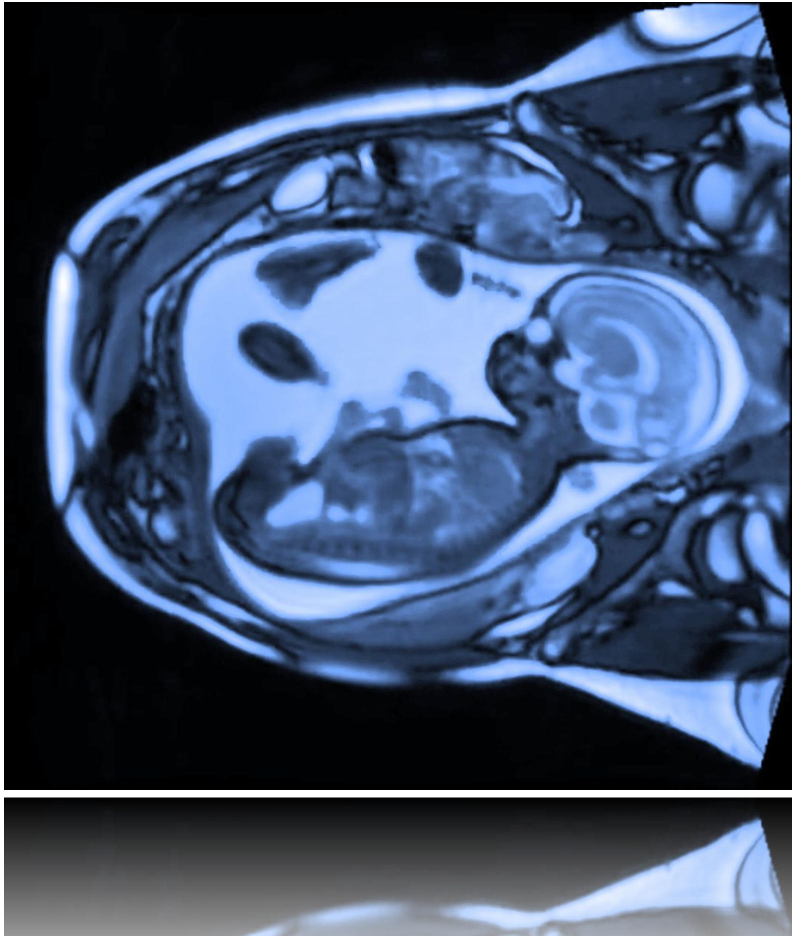
AmnioML: amniotic fluid segmentation and volume prediction with uncertainty quantification



Thiago Rodrigo Ramos - UFSCar
July 2024

Introduction

- This project was conducted during my Ph.D. in Centro PI - IMPA
- It was a collaboration with Dasa, IMPA Professors and other students
- **GOAL:** Segmentation and Uncertainty Quantification of amniotic fluid

(SHOW VIDEO)

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

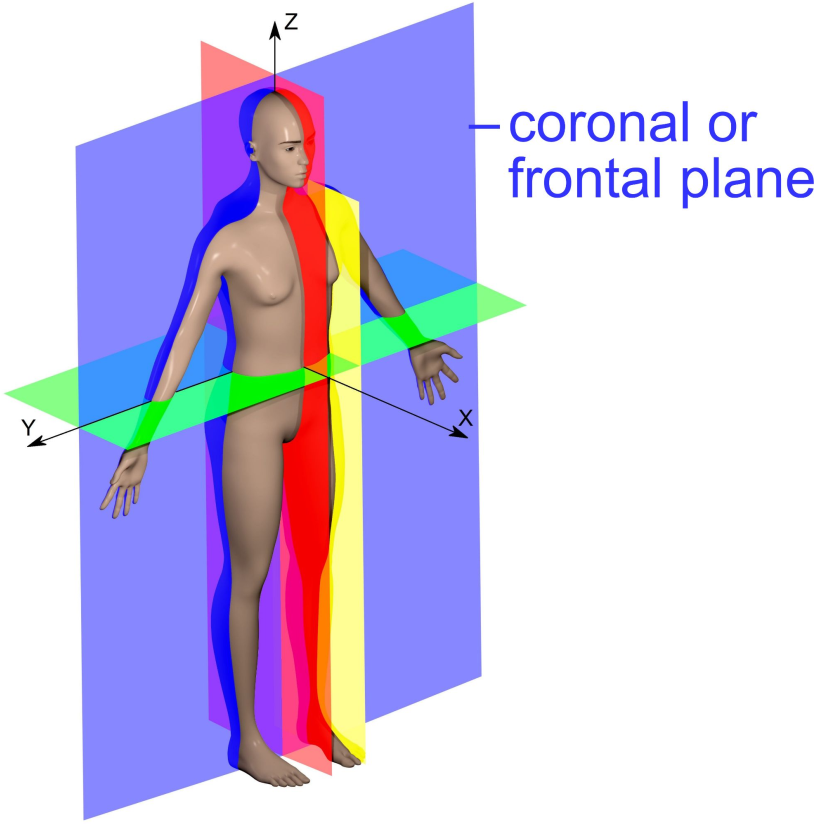
- Structured
- Verifiable
- Consistency checks:
  - Repetition
  - Dimension
  - Affine transformations (header)

Data 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

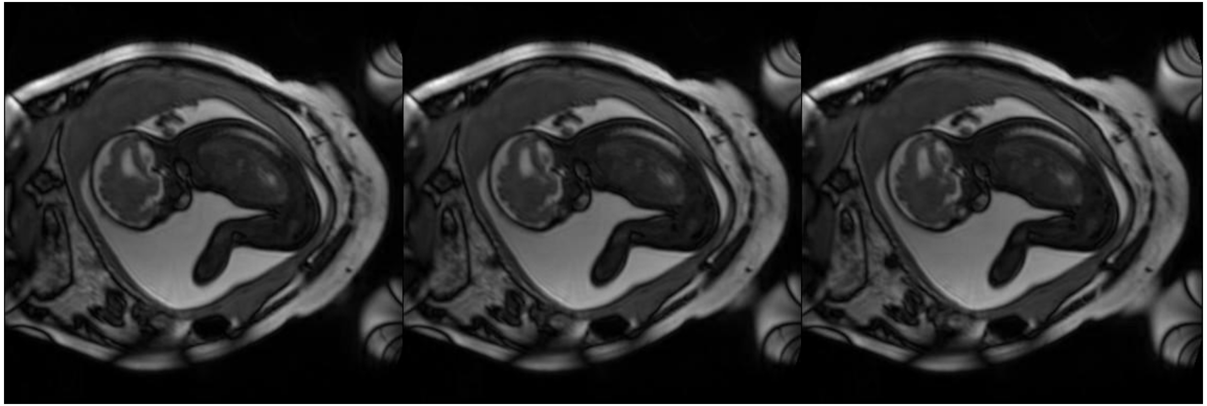
- 2D models performed better
- To construct an input:
  - Slice exams by planes parallel to the coronal plane
  - Select 3 consecutive slices
  - Normalize by the maximum of the aforementioned selection



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

Example of Input and Target

Input



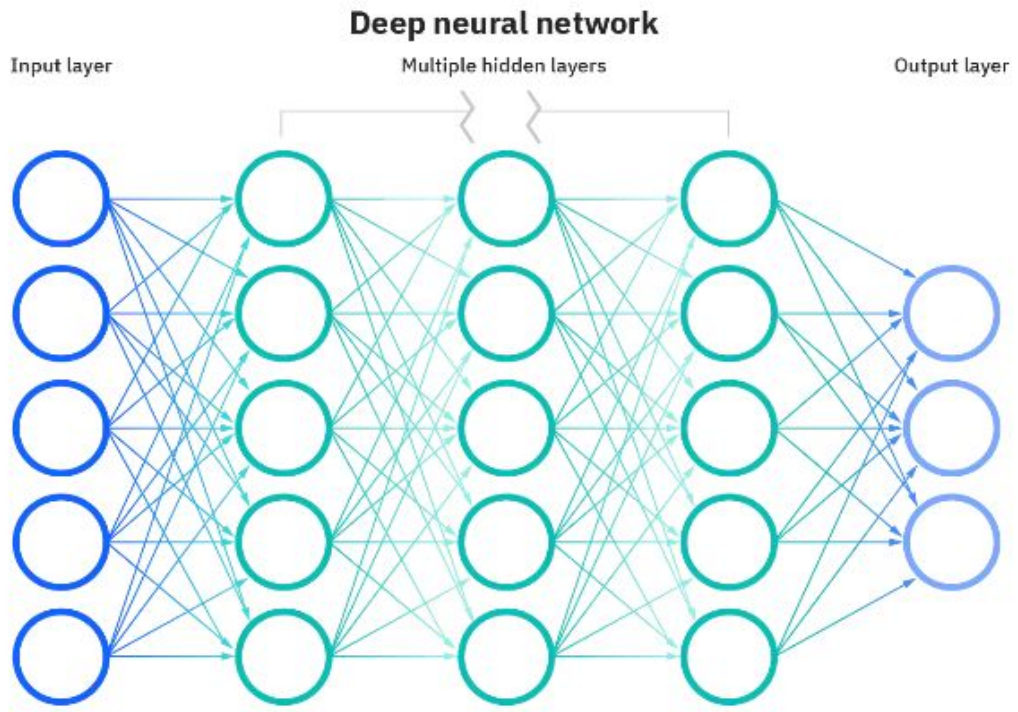
Target

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 <u>final</u> model evaluation

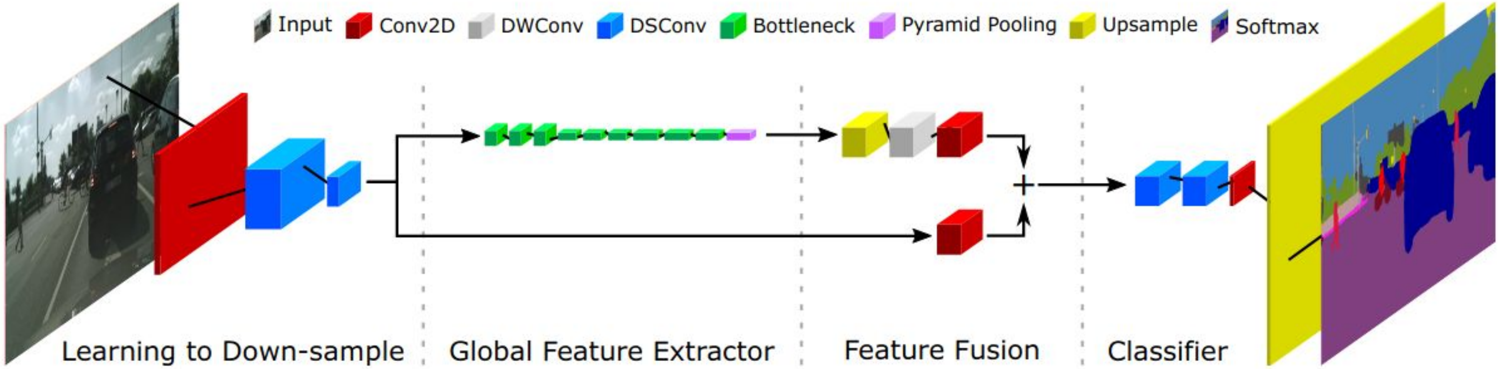
Neural Networks



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

Neural Networks: Architectures

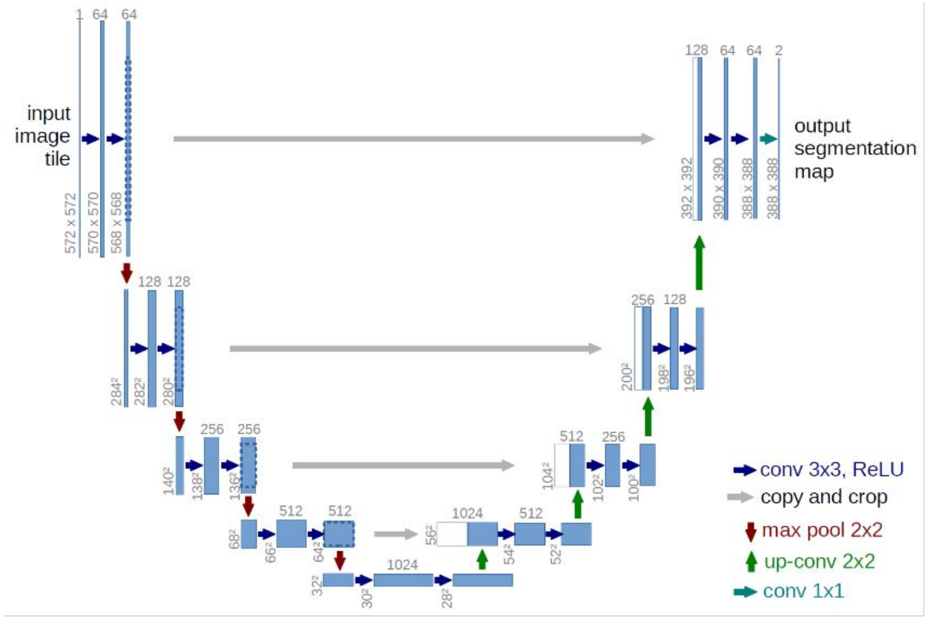
- Fast-SCNN



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

Neural Networks: Architectures

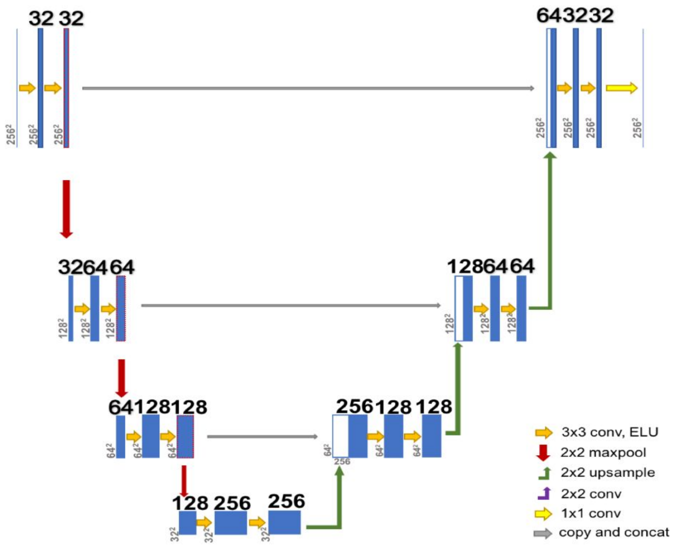
- U-Net



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

Neural Networks: Architectures

- Small U-Net



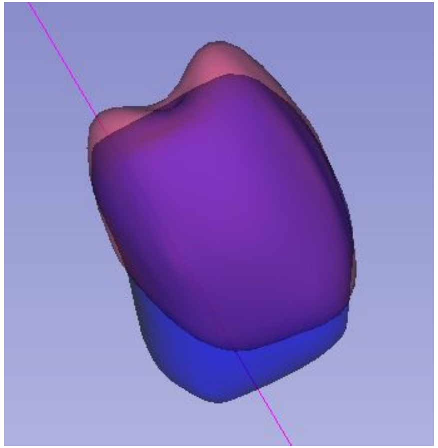
https://github.com/shreyaspadhy/UNet-Zoo

3. Evaluation

Dice coefficient

Notation

- y: medical segmentation
- ŷ: algorithm segmentation
- D(y, ŷ): Dice coefficient
    - Higher is better
    - Maximum value: 1
    - Minimum value: 0



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

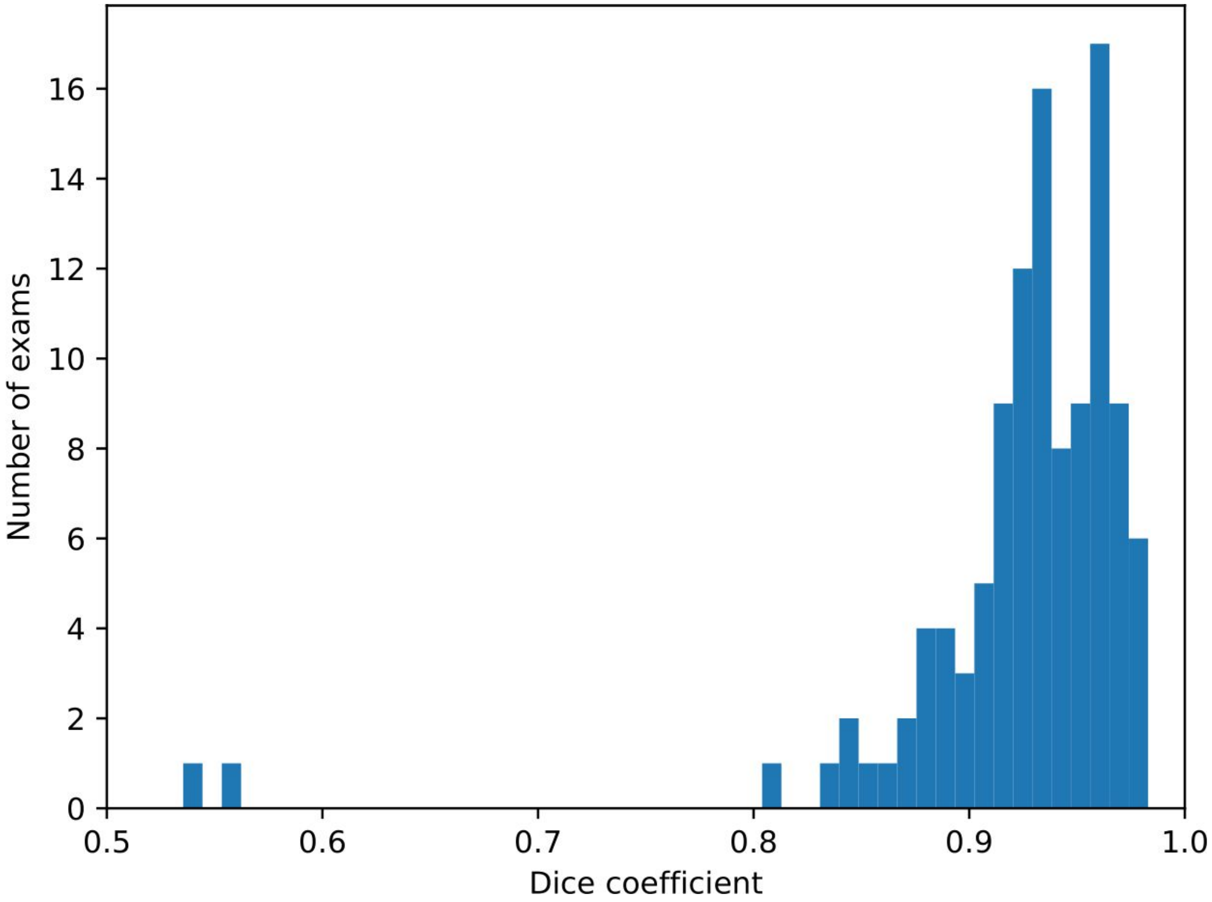
Results

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

Model	Soft Dice	BCE	AC+BCE
U-Net	0.908 ± 0.10	$\mathbf{0.924} \pm \mathbf{0.06}$	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

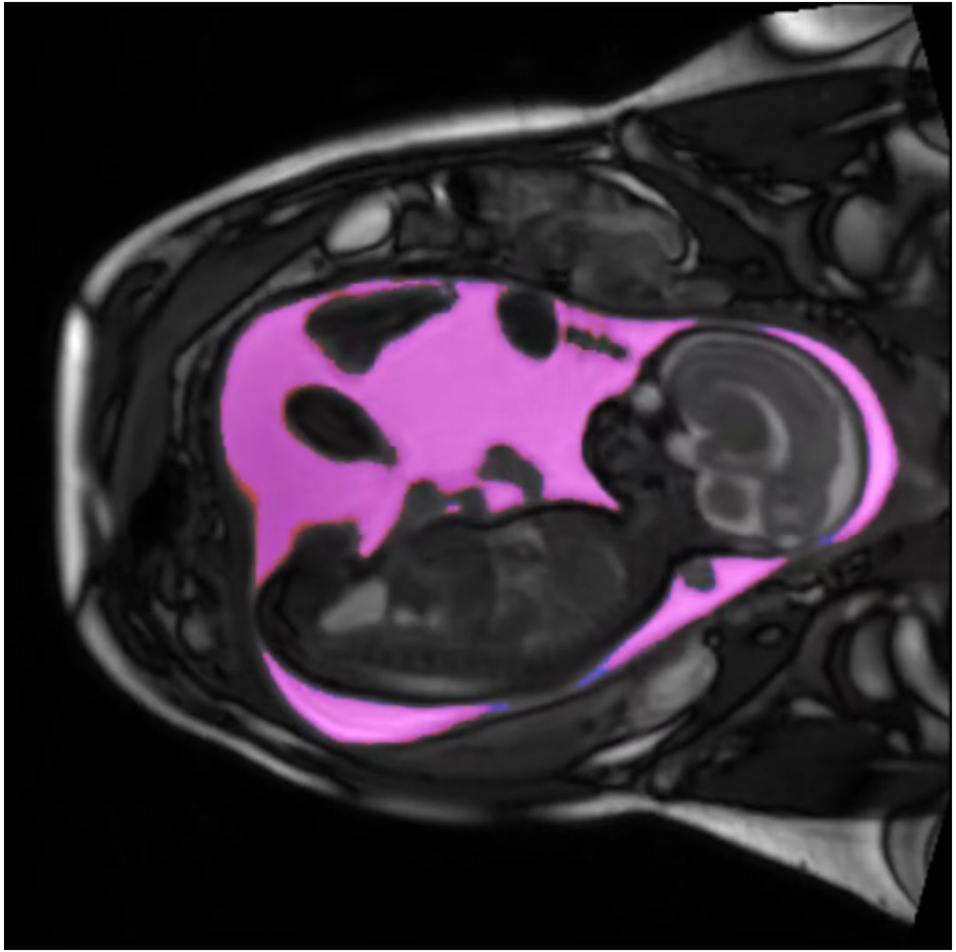
Histogram of best-performing model



Visualization

Image

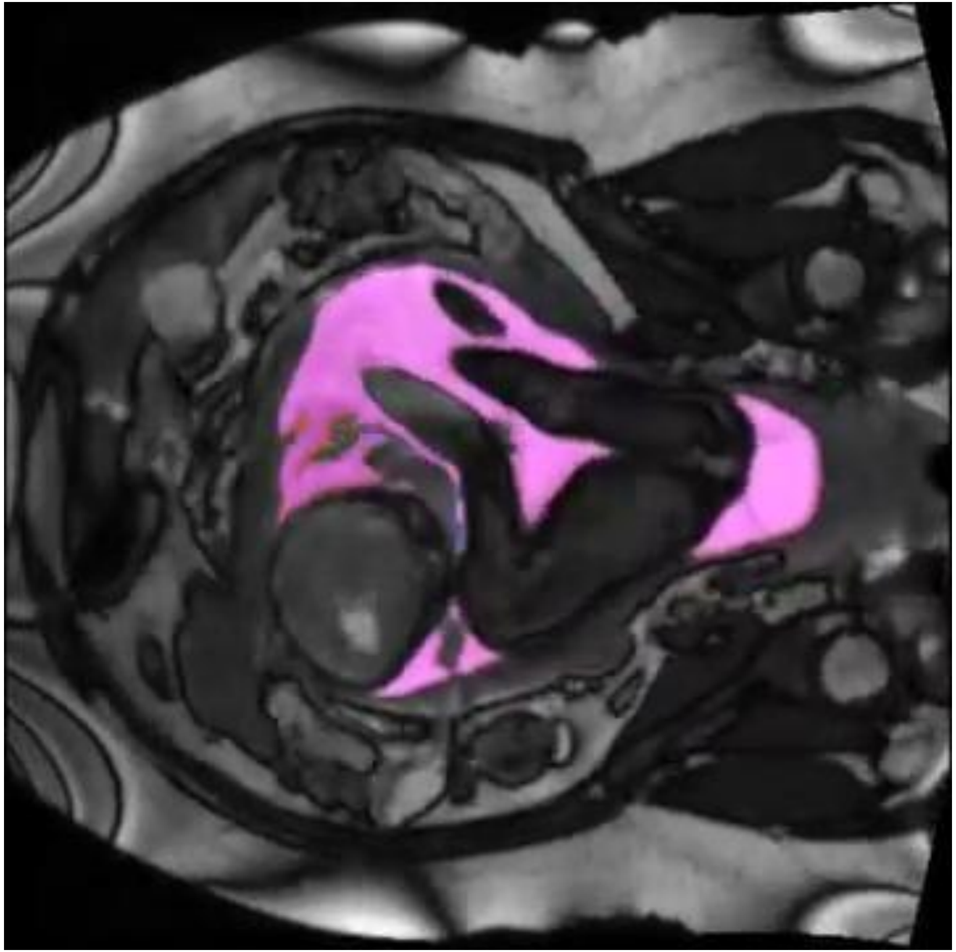
- Magenta
  - Region of segmentation correctly located by the algorithm.
- Red
  - Excessive region produced by algorithm but not in the segmentation.
- Blue
  - Region of segmentation not located by the algorithm.



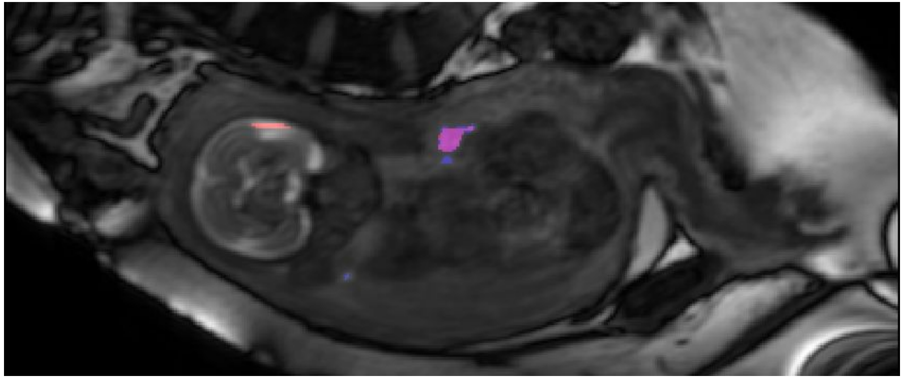
Visualization

Video

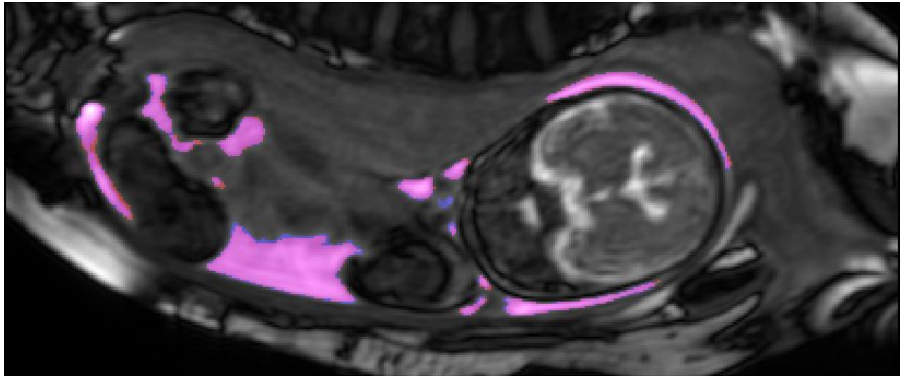
- Magenta
  - Region of segmentation correctly located by the algorithm.
- Red
  - Excessive region produced by algorithm but not in the segmentation.
- Blue
  - Region of segmentation not located by the algorithm.



Hard and typical cases

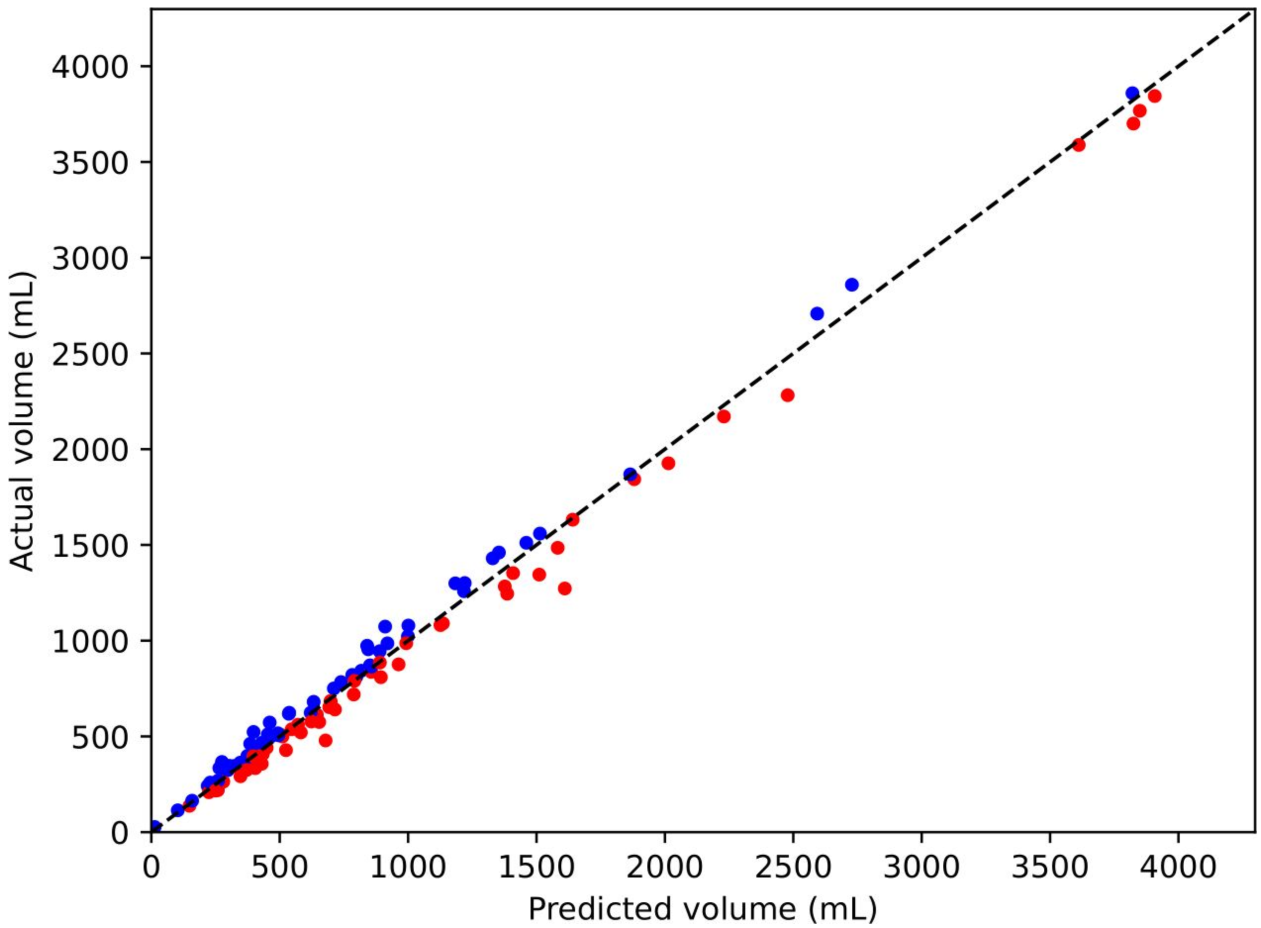


Dice of corresponding exam: ~ 0.54



Dice of corresponding exam: ~ 0.94

Volume



Actual volume (mL)
Predicted volume (mL)
0
500
1000
1500
2000
2500
3000
3500
4000

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.

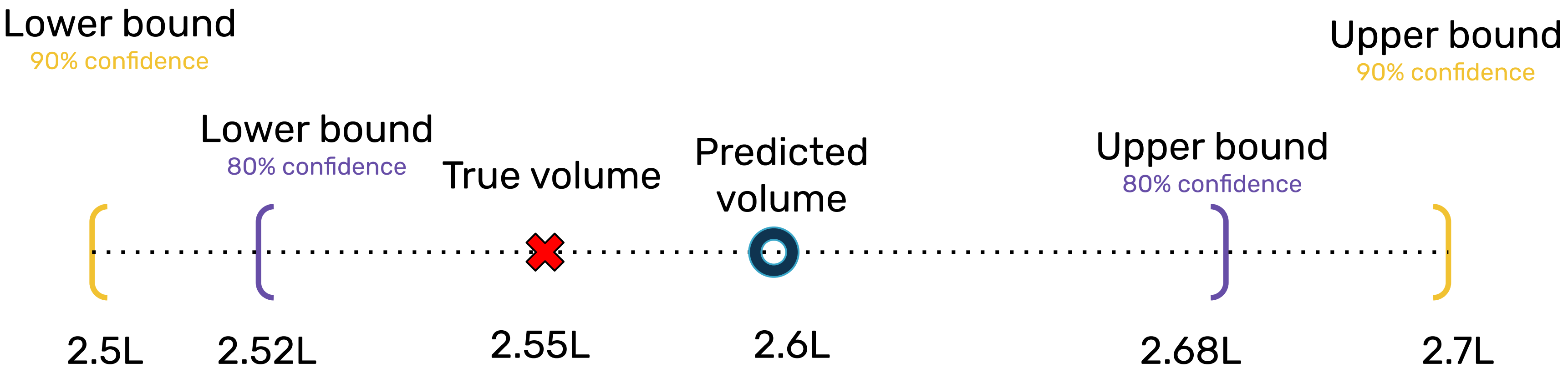
4. Uncertainty quantification

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

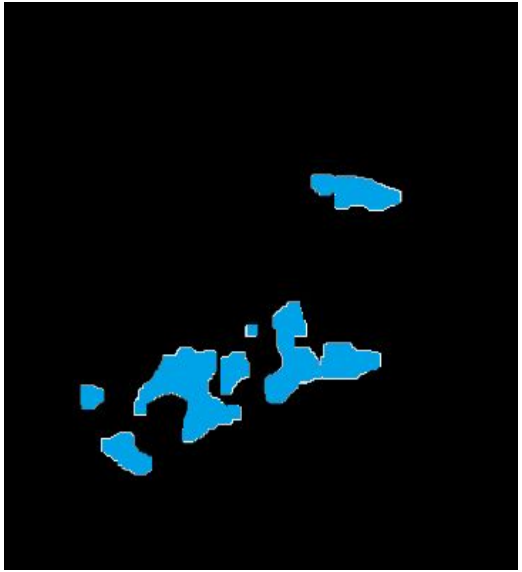
Volume-predictive intervals

What does it looks like?



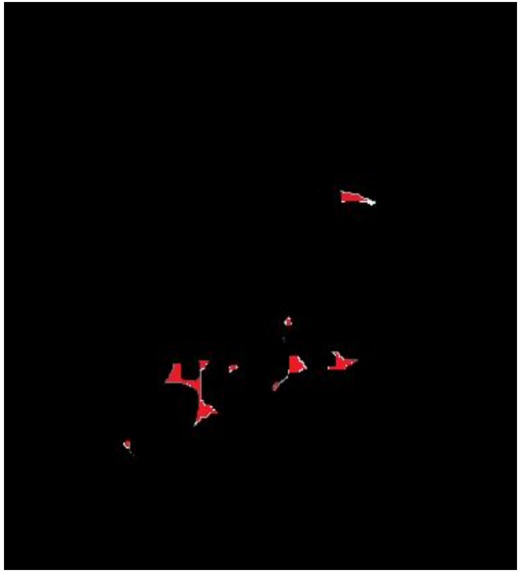
Shape-predictive regions

What does it looks like?



Upper
bound

Segmentation



Lower
bound

Volume-predictive intervals

Standard

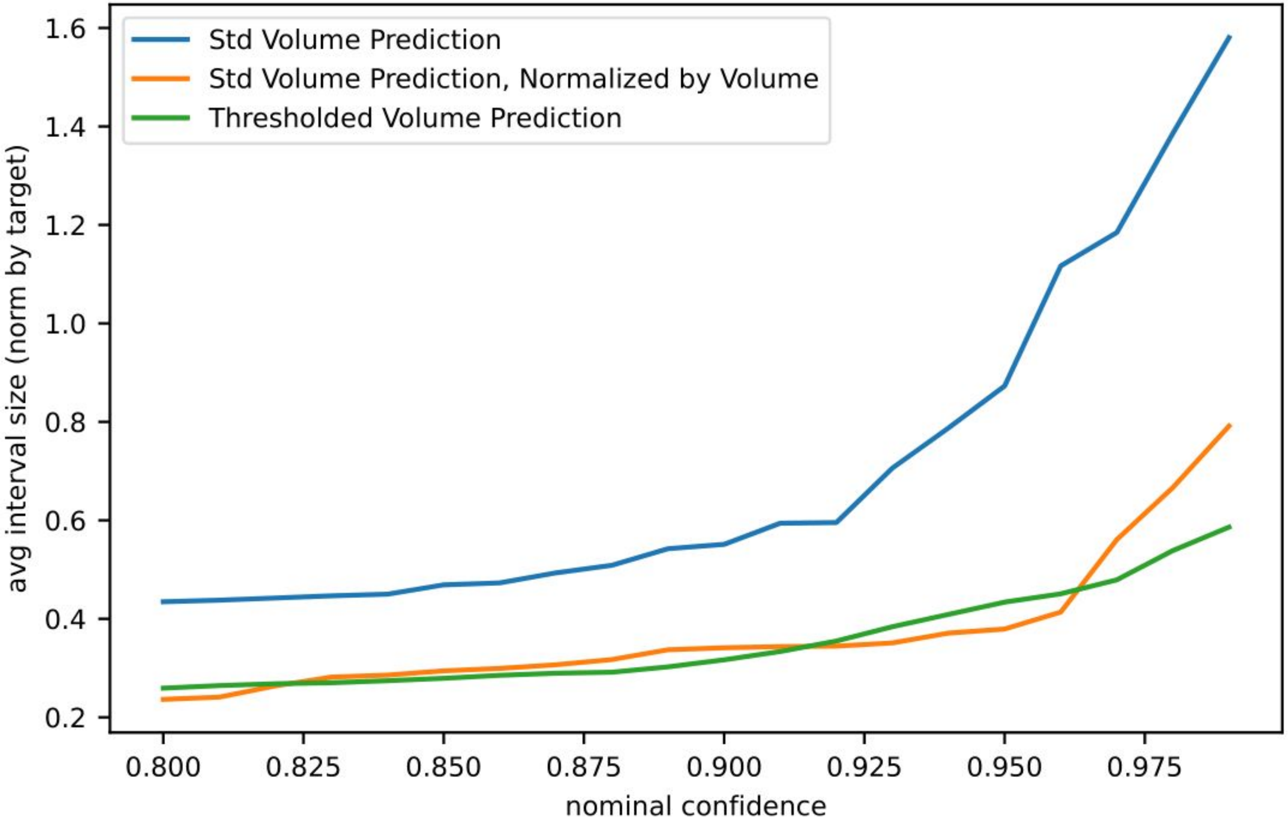
1. Choose the interval confidence (p%)
2. Calculate the distance between the true volume and the predicted volume
3. Choose the interval radius (r) as the number that is bigger than p% of all calculated distances
4. Define the confidence interval as the predicted volume ± r

Normalized by volume

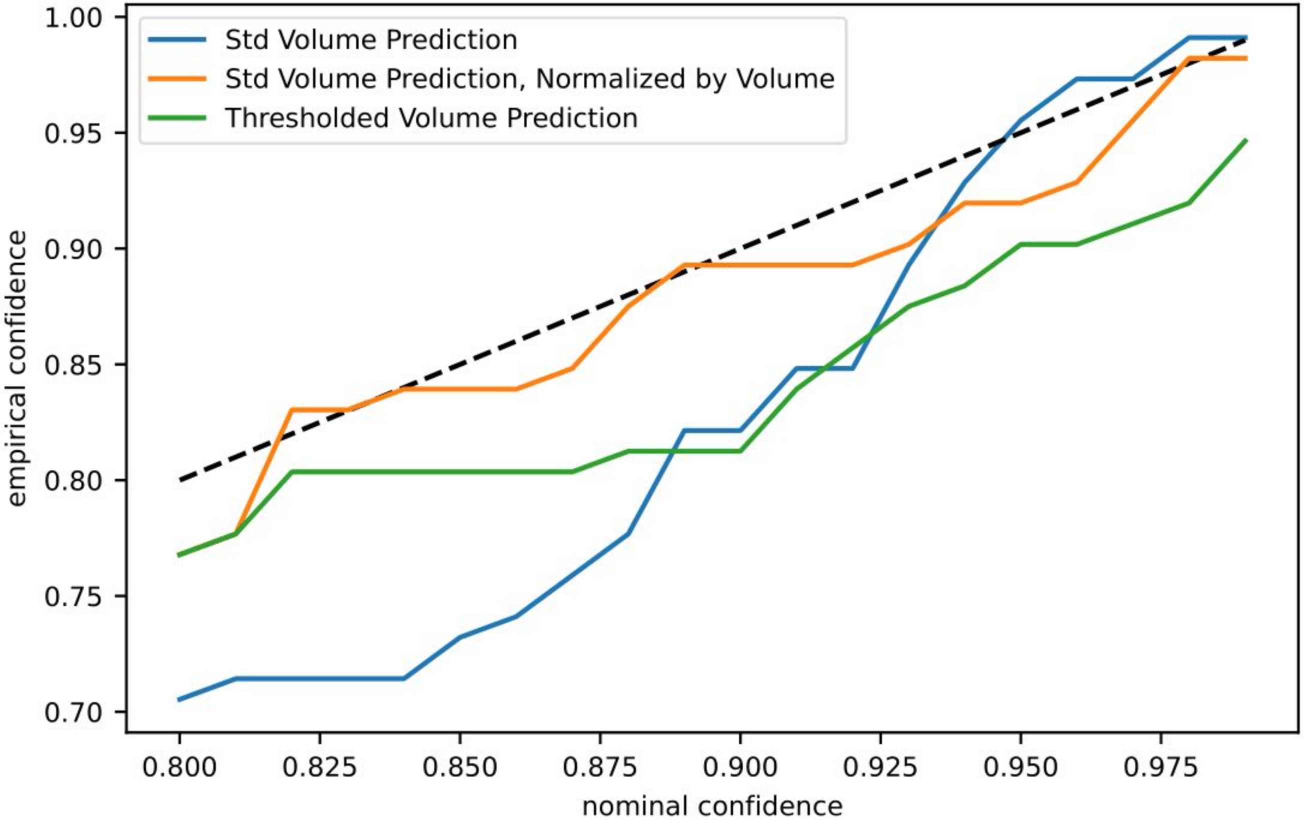
1. Choose the interval confidence (p%)
2. Calculate the percentages of the errors in comparison with the predicted volumes
3. Choose the error percentual (error%) as the number that is bigger than p% of all calculated error percentages
4. Define the interval radius (r) as error% of the predicted volume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

Performance of volume-prediction intervals:

Interval length

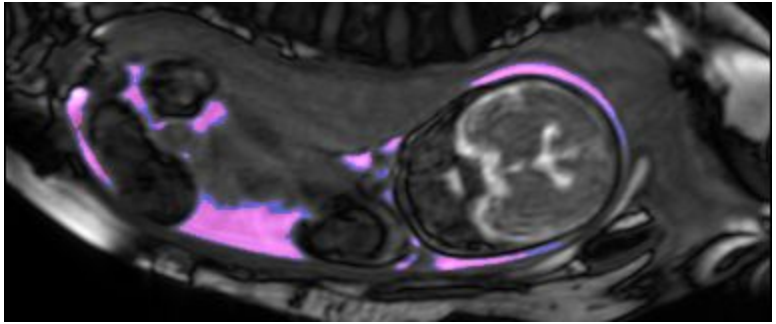


Performance of volume-prediction intervals:
Empirical vs nominal confidence

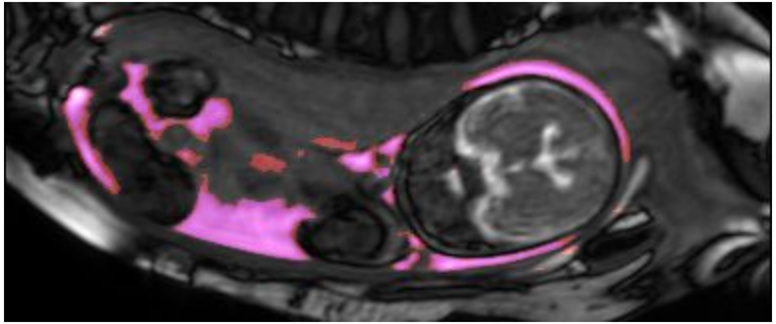


Results of shape-predictive regions:

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.

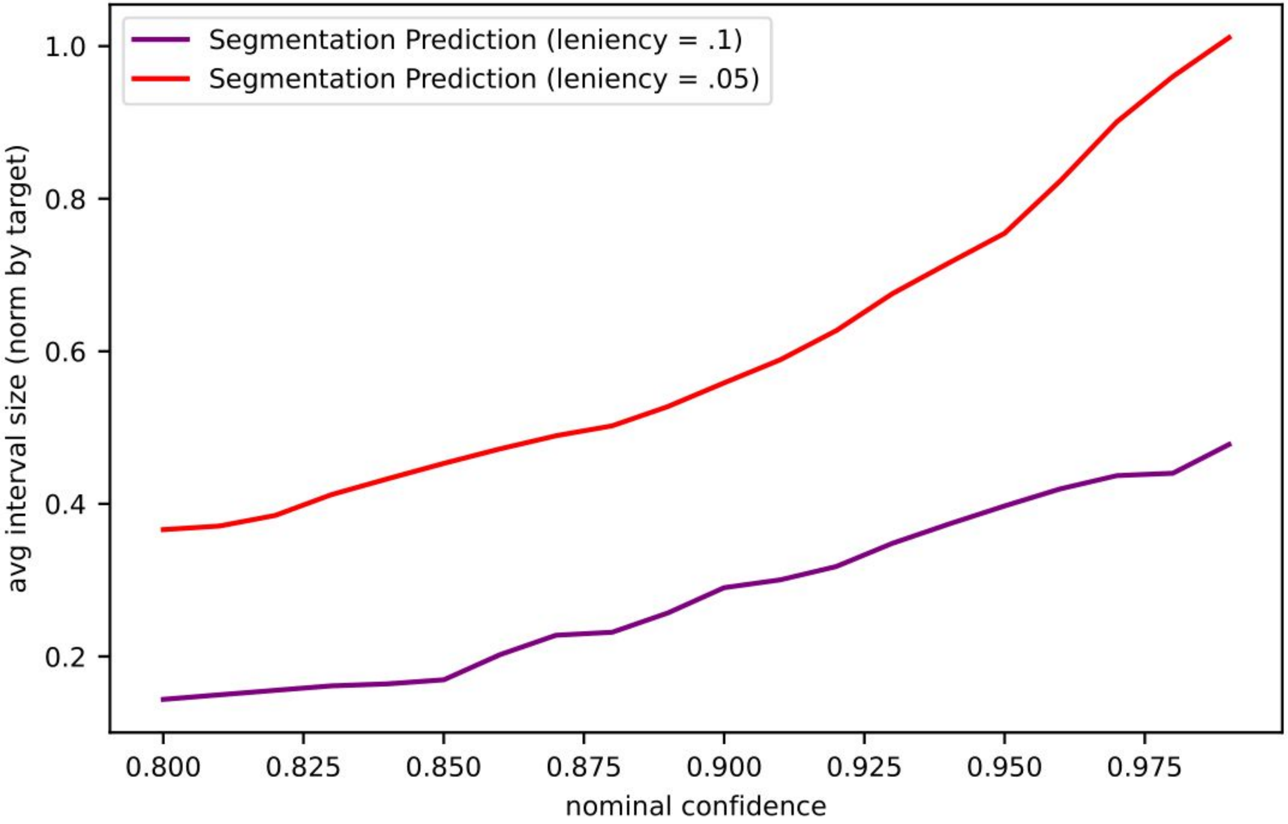


- Region in the **upper bound** and not in the **segmentation** in red.
- Region in the **segmentation** and in the **upper bound** in magenta.

Confidence = 90%
Leniency = 5%

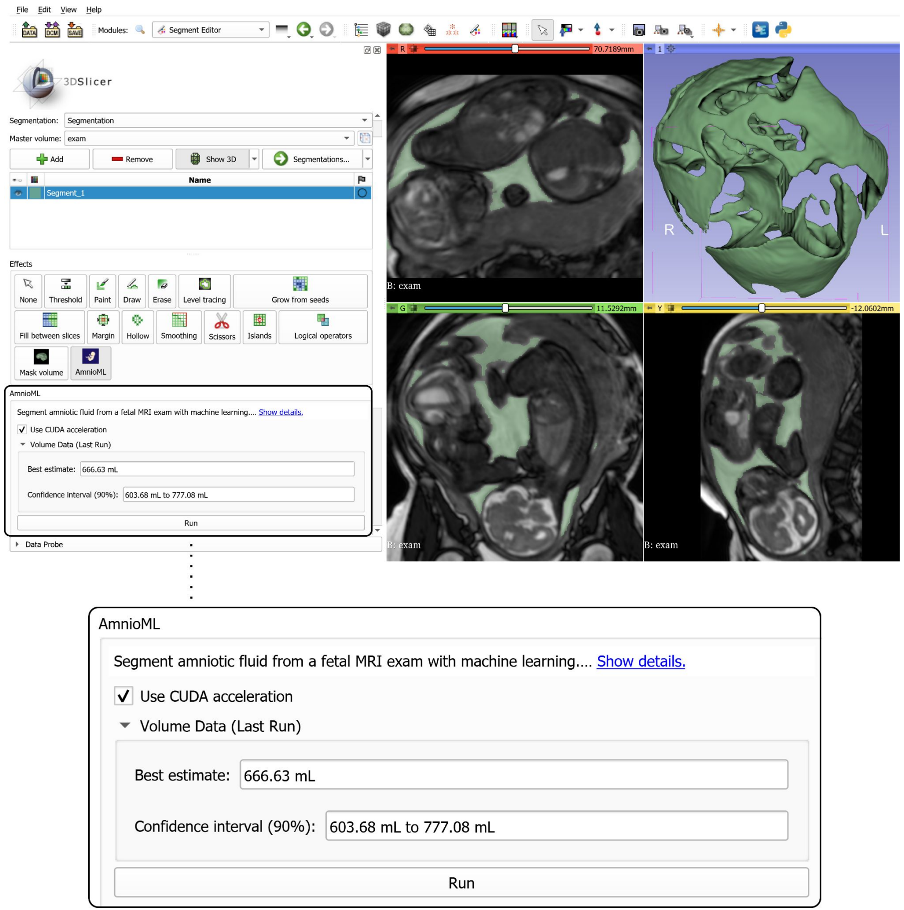
Performance of shape-predictive regions:

How tight is it?

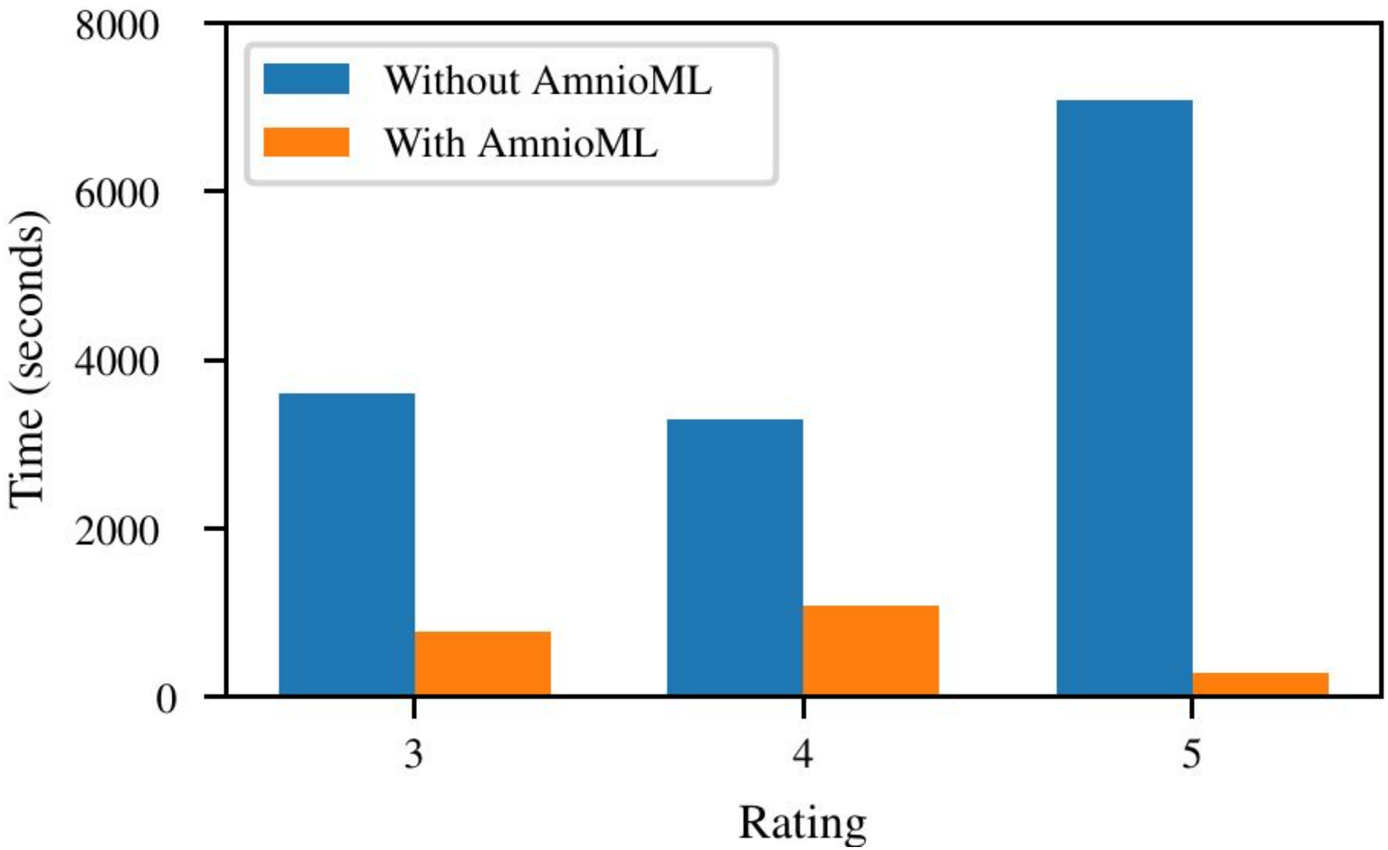


5. AmnioML

AmnioML's graphical user interface



Performance in practice



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

AmnioML Github Repository