

An Empirical Study on the Fairness of Foundation Models for Multi-Organ Image Segmentation (Supplementary Materials)

Paper ID: 1289

Algorithm 1 Heatmap generation for visualizing segmentation errors in sub-regions of organs (in producing Fig.3 in the main paper)

```

1: Input: The set of model prediction results termed prediction_list and annotation
   termed annotation_list with the number of points termed N in point cloud.
2: Output: Averaged organ template with segmentation error distribution.
3: Initialization Point_cloud_list, Distance_list.
4: for index,item in enumerate(annotation_list): do
5:   Annotation_cloud = To_Cloud(annotation_list[index], N)
6:   Prediction_cloud = To_Cloud(prediction_list[index], N)
7:   Cloud_distance = Euclidean(Annotation_cloud, Prediction_cloud)
8:   Point_cloud_list.append(Annotation_cloud)
9:   Distance_list.append(Cloud_distance)
10: end for
11: Reference_cloud = Point_cloud[0]
12: Registration of the point clouds regarding the reference
13: Center_list, Clusters = K-means(Point_cloud_registration, N)
14: for index,item in enumerate(Clusters): do
15:   Averaging the distance vectors for the point in Clusters[index]
16: end for
17: Calculating the point to point distance across Center_list
18: Template = Remove_outlier(Center_list, Distance_list, 95%)
19: Generating a heatmap based on the distance vectors of the template

```

Table 1. The result of t-test and pearson correlation analysis for fairness assessment with multivariate analysis

Models	Attribute	Liver	Kidney	Spleen	Lung	Aorta
nnU-Net	Gender	0.091	0.028	0.338	0.002	0.734
	Age	0.483	0.448	0.569	0.992	0.032
	BMI	0.939	<0.0001	0.424	0.008	0.325
SAM	Gender	<0.0001	<0.0001	0.008	<0.0001	<0.0001
	Age	<0.0001	<0.0001	0.129	<0.0001	<0.0001
	BMI	<0.0001	<0.0001	0.067	<0.0001	0.391
Medical SAM	Gender	0.002	0.013	<0.0001	0.004	<0.0001
	Age	<0.0001	0.017	0.166	<0.0001	0.537
	BMI	<0.0001	<0.0001	0.113	<0.0001	0.008
SAT	Gender	0.635	<0.0001	<0.0001	0.175	0.239
	Age	0.124	0.035	0.149	0.137	<0.0001
	BMI	0.0003	0.673	<0.0001	0.037	0.489

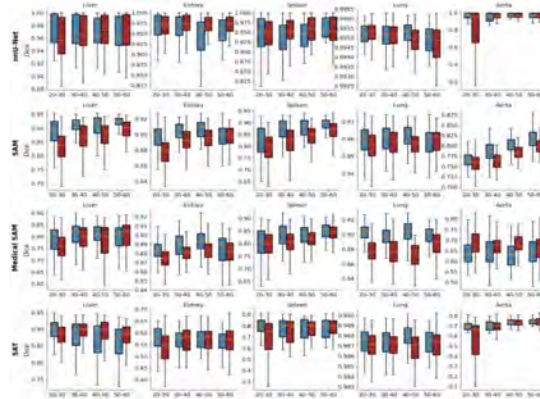


Fig. 1. Segmentation performance (means and standard deviations) for subject groups specified by combinations of gender and age attributes. Blue box: male. Red box: female.

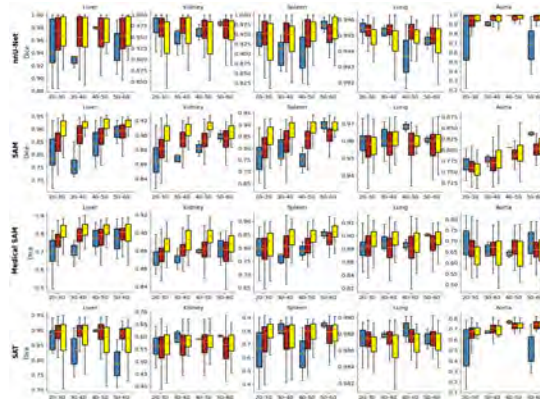


Fig. 2. Segmentation performance (means and standard deviations) for subject groups specified by combinations of age and BMI attributes. Blue box: underweight. Red box: healthy. Yellow box: overweight.