Aggregated Attributions for Explanatory Analysis of 3D Segmentation Models

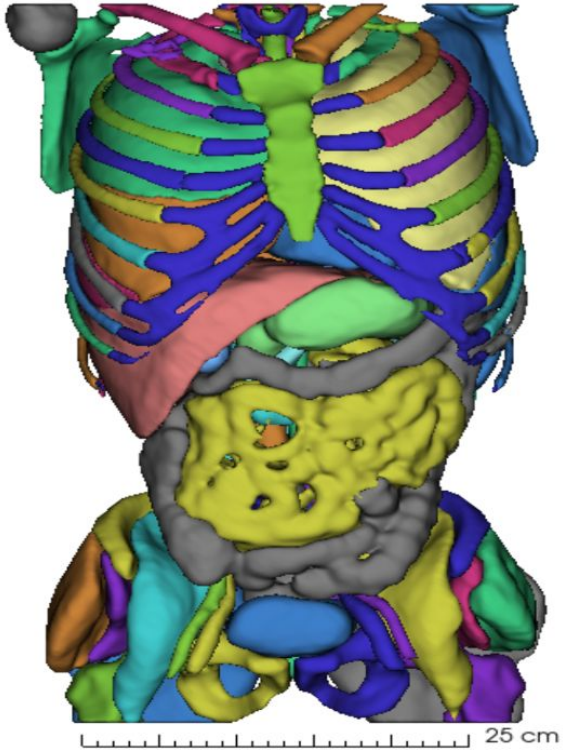
Maciej Chrabąszcz[1,2], Hubert Baniecki[2,3], Piotr Komorowski[3], Szymon Płotka[4], Przemysław Biecek[2,3]

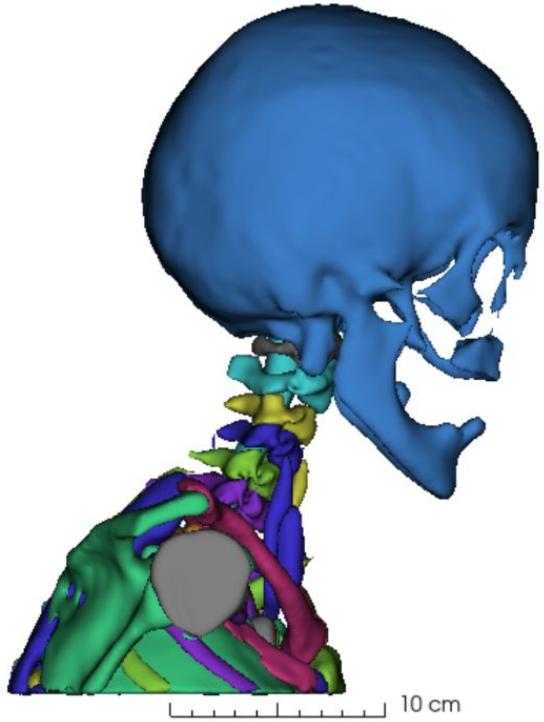
1 - NASK - National Research Institute
2 - Warsaw University of Technology
3 - University of Warsaw
4 - University of Amsterdam

Motivation

3D segmentation explanations face critical challenges in medical imaging:

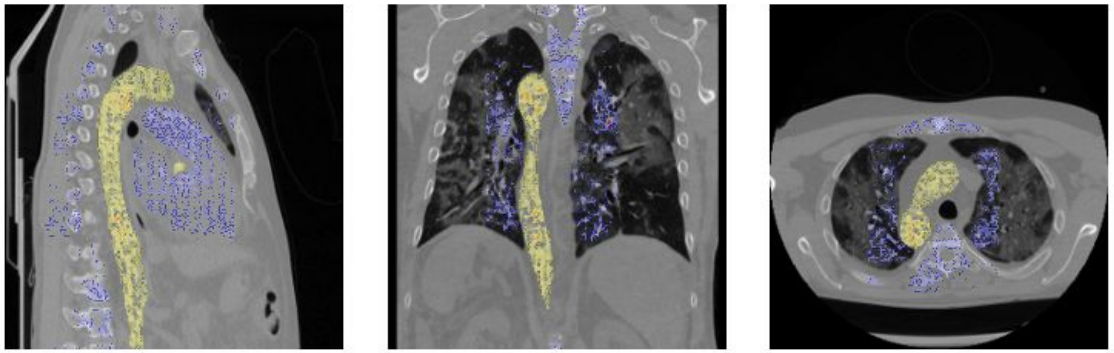
1) High dimensionality of input/output
2) Limited explainability
3) Potential biases
4) Trust issues with black-box models
5) High risk domain



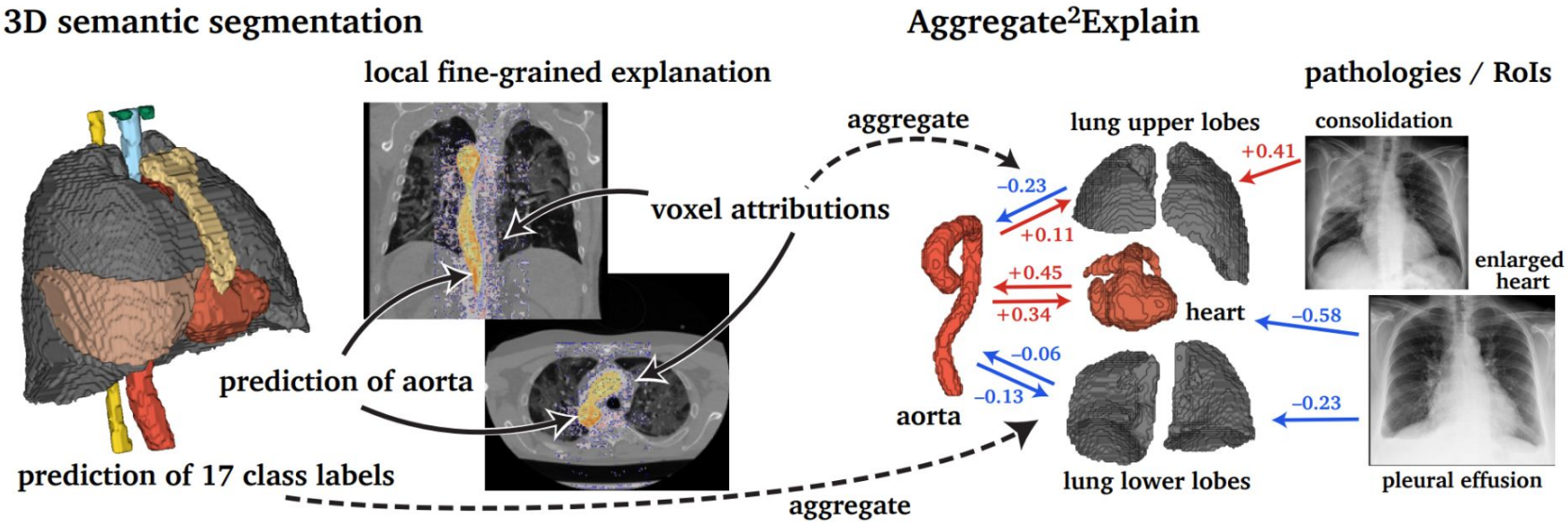


Problems with attributions for 3D segmentation models.

- Many attributions methods work for scalar outputs e.g. image classification, whereas segmentation models output multidimensional outputs.
- Attributions on their own take a lot of time to analyze how the model works.
- Attributions for 3D inputs are time-consuming to analyze, e.g. 512^3 voxels $\approx 10^8$ features.



Agg^{2}Exp Methodology



How to calculate attributions for segmentation

Because traditional attribution methods were develop for single output we create a proxy which transforms our multidimensional output into a scalar using $\boldsymbol{y}_c$ matrix for class $\boldsymbol{c}$ of interest:

$$\hat{\mathbf{y}}_c^{(i)} := \begin{cases} 1, & \text{if } \operatorname{argmax}_j \hat{\mathbf{y}}^{(i,j)} = c, \\ 0, & \text{otherwise.} \end{cases} \qquad (1)$$

Having calculated this we are able to calculate a proxy output which is a scalar:

$$\hat{f}_c(\mathbf{x}) := \mathcal{A}(f_c(\mathbf{x})) = f_c(\mathbf{x}) \odot \hat{\mathbf{y}}_c, \qquad (2)$$

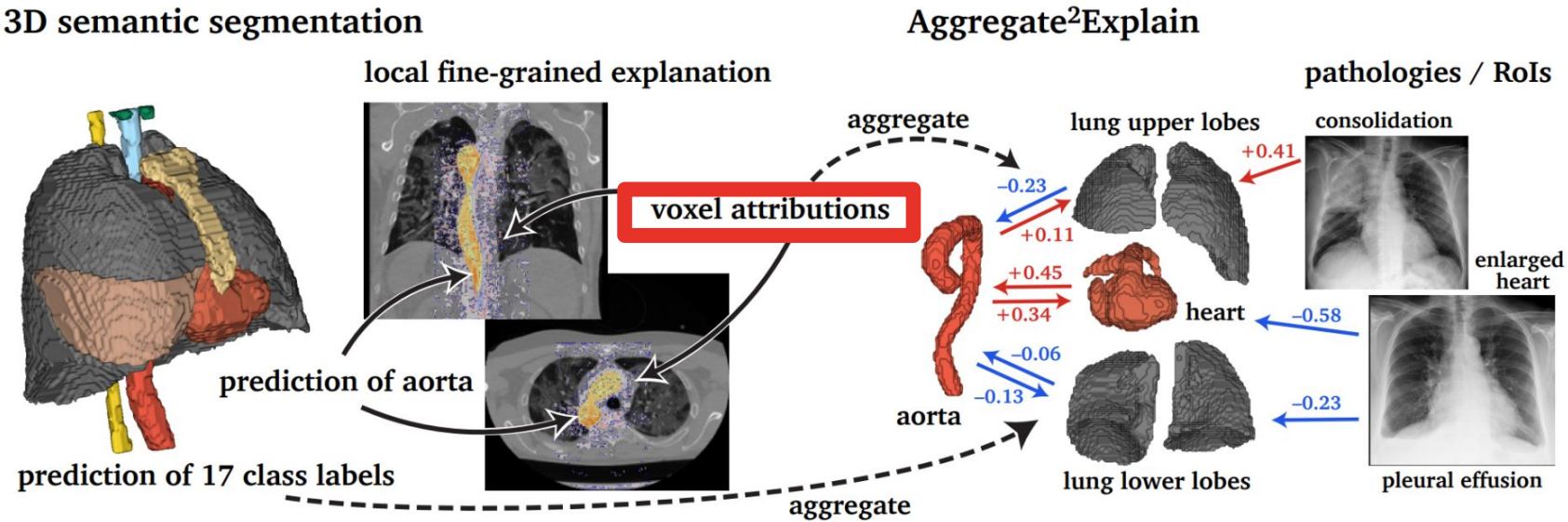
Gradient attributions for segmentation

Having calculated proxy we can now easily calculate gradient-based and perturbation-based attributions previously prepared for scalar output e.g.

$$g_{\mathrm{VG}}(\mathbf{x}; \hat{f}_c) := \frac{\partial \hat{f}_c(\mathbf{x})}{\partial \mathbf{x}}. \tag{3}$$

$$g_{\mathrm{SG}}(\mathbf{x}; \hat{f}_c, n, \sigma^2) := \frac{1}{n} \sum_{i=1}^{n} g_{\mathrm{VG}}(\mathbf{x} + \mathcal{N}(0, \sigma^2); \hat{f}_c). \tag{4}$$

Agg^{2}Exp methodology



Aggregation of attributions for RoI importance

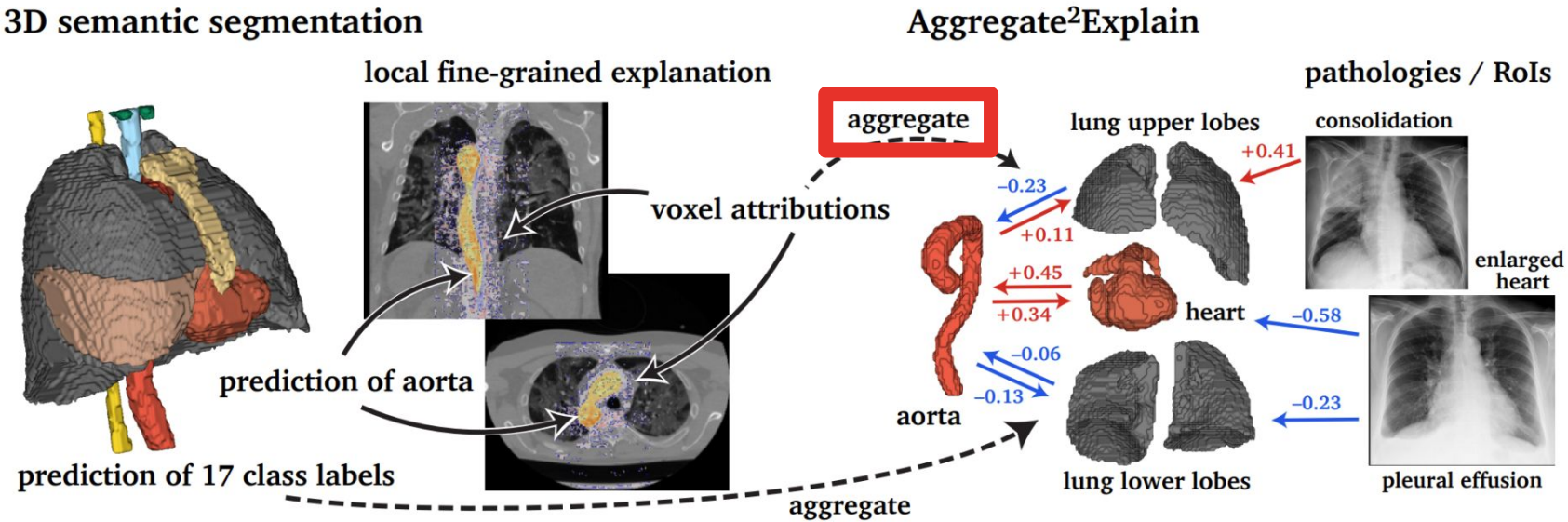
To overcome this challenge of high complexity stemming from the multiplication of input and output dimensions, we propose to aggregate voxel attributions into simpler explanations that are easier to visualize and analyze.

First you select RoI (in our case segmentation regions) to which you limit your attribution. Then you can Calculate this RoI importance for segmentation of class ***a***:

$$e_{b\to a} := \frac{\|\mathbf{e}_{b\to a}\|_1}{\|\mathbf{e}_a\|_1},$$

Having calculated local RoI importance one can calculate *average* of specific RoIs importance for all patients yielding a Global RoI importance

Agg^{2}Exp methodology



Experimental Setup

- 2 Datasets
  - TotalSegmentator v2 (public)
  - B50 (private clinical dataset)
- Swin-Unter model trained on 16 thorax anatomical structures
- 4 attributions methods
  - Vannilla Gradient
  - SmoothGrad
  - Integrated Gradients
  - KernelShap

Attributions methods evaluation

We evaluated selected attribution methods using *Faithfulness, Sensitivity, Complexity* and *Efficiency* metrics.

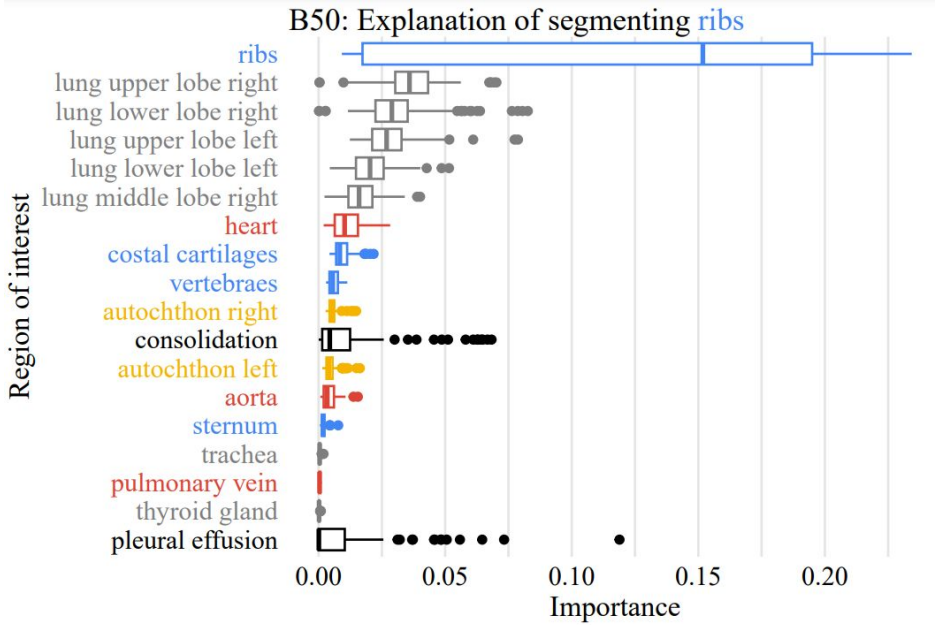
Metric	**What it measures**
Faithfulness	Measures whether voxels indicated as important by an attribution method are indeed important to the model's prediction.
Sensitivity	Determines how close attribution explanations are for similar inputs.
Complexity	Fraction of voxels whose attribution scores exceed a specified threshold.
Efficiency	Seconds needed to create attribution.

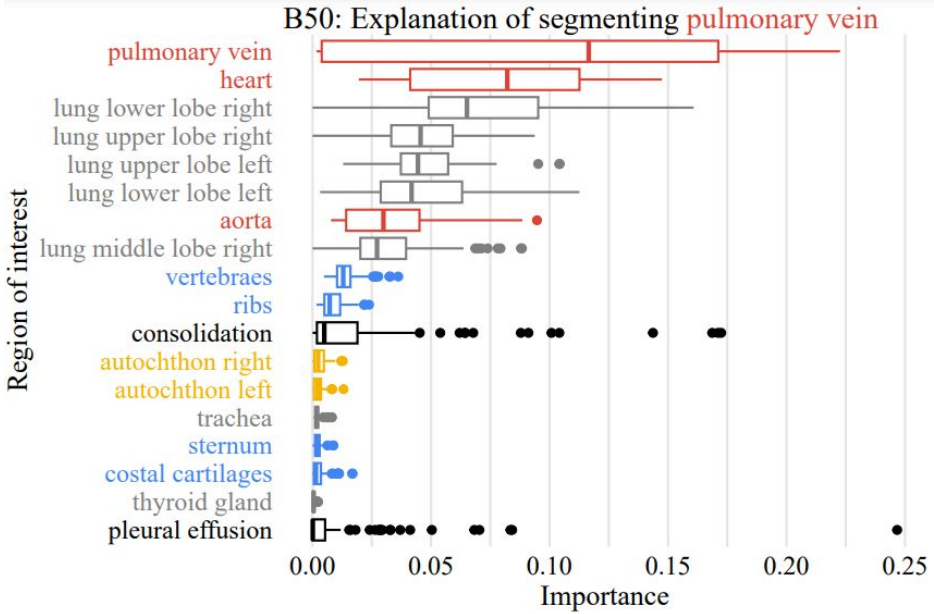
Attributions methods evaluation

We evaluated selected attribution methods using *Faithfulness, Sensitivity, Complexity* and *Efficiency* metrics.

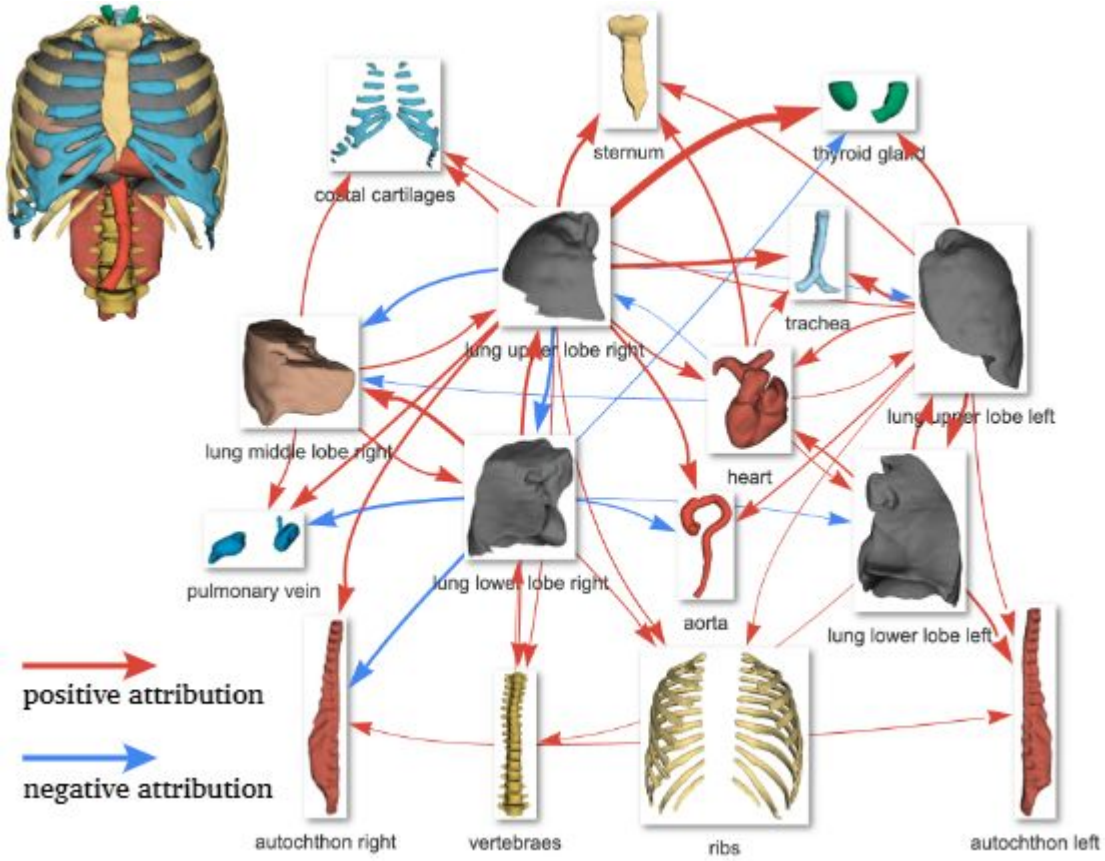
Dataset	Attribution method g	Faithfulness $\mu_F \uparrow$	Sensitivity $\mu_S \downarrow$	Complexity $\mu_C \downarrow$	Efficiency [s] $\downarrow$
TSV2	KernelSHAP (cubes)	$0.038_{\pm 0.11}$	$0.157_{\pm 0.10}$	$0.525_{\pm 0.23}$	$4278_{\pm 3013}$
	KernelSHAP (semantic)	$0.158_{\pm 0.21}$	$0.346_{\pm 0.49}$	$0.924_{\pm 0.05}$	$857_{\pm 604}$
	Vanilla Gradient	$0.383_{\pm 0.14}$	$1.130_{\pm 0.26}$	$0.001_{\pm 0.00}$	$12_{\pm 7}$
	Integrated Gradients	$0.213_{\pm 0.13}$	$0.848_{\pm 0.23}$	$0.001_{\pm 0.00}$	$210_{\pm 148}$
	SmoothGrad	$0.331_{\pm 0.14}$	$0.773_{\pm 0.20}$	$0.001_{\pm 0.00}$	$209_{\pm 148}$
B50	KernelSHAP (cubes)	$0.009_{\pm 0.11}$	$0.341_{\pm 0.38}$	$0.681_{\pm 0.11}$	$10427_{\pm 1751}$
	KernelSHAP (semantic)	$0.040_{\pm 0.11}$	$0.181_{\pm 0.29}$	$0.918_{\pm 0.04}$	$2087_{\pm 349}$
	Vanilla Gradient	$0.411_{\pm 0.14}$	$1.244_{\pm 0.33}$	$0.000_{\pm 0.00}$	$27_{\pm 5}$
	Integrated Gradients	$0.248_{\pm 0.12}$	$0.943_{\pm 0.21}$	$0.000_{\pm 0.00}$	$510_{\pm 86}$
	SmoothGrad	$0.309_{\pm 0.13}$	$0.874_{\pm 0.21}$	$0.000_{\pm 0.00}$	$509_{\pm 86}$

Distributions of explanations

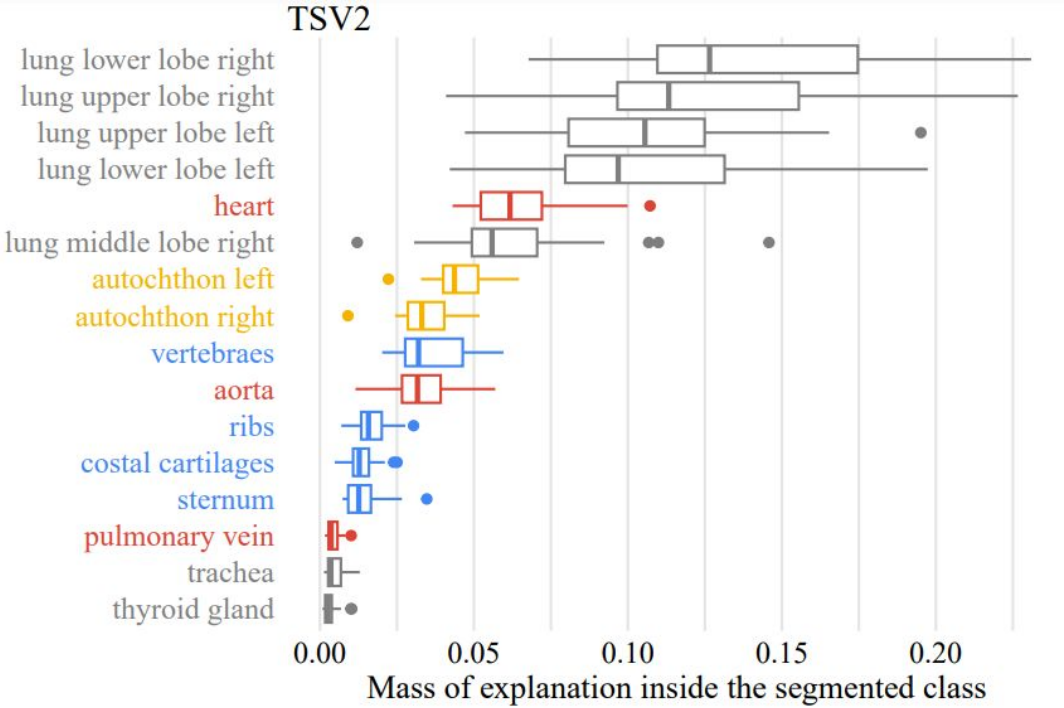




Global analysis of our model



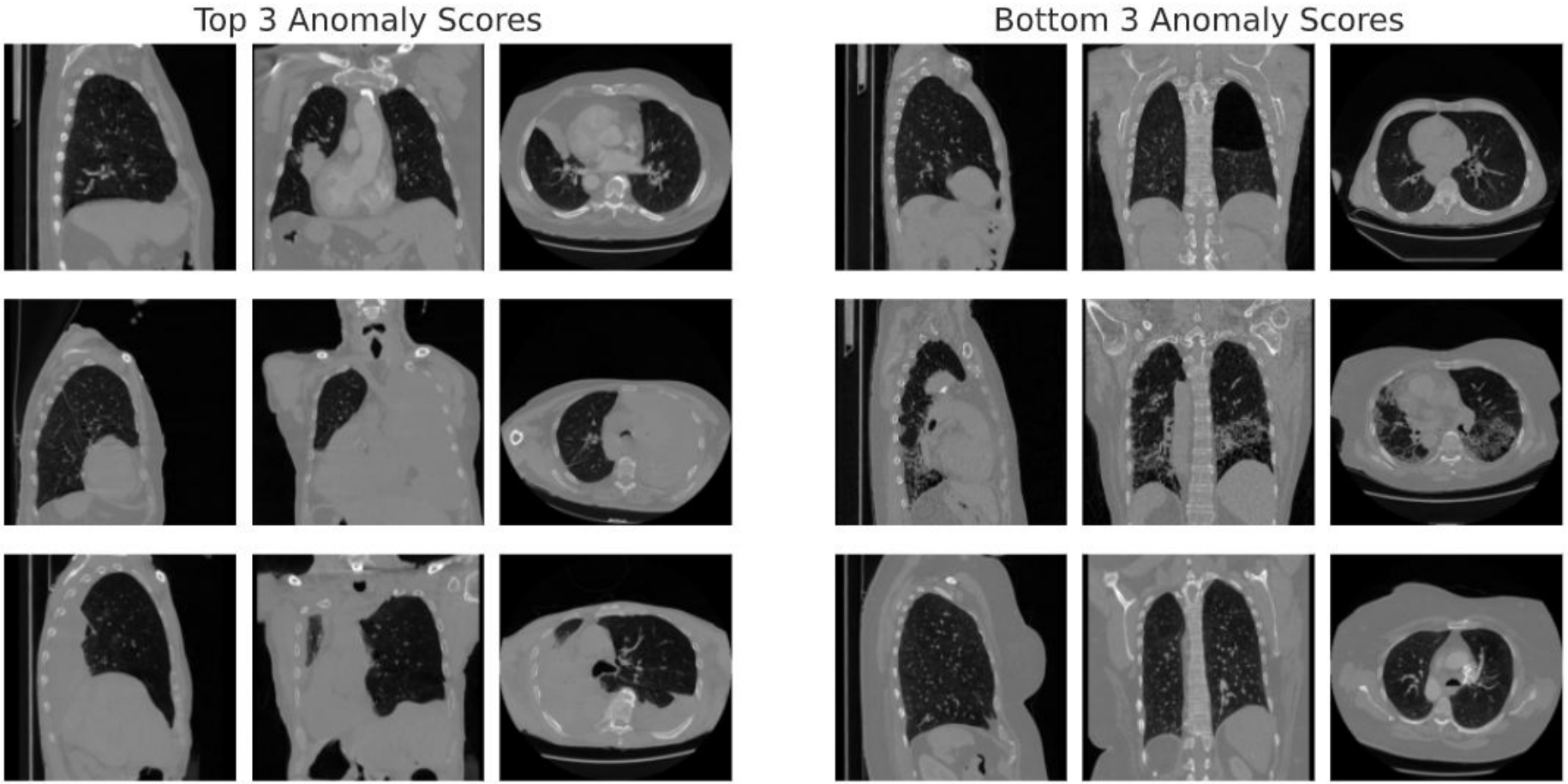
Does our model use contextual information for segmentation?



Detecting anomalies

We decided to try to find anomalies using our aggregated explanations in both TSV2 and B50 datasets. For each of our classes we have 17 features which we use for finding anomalies related to specific class by using Isolation Forest.

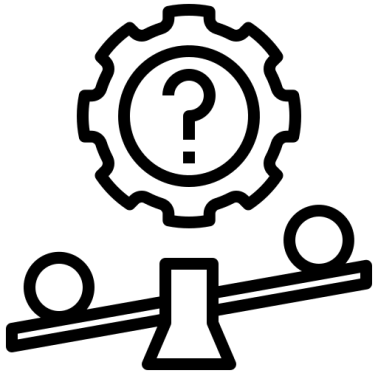
We conduct the Spearman's rank test between the anomaly score returned by IFs and the Dice score from TSV2 test data for each class label, which shows a statistically significant correlation for some labels.


Top 3 Anomaly Scores
Bottom 3 Anomaly Scores

Checking for Gender and Sex Bias

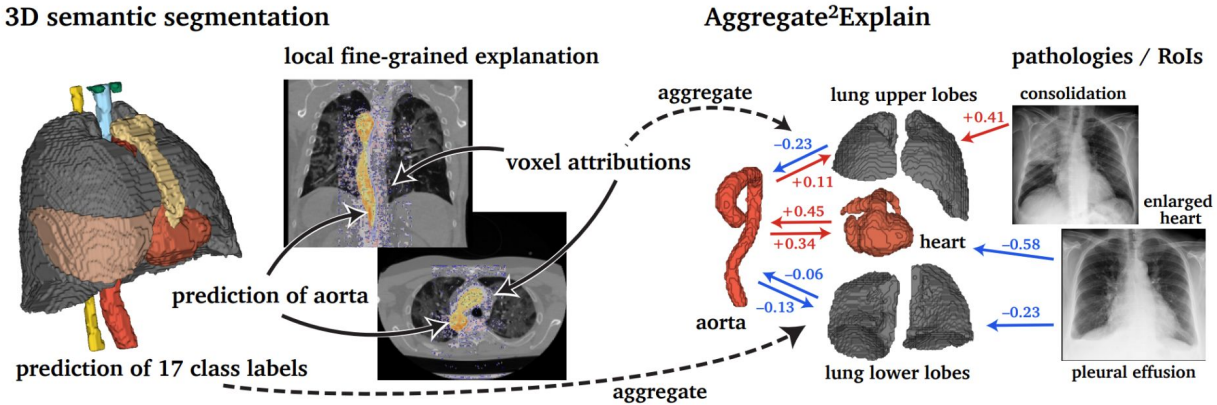
By having some patients flagged as anomalies by having access to metadata we analyzed whether our model doesn’t behave abnormally for specific subpopulations.

To our content we found no statistical evidence for some subgroups being overrepresented in outliers group compared to inliers group.



Summary

- AGG²EXP enables practical analysis of 3D segmentation models.
- Greatly reduces explanation complexity by creating easily interpretable RoIs importance.
- Enables detection of model biases and abnormal behavior.



Questions?