Uncertainty Quantification for Amniotic Fluid Segmentation and Volume Prediction

Daniel Csillag, Lucas Monteiro, Thiago Ramos, João Vitor Romano, Rodrigo Schuller, Roberto I. Oliveira, Roberto B. Seixas, Paulo Orenstein
Instituto de Matemática Pura e Aplicada, Rio de Janeiro, Brazil

Summary

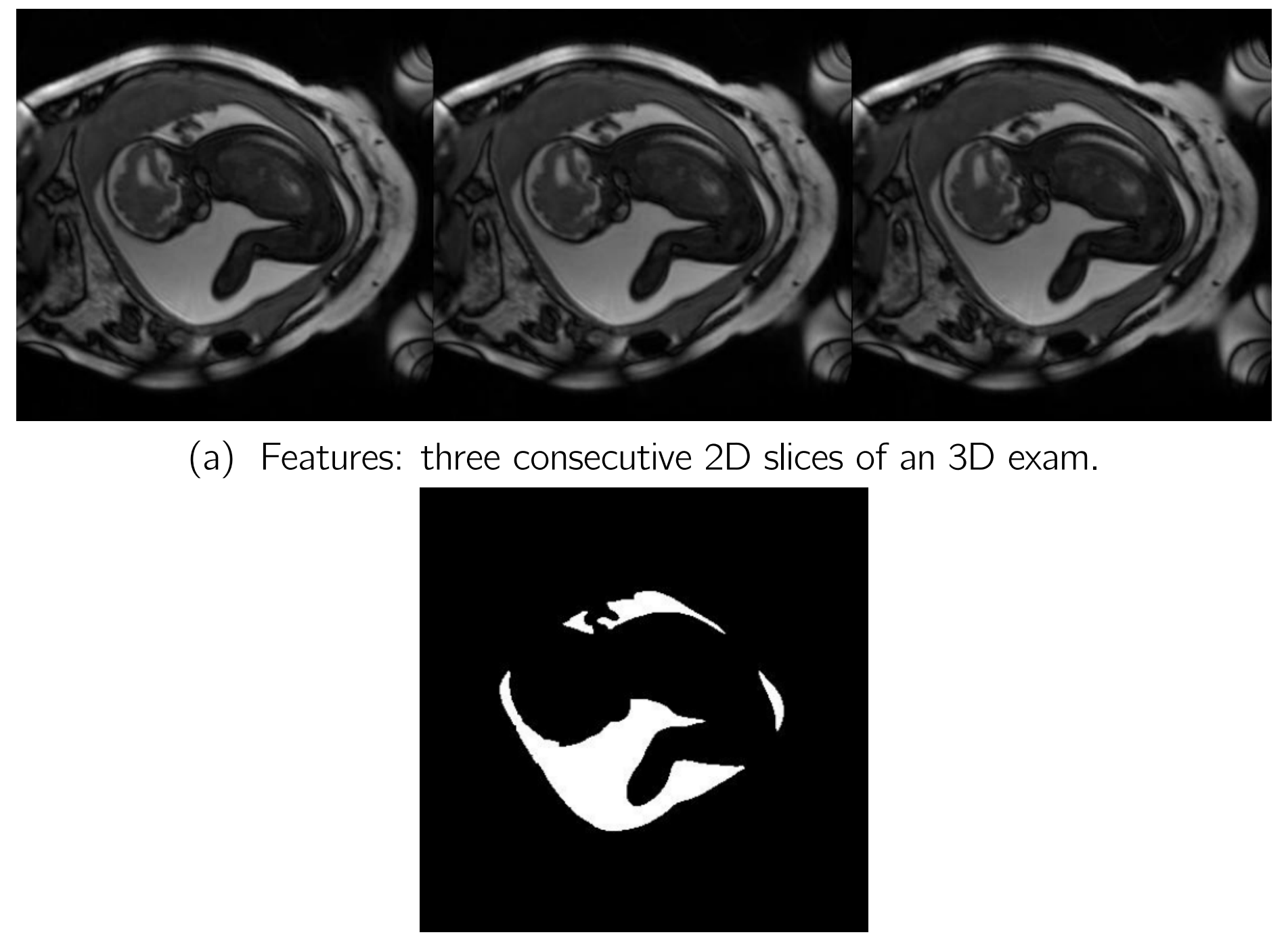
In many medical segmentation tasks, it is crucial to provide valid confidence intervals to machine learning predictions. In the case of segmenting amniotic fluid using fetal MRIs, this allows doctors to better understand and control the segmentation masks, bound the fluid volume, and statistically detect anomalies such as cysts. Our goals in this work are:

- Propose and evaluate different ways of creating confidence intervals for segmentation masks and volume predictions using tools from the field of conformal prediction.
- Show that simple but well-suited modifications of current methods, such as volume normalization and tuning of a leniency hyperparameter, lead to significant improvements, resulting in more consistent coverage and narrower confidence sets

These advances are thoroughly illustrated in the amniotic fluid segmentation problem.

New Fetal MRI Dataset

We evaluate 652 fetal MRIs performed by the same fetal radiologist. 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. The amniotic fluid (AF) of the MRI scans were manually segmented by two specialists, under the supervision of a third specialist and the radiologist that performed the exams. Whenever one of the supervisors disagreed with the segmentation, it was either refined or discarded. For approved segmentations, the AF was highlighted.



(a) Features: three consecutive 2D slices of an 3D exam.

(b) Target: segmented Amniotic Fluid of the middle slice.

Fig 1: Example of 2D exam slices and its highlighted amniotic fluid.

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.

Model Evaluation

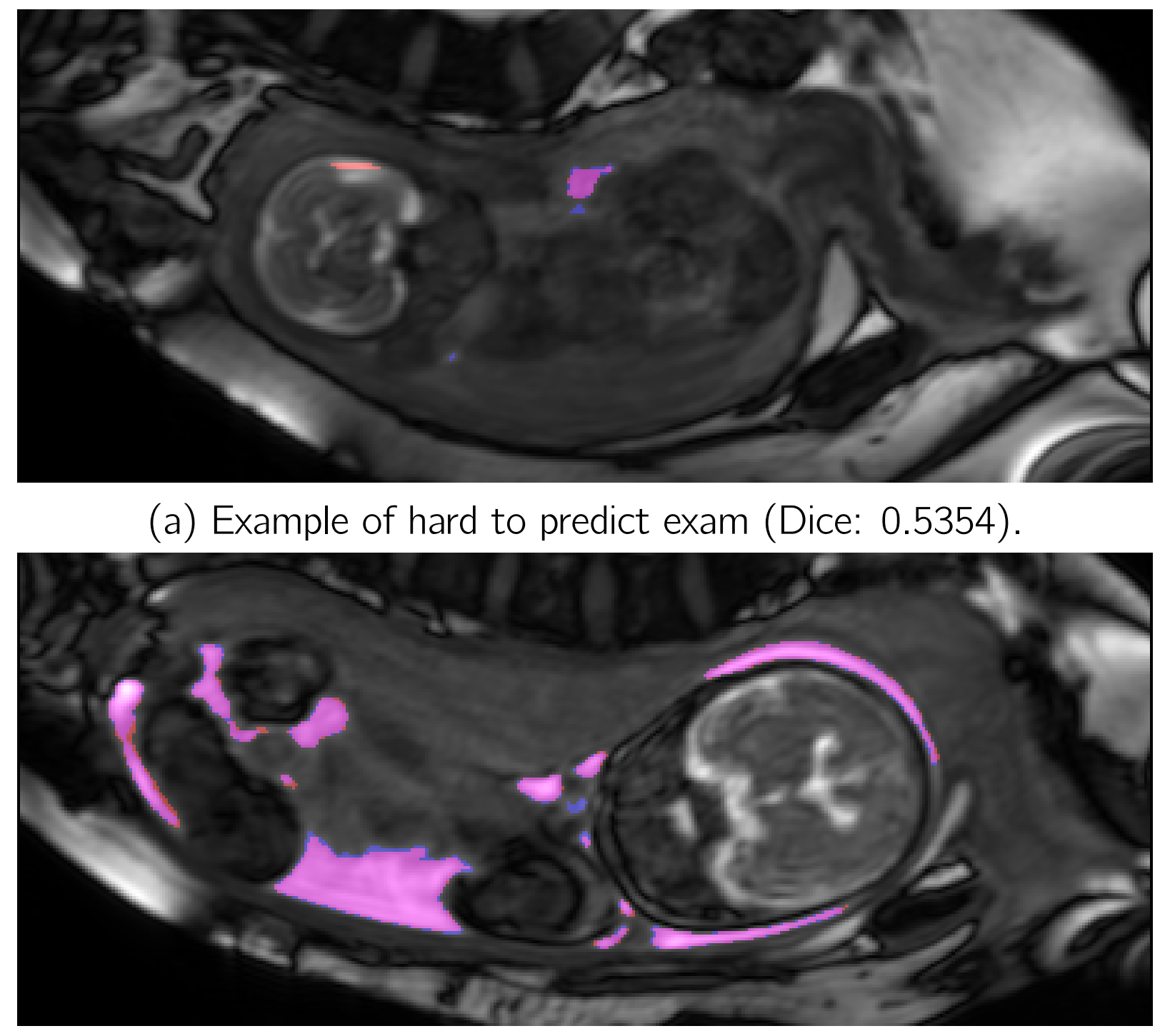
The neural network architectures tested were: U-Net, with 17 million parameter; Small U-Net, 1.9 million parameter; and Fast-SCNN, 1.1 million parameters

The losses considered were Soft Dice and binary cross entropy (BCE). As we observed that mis-segmentation happens mostly close to the border of AF, the Active Contour (AC) loss was used in conjunction with BCE.

The results are shown bellow:

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

Table 1: Average test Dice coefficient and standard deviation.



(a) Example of hard to predict exam (Dice: 0.5354).

(b) Example of typical exam (Dice: 0.9352).

Fig 2: The region correctly segmented by the U-Net using BCE is in magenta, while blue indicates missing regions from the predictions and red indicates excessive segmentation.

Volume-predictive Regions

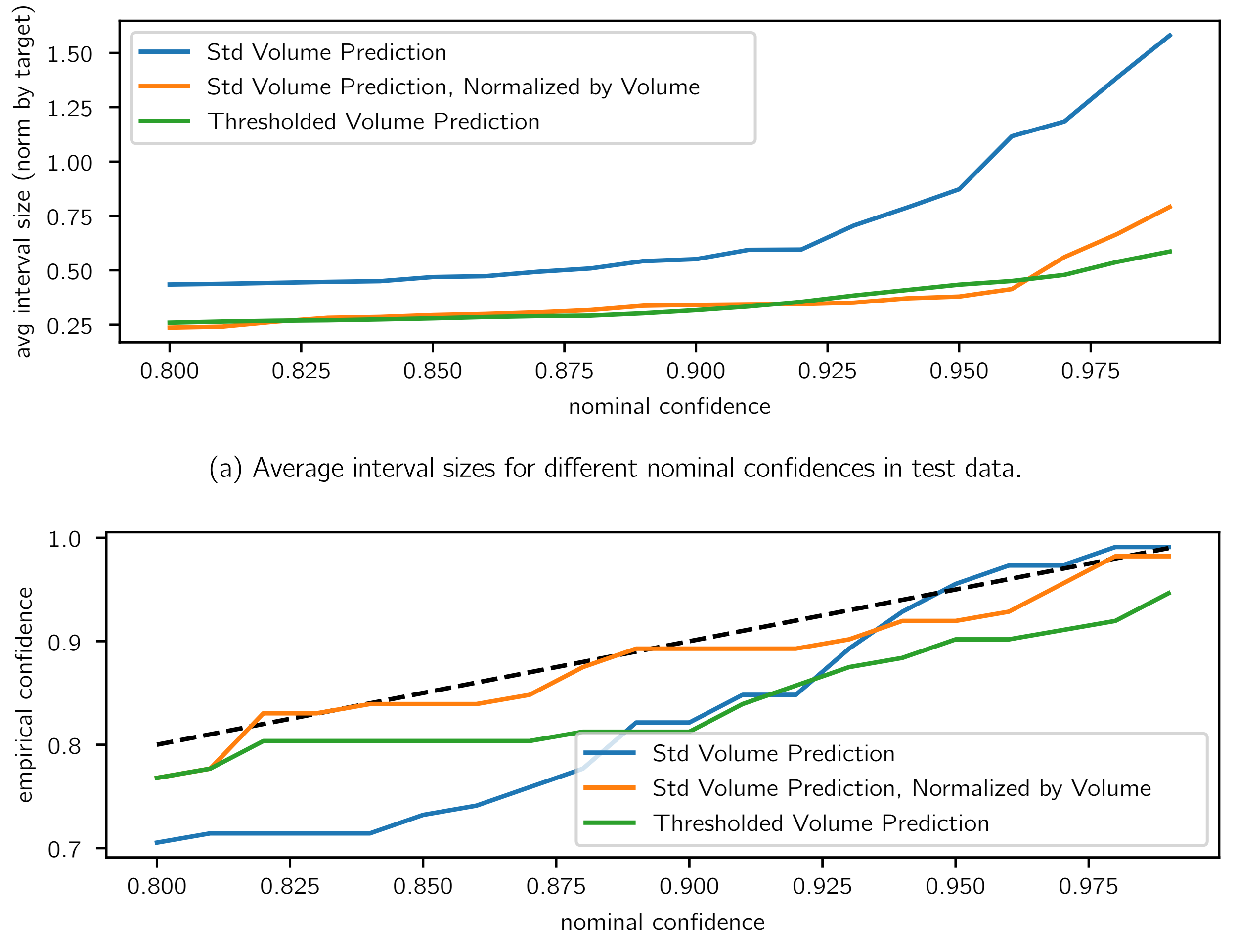
We consider different ways of generating confidence intervals for AF volume estimates.

Standard Volume Prediction: Our first method gives confidence intervals whose lengths on each test point are of the form $2g(\mathcal{M}(X_j)) \cdot$ radius where X_j is a test point and radius is a X_j-independent value chosen from validation data.

The case $g \equiv 1$ is standard for building constant-length predictive intervals. We propose taking $g(\mathcal{M}(X_j)) = \text{Vol}(\mathcal{M}(X_j))$. This simple but powerful modification makes the interval lengths adaptive to the wide range of volumes in the data.

Thresholded Volume Prediction: We consider a Thresholded Volume Prediction algorithm, 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.

Results: Generally, the proposed Standard Volume Prediction algorithm normalized by $g(\mathcal{M}(X_j)) = \text{Vol}(\mathcal{M}(X_j))$ yields the best results of the methods considered.



(a) Average interval sizes for different nominal confidences in test data.

(b) Empirical vs. nominal coverage in test data.

Fig 3: The Standard Volume Prediction normalized by volume, in orange, and Thresholded Volume Prediction, in green, have similar average lengths, but the former has more consistent coverage.

Shape-predictive Regions

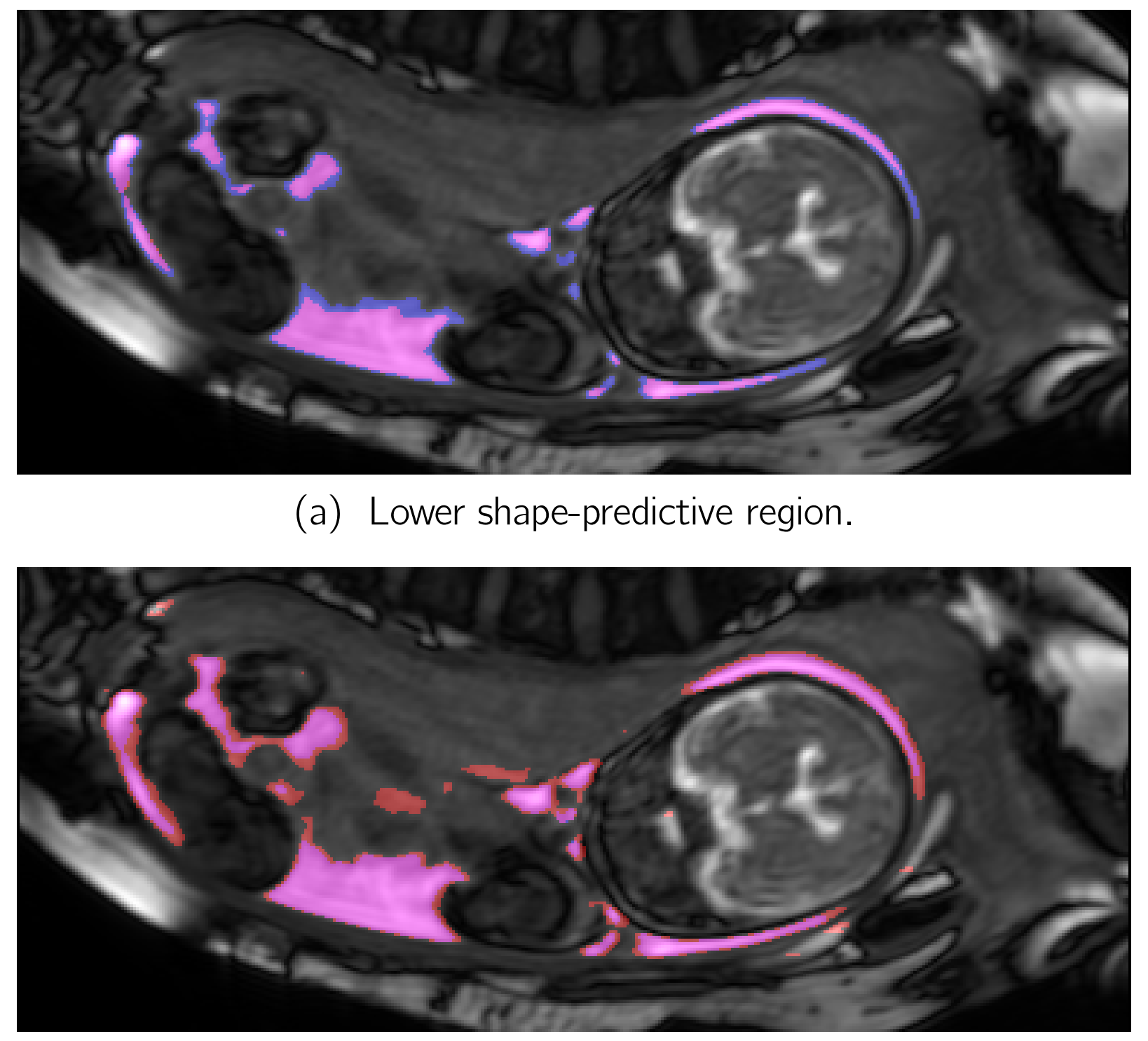
To obtain shape-predictive regions that can upper and lower bound the segmentation masks, we consider an algorithm which besides the level $0 < \alpha < 1$, also takes another user-specified leniency parameter $\lambda \geq 0$.

Leniency: For a positive leniency, we construct a confidence set $\mathcal{R}_\alpha(X_j)$ for an segmented exam Y_i as following: we say that $Y_j \in \mathcal{R}_\alpha(X_j)$ if

$$\max\{\text{Vol}\left(Y_j \backslash \mathcal{U}_{\alpha,\lambda}(X_j)\right), \text{Vol}\left(\mathcal{L}_{\alpha,\lambda}(X_j) \backslash Y_j\right)\} \leq \lambda \text{Vol}\left(Y_j\right).$$

Here, the lower and upper masks, $\mathcal{L}_{\alpha,\lambda}$, $\mathcal{U}_{\alpha,\lambda}$, are obtained from $\mathcal{M}(X_j)$ by thresholding at values learned from the validation data.

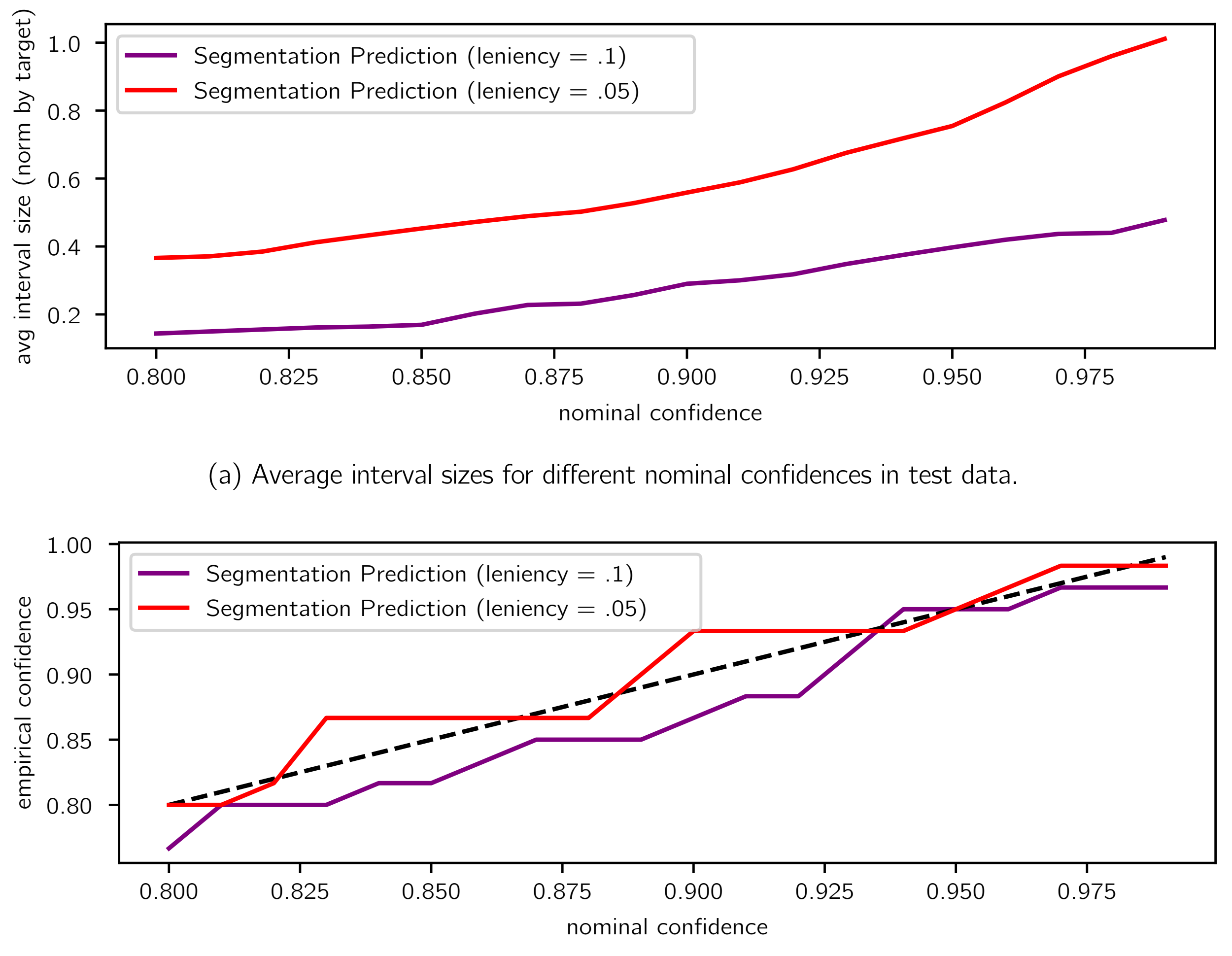
Intuitively, the idea is to add some slack to the region $\mathcal{R}_\alpha$ so the containment over lower and upper masks, $\mathcal{L}_{\alpha,\lambda}(X_j) \subset Y_j \subset \mathcal{U}_{\alpha,\lambda}(X_j)$, does not have to be exact. In practice, there is often some degree of subjectivity on how specialists segment the borders of AF.



(a) Lower shape-predictive region.

(b) Upper shape-predictive region.

Fig 4: Shape-predictive regions for $\alpha = 0.1$ and leniency $\lambda = 0.05$. Magenta indicates the region correctly segmented, while blue denotes missing segmentation and red indicates the excess segmentation.



(a) Average interval sizes for different nominal confidences in test data.

(b) Empirical vs. nominal coverage in test data.

Fig 5: Larger leniency leads to narrower intervals since there is more flexibility when constructing the shape-predictive regions. Smaller leniency leads to more consistent empirical confidence.

Results: There is trade-off when choosing the leniency parameter: smaller values lead to wider intervals but slightly more consistent empirical confidence, as shown in Figure 5b. Overall, not allowing any leniency would lead to disproportionately large intervals, and seems to be too harsh an inclusion notion for medical segmentation.

Algorithms

The pseudo-code for the algorithms considered are below.

```
Algorithm 1: Standard Volume Prediction
Input: model $\mathcal{M}$, validation set $\{(X_i, Y_i)\}_{i=n+1}^{m}$,
       test set $\{(X_i, Y_i)\}_{i=m+1}^{\ell}$, 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 $\left(\frac{|\text{Vol}(\mathcal{M}(X_i)_{\geq .5}) - \text{Vol}(Y_i)|}{g(\mathcal{M}(X_i)_{\geq .5})}\right)$ to radii
end
radius $\leftarrow$ $(1 - \alpha)$-quantile of radii
for $j \in \{m+1, ..., \ell\}$ do
    $dv \leftarrow g(\mathcal{M}(X_j)_{\geq .5}) \cdot$ radius ;
    lower volume $\leftarrow \text{Vol}(\mathcal{M}(X_j)_{\geq .5}) - dv$ ;
    upper volume $\leftarrow \text{Vol}(\mathcal{M}(X_j)_{\geq .5}) + dv$ ;
    $\mathcal{I}_\alpha(X_j) \leftarrow$ [lower volume, upper volume]
end
Output: $\mathcal{I}_\alpha(X_j)$, $j = m+1, \ldots, \ell$
```

```
Algorithm 2: Thresholded Volume Prediction
Input: model $\mathcal{M}$, validation set $\{(X_i, Y_i)\}_{i=n+1}^{m}$,
       test set $\{(X_i, Y_i)\}_{i=m+1}^{\ell}$ 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 $j \in \{m+1, ..., \ell\}$ do
    lower volume $\leftarrow \text{Vol}(\mathcal{M}(X_j)_{\geq \text{lower bound}_t})$ ;
    upper volume $\leftarrow \text{Vol}(\mathcal{M}(X_j)_{\geq \text{upper bound}_t})$ ;
    $\mathcal{I}_\alpha(X_j) \leftarrow$ [lower volume, upper volume]
end
Output: $\mathcal{I}_\alpha(X_j)$, $j = m+1, \ldots, \ell$
```

```
Algorithm 3: Segmentation Prediction
Input: model $\mathcal{M}$, validation set $\{(X_i, Y_i)\}_{i=n+1}^{m}$, test set $\{(X_i, Y_i)\}_{i=m+1}^{\ell}$, 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 $= \lambda_{\text{lower}}$-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 $j \in \{m+1, ..., \ell\}$ do
    $\mathcal{U}_{\alpha,\lambda}(X_j) \leftarrow \mathcal{M}(X_j)_{\geq \text{upper bound}_t}$; $\mathcal{L}_{\alpha,\lambda}(X_j) \leftarrow \mathcal{M}(X_j)_{\geq \text{lower bound}_t}$; $\mathcal{R}_{\alpha,\lambda}(X_j) = (\mathcal{L}_{\alpha,\lambda}(X_j), \mathcal{U}_{\alpha,\lambda}(X_j))$
end
Output: $\mathcal{R}_{\alpha,\lambda}(X_j)$, $j = m+1, \ldots, \ell$
```

Conclusions and Takeaways

Simple modifications to standard conformal predictions yield accurate and useful confidence sets on both medical image segmentations and volume estimates:

- **For volume-predictive regions**: Using volume-predictive intervals with adaptive sizes leads to narrower intervals than the standard normalization, while maintaining proper coverage (Figure 3).
- **For shape-predictive regions**: The effective use of a leniency parameter in shape-predictive regions give good upper and lower confidence sets that come with theoretical guarantees (Figure 4), and have the potential to visually aid radiologists when performing MRI segmentation (Figure 5).

Main References

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

[2] 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 CVPR*, pp. 11632–11640, 2019.

[3] 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`.

[4] Ronneberger, O., Fischer, P., and Brox, T. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing andComputer-Assisted Intervention (MICCAI)*, volume 9351 of LNCS, pp. 234–241. Springer, 2015. URL `http://jmlr.org/papers/v9/shafer08a.html`
(available on arXiv:1505.04597 [cs.CV]).