liquid surrounding the fetus during gestation, is a leading indicator of a pregnancy's health. For instance, it cushions the fetus against mechanical trauma, serves as a reservoir of fluids and nutrients, affects the development of lungs and limbs and provides antibacterial protection (Beloosesky and Ross 2018). Abnormal AF volumes are linked to several pregnancy complications (Chamberlain et al. 1984a; Moore 2011; Moore and 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. On the other hand, reduced AF volume, or oligohydramnios, occurs in 12% of cases, and corresponds to a 15 to 50-fold increase in perinatal morbidity and mortality. Thus, fast and accurately predicting AF volume is crucial to allow doctors to act as soon as possible to ensure a healthy pregnancy.

While doctors oftentimes estimate AF volume via a visual inspection of ultrasound exams, precise estimation typically requires the use of higher-quality magnetic resonance images (MRI) (Hellinger and Epelman 2010; Kubik-Huch et al. 2001; Moschos et al. 2017; Prayer, Brugger, and Prayer 2004). After obtaining the images, medical experts segment the areas in each slice of the MRI containing AF (see Figure 1) and, from this segmentation, the AF volume can be estimated. This process is laborious, long and requires specialized training. Moreover, it can take many hours, if not days, at which point the pregnancy might have changed altogether.

The main objectives of our work were to:

- Segment the amniotic fluid in fetal MRI exams with accuracy equal or superior to trained doctors;
- Drastically reduce segmentation time;
- Provide uncertainty quantification for segmentations and volume estimates given a prescribed coverage level;
- Allow doctors to dynamically set the coverage level;
- Provide an integrated and easy-to-use interface;
- Ensure faster diagnostics and improve patients care.

Given such demanding and nontrivial objectives, it was important to leverage modern AI solutions, develop custom techniques and combine them in innovative ways to achieve our goals. In order to reach human-level segmentation accuracy, we trained a state-of-the art convolutional neural network based on the U-Net architecture and optimized it to the AF setting. Although searching for the ideal architecture and hyperparameters is time-consuming, inference is very fast. Uncertainty quantification was achieved via conformal prediction (Vovk, Gammerman, and Shafer 2005) and novel ramifications we developed specifically for this task, all of which allow doctors to choose a desired coverage level. This empowers doctors to set custom alarms that can automatically alert when a patient needs further exams. Last but not least, the full solution was bundled as a plugin for 3D Slicer, a popular software for medical segmentation.

Our main contributions in this paper are:

- **AmnioML:** A fast and accurate solution capable of automatically segmenting AF and estimating volume from fetal MRIs. Both point and interval estimates for the volume are provided. Real-world use of AmnioML has shown it is capable of significantly improving doctors' ability to track AF changes through a pregnancy;

- **Fetal MRI Dataset:** A novel dataset with 853 pairs of MRI exams and their corresponding AF segmentation, along with other non-identifying patient information;
- **Fetal MRI Benchmarks:** To evaluate AmnioML, we construct, evaluate and compare several AF segmentation models based on recent neural network architectures, and benchmark their performance on our Fetal MRI Dataset;
- **Conformal Prediction (CP):** We introduce a CP tool tailored to the medical segmentation task, and compare it to different methods for AF volume estimation in terms of empirical coverage and interval length.

Related Work

Since AF plays an essential role in human fetal growth and development (Beloosesky and Ross 2018; Moore 2011), many studies have tried to relate AF volume to perinatal outcome (Baron, Morgan, and Garite 1995; Bakhsh et al. 2021). Qualitative results show that decreased or increased AF volumes were significantly related to incidences of major congenital anomaly and intrauterine growth retardation (Chamberlain et al. 1984a,b), with severe oligohydramnia linked to perinatal mortality (Bastide et al. 1986). There are also quantitative results providing normative data for AF volume in pregnancies (Queenan et al. 1972; Moore and Cayle 1990).

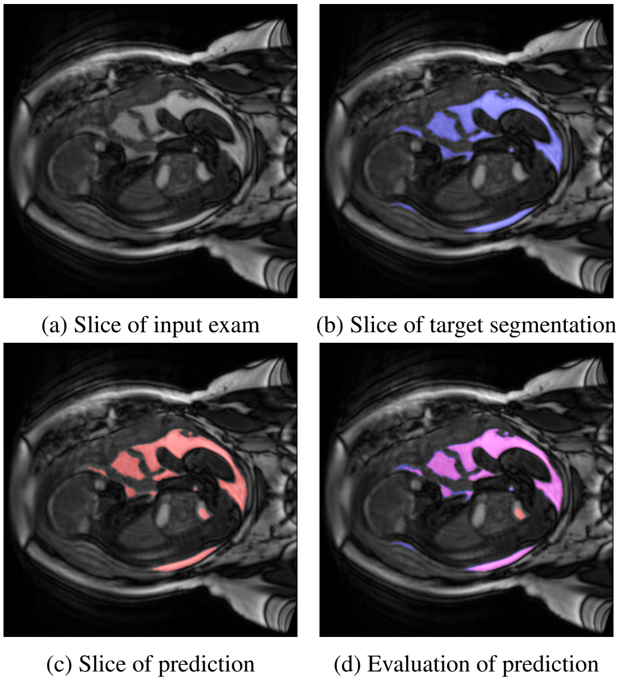
Deep learning methods have been widely explored in medical applications (Looney et al. 2021; Ayu et al. 2021; Lee, Yamanakkanavar, and Choi 2020). Such tools often rely on highly optimized convolutional neural network architectures (Liu et al. 2017), which we also use for our model (Ronneberger, Fischer, and Brox 2015) or as baselines (Chen et al. 2018a,b; Poudel, Liwicki, and Cipolla 2019; Shang et al. 2020; Khosravan et al. 2019; Zhao et al. 2017; Zhou et al. 2020). Typically, these applications employ ultrasound images (Looney et al. 2021; Ayu et al. 2021; Cho et al. 2021), while we use MRIs due to their higher resolution, allowing a more precise volume estimate, following (Prayer, Brugger, and Prayer 2004; Kubik-Huch et al. 2001; Hellinger and Epelman 2010; Shen et al. 2022).

Finally, quantifying the uncertainty in machine learning estimates has received much attention lately, particularly in medical applications (Edupuganti et al. 2021). In this paper, we make use of tools from the field of conformal prediction (Bates et al. 2021; Barber et al. 2020; Lei et al. 2018). We show that theoretical guarantees from CP (Vovk, Gammerman, and Shafer 2005; Shafer and Vovk 2008) are very useful in this setting and provide efficient algorithms to aid medical professionals in quantifying the uncertainty present in machine learning predictions.

Notation

The set $\{(X_i, Y_i)\}_{i=1}^n$ will denote an independent and identically distributed sample of pairs of 3D exams X_i and their corresponding medical segmentation Y_i (see Figure 1), with v voxels. Data indices will be partitioned into $I_{\text{train}} = \{1, \ldots, n_{\text{train}}\}$, $I_{\text{cal}} = \{n_{\text{train}} + 1, \ldots, n_{\text{train}} + n_{\text{cal}}\}$ and $I_{\text{test}} = \{n_{\text{train}} + n_{\text{cal}} + 1, \ldots, n_{\text{train}} + n_{\text{cal}} + n_{\text{test}}\}$, with $n_{\text{train}} + n_{\text{cal}} + n_{\text{test}} = n$ and $n_{\text{train}}, n_{\text{cal}}, n_{\text{test}} \in \mathbb{N}_+$. Any model $\mathcal{M}$ will always be trained on $\{(X_i, Y_i)\}_{i \in I_{\text{train}}}$ and evaluated on $\{(X_i, Y_i)\}_{i \in I_{\text{test}}}$. The purpose of the calibration set $\{(X_i, Y_i)\}_{i \in I_{\text{cal}}}$ will be to quantify the uncertainty of any given trained model and aid in the generation of predictive sets for unseen pairs. Hyperparameter optimization is performed on a subset of I_{train} named validation set.

Table 1 contains further notation used throughout the text.


(a) Slice of input exam (b) Slice of target segmentation

(c) Slice of prediction (d) Evaluation of prediction

Figure 1: Example of an MRI slice, displaying (a) the input slice, (b) the target (blue), (c) AmnioML's prediction (red) and (d) comparison of prediction and target, where magenta indicates agreement between prediction and target, blue indicates missing regions and red excessive segmentation.

Dataset

The Fetal MRI Dataset includes $n = 853$ fetal MRI exams performed from January 2015 to December 2021, with gestational age between 19 and 38 weeks; over 65% of the subjects present some degree of pathology, such as malformations, obstructions and tumors. The exams include AF segmentation masks, the gestational week and any patient pathology present. Exams that display significant motion artifacts were included for both training and testing purposes.

Symbol	Description	Domain
X	3D MRI exam	$[0, 1]^v$
Y	AF segmentation mask	$\{0, 1\}^v$
Vol(Y)	Exam volume	$\mathbb{R}_{\geq 0}$
$\mathcal{M}(X)$	Trained model output	$[0, 1]^v$
$\mathcal{M}(X)_{\geq t}$	Model output thresholded at t	$\{0, 1\}^v$
$A \odot B$	Element-wise multiplication	$\mathbb{R}^v \times \mathbb{R}^v \to \mathbb{R}^v$

Table 1: Notation used in the text.

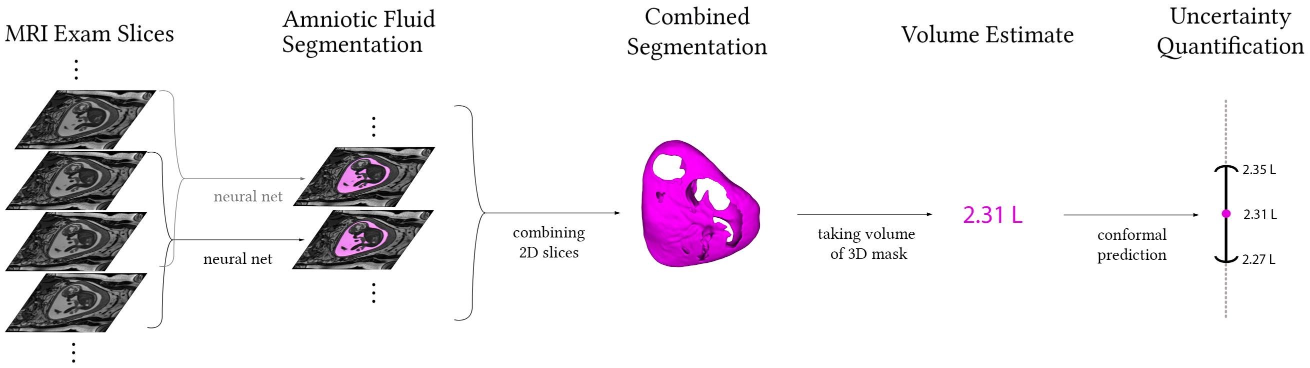


Figure 2: AmnioML pipeline: given a fetal MRI exam, amniotic fluid (AF) in each slice is segmented using a U-Net based neural network and combined to form a 3D mask. AmnioML outputs the AF volume as well as predictive intervals, using conformal prediction.

The MRI images $\{X_i\}_{i=1}^n$ are 3D images with varying number of voxels v, produced by a 1.5-T scanner, with 3D reconstruction protocol T2-weighted true fast imaging with a steady-state precession (TrueFisp) sequence in the sagittal plane, FOV 380mm, voxel size: $1 \times 1 \times 1\text{mm}$ and an acquisition time of 0.26s. Maternal sedation was not used to capture the images, which can impact the quality of the images due to fetal movements.

The MRI exams were carefully segmented by two medical specialists, under the supervision of a third specialist, along with the radiologist that performed the exams. Figure 1 displays a single slice of an MRI exam, along with the target (i.e., the human-annotated segmentation for the AF). Whenever one of the supervisors disagreed with the proposed segmentation, it was either refined or discarded.

The set of exam and segmentation pairs $\{(X_i, Y_i)\}_{i=1}^n$, with $n = 853$, was divided into three disjoint sets: $\{(X_i, Y_i)\}_{i=1}^{n_{\text{train}}}$ for training, $\{(X_i, Y_i)\}_{i=n_{\text{train}}+1}^{n_{\text{train}}+n_{\text{cal}}}$ for calibration and $\{(X_i, Y_i)\}_{i=n_{\text{train}}+n_{\text{cal}}+1}^{n}$ for test, with $n_{\text{train}} = 656$, $n_{\text{cal}} = 105$ and $n_{\text{test}} = 92$. Different exams from the same pregnancy were included in the same data fold to avoid any potential snooping through maternal fixed effects.

All exams provided were curated before being added to the Fetal MRI Dataset. A system of filters ran through the exams to remove dimension mismatches and pairs (X_i, Y_i) whose affine transformations from voxel to real-world coordinates were inconsistent. The resulting Fetal MRI Dataset is available at https://w3.impa.br/%7E daniel.csillag/projects/amnioml/dataset. Each exam and corresponding segmentation are stored as nrrd files with gzip compression, inside train, calibration and test sets folders, with respective sizes 8.2GB, 1.2GB and 1.2GB. Also included are: a csv file linking exams to each patient id, and another csv file containing exam metadata, such as the patient's gestational week and relevant pathologies.

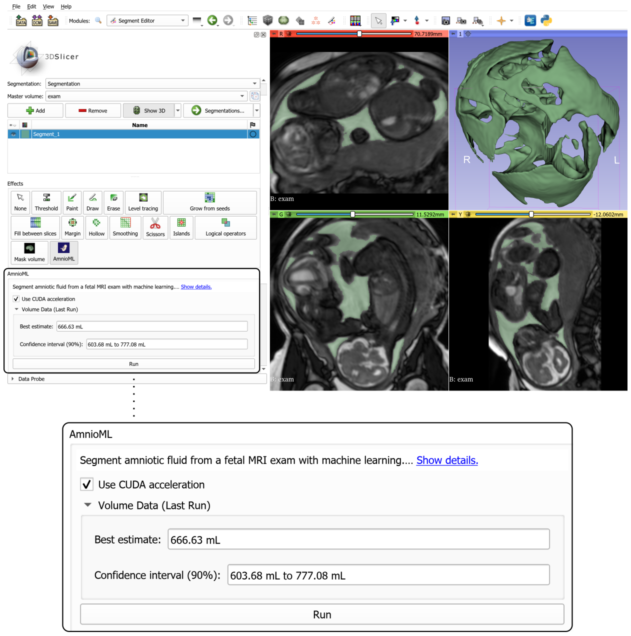
AmnioML

We developed AmnioML, a tool for automatic AF segmentation and volume estimation, with valid predictive intervals and segmentation masks (see Figure 2). Our solution leverages several advances in deep learning to provide an algorithm that can run under 6 seconds in most GPUs and with average Dice coefficient (Eq. 1) when segmenting AF in MRI slices superior to 0.9, displaying wide agreement with medical experts, but in much shorter time. Volume estimates are produced from the 3D segmentation mask, with mean absolute error of around 55mL — small enough to properly identify high-risk AF levels. By building on top of conformal prediction theory, AmnioML also produces valid predictive intervals for volume, which are crucial in quantifying the uncertainty associated to the estimates. Providing predictive sets for these methods allows doctors to better control and interpret the resulting segmentation masks and volume estimates, detect important anomalies and image artifacts, and be automatically alerted to abnormal levels of AF.

AmnioML is available as a 3D Slicer plugin, a popular software for medical segmentation, though we also make its source code available. In real-world usage, medical specialists have used AmnioML as an aid in their manual segmentation, reporting significant speedups of up to 20x; in most cases, no human post-processing was deemed necessary.

To train and test AmnioML, we collect 853 fetal MRI exams, annotated by medical professionals, and make the resulting dataset publicly available. Annotations include the gestational week, pathologies exhibited throughout the pregnancy and AF segmentation masks. Exams were collected from multiple patients, with gestational age between 19 to 38 weeks, and with over 65% of pregnancies displaying some degree of pathology. We anticipate this dataset can be used for other important fetal tasks, such as brain and lung segmentation, as well as fetus' weight estimation.

The AmnioML tool outputs both point and interval estimates for volume prediction (see Figure 2). AmnioML comprises a convolutional neural network developed in PyTorch, custom-built conformal methods built in Python for uncertainty quantification and a 3D Slicer plugin for on-the-fly deployment, also in Python. The source code is available at https://github.com/dccsillag/amnioml.



AmnioML

Segment amniotic fluid from a fetal MRI exam with machine learning.... Show details.

☑ Use CUDA acceleration

▾ Volume Data (Last Run)

Best estimate: 666.63 mL

Confidence interval (90%): 603.68 mL to 777.08 mL

Run

Figure 3: GUI for AmnioML's 3D Slicer plugin, outputting the segmentation masks and the combined AF solid, best volume estimate and the accompanying predictive interval.

Methodology

Segmentation Model

To create an estimate of the volume of AF in an MRI exam, each 2D slice of the exam (along the sagittal axis) is segmented using a neural network. Then, the slices' segmentations are combined into a 3D object and the volume estimated. Segmenting slices instead of entire exams allows for many more training samples, lower memory requirements and faster training time.

Each MRI exam $X \in [0,1]^v$, with $v = v_1 \times v_2 \times v_3$, is decomposed into v_3 slices (see Figure 1 for such a slice). To segment a single slice in the exam, say slice v_3^*, slices $v_3^* - 1$, v_3^* and $v_3^* + 1$ are used as features to provide motion context (using fewer or more slices, even further spread out, did not aid prediction in the validation set). Each feature set of three slices is reshaped to have dimensions $256 \times 256 \times 3$, and are used as inputs to segment the middle slice v_3^*.

A U-Net neural network (Ronneberger, Fischer, and Brox 2015) is used to predict the AF, trained with 17 million parameters, binary cross-entropy loss, Adam optimizer with learning rate of 10^{-3}, and a maximum of 100 epochs with early stopping of patience 7 on the validation loss. The output of the U-Net is thresholded at 0.5 to yield a segmentation mask. Hyperparameters were tuned via grid search and picked by the Dice loss on the validation set. The model was trained using PyTorch on 8 TPUs v3 under 6 hours.

After segmenting each slice in the exam, the 2D segmentations are stacked into a 3D object (see Figure 2). Adding the voxels in the 3D mask yield the estimate for the AF.

Volume Uncertainty Quantification

To compute a prediction interval for the volume, Algorithm 1 is employed. It is motivated by conformal prediction tools and is split into two consecutively-run procedures.

The calibration part selects the best pair of upper and lower-bound thresholds, $u, l \in \mathbb{R}$ such that the following holds for at least $\lceil (1-\alpha) n_{\text{cal}} \rceil$ exams from the calibration dataset:

$$\text{Vol}(Y_i) \in \left[\text{Vol}(\mathcal{M}(X_i)_{\geq u}), \text{Vol}(\mathcal{M}(X_i)_{\geq l})\right].$$

Then, the prediction step applies the precomputed pair of thresholds (u, l) to a particular prediction $\mathcal{M}(X_j)$, and returns the $(1-\alpha)$-predictive interval

$$\mathcal{I}_\alpha^{\text{tvp}} = \left[\text{Vol}(\mathcal{M}(X_j)_{\geq u}), \text{Vol}(\mathcal{M}(X_j)_{\geq l})\right].$$

Note that, for any given α, the calibration step runs only once, after the model has trained. The bounds u and l are then stored and repeatedly applied in the prediction step. Thus, once AmnioML is deployed, the calibration step does not have to be rerun.

Theorem 1, based on the conformal prediction literature, shows that Algorithm 1 produces intervals with valid theoretical coverage.

Theorem 1. *For a coverage level $1-\alpha$, take $\mathcal{I}_\alpha^{\text{tvp}}$ as in Algorithm 1, using $\{(X_i, Y_i)\}_{i \in \text{cal}}$ as calibration set. Then, for any $j \in I_{\text{test}}$,*

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

Hence, for any new MRI exam, AmnioML is able to quickly produce intervals at any coverage level $1-\alpha$ picked by the user. The proof of Theorem 1 resembles the standard techniques for split conformal prediction, but exploits the structure of the segmentation masks thresholded at higher and lower levels. For the complete proof, see https://github.com/dccsillag/amnioml/blob/main/proofs.pdf.

Algorithm 1: Thresholded Volume Prediction

```
Procedure calibrate(model $\mathcal{M}$, calibration set $\{(X_i, Y_i)\}_{i \in I_{\text{cal}}}$, coverage $1-\alpha \in (0,1)$)
    thresholds ← [ ]
    for $i \in I_{\text{cal}}$ do
        $p$ ← proportion of zeros in $Y_i$
        best threshold ← $p$-quantile($\mathcal{M}(X_i)$)
        append best threshold to thresholds
    end
    $\phi_u = \lceil (n_{\text{cal}} + 1)(1 - \alpha/2) \rceil / n_{\text{cal}}$
    $\phi_l = \lceil (n_{\text{cal}} + 1)(\alpha/2) \rceil / n_{\text{cal}}$
    $u$ ← $\phi_u$-quantile of thresholds
    $l$ ← $\phi_l$-quantile of thresholds
    return $l, u$
Procedure predict(prediction $\mathcal{M}(X)$, $l$, $u$)
    lower volume ← $\text{Vol}(\mathcal{M}(X)_{\geq u})$
    upper volume ← $\text{Vol}(\mathcal{M}(X)_{\geq l})$
    return [lower volume, upper volume]
```

Baselines

This subsection details segmentation and volume prediction baselines against which AmnioML is compared, as well as traditional split conformal prediction techniques that did not deliver great results and prompted us to develop our novel solution, which borrows from the conformal literature but focuses on the AF problem (Algorithm 1).

Several different highly regarded semantic segmentation architectures were considered as alternatives:

- Fast-SCNN (Poudel, Liwicki, and Cipolla 2019): small but efficient segmentation network;
- DeepLabV3 (Chen et al. 2018a): developed for multi-scale context features;
- DeepLabV3+ (Chen et al. 2018b): improvement over DeepLabV3 with a decoder module for better boundary segmentation;
- MANet (Shang et al. 2020): promising approach for different target sizes;
- PAN (Khosravan et al. 2019): developed for pancreas segmentation;
- PSPNet (Zhao et al. 2017): advanced usage of local and global context; and
- U-Net++ (Zhou et al. 2020): an ensemble of U-Nets of varying parameters.

These models generally employ between 20 and 30 million parameters (see Table 2). For each of these models, the same slice preprocessing strategy described in the beginning of this section was employed. Models were trained using the Adam optimizer, for a maximum of 100 epochs, with early stopping of patience 7. Hyperparameters were tuned via grid search to optimize the validation Dice coefficient.

Regarding volume estimation as a regression problem, it is straightforward to apply traditional split conformal prediction (Papadopoulos et al. 2002; Vovk, Gammerman, and Shafer 2005; Lei et al. 2018) to generate valid predictive intervals. Indeed, we evaluated split CP with several nonconformity scores, but our custom method (Algorithm 1) outperformed them in terms of average interval size for high levels of nominal coverage.

Model	# of Parameters	Avg. Dice Coeff.
Fast-SCNN	$1.1 \cdot 10^6$	0.83 ± 0.10
U-Net	$1.7 \cdot 10^7$	$\mathbf{0.91 \pm 0.06}$
PSPNet	$2.1 \cdot 10^7$	0.87 ± 0.8
PAN	$2.1 \cdot 10^7$	$\mathbf{0.91 \pm 0.06}$
DeepLabV3+	$2.2 \cdot 10^7$	0.90 ± 0.07
DeepLabV3	$2.6 \cdot 10^7$	0.87 ± 0.08
U-Net++	$2.6 \cdot 10^7$	$\mathbf{0.91 \pm 0.05}$
MANet	$3.1 \cdot 10^7$	$\mathbf{0.91 \pm 0.07}$

Table 2: Performance comparison of segmentation models on the test set, sorted by number of parameters. Best Dice coefficients are highlighted.

Evaluation

Segmentation and Volume Estimation

The metric used to evaluate a segmentation mask $\mathcal{M}(X)_{\geq .5}$ is the Dice coefficient,

$$\text{Dice}(\mathcal{M}(X)_{\geq .5}, Y) := 2 \cdot \frac{\text{Vol}(\mathcal{M}(X)_{\geq .5} \odot Y)}{\text{Vol}(\mathcal{M}(X)_{\geq .5}) + \text{Vol}(Y)}, \quad (1)$$

which measures the degree to which the algorithm's masks intersects with the ones proposed by the medical specialists. Here, $\odot$ denotes the intersection (or the element-wise product) of two masks. Note that, besides being a popular metric for segmentation, the Dice coefficient is naturally related to the volume of the segmented region which, in the case of AF, is our ultimate goal.

We evaluate AmnioML's U-Net segmentation model against the baselines described in the previous section. All models underwent hyperparameter tuning in the validation set, specifically tailoring to the Dice coefficient.

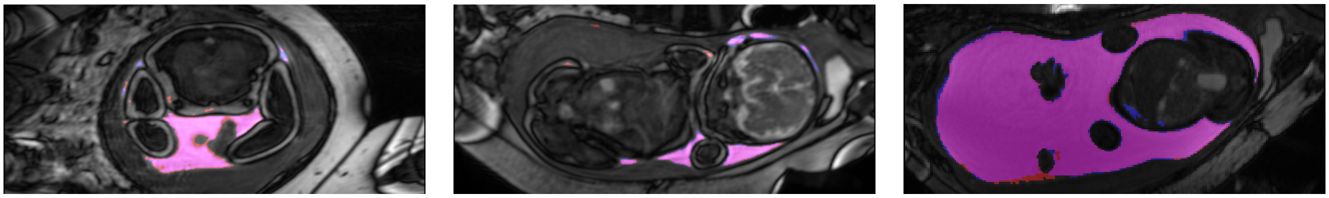
Table 2 illustrates their test set performance, in terms of average Dice coefficient of the segmentation masks over all slices in all test exams. U-Net, as employed by AmnioML, had the best average Dice coefficient tied with the PAN, UNet++ and MANet architectures. Still, the U-Net uses fewer parameters so it is faster to train and, crucially for medical applications, has lower inference time. AmnioML's network runs under 6s in GPUs and 35s in a modern CPU.

While AmnioML is quite capable at segmenting the AF, it is medically important to understand where most of its errors happen. Figure 4 illustrates the typical situation, where mistakes happen along the segmentation's borders. This is encouraging for two reasons: (i) there are degrees of subjectivity in segmenting the borders even for humans, due to artifacts of motion, noise or insufficient image clarity; (ii) the borders contribute relatively little to the overall volume. To further quantify the extent to which AmnioML's errors happen along the borders, we note that a 2-voxel dilation along the masks' borders reduces the missing volume by 71% on average, and a 2-voxel erosion reduces the excess volume by 87% on average.

It is possible to create an estimate for the volume by adding the values of all voxels in the segmentation mask. Figure 5 shows how AmnioML's volume predictions fare against the medical specialists. There is general agreement between them, with mean absolute error of only 56mL.

Uncertainty Quantification Evaluation

Predictive intervals for Algorithm 1 are assessed through two main criteria: (i) the average interval length, and (ii) comparisons between nominal and empirical coverages. A good algorithm is expected to produce tight intervals, with empirical coverage close to the theoretical value of $1 - \alpha$. Figure 6 (left) shows the average interval size for volume predictions (normalized by target volume). As expected, higher coverage levels require larger interval sizes. Still, note AmnioML's Algorithm 1 enjoys small normalized interval lengths for the entire range of coverage levels tested. Standard split CP sees a significant increase in interval sizes for high confidences, above 0.95.



(a) Example of hard exam (Dice: 0.83). (b) Example of typical exam (Dice: 0.92). (c) Example of an easy exam (Dice: 0.96).

Figure 4: The region correctly segmented by the AmnioML is in magenta, with blue indicating missing regions and red excessive segmentation. Typically, errors occur along the mask's borders.

Figure 6 (right) illustrates how the empirical coverage achieved matches the prescribed coverage level, as expected from Theorem 1.

To further ascertain the effectiveness of AmnioML's Algorithm 1, Figure 5 presents a visual comparison between AF volume predicted by AmnioML and by the medical specialists' segmentations, along with the prediction intervals of Algorithm 1. The diagonal dashed line indicates perfect prediction. Note that all points are close to the dashed line and that the intervals are small.

Application Use and Payoff

To evaluate the quality of our plugin in a professional setting, 80 new predictions were performed by a medical specialist employing AmnioML, and each was rated from 1 to 5 via the following scale:

1. Worse than automatic thresholding;
2. Same quality as automatic thresholding;
3. A lot of manual adjustments were necessary;
4. A few manual adjustments were necessary; and
5. No manual adjustments were necessary.

Ratings (1) and (2) compare AmnioML against thresholding, a popular first-step color filtering technique commonly used in fetal segmentation but requiring extensive refinement. Ratings (3)-(5) indicate the level of manual work required to post-process AmnioML's automatic segmentation beyond what is provided by simple thresholding.

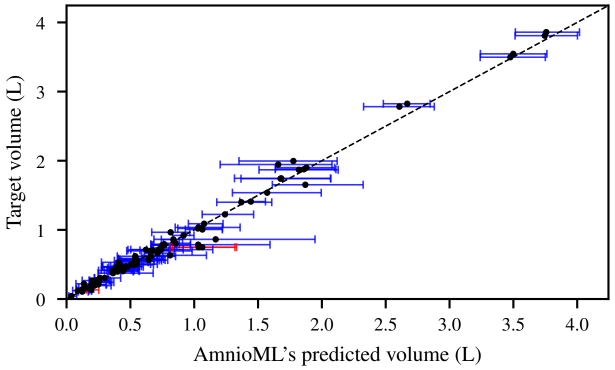


Figure 5: Comparison between AmnioML's volume prediction and target, at 90% coverage. Blue lines indicate a target volume inside AmnioML's predictive interval.

Eighty predictions from different patients were evaluated using the aforementioned scale. The ratings (1), (2), (3), (4) and (5) were tallied 0, 1, 8, 21 and 50 times respectively. Crucially, close to 90% of the AmnioML predictions required few or no manual refinements, with over 60% requiring no further human input whatsoever.

We also measured the decrease in segmentation time reported by medical experts with the help of AmnioML. On average, radiologists were 20.7 times faster (see Figure 7), reducing the time an expert has spent on an exam from an average of around 45 minutes down to 2 minutes. Likewise, harder tasks that could previously take hours or days were now finished in a few minutes. Perhaps more importantly, segmenting the MRI exams cannot be done within the time of typical medical appointment, and patients often wait for days to hear the results of the MRI. By integrating AmnioML with MRI machines, it will be possible to give a meaningful estimate of AF fluid within the same consultation. In particular, this could allow the doctor to quickly refer the patient for further exams, if necessary.

Deployment and Maintenance

AmnioML was deployed as an offline plugin of type *Segmentation Effect* for the open-source tool 3D Slicer (Fedorov et al. 2012), which has widespread use in the medical segmentation community. Developing it as an offline tool also allowed us to forgo important privacy considerations that could change from country to country. AmnioML's GUI was designed with simplicity in mind, and has two interactive elements: a checkbox for CUDA use and a run button that performs the segmentation (Figure 3).

The interface allows users to send textual feedback. For instance, with the user's consent, whenever an AmnioML's segmentation is rated poorly, we receive a notification including a hash of the exam, its rating and the user review. This has allowed us to continually improve our system. One such example included a cyst that our algorithm had not learned to differentiate from regular AF, and prompted us to start working on identifying such bodies of fluid.

Designing AmnioML as a plugin integrates it with other processing pipelines available at 3D Slicer's extensive plugin ecosystem and ensures that the disposition of the segmentation is coherent across different UI elements. We also released a version of AmnioML that runs independently of

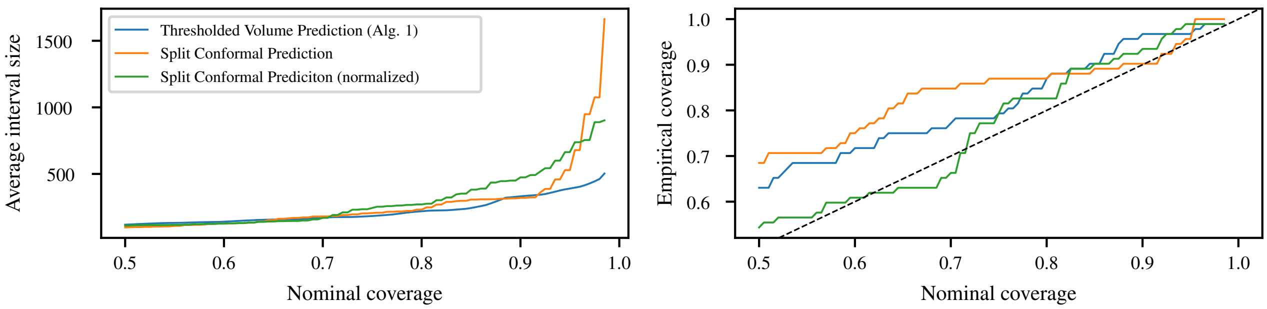


Figure 6: Uncertainty Quantification algorithms compared by average predictive interval length (left) and empirical coverage (right). Thresholded Volume Prediction (Algorithm 1) produces tighter predictive intervals compared to standard Split Conformal Prediction, while still attaining the prescribed coverage levels.

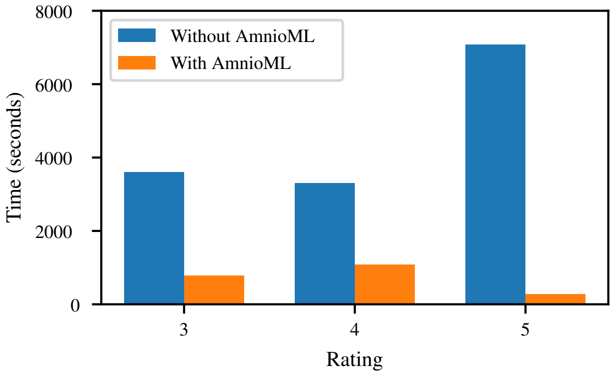


Figure 7: Segmentation times with and without the aid of AmnioML, according to the rating of the predictions (ratings of 1 and 2 were negligible and are not displayed). The average time reduction was 20×.

3D Slicer for broader outreach; this way, clinical diagnostics companies are free to deploy it internally and let their doctors and radiologists install our tool in whatever environment they use for medical segmentation.

On average in the test dataset, a GTX 1060 GPU (with CUDA 10.2) segments an entire MRI exam in 5.49s, and a Ryzen 5600X CPU executes the same operation in 34.76s. This performance profile makes running AmnioML offline viable, which, in turn, makes it simpler to maintain, privacy-friendly and more portable.

AmnioML was developed with the help of medical experts at DASA, one of the largest clinical diagnostics companies in the world, and the largest in South America. Partnering with them has allowed us to receive a stream of segmented fetal MRIs, as well as expert advice on the main difficulties of performing AF segmentation. We are in the process of implementing AmnioML company-wide, reaching over 40 regular users and over 800 exams per month. While we do not keep track of our userbase for privacy purposes, we expect to reach thousands of exams per month being segmented with AmnioML's aid.

Maintenance

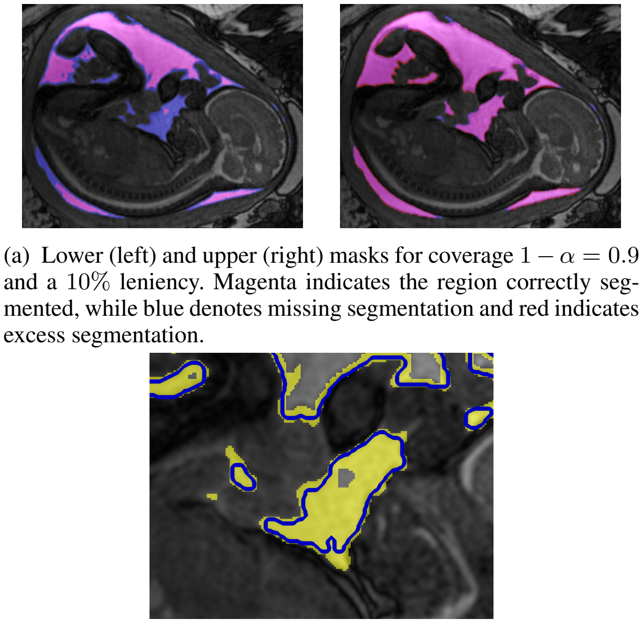
In order to ease upgrades to AmnioML's underlying model, its prediction module was developed as a standalone executable with accompanying libraries. The advantage of this separation is twofold: enhancements of the segmentation procedure can be easily deployed without changes to the frontend, and changes to 3D Slicer's Python environment do not affect the prediction procedure.

Upcoming Features

Besides providing intervals on volume alone, we also developed an algorithm that allows for predictive regions via upper and lower segmentation masks that cover the ground-truth segmentation at any prescribed coverage level, given a leniency parameter. Figure 8a presents an example of a single slice of a typical upper and lower shape-predictive regions generated by our novel algorithm, calibrated at the 90% level with a leniency of 0.1. The leniency parameter allows a percentage of the true segmentation not to be contained in the predictive regions, ensuring much tighter masks at a small further miscoverage penalty. The resulting upper and lower masks are quite similar and most of the errors happen near the ground truth's border.

Through feedback from our users, we have also identified challenging situations for AmnioML's AF segmentation. These include the presence of cysts and gastroschisis. We plan on collecting specific fetal MRI exams to properly address these situations in an upcoming release. While we are hopeful that such pathological data might be enough to guide the algorithm in these situations, we also contemplate developing filters for such specific scenarios. For instance, cysts are usually characterized as bodies of water within the uterus that do not connect with the AF, so it should be possible to identify them before running AmnioML, for example.

By integrating AmnioML to MRI machines, we would like to develop an alarm system to identify abnormal AF volumes before a patient leaves the machine. This way, immediate action could be taken, e.g., subsequent exams performed. Given our customizable interface, doctors should be able to set desired coverage levels for each exam and define exactly what constitutes abnormality.



(a) Lower (left) and upper (right) masks for coverage $1-\alpha=0.9$ and a 10% leniency. Magenta indicates the region correctly segmented, while blue denotes missing segmentation and red indicates excess segmentation.

(b) Uncertainty regions for a 10% leniency and coverage $1-\alpha=0.9$. Displayed in blue is the contour of the predicted segmentation, and in yellow the uncertainty region.

Figure 8: An overview of the regions given by our upcoming shape-predictive conformal algorithm.

Conclusion

This paper introduces AmnioML, a machine learning solution to automatically predict AF volume and quantify the uncertainty in this estimate, and discusses its successful deployment in a medical setting.

A Fetal MRI dataset consisting of 853 exams, along with gestational age, pathologies and corresponding amniotic fluid segmentations is also presented and publicly released. The exams were preprocessed to filter problems such as exam mismatches, and are already divided into train, calibration and test splits, such that no exams from the same mother are present in different folds.

AmnioML is able to to successfully segment unseen exams with average Dice coefficient of 0.91, even in the case of multiple pregnancies or medical conditions, and does so in a few seconds. AmnioML is based on a U-Net architecture for its segmentation task, and benchmarks show that this choice is able to achieve the best Dice coefficient among many recent architectures proposed in the field of medical segmentation. An analysis of AmnioML's segmentation errors showed that they mostly occur along the borders, where there is some degree of subjectivity even for humans. Finally, the speed at which segmentations can be computed are orders of magnitude faster than a medical specialist, which is crucial to reduce wait times and faster exam referral.

Beyond simple volume estimates, it is crucial that doctors can quantify the uncertainty associated with AmnioML's predictions, and assess some of the risks inherent in the pregnancy. For this reason, AmnioML also includes predictive intervals for the volume estimates. Doctors and radiologists can tune the desired coverage level, empowering them in interpreting AmnioML's prediction, as well as guiding its decisions. Different tools were employed and benchmarked, and our results show AmnioML's Algorithm 1 stands out due to its tight intervals and reliable coverage in practice.

AmnioML was deployed as a 3D Slicer plugin, and also as a standalone package. In close collaboration with medical teams, AmnioML was shown to reduce the task of manually segmenting AF by 20×, with the majority of automatic segmentations requiring no further post-processing. We hope it will become a valuable tool to doctors and radiologists.

Finally, AmnioML's success paves the way for similar endeavors. Amniotic fluid segmentation is of great medical interest, but MRI exams also provide a detailed view of the fetus or fetuses. Thus, a related line of work is to explore the field of fetal brain segmentation. In this direction, we expect that the public release of the Fetal MRI Dataset will provide a new benchmark dataset for such important tasks.

Acknowledgments

We would like to thank Dr. Heron Werner, Gerson Ribeiro and their medical team for invaluable help in acquiring, segmenting and processing the fetal MRI exams. We are also indebted to Google's TPU Research Cloud for the TPU credits used in this project.

References

Ayu, P. D. W.; Hartati, S.; Musdholifah, A.; and Nurdiati, D. S. 2021. Amniotic fluid segmentation based on pixel classification using local window information and distance angle pixel. *Applied Soft Computing*, 107: 107196.

Bakhsh, H.; Alenizy, H.; Alenazi, S.; Alnasser, S.; Alanazi, N.; Alsowinea, M.; Alharbi, L.; and Alfaifi, B. 2021. Amniotic fluid disorders and the effects on prenatal outcome: a retrospective cohort study. *BMC pregnancy and childbirth*, 21(1): 1–7.

Barber, R. F.; Candès, E. J.; Ramdas, A.; and Tibshirani, R. J. 2020. The limits of distribution-free conditional predictive inference. arXiv:1903.04684.

Baron, C.; Morgan, M. A.; and Garite, T. J. 1995. The impact of amniotic fluid volume assessed intrapartum on perinatal outcome. *American journal of obstetrics and gynecology*, 173(1): 167–174.

Bastide, A.; Manning, F.; Harman, C.; Lange, I.; and Morrison, I. 1986. Ultrasound evaluation of amniotic fluid: outcome of pregnancies with severe oligohydramnios. *Am J Obstet Gynecol*, 154(4): 895–900.

Bates, S.; Angelopoulos, A.; Lei, L.; Malik, J.; and Jordan, M. 2021. Distribution-Free, Risk-Controlling Prediction Sets. *J. ACM*, 68(6).

Beloosesky, R.; and Ross, M. G. 2018. Amniotic Fluid. In Skinner, M. K., ed., *Encyclopedia of Reproduction (Second Edition)*, 380–386. Oxford: Academic Press, second edition edition. ISBN 978-0-12-815145-7.

Chamberlain, P.; Manning, F.; Morrison, I.; Harman, C.; and Lange, I. 1984a. 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.

Chamberlain, P. F.; Manning, F. A.; Morrison, I.; Harman, C. R.; and Lange, I. R. 1984b. Ultrasound evaluation of amniotic fluid volume. II. The relationship of increased amniotic fluid volume to perinatal outcome. *Am J Obstet Gynecol*, 150(3): 250–254.

Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K. P.; and Yuille, A. L. 2018a. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40: 834–848.

Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; and Adam, H. 2018b. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. *ArXiv*, abs/1802.02611.

Cho, H. C.; Sun, S.; Min Hyun, C.; Kwon, J.-Y.; Kim, B.; Park, Y.; and Seo, J. K. 2021. Automated ultrasound assessment of amniotic fluid index using deep learning. *Medical Image Analysis*, 69: 101951.

Edupuganti, V.; Mardani, M.; Vasanawala, S.; and Pauly, J. 2021. Uncertainty Quantification in Deep MRI Reconstruction. *IEEE Transactions on Medical Imaging*, 40(1): 239–250.

Fedorov, A.; Beichel, R.; Kalpathy-Cramer, J.; Finet, J.; Fillion-Robin, J.-C.; Pujol, S.; Bauer, C.; Jennings, D.; Fennessy, F.; Sonka, M.; Buatti, J.; Aylward, S.; Miller, J. V.; Pieper, S.; and Kikinis, R. 2012. 3D Slicer as an image computing platform for the Quantitative Imaging Network. *Magnetic Resonance Imaging*, 30(9): 1323–1341.

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

Khosravan, N.; Mortazi, A.; Wallace, M.; and Bagci, U. 2019. PAN: Projective Adversarial Network for Medical Image Segmentation. *ArXiv*, abs/1906.04378.

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

Lee, B.; Yamanakkanavar, N.; and Choi, J. Y. 2020. Automatic segmentation of brain MRI using a novel patch-wise U-net deep architecture. *PLOS ONE*, 15(8): 1–20.

Lei, J.; G'Sell, M.; Rinaldo, A.; Tibshirani, R. J.; and Wasserman, L. 2018. Distribution-Free Predictive Inference for Regression. *Journal of the American Statistical Association*, 113(523): 1094–1111.

Liu, W.; Wang, Z.; Liu, X.; Zeng, N.; Liu, Y.; and Alsaadi, F. E. 2017. A survey of deep neural network architectures and their applications. *Neurocomputing*, 234: 11–26.

Looney, P.; Yin, Y.; Collins, S. L.; Nicolaides, K. H.; Plasencia, W.; Molloholli, M.; Natsis, S.; and Stevenson, G. N. 2021. Fully Automated 3-D Ultrasound Segmentation of the Placenta, Amniotic Fluid, and Fetus for Early Pregnancy Assessment. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, 68(6): 2038–2047.

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

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

Moschos, E.; Güllmar, D.; Fiedler, A.; John, U.; Renz, D.; Waginger, M.; Schleußner, E.; Schlembach, D.; Schneider, U.; and Mentzel, H. 2017. 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.

Papadopoulos, H.; Proedrou, K.; Vovk, V.; and Gammerman, A. 2002. Inductive Confidence Machines for Regression. In Elomaa, T.; Mannila, H.; and Toivonen, H., eds., *Machine Learning: ECML 2002*, 345–356. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN 978-3-540-36755-0.

Poudel, R. P. K.; Liwicki, S.; and Cipolla, R. 2019. Fast-SCNN: Fast Semantic Segmentation Network. In *BMVC*, 289. BMVA Press.

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

Queenan, J. T.; Thompson, W.; Whitfield, C.; and Shah, S. I. 1972. Amniotic fluid volumes in normal pregnancies. *American Journal of Obstetrics and Gynecology*, 114(1): 34–38.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 9351 of *LNCS*, 234–241. Springer. (available on arXiv:1505.04597 [cs.CV]).

Shafer, G.; and Vovk, V. 2008. A Tutorial on Conformal Prediction. *Journal of Machine Learning Research*, 9(12): 371–421.

Shang, R.; Zhang, J.; Jiao, L.; Li, Y.; Marturi, N.; and Stolkin, R. 2020. Multi-scale Adaptive Feature Fusion Network for Semantic Segmentation in Remote Sensing Images. *Remote. Sens.*, 12: 872.

Shen, L.; Zheng, J.; Lee, E. H.; Shpanskaya, K.; McKenna, E. S.; Atluri, M. G.; Plasto, D.; Mitchell, C.; Lai, L. M.; Guimaraes, C. V.; Dahmoush, H.; Chueh, J.; Halabi, S. S.; Pauly, J. M.; Xing, L.; Lu, Q.; Oztekin, O.; Kline-Fath, B. M.; and Yeom, K. W. 2022. Attention-guided deep learning for gestational age prediction using fetal brain MRI. *Scientific Reports*, 12(1).

Vovk, V.; Gammerman, A.; and Shafer, G. 2005. *Algorithmic Learning in a Random World*. Springer-Verlag. ISBN 0387001522.

Zhao, H.; Shi, J.; Qi, X.; Wang, X.; and Jia, J. 2017. Pyramid Scene Parsing Network. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6230–6239.

Zhou, Z.; Siddiquee, M. M. R.; Tajbakhsh, N.; and Liang, J. 2020. UNet++: Redesigning Skip Connections to Exploit Multiscale Features in Image Segmentation. *IEEE Transactions on Medical Imaging*, 39: 1856–1867.