



Centro Pi
Centro de Projetos
e Inovação IMPA

Uncertainty Quantification for Amniotic Fluid Segmentation and Volume Prediction

Instituto de Matemática Pura e Aplicada
June 2021





Centro Pi
Centro de Projetos
e Inovação IMPA

Summary

1. Dataset
2. Models
3. Evaluation
4. Uncertainty quantification



1. Dataset





Segmented Fetal MRI Exams - Data Acquisition

- 652 segmented fetal MRI exams
 - AF segmented by specialists
- 80% of the subjects with some degree of pathology
- Gestational age between 19 and 38 weeks
 - High variation of AF volumes
- MRI images produced using a 1.5-T scanner
 - TrueFisp image reconstruction protocol
 - FOV 380 mm
 - Voxel size ~1 x 1 x 1 mm
 - Acquisition time 0.24 s



Data Storage

List of Files

- Unorganized
- Susceptible to data corruption
- Prone to inconsistencies

Database

- Structured
- Verifiable
- Consistency checks:
 - Repetition
 - Dimension
 - Affine transformations (header)



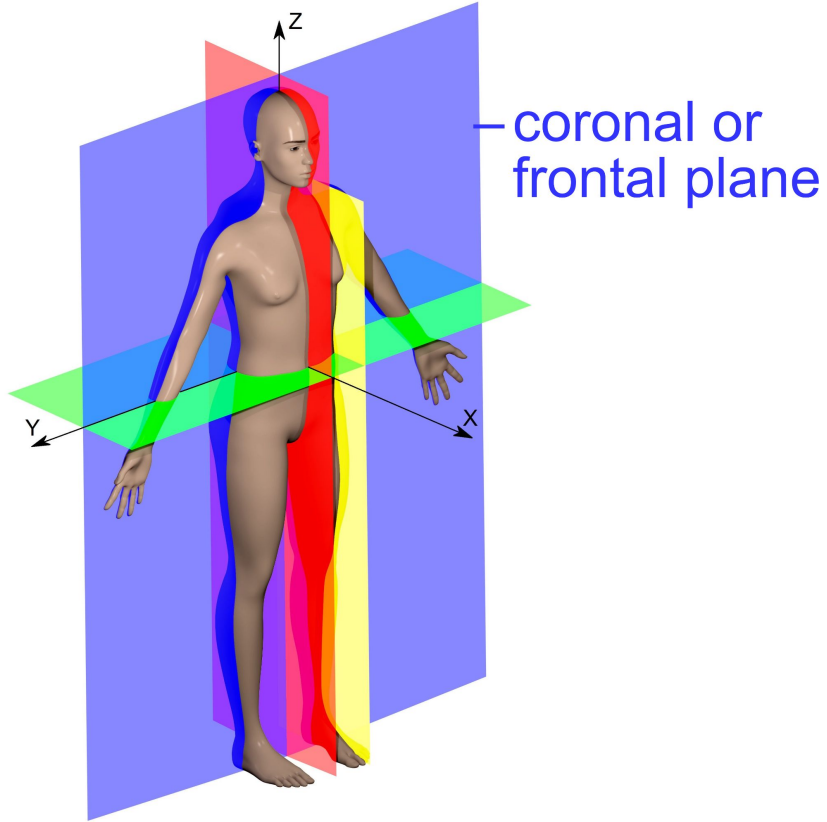
Examples of Inconsistencies

- Repetitions
 - Multiple segmentations for the same exam
 - Repeated pairs of exams and segmentations
- Dimension mismatch
 - Segmentation was cropped and needed re-alignment
- Mismatch of pair exam/segmentation
 - The segmentation didn't correspond to the given exam



From 3D to 2D - Slicing

- 2D models performed better
- To construct an input:
 - Slice exams by planes parallel to the coronal plane
 - Select 3 consecutive slices
 - Normalize by the maximum of the aforementioned selection

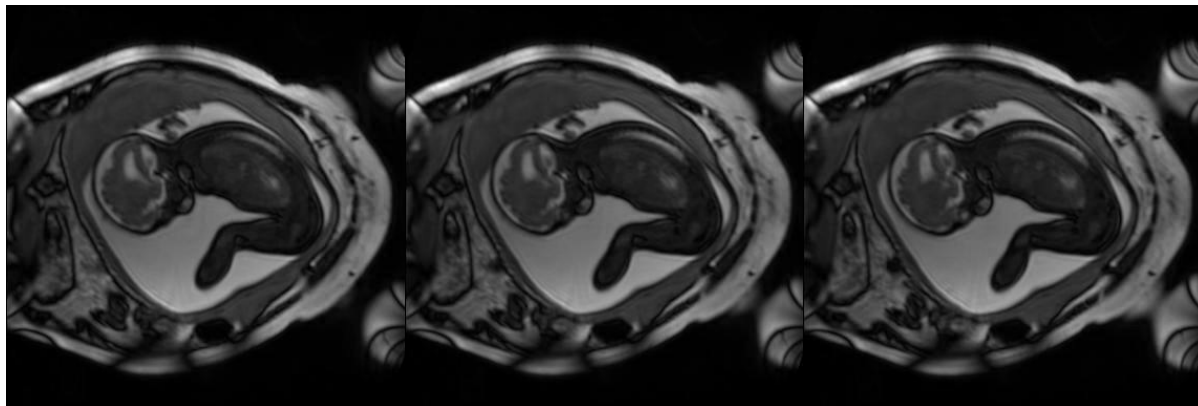


Picture adapted from https://commons.wikimedia.org/wiki/File:Human_anatomy_planes,_labeled.jpg



Example of Input and Target

Input



Target



2. Models



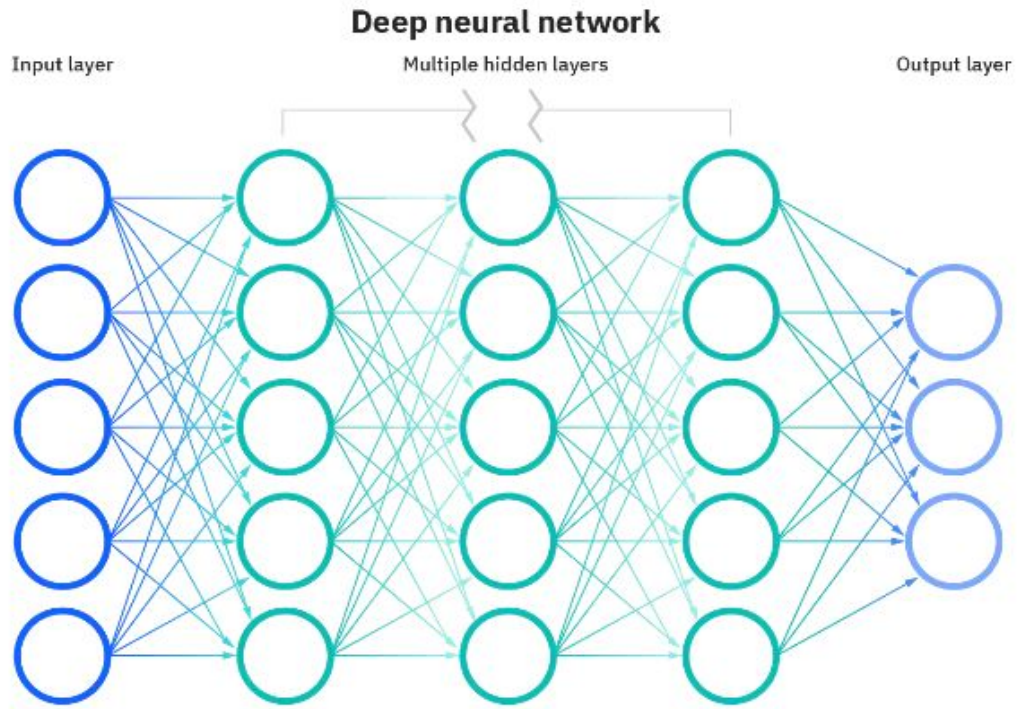


Supervised Learning

- Split data in three sets: Train, Validation, Test
- Train (420 exams):
 - Examples used during learning process
 - Used to fit a model
- Validation (120 exams):
 - Provides unbiased evaluation during training process.
 - Used for hyperparameters tuning
 - Used to calculate confidence interval/regions
- Test (112 exams):
 - Used for final model evaluation



Neural Networks: Introduction



<https://www.ibm.com/cloud/learn/neural-networks>



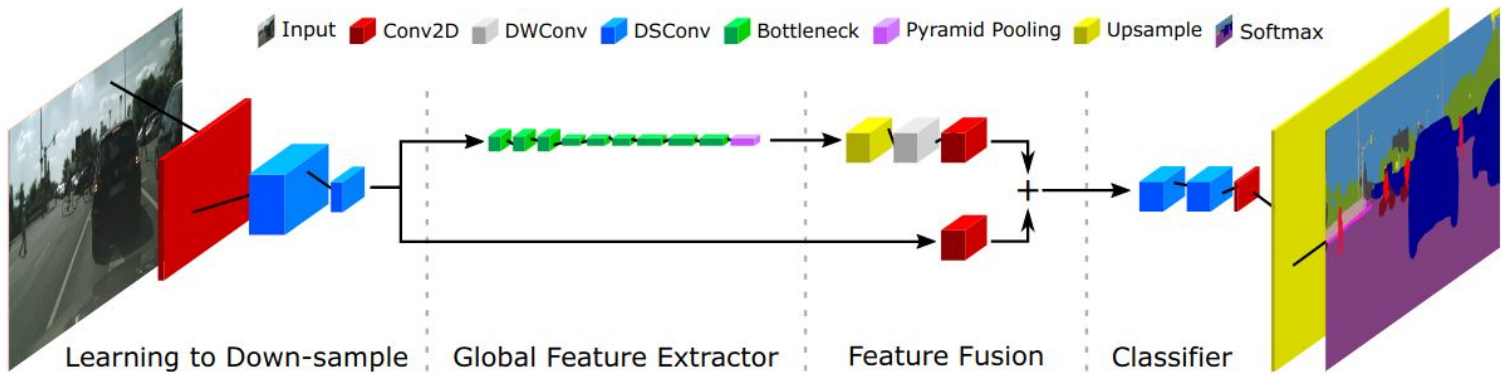
Neural Networks: Introduction

- Advantages:
 - Learn and model non-linear and complex relationships
 - Powerful predictions
 - Does not impose any restrictions on the input variables
- Applications:
 - Voice recognition
 - Artificial intelligence
 - Image classification
 - Image segmentation



Neural Networks: Architectures

- Fast-SCNN

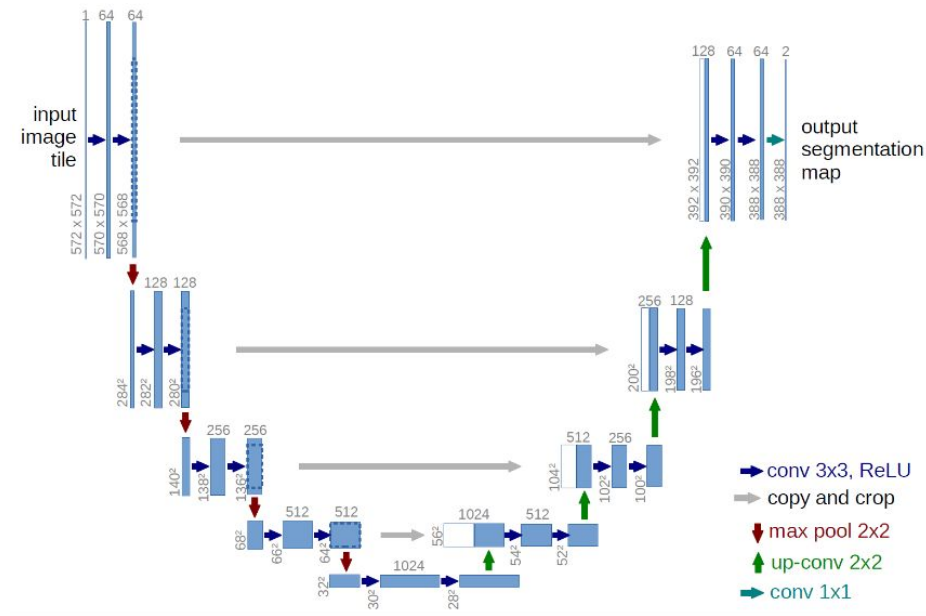


Fast-SCNN: Fast Semantic Segmentation Network - Rudra P K Poudel et al



Neural Networks: Architectures

- U-Net

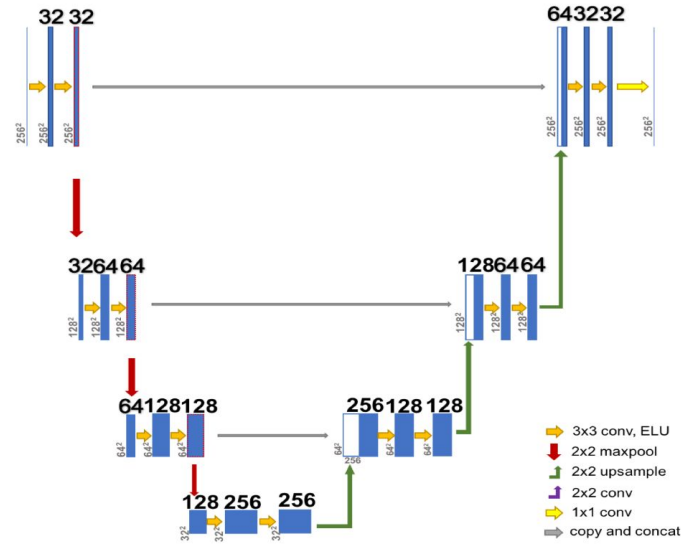


U-Net: Convolutional Networks for Biomedical Image Segmentation - O. Ronneberger et al



Neural Networks: Architectures

- Small U-Net




<https://github.com/shreyaspadhy/UNet-Zoo>



Neural Networks: Convolutional Layers

- Operation between neurons is a kernel convolution
- Examples:

Identity	$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$	
-----------------	---	---

[https://en.wikipedia.org/wiki/Kernel_\(image_processing\)](https://en.wikipedia.org/wiki/Kernel_(image_processing))



Neural Networks: Hyperparameters

- Optimizer: Adam, with learning rate 0.001
- Batch size: 4
- Maximum epoch: 100
- Early Stopping: 7 iterations



Neural Networks: Loss functions

- Statistics of data we want to minimize in each train iteration
- Examples:
 - Binary Cross Entropy: Works as a probability distribution distance
 - Dice: Measure how close segmentations and predictions are
 - Active contour: Incorporates area and size information
 - Learning Active Contour Models for Medical Image Segmentation - Xu Chen et al



Future Ideas

- Pre/Post processing:
 - Filters, Edge detection
 - Background Removal
 - Normalization
 - Removal of Extraneous Connected Components
- New Models:
 - ResNet
 - Utilize Total Variation norm
- Theory:
 - Use details of our data to improve results

3. Evaluation

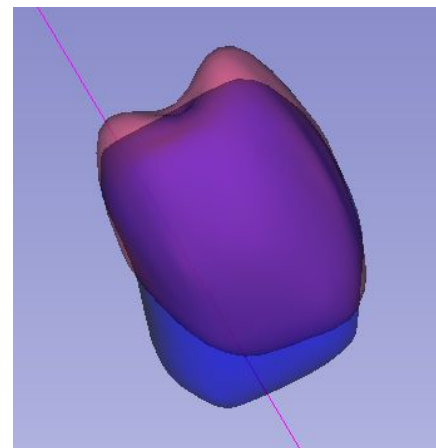




Dice coefficient

Notation

- y : medical segmentation
- \hat{y} : algorithm segmentation
- $D(y, \hat{y})$: Dice coefficient
 - Higher is better
 - Maximum value: 1
 - Minimum value: 0



$$D(y, \hat{y}) = \frac{2|y \cap \hat{y}|}{|y| + |\hat{y}|}$$



Dice coefficient

Extreme cases:

- Ground truth = Prediction = \emptyset
- Ground truth and prediction are exactly the same
- Ground truth and prediction do not overlap

$$D(y, \hat{y}) = \frac{2|y \cap \hat{y}| + \varepsilon}{|y| + |\hat{y}| + \varepsilon}$$



Results

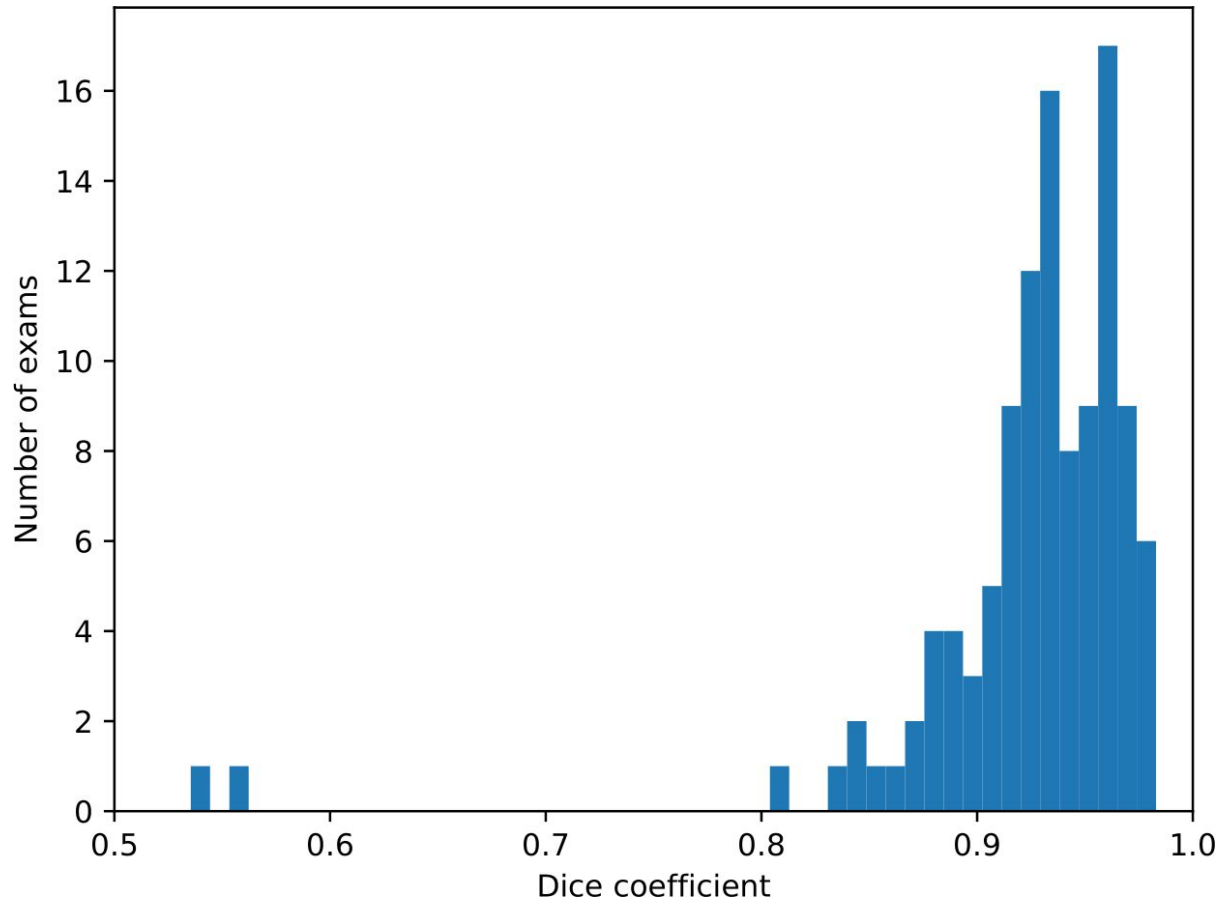
Average Dice coefficient and standard deviation across 112 exams (test set)

Model	Soft Dice	BCE	AC+BCE
U-Net	0.908 ± 0.10	0.924 ± 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



Results

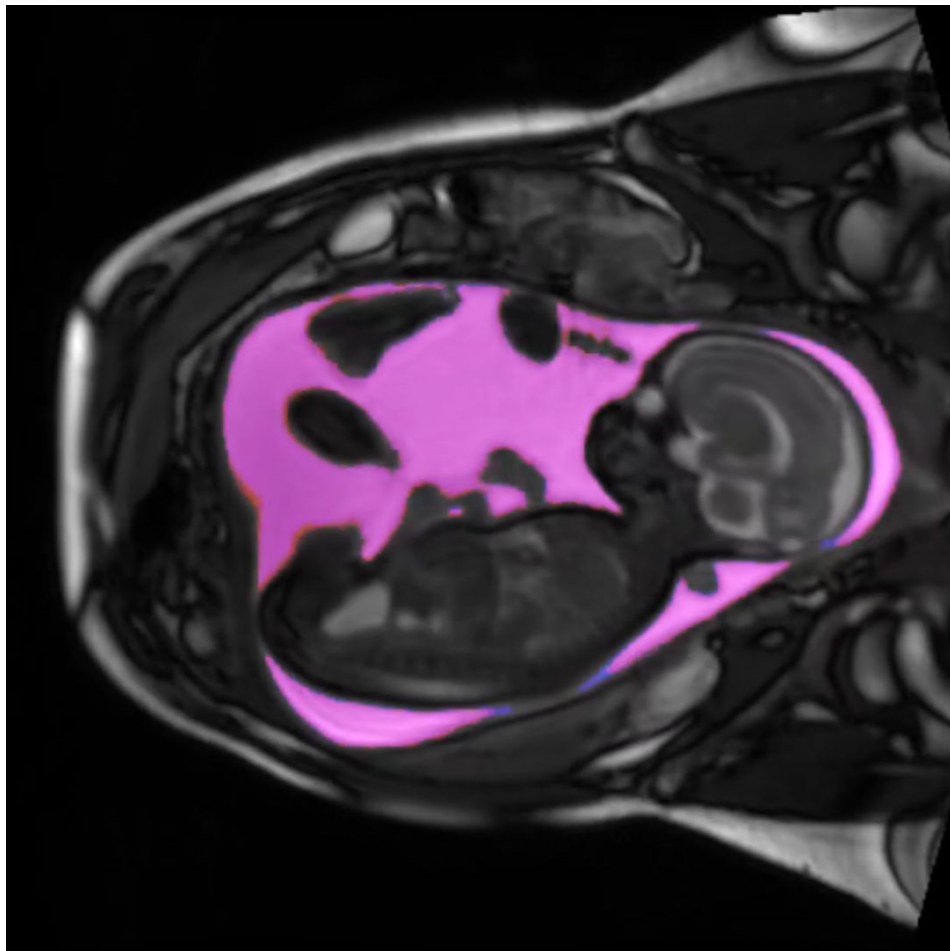
Histogram of best-performing model





Visualization Image

- **Magenta**
 - Region of segmentation correctly located by the algorithm.
- **Red**
 - Excessive region produced by algorithm but not in the segmentation.
- **Blue**
 - Region of segmentation not located by the algorithm.

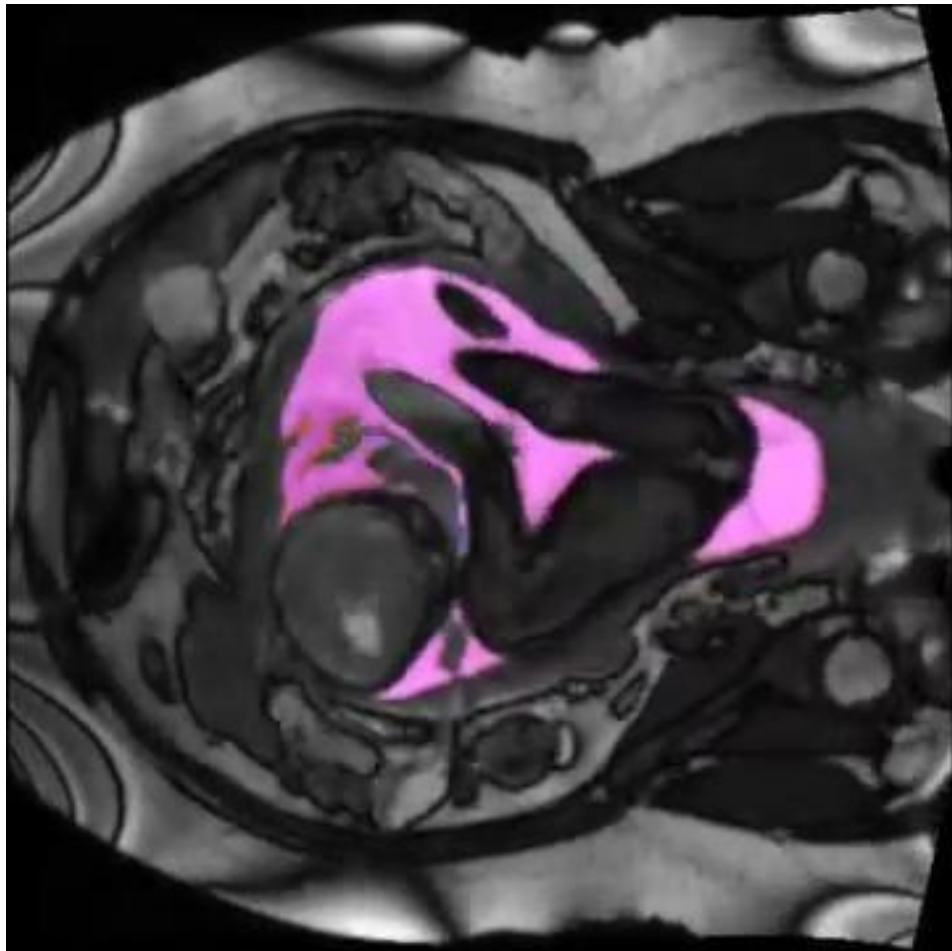




Visualization

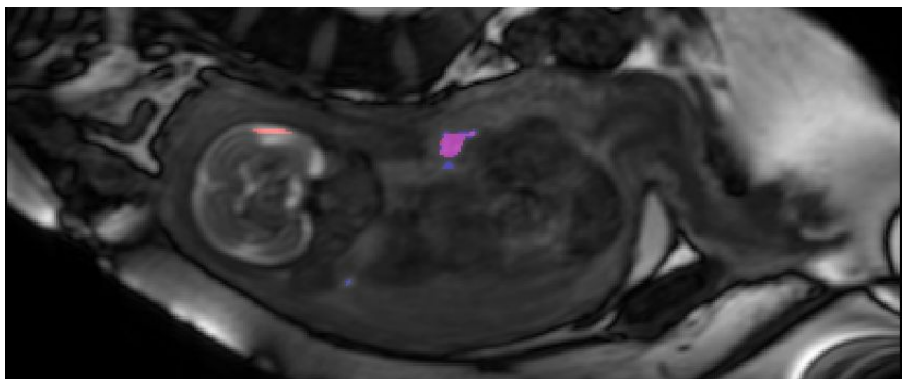
Video

- **Magenta**
 - Region of segmentation correctly located by the algorithm.
- **Red**
 - Excessive region produced by algorithm but not in the segmentation.
- **Blue**
 - Region of segmentation not located by the algorithm.

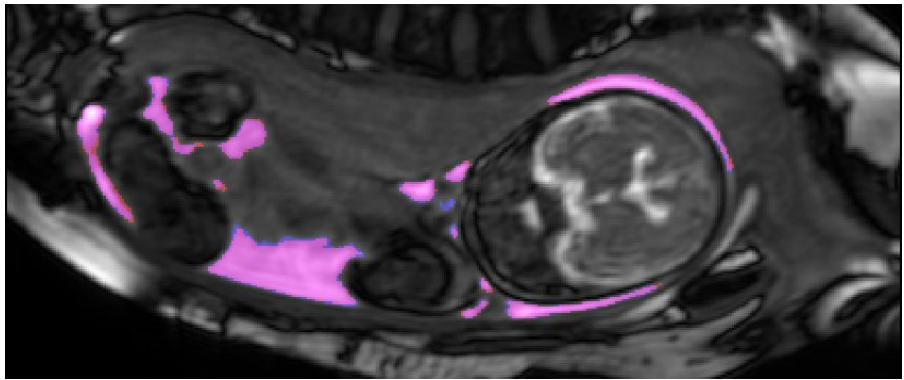




Hard and typical cases



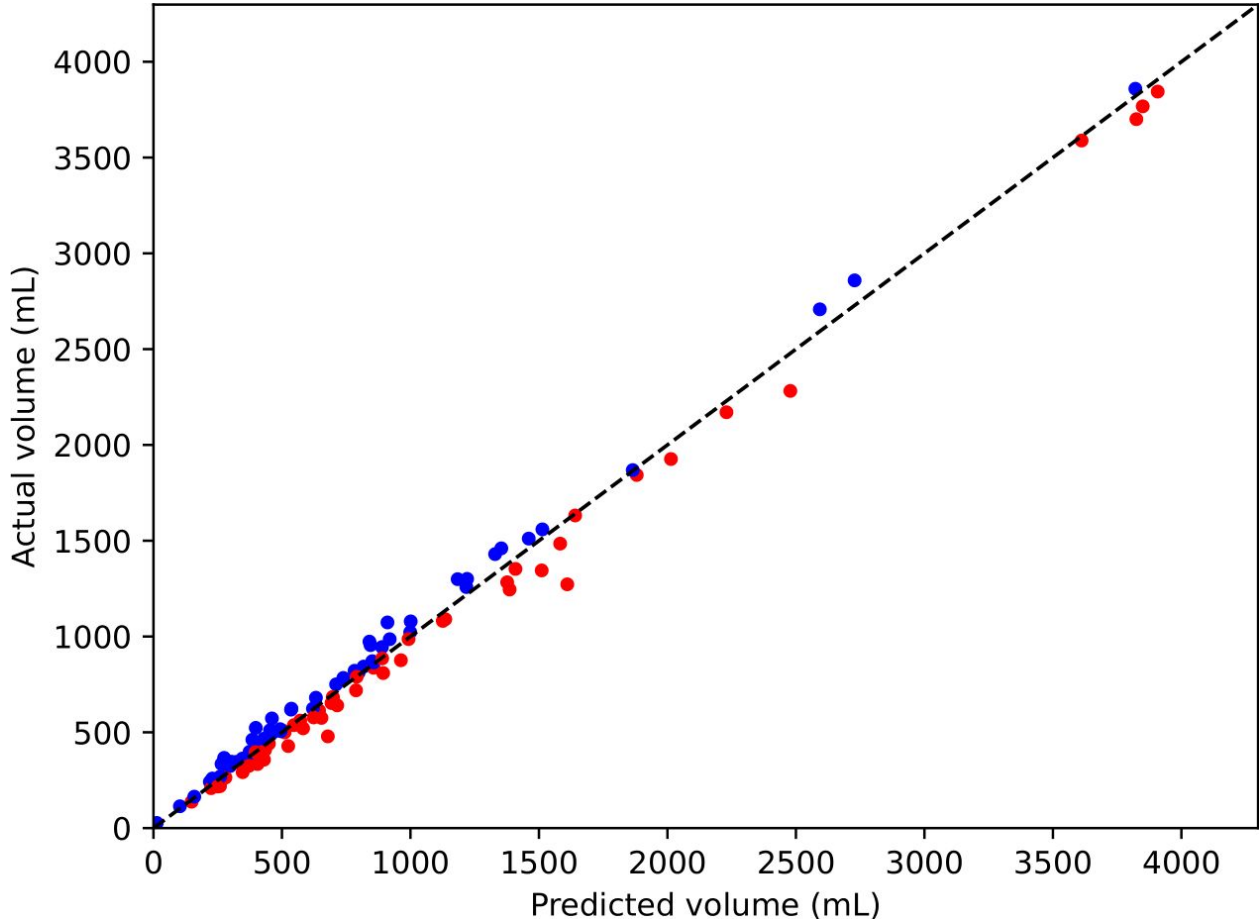
Dice of corresponding exam: ~ 0.54



Dice of corresponding exam: ~ 0.94



Volume





Volume evaluation

Amniotic fluid (mL)	Predicted class		
	Previous	Correct	Following
0 – 200	0	5	0
200 – 400	3	25	0
400 – 600	2	16	2
600 – 800	2	11	3
800 – 1000	2	12	0
1000 – 1250	3	3	0
1250 – 1500	1	4	1
1500 – 2000	0	4	3
2000 – 3000	0	4	1
3000 – 4000	0	5	0

- Volume discretization
 - Usually: low, mid, high
 - This analysis: 10 classes
- No mistake further than 1 class apart
- Correct class: close to 80%



Volume evaluation

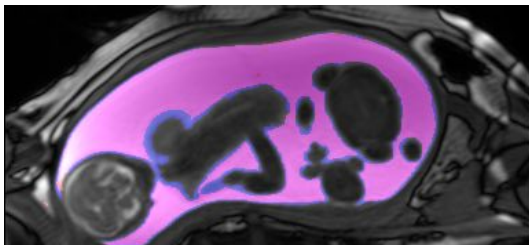
Amniotic fluid (mL)	Predicted class		
	Previous	Correct	Following
0 – 200	0	5	0
200 – 400	3	25	0
400 – 600	2	16	2
600 – 800	2	11	3
800 – 1000	2	12	0
1000 – 1250	3	3	0
1250 – 1500	1	4	1
1500 – 2000	0	4	3
2000 – 3000	0	4	1
3000 – 4000	0	5	0

- Even with large number of classes, correct 80% of the time.
- No mistake further than 1 class apart.

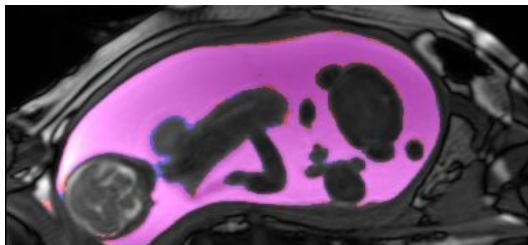


Amniotic fluid boundary

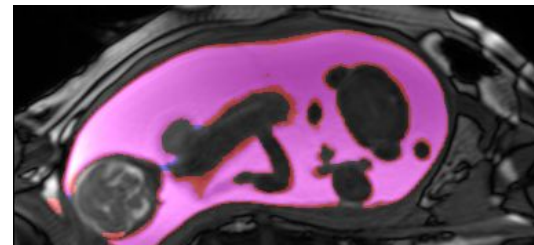
- Hardest region for humans to perform segmentation
 - Inherent uncertainty
- Quantification via dilation and erosion of prediction masks
 - Dilated masks cover 83% of **missing segmentation** on average
 - Eroded masks cover 87% of **excessive segmentation** on average



Eroded prediction mask



Original



Dilated prediction mask

4. Uncertainty quantification





Importance of uncertainty quantification

- Point prediction is important but not sufficient for medical goals
- Goal: provide intervals to quantify the certainty of our estimates
 - For volume: "we are 90% sure the true volume is between 2.5 and 2.7L"
 - For shape: "we are 95% sure the true segmentation is inside of this shape"
- We study multiple ways to create such intervals with theoretical guarantees
- This is important because of irreducible uncertainty in the medical segmentation



Methods for uncertainty quantification

Volume-predictive intervals

- Point prediction is important but not sufficient for medical goals
 - we would like to provide intervals to quantify the certainty of our estimates
- Diagnostic certainty
 - With high probability we want to be sure that the patient is inside the right volume class
- Uncertainty quantification
 - How close to the real AF volume is the predicted Af volume

Shape-predictive regions

- Developing shape-predictive regions
 - This is important because of the intrinsic uncertainty in the edges of the segmentation
- Uncertainty quantification
 - Shows that our errors are made in the edge regions where even human segmentations have a degree of uncertainty.



The road so far

- Data
 - 2D is better
 - We are taking care of duplicate exams
 - Normalization is subtle, but we know how to do it
- Model
 - Neural Networks are very important for AF segmentation
 - NN architecture need to be generalizable and efficient
 - We need to choose the best loss function for our problem
- Evaluation
 - Dice coefficient was the chosen metric for evaluation
 - U-Net with BCE as loss function was the best model
 - Inherent uncertainty makes a Dice coefficient of 1 highly unlikely



Volume-predictive intervals

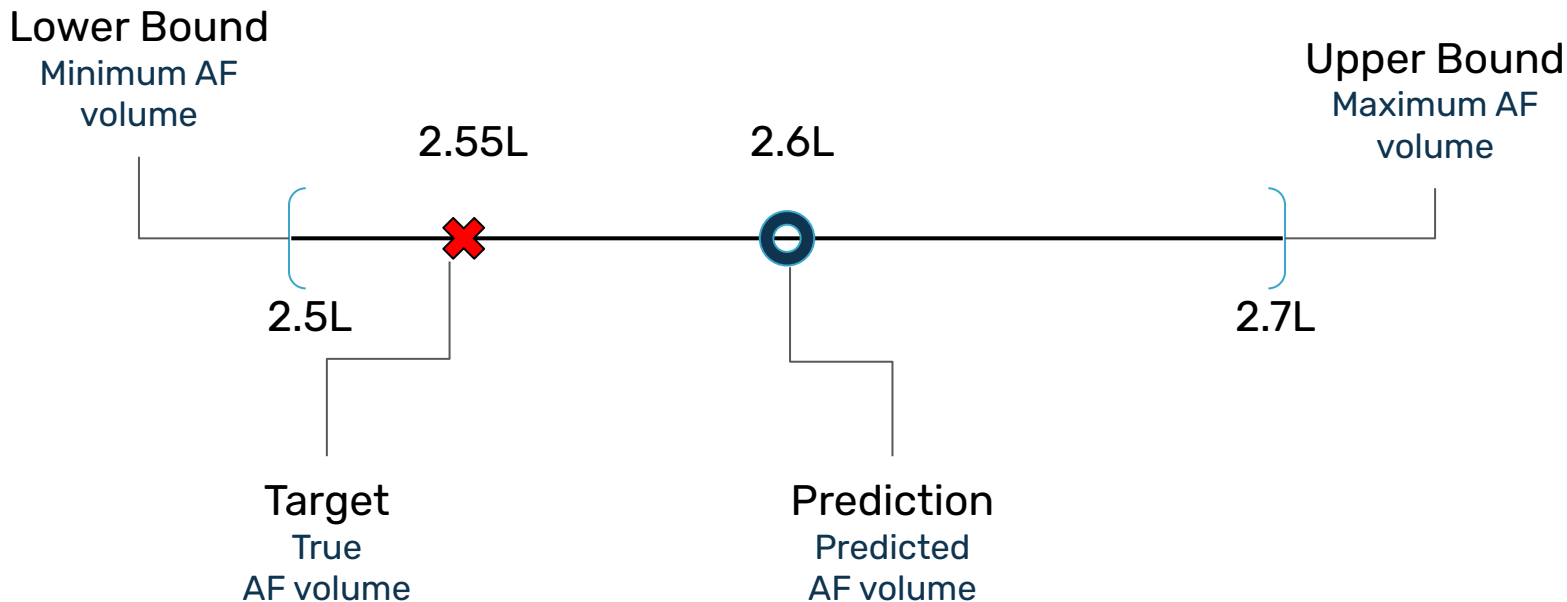
Why is it important?

- Point prediction is important but not sufficient for medical goals
 - We would like to provide intervals to quantify the certainty of our estimates
- Diagnostic certainty
 - With high probability, guarantee that the subject is inside the right volume class
- Uncertainty quantification
 - How close to the real AF volume is the predicted AF volume?



Volume-predictive intervals

How does it look like?





Volume-predictive intervals

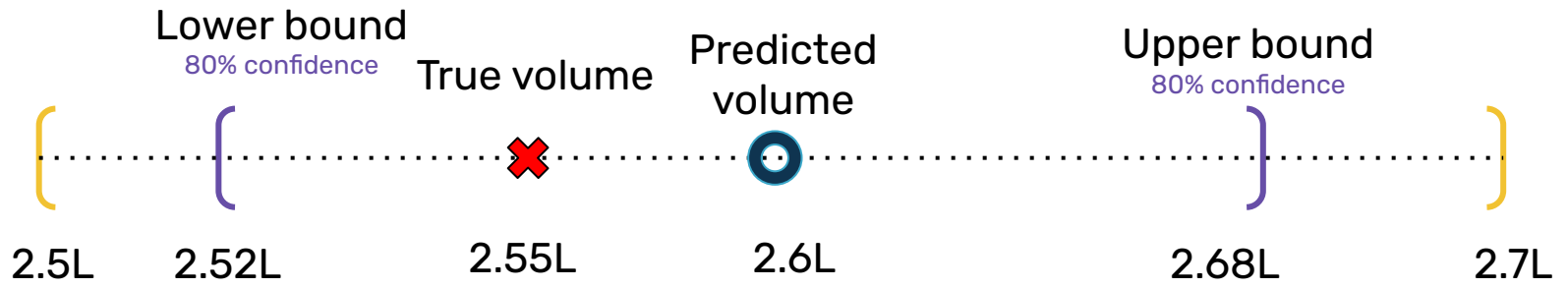
What does it look like?

Lower bound

90% confidence

Upper bound

90% confidence





Shape-predictive regions

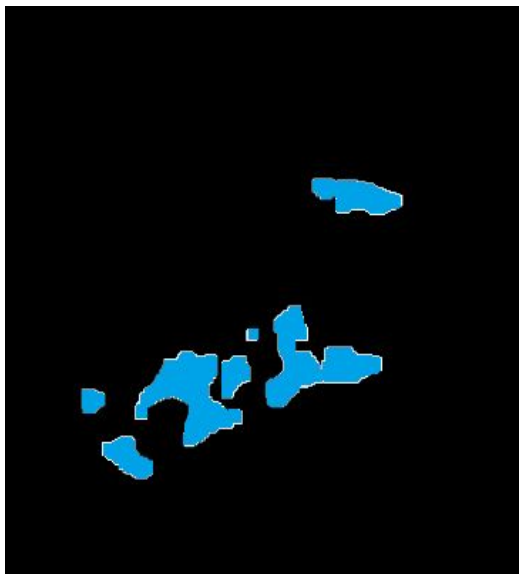
Why is it important?

- Developing Shape-predictive regions
 - This is important because of the intrinsic uncertainty in the edges of the segmentation
- Uncertainty quantification
 - Shows regions with high probability to be a right prediction (true positive)
 - Shows regions with high probability to be a wrong prediction (false negative)
 - Shows that our errors are made in the edge regions where even human segmentations have a degree of uncertainty.



Shape-predictive regions

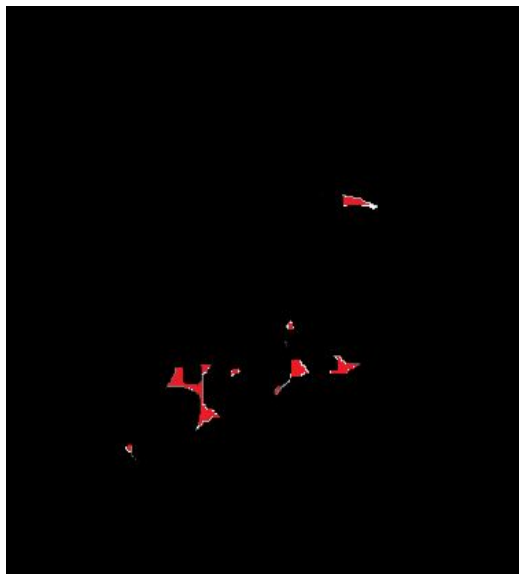
What does it look like?



Upper bound



Segmentation



Lower bound



Methods for uncertainty quantification

Let $(X_j, Y_j)_{m+1}^l$ be a point in the **test sample**.

Volume-predictive intervals

- A *volume-predictive interval at confidence level $1 - \alpha$* is a map $X \rightarrow \mathcal{I}_\alpha(X) \subset \mathbb{R}$ such that:

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

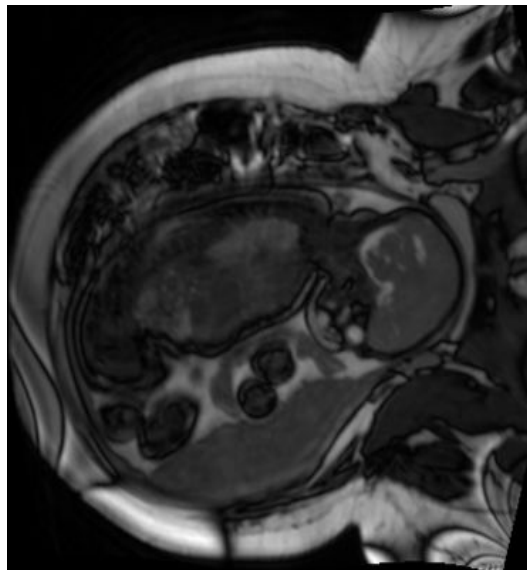
Shape-predictive regions

- A *Shape-predictive region at confidence level $1 - \alpha$* is another map $X \rightarrow \mathcal{C}_\alpha(X) \subset \mathbb{R}^3$ such that:

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



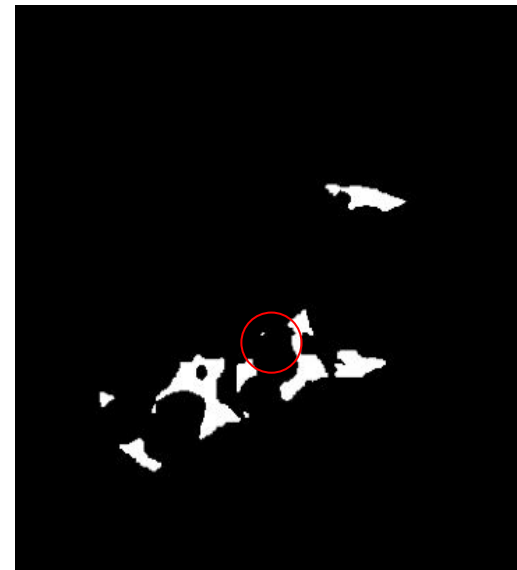
Revisiting the trained model



Exam



Probability mask
voxel value between 0 and 1



Predicted Segmentation
voxel value is 0 or 1



Volume-predictive intervals

Normalized Standard Volume Prediction Standard ($g = 1$) and Normalized ($g = \text{Vol}(\cdot)$)

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 $g : \mathfrak{S}(\mathcal{M}) \rightarrow \mathbb{R}_+$.

radii $\leftarrow []$

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

 append $\left(\frac{|\text{Vol}(\mathcal{M}(X)_{\geq .5}) - \text{Vol}(Y_i)|}{g(\mathcal{M}(x_i)_{\geq .5})} \right)$ to radii

end

radius $\leftarrow (1 - \alpha)$ -quantile of radii

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

$dv \leftarrow g(\mathcal{M}(X_j)_{\geq .5}) \cdot \text{radius}$
 lower volume $\leftarrow \text{Vol}(\mathcal{M}(X)_{\geq .5}) - dv$
 upper volume $\leftarrow \text{Vol}(\mathcal{M}(X)_{\geq .5}) + dv$
 $\mathcal{I}_\alpha(X_j) \leftarrow [\text{lower volume}, \text{upper volume}]$

end

Thresholded Volume Prediction

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 $i \in \{n + 1, \dots, m\}$ **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 $j \in \{m + 1, \dots, l\}$ **do**

 lower volume $\leftarrow \text{Vol}(\mathcal{M}(X_j)_{\geq \text{lower bound}_t})$
 upper volume $\leftarrow \text{Vol}(\mathcal{M}(X_j)_{\geq \text{upper bound}_t})$
 $\mathcal{I}_\alpha(X_j) \leftarrow [\text{lower volume}, \text{upper volume}]$

end



Volume-predictive intervals

Standard

1. Choose the interval confidence ($p\%$)
2. Calculate the distance between the true volume and the predicted volume
3. Choose the interval radius (r) as the number that is bigger than $p\%$ of all calculated distances
4. Define the confidence interval as the predicted volume $\pm r$

Normalized by volume

1. Choose the interval confidence ($p\%$)
2. Calculate the percentages of the errors in comparison with the predicted volumes
3. Choose the error percentual ($\text{error}\%$) as the number that is bigger than $p\%$ of all calculated error percentages
4. Define the interval radius (r) as $\text{error}\%$ of the predicted volume
5. Define the confidence interval as the predicted volume $\pm r$



Volume-predictive intervals

Thresholded

1. Choose the interval confidence (p%)
2. Calculate, for each ground true segmentation, the proportion of AF (v%)
3. For each predicted AF segmentation calculate the predicted threshold, i.e, the threshold that is bigger than v% of all threshold in the probability mask
4. Choose the lower threshold (t_{lower}) as the number that is bigger than (1-p/2)% of all threshold in the probability mask
5. Choose the upper threshold (t_{upper}) as the number that is smaller than (1-p/2)% of all threshold in the probability mask
6. Define the confidence interval as the set of volume between the volume of the NN output thresholded by the lower threshold and the volume of the NN output thresholded by the upper threshold





Volume-predictive intervals

Standard

- Constant interval lengths
- Does not use model output

Normalized by volume

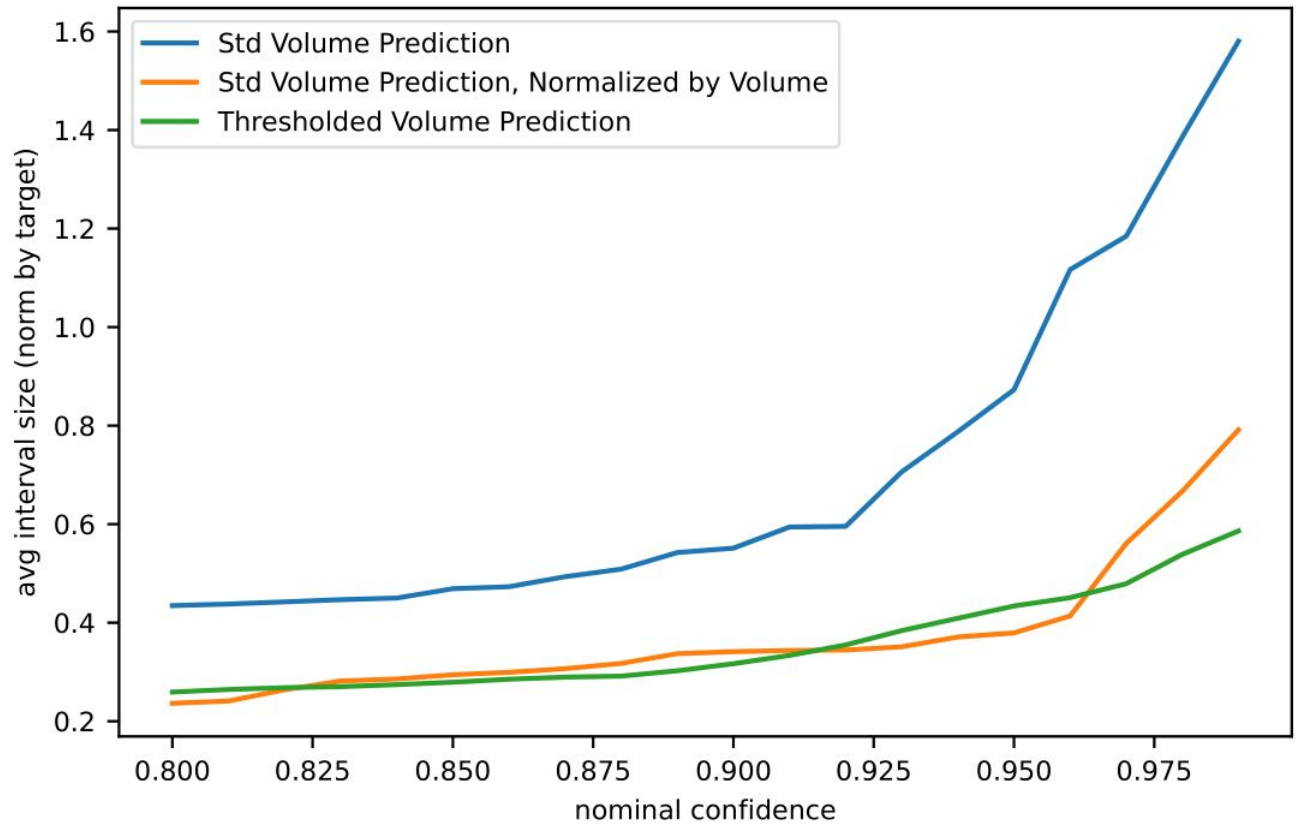
- Interval lengths adaptative to the wide range of volumes in the data
- Does not use model output

Thresholded

- Interval lengths adaptative to the wide range of volumes in the data
- Uses model output

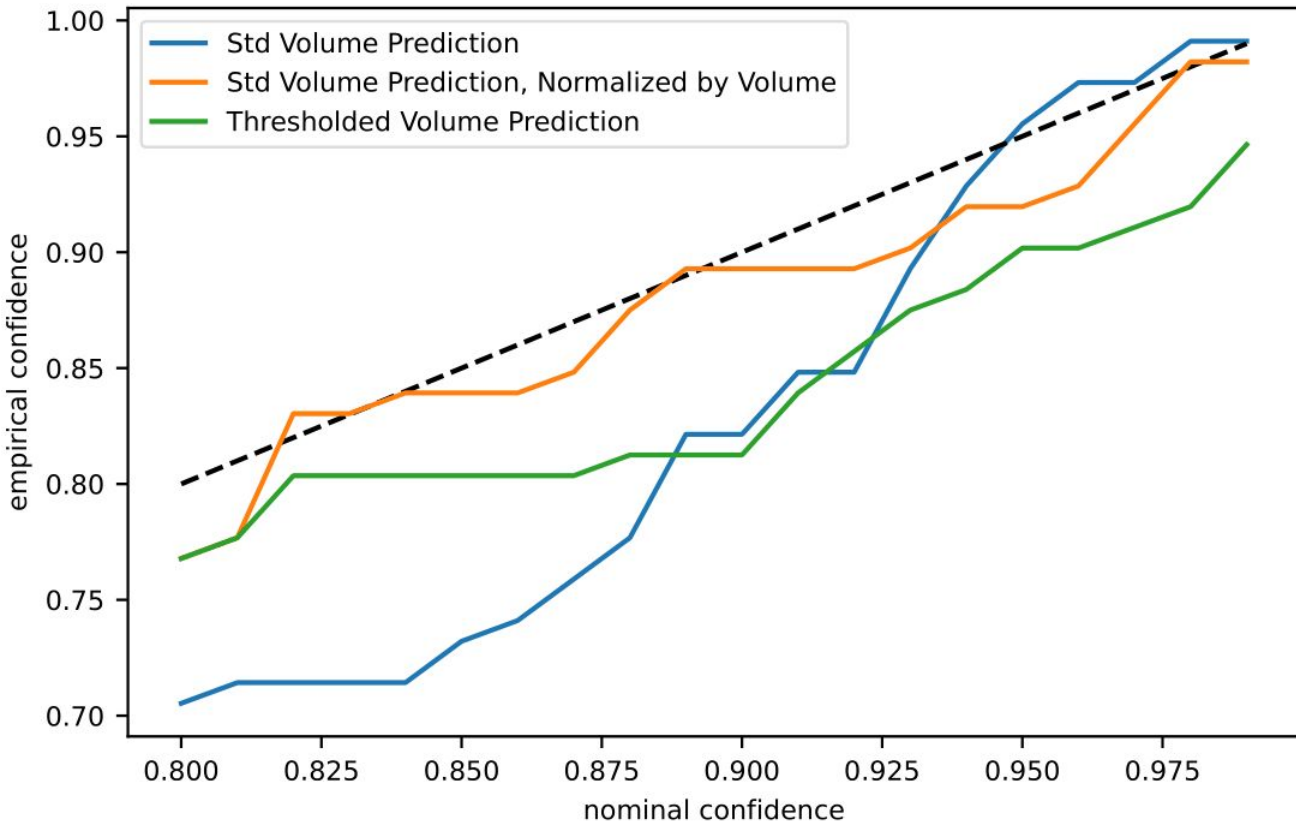


Performance of volume-prediction intervals: Interval length





Performance of volume-prediction intervals: Empirical vs nominal confidence





Performance of volume-prediction intervals: Why is better? Pearson's correlation coefficient.

Method	Pearson's correlation (interval length vs prediction error)
Standard Volume Prediction (90%)	-
Normalized Volume Prediction (90%)	0.2175
Threshold Volume Prediction (90%)	0.5946



Performance of volume-prediction intervals: Why is better? Pearson's correlation coefficient.

Methods	Interval size	Confidence generalization	Pearson's correlation
Standard	Big for all confidences	Big for high confidences but not for small ones	-
Normalized by volume	Big for high confidences	Big for all confidences	Regular
Threshold	The smallest (especially for high confidences)	Regular for all confidences	The highest



Shape-predictive regions

Segmentation prediction

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 $i \in \{n + 1, \dots, m\}$ **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 $j \in \{m + 1, \dots, l\}$ **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





Shape-predictive regions

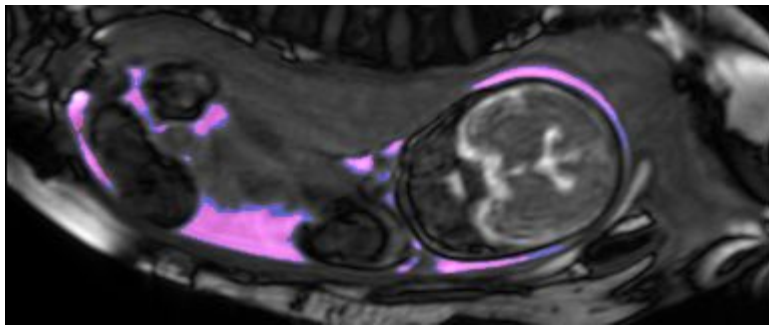
Segmentation prediction

1. Choose the interval confidence ($p\%$), and the leniency ($l\%$)
2. For each predicted AF segmentation calculate the upper/lower predicted threshold, i.e, the threshold that makes the thresholded NN output cover/be inside the ground truth segmentation (up to a leniency proportional to $l\%$).
3. Choose the upper threshold as the number that is smaller than $(1-p/2)\%$ of all threshold in the probability mask
4. Choose the lower threshold as the number that is bigger than $(1-p/2)\%$ of all threshold in the probability mask
5. Define the confidence interval as the region between the NN output thresholded by the lower threshold and the NN output thresholded by the upper threshold

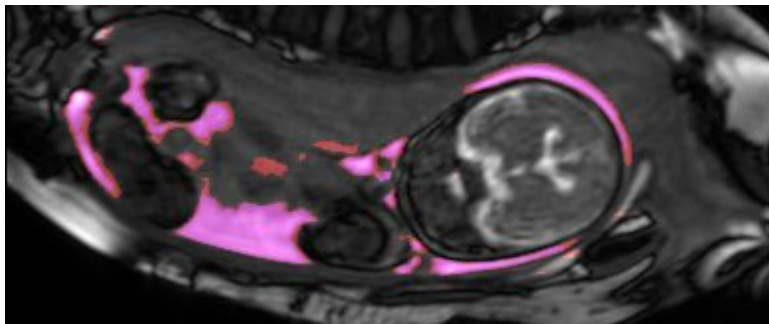




Results of shape-predictive regions: Segmentation prediction



- Region in the **segmentation** and **not** in the **lower bound** in **blue**.
- Region in the **segmentation** and in the **lower bound** in **magenta**.

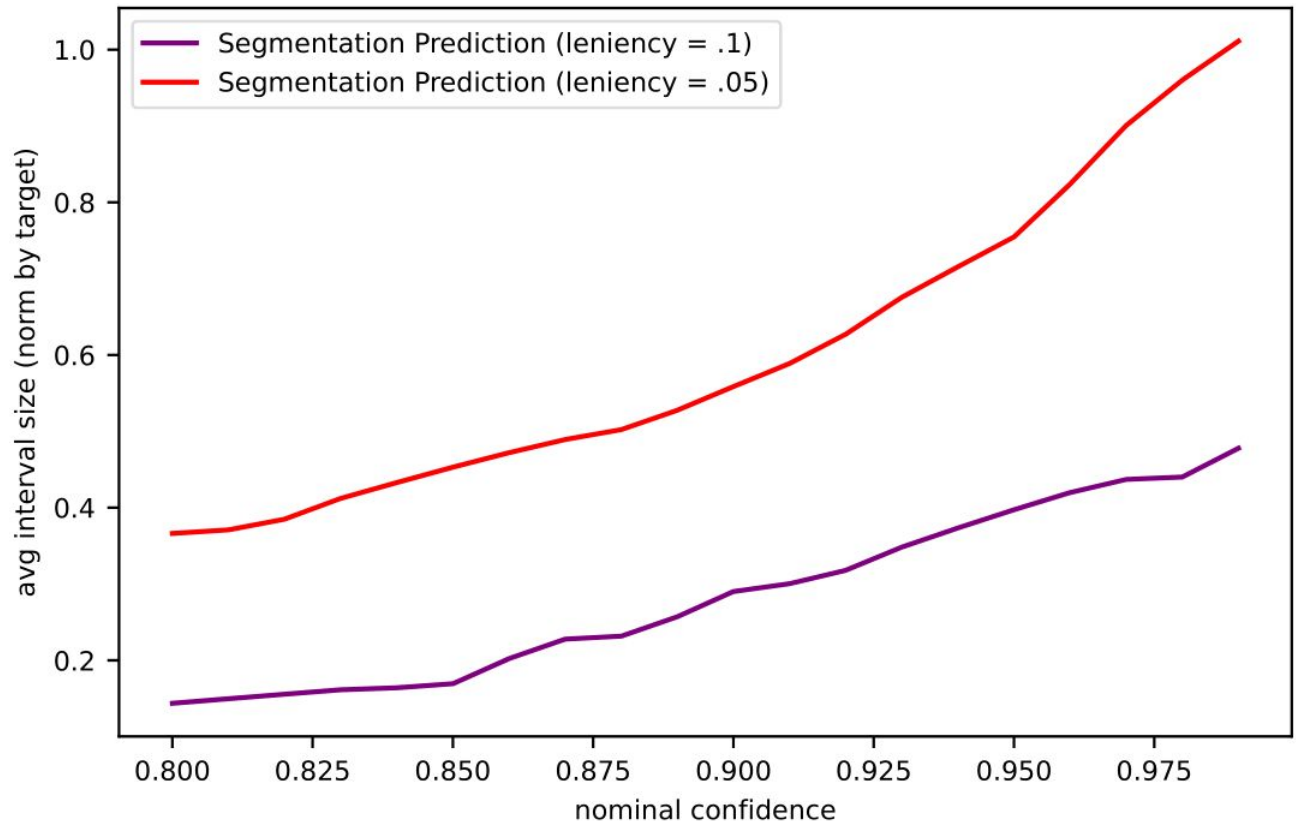


- Region in the **upper bound** and **not** in the **segmentation** in **red**.
- Region in the **segmentation** and in the **upper bound** in **magenta**.

Confidence = 90%
Leniency = 5%



Performance of shape-predictive regions: How tight is it?



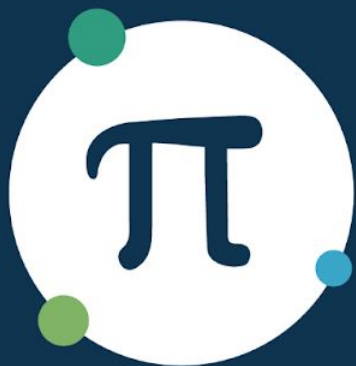


Performance of shape-predictive regions:
Why not to use it as a volume interval?



Conclusion

- U-Net with BCE as loss function is the best model for AF segmentation
 - High dice coefficient (>90%) for the vast majority of subjects
 - Each segmentation takes 5 seconds on a GPU
- Threshold Volume Prediction is the best method to create confidence intervals for AF volume
 - Threshold Volume Prediction build tight confidence intervals, with the length highly correlated with the prediction error, and great confidence generalization.
- The Segmentation Prediction method gives tight confidence regions for the AF shape.
- With these tools, it is possible to automate the segmentation and volume estimation of AF with theoretical guarantees and empirical validation



Centro Pi
Centro de Projetos
e Inovação IMPA

centropi@impa.br