Explaining the Black-box Smoothly- A Counterfactual Approach

Sumedha Singla, Brian Pollack, Stephen Wallace and Kayhan Batmanghelich

***Abstract*— We propose a BlackBox *Counterfactual Explainer* that is explicitly developed for medical imaging applications. Classical approaches (*e.g.,* saliency maps) assessing feature importance do not explain *how* and *why* variations in a particular anatomical region are relevant to the outcome, which is crucial for transparent decision making in healthcare application. Our framework explains the outcome by gradually *exaggerating* the semantic effect of the given outcome label. Given a query input to a classifier, Generative Adversarial Networks produce a progressive set of perturbations to the query image that gradually changes the posterior probability from its original class to its negation. We design the loss function to ensure that essential and potentially relevant details, such as support devices, are preserved in the counterfactually generated images. We provide an extensive evaluation of different classification tasks on the chest X-Ray images. Our experiments show that a counterfactually generated visual explanation is consistent with the disease's clinical relevant measurements, both quantitatively and qualitatively.**

***Index Terms*— Explainable AI, Interpretable Machine Learning, Counterfactual Reasoning, Chest X-Ray diagnosis explanation**

I. Introduction

Machine learning, specifically Deep Learning (DL), is being increasingly used for sensitive applications such as Computer-Aided Diagnosis [1] and other tasks in the medical imaging domain [2], [3]. However, for real-world deployment [4], the decision-making process of these models should be explainable to humans to obtain their trust in the model [5], [6]. Explainability is essential for auditing the model [7], identifying various failure modes [8], [9] or hidden biases in the data or the model [10], and for obtaining new insights from large-scale studies [11]. Current explanation methods focus on highlighting the important regions (*where*) for the classification decisions. The location information alone is insufficient for applications in medical imaging. A thorough explanation should explain *what* imaging features are present in those locations and *how* these features can be modified to change the classification decision. In this paper, we provide counterfactual explanations. A visual explanation is derived by gradually transforming the input image into its perturbation, where the model's decision has flipped.

S. Singla is with the Compuer Science Department at the University of Pittsburgh, Pittsburgh, PA 15206, USA (email: sus98@pitt.edu)

B. Pollack, and K. Batmanghelich are with the Department of Biomedical Informatics, the University of Pittsburgh (email: brp98@pitt.edu, kayhan@pitt.edu)

S. Wallace is with the University of Pittsburgh Medical School (e-mail: wallacesr2@upmc.edu).

Post-hoc *explanation* is a popular approach that aims to improve human understanding of a pre-trained model. Our work broadly relates to the following post-hoc methods:

Feature Attribution methods provide an explanation as a saliency map that reflects the importance of each input component (*e.g.,* pixel) to the classification decision. *Gradient-based* approaches [12]–[18] produce a saliency map by computing the gradient of the classifier's output with respect to the input components. Such methods are often applied to the medical imaging studies, *e.g.,* chest x-rays [19], skin imaging [20], brain MRI [21] and retinopathy [22].

Perturbation-based methods identify salient regions by directly manipulating the input image and analyzing the resulting changes in the classifier's output. Such methods aim to modify specific pixels or regions in an input image, either by masking with constant values [23] or with random noise, occluding [24], localized blurring [25], or in-filling [26]. Especially for medical images, such perturbations may introduce anatomically implausible features or textures. Our explanation framework enforces consistency between the perturbed data and the real data distribution to ensure that the perturbation is plausible and realistic-looking.

Counterfactual Explanations are a type of contrastive [27] explanation that are generated by perturbing the real data such that the classifier's prediction is flipped. Similar to our method, generative models like GANs and variational autoencoders (VAE) are used to compute interventions that generate realistic counterfactual explanations [28]–[34]. Much of this work is limited to simpler image datasets like MNIST, celebA [30]–[32] or simulated data [33]. For more complex natural images, previous studies [26], [34] focused on finding and in-filling salient regions to generate counterfactual images. In contrast, our explanation model doesn't require any re-training for generating explanations for a new image at inference time. In another line of work [35], [36] provide counterfactual explanations that explain both the predicted and the counter class. Further [37], [38] used a cycle-GAN [39] to perform image-to-image translation between normal and abnormal images. Such methods are independent of the classifier. In contrast, our model uses special loss to enable image perturbation that is consistent with the classifier.

Recently, researchers have focused on providing explanations in the form of human-defined concepts [40]–[42]. In medical imaging, such methods have been adopted to derive an explanation for breast mammograms [43], breast histopathology [44] and cardiac MRIs [45]. We used such human-defined concepts to quantify the differences in real images and their

respective counterfactual explanations.

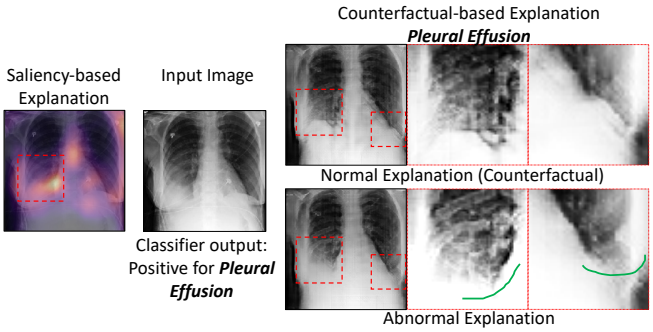


Fig. 1. Counterfactual explanation shows where" in the image the classifier is paying attention and "what" image-features in those regions are associated with the disease. For Pleural Effusion, we can observe the appearance of the meniscus (green) in anabnormal image as compared to the normal counterfactual image.

Fig. 1 shows an example of a saliency map generated by a *generic* explanation model. Saliency maps are inconclusive when different diagnoses affect the same anatomical regions. For example, both pleural effusion and edema may localize in the lower lung lobe region, highlighted in Fig. 1. In contrast, our explanation framework generates a perturbation of the input image, such that the classifier's prediction for the new image is shifted by δ. One can view δ as a "tuning knob" to gradually perturb the input image and traverse the decision boundary from one extreme (normal) to another (abnormal). In Fig. 1, we compared the images generated for the two extremes to identify the salient regions and zoomed-in those regions to understand *how* the image features have transformed to flip the classification decision for pleural effusion.

We adopted a conditional Generative Adversarial Network (cGAN) as our explanation framework to learn the desired perturbation over the input image [46]. However, using cGAN is challenging, as GANs with an encoder may ignore small or uncommon details during image generation [47]. This is particularly important in our application, as the missing information includes foreign objects such as a pacemaker that influence human users' perception. To address this issue, we stipulate when the input image has reconstructed the shape of the anatomy and that foreign objects are preserved. We achieve this by incorporating semantic segmentation and object detection into our loss function.

Our contributions are summarized as follows:

1) We developed a framework to generate a counterfactual visual explanation for a black-box classifier. Our conditional GAN-based approach generates a realistic sequence of images that gradually exaggerate the disease effect.
2) Our method accounts for subtleties of medical imaging by incorporating context from a semantic segmentation and a foreign object detection network.
3) We evaluated our method extensively on various tasks on a chest x-ray dataset.
4) We proposed a quantitative metric based on clinical knowledge for the evaluation of counterfactual explanations.

II. Method

In this paper, we assume that we are given a pre-trained function f, *i.e.,* a *black-box* that accepts the input image, $\mathbf{x}$, and outputs the posterior probability of the classifier, $f(\mathbf{x}) \in [0, 1]$. Also, we assume the gradient of the function $\nabla_{\mathbf{x}} f(\mathbf{x})$, can be computed. To avoid notation clutter, we focus on binary classification throughout this section. However, the proposed method is general and can be used for multi-class or multi-label settings.

Our goal is to learn an *explanation* function $\mathbf{x}_\delta \triangleq \mathcal{I}_f(\mathbf{x}, \delta)$, that perturbs the input image $\mathbf{x}$ and outputs a new image $\mathbf{x}_\delta$ such that the prediction from f is changed by the desired amount δ, *i.e.,* $f(\mathbf{x}_\delta) - f(\mathbf{x}) = \delta$. This formulation allows us to view δ as a "knob" that gradually perturb the input image to achieve visually perceptible differences in $\mathbf{x}$ while crossing the decision boundary given by function f. Figure 2 summarizes our framework. We design the explanation function to satisfy the following properties:

(A) **Data consistency**: $\mathbf{x}_\delta$ should resemble data instance from input space $\mathcal{X}$ *i.e.,* if input space comprises chest x-rays, $\mathbf{x}_\delta$ should look like a chest x-ray with minimum artifacts or blurring.

(B) **Classification model consistency**: $\mathbf{x}_\delta$ should produce the desired output from the classifier f, *i.e.,* $f(\mathcal{I}(\mathbf{x}, \delta)) \approx f(\mathbf{x}) + \delta$.

(C) **Context-aware self-consistency**: To be self-consistent, the explanation function should satisfy three criteria (1) Reconstructing the input image by setting $\delta = 0$ should return the input image, *i.e.,* $\mathcal{I}_f(\mathbf{x}, 0) = \mathbf{x}$. (2) Applying a reverse perturbation on the explanation image $\mathbf{x}_\delta$ should recover $\mathbf{x}$, *i.e.,* $\mathcal{I}_f(\mathbf{x}_\delta, -\delta) = \mathbf{x}$. (3) Achieving the aforementioned reconstructions while preserving anatomical shape and foreign objects (*e.g.,* pacemaker) in the input image.

Next, we will discuss each property in detail.

A. Data consistency

We formulated the explanation function, $\mathcal{I}_f(\mathbf{x}, \delta)$, as an image encoder $E(\cdot)$ followed by a conditional GAN (cGAN) [48], with δ as the condition. The encoder enables the transformation of a given image, while the GAN framework generates realistic-looking transformations as an explanation image. The cGAN is a variant of GAN that allows the conditional generation of the data by incorporating extra information as the context. Like GANs, cGANs are composed of two deep networks, generator $G(\cdot)$ and discriminator $D(\cdot)$. The $G(\cdot)$ network learns to transform samples drawn from a canonical distribution such that $D(\cdot)$ network fails to distinguish the generated data from the real data. The G, D are trained adversarially by optimizing the following objective function,

$$\begin{aligned}\mathcal{L}_{\text{cGAN}}(D, G) = \mathbb{E}_{\mathbf{x},\mathbf{c}\sim P(\mathbf{x},\mathbf{c})}\big[\log\big(D(\mathbf{x}, \mathbf{c})\big)\big] + \\ \mathbb{E}_{\mathbf{z}\sim P_{\mathbf{z}},\mathbf{c}\sim P_{\mathbf{c}}}\big[\log\big(1 - D(G(\mathbf{z}, \mathbf{c}), \mathbf{c})\big)\big]\end{aligned} \tag{1}$$

where $\mathbf{c}$ denotes a condition and $\mathbf{z}$ is noise sampled using a uniform distribution $P_{\mathbf{z}}$. In our formulation, $\mathbf{z}$ is the latent representation of the input image $\mathbf{x}$, learned by the encoder $E(\cdot)$.

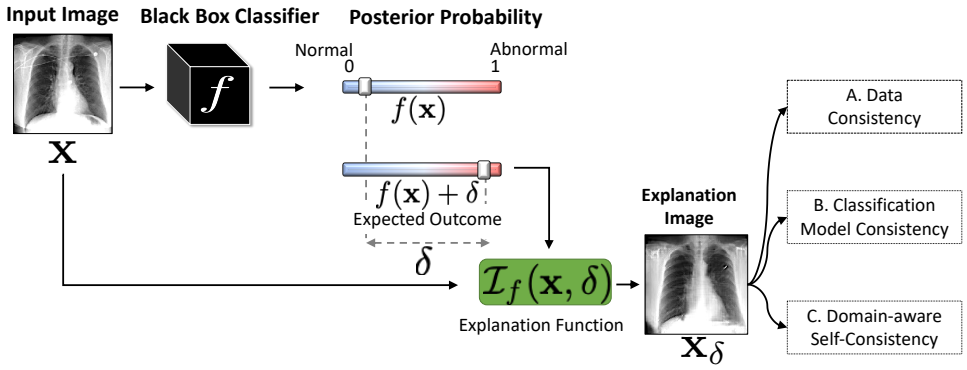


Fig. 2. Explanation function $\mathcal{I}_f(\mathbf{x},\delta)$ for classifier f. Given an input image $\mathbf{x}$, we generates a perturbation of the input, $\mathbf{x}_\delta$ as explanation, such that the posterior probability, f, changes from its original value, $f(\mathbf{x})$, to a desired value $f(\mathbf{x})+\delta$ while satisfying the three consistency constraints.

We model δ as the condition, by defining a discretizing function $c_f(\cdot)$ that maps the posterior probability of the classifier $f \in [0,1]$ to $\lfloor \frac{1}{\delta} \rfloor$ equally-sized bins of width δ. Hence, the explanation function learns to transform the input image, $\mathbf{x}$, which is in bin $c_f(\mathbf{x},0)$, to a perturbed image, $\mathbf{x}_\delta$, with prediction $f(\mathbf{x})+\delta$, which corresponds to bin number $c_f(\mathbf{x},\delta)$. Finally, the explanation function is defined as,

$$\mathbf{x}_\delta = \mathcal{I}_f(\mathbf{x},\delta) = G(E(\mathbf{x}), c_f(\mathbf{x},\delta)). \tag{2}$$

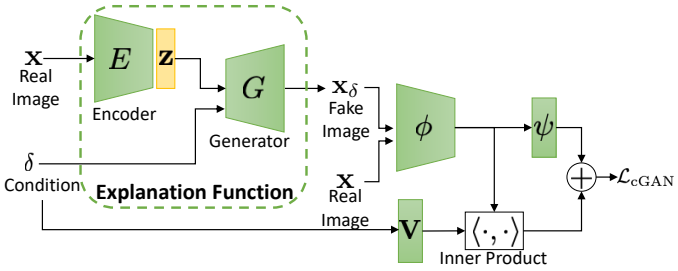


Fig. 3. The explanation function is a conditional-GAN with an encoder. The discriminator evaluates the similarity between real and fake data and the correspondence between fake data and the condition.

For the discriminator in cGAN, we adapted the loss function from Projection GAN [48] based on our application. As $c_f(\mathbf{x},\delta)$ is discrete, we can view its as a one-hot vector $\mathbf{c}$. The loss function of projection cGAN has two terms. The first term is the distribution ratio between marginals *i.e.,* the real data distribution $p_{\text{data}}(\mathbf{x})$ and the learned distribution of the generated data $q(\mathbf{x})$. The second term is the distribution ratio between conditionals. It evaluates the correspondence between the generated image and the condition. This formulation allows us to skip calculating q as we are only interested in the ratio. The overall loss function is as follows,

$$\begin{aligned}\mathcal{L}_{\text{cGAN}}(D,\hat{G})(\mathbf{x},\mathbf{c}) &= \log\frac{p_{\text{data}}(\mathbf{x})}{q(\mathbf{x})} + \log\frac{p_{\text{data}}(\mathbf{c}|\mathbf{x})}{q(\mathbf{c}|\mathbf{x})} \\ &:= r(\mathbf{x}) + r(\mathbf{c}|\mathbf{x}) \\ &:= \psi(\boldsymbol{\phi}(\hat{G}(\mathbf{z});\theta_\phi);\theta_\psi) + \mathbf{c}^T\mathbf{V}\boldsymbol{\phi}(\mathbf{x};\theta_\phi),\end{aligned} \tag{3}$$

where $\mathcal{L}_{\text{cGAN}}(D,\hat{G})$ indicates the loss function in Eq. 1 when $\hat{G}$ is fixed. $\boldsymbol{\phi}(\cdot)$ is an image feature extractor that become modulated on the embedding of the condition, $\mathbf{c}$, defined by the embedding matrix $\mathbf{V}$. The inner product computes the similarity between the latent representation and the condition. Function $\psi(\cdot)$ outputs a scalar value as loss. We modified $r(\mathbf{c}|\mathbf{x})$ to make it consistent with our formulation, in the next section. The parameters $\theta = \{\mathbf{V},\theta_\phi,\theta_\psi\}$ are learned through adversarial training.

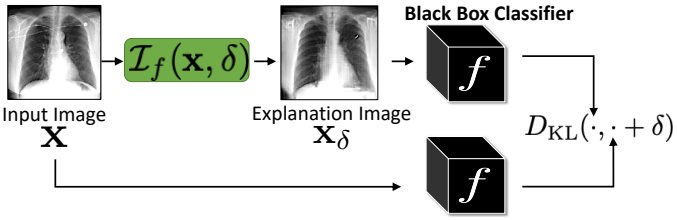


Fig. 4. To enforce consistency with the classifier f, we minimize a KullbackLeibler (KL) divergence between the actual $f(\mathbf{x}_\delta)$ and the desired $f(\mathbf{x})+\delta$ prediction from f. $\mathcal{I}_f(\mathbf{x},\delta)$ is the explanation function in Fig. 3.

B. Classification model consistency

The bin-index $c_f(\mathbf{x},\delta)$ is an ordinal-categorical variable, *i.e.,* $c_f(\mathbf{x},\delta_1) < c_f(\mathbf{x},\delta_2)$ when $\delta_1 < \delta_2$. We adapted Eq. 3 to account for a categorical variable as the condition, by modifying the second term to support ordinal multi-class regression. Specifically, we replaced a single one-hot vector for the condition $\mathbf{c}$, with $\lfloor \frac{1}{\delta} \rfloor - 1$ binary classification terms [49]. The ith binary attribute represents the test $i < n$ where $\mathbf{c} = n$. The modified loss function is as follows:

$$r(\mathbf{c}=n|\mathbf{x}) := \sum_{i<n} \mathbf{v}_i^T \boldsymbol{\phi}(\mathbf{x}), \tag{4}$$

Along with conditional loss for the discriminator, we need additional regularization for the generator to ensure that the actual classifier's outcome, *i.e.,* $f(\mathbf{x}_\delta)$, is very similar to the desired outcome, *i.e.,* $f(\mathbf{x})+\delta$. To ensure this compatibility with f, we further constrain the generator to minimize the KullbackLeibler (KL) divergence that encourages the classifier's score for $\mathbf{x}_\delta$ to differ from $\mathbf{x}$ by a margin of δ (*see* Fig. 4). Our final condition-aware loss is as follows,

$$\mathcal{L}_f(D,G) := r(\mathbf{c}|\mathbf{x}) + D_{\text{KL}}(f(\mathbf{x}_\delta)||f(\mathbf{x})+\delta), \tag{5}$$

Here, the first term evaluates a conditional probability associated with the generated image given the condition $\mathbf{c}$ and is a function of both G and D. The second term uses a KL divergence to compare the actual posterior probability for new image $f(\mathbf{x}_\delta)$ against the desired prediction distribution $f(\mathbf{x}) + \delta$. It influences only the G. Please note that, the term $r(\mathbf{x})$ is not appearing in Eq. 5 as it is independent of the condition $\mathbf{c}$ or δ.

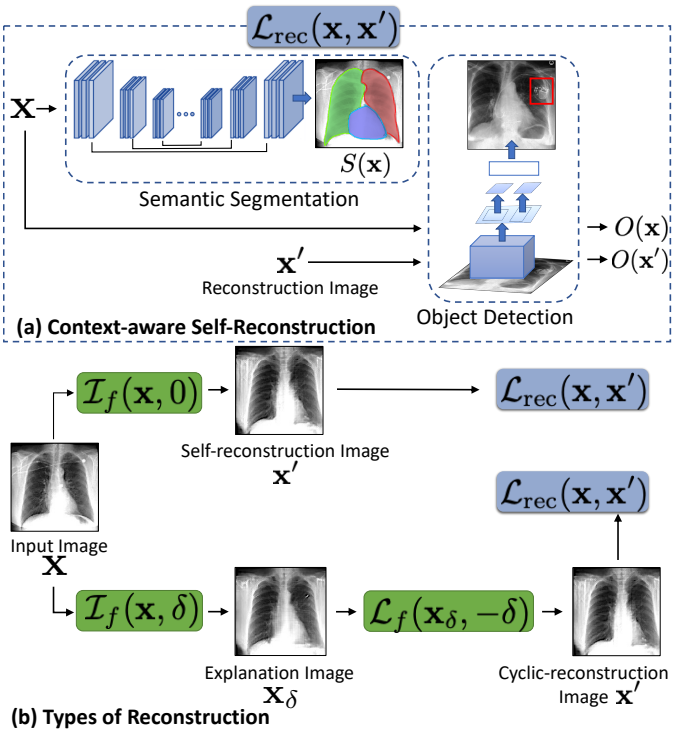


Fig. 5. (a) A domain-aware self-reconstruction loss with pre-trained semantic segmentation $\boldsymbol{S}(\mathbf{x})$ and object detection $\boldsymbol{O}(\mathbf{x})$ networks. (b) The self and cyclic reconstruction should retain maximum information from $\mathbf{x}$. Note, explanation image $\mathbf{x}_\delta$ may differ from input image, $\mathbf{x}$.

C. Context-aware self consistency

A valid explanation image is a minor modification of the input image and should preserve the inputs' identity *i.e.,* patient-specific information such as the anatomy shape. While images generated by a GAN are shown to be realistic looking [50], GAN with an encoder may ignore small or uncommon details in the input image [47]. To preserve these features, we propose a context-aware reconstruction loss (CARL) that exploits extra information from the input domain to refine the reconstruction results. This additional information comes as semantic segmentation and detection of any foreign object present in the input image. The CARL is defined as,

$$\mathcal{L}_{\text{rec}}(\mathbf{x}, \mathbf{x}') = \sum_j \frac{S_j(\mathbf{x}) \odot ||\mathbf{x} - \mathbf{x}'||_1}{\sum S_j(\mathbf{x})} + D_{\text{KL}}(O(\mathbf{x})||O(\mathbf{x}')). \tag{6}$$

Here, $S(\cdot)$ is a pre-trained semantic segmentation network that produces a label map for different regions in the input domain. $O(\mathbf{x})$ is a pre-trained object detector that, given an input image $\mathbf{x}$, output a binary mask $O(\mathbf{x})$, highlighting the region where FO is present. In Eq. 6, we used KL divergence to compare the probability mask created by $O(\cdot)$ over the input $\mathbf{x}$ and the reconstructed $\mathbf{x}'$ image. Rather than minimizing a distance such as ℓ_1 over the entire image, we minimize the reconstruction loss for each segmentation-label (j). Such a loss heavily penalizes differences in small regions to enforce local consistency.

Finally, we used the CAR loss to enforce two important properties of the explanation function:

1) If $\delta = 0$, the self-reconstructed image should resemble the input image.
2) For $\delta \neq 0$, applying a reverse perturbation on the explanation image $\mathbf{x}_\delta$ should recover the initial image *i.e.,* $\mathbf{x} \approx \mathcal{I}_f(\mathcal{I}_f(\mathbf{x}, \delta), -\delta)$.

We enforce these two properties by the following loss,

$$\mathcal{L}_{\text{identity}}(E, G) = \mathcal{L}_{\text{rec}}(\mathbf{x}, \mathcal{I}_f(\mathbf{x}, 0)) + \mathcal{L}_{\text{rec}}(\mathbf{x}, \mathcal{I}_f(\mathcal{I}_f(\mathbf{x}, \delta), -\delta)). \tag{7}$$

where $\mathcal{L}_{\text{rec}}(\cdot)$ is defined in Eq. 6. We minimize this loss only while reconstructing the input image (either by performing self or cyclic reconstruction). For the explanation image, $\mathbf{x}_\delta$, with a bin number different from the input image, we didn't enforce the reconstruction loss to ensure that the explanation function is not biased towards foreign objects or region-specific details.

D. Objective function

The overall objective function is

$$\min_{E,G} \max_{D} \lambda_1 \mathcal{L}_{\text{cGAN}}(D, G) + \lambda_2 \mathcal{L}_f(D, G) + \lambda_3 \mathcal{L}_{\text{identity}}(E, G) \tag{8}$$

where λ's are the hyper-parameters to balance each of the loss terms. The encoder $E(\cdot)$ and generator $G(\cdot)$ network follows ResNet [51] architecture. $G(\cdot)$ processes the latent representation to create a new image while incorporating condition information using conditional batch normalization (cBN). For discriminator $D(\cdot)$ network, we adapted the architecture from SNGAN [52]. The model is trained end-to-end to learn parameters for the three networks. Please note that the parameters for the classifier remained fixed throughout the training process. We optimized the adversarial hinge loss for the cGAN training. We set the loss hyper-parameters as $\lambda_1 = 1.0$, $\lambda_2 = 1.0$ and $\lambda_3 = 0.5$. We used the Adam optimizer [53], with default hyper-parameters set to $\alpha = 0.0002, \beta_1 = 0, \beta_2 = 0.9$.

III. Experiments

In this section, we evaluate our method using a chest x-ray dataset. We performed three sets of experiments:

(1) We evaluated our model on the three desiderata of valid explanations, defined in the method section. We compared our counterfactual explanations with closest existing methods such as xGEM [54] and CycleGAN [37], [38]. We considered the following three evaluation metrics: Fréchet Inception Distance (FID) score to assess visual quality, counterfactual validity (CV) score to quantify compatibility with the classifier, and foreign object preservation (FOP) score to evaluate the retention of patient-specific information in the explanations.

(2) We compared against the saliency-based methods to provide *post-hoc* model explanation. While our method does

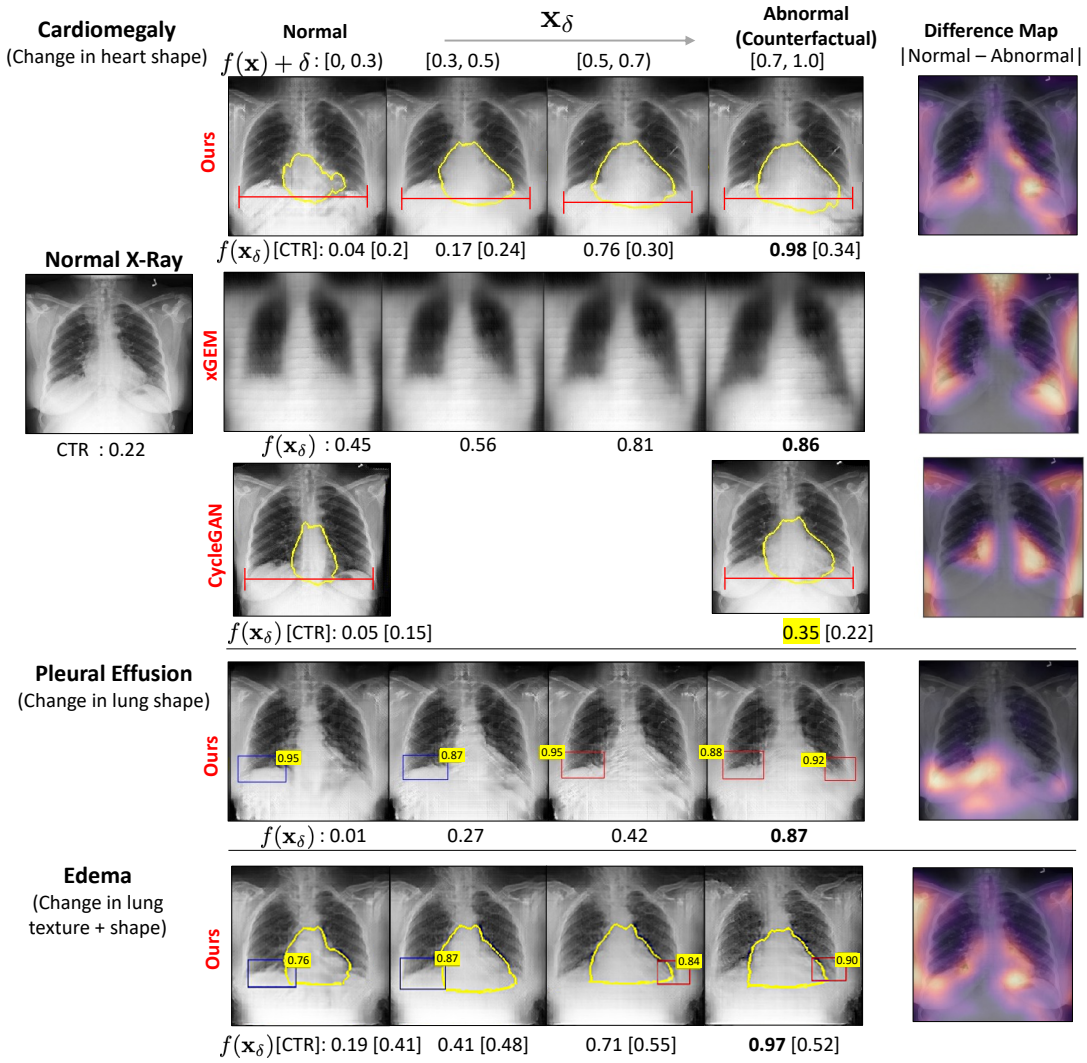


Fig. 6. Qualitative comparison of the counterfactual explanations generated for three classes, cardiomegaly (first row), pleural effusion (PE) (middle row), and edema (last row). The bottom labels are the classifier's predictions for the specific class. The yellow color highlight the prediction where counterfactual fails to flip the decision. The last column shows the difference map between normal and abnormal explanations. For cardiomegaly and edema, we are reporting cardiothoracic ratio (CTR) calculated from the heart segmentation (yellow) and thoracic diameter (red). For PE and edema, we show the bounding box (BB) for normal (blue) and abnormal (red) costophrenic (CP) recess. The number on blue-BB is the Score for detecting a normal CP recess (SCP). The number on red-BB is 1-SCP. For cardiomegaly, we are also showing the corresponding counterfactual explanations for xGEM and cycleGAN.

not produce a saliency map, we approximate it as a difference map between the explanations generated for the two extremes of the decision boundary.

(3) We used two clinical metrics, namely, cardiothoracic ratio (CTR) and the Score for detecting a normal Costophrenic recess (SCP) to demonstrate the clinical relevance of our explanations. CTR is associated with cardiomegaly, and SCP is indicative of pleural effusion (PE).

Experimental setup: We performed our experiments on MIMIC-CXR [55], which is a multi-modal dataset consisting of 377K chest X-ray images and 227K reports from 65K patients. Images are provided with binarized labels over fourteen radio-graphic observations, namely, enlarged cardiomediastinum, cardiomegaly, lung-lesion, lung-opacity, edema, consolidation, pneumonia, atelectasis, pneumothorax, pleural effusion, pleural other, fracture, support devices and no-finding. The images are preprocessed using a standard pipeline involving cropping, re-scaling, and intensity normalization. Following the prior work on diagnosis classification [56], we used DenseNet-121 [57] architecture as the classification model, which performs multi-label classification over fourteen labels, given the frontal view chest x-ray images. The model is trained on 80% of the images. The rest 50K images are further divided into 3:2 to create a training-testing dataset for the explanation model. No data augmentation was used for training the explanation model. Our experiment learns three explanation models to explain the three target labels: cardiomegaly, pleural effusion, and edema.

In our experiments, we set $\delta = 0.1$, and divide $f(\mathbf{x})[y] \in [0, 1]$ into ten equally size bins of width 0.1. Here, y is a target label. Next, we map each input image to a bin-index depending on the classification prediction $f(\mathbf{x})[y]$. From the

TABLE I

THE FID SCORE QUANTIFIES THE VISUAL APPEARANCE OF THE EXPLANATIONS. THE COUNTERFACTUAL VALIDITY (CV) SCORE IS THE FRACTION OF EXPLANATIONS THAT HAVE AN OPPOSITE PREDICTION COMPARED TO THE INPUT IMAGE.

	Cardiomegaly			**Pleural Effusion**			**Edema**		
	Ours	xGEM	CycleGAN	Ours	xGEM	CycleGAN	Ours	xGEM	CycleGAN
FID score									
Normal ($f(\mathbf{x}), f(\mathbf{x}_\delta) < 0.2$)	166	384	**30**	146	347	**37**	149	376	**72**
Abnormal ($f(\mathbf{x}), f(\mathbf{x}_\delta) > 0.8$)	137	316	**56**	122	355	**35**	102	274	**77**
Counterfactual Validity Score									
Real ($f(\mathbf{x}) \in [0, 1]$)	**0.91**	**0.91**	0.43	**0.97**	**0.97**	0.49	**0.98**	0.66	0.57

training set of the explanation model, we sample images such that each bin has 2500 to 3000 images. We created a similar non-overlapping dataset to test the explanation model with 700 to 1000 images in each bin. All the results are computed on this testing dataset.

For semantic segmentation, we adopted a 2D U-Net [58] to mark the lung and the heart contour in a chest x-ray. The network is trained on 385 chest x-rays and masks from Japanese Society of Radiological Technology (JSRT) [59] and Montgomery [60] datasets. We trained a Fast Region-based CNN [61] network for detecting foreign objects (FO) such as pacemaker and hardware in a chest x-ray. We manually created a training dataset of 300 x-rays by collecting bounding box annotations for FO. We further trained two detectors for identifying normal and abnormal costophrenic (CP) recess regions in the chest x-ray. We identify an abnormal CP recess through a positive mention for *"blunting of the 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. The detailed architecture for all modules is provided in the **Supplementary Material** (SM).

A. Desiderata of explanation function

In this section, we evaluate our method on three desiderata of a valid counterfactual [62]. First, *Data consistency:* A counterfactual should be realistic-looking *i.e.,* it should be very similar to the input image. Second, *Classification model consistency:* A counterfactual should flip the classification decision for the input image. Third, *Identify preservation:* A counterfactual should preserve patient-specific details such as foreign objects.

1) Data consistency: A counterfactual explanation is a minimal but perceptible modification of the input x-ray image that flips the classification decision. Given an input image, our model generates a series of images $\mathbf{x}_\delta$ as explanations that eventually flip the classification decision. We create multiple explanations by gradually changing δ such that $f(\mathbf{x}) + \delta$ is in range $[0, 1]$. In Fig. 6, the left-most image is the input x-ray of a normal subject. In the middle, we showed the explanation images for the three target diseases, cardiomegaly, PE, and edema. The last column presented a pixel-wise difference map between normal and abnormal explanations. The heatmaps highlight the regions that changed the most during the transformation. For **cardiomegaly**, we reported the cardiothoracic ratio (CTR). It is calculated as the ratio of the cardiac diameter extracted from the heart contour (yellow) and the thoracic diameter (red). CTR aids in the detection of enlargement of the cardiac silhouette. We observed a gradual increase in posterior probability $f(\mathbf{x}_\delta)$ (bottom label) as we transformed from normal to an abnormal counterfactual image. During this transformation, the CTR increased with corresponding changes in the heart shape. For **PE**, we showed the results of an object detector as bounding-box (BB) over the normal (blue) and abnormal (red) CP recess regions. The number on the top-right of the blue-BB is the Score for detecting a normal CP recess (SCP). The number on red-BB is 1-SCP. The CP recess is the potential area to be analyzed for PE [63]. As we go from left to right, the normal CP recess changed into an abnormal CP recess with a high detection score. In **edema**, we observed changes in both CTR and SCP. The counterfactual transformation is associated with an increasing CTR and blurring of the left CP recess region, as highlighted in the difference map. These findings are consistent with radiological signs for cardiogenic edema [64]. We also present a comparison against xGEM and cycleGAN for cardiomegaly. xGEM created blurry images while cycleGAN creates realistic images, but the abnormal-counterfactual failed to flip the classification outcome.

Quantitatively evaluation: We evaluated the visual quality of our explanations by computing FID score [65]. FID quantifies the visual similarity between the real images and the synthetic counterfactuals. FID computes the distance between the activation distributions of the real image $\mathbf{x}$ and the synthetic explanations $\mathbf{x}_\delta$ as,

$$\text{FID}(\mathbf{x}, \mathbf{x}_\delta) = ||\mu_\mathbf{x} - \mu_{\mathbf{x}_\delta}||_2^2 + \text{Tr}(\Sigma_\mathbf{x} + \Sigma_{\mathbf{x}_\delta} - 2(\Sigma_\mathbf{x}\Sigma_{\mathbf{x}_\delta})^{\frac{1}{2}}), \quad (9)$$

where μ's and Σ's are mean and covariance of the activation vectors derived from the penultimate layer of a pre-trained Inception v3 network [65]. We examined real and fake (*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 ($f(\mathbf{x}) > 0.8$). In Table. I, we compared three counterfactual-generating algorithms: ours, xGEM, and cycleGAN, and reported the FID for each group. Our model creates realistic-looking counterfactuals compared to xGEM. The cycleGAN model generates the most realistic-looking images with the lowest FID score (< 80). However, in the next section, we will show that the explanations generated by cycleGAN may not flip the classification decision and hence, may fail to provide a valid counterfactual explanation.

2) Classification model consistency: In this experiment, we quantify the strength of different counterfactual-generating

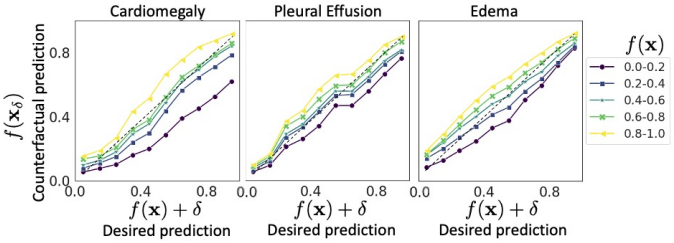


Fig. 7. The plot of desired outcome, $f(\mathbf{x})+\delta$, against actual response of the classifier on generated explanations, $f(\mathbf{x}_\delta)$. Each line represents a set of input images with classification prediction $f(\mathbf{x})$ in a given range. Plots for xGEM and cycleGAN are shown in SM-Fig. 4.

algorithms in creating explanations consistent with the classification model and, thus, successfully flips the classification decision. In the last row of Table. I, we report results on counterfactual validity (CV) score. Mothilal *et al.* [62] proposed CV score as the fraction of counterfactual explanations that corresponds to the opposing end of the prediction spectrum *i.e.,* if the input image is predicted as normal, the generated explanation is predicted as abnormal by the classifier. For all three target diseases, our model created the highest percentage of counterfactually valid explanations. CycleGAN achieved a low CV score, thus creating explanations that are frequently inconsistent with the classifier. Next, we quantify this consistency at every step of the transformation. We divided the classifier's prediction range of $[0, 1]$ into ten equally sized bins. For each bin, we generated an explanation image by choosing an appropriate expected classification output, $f(\mathbf{x}) + \delta$. We further divided the input image space into five groups based on their initial prediction *i.e.,* $f(\mathbf{x})$. In Fig 7, we represented each group as a line and plotted the average response of the classifier *i.e.,* $f(\mathbf{x}_\delta)$ for explanations in each bin against the expected classification outcome *i.e.,* $f(\mathbf{x}) + \delta$. The positive slope of the line-plot, parallel to $y = x$ line at 45° confirms that starting from images with low $f(\mathbf{x})$, our model creates fake images such that $f(\mathbf{x}_\delta)$ is high and vice-versa.

TABLE II
THE FOREIGN OBJECT PRESERVATION (FOP) SCORE AND LATENT-SPACE CLOSENESS (LSC) SCORE FOR OUR MODEL WITH AND WITHOUT THE CONTEXT-AWARE RECONSTRUCTION LOSS (CARL). FOP SCORE DEPENDS ON THE PERFORMANCE OF FO DETECTOR.

Foreign Object	**LSC score** CARL better than ℓ_1	**FOP score** Ours with CARL	Ours with ℓ_1
Pacemaker	0.79	0.52	0.40
Hardware	0.87	0.63	0.32

3) Identity preservation: Ideally, a counterfactual explanation should differ in semantic features associated with the target class while retaining unique properties of a patient, such as foreign objects (FO). FO provide critical information to identify the patient in an x-ray. The disappearance of FO in explanation images creates a distraction and increases confusion that explanation images show a different patient.

In this experiment, we compared explanations generated using CARL against those generated using simple ℓ_1 reconstruction loss on two identity constraints. First, we used latent-space closeness (LSC) score to quantify the similarity between the explanation images and the query image in a latent space. We derived LSC score as the fraction of the images where explanation image derived using CARL ($\mathbf{x}_\delta^{\text{CARL}}$) is closest to the query image $\mathbf{x}$ as compared to explanations generated using ℓ_1 loss *i.e.,* $\mathbf{x}_\delta^{\ell_1}$. We calculated similarity as the euclidean distance between the embedding for the query and explanation images. LSC score is defined as,

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

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

Second, we compared the two reconstruction losses in their ability to preserve FO in explanation images. We calculated FO preservation (FOP) score as the fraction of real images, with successful detection of FO, in which FO was also detected in the corresponding explanation image $\mathbf{x}_\delta$. Our model with CARL loss obtained a higher FOP score, as shown in Table II. The detector network has an accuracy of 80%. Fig. 8 presents examples of counterfactual explanations generated by our model with and without the CARL.

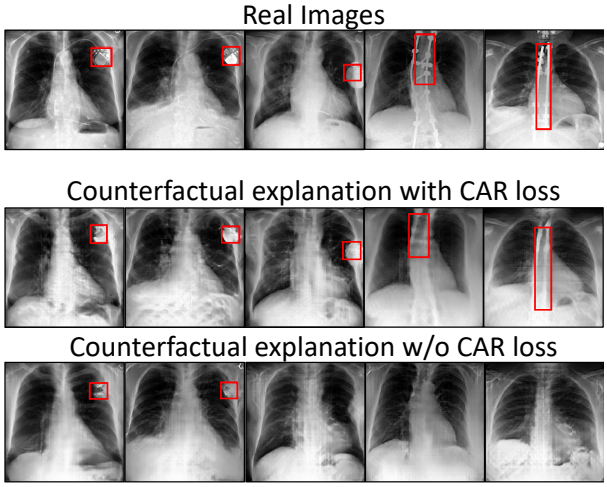


Fig. 8. Fidelity of generated images with respect to preserving FO. The top row shows real images with pacemaker or hardware. The middle shows counterfactual images generated by our model while using context-aware reconstruction (CAR) loss. The bottom row shows the explanation images, without the CAR loss.

B. Comparison with Saliency maps

Popular existing approaches for model explanation consist of gradient-based methods that provide a qualitative explanation in the form of saliency maps [56], [66]. To compare against such methods, we approximated a saliency map as an absolute difference map between the explanations generated for the two extremes (normal with $f(\mathbf{x}_\delta) < 0.2$ and abnormal $f(\mathbf{x}_\delta) > 0.8$) of the decision function f. For proper comparison, we considered the absolute values of the saliency maps and normalized them in the range $[0, 1]$. In Fig. 9 we

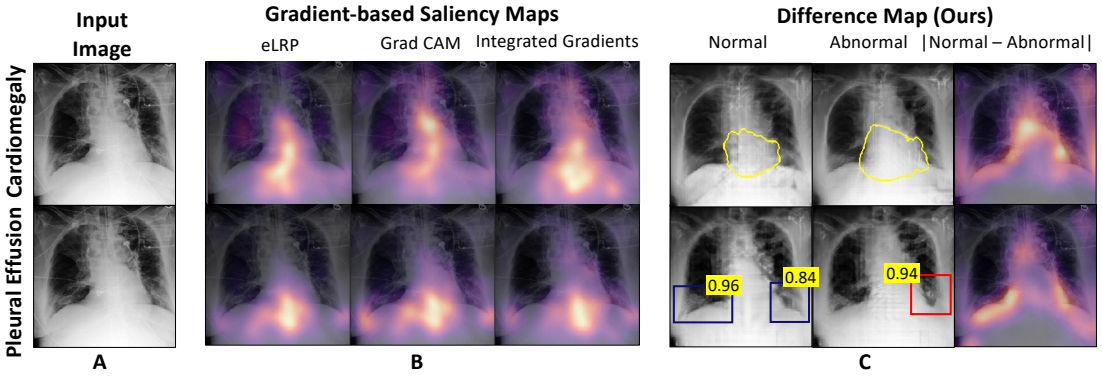


Fig. 9. Comparison of our method against different gradient-based methods. A: Input image; B: Saliency maps from existing works; C: Our simulation of saliency map as difference map between the normal and abnormal explanation images. More examples are shown in SM-Fig. 6, 7.

show an example of an input image, where the gradient-based saliency maps for two target classes highlight almost the same region. In contrast, our difference map localized disease to specific regions in the chest. Fig. 9.C, shows the two extreme explanation images and the corresponding difference map, derived for input images shown in Fig. 9.A.

Further, we used the *deletion* evaluation metric to quantitatively compare the different methods [67]. The metric quantifies how the probability of the target class changes as important pixels are removed from an image. For a given image, we plot the change in classification prediction as a function of the fraction of removed pixels to create the deletion curve (SM-Fig.11 shows an example). A low area under the deletion curve (AUDC) signifies a sharp drop in the probability as more pixels are removed. To remove pixels from an image, we selectively impaint the region based on its surroundings.

TABLE III
QUANTITY COMPARISON OF OUR METHOD AGAINST GRADIENT-BASED METHODS. MEAN AREA UNDER THE DELETION CURVE (AUDC), PLOTTED AS A FUNCTION OF THE FRACTION OF REMOVED PIXELS. A LOW AUDC SHOWS A SHARP DROP IN PREDICTION ACCURACY AS MORE PIXELS ARE DELETED.

Method	**Cardiomegaly**	**Pleural Effusion**	**Edema**
Ours	**0.040±0.04**	**0.023±0.02**	0.083±0.05
eLRP	0.071±0.05	0.033±0.02	0.055±0.03
Grad-CAM	0.045±0.04	0.058±0.05	**0.035±0.02**
Integrated Gradients	0.058±0.06	0.046±0.05	0.077±0.04

In Table III, we report the mean AUDC over a sample of 500 images. The images were selected such that the $f(\mathbf{x}) > 0.9$ for the target-disease. Our model achieved the lowest AUC in deletion-by-impainting for cardiomegaly and pleural effusion. The results show that the regions modified by our explanation model are important for the classification decision.

C. Disease-specific evaluation

Quantifying the clinical relevance of an explanation is a challenging task. We evaluated the clinical relevance in terms of radiographic features that are clinically used to characterize a disease. Specifically, we examined the following two metrics,

1) Cardio Thoracic Ratio (CTR): The CTR is the ratio of the cardiac diameter to the maximum internal diameter of the thoracic cavity. A CTR ratio greater than 0.5 is an abnormal finding associated with cardiomegaly [68]–[70]. We followed the approach in [71] to calculate the CTR from a chest x-ray. In the absence of ground truth lung and heart segmentation on the MIMIC-CXR dataset, we used a segmentation network trained on open-sourced supervised datasets [60], [72] to obtain segmentation. We calculated heart diameter as the distance between the leftmost and rightmost points from the lung centerline on the heart segmentation. The thoracic diameter is calculated as the horizontal distance between the widest points on the lung mask.

2) Costophrenic recess: The fluid accumulation in costophrenic (CP) recess may lead to the diaphragm's flattening and the associated blunting of the angle between the chest wall and the diaphragm arc, called costophrenic angle (CPA). The blunting of CPA is an indication of pleural effusion [72], [73]. Marking the CPA angle on a chest x-ray requires expert supervision, while annotating the CP region with a bounding box is a much simpler task (*see* SM-Fig. 1). We learned an object detector to identify normal or abnormal CP recess in the chest x-rays and used the Score for detecting a normal CP recess (SCP) as our evaluation metric.

Next, we evaluated the extent to which the counterfactual explanations adhere to the clinical understanding of a disease. We performed a statistical test to quantify the differences in real images and their corresponding counterfactuals based on the two clinical metrics. 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 chest x-rays that are predicted as normal by the classifier f. (2) A *real-abnormal* group defined as $\mathcal{X}^a = \{\mathbf{x}; f(\mathbf{x}) > 0.8\}$. For $\mathcal{X}^n$ we generated a counterfactual group as, $\mathcal{X}^a_{cf} = \{\mathbf{x} \in \mathcal{X}^n; f(\mathcal{I}_f(\mathbf{x}, \delta)) > 0.8\}$. Similarly for $\mathcal{X}^a$, we derived a counterfactual group as $\mathcal{X}^n_{cf} = \{\mathbf{x} \in \mathcal{X}^a; f(\mathcal{I}_f(\mathbf{x}, \delta)) < 0.2\}$.

In Fig. 10, we showed the distribution of differences in CTR for cardiomegaly and SCP for PE in a pair-wise comparison between real (normal/abnormal) images and their respective counterfactuals. Patients with cardiomegaly have higher CTR as compared to normal subjects. Hence, one should expect CTR($\mathcal{X}^n$) < CTR($\mathcal{X}^a_{cf}$) and likewise CTR($\mathcal{X}^a$) > CTR($\mathcal{X}^n_{cf}$). Consistent with clinical knowledge, in Fig. 10, we observe a negative mean difference for CTR($\mathcal{X}^n$) − CTR($\mathcal{X}^a_{cf}$) (a

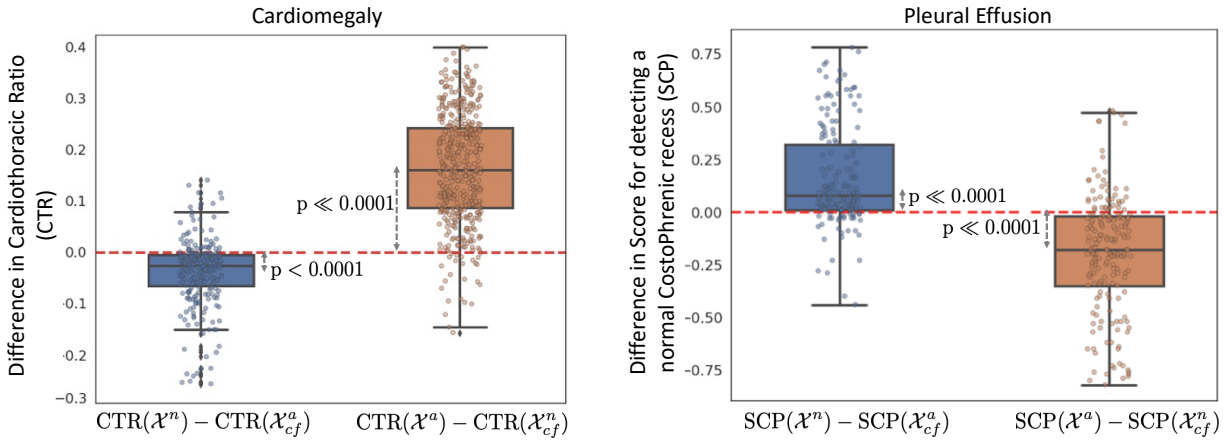


Fig. 10. Box plots to show distributions of pairwise differences in clinical metrics such as CTR for cardiomegaly and the Score of normal CP recess (SCP) for pleural effusion, before (real) and after (counterfactual) our generative counterfactual creation process. The mean value corresponds to the average causal effect of the clinical-metric on the target disease. The low p-values for the dependent t-test statistics confirm the statistically significant difference in the distributions of metrics for real and counterfactual images. The mean and standard deviation for the statistic tests are summarized in SM-Table 1.

p-value of < 0.0001) and a positive mean difference for CTR($\mathcal{X}^a$) − CTR($\mathcal{X}^n_{cf}$) (with a p-value of $\ll 0.0001$). The low p-value in the dependent t-test statistics supports the alternate hypothesis that the difference in the two groups is statistically significant, and this difference is unlikely to be caused by sampling error or by chance.

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($\mathcal{X}^a_{cf}$) and likewise SCP($\mathcal{X}^a$) < SCP($\mathcal{X}^n_{cf}$). Consistent with our expectation, we observe a positive mean difference for SCP($\mathcal{X}^n$) − SCP($\mathcal{X}^a_{cf}$) (with a p-value of $\ll 0.0001$) and a negative mean difference for SCP($\mathcal{X}^a$) − SCP($\mathcal{X}^n_{cf}$) (with a p-value of $\ll 0.0001$). A low p-value confirmed the statistically significant difference in SCP for real images and their corresponding counterfactuals.

IV. Discussion And Conclusion

We provided counterfactual explanations for classification models that are trained for clinical applications. Our framework explains the decision by gradually transforming the input image to its counterfactual, such that the classifier's prediction is flipped. To generate such an explanation, we have formulated and evaluated our framework on three properties of a valid transformation: data consistency, classification model consistency, and self-consistency. Our results in Section III-A showed that our framework adheres to all three properties and creates a realistic-looking explanation that produced a desired outcome from the classification model while retaining maximum patient-specific information.

We compared against two other generative methods for the model explanation, namely, xGEM and cycleGAN. CycleGAN produced the most visually appealing x-ray images with a high FID score. However, during training, the objective function of cycleGAN does not incorporate the external black-box classifier. Consequently, we observe a low counterfactual validity (CV) score in Table I. And in Fig. 6, cycleGAN counterfactual images failed to flip the classification decision for cardiomegaly. In contrast, 90% of the explanations generated by our model successfully flipped the classification decision.

The xGEM explanations are well-grounded with the classifier, with ahigh CV score. However, unless explicitly imposed, the explanation image from xGEM does not look realistic. The expressiveness of the generator limits the visual quality of images. xGEM adopted a variational autoencoder (VAE) as the generator. VAE uses a Gaussian likelihood (ℓ_2 reconstruction), an unrealistic assumption for image data, and is known to produce over-smoothed images [74]. In contrast, our model uses an implicit likelihood assumption as in GAN [48] that results in more realistic explanation images.

We also compared our method against popular saliency-map-based explanations. A good explanation model elaborates the classifier's reasoning by providing different explanations for different decisions *i.e.,* classes. However, for medical images, saliency maps may highlight almost the same region for different diseases, resulting in misleading and inconclusive explanations (*see* Fig. 9). In contrast, our counterfactual explanations provide additional information to clarify *how* input features in the important regions could be modified to change the prediction decision. Our difference map localizes disease to specific regions in the chest, and these regions align with the clinical knowledge of the disease. In Fig. 9 our difference map focused on the heart region for cardiomegaly and the CP recess region for PE.

From a clinical perspective, we demonstrated the usability of our explanations by quantifying the counterfactual changes in terms of disease-specific radiographic features such as CTR and SCP. Our explanations showed that the classification decision is consistent with the medical knowledge of the disease. For example, changes associated with an increased posterior probability for cardiomegaly also resulted in an increased CTR. Similarly, for PE, a healthy CP recess with a sharp diaphragm arc and a high SCP transformed into an abnormal CP recess with blunt CPA, as the posterior probability for PE increases (*see* Fig. 6 and Fig. 10).

To the best of our knowledge, ours is the first attempt to

quantify a model explanation in terms of clinical metrics. At the same time, our evaluation has certain limitations. Our automatic pipeline to compute CTR suffers from inaccuracies, in the absence of ground truth for lung and heart segmentation. Also, the object detector used for detecting normal and abnormal CP recess has a sub-optimal performance. This contributed to the large variance in difference plots in Fig. 10. Nevertheless, on a population level, CTR and SCP successfully captured the difference between normal and abnormal chest x-rays. One may argue using features such as CTR and SCP to perform disease classification. But models based on these features will also suffer from similar inaccuracies due to imperfect segmentation or detection, resulting in poor performance and generalization compared to the deep learning methods.

We acknowledge that there are areas of improvement in our counterfactual explanations. The GAN architecture is not perfect in preserving small details such as breasts and foreign objects (FO) in generated images. This behavior is consistent with a similar finding in computer vision [47]. In comparison to a simple distance-based reconstruction loss, our revised context-aware reconstruction loss (CARL) helped in preserving details such as a pacemaker (*see* Table II). However, even with CARL, the FO preservation score is not perfect. A possible reason for this gap is the limited capacity of the object detector. To the best of our knowledge, there is no publicly available FO detector for a chest x-ray. Hence, we trained an object detector on a manually annotated dataset.

Further, a resolution of 256×256 for counterfactually generated images is smaller than a standard chest x-ray. Small resolution limits the evaluation for fine details by both the algorithm and the interpreter. Our formulation of cGAN uses conditional-batch normalization (cBN) to encapsulate condition information while generating images. For efficient cBN, the mini-batches should be class-balanced. To accommodate high-resolution images with smaller batch sizes, we have to decrease the number of conditions to ensure class-balanced batches. Fewer conditions resulted in a coarse transformation with abrupt changes across explanation images. In our experiments, we selected the smallest δ, which created a class-balanced batch that fits in GPU memory and resulted in stable cGAN training. However, with the advent of larger-memory GPUs, we intend to apply our methods to higher resolution images in future work; and assess how that impacts interpretation by clinicians.

Defining clinical metrics for different diseases is a challenging task. For example, edema is a complex disease. It may appear as different radiographic concepts (*e.g.,* cephalization, peribronchial cuffing, perihilar batwing appearance, and opacities *etc.*) in different patients [75]. Transforming a healthy chest x-ray to a counterfactual image for edema may introduce changes in multiple such concepts. Future research should determine appropriate metrics to quantify and understand these concepts. Manual annotation is one solution for obtaining ground truth to train models that can identify concepts. Efforts should be made to reduce the dependency on manual labeling as it is expensive and not scalable.

To conclude, the counterfactually generated images in this study identified commonly utilized radiographic findings that clinicians use to diagnose and to grade the presence of pathology. In particular, the system did this well for cardiomegaly and pleural effusions and was corroborated by an experienced radiology resident physician. By providing visual explanations for deep learning decisions, radiologists better understand the causes of artificial intelligence decision-making. This is essential to lessen physicians' concerns regarding the "BlackBox" nature by an algorithm and build needed trust for incorporation into everyday clinical workflow. As an increasing amount of artificial intelligence algorithms offer the promise of everyday utility, counterfactually generated images are a promising conduit to building trust among diagnostic radiologists.

By providing counterfactual explanations, our work opens up many ideas for future work. Our framework showed that valid counterfactuals can be learned using an adversarial generative process that is regularized by the classification model. However, counterfactual reasoning is incomplete without a causal structure and explicitly modeling of the interventions. An interesting next step should explore incorporating or discovering plausible causal structures and creating explanations grounded with them.

REFERENCES

[1] A. Hosny, C. Parmar, J. Quackenbush, L. H. Schwartz, and H. J. Aerts, "Artificial intelligence in radiology," *Nature Reviews Cancer*, vol. 18, pp. 500–510, 8 2018.

[2] P. Rajpurkar, J. Irvin, R. L. Ball, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, and et al., "Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists," *PLOS Medicine*, vol. 15, pp. 1–17, 11 2018.

[3] A. Rodriguez-Ruiz, K. Lång, A. Gubern-Mérida, M. Broeders, G. Gennaro, P. Clauser, T. Helbich, M. Chevalier, T. Tan, T. Mertelmeier, M. Wallis, I. Andersson, S. Zackrisson, R. Mann, and I. Sechopoulos, "Stand-alone artificial intelligence for breast cancer detection in mammography: Comparison with 101 radiologists," *Journal of the National Cancer Institute*, vol. 111, 03 2019.

[4] F. Wang, R. Kaushal, and D. Khullar, "Should health care demand interpretable artificial intelligence or accept "black Box" Medicine?" *Annals of Internal Medicine*, vol. 172, pp. 59–61, 1 2020.

[5] A. Gastounioti and D. Kontos, "Is It Time to Get Rid of Black Boxes and Cultivate Trust in AI?" *Radiology: Artificial Intelligence*, vol. 2, p. e200088, 5 2020.

[6] H. Jiang, B. Kim, M. Guan, and M. Gupta, "To trust or not to trust a classifier," *International Conference on Advances in Neural Information Processing Systems (NeurIPS)*, vol. 31, 2018.

[7] J. Winkler, C. Fink, F. Toberer, A. Enk, T. Deinlein, R. Hofmann-Wellenhof, L. Thomas, A. Lallas, A. Blum, W. Stolz, and H. Haenssle, "Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition," *JAMA Dermatology*, vol. 155, 08 2019.

[8] L. Oakden-Rayner, J. Dunnmon, G. Carneiro, and C. Ré, "Hidden stratification causes clinically meaningful failures in machine learning for medical imaging," *ACM Conference on Health, Inference, and Learning*, vol. 2020, pp. 151–159, 04 2020.

[9] Z. Eaton-Rosen, F. Bragman, S. Bisdas, S. Ourselin, and M. J. Cardoso, "Towards safe deep learning: Accurately quantifying biomarker uncertainty in neural network predictions," *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 691–699, 6 2018.

[10] A. Larrazabal, N. Nieto, V. Peterson, D. Milone, and E. Ferrante, "Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis," *Proceedings of the National Academy of Sciences*, vol. 117, p. 201919012, 05 2020.

[11] J. Rubin, D. Sanghavi, C. Zhao, K. Lee, A. Qadir, and M. Xu-Wilson, "Large Scale Automated Reading of Frontal and Lateral Chest X-Rays using Dual Convolutional Neural Networks," *arXiv e-prints*, p. arXiv:1804.07839, 4 2018.

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

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

[14] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLOS ONE*, vol. 10, pp. 1–46, 07 2015.

[15] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," *34th International Conference on Machine Learning (ICML)*, vol. 70, pp. 3145–3153, 04 2017.

[16] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic Attribution for Deep Networks," *34th International Conference on Machine Learning (ICML)*, vol. 70, pp. 3319–3328, 2017.

[17] S. M. Lundberg, P. G. Allen, and S.-I. Lee, "A unified approach to interpreting model predictions," *31st International Conference on Advances in Neural Information Processing Systems (NeurIPS)*, p. 4768–4777, 2017.

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

[19] P. Rajpurkar, J. A. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, M. Lungren, and A. Ng, "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning," *arXiv e-prints*, p. arXiv:1711.05225, 11 2017.

[20] K. Young, G. Booth, B. Simpson, R. Dutton, and S. Shrapnel, "Deep neural network or dermatologist?" *International Workshop on Interpretability of Machine Intelligence in Medical Image Computing (IMIMIC)*, vol. 11797 LNCS, pp. 48–55, 10 2019.

[21] F. Eitel and K. Ritter, "Testing the robustness of attribution methods for convolutional neural networks in mri-based alzheimer's disease classification," *International Workshop on Interpretability of Machine Intelligence in Medical Image Computing (IMIMIC)*, vol. 11797 LNCS, pp. 3–11, 10 2019.

[22] R. Sayres, A. Taly, E. Rahimy, K. Blumer, D. Coz, N. Hammel, J. Krause, A. Narayanaswamy, Z. Rastegar, D. Wu, S. Xu, S. Barb, A. Joseph, M. Shumski, J. Smith, A. B. Sood, G. S. Corrado, L. Peng, and D. R. Webster, "Using a Deep Learning Algorithm and Integrated Gradients Explanation to Assist Grading for Diabetic Retinopathy," *Ophthalmology*, vol. 126, pp. 552–564, 4 2019.

[23] P. Dabkowski and Y. Gal, "Real time image saliency for black box classifiers," *31st International Conference on Advances in Neural Information Processing Systems (NeurIPS)*, p. 6970–6979, 2017.

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

[25] R. C. Fong and A. Vedaldi, "Interpretable Explanations of Black Boxes by Meaningful Perturbation," *IEEE International Conference on Computer Vision (ICCV)*, 10 2017.

[26] C.-H. Chang, E. Creager, A. Goldenberg, and D. Duvenaud, "Explaining Image Classifiers by Counterfactual Generation," *International Conference on Learning Representations (ICLR)*, 2019.

[27] A. Dhurandhar, P.-Y. Chen, R. Luss, C.-C. Tu, P. Ting, K. Shanmugam, and P. Das, "Explanations based on the missing: Towards contrastive explanations with pertinent negatives," *International Conference on Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[28] P. Samangouei, A. Saeedi, N. Liam, and S. Nathan, "ExplainGAN: Model Explanation via Decision Boundary Crossing Transformations," *IEEE European Conference on Computer Vision (ECCV)*, pp. 681–696, 2018.

[29] S. Joshi, O. Koyejo, W. Vijitbenjaronk, B. Kim, and J. Ghosh, "Towards Realistic Individual Recourse and Actionable Explanations in Black-Box Decision Making Systems," *arXiv e-prints*, p. arXiv:1907.09615, 2019.

[30] S. Liu, B. Kailkhura, D. Loveland, and Y. Han, "Generative Counterfactual Introspection for Explainable Deep Learning," *IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pp. 1–5, 2019.

[31] D. Mahajan, C. Tan, and A. Sharma, "Preserving Causal Constraints in Counterfactual Explanations for Machine Learning Classifiers," *CausalML: Machine Learning and Causal Inference for Improved Decision Making Workshop, NeurIPS*, 12 2019.

[32] A. Van Looveren and J. Klaise, "Interpretable Counterfactual Explanations Guided by Prototypes," *arXiv e-prints*, p. arXiv:1907.02584, 2019.

[33] A. Parafita Martinez and J. Vitria Marca, "Explaining visual models by causal attribution," *IEEE International Conference on Computer Vision Workshop (ICCVW)*, pp. 4167–4175, 10 2019.

[34] C. Agarwal and A. Nguyen, "Explaining image classifiers by removing input features using generative models," *Asian Conference on Computer Vision (ACCV)*, 11 2020.

[35] P. Wang and N. Vasconcelos, "SCOUT: Self-Aware Discriminant Counterfactual Explanations," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6 2020.

[36] Y. Goyal, Z. Wu, J. Ernst, D. Batra, D. Parikh, and S. Lee, "Counterfactual Visual Explanations," *36th International Conference on Machine Learning (ICML)*, vol. 97, pp. 2376–2384, 2019.

[37] A. Narayanaswamy, S. Venugopalan, D. R. Webster, L. Peng, G. S. Corrado, P. Ruamviboonsuk, P. Bavishi, M. Brenner, P. C. Nelson, and A. V. Varadarajan, "Scientific Discovery by Generating Counterfactuals using Image Translation," *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 273–283, 7 2020.

[38] A. J. DeGrave, J. D. Janizek, and S.-I. Lee, "AI for radiographic COVID-19 detection selects shortcuts over signal," *medRxiv*, p. 10.1101/2020.09.13.20193565, 10 2020.

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

[40] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, and R. Sayres, "Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV)," *International Conference on Machine Learning (ICML)*, vol. 6, pp. 4186–4195, 11 2017.

[41] D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba, "Network dissection: Quantifying interpretability of deep visual representations," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6541–6549, 2017.

[42] B. Zhou, Y. Sun, D. Bau, and A. Torralba, "Interpretable basis decomposition for visual explanation," *European Conference on Computer Vision (ECCV)*, vol. 11212 LNCS, pp. 122–138, 2018.

[43] H. Yeche, J. Harrison, and T. Berthier, "UBS: A dimension-agnostic metric for concept vector interpretability applied to radiomics," *International Workshop on Interpretability of Machine Intelligence in Medical Image Computing (IMIMIC)*, pp. 12–20, 2019.

[44] M. Graziani, V. Andrearczyk, M. M. S., and H. Müller, "Concept attribution: Explaining CNN decisions to physicians," *Computers in Biology and Medicine*, vol. 123, p. 103865, 8 2020.

[45] J. R. Clough, I. Oksuz, E. Puyol-Antón, B. Ruijsink, A. P. King, and J. A. Schnabel, "Global and local interpretability for cardiac MRI classification," *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 656–664, 2019.

[46] S. Singla, B. Pollack, J. Chen, and K. Batmanghelich, "Explanation by Progressive Exaggeration," *International Conference on Learning Representations (ICLR)*, 2019.

[47] D. Bau, J.-Y. Zhu, J. Wulff, W. Peebles, H. Strobelt, B. Zhou, and A. Torralba, "Seeing what a gan cannot generate," *IEEE International Conference on Computer Vision (ICCV)*, pp. 4502–4511, 2019.

[48] T. Miyato and M. Koyama, "cGANs with Projection Discriminator," *International Conference on Learning Representations (ICLR)*, 2018.

[49] E. Frank and M. Hall, "A simple approach to ordinal classification," *European Conference on Machine Learning*, pp. 145–156, 2001.

[50] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," *IEEE Conference on Computer Vision Pattern Recoginition (CVPR)*, pp. 4401–4410, 2019.

[51] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *IEEE Conference on Computer Vision Pattern Recoginition (CVPR)*, 12 2016.

[52] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," *International Conference on Learning Representations (ICLR)*, 2 2018.

[53] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *International Conference on Learning Representation (ICLR)*, 12 2015.

[54] S. Joshi, O. Koyejo, B. Kim, and J. Ghosh, "xGEMs: Generating Examplars to Explain Black-Box Models," *arXiv e-prints*, p. arXiv:1806.08867, 2018.

[55] A. E. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C. Y. Deng, R. G. Mark, and S. Horng, "MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports," *Scientific data*, vol. 6, p. 317, 12 2019.

[56] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, and et al., "Chexpert: A large chest radiograph dataset

with uncertainty labels and expert comparison," *33rd AAAI Conference on Artificial Intelligence*, pp. 590–597, 2019.
[57] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," *30th IEEE Conference on Computer Vision Pattern Recoginition (CVPR)*, pp. 2261–2269, 8 2016.
[58] O. Ronneberger, P. Fischer, and T. Brox, "U-net convolutional networks for biomedical image segmentation," *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, vol. 9351, pp. 234–241, 2015.
[59] B. van Ginneken, M. B. Stegmann, and M. Loog, "Segmentation of anatomical structures in chest radiographs using supervised methods: a comparative study on a public database," *Medical Image Analysis*, vol. 10, pp. 19–40, 2006.
[60] S. Jaeger, S. Candemir, S. Antani, Y.-X. J. Wáng, P.-X. Lu, and G. Thoma, "Two public chest x-ray datasets for computer-aided screening of pulmonary diseases." *Quantitative imaging in medicine and surgery*, vol. 4, pp. 475–477, 12 2014.
[61] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *International Conference on Advances in Neural Information Processing Systems (NeurIPS)*, vol. 28, 2015.
[62] R. K. Mothilal, A. Sharma, and C. Tan, "Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations," *Conference on Fairness, Accountability, and Transparency (FAT)*, pp. 607–617, 5 2020.
[63] "Pleural effusion imaging: Overview, radiography, computed tomography." https://emedicine.medscape.com/article/355524-overview
[64] M. A. Iqbal and M. Gupta, "Cardiogenic pulmonary edema," *StatPearls*, 7 2020.
[65] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *International Conference on Advances in Neural Information Processing Systems (NeurIPS)*, 12 2017.
[66] F. Pasa, V. Golkov, F. Pfeiffer, D. Cremers, and D. Pfeiffer, "Efficient Deep Network Architectures for Fast Chest X-Ray Tuberculosis Screening and Visualization," *Scientific Reports*, vol. 9, pp. 1–9, 12 2019.
[67] V. Petsiuk, A. Das, and K. Saenko, "Rise: Randomized input sampling for explanation of black-box models," *British Machine Vision Conference (BMVC)*, 2018.
[68] Y. Mensah, K. Mensah, S. Asiamah, H. Gbadamosi, E. Idun, W. Brakohiapa, and A. Oddoye, "Establishing the Cardiothoracic Ratio Using Chest Radiographs in an Indigenous Ghanaian Population: A Simple Tool for Cardiomegaly Screening," *Ghana medical journal*, 9 2015.
[69] O. A. Centurión, K. Scavenius, L. Miño, and O. R. Sequeira, "Evaluating Cardiomegaly by Radiological Cardiothoracic Ratio as Compared to Conventional Echocardiography," *Journal of Cardiology & Current Research (JCCR)*, vol. 9, 6 2017.
[70] K. Dimopoulos, G. Giannakoulas, I. Bendayan, E. Liodakis, R. Petraco, G.-P. Diller, M. F. Piepoli, L. Swan, M. Mullen, N. Best, P. A. Poole-Wilson, D. P. Francis, M. B. Rubens, and M. A. Gatzoulis, "Cardiothoracic ratio from postero-anterior chest radiographs: A simple, reproducible and independent marker of disease severity and outcome in adults with congenital heart disease," *International Journal of Cardiology*, vol. 166, 6 2013.
[71] I. Chamveha, T. Promwiset, T. Tongdee, P. Saiviroonporn, and W. Chaisangmongkon, "Automated Cardiothoracic Ratio Calculation and Cardiomegaly Detection using Deep Learning Approach," *arXiv e-prints*, p. arXiv:2002.07468, 2 2020.
[72] P. Maduskar, L. Hogeweg, R. Philipsen, and B. van Ginneken, "Automated localization of costophrenic recesses and costophrenic angle measurement on frontal chest radiographs," *Medical Imaging: Computer-Aided Diagnosis*, 3 2013.
[73] P. Maduskar, R. H. Philipsen, J. Melendez, E. Scholten, D. Chanda, H. Ayles, C. I. Sánchez, and B. van Ginneken, "Automatic detection of pleural effusion in chest radiographs," *Medical Image Analysis*, 2 2016.
[74] H. Huang, Z. Li, R. He, Z. Sun, and T. Tan, "IntroVAE: Introspective Variational Autoencoders for Photographic Image Synthesis," *International Conference on Advances in Neural Information Processing Systems (NeurIPS)*, pp. 10 236–10 245, 2018.
[75] E. N. Milne, M. Pistolesi, M. Miniati, and C. Giuntini, "The radiologic distinction of cardiogenic and noncardiogenic edema," *American Journal of Roentgenology*, vol. 144, no. 5, pp. 879–894, 1985.
[76] D. M. Hansell, A. A. Bankier, H. MacMahon, and et al., "Fleischner Society: Glossary of terms for thoracic imaging," *Radiology*, 3 2008.
[77] K. Wada, "labelme Image Polygonal Annotation with Python," https://github.com/wkentaro/labelme, 2016.
[78] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *arXiv e-prints*, p. arXiv 1409.1556, 6 2014.

SUPPLEMENTARY MATERIAL

A. Summarization of the notation

Table. IV summarizes the notation used in the manuscript.

B. Implementation Details

1) 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 a large, publicly available datasets of chest x-ray images, MIMIC-CXR [55]. MIMIC-CXR dataset is a multi-modal dataset consisting of 473K chest X-ray images and 206K reports from 63K patients. We considered only frontal (posteroanterior PA or anteroposterior AP) view chest images. 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 [56]. The labeler first defines some thoracic observations using a radiology lexicon [76]. 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.

2) 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 pixels. The intensities were normalized to have values between 0 and 1. Following the approach in prior work [11], [19], [56] on diagnosis classification, we used DenseNet-121 [57] 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 in a post-hoc manner 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.

3) Explanation Function: The explanation function is a conditional GAN with an encoder. We used a ResNet [51] architecture for the Encoder, Generator, and Discriminator. The details of the architecture are given in Table V. 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. 12.a. For the generator network, we follow the details in [52] and replace the BN layer in encoder-ResBlock with conditional BN (cBN) to encode the condition (*see* Fig. 12.b.). The architecture for the generator have five ResBlocks, each ResBlock performed up-sampling through the nearest neighbor interpolator. For the

TABLE IV
SUMMARIZATION OF THE NOTATION

Notation	Description
$\mathcal{X}$	Input image space
$\mathbf{x} \in \mathcal{X}$	Input image
$f : \mathcal{X} \to \mathcal{Y}$	Pre-trained classification function
$f(\mathbf{x}) \in [0, 1]$	Classifier's output
δ	Desired change in classifier's output
$\mathbf{x}_\delta$	Explanation image
$f(\mathbf{x}_\delta)$	Classifier's output for the explanation image
$\mathcal{I}_f(\mathbf{x}, \delta)$	Explanation function
$E(\cdot)$	Image encoder
$\mathbf{z}$	Latent representation of the input image
$\mathbf{c}$	Condition for cGAN
$c_f(\mathbf{x}, \delta)$	Discretizing function that maps $f(\mathbf{x}) + \delta$ 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
$\phi(\mathbf{x}; \theta_\phi)$	Image feature extractor; part of the discriminator function
$\psi(\cdot)$	Loss function over image features

TABLE V
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, $\mathbf{y}$
Generator-ResBlock up 512, $\mathbf{y}$
Generator-ResBlock up 256, $\mathbf{y}$
Generator-ResBlock up 128, $\mathbf{y}$
Generator-ResBlock up 64, $\mathbf{y}$
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($\mathbf{y}$)
Inner Product (GSP, Embed($\mathbf{y}$)) $\to \mathbb{R}^1$
Add(SN-Dense(GSP) $\to \mathbb{R}^1$, Inner Product)

discriminator, we used spectral normalization (SN) [48] in Discriminator-ResBlock and performed down-sampling after the second *conv* of the ResBlock as shown in Fig. 12.c. For the optimization, we used Adam optimizer [53], 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 set hyper-parameter δ to a fix value and divide the posterior distribution for the target class, $f(\mathbf{x}) \in [0, 1]$ into $\lfloor \frac{1}{\delta} \rfloor$ equally-sized bins. The cGAN is then trained on $\lfloor \frac{1}{\delta} \rfloor$ conditions. For efficient training, cBN requires class-balanced batches. A smaller value for δ 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 lower value for δ. A large value of δ is equivalent to fewer bins, resulting in a coarse transformation which leads to abrupt changes across explanation images. In our experiments, we used $\delta = 0.1$, which is equivalent to ten bins with a batch size of 32. We experimented with different values of δ and selected the smallest δ, which created a class-balanced batch that fits in GPU memory and resulted in stable cGAN training.

4) Semantic Segmentation: We adopted a 2D U-Net [58] to perform semantic segmentation, to mark the lung and the heart contour in a chest x-ray. The network optimizes a multi-categorical cross-entropy loss function, defined as,

$$\mathcal{L}_\theta := \sum_s \sum_i \mathbb{1}(y_i = s) \log(p_\theta(x_i)), \tag{10}$$

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 θ are the learned parameters. The network is trained on 385 chest x-rays and corresponding masks from Japanese Society of Radiological Technology (JSRT) [59] and Montgomery [60] datasets.

5) Object Detection: We trained an object detector network to identify medical devices in the chest x-ray. 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 chest x-ray and do not requires expert knowledge for manual labeling. Using the CheXpert labeler, we extracted 300 chest x-rays images with positive mentions for each observation. The extracted x-rays are then manually annotated with bounding box an-

notations marking the presence of foreign objects using the LabelMe [77] annotation tool. Next, we trained an object detector based on fFast Region-based CNN [61], which used VGG-16 model [78], trained on MIMIC-CXR dataset as its foundation. We used this object detector to enforce our novel context-aware reconstruction loss (CARL).

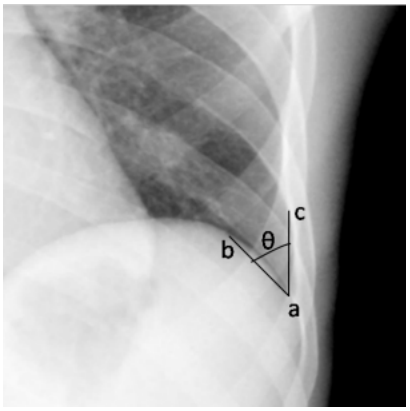


Fig. 11. The costophrenic angle (CPA) on a chest x-ray is marked as the angle formed by, (a) costophrenic angle point, (b) hemidiaphragm point and (c) lateral chest wall point, as shown by Maduskar *et al.*in [73]

We trained similar detectors for identifying normal and abnormal CP recess region in the chest x-ray. 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. To train the object detector we extracted 300 chest x-rays with positive mention of respective terms for normal and abnormal CP recess.

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 [72], [73]. The angle between the chest wall and the diaphragm arc is called costophrenic angle (CPA). Marking the CPA angle on a chest x-ray 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. 11. Learning automatic marking of CPA angle requires expert annotation and is prone to error. Hence, rather than marking 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 the chest x-rays and used the Score for detecting a normal CP recess (SCP) as our evaluation metric.

6) xGEM: We refer to work by Joshi *et al.* [29] 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

$$\tilde{\mathbf{x}} = \mathcal{G}_\theta(\arg\min_{\mathbf{z}\in\mathbb{R}^d} \mathcal{L}(\mathbf{x}, \mathcal{G}_\theta(\mathbf{z})) + \lambda\ell(f(\mathcal{G}_\theta(\mathbf{z})), y^{'})) \qquad (11)$$

where the first terms is an ℓ_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 (ℓ_2 reconstruction), an unrealistic assumption for image data. Hence, vanilla VAE is known to produce over-smoothed images [74]. 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.

7) cycleGAN: We refer to the work by Narayanaswamy *et al.* [37] and DeGrave *et al.* [38] 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 strong 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.

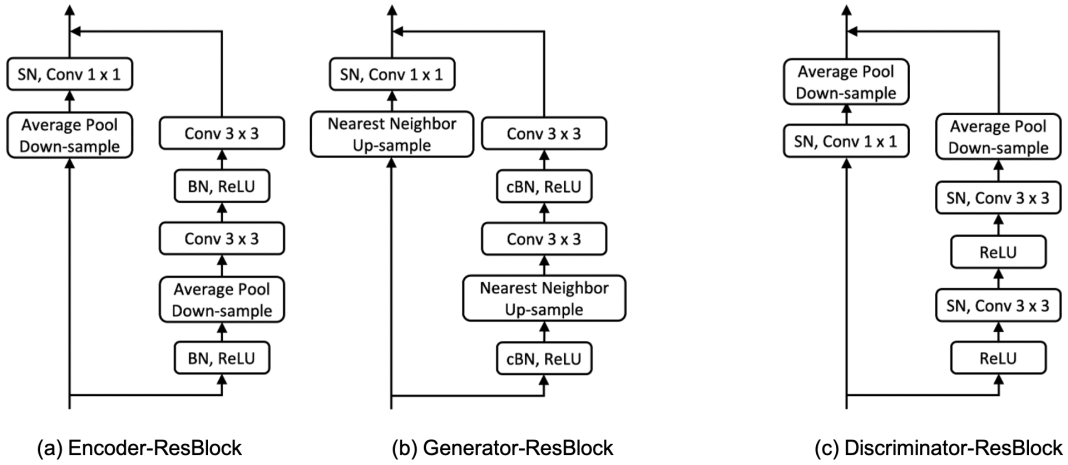
TABLE VI
COMPARISON OF OUR METHOD AGAINST XGEM AND CYCLEGAN ON ESSENTIAL PROPERTIES OF A COUNTERFACTUAL EXPLANATION.

Method	**Realistic-looking**	**Flipping classification decision**
Ours	✓	✓
xGEM	×	✓
cycleGAN	✓	×

C. Extended data consistency results

A counterfactual explanation is a perturbation of input image such that the decision of the classifier is flipped. For example, consider a chest x-ray with a positive classification decision for cardiomegaly. A counterfactual explanation provides *what-if* scenario such that a minimal but perceptible modification is applied to the x-ray image (*what*), and the resulting image is negative for cardiomegaly. To create such an explanation, a counterfactual should satisfy two essential properties; first, a counterfactual should be very similar to the input. Second, the counterfactual should produce an opposite outcome when processed by the classifier.

In Fig. 13 and Fig. 14, we show the results to visualize our explanations and compared it against xGEM and cycleGAN method. The results are an extension of main-Fig.6. We can observe the explanation images generated by xGEM are blurred and lacks the realistic-looking appeal of an x-ray image. Consistent with this observation, earlier in our results Table. 1, xGEM has a high FID *i.e.,* the explanation images are significantly different from the real x-ray images. The bottom labels in Fig. 13 are the classifier's prediction for the specific disease. For cycleGAN, the results demonstrate an example



(a) Encoder-ResBlock (b) Generator-ResBlock (c) Discriminator-ResBlock

Fig. 12. Architecture of the ResBlocks used in all experiments.

TABLE VII
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% Confidence Interval Lower	Upper	t	df	p-value
Cardiomegaly	$\mathcal{X}^n$	$\mathcal{X}^a_{cf}$	**-0.03**	0.07	-0.03	-0.01	-4.4	304	< 0.0001
(CTR)	$\mathcal{X}^a$	$\mathcal{X}^n_{cf}$	**0.14**	0.12	0.13	0.15	24.7	513	$\ll 0.0001$
Pleural effusion	$\mathcal{X}^n$	$\mathcal{X}^a_{cf}$	**0.13**	0.22	0.06	0.13	5.9	217	$\ll 0.0001$
(SCP)	$\mathcal{X}^a$	$\mathcal{X}^n_{cf}$	**-0.19**	0.27	-0.18	-0.09	-6.7	216	$\ll 0.0001$
			Un-Paired Differences Mean Real Group	Mean Counterfactual Group	95% Confidence Interval Lower	Upper	t	df	p-value
Cardiomegaly	$\mathcal{X}^n$	$\mathcal{X}^n_{cf}$	**0.46**	0.42	0.02	0.06	5.2	817	< 0.0001
(CTR)	$\mathcal{X}^a$	$\mathcal{X}^a_{cf}$	**0.56**	0.50	0.04	0.07	9.9	817	$\ll 0.0001$
Pleural effusion	$\mathcal{X}^n$	$\mathcal{X}^n_{cf}$	**0.69**	0.61	0.18	0.27	9.3	433	$\ll 0.0001$
(SCP)	$\mathcal{X}^a$	$\mathcal{X}^a_{cf}$	0.42	**0.56**	-0.32	-0.21	-9.7	433	$\ll 0.0001$

where the counterfactual image doesn't have an opposing prediction as compared to the input image. In Fig. 13, in cardiomegaly and edema the counterfactual image obtained by cycleGAN have almost the same prediction ($f(\mathbf{x}_\delta) < 0.5$) as compared to input normal x-ray ($f(\mathbf{x}) < 0.5$). Overall, this finding is consistent with the low counterfactual validity score in Table. 1. We summarize the comparison between three methods in Table VI. As compared to our model, both, xGEM and cycleGAN failed on atleast one essential property of a valid counterfactual explanation.

Next, we quantify the consistency between our explanations and the classification model at every step of the transformation. We generated multiple, progressively changing explanations for xGEM 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. In Fig. 15, we plotted the average response of the classifier *i.e.,* $f(\mathbf{x}_\delta)$ for explanations in each bin against the expected classification outcome *i.e.,* $f(\mathbf{x})+\delta$. The figure shows an extension of the results in main-Fig.7. The positive slope of the line-plot, parallel to $y = x$ line at 45° confirms that starting from images with low $f(\mathbf{x})$, our model creates fake images such that $f(\mathbf{x}_\delta)$ 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 $f(\mathbf{x}_\delta) < 0.5$, hence, cycleGAN model fails to flip the classification decision, as also evident in low CV score in main-Table.1.

Further, we provide addition plots with different measurements on each axis. In Fig 16, we plot $f(\mathbf{x})$ versus $f(\mathbf{x}_\delta) - \delta$ and color each point based on δ. Ideally, $f(\mathbf{x}_\delta) - \delta \approx f(\mathbf{x})$. Our model achieved maximum r^2 coefficient for regression. When $f(\mathbf{x})$ is large, δ is mostly negative (darker shades) and vice-versa. For cardiomegaly and edema, xGEM achieves a progressive transformation, but it doesn't cover the entire prediction range.

D. Evaluating class discrimination

In multi-label settings, multiple labels can be true for a given image. A multi-label setting is common in chest x-ray diagnosis. For example, cardiomegaly and pleural effusion are associated with cardiogenic edema and frequently co-occur in a chest x-ray. Please note that our classification model is also trained in a multi-label setting where the fourteen radiological findings may co-occur in a chest x-ray. In this evaluation,

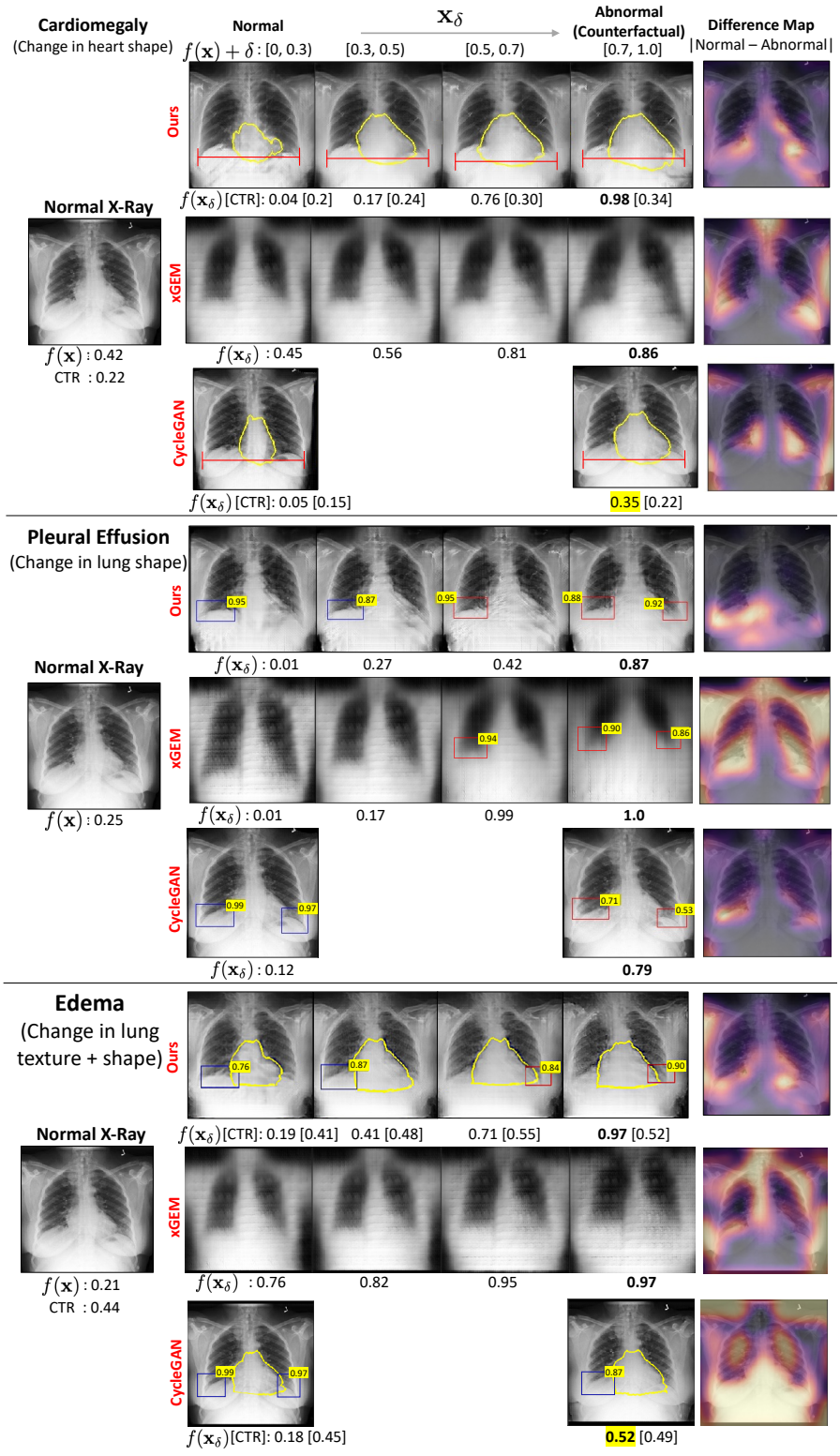


Fig. 13. The transformation of a normal chest x-ray into the counterfactual explanations for three classes, cardiomegaly (first row), pleural effusion (PE) (middle row) and edema (last row). The bottom labels are the classifier's prediction for the specific class. The yellow color highlight the prediction where counterfactual fails to flip the decision. The last column shows the difference map between normal and abnormal explanation. For cardiomegaly and edema, we are reporting cardio thoracic ratio (CTR) calculated from the heart segmentation (yellow) and thoracic diameter (red). For PE and edema, we show the bounding-box (BB) for normal (blue) and abnormal (red) costophrenic (CP) recess. The number on blue-BB is the Score for detecting a normal CP recess (SCP). The number on red-BB is 1-SCP.

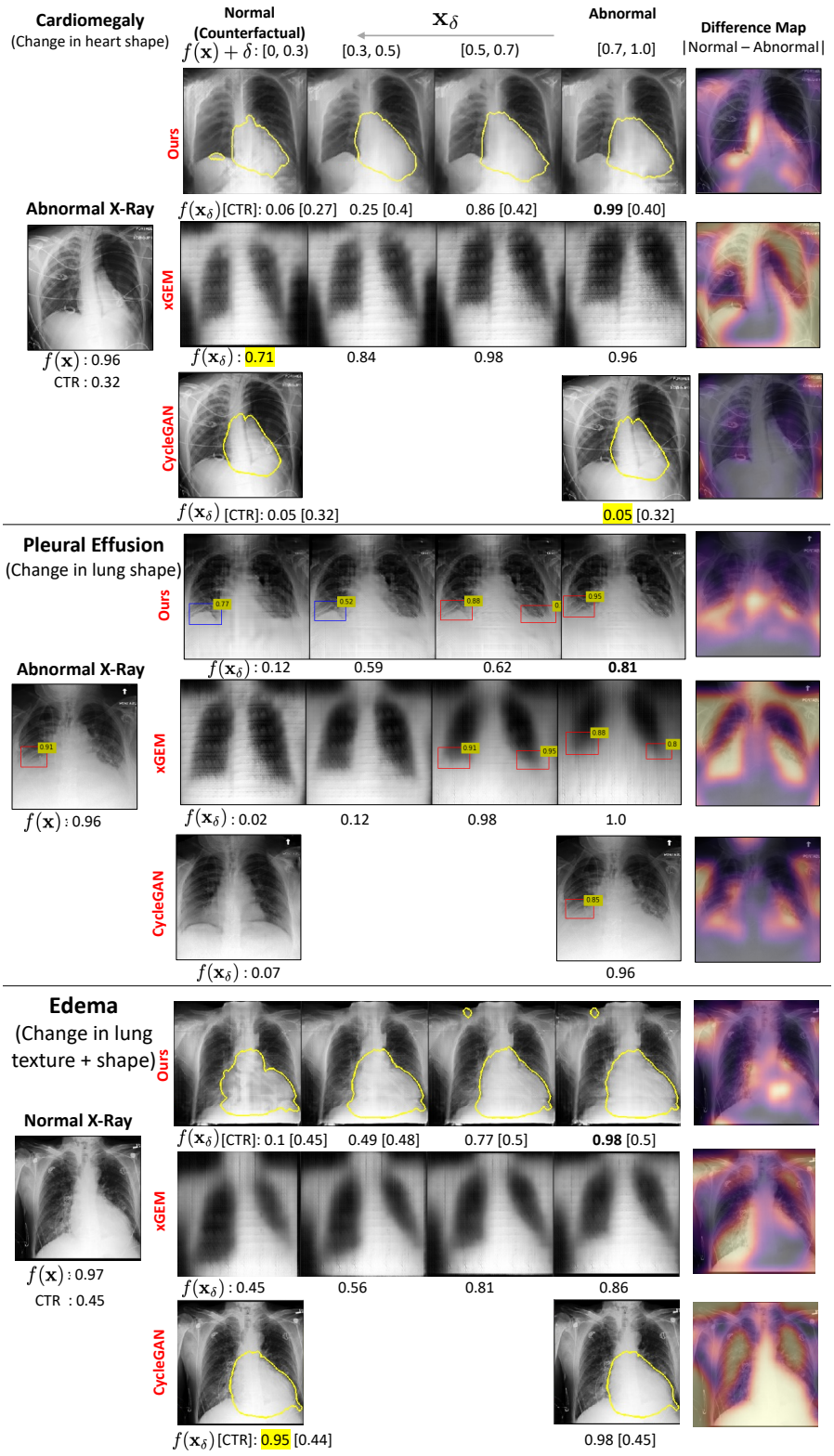


Fig. 14. The transformation of an abnormal chest x-ray into the counterfactual explanations for three classes, cardiomegaly (first row), pleural effusion (PE) (middle row) and edema (last row). The bottom labels are the classifier's prediction for the specific class. The yellow color highlight the prediction where counterfactual fails to flip the decision. The last column shows the difference map between normal and abnormal explanation. For cardiomegaly and edema, we are reporting cardio thoracic ratio (CTR) calculated from the heart segmentation (yellow) and thoracic diameter (red). For PE and edema, we show the bounding-box (BB) for normal (blue) and abnormal (red) costophrenic (CP) recess. The number on blue-BB is the Score for detecting a normal CP recess (SCP). The number on red-BB is 1-SCP.

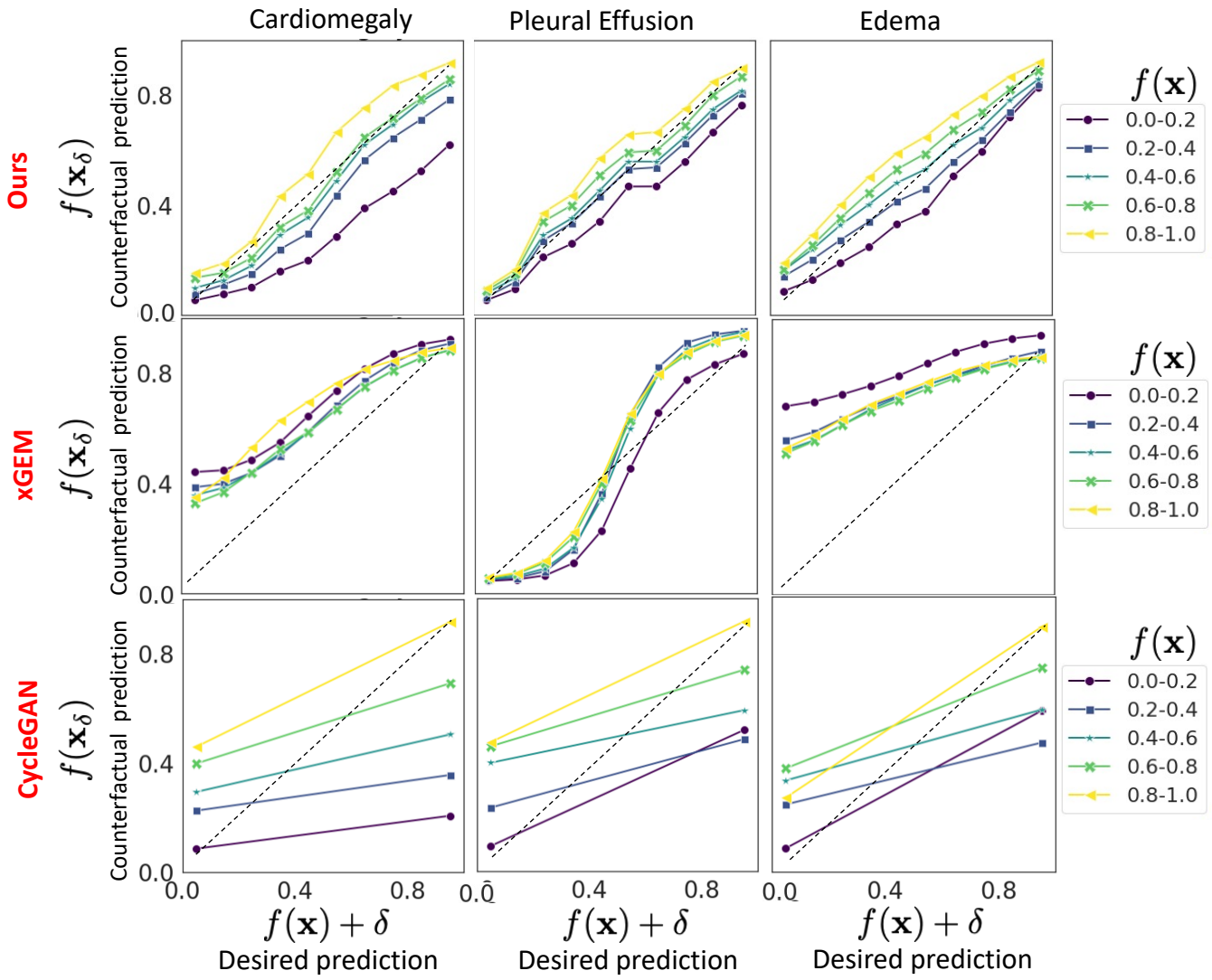


Fig. 15. The plot of desired outcome, $f(\mathbf{x}) + \delta$, against actual response of the classifier on generated explanations, $f(\mathbf{x}_\delta)$. Each line represents a set of input images with classification prediction $f(\mathbf{x})$ in a given range. Dashed line represents $y = x$ line.

we demonstrate the sensitivity of our generated explanations to the class being explained. We considered three classes, or diseases, cardiomegaly, pleural effusion, and edema. For each target class, we trained one explanation model. Ideally, an explanation model trained to explain a target class should produce explanations consistent with the query image on all the other classes besides the target. Fig. 17 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 target class and small for the rest of the classes. In Fig. 17, each column represents one class, and each row is one run of our method to explain a given target class. The diagonal values also represent the counterfactual validity (CV) score reported in main-Table.1.

E. 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.

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. 20. 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. We have added a new section in supplementary material to discuss this experiment.

Next, we present the extended results for the identity preservation experiment. We calculated the FO preservation (FOP) score to demonstrate the importance of CARL loss in

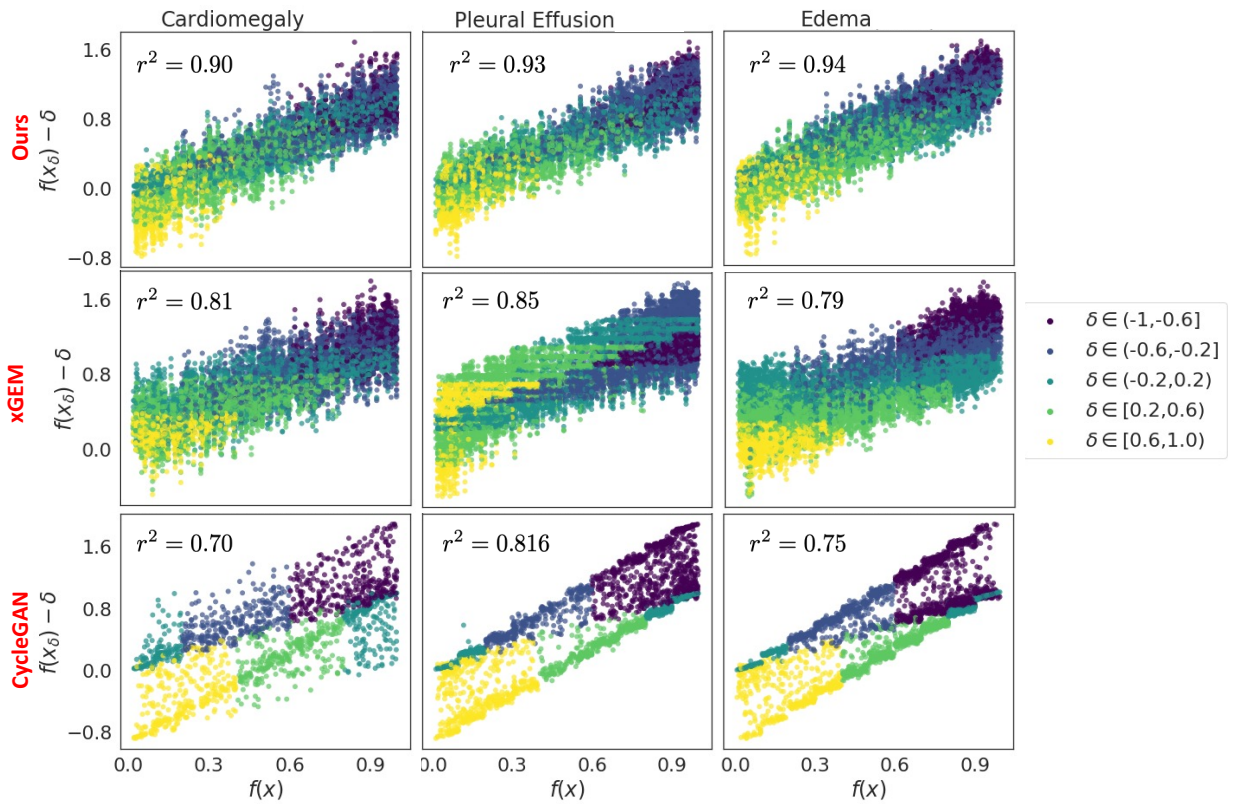


Fig. 16. The plot between prediction decision for the fake explanation images $f(\mathbf{x}_\delta)$ minus delta and prediction decision for the real images $f(\mathbf{x})$. Ideally $f(\mathbf{x}_\delta) - \delta \approx f(\mathbf{x})$.

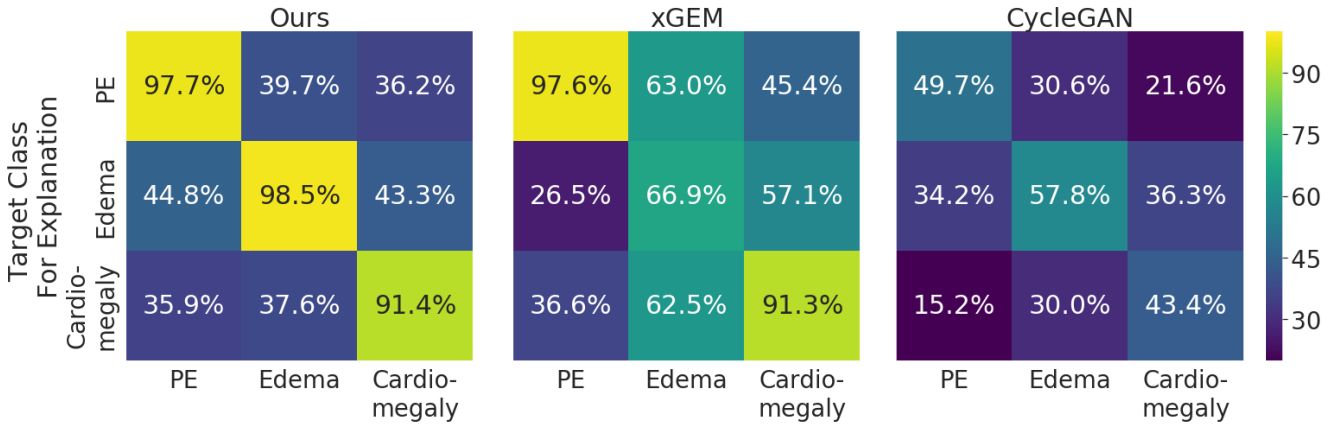


Fig. 17. Each cell is the fraction of the generated explanations, that have flipped in a class as compared to the query image. The x-axis shows the classes in a multi-label setting, and the y-axis shows the target class for which an explanation is generated. Note: This is not a confusion matrix.

preserving patient-specific details such as a pacemaker. We considered real images with successful detection of FO and reported the FOP score as the fraction of these images in which FO was also detected in the corresponding CE. In Table VIII, we provide FOP score our method and cycleGAN method. CycleGAN is good at preserving small details in the explanation images, as evident in its high FOP score. But in previous experiments, we have shown that even though images created by cycleGAN are the most realistic (with the lowest FID), they are not valid counterfactuals as they fail to flip the classification decision with a low CV score.

TABLE VIII
THE FOREIGN OBJECT PRESERVATION (FOP) SCORE FOR DIFFERENT MODELS. FOP SCORE DEPENDS ON THE PERFORMANCE OF FOREIGN OBJECT DETECTOR.

Foreign Object	**Ours** with CARL	**Ours** with ℓ_1	**CycleGAN**
Pacemaker	0.52	0.40	0.91
Hardware	0.63	0.32	0.89

F. Extended results for saliency maps

Our method doesn't produce a saliency map by default. We approximated a saliency map as an absolute difference

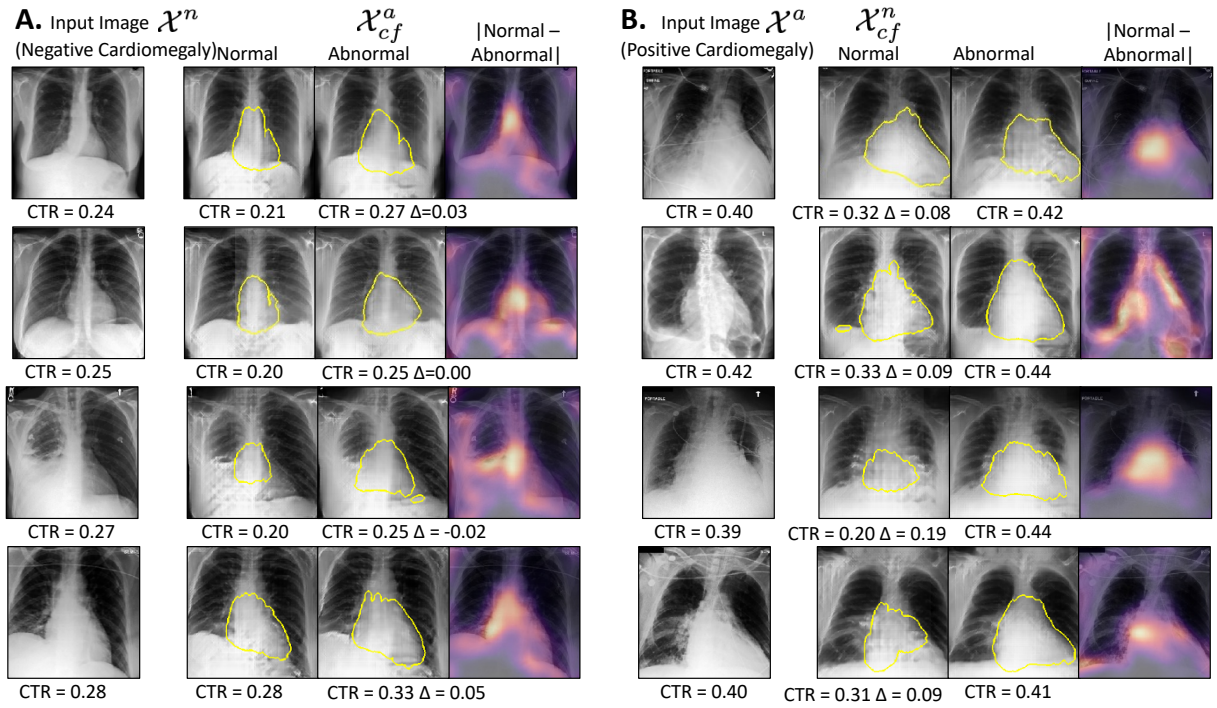


Fig. 18. Extended results for explanation produced by our model for **Cardiomegaly**. For each image, we generate a normal and an abnormal explanation image. We show pixel-wise difference of the two generated images as the saliency map. In column A.(B.), we show input images negatively (positively) classified 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.

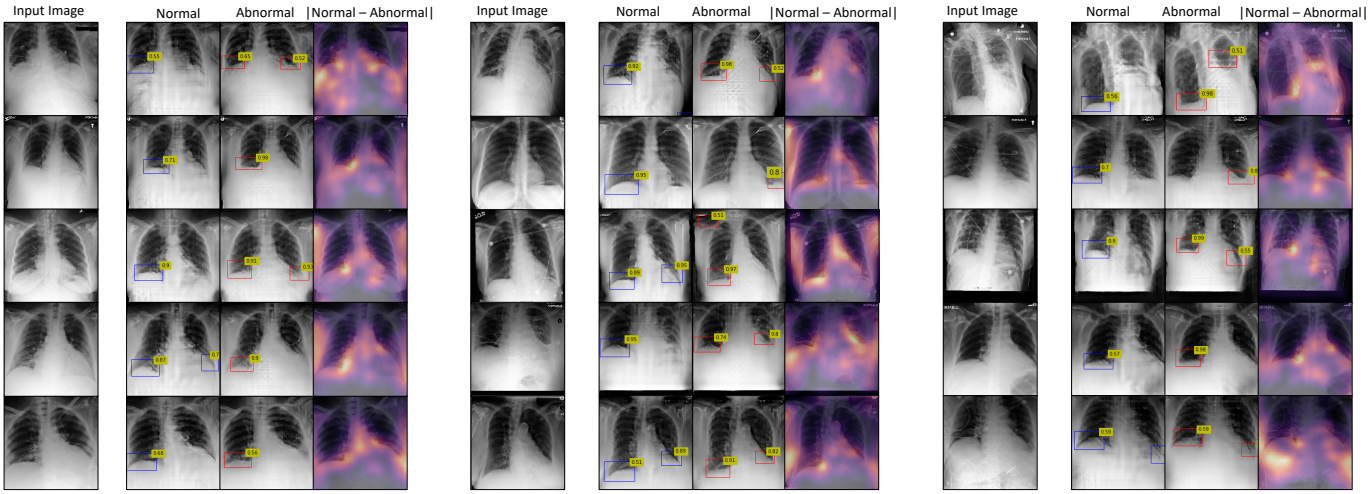


Fig. 19. Extended results for explanation produced by our model for **Pleural Effusion.** For each input image, we produce a normal and abnormal image as an explanation and take their pixel-wise difference to extract the saliency map.

map between the explanations generated for the two extremes (normal with $f(\mathbf{x}_\delta) < 0.2$ and abnormal $f(\mathbf{x}_\delta) > 0.8$) of the decision function f. In Fig. 18, we showed the two extreme explanation images and the corresponding difference map, derived for input images shown in the column A and B. We highlight the heart contour in yellow and reported CTR values. The difference maps mostly highlight the heart region for cardiomegaly. Next, we show extended results for pleural effusion (PE). For PE, our difference map highlights the CP recess region, as shown in Fig, 19.

Further, we used the *deletion* evaluation metric to quantitatively compare the saliency maps generated from gradient-based methods against our difference map [67]. 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. 21, we show an example of deletion-by-impainting. For generating results in main-Table.3, 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

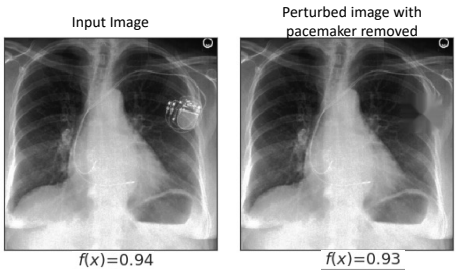


Fig. 20. An example of input image before and after removing the pacemaker.

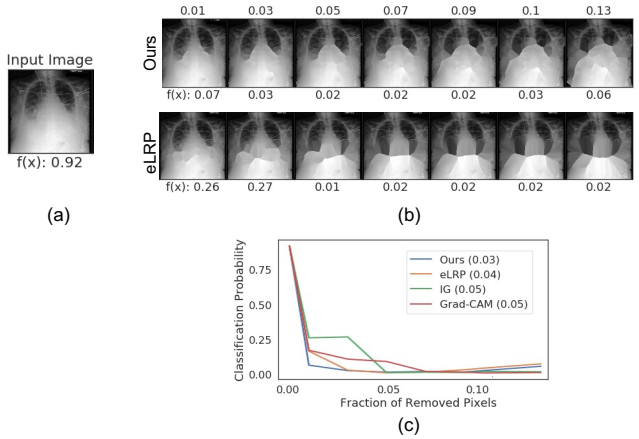


Fig. 21. Deletion-by-impainting: (a) input image. (b) transformation of the input image as important pixels are deleted, and the resulting patches are in-filled base on the surrounding context. The importance is derived from the saliency map produced from our (top-row) and gradient-based (bottom-row) 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.

a chest x-ray. 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.

G. 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 chest x-rays that are predicted as normal by the classifier f. (2) A *real-abnormal* group defined as $\mathcal{X}^a = \{\mathbf{x}; f(\mathbf{x}) > 0.8\}$. For $\mathcal{X}^n$ we generated a counterfactual group as, $\mathcal{X}^a_{cf} = \{\mathbf{x} \in \mathcal{X}^n; f(\mathcal{I}_f(\mathbf{x}, \delta)) > 0.8\}$. Similarly for $\mathcal{X}^a$, we derived a counterfactual group as $\mathcal{X}^n_{cf} = \{\mathbf{x} \in \mathcal{X}^a; f(\mathcal{I}_f(\mathbf{x}, \delta)) < 0.2\}$.

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 $\mathcal{X}^a_{cf}$), ($\mathcal{X}^a$ and $\mathcal{X}^n_{cf}$) and the independent two-sample t-test statistics for normal ($\mathcal{X}^n$, $\mathcal{X}^n_{cf}$) and abnormal ($\mathcal{X}^a$, $\mathcal{X}^a_{cf}$) 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 ($\mathbf{x}_\delta$), after intervention.

Table VII provides the extended results for the Fig. 10. Patients with cardiomegaly have higher CTR as compared to normal subjects. Hence, one should expect CTR($\mathcal{X}^n$) $<$ CTR($\mathcal{X}^a_{cf}$) and likewise CTR($\mathcal{X}^a$) $>$ CTR($\mathcal{X}^n_{cf}$). Consistent with clinical knowledge, in Table. VII, we observe a negative mean difference of -0.03 for CTR($\mathcal{X}^n$) $-$ CTR($\mathcal{X}^a_{cf}$) (a p-value of < 0.0001) and a positive mean difference of 0.14 for CTR($\mathcal{X}^a$) $-$ CTR($\mathcal{X}^n_{cf}$) (with a p-value of $\ll 0.0001$). On a population-level CTR was successful in capturing the difference between normal and abnormal chest x-rays. 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($\mathcal{X}^a$) = 0.56. The low p-values supports the alternate hypothesis that the difference in the two groups is statistically significant.

Further, in Fig 18.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($\mathcal{X}^a_{cf}$) $= -0.03$. In contrast, when we generate counterfactual normal (sixth column) from real abnormal images (positive for cardiomegaly, Fig 18.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($\mathcal{X}^n_{cf}$) $= 0.14$.

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($\mathcal{X}^a_{cf}$) and likewise SCP($\mathcal{X}^a$) $<$ SCP($\mathcal{X}^n_{cf}$). Consistent with our expectation, in Table. VII, we observe a positive mean difference of 0.13 for SCP($\mathcal{X}^n$) $-$ SCP($\mathcal{X}^a_{cf}$) (with a p-value of $\ll 0.0001$) and a negative mean difference of -0.19 for SCP($\mathcal{X}^a$) $-$ SCP($\mathcal{X}^n_{cf}$) (with a p-value of $\ll 0.0001$). On a population-level SCP was successful in capturing the difference between normal and abnormal chest x-rays 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($\mathcal{X}^a$) = 0.42. A low p-value confirmed the statistically significant difference in SCP for real images and their corresponding counterfactuals.