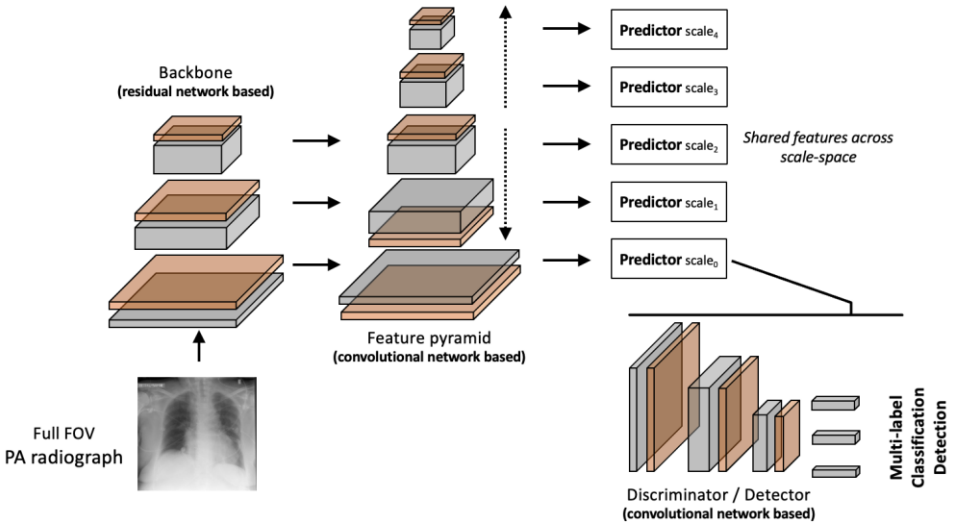
Appendix A: Algorithm Details

Note: same set of pre-processing steps and network architecture were used for both Systems A and B.

Pre-processing: Using bilinear interpolation for resampling, input images are rescaled to an isotropic resolution of 1025×1025 pixels with letterboxing. Pixel values are linearly remapped to achieve a robust brightness and contrast normalization. Given an arbitrary chest radiograph, i.e., image array I, we denote its pixel value histogram function as $h(x; I)$ with a bandwidth of 256 discrete values. Gaussian smoothing and median filtering are applied to h to significantly reduce noise (caused, e.g., by white text overlay) and account for long function tails that affect the brightness of the image at given intensity windows. Based on the processed function h, we determine two bounds: b_{low} and b_{high}; which represent a tight intensity window for image I. The value b_{low} indicates the lowest bin and b_{high} the highest bin along the image intensity histogram. We apply the following normalization: $I = (I - b_{low})/(b_{high} - b_{low})$. In addition, for training we apply the following augmentations: left/right image flip, random cropping and scaling, random rotations and (inverse) gamma transforms. The entire image is provided as input to the network and is processed in a single-shot inference.

Architectural Design: The architecture is displayed in and comprises an initial feature extractor acting as candidate generator into an abstract feature space, followed by a multi-scale discriminator sub-network used to compute probabilities on whether the abnormalities are present or not (in an image sub-region of interest). The architecture is fully convolutional and processes the entire image content in one single inference iteration, while analyzing its content on multiple levels of scales. As such, the architecture is capable of implicitly capturing both global as well as local comorbidities present in the image. The hyperparameters are optimized based on recommendations from the literature and using a fixed grid search with the goal to optimize performance. The architecture is inspired from [15].

Training is conducted in an end-to-end stage and in a multi-class setting [16] [17]. The loss function is based on summation of three elements: 1) a classification loss based on the focal loss described in detail in [15]; 2) a bounding box coordinate regression loss based on an intersection-over-union based metric; and 3) a center-ness loss designed to reduce outlier detections which is based on a weighted binary cross entropy loss. A batch-size of 8 is used for training, and multiple validation sets are used to track the system performance during training and perform early stopping.


Backbone
(residual network based)
Predictor scale4
Predictor scale3
Predictor scale2
Predictor scale1
Predictor scale0
Shared features across scale-space
Feature pyramid
(convolutional network based)
Full FOV
PA radiograph
Multi-label Classification Detection
Discriminator / Detector
(convolutional network based)