Purpose: Images that lie outside a model’s training distribution pose a significant barrier to the widespread clinical deployment of deep learning-based segmentation models. We aim for real-time, interpretable identification of these out-of-distribution (OOD) images during the clinical implementation process.

Methods: Our in-distribution dataset included 3,234 liver-containing computed tomography (CT) scans from 456 patients. Five-fold cross-validation was performed with random 80/20 train/test splits. Our OOD test data consisted of brain, head and neck, lung, cervix, and abnormal liver CTs. A StyleGAN2-ADA architecture was employed to model the training distribution. Images were reconstructed using backpropagation. Reconstructions were evaluated using the Wasserstein distance and mean squared error. OOD detection was evaluated with the area under the receiver operating characteristic curve (AUC). By comparing original images to their in-distribution reconstructions, our method is not only unsupervised, but also interpretable.

Results: As hypothesized, the generative adversarial network (GAN) was only successful in reconstructing in-distribution liver images. As a result, non-liver images were detected with a mean 0.94 AUC. Interestingly, the GAN attempted to model non-liver images by manipulating abdominal features. Additionally, the GAN was unable to reconstruct liver abnormalities, such as needles and ascites, which have historically caused segmentation models to fail. As such, it was able to detect images a trained segmentation model would likely fail on with a mean 0.70 AUC. Interpretability was achieved by visualizing the patch in the reconstructed image with the largest reconstruction error.

Conclusion: The proposed OOD detection paradigm can distinguish between liver and non-liver CTs. Additionally, the paradigm provides an interpretable way to detect abnormal aspects of liver CTs. It can warn clinicians when a deep learning-based model is likely to fail. Other potential applications include prioritizing images for labeling when developing datasets and flagging images for review when considerable amounts of data are segmented in retrospective studies.