

## Appendix S1. Medical equipment data in final segmentation training

Using the 4,593 chest radiographs (CRs) of general patients, we trained a nnU-Net model with default settings for lesion segmentation. The training was conducted for 1,000 epochs using the Pytorch framework on an NVIDIA TITAN RTX 24GB GPU. The 250 images from the additional CR dataset, containing only labeled medical devices from patients in the intensive care unit (ICU), were then processed using the trained segmentation model. This inference generated 250 CR images with both pulmonary lesions and medical devices altogether.

We did not evaluate or optimize the lesion segmentation performance specifically for this 250 CR dataset, as it represents a relatively small proportion of the entire training dataset in terms of lesion data. Its primary purpose is to provide the final model with explicit information on medical devices rather than lesion-related information.

By including this dataset in our final training dataset, we aimed to enable the model to semantically learn to better differentiate medical equipment from lesions in ICU and emergency department (ED) datasets, thereby reducing the misclassification of medical devices as lesions, resulting in better segmentation performance in general. As shown in Table 1, this additional dataset contributed to improving the performances of our final segmentation model by addressing this challenge effectively.

## Appendix S2. Details of the segmentation model

### Input preprocessing

The CRs in the training/tuning/temporal validation datasets were all preprocessed with the following steps: Firstly, intensi-

ty-based clipping was implemented, where the top and bottom 0.5% of pixel intensities were clipped. This was done to mitigate the influence of high-intensity outliers, such as L&R markers or other unexpected artifacts on the CR. Secondly, min-max normalization was conducted to scale pixel values of the images within the range of 0–1. Both steps were done to ensure consistency and facilitated convergence during the segmentation model training explained after. Other preprocessing and augmentations were conducted in accordance with the nnU-Net methodology, accounting for median shape, distribution of spacings, intensity distribution, and image modality within the training dataset.

### Model structure and training details

The model employed in this study is a modified nnU-Net with two auxiliary classifiers incorporated between the encoder and segmentation decoder. The auxiliary classifiers used here are quite simple: two fully connected layers with a ReLU activation in between. These auxiliary classifiers are intended to enable the shared encoder to progressively focus on features related to lesion type and presence during its training, thereby passing more pertinent information to the segmentation decoder. The total loss was calculated as the sum of the original nnU-Net segmentation loss and the losses from the auxiliary classifiers, lesion presence and lesion type classifier, as follows:

$$L_{\text{total}} = L_{\text{seg}} + (L_{\text{lesion\_presence}} + L_{\text{lesion\_type}}) \quad (1)$$

Severe augmentations, including Gaussian noise and Gaussian blur transformations, were adopted to enhance segmentation performance on the noisier ICU/ED CRs.

During inference, we excluded regions with fewer than 50 pixels in the predicted mask to eliminate insignificant noise, as most mask sizes exceed  $2000 \times 2000$  pixels,

with some surpassing 3000 pixels on one side. This threshold was chosen based on the observation that smaller regions often represent false positives or artifacts rather than meaningful predictions.

## Appendix S3. Optimal threshold determination for assessment decision-making

To select the optimal threshold, we employed Youden's J statistic, which measures the effectiveness of a threshold in terms of maximizing the true positive rate while minimizing the false positive rate. For each threshold  $t_i$ , where  $t_i$  increases by 0.01 from 0.00 to 1.00, the true positive ratio (TPR) and false positive ratio (FPR) can be computed, and the value of J at  $t_i$  is calculated as:

$$J(t_i) = TPR(t_i) - FPR(t_i) \quad (2)$$

The optimal threshold can be determined by identifying the highest value of J, representing the most balanced threshold between sensitivity and specificity:

$$t_{\text{optimal}} = \operatorname{argmax}_i J(t_i) \quad (3)$$

If multiple values of  $t_{\text{optimal}}$  exist, we selected the lowest threshold value to increase sensitivity and reduce the chance of false negatives:

$$t_{\text{selected}} = \min(t_{\text{optimal},i}), \text{ for } i = 1, 2, \dots, n \quad (4)$$

This approach minimizes the misclassification of true positives as negatives, which is crucial in settings such as the ICU/ED, where timely intervention is essential. We have inferred the tuning set using the trained segmentation model and then calculated the optimal threshold accordingly.