Miyato, T., Kataoka, T., Koyama, M., Yoshida, Y., 2018. Spectral normalization for generative adversarial networks. International Conference on Learning Representations (ICLR) .

Miyato, T., Koyama, M., 2018. cGANs with Projection Discriminator. International Conference on Learning Representations (ICLR) .

Mothilal, R.K., Sharma, A., Tan, C., 2020. Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations. Conference on Fairness, Accountability, and Transparency (FAT) , 607–617.

Narayanaswamy, A., Venugopalan, S., Webster, D.R., Peng, L., Corrado, G.S., Ruamviboonsuk, P., Bavishi, P., Brenner, M., Nelson, P.C., Varadarajan, A.V., 2020. Scientific Discovery by Generating Counterfactuals using Image Translation. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI) , 273–283.

Oakden-Rayner, L., Dunnmon, J., Carneiro, G., Ré, C., 2020. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. ACM Conference on Health, Inference, and Learning 2020, 151–159. doi:10.1145/3368555.3384468.

Parafita Martinez, A., Vitria Marca, J., 2019. Explaining visual models by causal attribution. IEEE International Conference on Computer Vision Workshop (ICCVW) , 4167–4175doi:10.1109/ICCVW.2019.00512.

Pasa, F., Golkov, V., Pfeiffer, F., Cremers, D., Pfeiffer, D., 2019. Efficient Deep Network Architectures for Fast Chest X-Ray Tuberculosis Screening and Visualization. Scientific Reports 9, 1–9.

Petsiuk, V., Das, A., Saenko, K., 2018. Rise: Randomized input sampling for explanation of black-box models. British Machine Vision Conference (BMVC) .

Rajpurkar, P., Irvin, J., Ball, R.L., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., et al., 2018. Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. PLOS Medicine 15, 1–17.

Rajpurkar, P., Irvin, J.A., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., Lungren, M., Ng, A., 2017. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. arXiv e-prints , arXiv:1711.05225.

Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. International Conference on Advances in Neural Information Processing Systems (NeurIPS) 28.

Rodriguez-Ruiz, A., Lång, K., Gubern-Mérida, A., Broeders, M., Gennaro, G., Clauser, P., Helbich, T., Chevalier, M., Tan, T., Mertelmeier, T., Wallis, M., Andersson, I., Zackrisson, S., Mann, R., Sechopoulos, I., 2019. Stand-alone artificial intelligence for breast cancer detection in mammography: Comparison with 101 radiologists. Journal of the National Cancer Institute 111. doi:10.1093/jnci/djy222.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net convolutional networks for biomedical image segmentation. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI) 9351, 234–241.

Rubin, J., Sanghavi, D., Zhao, C., Lee, K., Qadir, A., Xu-Wilson, M., 2018. Large Scale Automated Reading of Frontal and Lateral Chest X-Rays using Dual Convolutional Neural Networks. arXiv e-prints , arXiv:1804.07839arXiv:1804.07839.

Samek, W., Binder, A., Montavon, G., Lapuschkin, S., Müller, K.R., 2017. Evaluating the visualization of what a deep neural network has learned. IEEE Transactions on Neural Networks and Learning Systems 28, 2660–2673. doi:10.1109/TNNLS.2016.2599820.

Sayres, R., Taly, A., Rahimy, E., Blumer, K., Coz, D., Hammel, N., Krause, J., Narayanaswamy, A., Rastegar, Z., Wu, D., Xu, S., Barb, S., Joseph, A., Shumski, M., Smith, J., Sood, A.B., Corrado, G.S., Peng, L., Webster, D.R., 2019. Using a Deep Learning Algorithm and Integrated Gradients Explanation to Assist Grading for Diabetic Retinopathy. Ophthalmology 126, 552–564.

Seah, J.C., Tang, C.H., Buchlak, Q.D., Holt, X.G., Wardman, J.B., Aimoldin, A., Esmaili, N., Ahmad, H., Pham, H., Lambert, J.F., et al., 2021. Effect of a comprehensive deep-learning model on the accuracy of chest x-ray interpretation by radiologists: a retrospective, multireader multicase study. The Lancet Digital Health 3, e496–e506.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. IEEE International Conference on Computer Vision (ICCV) , 618–626.

Shrikumar, A., Greenside, P., Kundaje, A., 2017. Learning important features through propagating activation differences. 34th International Conference on Machine Learning (ICML) 70, 3145–3153.

Simonyan, K., Vedaldi, A., Zisserman, A., 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. Computing Research Repository abs/1312.6034.

Simonyan, K., Zisserman, A., 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv e-prints , arXiv 1409.1556.

Singla, S., Pollack, B., Chen, J., Batmanghelich, K., 2019. Explanation by Progressive Exaggeration. International Conference on Learning Representations (ICLR) .

Singla, S., Wallace, S., Triantafillou, S., Batmanghelich, K., 2021. Using causal analysis for conceptual deep learning explanation. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI) , 519–528.

Springenberg, J.T., Dosovitskiy, A., Brox, T., Riedmiller, M.A., 2015. Striving for Simplicity: The All Convolutional Net. International Conference on Learning Representations (ICLR-workshop track) .

Sundararajan, M., Taly, A., Yan, Q., 2017. Axiomatic Attribution for Deep Networks. 34th International Conference on Machine Learning (ICML) 70, 3319–3328.

Tonekaboni, S., Joshi, S., McCradden, M.D., Goldenberg, A., 2019. What Clinicians Want: Contextualizing Explainable Machine Learning for Clinical End Use URL: http://arxiv.org/abs/1905.05134.

Van Looveren, A., Klaise, J., 2019. Interpretable Counterfactual Explanations Guided by Prototypes. arXiv e-prints , arXiv:1907.02584.

Wada, K., 2016. labelme Image Polygonal Annotation with Python. https://github.com/wkentaro/labelme.

Wang, F., Kaushal, R., Khullar, D., 2020. Should health care demand interpretable artificial intelligence or accept "black Box" Medicine? Annals of Internal Medicine 172, 59–61. doi:10.7326/M19-2548.

Wang, P., Vasconcelos, N., 2020. SCOUT: Self-Aware Discriminant Counterfactual Explanations. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) .

Winkler, J., Fink, C., Toberer, F., Enk, A., Deinlein, T., Hofmann-Wellenhof, R., Thomas, L., Lallas, A., Blum, A., Stolz, W., Haenssle, H., 2019. Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition. JAMA Dermatology 155. doi:10.1001/jamadermatol.2019.1735.

Young, K., Booth, G., Simpson, B., Dutton, R., Shrapnel, S., 2019. Deep neural network or dermatologist? International Workshop on Interpretability of Machine Intelligence in Medical Image Computing (IMIMIC) 11797 LNCS, 48–55.

Zhou, B., Khosla, A., Àgata Lapedriza, Oliva, A., Torralba, A., 2015. Object detectors emerge in deep scene cnns. International Conference on Learning Representations (ICLR) .

Zhu, J.Y., Park, T., Isola, P., Efros, A.A., 2017. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. IEEE International Conference on Computer Vision (ICCV) .

## 6. Supplementary Material

### *6.1. Human evaluation*

In our human evaluation study, we asked the following 15 questions for each CXR:

1. Please provide your diagnosis for Cardiomegaly. Answers: Negative, mild, positive, not sure.
2. (Only assessment) Do you agree with the AI system assessment for Cardiomegaly? Answers: yes, no
3. (Only assessment) I understand how the AI system made the above assessment for Cardiomegaly. Answers: 5-point Likert scale.
4. (Assessment + SM) The heat-map is highlighting <blank> important/relevant regions for Cardiomegaly. Answers: all, most, some, a few, none.
5. (Assessment + SM) I understand how the AI system made the above assessment for Cardiomegaly. Answers: 5-point Likert scale.

6. (Assessment + cycleGAN) The changes in the video are related to Cardiomegaly. Answers: 5-point Likert scale.
7. (Assessment + cycleGAN) I understand how the AI system made the above assessment for Cardiomegaly. Answers: 5-point Likert scale.
8. (Assessment + cycleGAN) Images in the video look like a chest x-ray. Answers: 5-point Likert scale.
9. (Assessment + cycleGAN) The images in the video look like the chest x-ray from the subject. Answers: 5-point Likert scale.
10. (Assessment + ours) The changes in the video are related to Cardiomegaly. Answers: 5-point Likert scale.
11. (Assessment + ours) Changes in the anatomy in the highlighted regions in the heat-map will change the assessment of Cardiomegaly. Answers: 5-point Likert scale.
12. (Assessment + ours) I understand how the AI system made the above assessment for Cardiomegaly. Answers: 5-point Likert scale.
13. (Assessment + ours) Images in the video look like a chest x-ray. Answers: 5-point Likert scale.
14. (Assessment + ours) The images in the video look like the chest x-ray from the subject. Answers: 5-point Likert scale.
15. Which explanation helped you the most in understanding the assessment made by the AI system Answers: Explanation-1: Heat-map highlighting important regions for assessment, Explanation-2: A video showing the transformation from negative to positive decision, Explanation-3: Two images at the two extreme ends of the decision (positive and negative), none.

Next, we present the UI for different questions,

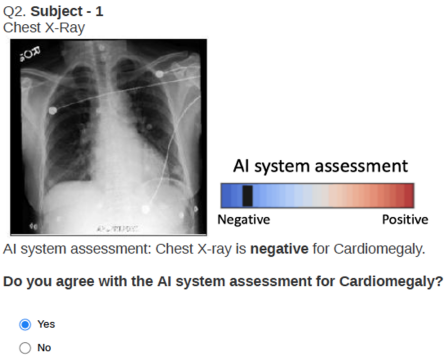


**Fig. 11. Question 2-3 showing the query CXR and the classifier's decision.**

### 6.2. *Summarizing the notation*

Table. 4 summarizes the notation used in the manuscript.

### 6.3. *Dataset*

We focus on explaining classification models based on deep convolution neural networks (CNN); most state-of-the-art performance models fall in this regime. We used large, publicly available datasets of chest x-ray (CXR) images, MIMIC-CXR (Johnson et al., 2019). MIMIC-CXR dataset is a multimodal dataset consisting of 473K CXR, and 206K reports from

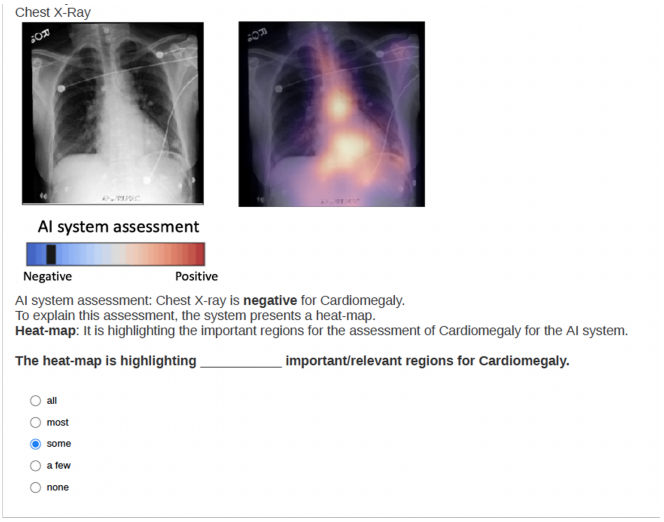


**Fig. 12. Question 4-5 showing the query CXR, the classifier's decision and the saliency map explanation.**

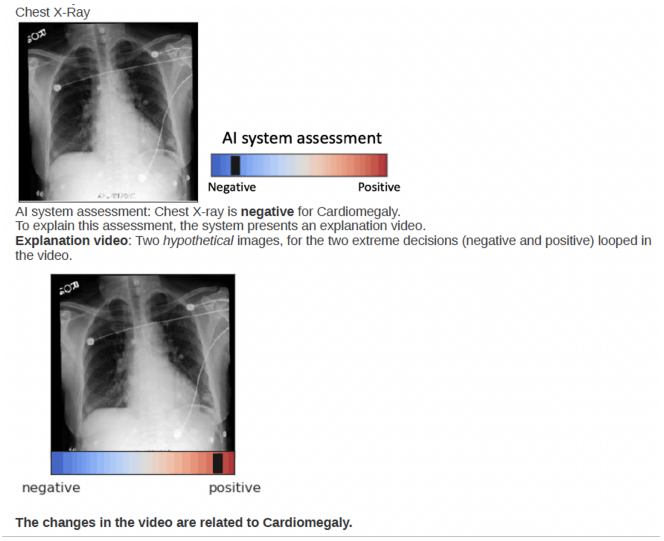


**Fig. 13. Question 6-9 showing the query CXR, the classifier's decision and the cycleGAN explanation.**

63K patients. We considered only frontal (posteroanterior PA or anteroposterior AP) view CXR. The datasets provide image-level labels for fourteen radio-graphic observations. These labels are extracted from the radiology reports associated with the x-ray exams using an automated tool called the Stanford CheXpert labeler (Irvin et al., 2019). The labeller first defines some thoracic observations using a radiology lexicon (Hansell et al., 2008). It extracts and classifies (positive, negative, or uncertain mentions) these observations by processing their context in the report. Finally, it aggregates these observations into fourteen labels for each x-ray exam. For the MIMIC-CXR dataset, we extracted the labels ourselves, as we have access to the reports.

### 6.4. *Classification Model*

To train the classifier, we considered the uncertain mention as a positive mention. We crop the original images to have the same height and width, then downsample them to 256 × 256

**Table 4. Summarizing the notation**

| Notation | Description |
|---|---|
| $\mathcal{X}$ | Input image space |
| $\mathbf{x} \in \mathcal{X}$ | Input image |
| $f : \mathcal{X} \rightarrow \mathcal{Y}$ | Pre-trained classification function |
| $f(\mathbf{x})[k] \in [0, 1]$ | Classifier's output for $k^{\text{th}}$ class |
| $\mathbf{c}$ | The condition used in cGAN, the desired classifier's output for $k^{\text{th}}$ class |
| $\mathbf{x_c}$ | Explanation image |
| $f(\mathbf{x_c})$ | Classifier's output for the explanation image |
| $\mathcal{I}_f(\mathbf{x}, \mathbf{c})$ | Explanation function |
| $E(\cdot)$ | Image encoder |
| $\mathbf{z}$ | Latent representation of the input image |
| $C(\mathbf{c})$ | Discretizing function that maps $\mathbf{c}$ to an integer |
| $G(\mathbf{z}, \mathbf{c})$ | Generator of cGAN |
| $D(\mathbf{x}, \mathbf{c})$ | Discriminator of cGAN |
| $p_{\text{data}}(\mathbf{x})$ | Real image data distribution |
| $q(\mathbf{x})$ | Learned data distribution by cGAN |
| $r(\mathbf{x})$ | Loss term of cGAN that measures similarity between real and learned data distribution |
| $r(\mathbf{c}\|\mathbf{x})$ | Loss term of cGAN that evaluates correspondence between generated images and condition |
| $\phi(\mathbf{x})$ | Image feature extractor; part of the discriminator function |

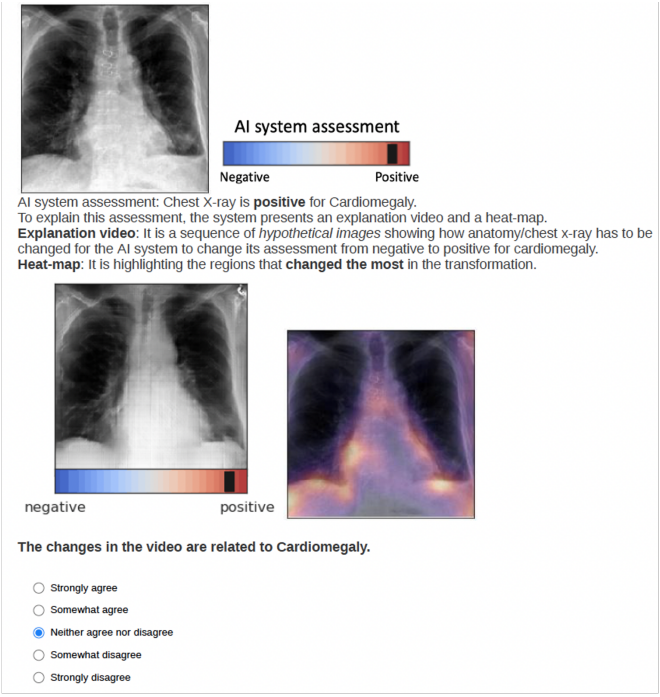


**Fig. 14. Question 10-14 showing the query CXR, the classifier's decision and our counterfactual explanation.**

pixels. The intensities were normalized to have values between 0 and 1. Following the approach in prior work (Rajpurkar et al., 2017; Rubin et al., 2018; Irvin et al., 2019) on diagnosis classification, we used DenseNet-121 (Huang et al., 2016) architecture as the classification model. In DenseNet, each layer implements a non-linear transformation based on composite functions such as Batch Normalization (BN), rectified linear unit (ReLU), pooling, or convolution. The resulting feature map at each layer is used as input for all the subsequent layers, leading to a highly convoluted multi-level multi-layer non-linear convolutional neural network. We aim to explain such a model post-hoc without accessing the parameters learned by any layer or knowing the architectural details. Our proposed approach can be used for explaining any DL based neural network.

### *6.5. Explanation Function*

The explanation function is a conditional GAN with an encoder. We used a ResNet (He et al., 2016) architecture for the Encoder, Generator, and Discriminator. The details of the architecture are given in Table 5. For the encoder network, we used five ResBlocks with the standard batch normalization layer (BN). In encoder-ResBlock, we performed down-sampling (average pool) before the first *conv* of the ResBlock as shown in Fig. 16.a. For the generator network, we follow the details in (Miyato et al., 2018) and replace the BN layer in encoder-ResBlock with conditional BN (cBN) to encode the condition (*see* Fig. 16.b.). The architecture for the generator has five ResBlocks; each ResBlock performed up-sampling through the nearest neighbour interpolator. For the discriminator, we used spectral normalization (SN) (Miyato and Koyama, 2018) in Discriminator-ResBlock and performed down-sampling after the second *conv* of the ResBlock as shown in Fig. 16.c. For the optimization, we used Adam optimizer (Kingma and Ba, 2015), with hyper-parameters set to $\alpha = 0.0002, \beta_1 = 0, \beta_2 = 0.9$ and updated the discriminator five times per one update of the generator and encoder.

For creating the training dataset, we divide the posterior distribution for the target class, $f(\mathbf{x}) \in [0, 1]$ into $N$ equally-sized bins. The cGAN is then trained on $N$ conditions. For efficient training, cBN requires class-balanced batches. A smaller value for $\delta$ results in more conditions for training cGAN, increasing cGAN complexity and training time. Also, we have to increase the batch size to ensure each condition is well represented in a batch. Hence, the GPU memory size bounds the high value for $N$. A small $N$ is equivalent to fewer conditions, resulting in a coarse transformation which leads to abrupt changes across explanation images. In our experiments, we used $N = 10$, with a batch size of 32. We experimented with different values of $N$ and selected the largest $N$, which created a class-balanced batch that fits in GPU memory and resulted in stable cGAN training.

**Table 5. Explanation Model (cGAN) Architecture**

(a) Encoder

| |
|---|
| Grayscale image $\mathbf{x} \in \mathbb{R}^{256\times256\times1}$ |
| BN, ReLU, 3×3 conv 64 |
| Encoder-ResBlock down 128 |
| Encoder-ResBlock down 256 |
| Encoder-ResBlock down 512 |
| Encoder-ResBlock down 1024 |
| Encoder-ResBlock down 1024 |

(b) Generator

| |
|---|
| Latent code $\mathbf{z} \in \mathbb{R}^{1024}$ |
| Generator-ResBlock up 1024, **y** |
| Generator-ResBlock up 512, **y** |
| Generator-ResBlock up 256, **y** |
| Generator-ResBlock up 128, **y** |
| Generator-ResBlock up 64, **y** |
| BN, ReLU, 3×3 conv 1 |
| Tanh |

(c) Discriminator

| |
|---|
| Grayscale image $\mathbf{x} \in \mathbb{R}^{256\times256\times1}$ |
| Discriminator-ResBlock down 64 |
| Discriminator-ResBlock down 128 |
| Discriminator-ResBlock down 256 |
| Discriminator-ResBlock down 512 |
| Discriminator-ResBlock down 1024 |
| Discriminator-ResBlock 1024 |
| ReLU, Global Sum Pooling (GSP) \| Embed(**y**) |
| Inner Product (GSP, Embed(**y**)) $\rightarrow \mathbb{R}^1$ |
| Add(SN-Dense(GSP) $\rightarrow \mathbb{R}^1$, Inner Product) |

### *6.6. Semantic Segmentation*

We adopted a 2D U-Net (Ronneberger et al., 2015) to perform semantic segmentation, to mark the lung and the heart contour in a CXR. The network optimizes a multi-categorical cross-entropy loss function, defined as,

$$\mathcal{L}_\theta := \sum_s \sum_i \mathbb{1}(y_i = s) \log(p_\theta(x_i)), \tag{13}$$

where $\mathbb{1}$ is the indicator function, $y_i$ is the ground truth label for i-th pixel. $s$ is the segmentation label with values (background, the lung or the heart). $p_\theta(x_i)$ denotes the output probability for pixel $x_i$ and $\theta$ are the learned parameters. The network is trained on 385 CXRs and corresponding masks from Japanese Society of Radiological Technology (JSRT) (van Ginneken et al., 2006) and Montgomery (Jaeger et al., 2014) datasets.

### *6.7. Object Detection*

We trained an object detector network to identify medical devices in a CXR. For the MIMIC-CXR dataset, we pre-processed the reports to extract keywords/observations that correspond to medical devices, including pacemakers, screws, and other hardware. Such foreign objects are easy to identify in a CXR and do not requires expert knowledge for manual labelling. Using the CheXpert labeller, we extracted 300 CXR images with positive mentions for each observation. The extracted x-rays are then manually annotated with bounding box annotations marking the presence of foreign objects using the LabelMe (Wada, 2016) annotation tool. Next, we trained an object detector based on fFast Region-based CNN (Ren et al., 2015), which used VGG-16 model (Simonyan and Zisserman, 2014), trained on the MIMIC-CXR dataset as its foundation. We used this object detector to enforce our novel context-aware reconstruction loss (CARL).

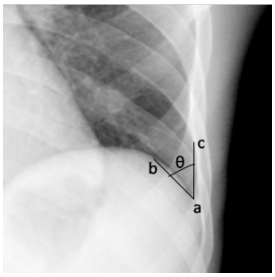


**Fig. 15. The costophrenic angle (CPA) on a CXR is marked as the angle formed by, (a) costophrenic angle point, (b) hemidiaphragm point and (c) lateral chest wall point, as shown by Maduskar *et al.*in (Maduskar et al., 2016)**

We trained similar detectors for identifying normal and abnormal CP recess regions in a CXR. We associated an abnormal CP recess with the radiological finding of a blunt CP angle as identified by the positive mention for *"blunting of costophrenic angle"* in the corresponding radiology report. For the normal-CP recess, we considered images with a positive mention for *"lungs are clear"* in the reports. We extracted 300 CXR images with positive mention of respective terms for normal and abnormal CP recess to train the object detector.

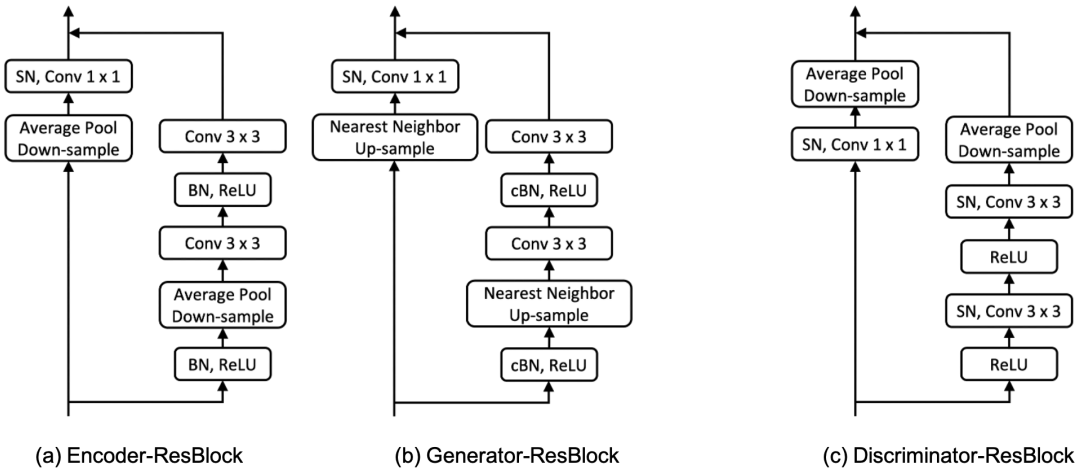


**Fig. 16. Architecture of the ResBlocks used in all experiments.**

Please note that the object detector for CP recess is only used for evaluation purposes, and they were not used during the training of the explanation function. In literature, the blunting of CPA is an indication of pleural effusion (Maduskar et al., 2013, 2016). The angle between the chest wall and the diaphragm arc is called the costophrenic angle (CPA). Marking the CPA angle on a CXR requires an expert to mark the three points, (a) costophrenic angle point, (b) hemidiaphragm point and (c) lateral chest wall point and then calculate the angle as shown in Fig. 15. Learning automatic marking of CPA angle requires expert annotation and is prone to error. Hence, rather than marking the CPA angle, we annotate the CP region with a bounding box which is a much simpler task. We then learned an object detector to identify normal or abnormal CP recess in a CXR and used the Score for detecting a normal CP recess (SCP) as our evaluation metric.

### *6.8. xGEM*

We refer to work by Joshi *et al.* (Joshi et al., 2019) for the implementation of xGEM. xGEM iteratively traverses the input image's latent space and optimizes the traversal to flip the classifier's decision to a different class. Specifically, it solves the following optimization

$$\tilde{\mathbf{x}} = \mathcal{G}_\theta(\arg\min_{\mathbf{z}\in\mathbb{R}^d} \mathcal{L}(\mathbf{x}, \mathcal{G}_\theta(\mathbf{z})) + \lambda\ell(f(\mathcal{G}_\theta(\mathbf{z})), y')) \qquad (14)$$

where the first terms is an $\ell_2$ distance loss for comparing real and generated data. The second term ensures that the classification decision for the generated sample is in favour of class $y'$ and $y' \neq y$ is a class other than original decision. Unless explicitly imposed, the explanation image does not look realistic. The explanation image is generated from an updated latent feature, and the expressiveness of the generator limits its visual quality. xGEM adopted a variational autoencoder (VAE) as the generator. VAE uses a Gaussian likelihood ($\ell_2$ reconstruction), an unrealistic assumption for image data. Hence, vanilla VAE is known to produce over-smoothed images (Huang et al., 2018). The VAE used is available at https://github.com/LynnHo/VAE-Tensorflow. All settings and architectures were set to default values. The original code generates an image of dimension 64x64. We extended the given network to produce an image with dimensions 256×256.

### *6.9. cycleGAN*

We refer to the work by Narayanaswamy *et al.* (Narayanaswamy et al., 2020) and DeGrave *et al.* (DeGrave et al., 2020) for the implementation details of cycleGAN. The network architecture for cycleGAN is replicated from the GitHub repository https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix.

For training cycleGAN, we consider two sets of images. The first set comprises 2000 images from the MIMIC-CXR dataset such that the classifier has a positive prediction for the presence of a target disease *i.e.,* $f(\mathbf{x}) > 0.9$, and the second set has the same number of images but with strong negative prediction *i.e.,* $f(\mathbf{x}) < 0.1$. We train one such model for each target disease.

### *6.10. Extended results for identity preservation*

A FO is critical in identifying the patient in an x-ray. FO's disappearance may lead to a false conclusion that removing FO resulted in the changed classification decision.

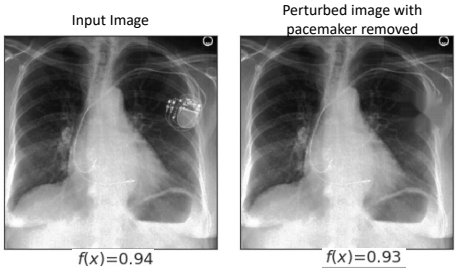


**Fig. 17. An example of input image before and after removing the pacemaker.**

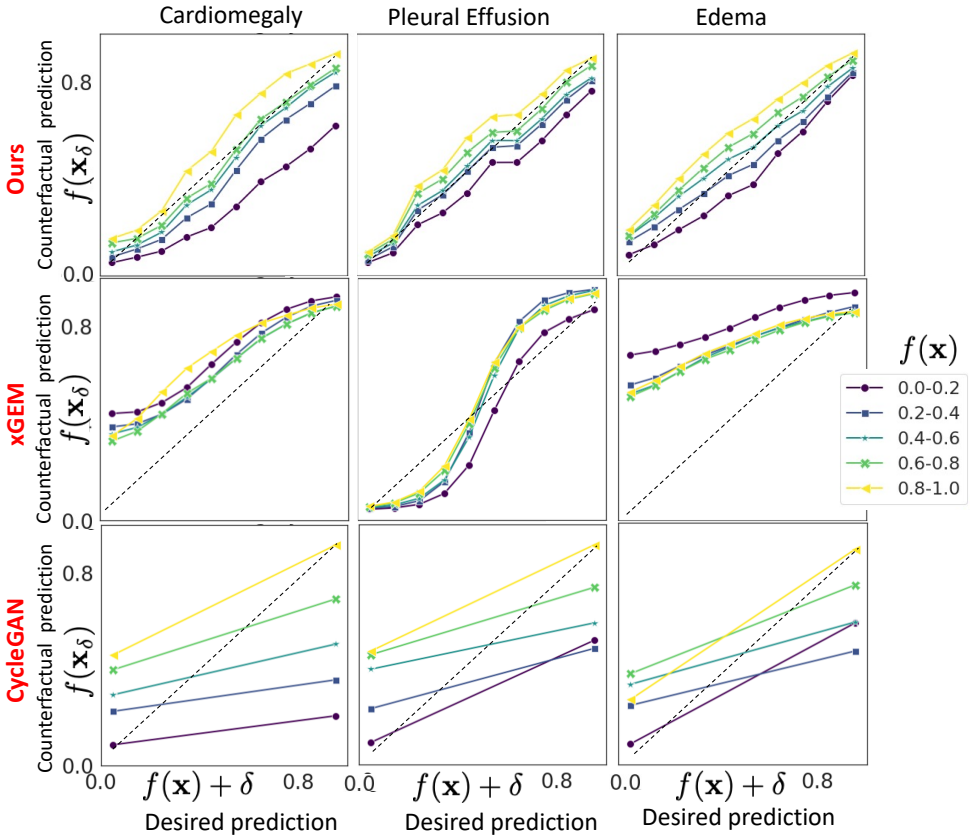


**Fig. 18. The plot of desired outcome, $f(\mathbf{x}) + \delta$, against actual response of the classifier on generated explanations, $f(\mathbf{x}_\delta)$. Each line represents a set of input images with classification prediction $f(\mathbf{x})$ in a given range. Dashed line represents $y = x$ line.**

### *6.10.1. Ablation study over pacemaker*

We performed an ablation study to investigate if a pacemaker is influencing the classifier's prediction for cardiomegaly. We consider 300 subjects that are positively predicted for cardiomegaly and have a pacemaker. We used our pre-trained object detector to find the bounding-box annotations for these images. Using the bounding-box, we created a perturbation of the input image by masking the pacemaker and in-filling the masked region with the surrounding context. An example of the perturbation image is shown in Fig. 17. We passed the perturbed image through the classifier and calculated the difference in the classifier's prediction before and after removing the pacemaker. The average change in prediction was negligible (0.03). Hence, pacemaker is not influencing classification decisions for cardiomegaly.

**Table 6. The latent-space closeness (LSC) score for our model with and without the context-aware reconstruction loss (CARL).**

| Foreign Object | **LSC score** CARL better than $\ell_1$ |
|---|---|
| Pacemaker | 0.79 |
| Hardware | 0.87 |

### *6.10.2. Latent space closeness (LCS)*

We compared the explanations generated using CARL against those generated using simple $\ell_1$ reconstruction loss on their similarity with the input images. To quantify the similarity between the explanation images and the query image in a latent space, we used latent-space closeness (LSC) score. LSC score is the fraction of the images where explanation image derived using CARL ($\mathbf{x}_\mathbf{c}^{\text{CARL}}$) is closest to the query image $\mathbf{x}$ as compared to explanations generated using $\ell_1$ loss *i.e.,* $\mathbf{x}_\mathbf{c}^{\ell_1}$. We calculated similarity as the euclidean distance between the embedding for the query and explanation images. LSC score is defined as,

$$LSC = \sum_{\mathbf{x}\in\mathcal{X},\mathbf{c}} \mathbb{1}\left(\langle E(\mathbf{x}), E(\mathbf{x}_\mathbf{c}^{\text{CARL}})\rangle < \langle E(\mathbf{x}), E(\mathbf{x}_\mathbf{c}^{\ell_1})\rangle\right)$$

where $E(\cdot)$ is a pre-trained feature extractor based on the Inception v3 network. Table 6 presents our results. A high LSC score, together with a high CV score (Fig. 19) shows that the query and counterfactual images are fundamentally same but differs only in features that are sufficient to flip the classification decision.

### *6.11. Extended classifier consistency results*

Our explanation framework gradually perturbs the input image to traverse the classification boundary from one extreme (negative) to another (positive). We quantify the consistency between our explanations and the classification model at every step of this transformation. We divided the prediction range $[0, 1]$ into ten equally sized bins. For each bin, we generated an explanation image by choosing an appropriate, $\mathbf{c} \in [0, 1]$. We

further divided the input image space into five groups based on their initial prediction *i.e.,* $f(\mathbf{x})$. In Fig. 18, we represented each group as a line and plotted the average response of the classifier *i.e.,* $f(\mathbf{x_c})$ for explanations in each bin against the expected outcome *i.e.,* $\mathbf{c}$. For xGEM, we generated multiple, progressively changing explanations by traversing the latent space. For each input image, we generated ten explanation images. For cycleGAN, we can generate only images at the two extreme ends of the decision boundary.

Fig. 18 shows our results. It an extension of the results in Fig. 5. The positive slope of the line-plot, parallel to $y = x$ line confirms that starting from images with low $f(\mathbf{x})$, our model creates fake images such that $f(\mathbf{x_c})$ is high and vice-versa. Thus, our model creates explanations that successfully flips the classification decision and, hence, represents the decision-making process of the classifier. In contrast, for cycleGAN model, if $f(\mathbf{x}) \in [0.0, 0.4]$ (blue line-plot), the resulting explanations have $f(\mathbf{x_c}) < 0.5$, hence, cycleGAN model fails to flip the classification decision, as also evident in low CV score in Table. 1.

### *6.12. Evaluating class discrimination*

In multi-label settings, multiple labels can be true for a given image. A multi-label setting is common in CXR diagnosis. For example, cardiomegaly and pleural effusion are associated with cardiogenic edema and frequently co-occur in a CXR. Please note that our classification model is also trained in a multi-label setting where the fourteen radiological findings may co-occur in a CXR. In this evaluation, we demonstrate the sensitivity of our generated explanations to the task being explained. We considered three diagnosis tasks, cardiomegaly, pleural effusion, and edema. For each task, we trained one explanation model. Ideally, an explanation model trained to explain a given task should produce explanations consistent with the query image on all the other classes besides the given task. Fig. 19 plots the fraction of the generated explanations, that have flipped in other classes as compared to the query image. Ideally, the fraction should be maximum for the given task and small for the rest of the classes. In Fig. 19, each column represents one task, and each row is one run of our method to explain a given task. The diagonal values also represent the counterfactual validity (CV) score reported in Table. 1.

### *6.13. Ablation Study*

Eq. 1 shows our final loss function: We have three types of loss functions: adversarial loss from cGAN $\mathcal{L}_{\text{cGAN}}(D, G)$, KL loss $\mathcal{L}_f(D, G)$, and CARL reconstruction loss $\mathcal{L}_{\text{identity}}(E, G)$. The three losses enforce the three properties of our proposed explanation function: data consistency, classifier consistency, and context-aware self-consistency, respectively. In the ablation study, we quantify the importance of each of these components by training different models, where on hyper-parameter is set to zero while rest are equivalent ($\lambda_1 = 1$, $\lambda_2 = 1$ and $\lambda_3 = 0.5$). For **data consistency**, we evaluate Fréchet Inception Distance (FID). FID score measures the visual quality of the generated explanations by comparing them with the real images. We examined real and synthetic (*i.e.,* generated explanations) images on the two extreme of the decision boundary, *i.e.,* a normal group ($f(\mathbf{x}) < 0.2$) and an abnormal group ($f(\mathbf{x}) > 0.8$). For **classifier consistency**, we reported results on counterfactual validity (CV) score. CV score is the fraction of counterfactual explanations that successfully flipped the classification decision i.e., if the input image is negative (normal) then the generated explanation is predicted as positive (abnormal) for the specific classification task. For **self consistency**, we calculated the FO preservation (FOP) score. FOP score is the fraction of real images, with successful detection of FO, in which FO was also detected in the corresponding explanation image $\mathbf{x}_\delta$. Table 7 summarizes our results. In the absence of adversarial loss from cGAN ($\lambda_1 = 0$), FID score is very high and the FOP score is low as the generated images looks very different from the real CXR images. When KL loss for classifier consistency is missing ($\lambda_2 = 0$), the CV score is poor as the generated explanations are derived without considering the classification function and hence they failed to flip the classification decision. In the absence of CARL loss ($\lambda_3 = 0$), the generated explanations are no longer for the same patient as in query CXR, hence FO in query CXR are absent in generated explanations, resulting in low FOP score.

### *6.14. Extended results for saliency maps*

Our method doesn't produce a saliency map by default. We approximated a saliency map as an absolute difference map between the explanations generated for the two extremes (normal with $f(\mathbf{x_c}) < 0.2$ and abnormal $f(\mathbf{x_c}) > 0.8$) of the decision function $f$. We show an example of saliency map generated by our method in Fig. 8. Fig. 21 shows our extended results.

We also compared the saliency maps generated by our model with popular gradients based methods. For quantitative evaluation, we consider the *deletion* evaluation metric (Petsiuk et al., 2018). The metric quantifies how the probability of the target-class changes as important pixels are removed from an image. To remove pixels from an image, we tried selectively impainting the region based on its surroundings. In Fig. 20, we show an example of deletion-by-impainting. For generating results in Table. **??**, we plot the deletion curve for 500 images, and calculated area under the deletion curve (AUDC) for each.

Please note that, as more pixels are removed, the modified images become unrealistic and visually appear different from a CXR. The behavior of the classifier on such images is inconsistent. Low AUDC demonstrates that all the methods are successful in localizing the important regions for classification. However, unlike saliency-based methods, our counterfactual explanation provides extra information on *what* image features in those relevant regions for classification and *how* those image features should be modified to flip the decision.

### *6.15. Disease-specific evaluation*

For quantitative analysis, we randomly sample two groups of real images (1) a *real-normal* group defined as $\mathcal{X}^n = \{\mathbf{x}; f(\mathbf{x}) < 0.2\}$. It consists of real CXR images that are predicted as normal by the classifier $f$. (2) A *real-abnormal* group defined as $\mathcal{X}^p = \{\mathbf{x}; f(\mathbf{x}) > 0.8\}$. For $\mathcal{X}^n$, we generated a counterfactual group as,

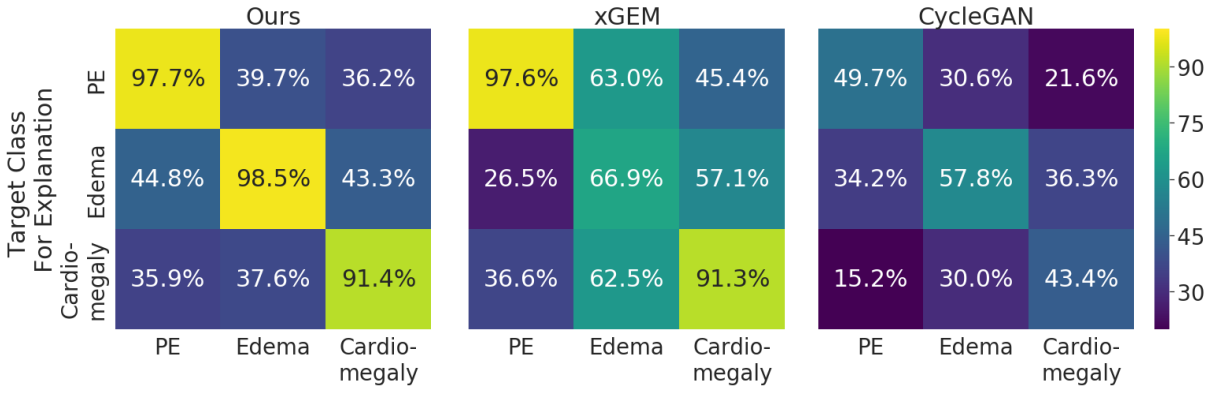


**Fig. 19. Each cell is the fraction of the generated explanations, that have flipped in a class as compared to the query image. The x-axis shows the classes in a multi-label setting, and the y-axis shows the target class for which an explanation is generated. Note: This is not a confusion matrix.**

**Table 7. FID score quantifies the visual appearance of the explanations. CV score is the fraction of explanations that have an opposite prediction compared to the input image. FOP score is the fraction of real images with FO, in which FO was also detected in the corresponding explanation image. In configuration with $\lambda_1 = 0$ there is no adversarial loss from cGAN, in $\lambda_2 = 0$ there is no KL -loss for classifier consistency and in $\lambda_3 = 0$ there is no context-aware self reconstruction loss.**

| | **Cardiomegaly** | | | | **Pleural Effusion** | | | | **Edema** | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Baseline | $\lambda_1 = 0$ | $\lambda_2 = 0$ | $\lambda_3 = 0$ | Baseline | $\lambda_1 = 0$ | $\lambda_2 = 0$ | $\lambda_3 = 0$ | Baseline | $\lambda_1 = 0$ | $\lambda_2 = 0$ | $\lambda_3 = 0$ |
| **FID score** | | | | | | | | | | | | |
| Normal | 166 | 200 | 174 | 160 | 146 | 210 | 150 | 149 | 149 | 169 | 153 | 155 |
| Abnormal | 137 | 189 | 138 | 140 | 122 | 178 | 120 | 130 | 102 | 170 | 109 | 120 |
| **Counterfactual Validity (CV) Score** | | | | | | | | | | | | |
| Real ($f(\mathbf{x}) \in [0, 1]$) | 0.91 | 0.89 | 0.43 | 0.92 | 0.97 | 0.93 | 0.43 | 0.97 | 0.98 | 0.95 | 0.45 | 0.91 |
| **Foreign Object Preservation (FOP) score** | | | | | | | | | | | | |
| Pacemaker | 0.52 | 0.2 | 0.55 | 0.19 | | | | | | | | |

$\mathcal{X}^p_{\mathbf{c}} = \{\mathbf{x} \in \mathcal{X}^n; f(\mathcal{I}_f(\mathbf{x}, \mathbf{c})) > 0.8\}$. Similarly for $\mathcal{X}^p$, we derived a counterfactual group as $\mathcal{X}^n_{\mathbf{c}} = \{\mathbf{x} \in \mathcal{X}^p; f(\mathcal{I}_f(\mathbf{x}, \mathbf{c})) < 0.2\}$.

Next, we quantify the differences in real and counterfactual groups by performing statistical tests on the distribution of clinical metrics such as cardiothoracic ratio (CTR) and the Score of normal Costophrenic recess (SCP). Specifically, we performed the dependent t-test statistics on clinical metrics for paired samples ($\mathcal{X}^n$ and $\mathcal{X}^p_{\mathbf{c}}$), ($\mathcal{X}^p$ and $\mathcal{X}^n_{\mathbf{c}}$) and the independent two-sample t-test statistics for normal ($\mathcal{X}^n$, $\mathcal{X}^n_{\mathbf{c}}$) and abnormal ($\mathcal{X}^p$, $\mathcal{X}^p_{\mathbf{c}}$) groups. The two-sample t-tests are statistical tests used to compare the means of two populations. A low p-value $< 0.0001$ rejects the null hypothesis and supports the alternate hypothesis that the difference in the two groups is statistically significant and that this difference is unlikely to be caused by sampling error or by chance. For paired t-test, the mean difference corresponds to the average causal effect of the intervention on the variable under examination. In our setting, intervention is a *do* operator on input image ($\mathbf{x}$), before intervention, resulting in a counterfactual image ($\mathbf{x}_{\mathbf{c}}$), after intervention.

Table 8 provides the extended results for the Fig. 9. Patients with cardiomegaly have higher CTR as compared to normal subjects. Hence, one should expect CTR($\mathcal{X}^n$) $<$ CTR($\mathcal{X}^p_{\mathbf{c}}$) and likewise CTR($\mathcal{X}^p$) $>$ CTR($\mathcal{X}^n_{\mathbf{c}}$). Consistent with clinical knowledge, in Table. 8, we observe a negative mean difference of -0.03 for CTR($\mathcal{X}^n$) $-$ CTR($\mathcal{X}^p_{\mathbf{c}}$) (a p-value of $< 0.0001$) and a positive mean difference of 0.14 for CTR($\mathcal{X}^p$) $-$ CTR($\mathcal{X}^n_{\mathbf{c}}$) (with a p-value of $\ll 0.0001$). On a population-level CTR was successful in capturing the difference between normal and abnormal CXRs. Specifically in un-paired differences, we observe a low mean CTR values for normal subjects *i.e.,* mean CTR($\mathcal{X}^n$) = 0.46 as compared to mean CTR for abnormal patients *i.e.,* mean CTR($\mathcal{X}^p$) = 0.56. The low p-values supports the alternate hypothesis that the difference in the two groups is statistically significant.

Further, in Fig 21.A, we show samples from input images that were predicted as negative for cardiomegaly ($\mathcal{X}^n$). In their counterfactual abnormal images (third column), we observe small changes in CTR are sufficient to flip the classification decision. This is consistent with a small mean difference CTR($\mathcal{X}^n$) - CTR($\mathcal{X}^a_{\mathbf{c}}$) = $-0.03$. In contrast, when we generate counterfactual normal (sixth column) from real abnormal images (positive for cardiomegaly, Fig 21.B), significant changes in CTR lead to flipping of the prediction decision. This observation is consistent with a large mean difference CTR($\mathcal{X}^a$) - CTR($\mathcal{X}^n_{\mathbf{c}}$) = 0.14.

By design, the object detector assigns a low SCP to any indication of blunting CPA or abnormal CP recess. Hence, SCP($\mathcal{X}^n$) $>$ SCP($\mathcal{X}^p_{\mathbf{c}}$) and likewise SCP($\mathcal{X}^p$) $<$ SCP($\mathcal{X}^n_{\mathbf{c}}$). Consistent with our expectation, in Table. 8, we observe a positive mean dif-

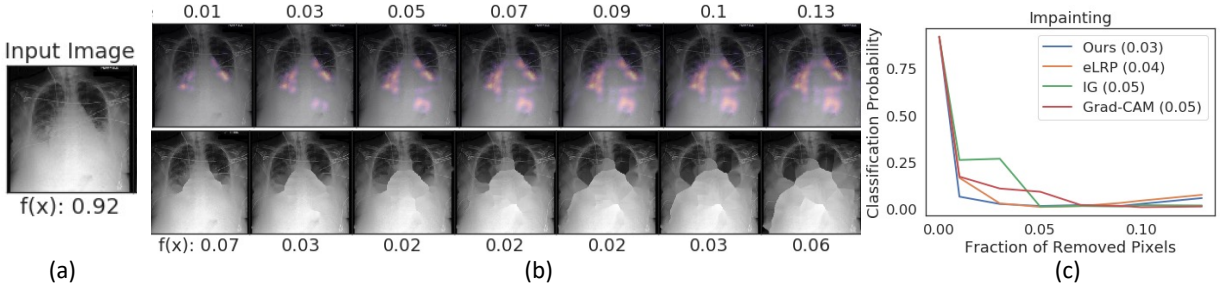


**Fig. 20. Deletion-by-impainting: (a) input image. (b) transformation of the input image as important pixels are deleted, and the resulting patches are in-filled base on the surrounding context. The importance is derived from the saliency map produced from our (top-row) and gradient-based (bottom-row) method. The top label shows the fraction of removed pixels. The bottom label shows the classification outcome for a target class. (c) The plot shows the change in classification prediction as a function of the fraction of removed pixels.**

**Table 8. Results of independent t-test. We compared the difference distribution of cardiothoracic ratio (CTR) for cardiomegaly and the Score for normal Costophrenic recess (SCP) for pleural effusion.**

| Target Disease | Real Group | Counterfactual Group | Paired Differences: Mean Difference | Std | 95% CI Lower | 95% CI Upper | t | df | p-value |
|---|---|---|---|---|---|---|---|---|---|
| Cardiomegaly (CTR) | $\mathcal{X}^n$ | $\mathcal{X}_{\mathbf{c}}^{n\to p}$ | **-0.03** | 0.07 | -0.03 | -0.01 | -4.4 | 304 | $< 0.0001$ |
| | $\mathcal{X}^p$ | $\mathcal{X}_{\mathbf{c}}^{p\to n}$ | **0.14** | 0.12 | 0.13 | 0.15 | 24.7 | 513 | $\ll 0.0001$ |
| Pleural effusion (SCP) | $\mathcal{X}^n$ | $\mathcal{X}_{\mathbf{c}}^{n\to p}$ | **0.13** | 0.22 | 0.06 | 0.13 | 5.9 | 217 | $\ll 0.0001$ |
| | $\mathcal{X}^p$ | $\mathcal{X}_{\mathbf{c}}^{p\to p}$ | **-0.19** | 0.27 | -0.18 | -0.09 | -6.7 | 216 | $\ll 0.0001$ |
| | | | **Un-Paired Differences** | | | | | | |
| | | | Mean Real Group | Mean Counterfactual Group | 95% CI Lower | 95% CI Upper | t | df | p-value |
| Cardiomegaly (CTR) | $\mathcal{X}^n$ | $\mathcal{X}_{\mathbf{c}}^{p\to n}$ | **0.46** | 0.42 | 0.02 | 0.06 | 5.2 | 817 | $< 0.0001$ |
| | $\mathcal{X}^p$ | $\mathcal{X}_{\mathbf{c}}^{n\to p}$ | **0.56** | 0.50 | 0.04 | 0.07 | 9.9 | 817 | $\ll 0.0001$ |
| Pleural effusion (SCP) | $\mathcal{X}^n$ | $\mathcal{X}_{\mathbf{c}}^{p\to n}$ | **0.69** | 0.61 | 0.18 | 0.27 | 9.3 | 433 | $\ll 0.0001$ |
| | $\mathcal{X}^p$ | $\mathcal{X}_{\mathbf{c}}^{n\to p}$ | 0.42 | **0.56** | -0.32 | -0.21 | -9.7 | 433 | $\ll 0.0001$ |

ference of 0.13 for SCP($\mathcal{X}^n$) − SCP($\mathcal{X}_{\mathbf{c}}^p$) (with a p-value of $\ll 0.0001$) and a negative mean difference of -0.19 for SCP($\mathcal{X}^p$) − SCP($\mathcal{X}_{\mathbf{c}}^n$) (with a p-value of $\ll 0.0001$). On a population-level SCP was successful in capturing the difference between normal and abnormal CXR for pleural effusion. Specifically in un-paired differences, we observe a high mean SCP values for normal subjects *i.e.,* mean SCP($\mathcal{X}^n$) = 0.69 as compared to mean SCP for abnormal patients *i.e.,* mean SCP($\mathcal{X}^p$) = 0.42.

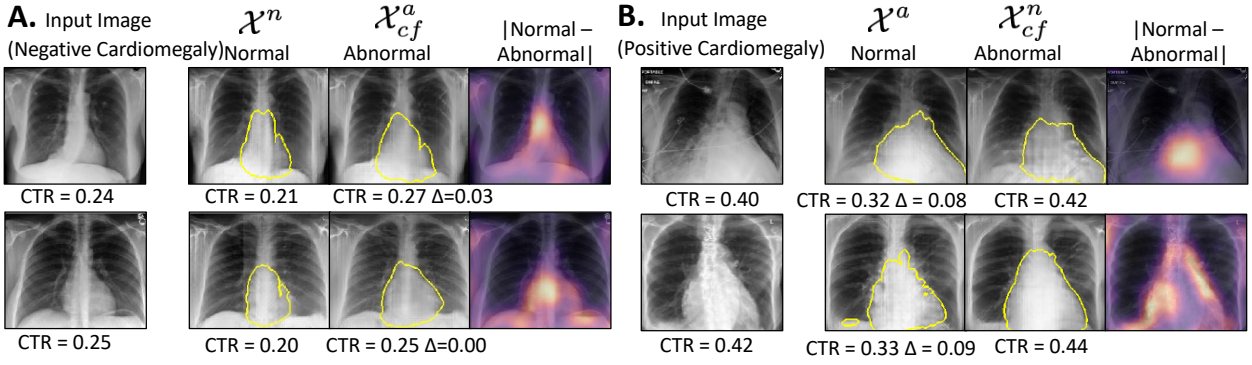


**Fig. 21. Extended results for explanation produced by our model for Cardiomegaly. For each image, we generate a normal and an abnormal explanation image. We show pixel-wise difference of the two generated images as the saliency map. In column A.(B.), we show input images negatively (positively) classified for Cardiomegaly. The yellow contour shows the heart boundary learned by a segmentation network. CTR is the cardiothoracic ratio. For column A, we observe a relatively minor change in CTR (Δ) between real and counterfactual images than in column B.**