Explaining the black-box smoothly—A counterfactual approach

Sumedha Singla [a,*], Motahhare Eslami [b], Brian Pollack [c,1], Stephen Wallace [d],
Kayhan Batmanghelich [c]

[a] *Computer Science Department at the University of Pittsburgh, Pittsburgh, PA, 15206, USA*
[b] *School of Computer Science, Human–Computer Interaction Institute, Carnegie Mellon University, USA*
[c] *Department of Biomedical Informatics, the University of Pittsburgh, Pittsburgh, PA, 15206, USA*
[d] *University of Pittsburgh Medical School, Pittsburgh, PA, 15206, USA*

ARTICLE INFO

ABSTRACT

We propose a BlackBox *Counterfactual Explainer*, designed to explain image classification models for medical applications. Classical approaches (*e.g.,*, saliency maps) that assess feature importance do not explain *how* imaging features in important anatomical regions are relevant to the classification decision. Such reasoning is crucial for transparent decision-making in healthcare applications. Our framework explains the decision for a target class by gradually *exaggerating* the semantic effect of the class in a query image. We adopted a Generative Adversarial Network (GAN) to generate a progressive set of perturbations to a query image, such that the classification decision changes from its original class to its negation. Our proposed loss function preserves essential details (*e.g.,* support devices) in the generated images.

We used counterfactual explanations from our framework to audit a classifier trained on a chest X-ray dataset with multiple labels. Clinical evaluation of model explanations is a challenging task. We proposed clinically-relevant quantitative metrics such as cardiothoracic ratio and the score of a healthy costophrenic recess to evaluate our explanations. We used these metrics to quantify the counterfactual changes between the populations with negative and positive decisions for a diagnosis by the given classifier.

We conducted a human-grounded experiment with diagnostic radiology residents to compare different styles of explanations (no explanation, saliency map, cycleGAN explanation, and our counterfactual explanation) by evaluating different aspects of explanations: (1) understandability, (2) classifier's decision justification, (3) visual quality, (d) identity preservation, and (5) overall helpfulness of an explanation to the users. Our results show that our counterfactual explanation was the only explanation method that significantly improved the users' understanding of the classifier's decision compared to the no-explanation baseline. Our metrics established a benchmark for evaluating model explanation methods in medical images. Our explanations revealed that the classifier relied on clinically relevant radiographic features for its diagnostic decisions, thus making its decision-making process more transparent to the end-user.

1. Introduction

Machine learning, specifically Deep Learning (DL), is being increasingly used for sensitive applications such as Computer-Aided Diagnosis (Hosny et al., 2018) and other tasks in the medical imaging domain (Rajpurkar et al., 2018; Rodriguez-Ruiz et al., 2019). However, for real-world deployment (Wang et al., 2020), the decision-making process of these models should be explainable to humans to obtain their trust in the model (Gastounioti and Kontos, 2020; Jiang et al., 2018). Explainability is essential for auditing the model (Winkler et al., 2019), identifying various failure modes (Oakden-Rayner et al., 2020;

Eaton-Rosen et al., 2018) or hidden biases in the data or the model (Larrazabal et al., 2020), and for obtaining new insights from large-scale studies (Rubin et al., 2018).

With the advancement of DL methods for medical imaging analysis, deep neural networks (DNNs) have achieved near-radiologist performance in multiple image classification tasks (Seah et al., 2021; Rajpurkar et al., 2017). However, DNNs are criticized for their "black-box" nature, *i.e.,* they fail to provide a simple explanation as to why a given input image produces a corresponding output (Tonekaboni et al., 2019). To address this concern, multiple model explanation
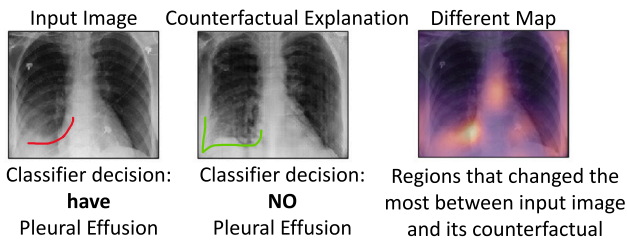
Fig. 1. Counterfactual explanation shows "where" + "what" minimum change must be made to the input to flip the classification decision. For Pleural Effusion, we can observe vanishing of the meniscus (red) in counterfactual image as compared to the query image.

techniques have been proposed that aim to explain the decision-making process of DNNs (Selvaraju et al., 2017; Cohen et al., 2021). The most common form of explanation in medical imaging is a class-specific heatmap overlaid on the input image. It highlights the most relevant regions (*where*) for the classification decision (Rajpurkar et al., 2017; Young et al., 2019). However, the location information alone is insufficient for applications in medical imaging. Different diagnoses may affect the same anatomical regions, resulting in similar explanations for multiple diagnosis, resulting in inconclusive explanations. A thorough explanation should explain *what* imaging features are present in those important locations, and *how* changing such features modifies the classification decision.

To address this problem, we propose a novel explanation method to provide a counterfactual explanation. A counterfactual explanation is a perturbation of the input image such that the classification decision is flipped. By comparing, the input image and its corresponding counterfactual image, the end-users can visualize the difference in important image features that leads to a change in classification decision. Fig. 1 shows an example. The input image is predicted as positive for pleural effusion (PE), while the generated counterfactual image is negative for PE. The changes are mostly concentrated in the lower lobe region, which is known to be clinically important for PE (Lababede, 2017). Counterfactual explanation is used to derive a pseudo-heat-map, highlighting the regions that change the most in the transformation (difference map in Fig. 1).

We demonstrate the performance of the counterfactual explainer on a chest X-ray (CXR) dataset. Rather than generating just one counterfactual image at the end of the prediction spectrum, our explanation function generates a series of perturbed images that gradually traverse the decision boundary from one extreme (negative decision) to another (positive decision) for a given target class. We adopted a conditional Generative Adversarial Network (cGAN) as our explanation function (Singla et al., 2019). We extend the cGAN to preserve small or uncommon details during image generation (Bau et al., 2019). Preserving such details is particularly important in our application, as the missing information may include support devices that may influence human users' perceptions. To this end, we incorporated semantic segmentation and object detection into our loss function to preserve the shape of the anatomy and foreign objects during image reconstruction. We evaluated the quality of our explanations using different quantitative metrics, including clinical measures. Further, we performed a clinical study with 12 radiology residents to compare the explanations for the proposed method and the baseline models.

1.1. Related work

Posthoc *explanation* is a popular approach that aims to improve human understanding of a pre-trained classifier. Our work broadly relates to the following posthoc methods:

Feature Attribution methods provide explanation by producing a saliency map that shows the importance of each input component (*e.g.*, pixel) to the classification decision. *Gradient-based* approaches (Simonyan et al., 2013; Springenberg et al., 2015; Bach et al., 2015; Shrikumar et al., 2017; Sundararajan et al., 2017; Lundberg et al., 2017; Selvaraju et al., 2017) 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.*, CXR (Rajpurkar et al., 2017), skin imaging (Young et al., 2019), brain MRI (Eitel and Ritter, 2019) and retinopathy (Sayres et al., 2019).

Perturbation-based methods identify salient regions by directly manipulating the input image and analyzing the resulting changes in the classifier's output. Such methods modify specific pixels or regions in an input image, either by masking with constant values (Dabkowski and Gal, 2017) or with random noise, occluding (Zhou et al., 2015), localized blurring (Fong and Vedaldi, 2017), or in-filling (Chang et al., 2019). Especially for medical images, such perturbations may introduce anatomically implausible features or textures. Our proposed method also generates a perturbation of the query image such that classification decision is flipped. But in contrast to the above methods, we enforce consistency between the perturbed data and the real data distribution to ensure that the perturbations are plausible and visually similar to the input.

Counterfactual Explanations are a type of contrastive (Dhurandhar et al., 2018) explanation that provides a useful way to audit the classifier and determine causal attributes that lead to the classification decision (Parafita Martinez and Vitria Marca, 2019; Singla et al., 2021). Similar to our method, generative models like GANs and variational autoencoders (VAE) are used to compute interventions that generate realistic counterfactual explanations (Cohen et al., 2021; Joshi et al., 2019). Much of this work is limited to simpler image datasets like MNIST, celebA (Liu et al., 2019; Van Looveren and Klaise, 2019) or simulated data (Parafita Martinez and Vitria Marca, 2019). For more complex natural images, previous studies (Chang et al., 2019; Agarwal and Nguyen, 2020) focused on finding and in-filling salient regions to generate counterfactual images. In contrast, our explanation function does not require any re-training for generating explanations for a new image at inference time. In another line of work (Wang and Vasconcelos, 2020; Goyal et al., 2019) provide counterfactual explanations that explain both the predicted and the counter class. Further, researcher (Narayanaswamy et al., 2020; DeGrave et al., 2020) has used a cycle-GAN (Zhu et al., 2017) model to perform image-to-image translation between normal and abnormal images. Such methods are independent of the classifier. In contrast, our framework uses a classifier consistency loss to enable image perturbation that is consistent with the classifier.

1.2. Contributions

In this paper, we propose a progressive counterfactual explainer, that explains the decision of a pre-trained image classifier. Our contributions are summarized as follows:

1. We developed a cGAN-based framework to generate progressively changing perturbations of the query image, such that classification decision changes from being negative to being positive for a given target class. We performed a thorough qualitative and quantitative evaluation of our explanation function to audit a classifier trained on a CXR dataset.
2. Our method preserved the anatomical shape and foreign objects such as support devices across generated images by adding a specialized reconstruction loss. The loss incorporates context from semantic segmentation and foreign object detection networks.
3. We proposed quantitative metrics based on clinical definition of two diseases (cardiomegaly and PE). We are one of the first methods to use such metrics for quantifying DNN model explanation. Specifically, we used these metrics to quantify statistical differences between the real images and their corresponding counterfactual images.

4. We are one of the first methods to conduct a thorough human-grounded study to evaluate different counterfactual explanations for medical imaging task. Specifically, we collected and compared feedback from diagnostic radiology residents, on different aspects of explanations: (1) understandability, (2) classifier's decision justification, (3) visual quality, (d) identity preservation, and (5) overall helpfulness of an explanation to the users.

2. Methodology

We consider a *black-box* image classifier f, with high prediction accuracy. We assume that f is a differentiable function and we have access to its value as well as its gradient with respect to the input $\nabla_{\mathbf{x}} f(\mathbf{x})$. We also assume access to either the training data for f, or an equivalent dataset with competitive prediction accuracy.

Notation: The classification function is defined as $f : \mathbb{R}^d \rightarrow \mathbb{R}^K$, where d is the dimensionality of the image space and K is the number of classes. The classifier produces point estimates for posterior probability of class k as $\mathbb{P}(y_k|\mathbf{x}) = f(\mathbf{x})[k] \in [0, 1]$.

Explanation function: We aim to explain the decision of function f for a target class k. We consider visual explanation of the black-box as a generative process that produces a plausible and realistic perturbation of the query image $\mathbf{x}$ such that the classification decision for class k is changed to a desired value $\mathbf{c}$. This idea allows us to view $\mathbf{c}$ as a "knob". By gradually changing the desired output $\mathbf{c}$ in range $[0, 1]$, we generate progressively changing perturbations of the query image $\mathbf{x}$, such that classification decision changes from being negative to being positive for a class k.

To achieve this, we propose an *explanation* function $\mathbf{x_c} \triangleq \mathcal{I}_{f_k}(\mathbf{x}, \mathbf{c}) : (\mathcal{X}, \mathbb{R}) \rightarrow \mathcal{X}$. This function takes two arguments: a query image $\mathbf{x}$ and the desired posterior probability $\mathbf{c}$ for the target class k. The explanation function generates a perturbed image $\mathbf{x_c}$ such that $f(\mathbf{x_c})[k] \approx \mathbf{c}$. For simplicity, we will drop k from subsequent notations. Fig. 2 summarizes our framework. We design the explanation function to satisfy the following properties:

(A) **Data consistency**: $\mathbf{x_c}$ should resemble data instance from input space *i.e.,* if input space comprises of CXRs, $\mathbf{x_c}$ should look like a CXR with minimum artifacts or blurring.

(B) **Classifier consistency**: $\mathbf{x_c}$ should produce the desired output from the classifier f, *i.e.,* $f(\mathcal{I}_f(\mathbf{x}, \mathbf{c})) \approx \mathbf{c}$.

(C) **Context-aware self-consistency**: On using the original decision as the condition, *i.e.,* $\mathbf{c} = f(\mathbf{x})$, the explanation function should reconstruct the query image. We forced this condition for self-consistency as $\mathcal{I}_f(\mathbf{x}, f(\mathbf{x})) = \mathbf{x}$ and for cyclic-consistency as $\mathcal{I}_f(\mathbf{x_c}, f(\mathbf{x})) = \mathbf{x}$. Further, we constrained the explanation function to achieve the aforementioned reconstructions while preserving anatomical shape and foreign objects (*e.g.,* pacemaker) in the input image.

Overall objective: Our explanation function $\mathcal{I}_f(\mathbf{x}, \mathbf{c})$ is trained end to end to learn parameters for three networks, an image encoder $E(\cdot)$, a conditional GAN generator $G(\cdot)$ and a discriminator $D(\cdot)$, to satisfy the above three properties while minimizing the following objective function:

$$\min_{E,G} \max_D \lambda_{cGAN} \mathcal{L}_{\text{cGAN}}(D, G) + \lambda_f \mathcal{L}_f(D, G) + \lambda_{rec} \mathcal{L}_{\text{rec}}(E, G) \quad (1)$$

where $\mathcal{L}_{\text{cGAN}}$ is a conditional GAN-based loss function that enforces data-consistency, $\mathcal{L}_f$ enforces classifier consistency through a KullbackLeibler (KL) divergence loss and $\mathcal{L}_{\text{rec}}$ is a reconstruction loss that enforces self-consistency. Hyper-parameters, $\lambda_{cGAN}, \lambda_f$ and λ_{rec} controls the balance between the terms. In the following sections, we will discuss each property and the associated loss term in detail.

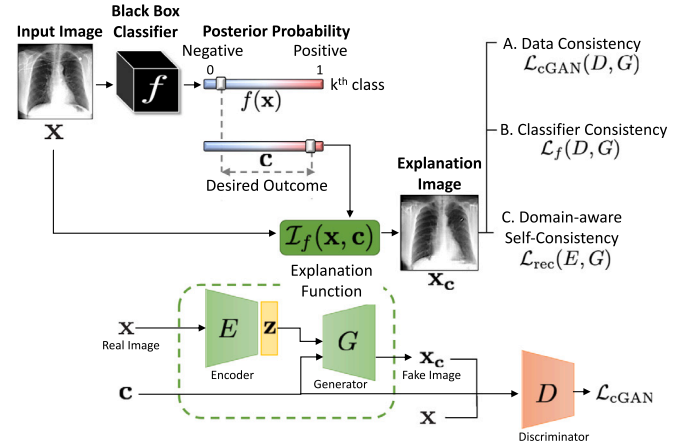**Fig. 2.** Explanation function $\mathcal{I}_f(\mathbf{x}, \mathbf{c})$ for classifier f. Given an input image $\mathbf{x}$, we generates a perturbation of the input, $\mathbf{x_c}$ as explanation, such that the posterior probability, f, changes from its original value, $f(\mathbf{x})$, to a desired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

2.1. Data consistency

We formulated the explanation function, $\mathcal{I}_f(\mathbf{x}, \mathbf{c})$, as an image encoder $E(\cdot)$ followed by a conditional GAN (cGAN) (Miyato and Koyama, 2018), with $\mathbf{c}$ 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, cGAN is composed of two deep networks, generator $G(\cdot)$ and discriminator $D(\cdot)$. The G, D are trained adversarially by optimizing the following objective function,

$$\mathcal{L}_{\text{cGAN}}(D, G) = \mathbb{E}_{\mathbf{x}, \mathbf{c} \sim P(\mathbf{x}, \mathbf{c})} \left[\log \left(D(\mathbf{x}, \mathbf{c}) \right) \right] + \\ \mathbb{E}_{\mathbf{z} \sim P_{\mathbf{z}}, \mathbf{c} \sim P_c} \left[\log \left(1 - D(G(\mathbf{z}, \mathbf{c}), \mathbf{c}) \right) \right] \quad (2)$$

where $\mathbf{c}$ denotes a condition and $\mathbf{z}$ is noise sampled from 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 $E(\cdot)$. Finally, the explanation function is defined as,

$$\mathcal{I}_f(\mathbf{x}, \mathbf{c}) = G(E(\mathbf{x}), \mathbf{c}). \quad (3)$$

For the discriminator in cGAN, we adapted the loss function from the Projection GAN (Miyato and Koyama, 2018). The Projection GAN imposes the following structure on the discriminator loss function:

$$\mathcal{L}_{\text{cGAN}}(D, G)(\mathbf{x}, \mathbf{c}) := r(\mathbf{x}) + r(\mathbf{c}|\mathbf{x}), \quad (4)$$

Here, $r(\mathbf{x})$ is the discriminator logit that evaluates the visual quality of the generated image. It is the discriminator's attempt to separate real images from the fakes images created by the generator. The second term evaluates the correspondence between the generated image $\mathbf{x_c}$ and the condition $\mathbf{c}$.

To represent the condition, the discriminator learns an embedding matrix $\mathbf{V}$ with N rows, where N is the number of conditions. The condition is encoded as an N-dimensional one-hot vector which is multiplied by the embedding-matrix to extract the condition-embedding. When $\mathbf{c} = n$, the conditional embedding is given as the nth row of the embedding-matrix ($\mathbf{v}_n$). The projection is computed as the dot product of the condition-embedding and the features extracted from the fake image,

$$\mathcal{L}_{\text{cGAN}}(D, G)(\mathbf{x}, \mathbf{c}) := r(\mathbf{x}) + \mathbf{v}_n^T \phi(\mathbf{x}), \quad (5)$$

where, n is the current class for the conditional generation and ϕ is the feature extractor.

In our use-case, the condition $\mathbf{c}$ is the desired posterior probability from the classification function f. $\mathbf{c}$ is a continuous variable with values in range $[0, 1]$. Projection-cGAN requires the condition to be a discrete variable, to be mapped to the embedding matrix $\mathbf{V}$. Hence, we discretize the range $[0, 1]$ into N bins, where each bin is one condition. One can view change from $f(\mathbf{x})$ to $\mathbf{c}$ as changing the bin index from the current value $C(f(\mathbf{x}))$ to $C(\mathbf{c})$ where $C(\cdot)$ returns the bin index.

2.2. Classifier consistency

Ideally, cGAN should generate a series of smoothly transformed images as we change condition $\mathbf{c}$ in range $[0, 1]$. These images, when processed by the classifier f should also smoothly change the classification prediction between $[0, 1]$. To enforce this, rather than considering bin-index $C(\mathbf{c})$ as a scalar, we consider it as an ordinal-categorical variable, i.e., $C(\mathbf{c}_1) < C(\mathbf{c}_2)$ when $\mathbf{c}_1 < \mathbf{c}_2$. Specifically, rather than checking one condition that desired bin-index is equal to some value n, $C(\mathbf{c}) = n$, we check $n - 1$ conditions that desired bin-index is greater than all bin-index which are less than n i.e., $C(\mathbf{c}) > i \forall i \in [1, n)$ (Frank and Hall, 2001).

We adapted Eq. (5) to account for a categorical variable as the condition, by modifying the second term to support ordinal multi-class regression. The modified loss function is as follows:

$$r(\mathbf{c} = n | \mathbf{x}) := \sum_{i < n} \mathbf{v}_i^T \boldsymbol{\phi}(\mathbf{x}), \tag{6}$$

Along with conditional loss for the discriminator, we need additional regularization for the generator to ensure that the actual classifier's outcome, i.e., $f(\mathbf{x_c})$, is very similar to the condition $\mathbf{c}$. To ensure this compatibility with f, we further constrain the generator to minimize the Kullback–Leibler (KL) divergence that encourages the classifier's score for $\mathbf{x_c}$ to be similar to $\mathbf{c}$. Our final condition-aware loss is as follows,

$$\mathcal{L}_f(D, G) := r(\mathbf{c} | \mathbf{x}) + D_{\mathrm{KL}}(f(\mathbf{x_c}) || \mathbf{c}), \tag{7}$$

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 minimize the KL divergence between the posterior probability for new image $f(\mathbf{x_c})$ and the desired prediction distribution $\mathbf{c}$. It influences only the G. Please note, the term $r(\mathbf{x})$ is not appearing in Eq. (7) as it is independent of $\mathbf{c}$.

2.3. 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 or any foreign objects (FO) if present. While images generated by GAN are shown to be realistic looking (Karras et al., 2019), GAN with an encoder may ignore small or uncommon details in the input image (Bau et al., 2019). To preserve such details, 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 FO present in the input image. The CARL is defined as,

$$\mathcal{L}_{\mathrm{rec}}(\mathbf{x}, \mathbf{x}') = \sum_j \frac{S_j(\mathbf{x}) \odot \|\mathbf{x} - \mathbf{x}'\|_1}{\sum S_j(\mathbf{x})} + D_{\mathrm{KL}}(O(\mathbf{x}) \| O(\mathbf{x}')). \tag{8}$$

Here, $S(\cdot)$ is a pre-trained semantic segmentation network that produces a label map for different regions in the input domain. We used a U-Net architecture for segmentation network as shown in Fig. 3.a. 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.
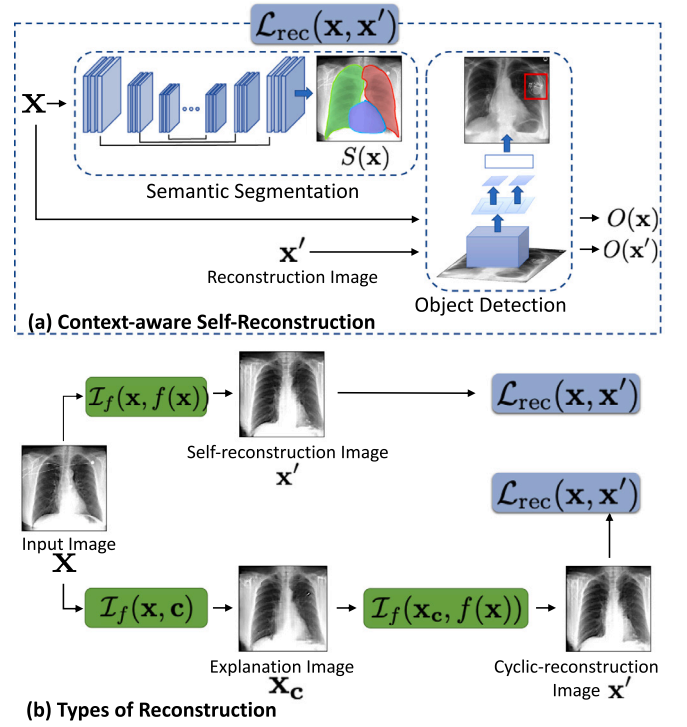

Fig. 3. (a) A domain-aware self-reconstruction loss with pre-trained semantic segmentation $S(\mathbf{x})$ and object detection $O(\mathbf{x})$ networks. (b) The self and cyclic reconstruction should retain maximum information from $\mathbf{x}$.

$O(\mathbf{x})$ is a pre-trained object detector that, given an input image $\mathbf{x}$, outputs a number of bounding boxes called region of interests (ROIs). For each bounding box, it outputs 2-d coordinates in the image where the box is located and an associated probability of presence of an object. Using the input image $\mathbf{x}$, we obtain the ROIs and associated $O(\mathbf{x})$, which is a probability vector, stating probability of finding an object in each ROI. For reconstructed image $\mathbf{x}'$, we reuse the ROIs obtained from image $\mathbf{x}$ and computed the associated probabilities for the reconstructed image as $O(\mathbf{x}')$. Next, we used KL divergence to quantify the difference between probability vectors as $D_{\mathrm{KL}}(O(\mathbf{x}) \| O(\mathbf{x}'))$, in Eq. (8).

Finally, we used the CAR loss to enforce two essential properties of the explanation function (see Fig. 3.b):

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

We enforce these two properties by the following loss,

$$\mathcal{L}_{\mathrm{rec}}(E, G) = \mathcal{L}_{\mathrm{rec}}(\mathbf{x}, \mathcal{I}_f(\mathbf{x}, f(\mathbf{x}))) + \mathcal{L}_{\mathrm{rec}}(\mathbf{x}, \mathcal{I}_f(\mathcal{I}_f(\mathbf{x}, \mathbf{c}), f(\mathbf{x}))). \tag{9}$$

where $\mathcal{L}_{\mathrm{rec}}(\cdot)$ is defined in Eq. (8). We minimize this loss only while reconstructing the input image (either by performing self or cyclic reconstruction). Please note, the classifier f and support networks $S(\cdot)$ and $O(\cdot)$ remained fixed during training.

3. Implementation and evaluation

3.1. Dataset

We performed our experiments on MIMIC-CXR (Johnson et al., 2019) dataset consisting of 377 K CXR images from 65 K patients. The dataset provides image-level labels for the presence of 14 observations,

namely, enlarged cardiomediastinum, cardiomegaly, lung-lesion, lung-opacity, edema, consolidation, pneumonia, atelectasis, pneumothorax, pleural effusion, pleural other, fracture, support devices and no-finding.

3.2. Implementation details

Classification model: We consider classification model that take as input a single-view chest radiograph and output the probability of each of the 14 observations. Following the prior work on diagnosis classification (Irvin et al., 2019), we used DenseNet-121 (Huang et al., 2016) architecture for the classifier. We use the Adam optimizer with default β-parameters of $\beta_1 = 0.9$, $\beta_2 = 0.999$ and learning rate 1×10^{-4} which is fixed for the duration of the training. We used a batch size of 16 images and train for 3 epochs, saving checkpoints every 4800 iterations.

The classifier is trained on 198 K (∼80%) frontal view CXR from 51 K patients and is test on a held-out set of 50 K images from 12 K non-overlapping patients. The images are resized to 256×256 and are pre-processed using a standard pipeline involving cropping, re-scaling, and intensity normalization. Our classification model achieved an AUC-ROC of 0.87 for Cardiomegaly, 0.95 for pleural effusion, and 0.91 for edema. These results are comparable to performance of the published model (Irvin et al., 2019).

Segmentation network: Semantic segmentation network $S(\cdot)$ is a 2D U-Net (Ronneberger et al., 2015) that marks the lung and the heart contour in a CXR. In the absence of ground truth lung and heart segmentation on the MIMIC-CXR dataset, we pre-trained the segmentation network trained on 385 CXRs from Japanese Society of Radiological Technology (JSRT) (van Ginneken et al., 2006) and Montgomery (Jaeger et al., 2014) datasets. The pre-trained segmentation network is used in our explanation function to enforce CARL loss and to compute Cardio Thoracic Ratio (CTR). Please refer SM-Sec 6.6 for details on segmentation network.

Object detector: We trained a Fast Region-based CNN (Ren et al., 2015) network as object detector $O(\cdot)$. We trained three independent detectors for three use-cases: detecting foreign objects (FO) such as pacemakers and hardware, detecting healthy costophrenic (CP) recess and detecting blunt CP recess.

For constructing a training dataset for this object detection, we first collect candidate CXRs for each object by parsing the radiology reports associated with the CXR to find positive mention for *"blunting of the costophrenic angle"* for blunt CP recess, and *"lungs are clear"* for healthy CP recess. For each object, we manually collect bounding box annotations for 300 candidate CXRs.

Explanation Function: Our explanation function is implemented using TensorFlow version 2.0 and is trained on NVIDID P100 GPU. Before training the explanation function, we assume access to the pre-trained classification function, that we aim to explain. We also assume access to pre-trained segmentation and object detection networks, that are used to enforce CARL loss.

In cGAN, we adapted a ResNet (He et al., 2016) architecture for the encoder, generator, and discriminator networks. For optimization, we used Adam optimizer (Kingma and Ba, 2015), with hyper-parameters set to $\alpha = 0.0002, \beta_1 = 0, \beta_2 = 0.9$ and updated the discriminator five times per one update of the generator, and the encoder.

In our experiments, we train three independent explanation functions, for explaining classifier's decision for three class labels; cardiomegaly, pleural effusion (PE), and edema. For training, we divide $f(\mathbf{x}) \in [0, 1]$ into $N = 10$ equally-sized bins and trained the cGAN with 10 conditions. To construct the training-set for the explanation function, we randomly sample images from the test-set of the classifier such that each condition (bin-index) have 2500–3000 images. Similarly, we created a non-overlapping (unique subjects) evaluation dataset, of 20 K images for the explanation function. We created one such dataset for each class label.

3.3. Evaluation

For evaluating the explanations, we randomly sample two groups of real images from the test set of the explanation function (1) a *real-negative* group defined as $\mathcal{X}^n = \{\mathbf{x}; f(\mathbf{x}) < 0.2\}$. It consists of real CXR that are predicted as negative by the classifier f for a target class k. (2) A *real-positive* group defined as $\mathcal{X}^p = \{\mathbf{x}; f(\mathbf{x}) > 0.8\}$. For $\mathcal{X}^n$, we generated a counterfactual group by setting condition $\mathbf{c} = 1.0$ as, $\mathcal{X}_{\mathbf{c}}^{n \to p} = \{\mathcal{I}_f(\mathbf{x}, \mathbf{c} = 1) \forall \mathbf{x} \in \mathcal{X}^n\}$. Similarly for $\mathcal{X}^p$, we derived a counterfactual group as $\mathcal{X}_{\mathbf{c}}^{p \to n} = \{\mathcal{I}_f(\mathbf{x}, \mathbf{c} = 0) \forall \mathbf{x} \in \mathcal{X}^p\}$. We create one such dataset for each target class k. Combining the two groups, our set of real images is $\mathcal{X} = \mathcal{X}^n \cup \mathcal{X}^p$ and corresponding set of counterfactual explanations is $\mathcal{X}_{\mathbf{c}} = \mathcal{X}_{\mathbf{c}}^{n \to p} \cup \mathcal{X}_{\mathbf{c}}^{p \to n}$. All the results are computed on this evaluation dataset.

We employ several metrics to quantify different aspects of a valid counterfactual explanation.

Frechet Inception Distance (FID) score: FID score quantifies the visual similarity between the real images and the synthetic counterfactuals. It computes the distance between the activation distributions as follow,

$$\text{FID}(\mathcal{X}, \mathcal{X}_{\mathbf{c}}) = \|\mu_{\mathbf{x}} - \mu_{\mathbf{x}_{\mathbf{c}}}\|_2^2 + \text{Tr}(\Sigma_{\mathbf{x}} + \Sigma_{\mathbf{x}_{\mathbf{c}}} - 2(\Sigma_{\mathbf{x}} \Sigma_{\mathbf{x}_{\mathbf{c}}})^{\frac{1}{2}}), \quad (10)$$

where μ's and Σ's are mean and covariance of the activation vectors derived from the penultimate layer of a Inception v3 network (Heusel et al., 2017) pre-trained on MIMIC-CXR dataset.

Counterfactual Validity (CV) score: CV score (Mothilal et al., 2020) is defined as the fraction of counterfactual explanations that successfully flipped the classification decision *i.e.,* if the input image is negative, then the explanation is predicted as positive for the target class. CV score is computed as,

$$\text{CV}(\mathcal{X}, \mathcal{X}_{\mathbf{c}}) = \frac{\mathbb{1}(\|f(\mathbf{x}) - f(\mathbf{x}_{\mathbf{c}})\| > \tau)}{|\mathcal{X}|} \quad (11)$$

where, τ is the margin between the two prediction distributions. We used $\tau = 0.8$ in our experiments.

Foreign Object Preservation (FOP) score: FOP score is the fraction of the real images, with successful detection of FO, in which FO was also detected in the corresponding explanation image $\mathbf{x}_{\mathbf{c}}$.

$$\text{FOP}(\mathcal{X}, \mathcal{X}_{\mathbf{c}}) = \frac{\mathbb{1}(O(\mathbf{x}_{\mathbf{c}}) > 0.5)}{|\mathcal{X}|} \quad (12)$$

where, $O(\mathbf{x})$ is the probability of finding a FO in image $\mathbf{x}$ as predicted by a pre-trained object detector. $O(\mathbf{x}) > 0.5 \forall \mathbf{x} \in \mathcal{X}$ *i.e.,* we consider images with positive detection of FO in set $\mathcal{X}$.

Next, we define two clinical metrics to quantify the counterfactual changes that leads to the flipping of the classifier's decision. Precisely, we translated the clinical definition of cardiomegaly and pleural effusion into metrics that can be computed using a CXR.

Cardio Thoracic Ratio (CTR): We used CTR as the clinical metric to quantify cardiomegaly. CTR is the ratio of the cardiac diameter to the maximum internal diameter of the thoracic cavity. A CTR ratio greater than 0.5 indicates cardiomegaly (Mensah et al., 2015; Centurión et al., 2017; Dimopoulos et al., 2013). We followed the approach in (Chamveha et al., 2020) to calculate CTR from a CXR. We use the pre-trained segmentation network $S(\cdot)$ to mark the heart and lung region. 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 the horizontal distance between the widest points on the lung mask.

Score for detecting a healthy Costophrenic recess (SCP): We first identify CP recess in a CXR and then classify it as healthy or blunt to quantify pleural effusion. The fluid accumulation in 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 blunt CPA is an indication of pleural effusion (Maduskar et al., 2013, 2016; Lababede, 2017). Marking the
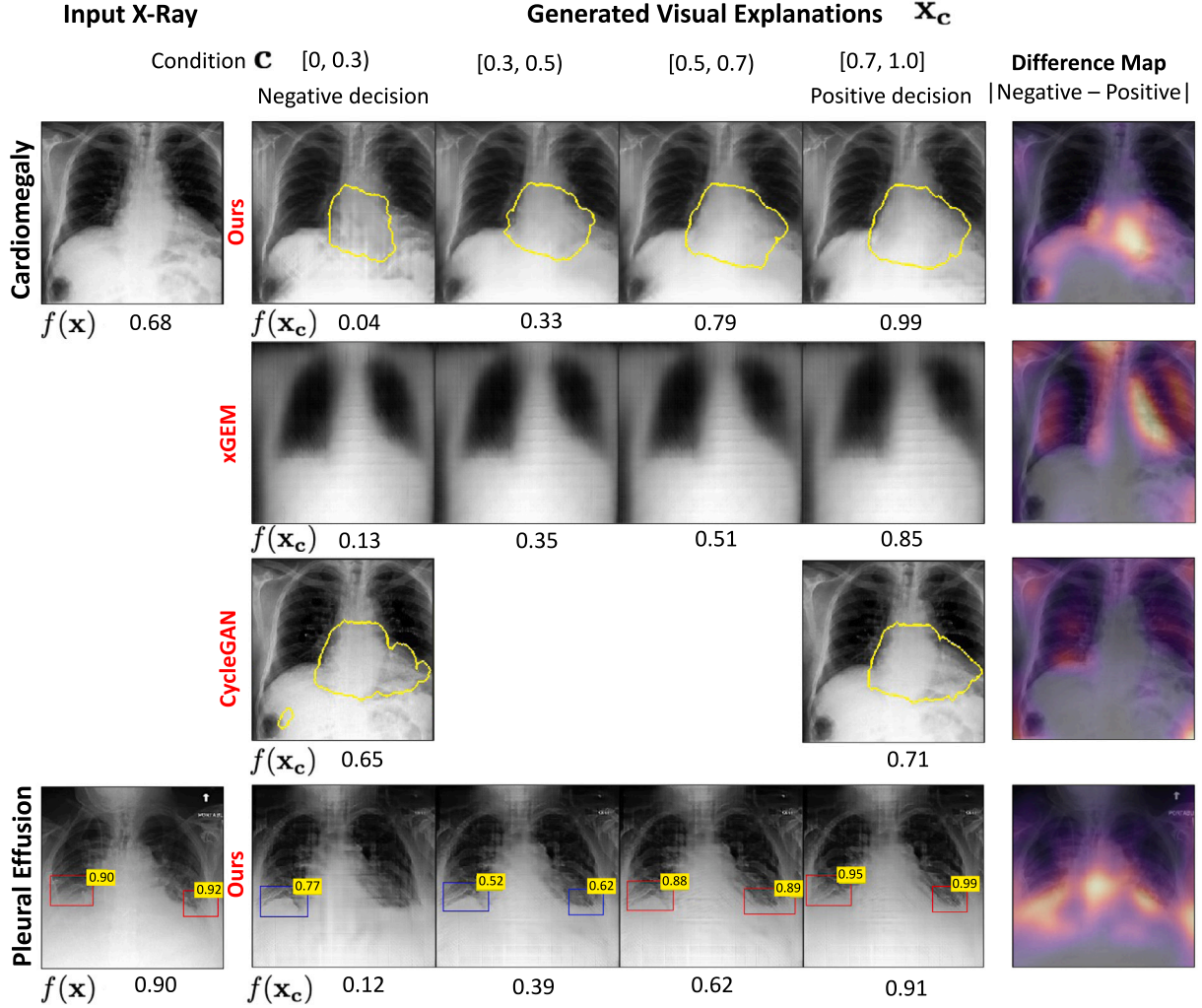
Fig. 4. Qualitative comparison of the counterfactual explanations generated for two classification tasks, cardiomegaly (first row) and pleural effusion (PE) (last row). The bottom labels are the classifier's predictions for the specific task. For the input image in first column, our model generates a series of images x_c as explanations by gradually changing c in range $[0, 1]$. The last column presented a pixel-wise difference map between the explanations at the two extreme ends *i.e.*, with condition $c = 0$ (negative decision) and with condition $c = 1$ (positive decision). The heatmap highlights the regions that changed the most during the transformation. For cardiomegaly, we show the heart border in yellow. For PE, we showed the results of an object detector as a bounding-box (BB) over the healthy (blue) and blunt (red) CP recess regions. The number on the top-right of the blue-BB is the Score for detecting a healthy CP recess (SCP). The number on red-BB is 1-SCP.

CPA angle on a CXR requires expert supervision, while annotating the CP region with a bounding box is a much simpler task (*see* SM-FIg. 15). We learned an object detector to identify healthy or blunt CP recess in the CXRs and used SCP as our evaluation metric.

4. Experiments and results

We performed four sets of experiments on CXR dataset:

(1) In Section 4.1, we evaluated the validity of our counterfactual explanations and compared them against xGEM (Joshi et al., 2018) and CycleGAN (Narayanaswamy et al., 2020; DeGrave et al., 2020).

(2) In Section 4.2, we compared against the saliency-based methods to provide *post-hoc* model explanation.

(3) In Section 4.3, we associate the counterfactual changes in our explanation with the clinical definitions of two diagnosis, cardiomegaly and pleural effusion.

(4) In Section 4.4, we present a clinical study that collects subjective feedback from radiology residents on three different explanation approaches, saliency maps, cycleGAN and ours.

4.1. Validity of counterfactual explanations

A valid counterfactual explanation resembles the query image while having perceivable differences that achieves an opposite classification decision as compared to the query image from the classifier. In Fig. 4, we present qualitative examples of counterfactual explanations from our method and compared them against those obtained from xGEM and CycleGAN.

4.1.1. Classifier consistency

In Fig. 4, we observe that the explanation images gradually flip their decision $f(x_c)$ (bottom label) as we go from left to right. Table 1 summarizes our results on CV score metric. A high CV score for our model confirms that the condition used in cGAN is successfully met and the generated explanations are successfully flipping the classification decision and hence, are consistent with the classifier.

On the other hand, cycleGAN achieves a CV score of about 50%, thus creating valid counterfactual images only half of the times. In a deployment scenario, a counterfactual explanation that fails to flip the classification decision would be rejected as an invalid, and hence half of the explanations provided by cycleGAN would be rejected.

Table 1

The counterfactual validity (CV) score is the fraction of explanations that have an opposite prediction compared to the input image. The FID score quantifies the visual appearance of the explanations. We have normalized the FID scores with respect to the best method (cycleGAN).

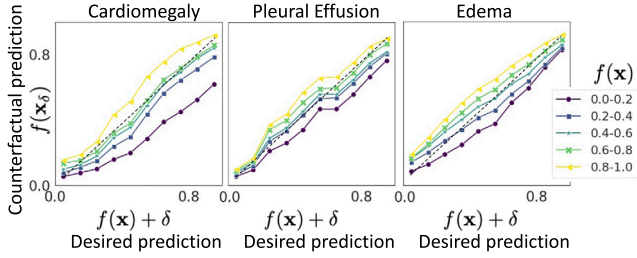|	Cardiomegaly			Pleural effusion			Edema		
	Ours	xGEM	CycleGAN	Ours	xGEM	CycleGAN	Ours	xGEM	CycleGAN
Counterfactual validity score									
$CV(\mathcal{X}, \mathcal{X}_c)$	**0.91**	**0.91**	0.43	**0.97**	**0.97**	0.49	**0.98**	0.66	0.57
FID score									
Negative group: $FID(\mathcal{X}^n, \mathcal{X}_c^{n \to p})$	4.3	12.5	1	3.7	9.2	1	1.9	5.0	1
Positive group: $FID(\mathcal{X}^p, \mathcal{X}_c^{p \to n})$	2.4	5.6	1	3.4	10.1	1	1.3	3.5	1

Fig. 5. The plot of condition, c (desired prediction), against actual response of the classifier on generated explanations, $f(x_c)$. Each line represents a set of input images with prediction $f(x)$ in a given range. Plots for xGEM and cycleGAN are shown in SM-Fig. 18.

Our formulation constraints the condition c to vary linearly with the actual prediction $f(x_c)$ *i.e.,* if we increase c in range $[0, 1]$ then the cGAN should create an image x_c such that condition c is met and the actual prediction $f(x_c)$ should also increase. Further, consider a scenario when $c = 1.0$. The expected behavior is $f(x_{c=1.0}) \approx 1.0$ and also, $f(x_{c=1.0}) > f(x_{c=0.9}) > f(x_{c=0.8}) > \cdots > 0$, where different x_c are generated using same x but different conditions.

To verify this behavior, we group images in the test-set of the explanation function, into five non-overlapping groups based on their original prediction $f(x)$. Next, for each image, we created 10 explanation images by discretizing the range $[0, 1]$ into 10 bins. In Fig. 5, we represented each input group as a line and plotted the average response of the classifier *i.e.,* $f(x_c)$ for explanations generated with a same condition c. The positive slope of the line-plot, parallel to $y = x$ line confirms that starting from images with low $f(x)$, our model creates fake images such that $f(x_c)$ is high and vice-versa.

4.1.2. Visual quality

Qualitatively, the counterfactual explanations generated by our method look visually similar to the query image (*see* Fig. 4). Table 1 reports the FID score for our method and compares them against xGEM and cycleGAN. Our approach achieved a lower FID score as compared to xGEM. xGEM has very high FID score, thus creating counterfactual images that are very different from the query image and hence are unsuitable for deployment. CycleGAN achieved the least FID score, thus generating the most realistic images as explanations.

4.1.3. Identity preservation

Ideally, a counterfactual explanation should differ in semantic features associated with the target task while retaining unique properties of the input, such as foreign objects (FOs). FO provide critical information to identify the patient in an X-ray. The disappearance of FO in explanation images may create confusion that explanation images show a different patient.

In this experiment, we quantify the strength of our revised CARL loss in preserving FO in explanation images compared to an image-level ℓ_1 reconstruction loss. In Table 2, we report the results on the FOP score metric. Our model with CARL obtained a higher FOP score. The FO

Table 2

The foreign object preservation (FOP) score for our model with and without the context-aware reconstruction loss (CARL). FOP score depends on the performance of FO detector.

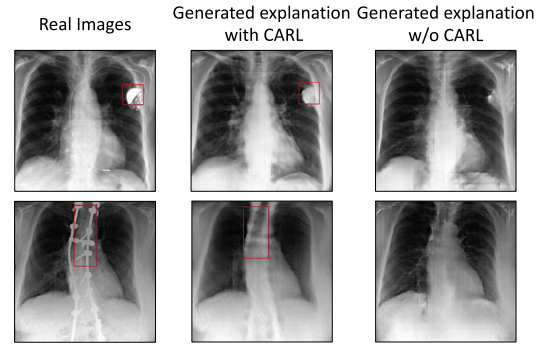| Foreign object (FO)	FOP score	
	Ours with CARL	Ours with ℓ_1
Pacemaker	**0.52**	0.40
Hardware	**0.63**	0.32

Fig. 6. Fidelity of generated images with respect to preserving FO.

detector network has an accuracy of 80%. Fig. 6 presents examples of counterfactual explanations generated by our model with and without the CARL. Our results confirm that CARL is an improvement over ℓ_1 reconstruction loss. We further provide a detailed ablation study over different components of our loss in SM-Sec.6.13.

4.2. Comparison with saliency maps

Popular existing approaches for post-hoc model explanation includes explaining using a saliency-map (Pasa et al., 2019; Irvin et al., 2019). To compare against such methods, we approximated a saliency map as a pixel-wise difference map between the explanations at the two extreme ends *i.e.,* with condition $c = 0$ (negative decision) and with condition $c = 1$ (positive decision). For proper comparison, we normalized the absolute values of the saliency maps in the range $[0, 1]$.

In clinical setting, multiple diagnosis may affect the same anatomical region. In this case, the saliency map may highlight same regions as important for multiple target tasks. Fig. 8 is showing one such example. Our method not only provides a saliency map, but also counterfactual images to demonstrate how image features in those relevant regions should be modify to change the classification decision.

Quantitatively evaluation: In this experiment, we quantitatively compare different methods for generating saliency maps, to show that important regions identified by these methods are actually relevant for classification decision. Specifically, we used the *deletion* evaluation metric (Petsiuk et al., 2018; Samek et al., 2017). For each image in set $\mathcal{X}^p$, we derived saliency maps using different methods. We used the saliency information to sort the pixels based on their importance.
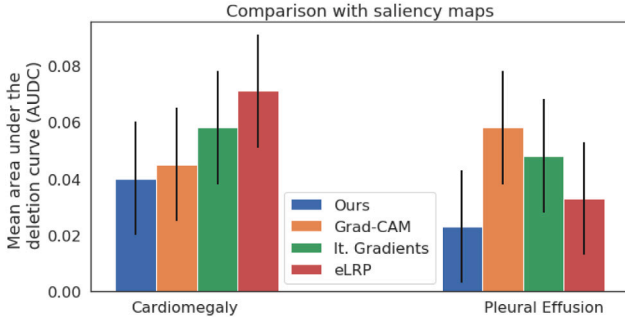
Fig. 7. Quantitative comparison of our method against gradient-based methods. Mean area under the deletion curve (AUDC), plotted as a function of the fraction of removed pixels. A low AUDC shows a sharp drop in prediction accuracy as more pixels are deleted.

Next, we gradually removed top $x\%$ of important pixels by selectively impainting the removed region based on its surroundings. We processed the resulting image through the classifier and measure the output probability. We repeated this process while gradually increasing the fraction of removed pixels.

For each image, we plotted the updated classification probability as a function of the fraction of removed pixels, to obtain the deletion curve and measure its area under the deletion curve (AUDC). A sharp decline in classification probability shows that the removed pixels were actually important for classification decision. A sharp decline results in smaller AUDC, and demonstrates the high sensitivity of the classifier in the salient regions. In Fig. 7, we reported the mean and standard deviation of the AUDC metric over the set $\mathcal{X}^p$. Our method achieved the lowest AUDC, confirming the high sensitivity of the classifier in the salient regions identified by our method.

4.3. Disease-specific evaluation

In this experiment, we demonstrated the clinical relevance of our explanations. We defined two clinical metrics, cardiothoracic ratio (CTR) for cardiomegaly and score of the normal costophrenic recess (SCP) for pleural effusion. We used these metrics to quantify the counterfactual changes between normal (negative diagnosis) and abnormal (positive diagnosis) populations, as identified by the given classifier. If the change in classification decision is associated with the corresponding change in clinical-metric, we can conclude that the classifier considers clinically relevant information in its diagnosis prediction.

We conducted a paired t-test to determine the effect of counterfactual perturbation on the clinical metric for the respective diagnosis. To perform the test, we considered the two groups of real images $\mathcal{X}^n$ and $\mathcal{X}^p$ and their corresponding counterfactual groups $\mathcal{X}_{\mathbf{c}}^{n\rightarrow p}$ and $\mathcal{X}_{\mathbf{c}}^{p\rightarrow n}$ respectively. In Fig. 9, we showed the distribution of differences in CTR for cardiomegaly and SCP for PE in a pair-wise comparison between real images and their respective counterfactuals. Patients with cardiomegaly have higher CTR as compared to normal subjects. Consistent with clinical knowledge, in Fig. 9, we observe a negative mean difference for CTR($\mathcal{X}^n$) - CTR($\mathcal{X}_{cf}^{n\rightarrow p}$) (a p-value of < 0.0001) and a positive mean difference for CTR($\mathcal{X}^a$) - CTR($\mathcal{X}_{cf}^n$) (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 high SCP to a healthy CP recess with no evidence of blunting CPW. Consistent with our expectation, we observe a positive mean difference for SCP($\mathcal{X}^n$) - SCP($\mathcal{X}_{\mathbf{c}}^{n\rightarrow p}$) (with a p-value of $\ll 0.0001$) and a negative mean difference for SCP($\mathcal{X}^p$) - SCP($\mathcal{X}_{\mathbf{c}}^{p\rightarrow n}$) (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.

4.4. Human evaluation

We conducted a human-grounded experiment with diagnostic radiology residents to compare different styles of explanations (no explanation, saliency map, cycleGAN explanation, and our counterfactual explanation) by evaluating different aspects of explanations: (1) understandability, (2) classifier's decision justification, (3) visual quality, (d) identity preservation, and (5) overall helpfulness of an explanation to the users.

Our results show that our counterfactual explanation was the only explanation method that significantly improved the users' understanding of the classifier's decision compared to the no-explanation baseline. In addition, our counterfactual explanation had a significantly higher classifier's decision justification than the cycleGAN explanation, indicating that the participants found a good evidence for the classifier's decision more frequently in our counterfactual explanation as compared to cycleGAN explanation.

Further, cycleGan explanation performed better in terms of visual quality and identity preservation. However, at times the cycleGAN explanations were identical to the query image, thus providing inconclusive explanations. Overall the participants found our explanation method the most helpful method in understanding the assessment made by the AI system in comparison to other explanation methods. Below, we describe the design of the study, the data analysis methods, along with the results of the experiment in detail.

4.4.1. Experiment design

We conducted an online survey experiment with 12 diagnostic radiology residents. Participants first reviewed an instruction script, which described the AI system developed to provide an autonomous diagnosis for CXR findings such as cardiomegaly. The study comprised of the radiologists evaluating six CXR images which were presented in random order to them. For selecting these six CXR, we first, divided the test-set of the explanation function for cardiomegaly into three groups, positive ($f(\mathbf{x}) \in [0.8, 1.0]$), mild ($f(\mathbf{x}) \in [0.4, 0.6]$) and negative ($f(\mathbf{x}) \in [0.0, 0.2]$). Next, we randomly selected two CXR images from each group. The six CXR images were anonymized as part of the MIMIC-CXR dataset protocol. For each image, we had the same procedure consisted of a diagnosis tasks, followed by four explanation conditions, and ended by a final evaluation question between the explanation conditions. Further details of the study design are includes in SM-Section 6.1.

Diagnosis: For each CXR image, we first asked a participant to provide their diagnosis for cardiomegaly. This question ensures that the participants carefully consider the imaging features that helped them diagnose. Subsequently, the participants were presented with the classifier's decision and were asked to provide feedback on whether they agreed.

Explanation Conditions: Next, the study provides the classifier's decision with the following explanation conditions:

1. **No explanation (Baseline):** This condition simply provides the classifier decision without any explanation, and is used as the control condition.
2. **Saliency map:** A heat map overlaid on the query CXR, highlighting essential regions for the classifier's decision.
3. **CycleGAN explanation:** A video loop over two CXR images, corresponding to the query CXR transformation with a negative and a positive decision for cardiomegaly.
4. **Our counterfactual explanation:** A video showing a series of CXR images gradually changing the classifier's decision from negative to positive.

Please note that after showing the baseline condition, we provided the other explanation conditions in random order to avoid any learning or biasing effects.

Evaluation metrics: Given the classifier's decision and corresponding explanation, we consider the following metrics to compare different explanation conditions:
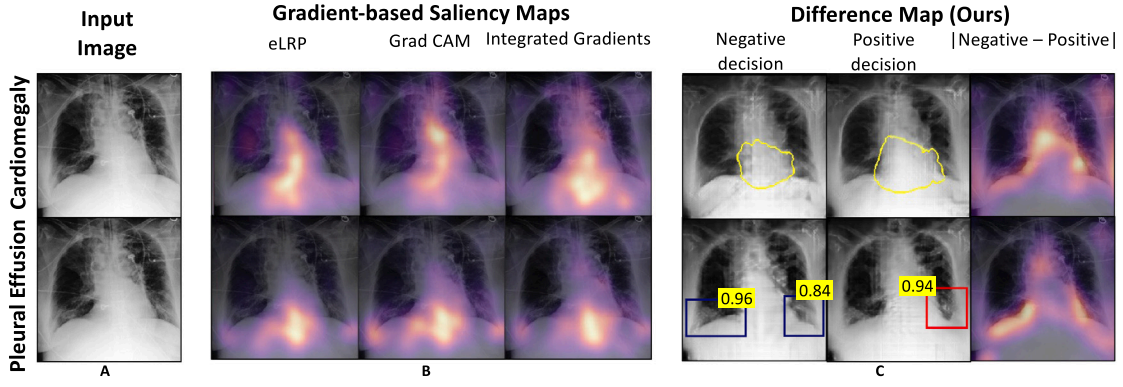
Fig. 8. Comparison of our method against different gradient-based methods. A: Input image; B: Saliency maps from existing works; C: Our simulation of saliency map as difference map between the normal and abnormal explanation images. More examples are shown in SM-Fig. 21.
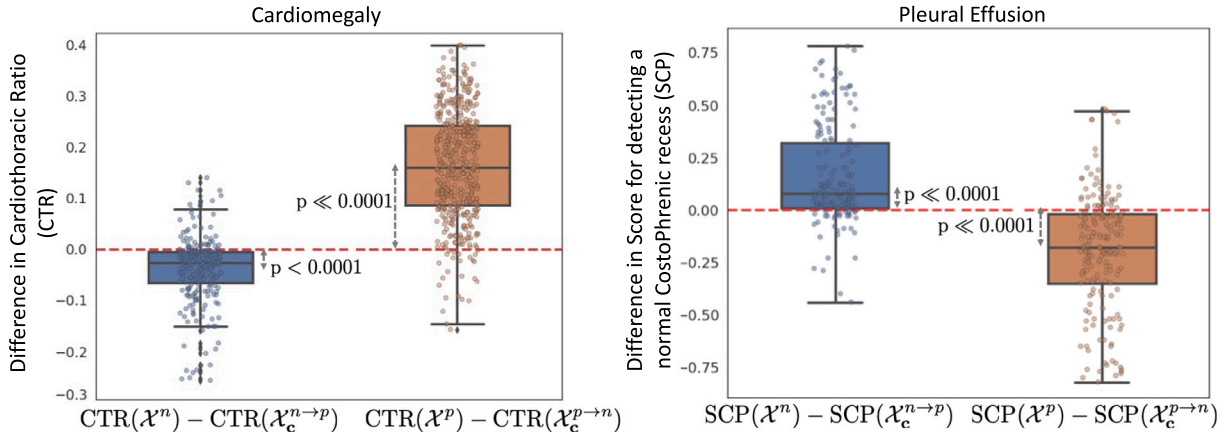

Fig. 9. Box plots to show distributions of pairwise differences in clinical metrics, CTR for cardiomegaly and the Score of normal CP recess (SCP) for pleural effusion, before (real) and after (counterfactual) our generative counterfactual creation process. The mean value corresponds to the average causal effect of the clinical-metric on the target task. 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 8.

1. **Understandability**: For each explanation condition, the study included a question to measure whether the end-user understood the classifier's decision, when explanation was provided. The participants were asked to rate agreement with *"I understand how the AI system made the above assessment for Cardiomegaly"*.

2. **Classifier's decision justification**: Human user's may perceive explanations as the reason for the classifier's decision. For the cycleGAN and our counterfactual explanation conditions, we quantify whether the provided explanation were actually related to the classification task by measuring the participants' agreement with *"The changes in the video are related to Cardiomegaly"*.

3. **Visual quality**: The study quantifies the proximity between the explanation images and the query CXR by measuring the participants' agreement with *"Images in the video look like a chest X-ray."*.

4. **Identity preservation**: The study also measures the extent to which participants think the explanation images correspond to the same subject as the query CXR by measuring the participants' agreement with *"Images in the video look like the chest X-ray from a given subject"*.

5. **Helpfulness:** For each CXR image, we asked the participants to select the most helpful explanation condition in understanding the classifier's decision, *"Which explanation helped you the most in understanding the assessment made by the AI system?"*. This evaluation metric directly compares the different explanation conditions.

All metrics, but the helpfulness metric were evaluated for agreement on a 5-point Likert scale, where one means *"strongly disagree"* and five means *"strongly agree"*.

Free-form Response: After each question, we also asked the participants a free-form question: *"Please explain your selection in a few words."* We used answers to these questions to triangulate our findings and complement our quantitative metrics by understanding our participants' thought-processes and reasoning.

Participants. Our participants include 12 diagnostic radiology residents who have completed medical school and have been in the residency program for one or more years. On average, the participants finished the survey in 40 min and were paid $100 for their participation in the study.

4.4.2. Data analysis

For each evaluation metric, the study asked the same question to the participants while showing them different explanations. For each question, we gather 72 responses (6 - number of CXR images × 12 - number of participants).

For the understandability and helpfulness metrics, we conducted a one-way ANOVA test to determine if there is a statistically significant difference between the mean metric scores for the four explanation conditions. Specifically, we built a one-way ANOVA with the metric as our dependent variable and analyzed agreement rating as the independent variable. If we found a significant difference in the ANOVA method, we ran Tukey's Honestly Significant Difference (HSD) posthoc test to perform a pair-wise comparison between different explanation conditions.
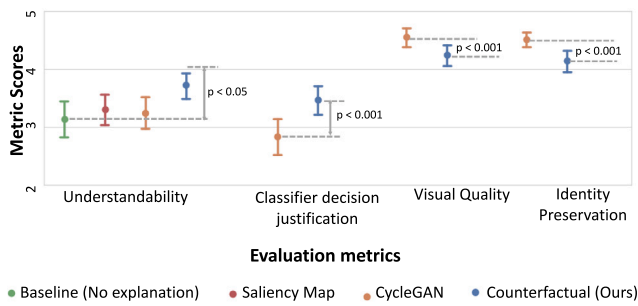
Fig. 10. Comparing the evaluation metrics of understandability, classifier's decision justification, visual quality, and identity preservation across the different explanation conditions.

We measured the classifier's decision justification, visual quality and identity preservation metrics only for the cycleGAN and our counterfactual explanations. We conducted paired t-tests to compare these evaluation metrics between these two explanation conditions. We also qualitatively analyzed the participants' free-form responses to find themes and patterns in their responses.

4.4.3. Results

Fig. 10 shows the mean score for the evaluation metrics of understandability, classifier's decision justification, visual quality, and identity preservation among the different explanation conditions. Below, we report the statistical analysis for these results, followed by analysis of the participants' free-form responses to understand the reasons behind these results.

Understandability: The results show that our counterfactual explanation was the most understandable explanation to the participants. A one-way ANOVA revealed that there was a statistically significant difference in the understandability metric between at least two explanation conditions ($F(3, 284) = [3.39]$, $p = 0.019$). The Tukey post-hoc test showed that the understandability metric for our counterfactual explanation was significantly higher than the no-explanation baseline ($p = 0.018$). However, there was no statistically significant difference in mean scores between other pairs of explanations (refer to Table 3, "Understandability" column). This finding indicates that providing our counterfactual explanations along with the classifier's decision made the algorithm most understandable to our clinical participants, while other explanation conditions, saliency map and cycleGAN failed to achieve significant difference from no-explanation baseline on the understandability metric. Next, we use responses from free-text question to supplement our findings.

For the no-explanation baseline, the primary reason for poor understanding was the absence of explanation ($n = 30$), (*e.g.,* they stated that *"there is no indication as to how the AI made this decision"*). Interestingly, many responses ($n = 23$) either associated their high understanding with the correct classification decision *i.e.,* participants understood the decision as the decision is correct (*"I agree, it is small and normal"*) or they assumed the AI-system is using similar reasoning as them to arrive at its decision (*"I assume the AI is just measuring the width of the heart compared to the thorax", "Assume the AI measured the CT ratio and diagnosed accordingly."*).

Participants' mostly found saliency maps to be correct but incomplete ($n = 23$), (*"Unclear how assessment can be made without including additional regions"*). Specifically, for cardiomegaly, the saliency maps were highlighting parts of the heart and not its border (*"Not sure how it gauges not looking at the border"*) or thoracic diameter (*"thoracic diameter cannot be assessed using highlighted regions of heat map"*). We observe a similar result in Fig. 8, where the heatmap focuses on the heart but not its border. Further, some participants expressed a concern that they didn't understand how relevant regions were used to derive

the decision (*"i understand where it examined but not how that means definite cardiomegaly"*).

For cycleGAN explanation, the primary reason for poor understanding was the minimal perceptible changes between the negative and positive images ($n = 3$), (*"There is no change in the video".*). In contrast, many participant's explicitly reported an improved understanding of the classifier's decision in the presence of our counterfactual explanations ($n = 33$), (*"I think the AI looking at the borders makes sense"., "i can better understand what the AI is picking up on with the progression video"*).

Classifier's decision justification: Our counterfactual explanation ($M = 3.46$; $SD = 1.12$) achieved a positive mean difference of 0.63 on this metric as compared to cycleGAN ($M = 2.83$; $SD = 1.33$), with $t(71) = 3.55$ and $p < 0.001$. This result indicates that the participants found a good evidence for the predicted class (cardiomegaly), much frequently in our counterfactual explanations as compared to cycleGAN.

Most responses ($n = 25$) explicitly mentioned visualizing changes related to cardiomegaly such as an enlarged heart in our explanation video as compared to cycleGAN ($n = 17$). In cycleGAN, many reported that changes in the explanation video was not perceptible ($n = 23$). Further, the participants reported changes in density, windowing level or other attributes which were not related to cardiomegaly (*"Decreasing the density does not impact how I assess for cardiomegaly"., "they could be or just secondary to windowing the radiograph"*). Such responses were observed in both cycleGAN ($n = 17$) and our explanation ($n = 17$). This indicates that the classifier may have associated such secondary information (short-cuts) with cardiomegaly diagnosis. A more in-depth analysis is required to quantify the classifiers' behavior.

Visual quality and identity preservation: We observe a negative mean difference of 0.31 and 0.37 between our and cycleGAN explanation methods in visual quality and identity preservation metrics, respectively. The mean score for visual quality was higher for cycleGAN ($M = 4.55$; $SD = 0.71$) as compared to our method ($M = 4.24$; $SD = 0.80$) with $t(71) = 3.49$ and $p < 0.001$. Similarly, the mean score for identity preservation was also higher for cycleGAN ($M = 4.51$; $SD = 0.56$) as compared to our method ($M = 4.14$; $SD = 0.78$) with $t(71) = 3.96$ and $p < 0.001$.

Most of the responses ($n = 69$) agreed that the CycleGAN explanation were marked as highly similar to the query CXR image. These results are consistent with our earlier results, that cycleGAN has better visual quality with a lower FID score (*see* Table 1). However, in some responses, the participants pointed out that the explanation images were almost identical to the query image (*"There's virtually no differences. This is within the spectrum of a repeat chest X-ray for instance".*). An explanation image identical to the query image provides no information about the classifier's decision. Further, similar looking CXR will also result in similar classification decision, and hence will fail to flip the classification decision. As a result, we also observed a lower agreement in the classifier consistency metric and a lower counterfactual validity score in Table 1 for cycleGAN.

Helpfulness: In our concluding question, *"Which explanation helped you the most in understanding the assessment made by the AI system?"*, 57% of the responses selected our counterfactual explanation as the most helpful method. A one-way ANOVA revealed that there was a statistically significant difference in the helpfulness metric between at least two explanation conditions ($F(3, 284) = [21.5]$, $p < 0.0001$). In pair-wise Tukey's HSD posthoc test, we found that the mean usefulness metric for our counterfactual explanations was significantly different from all the rest explanation conditions($p < 0.0001$). Table 3 ("Helpfulness" column) summarizes these results.

These results indicates that the participant's selected our counterfactual explanations as the most helpful form of explanation for understanding the classifier's decision.

Table 3

Results for one-way ANOVA for understandability metric, followed by Tukey's HSD post-hoc test between different levels of agreement. Note that the mean value for E4 (our counterfactual explanation) is the highest, indicating that our explanations helped users the most in understanding the classifier's decision.

Understandability $F(3, 284) = 3.39$ $p < 0.05$			Helpfulness $F(3, 284) = 21.5$ $p < 0.001$		
Explanation method		p	Explanation method		p
E1 (No explanation)	E2		E1	E2	
M = 3.14	E3		M = 0.05	E3	
SD = 1.39	E4	*	SD = 0.23	E4	***
E2 (Saliency Map)	E1		E2	E1	
M = 3.31	E3		M = 0.18	E3	
SD = 1.13	E4		SD = 0.39	E4	***
E3 (CycleGAN)	E1		E3	E1	
M = 3.24	E2		M = 0.16	E2	
SD = 1.19	E4		SD = 0.37	E4	***
E4 (Our counterfactual	E1	*	E4	E1	***
explanation) **M = 3.72**	E2		**M = 0.24**	E2	***
SD = 0.97	E3		SD = 0.42	E3	***

* $p < 0.05$.

*** $p < 0.0001$.

5. Discussion and conclusion

We provided a BlackBox *Counterfactual Explainer*, designed to explain image classification models for medical applications. Our framework explains the decision by gradually transforming the input image to its counterfactual, such that the classifier's prediction is flipped. We have formulated and evaluated our framework on three properties of a valid counterfactual transformation: data consistency, classifier consistency, and self-consistency. Our results in Section 4.1 showed that our framework adheres to all three properties.

Comparison with xGEM and cycleGAN: Our model creates visually appealing explanations that produce a desired outcome from the classification model while retaining maximum patient-specific information. In comparison, both xGEM and cycleGAN failed on at least one essential property. xGEM model fails to create realistic images with a high FID score. Furthermore, the cycleGAN model fails to flip the classifier's decision with a low CV score (~50%).

Further, we present a thorough comparison between cycleGAN and our explanation in a human evaluation study. The clinical experts' expressed high agreement that explanation images from cycleGAN were of high quality and they resembles the query CXR. But at the same time, users found the explanation images to be too similar to query CXR, and the cycleGAN explanations failed to provide the counterfactual reasoning for the decision.

In comparison, our explanation were most helpful in understanding the classification decision. Though the users reported inconsistencies in the visual appearance, but the overall sentiment looks positive and they selected our method as their preferred explanation method for improved understandability.

Clinical relevance of the explanations: From a clinical perspective, we demonstrated that the counterfactual changes associated with normal (negative) or abnormal (positive) classification decisions are also associated with corresponding changes in disease-specific metrics such as CTR and SCP. In our clinical study, multiple radiologist reported using CTR as the metric to diagnose cardiomegaly. As radiologist annotations are expensive, and it is not efficient to perform human evaluation on a large test set, our results with CTR calculations provides a quantitative way t evaluate difference in real and counterfactual populations.

We acknowledge that our GAN-generated counterfactual explanations may have missing details such as small wires. In our extended experiments, we found that the foreign objects such as pacemaker have minimal importance in the classification decision (*see* SM-Sec. 6.10.1).

We attempted to improve the preservation of such information through our revised context-aware reconstruction loss (CARL). 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 used to calculate the FOP score. Training a highly accurate FO detector is outside the scope of this study.

Further, a resolution of 256×256 for counterfactually generated images is smaller than a standard CXR. 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 must 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 largest N, 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.

To conclude, this study uses counterfactual explanations as a way to audit a given black-box classifier and evaluate whether the radiographic features used by that classifier have any clinical relevance. In particular, the proposed model did well in explaining the decision 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 its 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.

Declaration of competing interest

Data availability

Data will be made available on request.

Acknowledgments

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.media.2022.102721.

References

Agarwal, C., Nguyen, A., 2020. Explaining image classifiers by removing input features using generative models. Asian Conf. Comput. Vis..

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., Samek, W., 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLOS ONE 10, 1–46. http://dx.doi.org/10.1371/journal.pone.0130140.

Bau, D., Zhu, J.-Y., Wulff, J., Peebles, W., Strobelt, H., Zhou, B., Torralba, A., 2019. Seeing what a gan cannot generate. IEEE Int. Conf. Comput. Vis. 4502–4511.

Centurión, O.A., Scavenius, K., Miño, L., Sequeira, O.R., 2017. Evaluating cardiomegaly by radiological cardiothoracic ratio as compared to conventional echocardiography. J. Cardiol. Curr. Res. 9, http://dx.doi.org/10.15406/jccr.2017.09.00319.

Chamveha, I., Promwiset, T., Tongdee, T., Saivirooonporn, P., Chaisangmongkon, W., 2020. Automated cardiothoracic ratio calculation and cardiomegaly detection using deep learning approach. arXiv e-prints, arXiv:2002.07468.

Chang, C.-H., Creager, E., Goldenberg, A., Duvenaud, D., 2019. Explaining image classifiers by counterfactual generation. Int. Conf. Learn. Rep..

Cohen, J.P., Brooks, R., En, S., Zucker, E., Pareek, A., Lungren, M.P., Chaudhari, A., 2021. Gifsplanation via latent shift: A simple autoencoder approach to counterfactual generation for chest X-rays. Med. Imaging Deep Learn..

Dabkowski, P., Gal, Y., 2017. Real time image saliency for black box classifiers. In: 31st International Conference on Advances in Neural Information Processing Systems (NeurIPS). NIPS '17, ISBN: 9781510860964, pp. 6970–6979.

DeGrave, A.J., Janizek, J.D., Lee, S.-I., 2020. AI for radiographic COVID-19 detection selects shortcuts over signal. MedRxiv http://dx.doi.org/10.1101/2020.09.13.20193565.

Dhurandhar, A., Chen, P.-Y., Luss, R., Tu, C.-C., Ting, P., Shanmugam, K., Das, P., 2018. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. Int. Conf. Adv. Neural Inf. Process. Syst..

Dimopoulos, K., Giannakoulas, G., Bendayan, I., Liodakis, E., Petraco, R., Diller, G.-P., Piepoli, M.F., Swan, L., Mullen, M., Best, N., Poole-Wilson, P.A., Francis, D.P., Rubens, M.B., Gatzoulis, M.A., 2013. Cardiothoracic ratio from postero-anterior chest radiographs: A simple, reproducible and independent marker of disease severity and outcome in adults with congenital heart disease. Int. J. Cardiol. 166.

Eaton-Rosen, Z., Bragman, F., Bisdas, S., Ourselin, S., Cardoso, M.J., 2018. Towards safe deep learning: Accurately quantifying biomarker uncertainty in neural network predictions. Int. Conf. Med. Imag. Comput. Computer-Assisted Interv. 691–699.

Eitel, F., Ritter, K., 2019. Testing the robustness of attribution methods for convolutional neural networks in MRI-based alzheimer's disease classification. Int. Workshop Interpret. Mach. Intell. Med. Imag. Comput. 11797 LNCS, 3–11.

Fong, R.C., Vedaldi, A., 2017. Interpretable Explanations of Black Boxes by Meaningful Perturbation. IEEE Int. Conf. Comput. Vis. http://dx.doi.org/10.1109/ICCV.2017.371.

Frank, E., Hall, M., 2001. A simple approach to ordinal classification. Eur. Conf. Mach. Learn. 145–156.

Gastounioti, A., Kontos, D., 2020. Is It Time to Get Rid of Black Boxes and Cultivate Trust in AI? Radiol.: Artif. Intell. 2, e200088.

van Ginneken, B., Stegmann, M.B., Loog, M., 2006. Segmentation of anatomical structures in chest radiographs using supervised methods: a comparative study on a public database. Med. Image Anal. 10, 19–40. http://dx.doi.org/10.1016/j.media.2005.02.002.

Goyal, Y., Wu, Z., Ernst, J., Batra, D., Parikh, D., Lee, S., 2019. Counterfactual visual explanations. In: 36th International Conference on Machine Learning (ICML). vol. 97, pp. 2376–2384.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. IEEE Conf. Comput. Vis. Pattern Recoginit..

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S., 2017. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. Int. Conf. Adv. Neural Inf. Process. Syst..

Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L.H., Aerts, H.J., 2018. Artificial intelligence in radiology. Nat. Rev. Cancer 18, 500–510. http://dx.doi.org/10.1038/s41568-018-0016-5.

Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q., 2016. Densely connected convolutional networks. In: 30th IEEE Conference on Computer Vision Pattern Recoginition (CVPR). pp. 2261–2269.

Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., et al., 2019. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: 33rd AAAI Conference on Artificial Intelligence. pp. 590–597.

Jaeger, S., Candemir, S., Antani, S., Wáng, Y.-X.J., Lu, P.-X., Thoma, G., 2014. Two public chest X-ray datasets for computer-aided screening of pulmonary diseases.. Quant. Imaging Med. Surg. 4, 475–477. http://dx.doi.org/10.3978/j.issn.2223-4292.2014.11.20.

Jiang, H., Kim, B., Guan, M., Gupta, M., 2018. To trust or not to trust a classifier. Int. Conf. Adv. Neural Inf. Process. Syst. 31.

Johnson, A.E., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.Y., Mark, R.G., Horng, S., 2019. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Sci. Data 6, 317. http://dx.doi.org/10.1038/s41597-019-0322-0.

Joshi, S., Koyejo, O., Kim, B., Ghosh, J., 2018. xGEMs: Generating examplars to explain black-box models. arXiv e-prints, arXiv:1806.08867.

Joshi, S., Koyejo, O., Vijitbenjaronk, W., Kim, B., Ghosh, J., 2019. Towards realistic individual recourse and actionable explanations in black-box decision making systems. arXiv e-prints, arXiv:1907.09615.

Karras, T., Laine, S., Aila, T., 2019. A style-based generator architecture for generative adversarial networks. IEEE Conf. Comput. Vis. Pattern Recoginit. 4401–4410.

Kingma, D.P., Ba, J., 2015. Adam: A Method for Stochastic Optimization. Int. Conf. Learn. Rep..

Lababede, O., 2017. Pleural effusion imaging: Overview, radiography, computed tomography. URL https://emedicine.medscape.com/article/355524-overview.

Larrazabal, A., Nieto, N., Peterson, V., Milone, D., Ferrante, E., 2020. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. Proc. Natl. Acad. Sci. 117, 201919012. http://dx.doi.org/10.1073/pnas.1919012117.

Liu, S., Kailkhura, B., Loveland, D., Han, Y., 2019. Generative counterfactual introspection for explainable deep learning. IEEE Glob. Conf. Signal Inf. Process. 1–5. http://dx.doi.org/10.1109/GlobalSIP45357.2019.8969491.

Lundberg, S.M., Allen, P.G., Lee, S.-I., 2017. A unified approach to interpreting model predictions. In: 31st International Conference on Advances in Neural Information Processing Systems (NeurIPS). NIPS '17, pp. 4768–4777.

Maduskar, P., Hogeweg, L., Philipsen, R., van Ginneken, B., 2013. Automated localization of costophrenic recesses and costophrenic angle measurement on frontal chest radiographs. Med. Imaging: Computer-Aided Diagn..

Maduskar, P., Philipsen, R.H., Melendez, J., Scholten, E., Chanda, D., Ayles, H., Sánchez, C.I., van Ginneken, B., 2016. Automatic detection of pleural effusion in chest radiographs. Med. Image Anal..

Mensah, Y., Mensah, K., Asiamah, S., Gbadamosi, H., Idun, E., Brakohiapa, W., Oddoye, A., 2015. Establishing the cardiothoracic ratio using chest radiographs in an indigenous ghanaian population: A simple tool for cardiomegaly screening. Ghana Med. J..

Miyato, T., Koyama, M., 2018. cGANs with projection discriminator. Int. Conf. Learn. Rep..

Mothilal, R.K., Sharma, A., Tan, C., 2020. Explaining machine learning classifiers through diverse counterfactual explanations. Conf. Fairness Account. Transpar. 607–617.

Narayanaswamy, A., Venugopalan, S., Webster, D.R., Peng, L., Corrado, G.S., Ruamviboonsuk, P., Bavishi, P., Brenner, M., Nelson, P.C., Varadarajan, A.V., 2020. Scientific discovery by generating counterfactuals using image translation. Int. Conf. Med. Imag. Comput. Assisted Interv. 273–283.

Oakden-Rayner, L., Dunnmon, J., Carneiro, G., Ré, C., 2020. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. ACM Conf. Health Inf. Learn. 2020, 151–159. http://dx.doi.org/10.1145/3368555.3384468.

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

Pasa, F., Golkov, V., Pfeiffer, F., Cremers, D., Pfeiffer, D., 2019. Efficient deep network architectures for fast chest X-Ray tuberculosis screening and visualization. Sci. Rep. 9, 1–9.

Petsiuk, V., Das, A., Saenko, K., 2018. RISE: Randomized input sampling for explanation of black-box models. Br. Mach. Vis. Conf..

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

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

Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. Int. Conf. Adv. Neural Inf. Process. Syst. 28.

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

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net convolutional networks for biomedical image segmentation. Int. Conf. Med. Imag. Comput. Assisted Interv. 9351, 234–241.

Rubin, J., Sanghavi, D., Zhao, C., Lee, K., Qadir, A., Xu-Wilson, M., 2018. Large scale automated reading of frontal and lateral chest X-Rays using dual convolutional neural networks. arXiv e-prints, arXiv:1804.07839.

Samek, W., Binder, A., Montavon, G., Lapuschkin, S., Müller, K.-R., 2017. Evaluating the visualization of what a deep neural network has learned. IEEE Trans. Neural Netw. Learn. Syst. 28, 2660–2673. http://dx.doi.org/10.1109/TNNLS.2016.2599820.

Sayres, R., Taly, A., Rahimy, E., Blumer, K., Coz, D., Hammel, N., Krause, J., Narayanaswamy, A., Rastegar, Z., Wu, D., Xu, S., Barb, S., Joseph, A., Shumski, M., Smith, J., Sood, A.B., Corrado, G.S., Peng, L., Webster, D.R., 2019. Using a deep learning algorithm and integrated gradients explanation to assist grading for diabetic retinopathy. Ophthalmology 126, 552–564.

Seah, J.C., Tang, C.H., Buchlak, Q.D., Holt, X.G., Wardman, J.B., Aimoldin, A., Esmaili, N., Ahmad, H., Pham, H., Lambert, J.F., et al., 2021. Effect of a comprehensive deep-learning model on the accuracy of chest X-ray interpretation by radiologists: a retrospective, multireader multicase study. Lancet Digit. Health 3 (8), e496–e506.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. IEEE Int. Conf. Comput. Vis. 618–626.

Shrikumar, A., Greenside, P., Kundaje, A., 2017. Learning important features through propagating activation differences. In: 34th International Conference on Machine Learning (ICML). ICML '17, 70, pp. 3145–3153.

Simonyan, K., Vedaldi, A., Zisserman, A., 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. Comput. Res. Repos. arXiv:1312.6034.

Singla, S., Pollack, B., Chen, J., Batmanghelich, K., 2019. Explanation by progressive exaggeration. Int. Conf. Learn. Rep..

Singla, S., Wallace, S., Triantafillou, S., Batmanghelich, K., 2021. Using causal analysis for conceptual deep learning explanation. Int. Conf. Med. Imag. Comput. Assisted Interv. 519–528.

Springenberg, J.T., Dosovitskiy, A., Brox, T., Riedmiller, M.A., 2015. Striving for simplicity: The all convolutional net. In: International Conference on Learning Representations (ICLR-Workshop Track).

Sundararajan, M., Taly, A., Yan, Q., 2017. Axiomatic attribution for deep networks. In: 34th International Conference on Machine Learning (ICML). ICML '17, 70, pp. 3319–3328.

Tonekaboni, S., Joshi, S., McCradden, M.D., Goldenberg, A., 2019. What clinicians want: Contextualizing explainable machine learning for clinical end use. URL http://arxiv.org/abs/1905.05134.

Van Looveren, A., Klaise, J., 2019. Interpretable counterfactual explanations guided by prototypes. arXiv e-prints, arXiv:1907.02584.

Wang, F., Kaushal, R., Khullar, D., 2020. Should health care demand interpretable artificial intelligence or accept "black Box" Medicine? Ann. Intern. Med. 172, 59–61. http://dx.doi.org/10.7326/M19-2548.

Wang, P., Vasconcelos, N., 2020. SCOUT: Self-aware discriminant counterfactual explanations. IEEE Conf. Comput. Vis. Pattern Recognit..

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

Young, K., Booth, G., Simpson, B., Dutton, R., Shrapnel, S., 2019. Deep neural network or dermatologist? 11797 LNCS, pp. 48–55.

Zhou, B., Khosla, A., Lapedriza, À., Oliva, A., Torralba, A., 2015. Object detectors emerge in deep scene cnns. Int. Conf. Learn. Rep..

Zhu, J.-Y., Park, T., Isola, P., Efros, A.A., 2017. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. IEEE Int. Conf. Comput. Vis..