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 is relevant to the outcome, which is crucial for a 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. It is a visual explanation derived by gradually transforming the input image into its perturbation, where the model's decision has flipped.

There are two general approaches towards a model explanation: (1) developing *interpretable* models, (2) *post-hoc explaining* a pre-trained model. An interpretable model typically imposes certain simplifications such as, linear architecture, a limited number of rules in decision trees [12] or rule-based models [13] to ensure interpretability. Such models often compromise on predictive accuracy for interpretability. Post-hoc *explanation* aims to improve human understanding of a pre-trained model. Hence, the performance of the model is not compromised. Post-hoc explanation comprises of several broad approaches, such as approximating with simpler models [14], [15], understanding feature attribution [16] or importance [17], [18], to provide a local (image-level) or a global (target label-level) perspective on the decision. Most post-hoc methods produce a saliency map as an explanation. This paper focuses on a post-hoc explanation, which learns a function for a given target label and uses it to generate a local counterfactual explanation for an input image.

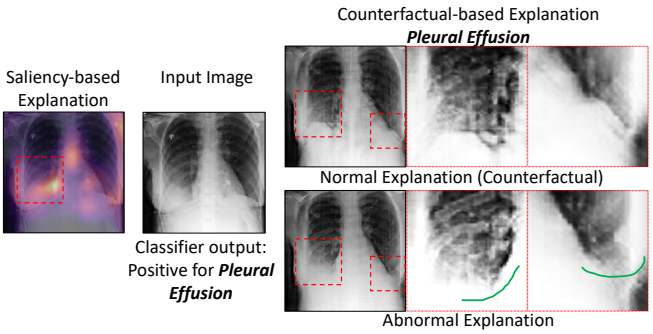

Fig. 1. Counterfactual explanation shows where" in the image the classifier is paying attention and "what" image-features in those regions are associated with the disease. For Pleural Effusion, we can observe the appearance of the meniscus (green) in abnormal image as compared the normal counterfactual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

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).

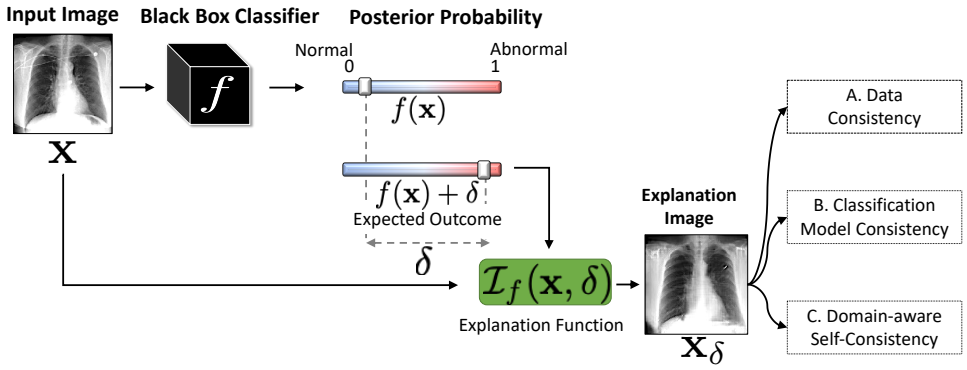

Fig. 2. Explanation function $\mathcal{I}_f(\mathbf{x},\delta)$ for classifier f. Given an input image $\mathbf{x}$, we generates a perturbation of the input, $\mathbf{x}_\delta$ as explanation, such that the posterior probability, f, changes from its original value, $f(\mathbf{x})$, to a desired value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

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 [19]. However, using cGAN is challenging, as GANs with an encoder may ignore small or uncommon details during image generation [20]. 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 counterfactual visual explanation for a black-box classifier. Our conditional GAN-based approach generates 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 developed quantitative metric based on clinical knowledge for 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.

In the following sections, 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) [21], with δ as the condition. The encoder enables transformation of a given image, while the GAN framework allows to generate realistic looking transformations as explanation image. GANs are implicit generative model, which learns the underlying data distribution $p_{\text{data}}(\mathbf{x})$ without explicitly parameterizing it. The cGAN is a variant of GAN that allows the conditional generation of the data by incorporating extra information as the context. Similar to 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 $E(\cdot)$.

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,

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

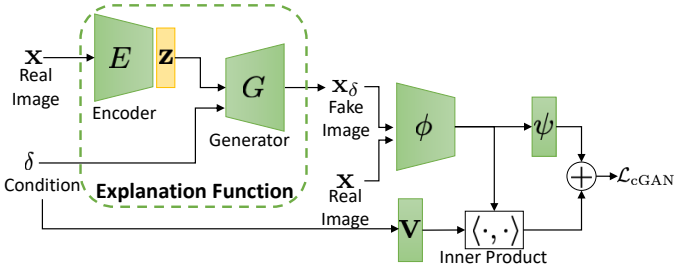

Fig. 3. The explanation function is a conditional-GAN with an encoder. The discriminator evaluates the similarity between real and fake data, and the correspondence between fake data and the condition.

For the discriminator in cGAN, we adapted the loss function from Projection GAN [21] 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.

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 [22]. The ith binary attribute represents the test $i < n$ where $\mathbf{c} = n$. The modified loss function is as follows:

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

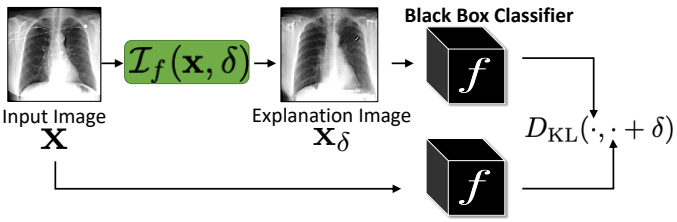

Fig. 4. To enforce consistency with the classifier f, we minimize a KullbackLeibler (KL) divergence between the actual $\boldsymbol{f}(\mathbf{x}_\delta)$ and the desired $\boldsymbol{f}(\mathbf{x}) + \boldsymbol{\delta}$ prediction from $\boldsymbol{f}$. $\boldsymbol{\mathcal{I}_f(\mathbf{x}, \delta)}$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

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 is a function of both G and D, the second term influences only the G.

C. Context-aware self consistency

A valid explanation image is a small modification of the input image, and should preserve the inputs' identity *i.e.,* patient-specific information such as the shape of the anatomy. While images generated by a GAN is shown to be realistic looking [23], GAN with an encoder may ignore small or uncommon details in the input image [20]. 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 extra information comes in the form of 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. $O(\cdot)$ is a pre-trained object detector to identify the presence of foreign objects in the input domain. Rather than minimizing a distance such as ℓ_1 over the entire image, we minimize the reconstruction loss for each segmentation-label

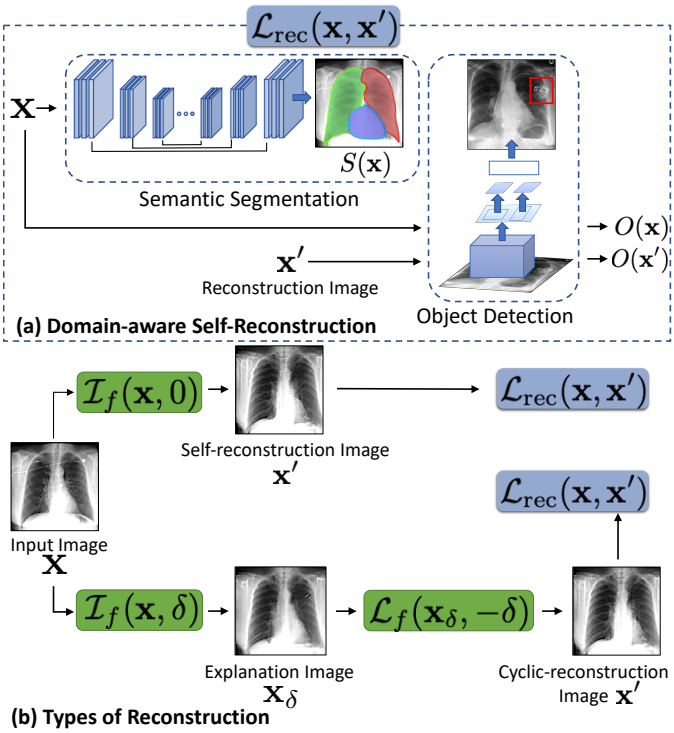

Fig. 5. (a) A domain-aware self-reconstruction loss with pre-trained semantic segmentation $\boldsymbol{S}(\mathbf{x})$ and object detection $\boldsymbol{O}(\mathbf{x})$ networks. (b) The self and cyclic reconstruction should retain maximum information from $\mathbf{x}$. Note, explanation image $\mathbf{x}_\delta$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

(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 for reconstruction of 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 generator for cGAN follows ResNet [24] architecture, with conditional batch normalization to encode the condition. We used hinge version of the adversarial loss for cGAN, trained using the Adam optimizer [25]. The model is trained end-to-end to generate explanation images.

III. Related Work

Our work broadly relates to model interpretation that generate post-hoc explanations of black-box classifiers.

1) Feature Attribution Explanations: Feature attribution methods provides an explanation as a saliency map that reflects the importance of each input component (*e.g.,* pixel) to the classification decision. *Gradient-based* approaches [16]–[18], [26]–[29] 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 [30], skin imaging [31], brain MRI [32] and retinopathy [33]. Saliency maps lack a clear interpretation and provide incomplete explanation especially when different diagnoses affect the same regions of the anatomy. In contrast our counterfactual explanation highlights important regions and also explains *what* imaging features present in those locations lead to a particular decision and *how* we can modify the features to change the decision.

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 [34] or with random noise, occluding [35], localized blurring [36], or in-filling [37]. The resulting explanation is a pixel- or patch-level manipulation of the input image, and is not a natural-looking image. Especially for medical images, such perturbations may introduce anatomically implausible features or textures. Our explanation framework enforces a consistency between the perturbed data and the real data distribution to ensure that the perturbation is plausible and realistic-looking.

2) Counterfactual Explanations: Counterfactual explanations are a type of contrastive [38] 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 [39]–[45]. Much of this work is limited to simpler image datasets like MNIST, celebA [41]–[43] or simulated data [44]. For more complex natural images, previous studies [37], [45] focused on finding and in-filling salient regions, in order to generate counterfactual images. In contrast, at inference time, our explanation model doesn't require any re-training for generating explanations for a new image. In another line of work [46], [47] provide counterfactual explanations that explains both the predicted and the counter class. Recently [48], [49] used a cycle-GAN [50] to perform image-to-image translation between normal and abnormal images. Those methods are not constrained by the classifier. Hence, cycle-GAN may end up learning features that do not reflect the true behaviour of the classifier. In contrast, our model uses special loss to enable image perturbation that is consistent with the classifier.

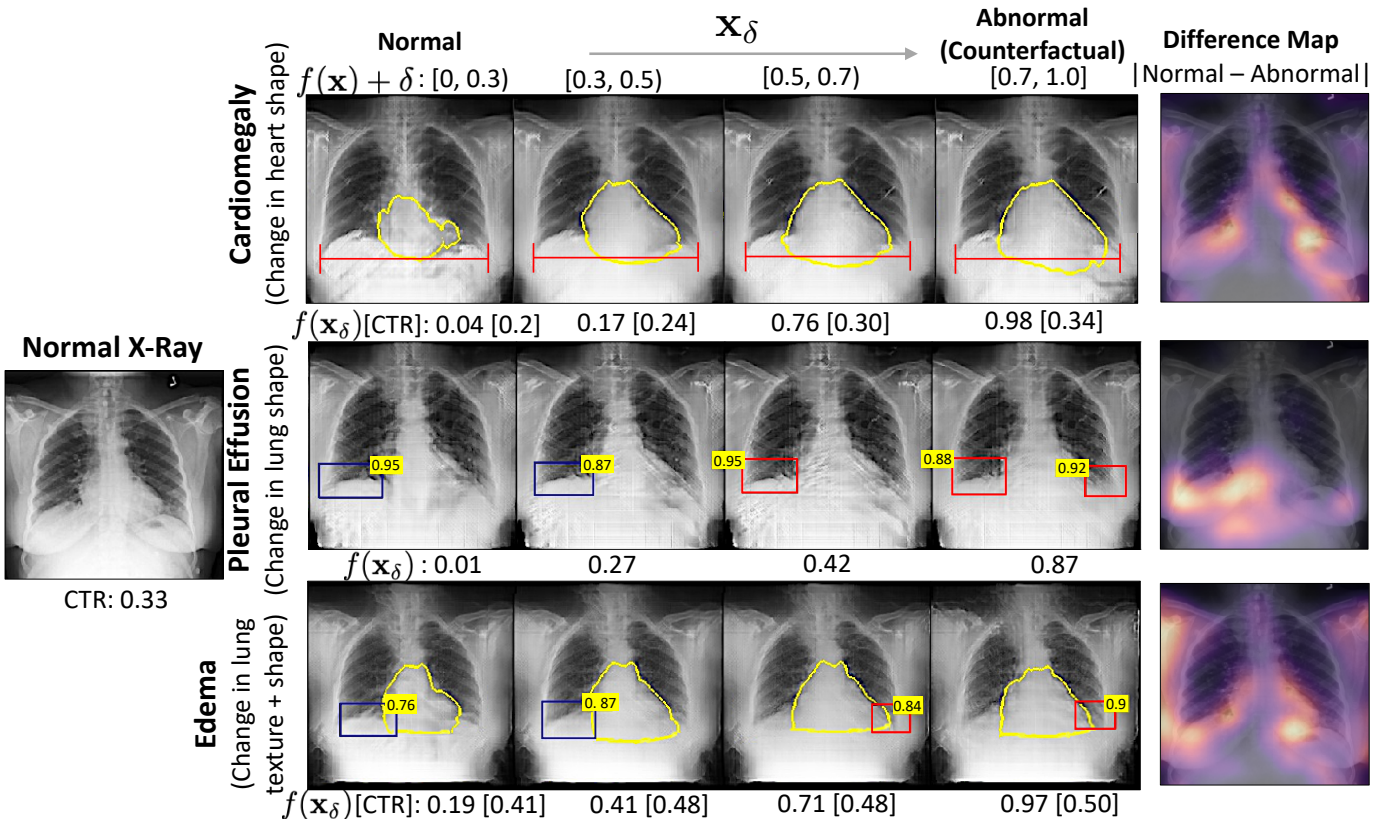

Fig. 6. Qualitative comparison of the counterfactual explanations generated for three classes, cardiomegaly (first row), pleural effusion (PE) (middle row) and edema (last row). The bottom labels are the classifier's prediction for the specific class. The last column shows the difference map between 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. Corresponding counterfactual explanations for xGEM and cycleGAN are shown in SM-Fig. 2.

IV. 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 [51] and CycleGAN [48], [49]. We considered 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 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 [52], which is a multi-modal dataset consisting of 473K chest X-ray images and 206K reports from 63K patients. The images are preprocessed using a standard pipeline involving cropping, re-scaling and intensity normalization. Following the prior work on diagnosis classification [53], we used DenseNet-121 [54] architecture as the classification model. It is trained to perform multi-label classification for fourteen radio-graphic observations, given the frontal view chest x-ray images. For semantic segmentation, we adopted a 2D U-Net [55] 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) [56] and Montgomery [57] datasets.

We trained a faster regional CNN [58] network, for detecting foreign objects such as pacemaker and hardware in a chest x-ray. The network learns to detect foreign objects by placing a bounding box over them. To create the training dataset, we extracted 300 x-rays with a positive mention of these objects in the corresponding radiology reports, and collected bounding box annotations to mark the ground truth. We further trained two more detectors to evaluate our explanations. Specifically, we trained detectors for identifying normal and abnormal costophrenic (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 the costophrenic angle"* in the corresponding radiology report. For the normal-CP recess, we considered

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 AS 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

images with a positive mention for *"lungs are clear"* in the reports. The detailed architecture for all modules can be found in the **Supplementary Material** (SM).

A. Desiderata of explanation function

1) Data consistency: Given an input image, our model generates a series of images $\mathbf{x}_\delta$ as explanations, by gradually changing $f(\mathbf{x}) + \delta$ 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 (first row), PE (middle row), and edema (last row). 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 [59]. 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 along with blurring of left CP recess region, as highlighted in the difference map. These findings are consistent with radiological signs for cardiogenic edema [60]. A comparison with counterfactual explation from xGEM and cycleGAN is shown in SM-Fig. 2.

Quantitatively evaluation: We evaluated the visual quality of our explanations by computing FID [61] score. It 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 [61]. We examined real and fake (*i.e.,* generated explanations) images on the two extreme of the decision boundary, *i.e.,* a normal group $(f(\mathbf{x}_\delta) < 0.2)$ and an abnormal group $(f(\mathbf{x}_\delta) > 0.8)$. In Table I, we reported the FID for each group. Our method achieved a lower FID score as compared to xGEM. At the same time, cycleGAN obtained the lowest FID score and hence, it generates the most realistic images; however the explanation images from cycelGAN are not consistent with the classifier.

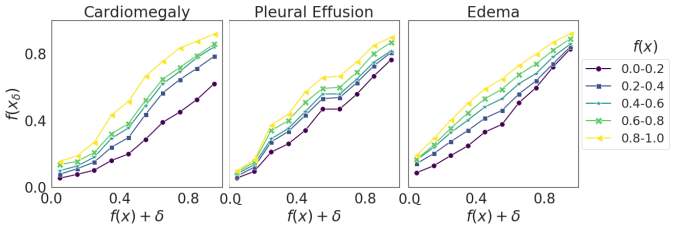

Fig. 7. The plot of expected outcome, $\boldsymbol{f(\mathbf{x}) + \delta}$, against actual response of the classifier on generated explanations, $\boldsymbol{f(\mathbf{x}_\delta)}$. Each line represents a set of input images with classification prediction $\boldsymbol{f(\mathbf{x})}$ in a given range.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

2) Classification model consistency: To quantify the consistency between the explanations and the classification model, we reported CV score in the last row of Table I. 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.,* $f(\mathbf{x}) + \delta$. An increasing monotonic trend validated that our explanation model represents the decision-making process of the classifier.

3) Identity preservation: A valid explanation should preserve small patient-specific details such as foreign objects (FO) including pacemaker and hardware. These objects are critical in identifying the patient in an x-ray. In this experiment,

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

Foreign Object	Ours with CARL	Ours w/o CARL
Pacemaker	**0.52**	0.40
Hardware	**0.63**	0.32

we demonstrate the importance of our proposed CARL in preserving such details. 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 counterfactual explanations. We performed an ablation study, where we replaced the CARL with a simple reconstruction loss based on ℓ_1 distance. Our model with CARL obtained a higher FOP score, as shown in Table II. The detector network have an accuracy of 80%. Fig. 8 presents examples of counterfactual explanations generated by our model with and without the CARL.

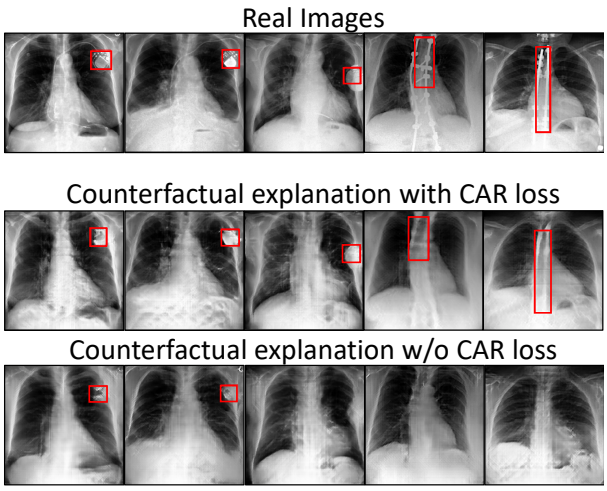

Fig. 8. Fidelity of generated images with respect to preserving foreign objects. The top row shows real images with pacemaker or hardware. The middle shows counterfactual images generated by our model while using context-aware reconstruction (CAR) loss. The bottom row shows the explanation images, without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

B. Comparison with Saliency maps

Popular existing approaches for model explanation consists of gradient-based methods that provide a qualitative explanation in the form of saliency maps [53], [63]. 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. In Fig. 9.C, we showed the two extreme explanation images and the corresponding difference map, derived for input images shown in Fig. 9.A. For proper comparison, we considered the absolute values of the saliency maps and normalized them in the range $[0, 1]$. We observed that the gradient-based saliency map highlighted almost the same region for all the three diseases. In contrast, our difference map localized disease to specific regions in the chest. For cardiomegaly, all the methods focused on the heart region. For PE, our difference map is confined to CP recess regions. And for edema, it mostly focuses on the right boundary of heart and lung extending till the hilar region. These regions are consistent with the clinical knowledge of the disease. In contrast, saliency maps are spread across the lower zone of the chest, without marking specific lung regions.

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 [64]–[66]. We followed the approach in [67] to calculate the CTR from a chest x-ray. We first segmented the lung and the heart region and then measured the thoracic and cardiac diameters to compute CTR.

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 [68], [69]. 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 Table. III, we observe a negative mean difference for CTR($\mathcal{X}^n$) − CTR($\mathcal{X}^a_{cf}$) (a 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.

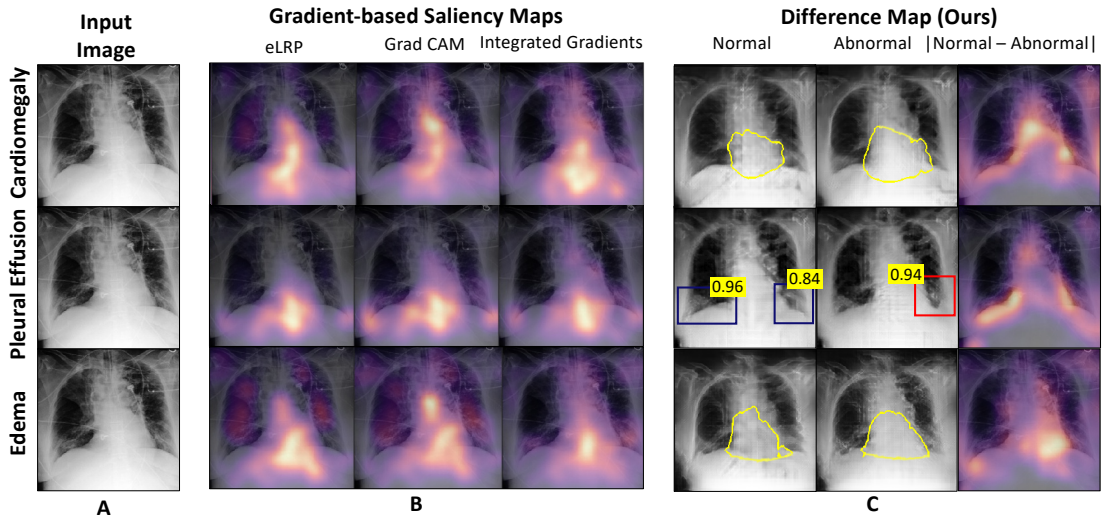

Fig. 9. Comparison of our method against different gradient-based methods. A: Input image; B: Saliency maps from existing works; C: Our simulation of saliency map as difference map between the normal and abnormal explanation images. More examples are shown in SM-Fig. 6, 7.

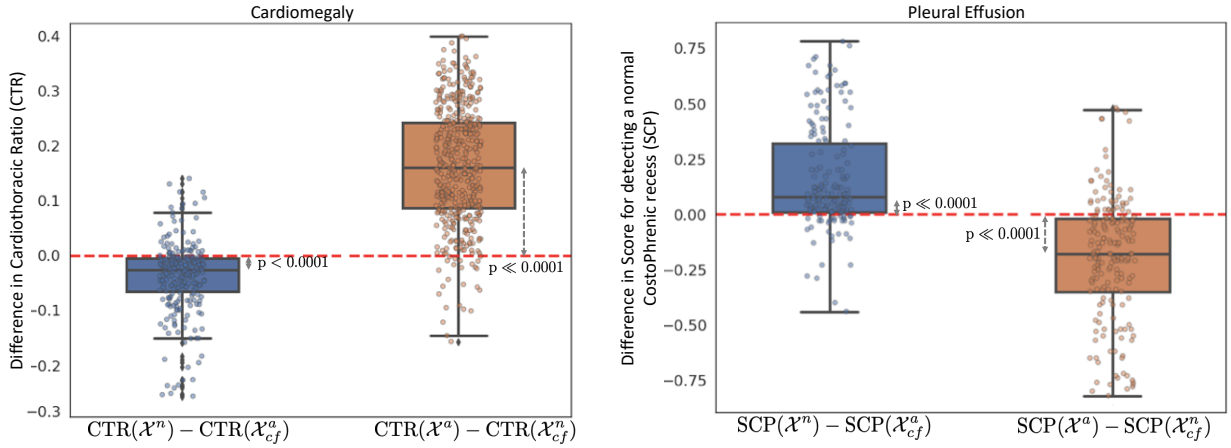

Fig. 10. Box plots to show distributions of pairwise differences in clinical-metrics such as CTR for cardiomegaly and the Score of normal CP recess (SCP) for pleural effusion, before (real) and after (counterfactual) our generative counterfactual creation process. The mean value corresponds to the average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

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.

V. 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 IV-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 other generative methods for model explanation such as xGEM and cycleGAN. CycleGAN produced the most visually appealing x-ray images. However, these images are not valid explanations as they failed to flip the classification decision about half of the time as shown in the last row of Table I. In contrast, 90% of the explanations generated by our model produced the desired outcome from the classifier (*see* Table I). The results suggest that producing realistic-looking images as explanations is not sufficient and hence, cycleGAN is not sufficient to capture the true behaviour of the classifier. The xGEM explanations are well-grounded with the classifier. But at the same time, the xGEM explanation images have a poor visual quality

which deemed then unsuited for clinical applications. xGEM adopted a variational autoencoder (VAE) to learn the data distribution. We suspect that the ℓ_2 reconstruction loss in VAE contributes to its low visual quality. Also, our revised context-aware reconstruction loss (CARL) helped in retaining small patient-specific details, which may otherwise get lost due to poor reconstruction. CARL improved the foreign object preservation (FOP) score and performed better than a simple distance-based reconstruction loss as reported in Table II. Though the FOP score is not perfect, part of the gap is due to the inaccuracies of the object detector. Please note that we constrained our explanation function to retain small details only during self reconstruction, *i.e.,* $\delta = 0$. While we kept the explanation image for $\delta \neq 0$ unconstrained to ensure that our explanations are not biased towards any FO.

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 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. In the absence of ground truth for lung and heart segmentation, our automatic pipeline to compute CTR suffers from inaccuracies. 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 were successful in capturing the difference between normal and abnormal chest x-rays. One may argue to use features such as CTR and SCP to perform disease classification. But models-based on these features will suffer from similar inaccuracies due to imperfect segmentation or detection, resulting in poor performance and generalization as compared to the deep learning methods.

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 [70]. 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. Though 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.

By providing counterfactual explanations, our work opens up many ideas for future work. Our framework showed that valid counterfactual 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 that are grounded with them.

REFERENCES

[1] A. Hosny, C. Parmar, J. Quackenbush, L. H. Schwartz, and H. J. Aerts, "Artificial intelligence in radiology," 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*.

[3] A. Rodriguez-Ruiz, K. Lång, A. Gubern-Merida, M. Broeders, G. Gennaro, and et al., "Stand-Alone Artificial Intelligence for Breast Cancer Detection in Mammography: Comparison With 101 Radiologists," *Journal of the National Cancer Institute*, 9 2019.

[4] F. Wang, R. Kaushal, and D. Khullar, "Should health care demand interpretable artificial intelligence or accept "black Box" Medicine?" 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*, 5 2020.

[6] H. Jiang, B. Kim, M. Guan, and M. Gupta, "To Trust Or Not To Trust A Classifier," in *Advances in Neural Information Processing Systems*, 2018, pp. 5541–5552.

[7] J. K. Winkler, C. Fink, F. Toberer, A. Enk, T. Deinlein, and et al., "Association between Surgical Skin Markings in Dermoscopic Images and Diagnostic Performance of a Deep Learning Convolutional Neural Network for Melanoma Recognition," *JAMA Dermatology*, 10 2019.

[8] L. Oakden-Rayner, J. Dunnmon, G. Carneiro, and C. Ré, "Hidden Stratification Causes Clinically Meaningful Failures in Machine Learning for Medical Imaging," in *the ACM Conference on Health, Inference, and Learning*, 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," in *Lecture Notes in Computer Science*, vol. 11070 LNCS, 6 2018.

[10] A. J. Larrazabal, N. Nieto, V. Peterson, D. H. Milone, and E. Ferrante, "Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis," *Proceedings of the NAS of the USA*, vol. 117, 6 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," 4 2018.

[12] A. T. Azar and S. M. El-Metwally, "Decision tree classifiers for automated medical diagnosis," *Neural Computing and Applications*, vol. 23, no. 7-8, pp. 2387–2403, 12 2013.

[13] S. Tsumoto, "Mining diagnostic rules from clinical databases using rough sets and medical diagnostic model," *Information Sciences*, vol. 162, no. 2, pp. 65–80, 5 2004.

[14] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why should i trust you?" Explaining the predictions of any classifier," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, vol. 13-17-August-2016, 8 2016, pp. 1135–1144.

[15] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems 30*, 2017, pp. 4765–4774.

[16] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic Attribution for Deep Networks," in *Proceedings of the 34th International Conference on Machine Learning*, 2017, pp. 3319–3328.

[17] 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," in *Proceedings of the IEEE ICCV*, 2017.

[18] S. M. Lundberg, P. G. Allen, and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," Tech. Rep.

[19] S. Singla, B. Pollack, J. Chen, and K. Batmanghelich, "Explanation by Progressive Exaggeration," in *ICLR*, 2019.

[20] D. Bau, J.-Y. Zhu, J. Wulff, W. Peebles, H. Strobelt, B. Zhou, and A. Torralba, "Seeing what a gan cannot generate," in *Proceedings of the IEEE ICCV*, 2019, pp. 4502–4511.

[21] T. Miyato and M. Koyama, "cGANs with Projection Discriminator," in *ICLR*, 2018.

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

[23] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE Conference on CVPR*, 2019, pp. 4401–4410.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on CVPR*, 12 2016.

[25] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," 12 2014.

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

[27] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. A. Riedmiller, "Striving for Simplicity: The All Convolutional Net," in *ICLR (workshop track)*, 2015.

[28] S. Bach, A. Binder, G. Montavon, and et al., "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PloS one*, vol. 10, no. 7, 2015.

[29] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *Proceedings of the 34th ICML-Volume 70*, 2017, pp. 3145–3153.

[30] P. Rajpurkar, J. Irvin, K. Zhu, and et al., "CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning," 11 2017.

[31] K. Young, G. Booth, B. Simpson, R. Dutton, and S. Shrapnel, "Deep neural network or dermatologist?" in *Lecture Notes in Computer Science*, vol. 11797 LNCS. Springer, 10 2019, pp. 48–55.

[32] F. Eitel and K. Ritter, "Testing the robustness of attribution methods for convolutional neural networks in MRI-based Alzheimer's disease classification," in *Lecture Notes in Computer Science*, vol. 11797 LNCS. Springer, 10 2019, pp. 3–11.

[33] R. Sayres, A. Taly, E. Rahimy, K. Blumer, and et al., "Using a Deep Learning Algorithm and Integrated Gradients Explanation to Assist Grading for Diabetic Retinopathy," *Ophthalmology*, vol. 126, 4 2019.

[34] P. Dabkowski and Y. Gal, "Real time image saliency for black box classifiers," in *Advances in Neural Information Processing Systems*, 2017, pp. 6967–6976.

[35] B. Zhou, A. Khosla, and et al., "Object detectors emerge in deep scene cnns," *Computing Research Repository*, 2014.

[36] R. C. Fong and A. Vedaldi, "Interpretable Explanations of Black Boxes by Meaningful Perturbation," in *Proceedings of the IEEE ICCV*, 2017.

[37] C.-H. Chang, E. Creager, A. Goldenberg, and D. Duvenaud, "Explaining Image Classifiers by Counterfactual Generation," 2019.

[38] A. Dhurandhar, P.-Y. Chen, R. Luss, and et al., "Explanations based on the missing: Towards contrastive explanations with pertinent negatives," in *Advances in Neural Information Processing Systems*, 2018.

[39] P. Samangouei, A. Saeedi, N. Liam, and S. Nathan, "ExplainGAN: Model Explanation via Decision Boundary Crossing Transformations," in *Computer Vision – ECCV 2018*, 2018, pp. 681–696.

[40] 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 preprint:1907.09615*, 2019.

[41] S. Liu, B. Kailkhura, D. Loveland, and Y. Han, "Generative Counterfactual Introspection for Explainable Deep Learning," *arXiv*, 2019.

[42] D. Mahajan, C. Tan, and A. Sharma, "Preserving Causal Constraints in Counterfactual Explanations for Machine Learning Classifiers," *arXiv preprint:1912.03277*, 12 2019.

[43] A. Van Looveren and J. Klaise, "Interpretable Counterfactual Explanations Guided by Prototypes," *arXiv arXiv:1907.02584*, 7 2019.

[44] A. Parafita Martinez and J. Vitria Marca, "Explaining visual models by causal attribution," in *ICCV Workshop*. Institute of Electrical and Electronics Engineers Inc., 10 2019, pp. 4167–4175.

[45] C. Agarwal and A. Nguyen, "Explaining an image classifier's decisions using generative models," 10 2019. http://arxiv.org/abs/1910.04256

[46] P. Wang and N. Vasconcelos, "SCOUT: Self-Aware Discriminant Counterfactual Explanations," in *Proceedings of the IEEE CVPR*, 6 2020.

[47] Y. Goyal, Z. Wu, J. Ernst, D. Batra, D. Parikh, and S. Lee, "Counterfactual Visual Explanations," in *ICML*, 2019, pp. 2376–2384.

[48] A. Narayanaswamy, S. Venugopalan, D. R. Webster, and et al., "Scientific Discovery by Generating Counterfactuals using Image Translation," *MICCAI*, 7 2020.

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

[50] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks," in *Proceedings of the IEEE ICCV*, 2017.

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

[52] 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, no. 1, p. 317, 12 2019.

[53] 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."

[54] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," *Proceedings - 30th IEEE Conference on CVPR*, vol. 2017-January, pp. 2261–2269, 8 2016.

[55] O. Ronneberger, P. Fischer, and T. Brox, "U-Net Convolutional Networks for Biomedical Image Segmentation," in *MICCAI*, 2015.

[56] 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, no. 1, pp. 19–40, 2006.

[57] 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, no. 6, pp. 475–477, 2014.

[58] S. Ren, K. He, and et al., "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," Tech. Rep., 2015.

[59] "Pleural Effusion Imaging: Overview, Radiography, Computed Tomography." https://emedicine.medscape.com/article/355524-overview

[60] M. A. Iqbal and M. Gupta, *Cardiogenic Pulmonary Edema*. StatPearls Publishing, 7 2020.

[61] M. Heusel, H. Ramsauer, T. Unterthiner, and et al., "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Advances in Neural Information Processing Systems*, 2017.

[62] R. K. Mothilal, A. Sharma, and C. Tan, "Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations," *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 607–617, 5 2019.

[63] 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, no. 1, pp. 1–9, 12 2019.

[64] 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.

[65] "Evaluating Cardiomegaly by Radiological Cardiothoracic Ratio as Compared to Conventional Echocardiography," *Journal of Cardiology & Current Research*, vol. 9, no. 2, 6 2017.

[66] K. Dimopoulos, G. Giannakoulas, and et al., "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*, 6 2013.

[67] I. Chamveha, T. Promwiset, T. Tongdee, P. Saiviroonporn, and W. Chaisangmongkon, "Automated Cardiothoracic Ratio Calculation and Cardiomegaly Detection using Deep Learning Approach," *arXiv:2002.07468*, 2 2020.

[68] P. Maduskar, L. Hogeweg, R. Philipsen, and B. van Ginneken, "Automated localization of costophrenic recesses and costophrenic angle measurement on frontal chest radiographs," in *Medical Imaging 2013: Computer-Aided Diagnosis*, 3 2013.

[69] 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.

[70] 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.

[71] D. M. Hansell, A. A. Bankier, H. MacMahon, and et al., "Fleischner Society: Glossary of terms for thoracic imaging," 3 2008.
[72] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral Normalization for Generative Adversarial Networks," *6th ICLR 2018*, 2 2018.
[73] K. Wada, "labelme Image Polygonal Annotation with Python," https://github.com/wkentaro/labelme, 2016.
[74] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *arXiv 1409.1556*, 6 2014.

SUPPLEMENTARY MATERIAL

A. 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 [52]. 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 [53]. The labeler first defines some thoracic observations using a radiology lexicon [71]. 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], [30], [53] on diagnosis classification, we used DenseNet-121 [54] 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 explainer function is a conditional GAN with an encoder. We used a ResNet [24] architecture for the Encoder, Generator, and Discriminator in Fig. 3. The image encoding learned by the encoder $E(\mathbf{x})$ is fed into the generator $G(\cdot)$, along with the condition $c_f(\mathbf{x}, \delta)$. In Generator, following the details in [72], we replace the ordinary batch normalization layer (BN) in the ResBlock with conditional BN to encode the condition. The $G(\cdot)$ have five ResNet blocks, where each block consists of cBN-ReLU-UpSampling-Conv3-cBN-ReLU-Conv3. cBN is the conditional batch normalization. ReLU is the activation function and Conv3 is the convolution filter. Upsampling is performed by the nearest neighbor interpolator. The encoder function has the same architecture, but without conditional BN and it down-samples an image. Downsampling is performed using an average pooling operator. The discriminator function has five ResNet blocks, where each block is ReLU-Conv3-SN-ReLU-Conv3-SN-DownSample. SN is spectral normalization [21] in the discriminator. We optimized the adversarial hinge loss for the GAN training. We used the Adam optimizer [25], 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.

4) Semantic Segmentation: We adopted a 2D U-Net [55] 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) [56] and Montgomery [57] 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 annotations marking the presence of foreign objects using the LabelMe [73] annotation tool. Next, we trained an object detector based on faster regional CNN [58], which used VGG-16 model [74], trained on MIMIC-CXR dataset as its foundation. We used this object detector to enforce our novel context-aware reconstruction loss (CARL).

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 [68], [69]. 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

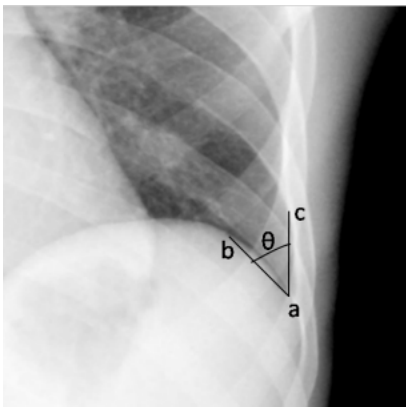

Fig. 11. The costophrenic angle (CPA) on a chest x-ray is marked as the angle formed by, (a) costophrenic angle point, (b) hemidiaphragm point and (c) lateral chest wall point, as shown by Maduskar *et al.*in [69]

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.* [40] for the implementation of xGEM. First, VAE is trained to generate face images. 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. The pre-trained VAE is then extended to incorporate the cross-entropy loss for flipping the label of the query image. The model evaluates the cross-entropy loss by passing the generated image through the classifier.

7) cycleGAN: We refer to the work by Narayanaswamy *et al.* [48] and DeGrave *et al.* [49] 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.

B. Extended data consistency results

In Fig. 12 and Fig. 13, we show the results to visualize our explanations and compared it against xGEM and cycleGAN method. The results are an extension of Fig.6 in the main manuscript. 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. 12 are the classifier's prediction for the specific disease. For cycleGAN, the results demonstrate an example where the counterfactual image is not faithful to the classifier *i.e.,* the explanation doesn't have an opposing prediction as compared to the input image. For example, in Fig. 12, 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.

Next, we quantify the consistency between our explanations and the classification model at every step of the transformation. Similar to Fig.7., 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 14, 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 Fig. 7. Although the trend is monotonically increasing in almost all cases, the slope of the line varies. A high slope covering the entire y-axis range of [0,1] shows that the generated explanations gradually transform and cross the decision boundary. For cardiomegaly and edema, xGEM achieves a progressive transformation, but it doesn't cover the entire prediction range. While cycleGAN, create counterfactual images at extreme ends with very similar classification prediction, as evident in small slope, hence these explanations are not consistent with the classifier.

C. 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, 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. 15 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. 15, 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 Table. 1. of the main manuscript.

D. Extended results for saliency maps

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

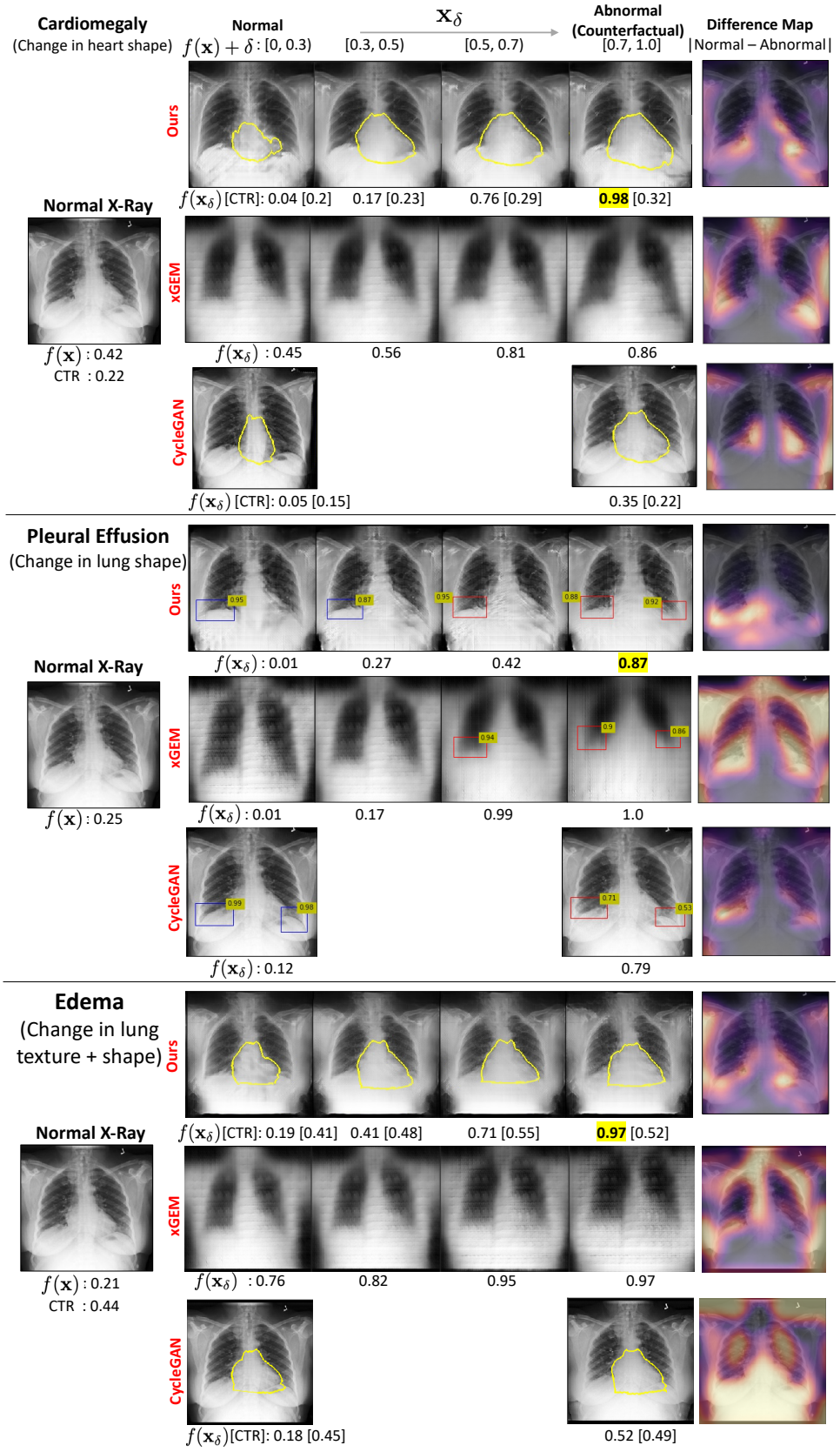

Fig. 12. The transformation of a normal chest x-ray into the counterfactual explanations for three classes, cardiomegaly (first row), pleural effusion (PE) (middle row) and edema (last row). The bottom labels are the classifier's prediction for the specific class. The last column shows the difference 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.

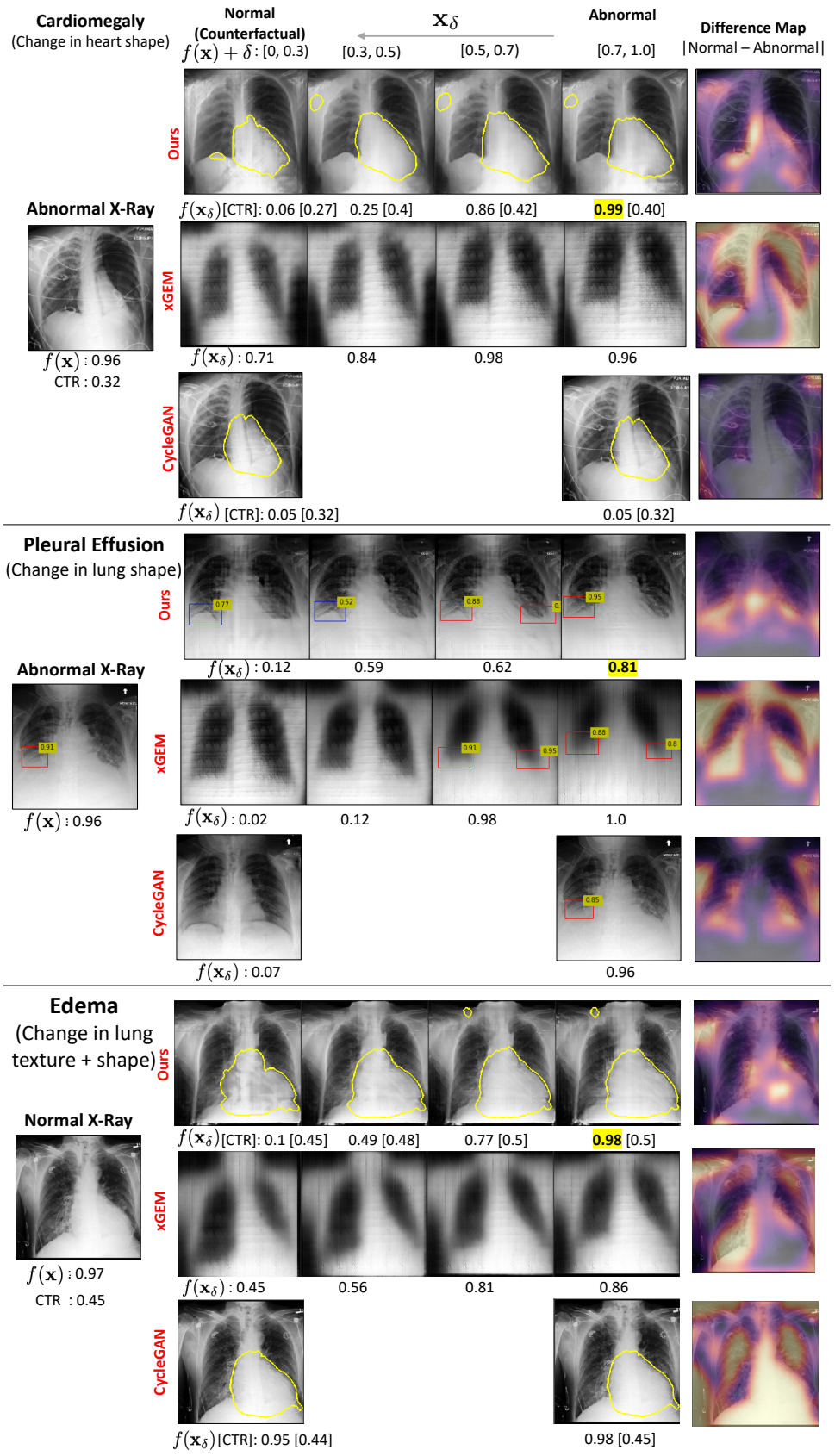

Fig. 13. The transformation of an abnormal chest x-ray into the counterfactual explanations for three classes, cardiomegaly (first row), pleural effusion (PE) (middle row) and edema (last row). The bottom labels are the classifier's prediction for the specific class. The last column shows the difference 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.

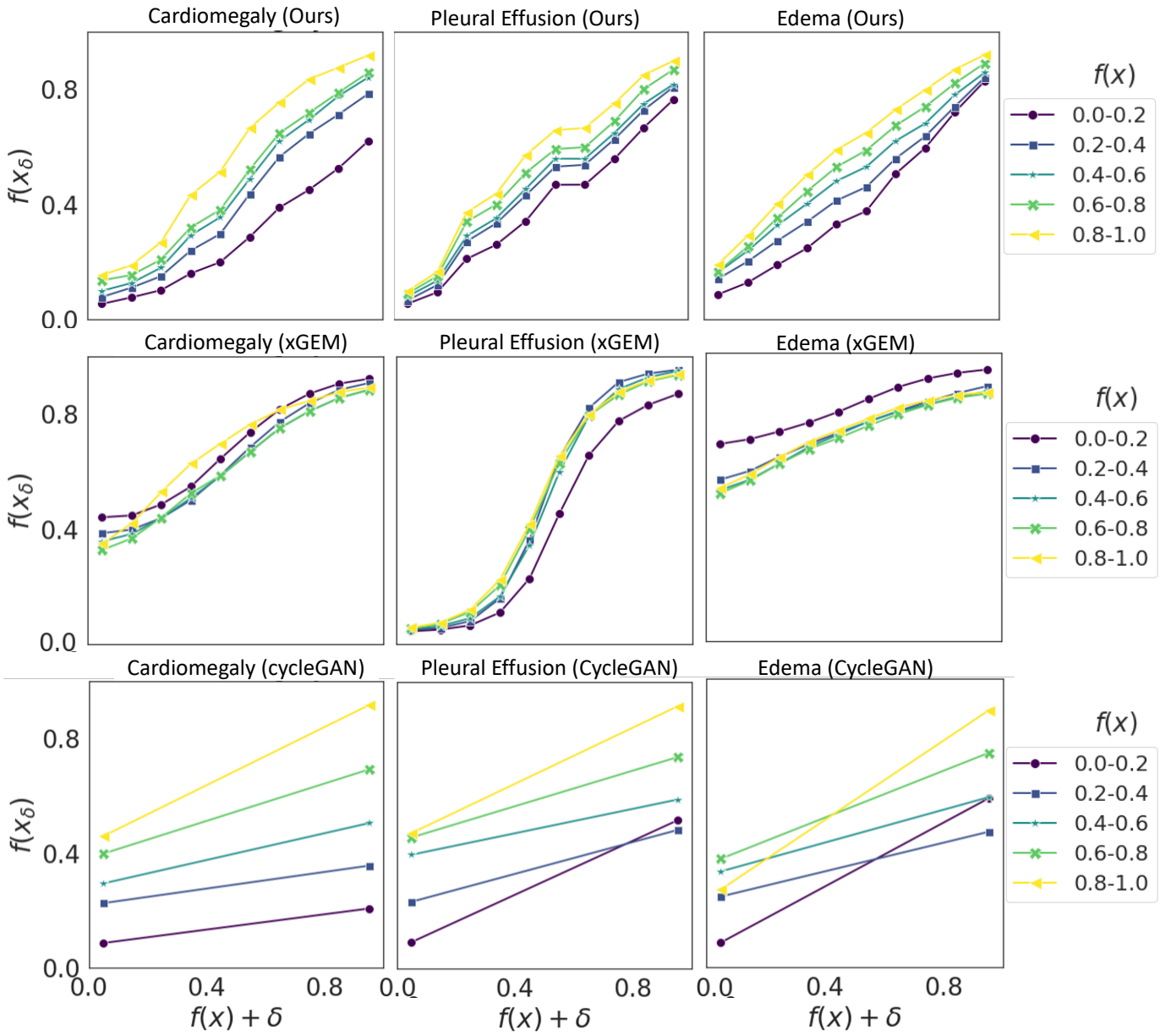

Fig. 14. The plot of expected outcome, $f(\mathbf{x})+\delta$, against actual response of the classifier on generated explanations, $f(\mathbf{x}_\delta)$. Each line represents a set of input images with classification prediction $f(\mathbf{x})$ in a given range.

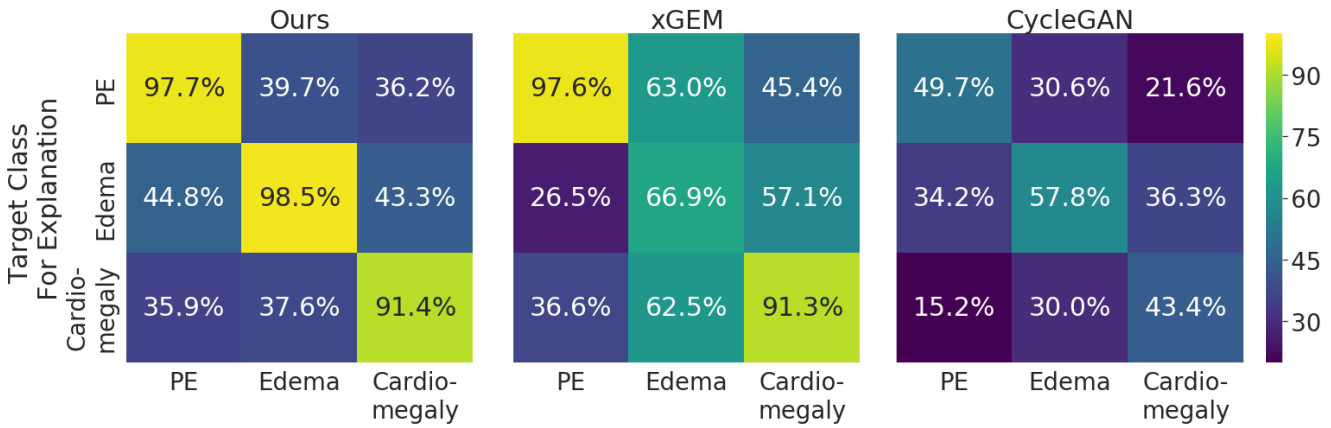

Fig. 15. Each cell is the fraction of the generated explanations, that have flipped in a class as compared to the query image. The x-axis shows the classes in a multi-label setting, and the y-axis shows the target class for which an explanation is generated. Note: This is not a confusion matrix.

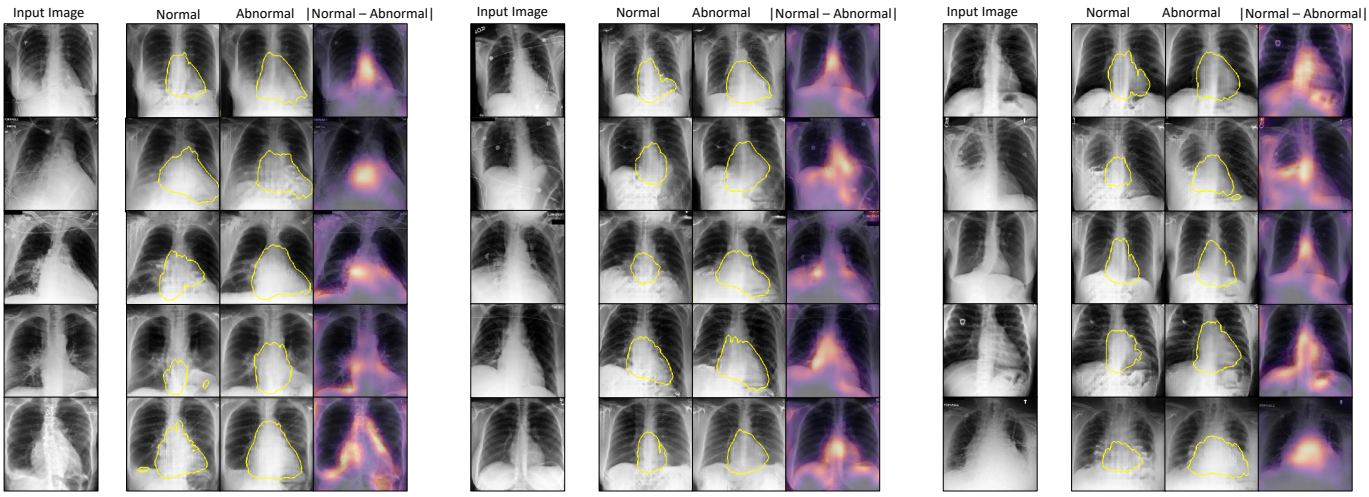

Fig. 16. Extended results for explanation produced by our model for **Cardiomegaly.** For each input image, we produce a normal and abnormal image as an explanation and take their pixel-wise difference to extract the saliency map.

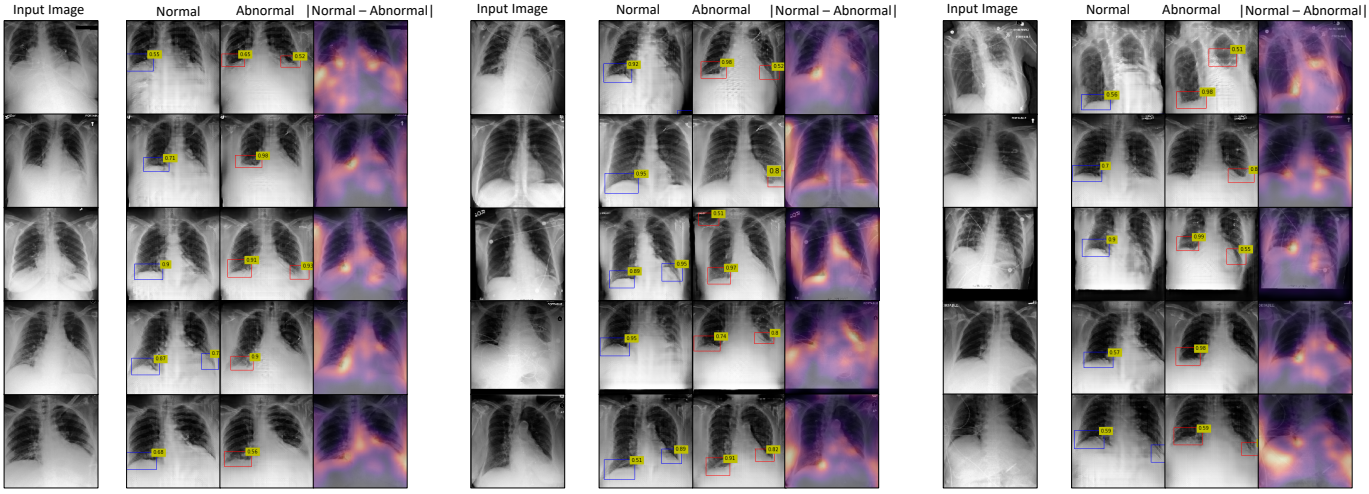

Fig. 17. Extended results for explanation produced by our model for **Pleural Effusion.** For each input image, we produce a normal and abnormal image as an explanation and take their pixel-wise difference to extract the saliency map.

TABLE III
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	$\ll 0.0001$

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. 16, we showed the two extreme explanation images and the corresponding difference map, derived for input images shown in the column on left. For proper comparison, we considered the absolute values of the saliency maps and normalized them in the range $[0, 1]$. For cardiomegaly, we highlight the heart contour in yellow. 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, 17.

E. 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 III 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. III, 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.

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. III, 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.