Innovation/Impact: A major challenge with clinically deploying deep learning models is their poor generalization. This problem is exacerbated in the medical field due to limited data availability, especially of rare cases. For instance, a deep learning-based liver segmentation model performed well on unseen test data of normal livers. However, the same model completely failed on images where novel information was presented (Figure 1). We intend to mitigate the potential consequences of poor generalization by detecting when out-of-distribution (OOD) information is presented to the network. The main contribution of our research is the interpretable detection of OOD images on which a trained segmentation model is likely to fail. Secondarily, we show that the Wasserstein distance (WD) outperforms the mean squared error (MSE) as a reconstruction metric in OOD detection in the medical domain.

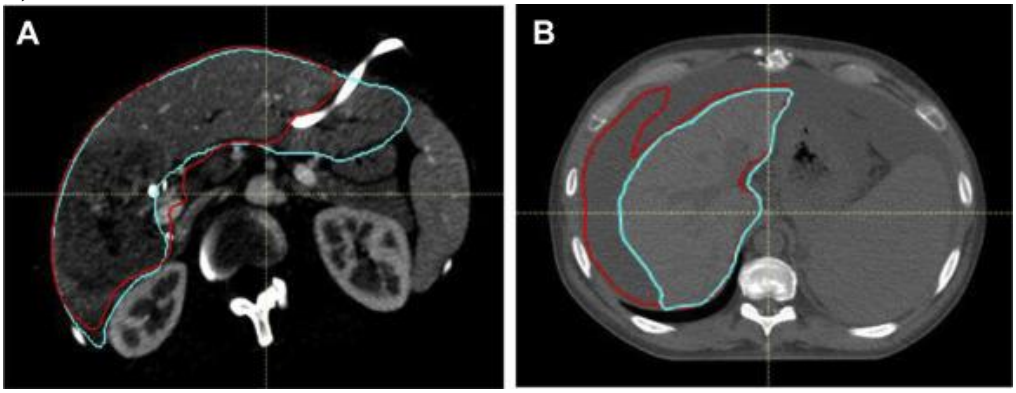

Figure 1: Failures of the liver segmentation model due to OOD information (A: stent, B: ascites). Blue is manual segmentation, red is the autosegmentation.

Key Results: The generative adversarial network (GAN)-based OOD detection paradigm detected non-liver images with high areas under the receiver operating characteristic curve (AUCs) (Table 1). The GAN modeled non-liver images by manipulating abdominal features (Figure 2), which resulted in large reconstruction errors. The liver images that were classified as OOD often contained underrepresented artifacts. For example, in the worst reconstruction of an in-distribution liver image in Figure 2, the network failed to model adjacent organs that had a high amount of contrast. Our paradigm was also able to detect liver images with abnormalities (Table 1). The GAN was completely unable to reconstruct needles and ascites (Figure 3). In Figure 2, by looking at the patch with the highest WD, we can successfully locate the abnormality. The WD-based AUC consistently outperformed the MSE-based AUC ($p < 0.01$, one-sided permutation test).

	AUC (±SD)	
Dataset	**WD-based**	**MSE-based**
Brain	**.98** (±.01)	.84 (±.04)
Cervix	**.88** (±.02)	.68 (±.02)
Head & Neck	**.96** (±.01)	.88 (±.02)
Lung	**.94** (±.02)	.86 (±.02)
Needles	**.73** (±.06)	.52 (±.02)
Ascites	**.66** (±.03)	.56 (±.02)

Table 1: The average AUCs, n=5, for each OOD dataset and reconstruction metric.

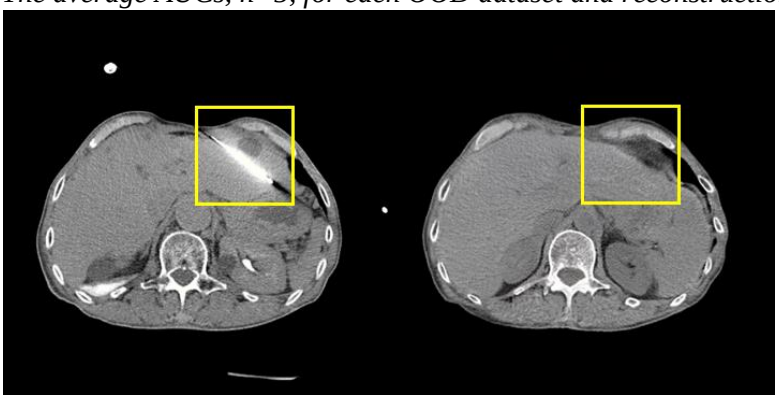

Figure 2: On the left is a real image with a needle. On the right is the reconstruction. The yellow box is the image patch with the highest WD.

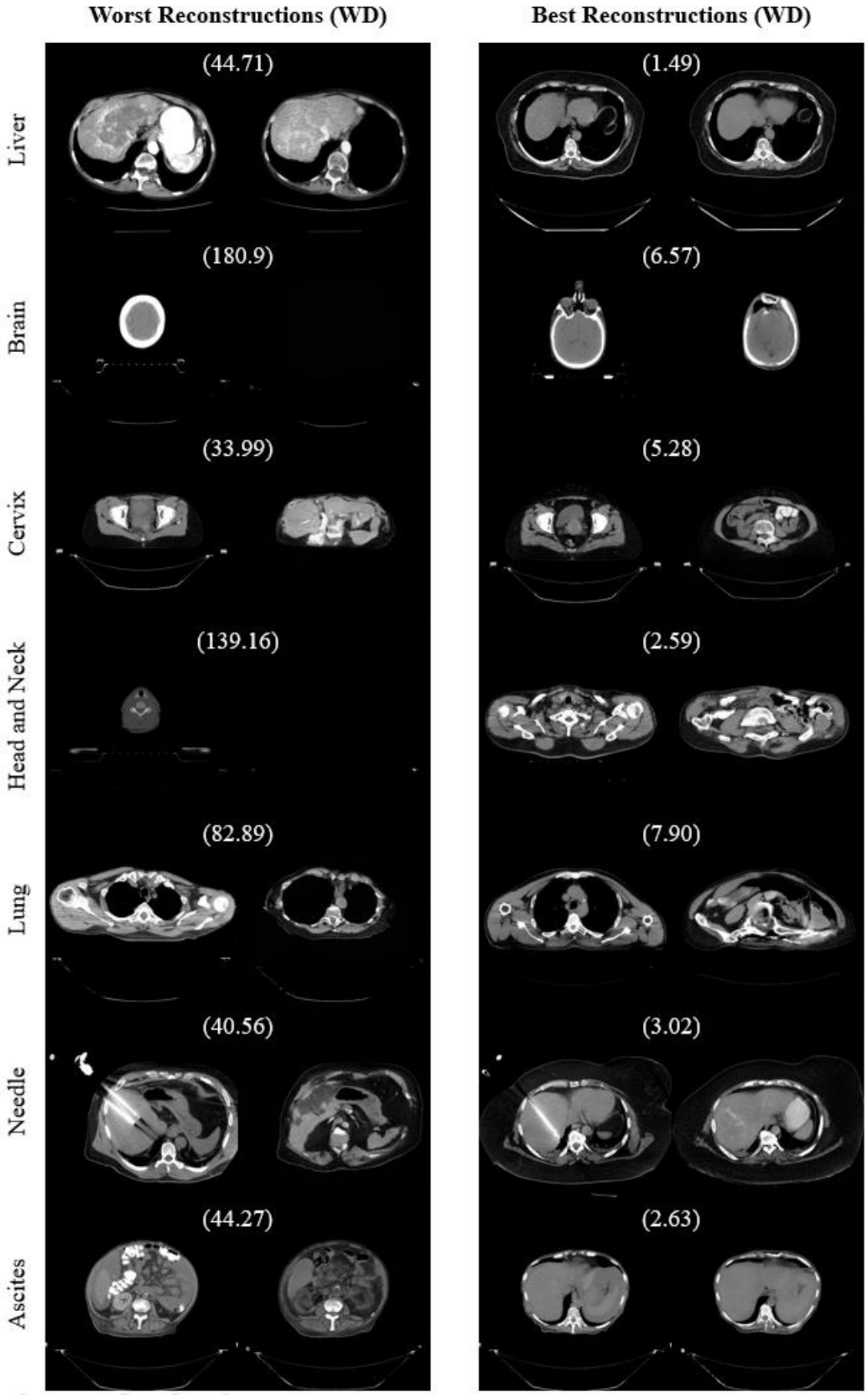

Figure 3: The best and worst reconstructions for each dataset (according to WD). For each pair, the left image is the original and the right image is the reconstruction. The number in parentheses is the WD between the pair.