Machine Learning
https://doi.org/10.1007/s10994-023-06333-w

Contrastive counterfactual visual explanations with overdetermination

Adam White[1] · Kwun Ho Ngan[1] · James Phelan[1] · Kevin Ryan[1] · Saman Sadeghi Afgeh[1] · Constantino Carlos Reyes-Aldasoro[1] · Artur d'Avila Garcez[1]

Received: 4 August 2022 / Revised: 3 March 2023 / Accepted: 28 March 2023

Abstract

A novel explainable AI method called CLEAR *Image* is introduced in this paper. CLEAR *Image* is based on the view that a satisfactory explanation should be contrastive, counterfactual and measurable. CLEAR *Image* seeks to explain an image's classification probability by contrasting the image with a representative contrast image, such as an auto-generated image obtained via adversarial learning. This produces a salient segmentation and a way of using image perturbations to calculate each segment's importance. CLEAR *Image* then uses regression to determine a causal equation describing a classifier's local input–output behaviour. Counterfactuals are also identified that are supported by the causal equation. Finally, CLEAR *Image* measures the fidelity of its explanation against the classifier. CLEAR *Image* was successfully applied to a medical imaging case study where it outperformed methods such as Grad-CAM and LIME by an average of 27% using a novel pointing game metric. CLEAR *Image* also identifies cases of causal overdetermination, where there are multiple segments in an image that are sufficient individually to cause the classification probability to be close to one.

Keywords Explainable AI · Counterfactual reasoning · Image classification · Overdetermination

Adam White and Kwun Ho Ngan are joint first authors.

✉ Adam White
Adam.White.3@city.ac.uk

✉ Kwun Ho Ngan
Kwun.Ngan.3@city.ac.uk

[1] City Data Science Institute - City, University of London, London, UK

Published online: 17 May 2023

1 Introduction

Data-driven AI for Computer Vision can achieve high levels of predictive accuracy, yet the rationale behind these predictions is often opaque. This paper proposes a novel explainable AI (XAI) method called CLEAR *Image* that seeks to reveal the causal structure implicitly modelled by an AI system, where the causes are an image's segments and the effect is the AI system's classification probability. The explanations are for single predictions and describe the local input–output behaviour of the classifier. CLEAR *Image* is based on the philosopher James Woodward's seminal analysis of causal explanation (Woodward, 2003), which develops Judea Pearl's manipulationist account of causation (Pearl, 2000). Together they constitute the dominant accounts of explanation in the philosophy of science. We argue that a successful explanation for an AI system should be contrastive, counterfactual and measurable.

According to Woodward, to explain an event E is "to provide information about the factors on which it depends and exhibit how it depends on those factors". This requires a *causal equation* to describe the causal structure responsible for generating the event. The causal equation must support a set of counterfactuals; a counterfactual specifies a possible world where, contrary to the facts, a desired outcome occurs. The counterfactuals serve to illustrate the causal structure and to answer a set of 'what-if-things-had-been-different' questions. In XAI, counterfactuals usually state minimal changes needed to achieve a desired alternative outcome.

A contrastive explanation answers the question *'Why E rather than F?'*. In the philosophy literature, F is referred to as E's foil. F comes from a contrast class of events that are alternatives to E, but which did not happen (Van Fraassen, 1980). The reason why explanations should be contrastive is captured by Hilton: "The key insight is to recognise that one does not explain events per se, but that one explains why the puzzling event occurred in the target cases but not in some counterfactual contrast case" (Hilton, 1990). When a person asks for an explanation, the relevant contrast class is often not explicitly conveyed but instead is implicit in the explanatory question. For example, when a priest asked Willie Sutton why he robbed banks, Sutton's reply 'Well that's where the money is' was not a satisfactory explanation because the priest's implicit contrast was 'not robbing' but Sutton took it to be 'robbing something else' (Garfinkel, 1982). An explanation identifies the salient causes that led to E occurring rather than F.

For Woodward, all causal claims are counterfactual and contrastive: 'to causally explain an outcome is always to explain why it, rather than some alternative, occurred'. Woodward's analysis is consistent with Miller's review of over 250 papers on explanation from philosophy, psychology and cognitive science (Miller, 2018). Miller states that perhaps his most important finding is that "Explanations are contrastive — they are sought in response to particular counterfactual cases... This has important social and computational consequences for explainable AI."

Woodward's theory of explanation stands in opposition to the multiple XAI methods that claim to provide counterfactual explanations (Verma et al., 2020), but which only provide statements of single or multiple counterfactuals. As this paper will illustrate, counterfactuals without a supporting causal equation will only provide incomplete explanations. Woodward's theory also stands in opposition to XAI methods such as LIME that only provide an equation, but do not provide counterfactuals.

CLEAR *Image* identifies cases of 'causal overdetermination'. The causal overdetermination of an event occurs when two or more sufficient causes of that event occur. An example from the philosophy literature is of two vandals who each throw a rock that simultaneously shatters a window, with each rock being sufficient to shatter the window. The shattering of the window is causally overdetermined (Schaffer, 2003). This causal structure may well be ubiquitous in learning systems. For example, there may be multiple patches in a medical image, any of which being sufficient by itself to cause a classification probability close to one. To the best of our knowledge, CLEAR *Image* is the first XAI method capable of identifying causal overdetermination.

CLEAR *Image* explains an image's classification probability by comparing the image with a corresponding contrast image. In this work, the contrast image is a synthetic image created by a generative adversarial network (GAN) (Goodfellow et al., 2014). The contrast between the two images can be reflected in the pixel differences and a difference mask is created by subtracting the original image from its corresponding GAN-generated image. For example, difference masks have been previously used to visualise the difference in synthetic image generation for face forgery detection (Cao et al., 2022) and for anomaly detection in medical images (Wolleb et al., 2020). These pixel differences are good segments to start a contrastive explanation with. However, as we will illustrate, segments identified from difference masks alone can vary significantly in their relevance to a classification; furthermore, other segments critical to the classification can often be absent from the mask. Therefore, CLEAR *Image* uses a novel segmentation method that combines information from the difference mask, the original image and the classifier's behaviour. After completing its segmentation, CLEAR *Image* identifies counterfactuals and then follows a process of perturbation, whereby segments of the original image are changed, and the change in outcome is observed to produce a regression equation. The regression equation is used to determine the contribution that each segment makes to the classification probability. The regression equation is a causal equation with each independent variable referring to whether a particular segment is a *direct cause* of the classification probability. As will be shown, the explanations provided by leading XAI methods LIME and Grad-CAM may not be reliable. CLEAR *Image*, therefore, measures the fidelity of its explanations, where fidelity refers to how closely an XAI method is able to mimic a classifier's behaviour. In summary, a CLEAR *Image* explanation specifies: segmentation importance scores, counterfactuals, a regression equation, segments leading to overdetermination and fidelity errors.

By providing both a statement of counterfactuals and a supporting causal equation, CLEAR *Image* seeks to satisfy Woodward's specification for an explanation.

CLEAR *Image* was evaluated in two case studies, both involving overdetermination. The first uses a multifaceted synthetic dataset, and the second uses chest X-rays. CLEAR *Image* outperformed XAI methods such as LIME and Grad-CAM by an average of 31% on the synthetic dataset and 27% on the X-ray dataset (see Sect. 4.4) based on a pointing game metric defined in this paper for the case of multiple targets.

The contribution of this paper is four-fold. We introduce an XAI method that:

- Generates contrastive, counterfactual and measurable explanations outperforming established XAI methods in a challenging image domain;
- Uses a GAN-generated contrast image determining a causal equation, segment importance scores and counterfactuals.
- Offers novel segmentation and pointing game algorithms for the evaluation of image explanations.
- Is capable of identifying causal overdetermination, i.e. the multiple sufficient causes for an image classification.

CLEAR *Image* is a substantial development of an earlier XAI method, (Counterfactual Local Explanations viA Regression), which only applies to tabular data (White and Garcez, 2020). New functionalities include: (i) the segmentation algorithm, (ii) generating perturbed images by infilling from the corresponding GAN image, (iii) a novel pointing game suitable for images with multiple targets, (iv) identification of sufficient causes and overdetermination, (v) measurement of fidelity errors for counterfactuals involving categorical features.

The remainder of the paper is organised as follows: Sect. 2 provides a summary of related work. Section 3 introduces the CLEAR *Image* method and algorithms. Section 4 details the experimental setup and discusses the results. Section 5 concludes the paper and indicates directions for future work.

2 Related work

This paper adopts the following notation: Let m be a machine learning system mapping each input instance x to a class label l with probability y. Each input instance x is an image that can be partitioned into S segments (regions) $\{s_1, \dots, s_n\}$. We use x' to denote a GAN-generated image derived from x such that $m(x') = l$ with probability y'.

The XAI methods most relevant to this paper can be broadly grouped into four types:

(i) **Counterfactual methods** Wachter et al. (2017) first proposed using counterfactuals as explanations of single machine learning predictions. Many XAI methods have attempted to generate ‘optimal’ counterfactuals; for example, Karimi et al. (2020) review sixty counterfactual methods. The algorithms differ in their constraints and the attributes referenced in their loss functions (Verma et al., 2020). Desiderata often include that a counterfactual is: (1) actionable – e.g. actions do not get recommended if they are physically infeasible, such as reducing a person’s age, (2) near to the original observation - common measures include Manhattan distance, L1 norm and L2 norm, (3) sparse – only changing the values of a small number of features, (4) plausible - e.g. the counterfactual must correspond to a high-density part of the training data, (5) efficient to compute. Karimi et al. (2021) argue that these methods are likely to identify counterfactuals that are either suboptimal or infeasible in terms of their actionability. This is because they do not take into account the causal structure that determines the consequences of the person’s actions. The underlying problem is that unless all of the person’s features are causally independent of each other, then when the person acts to change the value of one feature, other downstream dependents may also change. In Sect. 5 we will explain why this criticism does not apply to CLEAR *Image*. In this paper, we provide a different criticism of counterfactual methods: that they fail to provide satisfactory explanations because they do not provide a causal equation describing the local behaviour of the classifier they are meant to explain. Without this, they cannot identify: the relative importance of different features, how the features are taken to interact with each other, or the functional forms that the classifier is, in effect, applying to each feature. They will also fail to identify cases of overdetermination.

(ii) **Gradient-based methods** These provide saliency maps by backpropagating an error signal from a neural network’s output to either the input image or an intermediate layer. Simonyan et al. (2014) use the derivative of a class score for the image to assign an importance score to each pixel. Kumar et al. (2017)'s CLass-Enhanced Attention Response uses backpropagation to visualise the most dominant classes; this should not be confused with our method. A second approach modifies the back-propagation algorithm to produce sharper saliency maps, e.g. by suppressing the negative flow of gradients. Prominent examples of this approach (Springenberg et al., 2014; Zeiler and Fergus, 2014) have been found to be invariant to network re-parameterisation or the class predicted (Adebayo et al., 2018; Nie et al., 2018). A third approach (Selvaraju et al., 2017; Chattopadhay et al., 2018) uses the product of gradients and activations starting from a late layer. In Grad-CAM (Selvaraju et al., 2017), the product is clamped to only highlight positive influences on class scores.

(iii) **Perturbation based methods** Methods such as Occlusion (Zhou et al., 2016), Extremal Perturbation (Fong et al., 2019), FIDO (Chang et al., 2018b), LIME (Ribeiro et al., 2016) and Kernel SHAP (Lundberg and Lee, 2017) use perturbation to evaluate which segments of an image x are most responsible for x's classification probability y. The underlying idea is that the contribution that a segment s_i makes to y can be determined by substituting it with an uninformative segment s_i', where s_i' may be either grey, black or blurred (Zhou et al., 2016; Fong et al., 2019; Ribeiro et al., 2016) or in-painted without regard to any contrast class (Chang et al., 2018b). LIME and Kernel SHAP generate a dataset of perturbed images, which feeds into a regression model, which then calculates segment importance scores (LIME) or Shapley Values (Kernel SHAP). Extremal Perturbation uses gradient descent to determine an optimal perturbed version of an image that, for a fixed area, has the maximal effect on a network's output whilst guaranteeing that the selected segments are smooth. FIDO uses a variational Bernoulli drop to find a minimal set of segments that would change an image's class. In contrast to LIME, Kernel SHAP and Extremal Perturbation, FIDO uses a GAN to in-paint segments with 'plausible alternative values'; however, these values are not generated to belong to a chosen contrast class. Furthermore, segment importance scores are not produced.

There are three key problems with using perturbed images to explain a classification:

1. A satisfactory explanation must be contrastive; it must answer *'Why E rather than F?'* None of the above methods does this. Their contrasts are instead images of uninformative segments.
2. The substitution may fail to identify the contribution that s_i makes to y. For example, replacing s_i with black pixels can take the entire image beyond the classifier's training distribution. By contrast, blurring or uninformative in-painting might result in s_i' being too similar to s_i resulting in the contribution of s_i being underestimated.
3. A segmentation needs to be relevant to its explanatory question. Current XAI perturbation approaches produce radically different segmentations. FIDO and Extremal Perturbation identify 'optimal' segments that, when substituted by an uninformative segment, maximally affect the classification probability; by contrast, LIME uses a texture/intensity/colour algorithm (e.g. Quickshift (Vedaldi and Soatto, 2008)).

(iv) **Contrastive methods using GAN image synthesis** Generative adversarial network (GAN) (Goodfellow et al., 2014) has been widely applied for synthetic image generation. Image-to-image translation GANs enable a conditional transformation of an input image to a specified target. For example, CycleGAN (Zhu et al., 2017) and

StarGAN (Choi et al., 2018) translate images between different domain classes. StarGAN-V2 (Choi et al., 2020) improved the conditional image translation by incorporating a style vector instead resulting in a more scaleable and high-quality synthetic image generation across a variety of target conditions. Fixed-point GAN penalised any deviation of image for intra-domain translation with identity loss. DeScarGAN (Wolleb et al., 2020) incorporates this loss function in its own GAN architecture and has outperformed Fixed-point GAN in its case study for identifying and localising pathology from chest X-rays. The availability of synthetic images can alleviate the constraint of data scarcity typically found in specialised domains (e.g. medical imaging). Singh and Raza (2021), Osuala et al. (2022) has presented GAN's applicability in the medical domain.

While the adversarial training needed by GAN is known to be challenging for (i) maintaining training stability, (ii) reaching convergence and (iii) avoiding mode collapse (Arora et al., 2022; Mescheder et al., 2018; Salimans et al., 2016; Osuala et al., 2022), many examples of properly trained GAN have been achieved (Osuala et al., 2022). Kazeminia et al. (2020) provided numerous examples of employing GAN in medical image analysis. Chang et al. (2018a) introduced the fill-in the dropout region (FIDO) methods, wherein generative methods were applied for in-filling. This method however requires the generative model to recreate the missing regions based on the remaining unmasked features. Shih et al. (2020) emphasised the improvement in contrastive comparison using a GAN-generated contrast image in over earlier work using a uniform-value reference or a blurred input image. They modified the training of StarGAN model (Choi et al., 2018) and demonstrated that their GAN-generated images allowed more appropriate identification of attributing features and minimise errors that can be induced from other non GAN-generated alternatives. In situations where data is scarce, it is anticipated that the benefits of GAN-based synthetic image generation would outweigh the time and effort required to attain proper training in a GAN.

In Sect. 5 we will explain how CLEAR *Image* builds on the strengths of the above XAI methods but also addresses key shortcomings.

3 The CLEAR *Image* method

CLEAR *Image* is a model-agnostic XAI method that explains the classification of an image made by any classifier (see Fig. 1). It requires both an image x and a contrast image x' generated by a GAN. CLEAR *Image* segments x into $\{s_1, \ldots, s_n\} \in S$ and then applies the same segmentation to x' creating $\{s'_1, \ldots, s'_n\} \in S'$. It then determines the contributions that different subsets of S make to y by substituting with the corresponding segments of S'. CLEAR *Image* is GAN agnostic, allowing the user to choose the GAN architecture most suitable to their project. A set of 'image-counterfactuals' $\{c_1 \ldots c_k\}$ is also identified. Figures 1, 2, 3, 4 and 5 provide a running example of the CLEAR *Image* pipeline, using the same X-ray taken from the CheXpert dataset.

3.1 GAN-based image generation

To generate contrastive images, StarGAN-V2 (Choi et al., 2020) and DeScarGAN (Wolleb et al., 2020) have been found to be capable of generating the high-quality

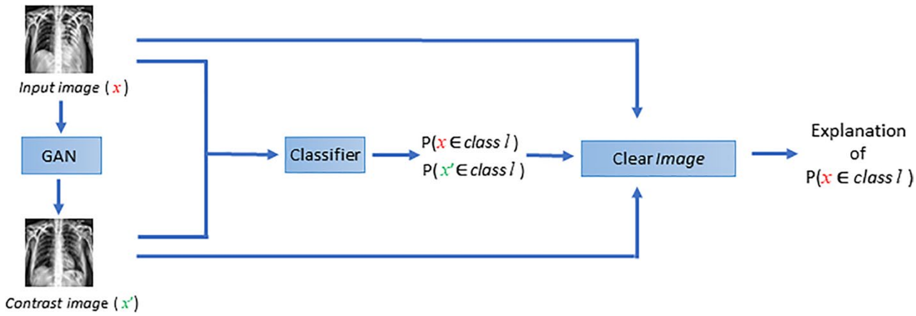


Fig. 1 The CLEAR *Image* pipeline. The GAN produces a contrast image. CLEAR *Image* explains the classification probability by comparing the input image with its contrast image. It produces a regression equation that measures segment scores, reports fidelity and identifies cases of overdetermination. In this example, class l is 'pleural effusion' and its contrast class l' is 'healthy'. Using our Densenet model, the X-ray shown in this figure had a probability of belonging to l equal to 1, and its contrast image had a probability of belonging to l equal to 0

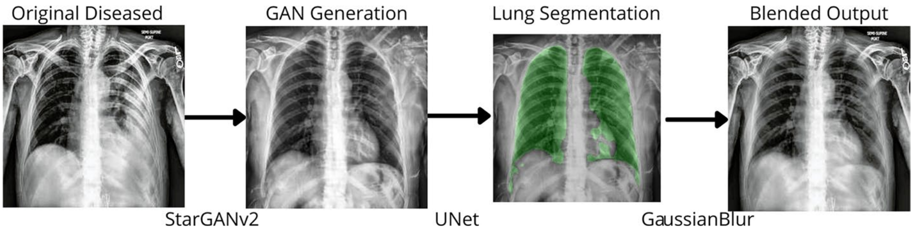


Fig. 2 The process of generating a contrast image. An original diseased image is first used to generate a healthy contrast image with a trained GAN model. In this example, StarGAN v2 is used as the architecture. The generated healthy lung airspace is then segmented using a U-Net segmentation model blended onto the original diseased image to produce the final image by applying Gaussian blur to minimise any edging effect around the segments

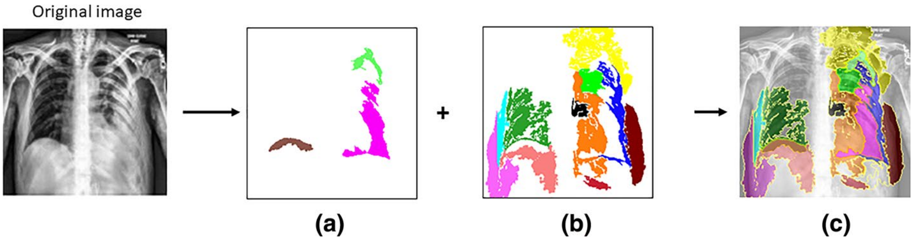


Fig. 3 The GAN-Augmented segmentation algorithm. There are three stages. First, segments are identified from the high-intensity differences between the original image x and its contrast image x' (**a**). Second, additional segments are identified from the regions of x corresponding to low-intensity differences between x and x' (**b**) Third, the segments from the two steps are combined (**c**)

images needed to identify the segments of pixel differences. These GANS are therefore deployed as the network architectures for our two case studies, the first using CheXpert, and the second using a synthetic dataset respectively. These established GAN networks

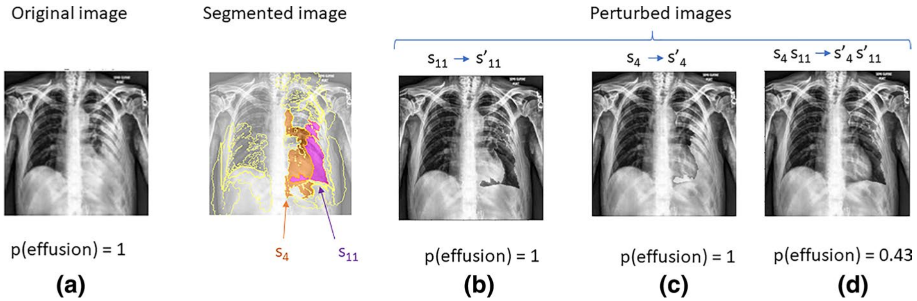
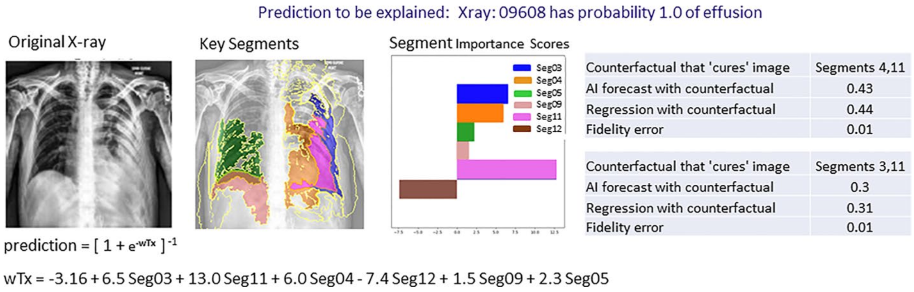


Fig. 4 Determining image-counterfactuals. In this example segments, s_4 and s_{11} are evaluated both separately and in combination. Substituting s_{11} with its corresponding contrast segment s'_{11} creates a perturbed image (**b**) with the same classification probability as the original image (**a**). The same applies with segment s_4 (**c**). However, substituting both segments s_4 and s_{11} results in a perturbed image (**d**) which has a classification probability of 0.43. Given a decision boundary at the probability of 0.5, **d** would be classified as a 'healthy' X-ray and would therefore be an image-counterfactual


Prediction to be explained: Xray: 09608 has probability 1.0 of effusion

Counterfactual that 'cures' image	Segments 4,11
AI forecast with counterfactual	0.43
Regression with counterfactual	0.44
Fidelity error	0.01

Counterfactual that 'cures' image	Segments 3,11
AI forecast with counterfactual	0.3
Regression with counterfactual	0.31
Fidelity error	0.01

prediction = $[\ 1 + e^{-wTx}\ ]^{-1}$

wTx = -3.16 + 6.5 Seg03 + 13.0 Seg11 + 6.0 Seg04 - 7.4 Seg12 + 1.5 Seg09 + 2.3 Seg05

Fig. 5 Extracts from a CLEAR *Image* report. The report identifies that substituting both segments 4 and 11 with the corresponding segments from its contrast image flips the classification probability to 'healthy' According to the logistic regression equation these substitutions would change the probability of the X-ray being classified as 'pleural effusion' to 0.44. However, when these segments are actually substituted and passed through the classifier, the probability changes to 0.43, hence the fidelity error is 0.01. CLEAR *Image* also identifies that substituting segments 3 and 11 also creates an image-counterfactual. Note that unlike methods such as GradCAM, CLEAR *Image* is able to identify segments that have a negative impact on a classification probability

demonstrate how the generated contrastive images can aid the overall CLEAR *Image* pipeline in our cases where contrast images are not available. Default training hyperparameters were applied unless otherwise stated. Details of model training and hyperparameters can be found in Appendix B. The source image was used as input for the Style Encoder instead of a specific reference image for StarGAN-V2. This ensures the generated style mimics that of the input source images. StarGAN-V2 is also not locally constrained (i.e. the network will modify all pixels in an image related to the targeted class, which will include irrelevant spurious regions of the image). A post-generation lung space segmentation step using a pre-trained U-Net model (Ronneberger et al., 2015) was therefore implemented. The original diseased lung space was replaced with the generated image, with a Gaussian Blur process to fuse the edge effect (see Fig. 2). This confines the feature identification space used by CLEAR *Image* to the lung space. It is an advantage of the CLEAR *Image* pipeline that it is possible to use pre-processing to focus the explanation on the relevant parts of x.

3.2 Generating contrastive counterfactual explanations

Definition 1 An **image-counterfactual** c_j from l to l' is an image resulting from a change in the values of one or more segments S of x to their corresponding values in S' such that class$(m(x)) = l$, class$(m(c_j)) = l'$ and $l \neq l'$. The change is minimal in that if any of the changed segments had remained at its original value, then class$(m(x))$ = class$(m(c_j))$.

CLEAR *Image* uses a regression equation to quantify the contribution that the individual segments make to y. It then measures the fidelity of its regression by comparing the classification probability resulting from each c_j with an estimate obtained from the regression equation.

Definition 2 **Counterfactual-regression fidelity error** Let $reg(c_j)$ denote the application of the CLEAR *Image* regression equation given image-counterfactual c_j.

$$\text{Counterfactual-regression fidelity error} = |reg(c_j) - y_{c_j}|.$$

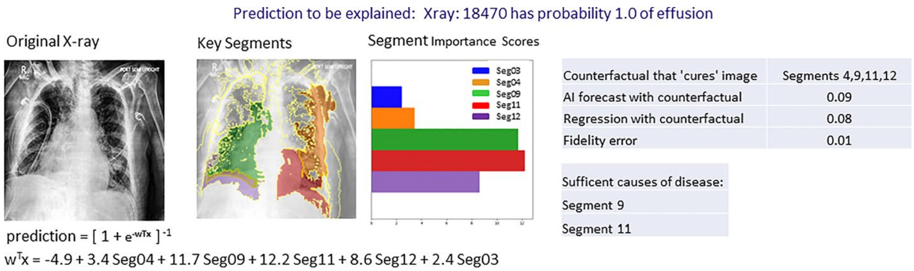
The following steps generate an explanation of prediction y for image x:

1. GAN-Augmented segmentation algorithm. This algorithm is based on our findings (in Section 5.4) that the segments (S_h) determined by analysing high-intensity differences between an image x and its corresponding GAN-generated image x' will often miss regions of x that are important to explaining x's classification. It is, therefore, necessary to supplement segments S_h with a second set of segments S_l confined to those regions of x corresponding to low-intensity differences between x and x'. S_l is created based on similar textures/intensities/colours solely within x.

   The pseudocode for our algorithm is shown in Algorithm 1. First, high and low thresholds (T_h and T_l) are determined by comparing the differences between x and x' using multi-Otsu; alternatively, the thresholds can be user-specified. T_h is then used to generate a set of segments, S_h. The supplementary segments S_l, are determined by applying the low threshold, T_l, to the low-intensity regions and then applying a sequence of connected component labelling, erosion and Felzenszwalb (Felzenszwalb and Huttenlocher, 2004). The combined set of segments, S_h and S_l, is checked to see if any individual segment is an image-counterfactual. If none is found, an iterative process is applied to gradually increase the minimum segment size parameter. The final set of segments (S, S') is subsequently created using the combined set (S_h, S_l) as shown in Fig. 3.
2. Determine x's image-counterfactuals. A dataset of perturbed images is created by selectively replacing segments of x with the corresponding segments of x' (see Fig. 4). A separate image is created for every combination in which either 1, 2, 3, or 4 segments are replaced. Each perturbed image is then passed through m to determine its classification probability. All image-counterfactuals involving changes in up to four segments are then identified. (The maximum number of perturbed segments in a counterfactual

is a user parameter; the decision to set it to 4 in our experiments was made as we found counterfactuals involving 5+ segments to have little additional explanatory value.)

3. Perform a stepwise logistic regression. A tabular dataset is created by using a {0,1} representation of the segments in each perturbed image from step 2. Consider a perturbed image x_{per}. This will be composed of a combination of segments s_i from the original image x and segments s'_i from the GAN contrast image x'. In order to represent x_{per} in tabular form, each segment of x_{per} that is from x is represented as a 1 and each segment of x_{per} that is from x' is represented as a 0. For example, if x_{per} consisted solely of $\{s'_1, s_2, s_3, s_4\}$, and had a classification probability from m equal to 0.75 of being 'pleural effusion', then this would be represented in tabular form as $\{0, 1, 1, 1, 0.75\}$. The table of representation vectors is the input to a weighted logistic regression in which those perturbed images that are image-counterfactuals are given a high weighting and act as soft constraints. The {0,1} representation of the segments are the independent variables and the classification probability is the dependent variable. Figures (5 and 6) provide examples of the resulting logistic equation and the calculation of classification probability.
4. Calculate segment importance scores. These are the regression coefficients for each segment from step 3.
5. Identify cases of causal overdetermination (see below).
6. Measure the fidelity of the regression by calculating fidelity errors (see Fig. 5) and goodness of fit statistics.
7. Iterate to the best explanation. In XAI there is often a trade-off between the interpretability of an explanation and its fidelity. For example, a regression equation that has two independent variables and no interaction terms is likely to be easier to interpret than a regression equation with more independent variables and several interaction terms. Because of its increased complexity, the latter regression equation might better mimic the local input–output behaviour of the AI system to be explained (i.e. it will have greater fidelity). CLEAR *Image* allows the user to adjust parameters such as (i) whether to include interaction terms (ii) the maximum number of independent variables in a regression. It then reports the fidelity of the resulting explanation. In this way, the user can iterate to the explanation that they judge provides the best trade-off between interpretability and fidelity.


Prediction to be explained: Xray: 18470 has probability 1.0 of effusion

Counterfactual that 'cures' image	Segments 4,9,11,12
AI forecast with counterfactual	0.09
Regression with counterfactual	0.08
Fidelity error	0.01

Sufficent causes of disease:
Segment 9
Segment 11

prediction = $[\,1 + e^{-w^T x}\,]^{-1}$

$w^T x = -4.9 + 3.4\,\mathrm{Seg04} + 11.7\,\mathrm{Seg09} + 12.2\,\mathrm{Seg11} + 8.6\,\mathrm{Seg12} + 2.4\,\mathrm{Seg03}$

Fig. 6 Overdetermination. The report identifies segments 9 and 11 as each sufficient to have caused the original X-ray to be classified as 'pleural effusion' with a probability greater than 0.99. Hence this is a case of causal overdetermination. The corresponding GAN-generated image x' has a classification probability ≈ 0 for pleural effusion. If a perturbed image x_{per} was created by substituting all the segments of the original image x with the corresponding segments of x' *except* for segment 9, then x_{per} would still have a classification probability for pleural effusion greater than 0.99. The same would apply if only segment 11 was substituted

```
Algorithm 1: GAN_Augmented_Segmentation
input                       : $x$ - diseased image, $x'$ - contrast image, $m$ - AI classifier
Segmentation Parameters:    $min\_num\_S_l$ - min number of $S_l$ segments,
                            $min\_seg\_size$ - min segment size,
                            $seg\_size\_increment$ - segment size increment

$T_h, T_l \leftarrow$ Determine_Thresholds($x, x'$)
$D_h, D_l \leftarrow$ Create_Difference_masks($x, x', T_h, T_l$)
$S_h \leftarrow$ Create_high_intensity_segments($D_h, min\_seg\_size$)
// Connected components, erosion and Felzenszwalb are now
   used to create the low-intensity segments
$S_l \leftarrow$ Create_low_intensity_segments($D_l, x, min\_seg\_size$)
$N_l \leftarrow$ $Count\_S_l\_Segments(S_l)$
$N_c \leftarrow$ $Count\_Single\_Segment\_Counterfactuals(m, S_h, S_l)$
// If there are no counterfactuals then increase the size
   of the $S_l$ segments
while $N_c = 0$ and $min\_num\_S_l < N_l$ do
    $min\_seg\_size$ += $seg\_size\_increment$
    $S_l \leftarrow$ Create_low_intensity_segments($D_l, x, min\_seg\_size$)
    $N_l \leftarrow$ $Count\_S_l\_Segments(S_l)$
    $N_c \leftarrow$ $Count\_Single\_Segment\_Counterfactuals(m, S_h, S_l)$
$S, S' \leftarrow$ Add_Segments($S_h, S_l, x, x'$)
return $S, S'$
```

For CLEAR *Image* an explanation is a tuple $< G;C;r;O,e >$, where G are segment importance scores, C are image-counterfactuals, r is a regression equation, O are the causes resulting in overdetermination, and e are fidelity errors. The regression equation is a causal equation with each independent variable (each referring to whether a particular segment is from x or x') being a *direct cause* of the classification probability. Figure 5 shows an extract from a CLEAR report. Pseudocode summarising how CLEAR *Image* generates an explanations is provided in Algorithm 2.

```
Algorithm 2: CLEAR Image
input  : $x$ - input image,
         $x'$ - contrast image,
         $m$ - AI classifier.
$S, S' \leftarrow$ GAN_Augmented_Segmentation($x, x', m$)
$D \leftarrow$ Create_Perturbed_Data($S, S', m$)
$C \leftarrow$ Find_Counterfactuals($S, S', m$)
$r \leftarrow$Find_Regression_Equation($D, C$)
$G \leftarrow$Extract_Segment_Scores($r$)
$O \leftarrow$Find_Overdetermination($r$)
$e \leftarrow$ Calculate_Fidelity($C, r$)
return explanation=$< G, C, r, O, e >$
```

The causal overdetermination of an effect occurs when multiple sufficient causes of that effect occur. By default, CLEAR *Image* only reports sufficient causes which each consist of a single segment belonging to S. Substituting a sufficient cause for its corresponding member in S' guarantees the effect. In the philosophy of science, it is generally taken that for an effect to be classified as overdetermined, it should be narrowly defined, such that all the sufficient causes have the same, or very nearly the same impact (Paul, 2009). Hence for the case studies, the effect is defined as $p(x \in \mathit{diseased}) > 0.99$, though the user may choose a different probability threshold. A sufficient cause changes a GAN-generated healthy image to a diseased image. This is in the opposite direction to CLEAR *Image*'s counterfactuals whose perturbed segments flip the classification to 'healthy'. Sufficient causes can be read off from CLEAR *Image*'s regression equation. Using the example in Fig. 6 with the logistic formula, a classification probability of > 0.99 requires $\boldsymbol{w}^T\boldsymbol{x} > 4.6$. The GAN healthy image corresponds to all the binary segment variables being equal to 0. Hence, $\boldsymbol{w}^T\boldsymbol{x}$ is equal to the intercept value of −4.9, giving a probability of $(1 + exp^{4.9})^{-1} \approx 0.01$. If a segment s'_i is now replaced by s_i, the corresponding binary variable changes to 1. Hence if segment 9 is infilled, then $\mathit{Seg09} = 1$ and $\boldsymbol{w}^T\boldsymbol{x} = 6.8$ ($i.e.$11.7 − 4.9). Similarly, infilling just segment 11 will make $\boldsymbol{w}^T\boldsymbol{x} > 4.6$. Either substitution is sufficient to guarantee $\boldsymbol{w}^T\boldsymbol{x} > 4.6$, irrespective of any other changes that could be made to the values of the other segment variables. Hence segments 9 and 11 are each a sufficient cause leading to overdetermination.

By contrast, XAI methods such as LIME and Kernel SHAP cannot identify cases of overdetermination. This is because they use simple linear regression instead of logistic regression. For example, suppose that an image has three segments: s_1, s_2, s_3. In the regression dataset, each segment infilled from x has a value of 1 and each segment infilled from x' has a value of 0. LIME/Kernel SHAP's regression equation will have the form: $y = k_1 s_1 + k_2 s_2 + k_3 s_3$. In the case of LIME, y is meant to be the classification probability and the regression coefficients (k_1, k_2, k_3) are the feature importance scores. Let us suppose there is overdetermination, with segments s_1 and s_2 each being a sufficient cause for x to be in a given class (e.g. 'pleural effusion') with more than 0.99 probability. Hence, the regression equation should set y to a value greater than 0.99 not only when $s_1 = s_2 = 1$, but also when either $s_1 = 1$ or $s_2 = 1$. This is clearly impossible with the above linear form (and the constraint that $y \leq 1$). Mutatis mutanda, the same argument applies for Kernel SHAP.

4 Experimental investigation

There are two case studies, the first using a synthetic dataset, the second analysing pleural effusion X-rays taken from the CheXpert dataset (Irvin et al., 2019). Transfer learning was used to train both a VGG-16 with batch normalisation and a DenseNet-121 classifier for each dataset. CLEAR *Image* was evaluated against Grad-CAM, Extremal Perturbations and LIME. The evaluation consisted of both a qualitative comparison of saliency maps and a comparison of pointing game and intersection over union (IoU) scores. CLEAR *Image*'s fidelity errors were also analysed (none of the other XAI methods measures fidelity).

4.1 Datasets

The synthetic dataset's images share some key characteristics found in medical imaging including: (i) different combinations of features leading to the same classification and (ii) irrelevant features. All images (healthy and diseased) contain a set of concentric circles, a large and a small ellipse. An image is 'diseased' if either: (1) the small ellipse is thin-lined, and the large ellipse contains a square or (2) there is a triangle, and the large ellipse contains a square. The dataset is an adaptation of Wolleb et al. (2020).

CheXpert is a dataset of chest X-rays with automated pathological label extraction through radiology reports, consisting of 224,316 radiographs of 65,240 patients in total. Images were extracted just for the classes 'pleural effusion' and 'no finding'. Mis-classified images and images significantly obstructed by supporting devices were manually filtered. A random frontal X-ray image per patient was collected. In total, a dataset of 2,440 images was used in this work for model training, validation and testing. Appendix A.2 details the data preparation process. A hospital doctor provided the ground truth annotation to the X-ray images with pleural effusion for our case study.

4.2 Evaluation metrics

This paper uses two complementary metrics to evaluate XAI methods. Both require annotated images identifying 'target' regions that should be critical to their classification. A pointing game produces the first metric, which measures how successfully a saliency map 'hits' an image's targets. Previously pointing games have been designed for cases where (i) images have single targets (ii) the saliency maps have a maximum intensity point (Fong et al., 2019; Zhang et al., 2018). By contrast, this paper's case studies have multiple targets, and the pixels within each CLEAR *Image* segment have the same value. We, therefore, formulated a novel pointing game. The pointing game partitions a 'diseased' image into 49 square segments, $\mathrm{P} = \{p_1 \ldots p_{49}\}$ and identifies which squares contain each of the targets. The corresponding saliency map is also partitioned, and each square is allocated a score equal to the average intensity of that square's pixels $\mathrm{Q} = \{q_1 \ldots q_{49}\}$. The pointing game then starts with the q_i of highest intensity and determines if the corresponding p_i contains a relevant feature. A successful match is a 'hit' and an unsuccessful match is a 'miss'. This process continues until every target has at least one hit. The score for an image is the number of hits over the number of hits plus misses. Pseudocode is provided in Algorithm 3.

The second metric is IoU. It is assumed that each pixel in a saliency map is classified as 'salient' if it is above 70^{th} percent of the maximum intensities in that map. IoU then measures the overlap between the 'salient' pixels $pix^{salient}$ and the pixels belonging to the image's targets pix^{target}: $IOU = (pix^{salient} \cap pix^{target})/(pix^{salient} \cup pix^{target})$. The chosen percentile was an empirically identified threshold to maintain a relatively high IoU score by balancing high intersection with pix^{target} and small union of pixel regions with a large enough $pix^{salient}$ (see Appendix A.1 for details).

Both metrics are useful but have counterexamples. For example, IoU would give too high a score to a saliency map that strongly overlapped with a large image target but completely missed several smaller targets that were also important to a classification. However, applied together, the two metrics provide a good indication of an XAI's performance.

Algorithm 3: Pointing Game

```
input : $x$ - input image, $A$ - annotated features
        $w$ - XAI saliency map
$P_A \leftarrow$ Square_Idx_of_Each_Feature($A, x$)
$Q \leftarrow$ Average_Intensity_Each_Square($w$)
$Q' \leftarrow$ Square_Idx_Sort_Highest_Intensity($Q$)
$hits \leftarrow 0$; $misses \leftarrow 0; h_A \leftarrow False$
foreach $q'_i \in Q'$ do
  // starting with highest
    foreach $a_j \in A$ do
        if $q'_i \in P_{a_j}$ then // square idx match
            $hits \leftarrow hits + 1$ ; $h_{a_j} \leftarrow True$
        else
            $misses \leftarrow misses + 1$
    if $\forall a_j$ $(h_{a_j} = True)$ then
      // exit once all features hit
        break
return $< hits, misses >$
```

4.3 Experimental runs

CLEAR *Image* was run using logistic regression with the Akaike information criterion; full testing and parameter values can be found in Appendix B.3. The test datasets consisted of 95 annotated X-rays and 100 synthetic images. The average running time for CLEAR *Image* was 20 s per image for the synthetic dataset and 38 s per image for the CheXpert dataset, running on a Windows i7-8700 RTX 2070 PC. Default parameter values were used for the other XAI methods, except for the following beneficial changes: Extremal Perturbations was run with 'fade to black' perturbation type, and using areas {0.025,0.05,0.1,0.2} with the masks summed and a Gaussian filter applied. LIME was run using Quickshift segmentation with kernel sizes 4 and 20 for the CheXpert and synthetic datasets respectively.

4.4 Experimental results

CLEAR *Image* outperforms the other XAI methods on both datasets (Fig. 7a). Furthermore, its fidelity errors are low, indicating that the regression coefficients are accurate for the counterfactually important segments (Fig. 7b). Figure 7c illustrates some of the benefits of using the 'Best Configuration', which uses GAN-augmented segmentation and infills using x'. This is compared with (i) segmenting with Felzenszwalb and infilling with x' (ii) segmenting with GAN-augmented but infilling with black patches (iii) segmenting with Felzenszwalb, infilling with black patches. Figure 8 illustrates how CLEAR *Image*'s use of GAN-augmented leads to a better explanation than just using a difference mask (e.g. CLEAR *Image*'s performance was similar for VGG-16 and DenseNet; therefore, only the DenseNet results are presented unless otherwise stated.

CLEAR *Image*'s regression equation was able to capture the relatively complex causal structure that generated the synthetic dataset. Figure 9 shows an example. A square (SQ) is a *necessary but insufficient* cause for being diseased. An image is labelled as diseased if

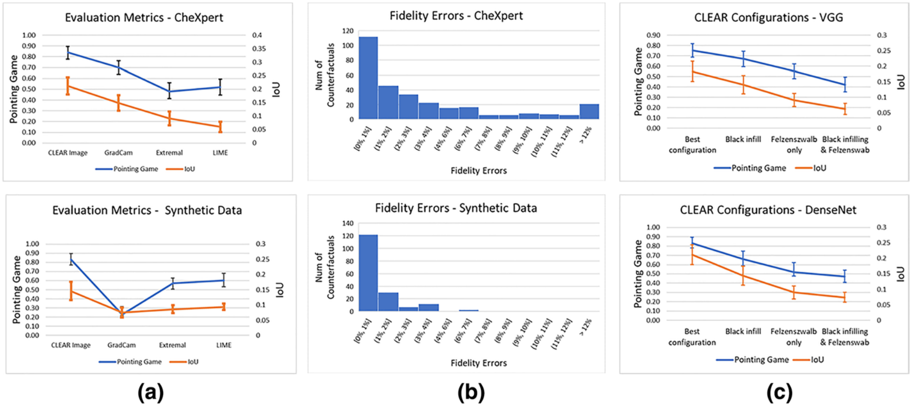


Fig. 7 Evaluation metrics. **a** Compares the performances of different XAI methods with the DenseNet models. **b** Shows the fidelity errors for the DenseNet models. **c** Compares the performances of different configurations of CLEAR *Image*. The bars show 95% confidence intervals

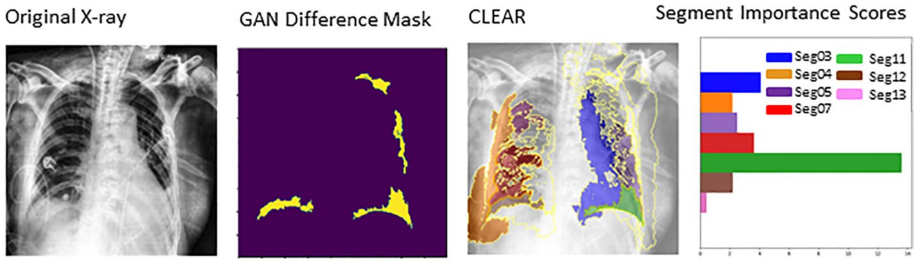
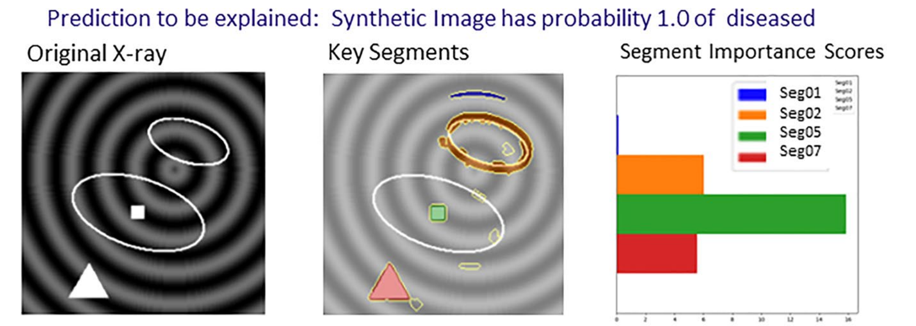


Fig. 8 GAN-Augmented Segmentation versus GAN difference mask. The difference mask identifies four segments but when CLEAR *Image* perturbs these, the two nearest to the top were found to be irrelevant. Of the other two segments, CLEAR *Image* identifies the segment it colors green to be far more important to the classification probability



prediction = $[\,1 + e^{-w^Tx}\,]^{-1}$ $w^Tx = -16.8 + 16.0\,\mathrm{Seg05} + 6.0\,\mathrm{Seg02} + 5.6\,\mathrm{Seg\ 07} + 0.1\,\mathrm{Seg01}$

Fig. 9 Extracts from a CLEAR *Image* report for a synthetic image. The regression equation shows that Seg05 is a *necessary but insufficient cause* of the X-ray being diseased

there is also either a triangle (TR) or the small ellipse is thin-lined (TE). When SQ, TR and TE are all present in a single image, there is a type of overdetermination in which TR and TE are each a sufficient cause *relative* to the 'image with SQ already present'. As before, a diseased image corresponds to the binary segment variables equalling one and a classification probability of being diseased > 0.99 requires $\boldsymbol{w}^T\boldsymbol{x} > 4.6$. This can only be achieved by Seg 5 (corresponding to SQ) plus at least one of Seg 2 or Seg 7 (TE, TR) being set to 1 (i.e. being present). Figure 10 compares the saliency maps for synthetic data.

For the CheXpert dataset, Fig. 11 illustrates how CLEAR *Image* allows for a greater appreciation of the pathology compared to 'broad-brush' methods such as Grad-CAM (please see Appendix A1 for further saliency maps). Nevertheless, the IoU scores highlight that the segmentation can be further improved. For CheXpert's counterfactuals, only 5% of images did not have a counterfactual with four or fewer s' segments. Most images required several s segments to be infilled before its classification flipped to 'healthy', 17% required one segment, 30% with two segments, 24% with three segments and 24% with four segments. 17% of the X-rays' were found to be causally overdetermined.

5 Discussion, conclusion and future work

With AI systems for images being increasingly adopted in society, understanding their implicit causal structures becomes paramount. Yet, the explanations provided by XAI methods cannot always be trusted, as the differences in the saliency maps of Fig. 11 exemplify. It is therefore important that XAI methods should measure their fidelity. By 'knowing when it does not know', it can alert the user when its explanations are unfaithful.

CLEAR *Image* recognises that a difference mask is only the starting point for an explanation. In the experiments reported in this paper, CLEAR *Image* uses a GAN-generated image both for infilling and as input to its own segmentation algorithm. As discussed

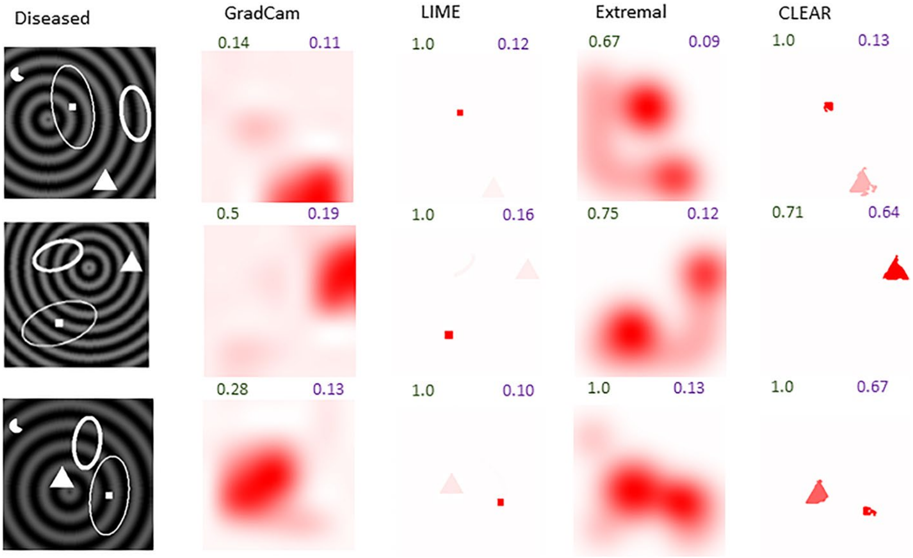


Fig. 10 Comparison of XAI methods on synthetic data. The pointing game scores are shown in green and the IoU scores are in purple. The maps illustrate how CLEAR *Image* and LIME are able to tightly focus on salient regions of an image compared to broadbrush methods such as Grad-CAM and Extremal. The significance of a patch is indicated by its red intensity

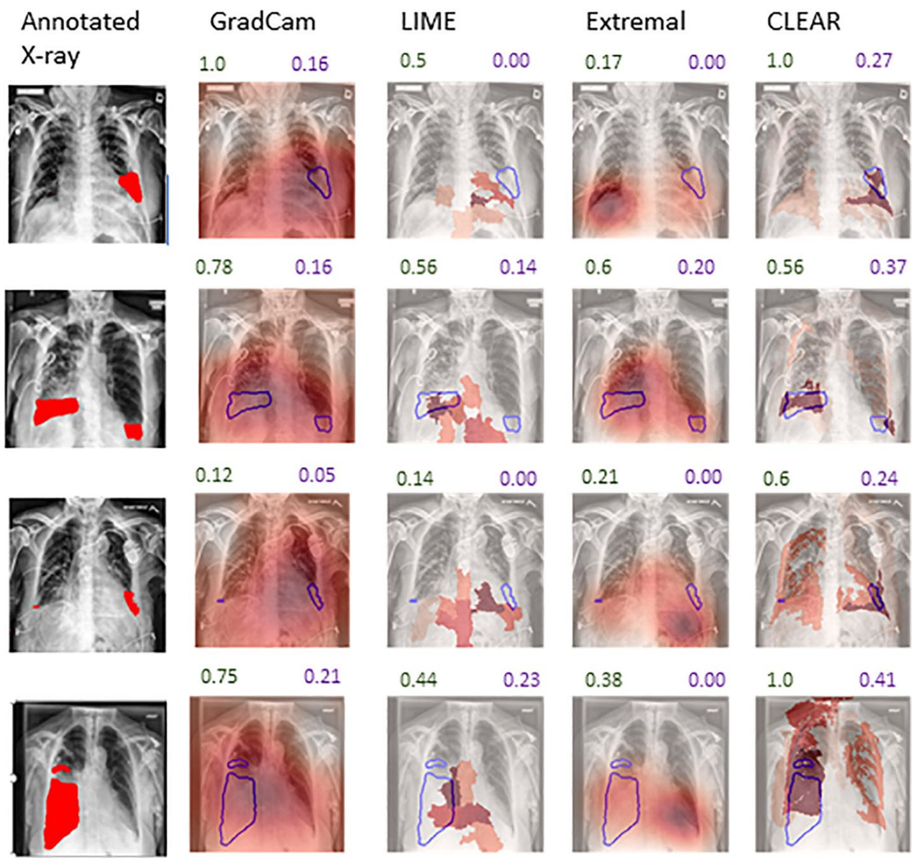


Fig. 11 Comparison of XAI methods on X-ray. The pointing game scores are shown in green and the IoU scores are in purple. The significance of a patch is indicated by the intensity of red against the blue-outlined annotated ground truth

below, other approaches are possible when the segmentation can be defined in advance with the use of prior knowledge as in the case of brain scans. This is under investigation.

We have shown that CLEAR *Image* can illuminate cases of causal overdetermination. Many other types of causal structures may also be ubiquitous in AI. For example, causal preemption and causal clustering are well documented within the philosophy of science (Baumgartner, 2009; Schaffer, 2004). The relevance of these to XAI creates an area of future work.

The examples in this paper help illustrate our claim that XAI counterfactual methods will often fail to provide satisfactory explanations of a classifier's local input–output behaviour. This is because a satisfactory explanation requires both counterfactuals and a supporting causal equation. It is only because CLEAR *Image* produces a causal equation that it is able to identify (a) segment importance scores, including identifying segments with negative scores (Fig. 5), (b) segments that are necessary but insufficient causes (Fig. 9), (c) cases of overdetermination (Fig. 6). Providing only counterfactuals is insufficient; imagine another science, say physics, treating a statement of counterfactuals as being an explanation, rather than seeking to discover the governing equation. Perhaps the primary benefit of XAI counterfactual methods is in suggesting sets of actions. But as we noted in Sect. 2 and argued in Karimi et al. (2021), such methods may identify counterfactuals that are suboptimal or infeasible in terms of their actionability. This criticism does not apply to CLEAR *Image* because CLEAR *Image*'s purpose is to explain the local input–output behaviour of a classifier, and the role of its counterfactuals is (i) to illustrate the classifier's causal structure (at the level of how much each segment can cause the classification probability to change) and (ii) answer contrastive questions. Hence, if the explanatory question is "why is this image classified as showing a zebra and not a horse?", CLEAR *Image* might highlight the stripes on the animal as being a cause of the classification. Whilst this might be a satisfactory explanation of the classification, it is, of course, not actionable.

Methods such as LIME and Kernel SHAP bear some similarity to CLEAR *Image* as they also use a dataset of perturbed images to feed a regression. However, these methods do not use a GAN-generated image and do not report fidelity. Also, these methods assume that a classification probability is a simple linear addition of its causes. This is incorrect for cases of causal overdetermination and CLEAR *Image*, therefore, uses a sigmoid function.

A key limitation for CLEAR *Image* is its reliance on using a contrast image, for both infilling and for guiding segmentation. The contrast image needs to be aligned with the target image so that the perturbed images are correctly infilled. In this paper's experiments, CLEAR *Image* uses a GAN-generated contrast image. But there is a data availability constraint for custom training of a GAN especially in specialized domains. Training stability and convergence, as well as mode collapse are also common concerns during GAN training. Nevertheless, it may still be possible to obtain contrast images through other means. For example in human neuroimaging, AI systems are often trained using registered and normalised MRI scans (Pölsterl et al., 2021). In such cases, a contrast image can be simply selected from images belonging to the required contrast class. In cases where a contrast image cannot be obtained then CLEAR can use the same infilling (black/blurred) and external segmentation methods used by LIME. CLEAR *Image* will then be expected to have similar fidelity as LIME but, critically, unlike LIME it will report its fidelity, so the user will know if the explanation corresponds to the underlying model.

Another possible limitation could be the understandability of CLEAR *Image* to non-technical users. A user study should now be carried out. These are time and resource consuming and need to be devised carefully by experts within specific application domains to produce sound and reliable results. Instead, we have focused on objective measures and evaluations of fidelity which in our view should precede any user study. Future work will also include adapting CLEAR *Image* to the multimodal neural networks now being used in human neuroimaging, where contrast images can be readily obtained without using a GAN. There are brain atlases for these registered images (e.g. https://atlas.brainnetome.org) which provide neurologically meaningful segments. Another area of work will be to extend our analysis of overdetermination to other types of causal structures.

Appendix A: Supplemental results for CheXpert dataset and associated data pre-processing

Supplementary qualitative results

Additional qualitative results for the CheXpert dataset are presented in this section (Figs. 12 and 13) where the most important segments (regions) identified by each XAI method is matched against the annotated ground truth. These are the pixels of saliency

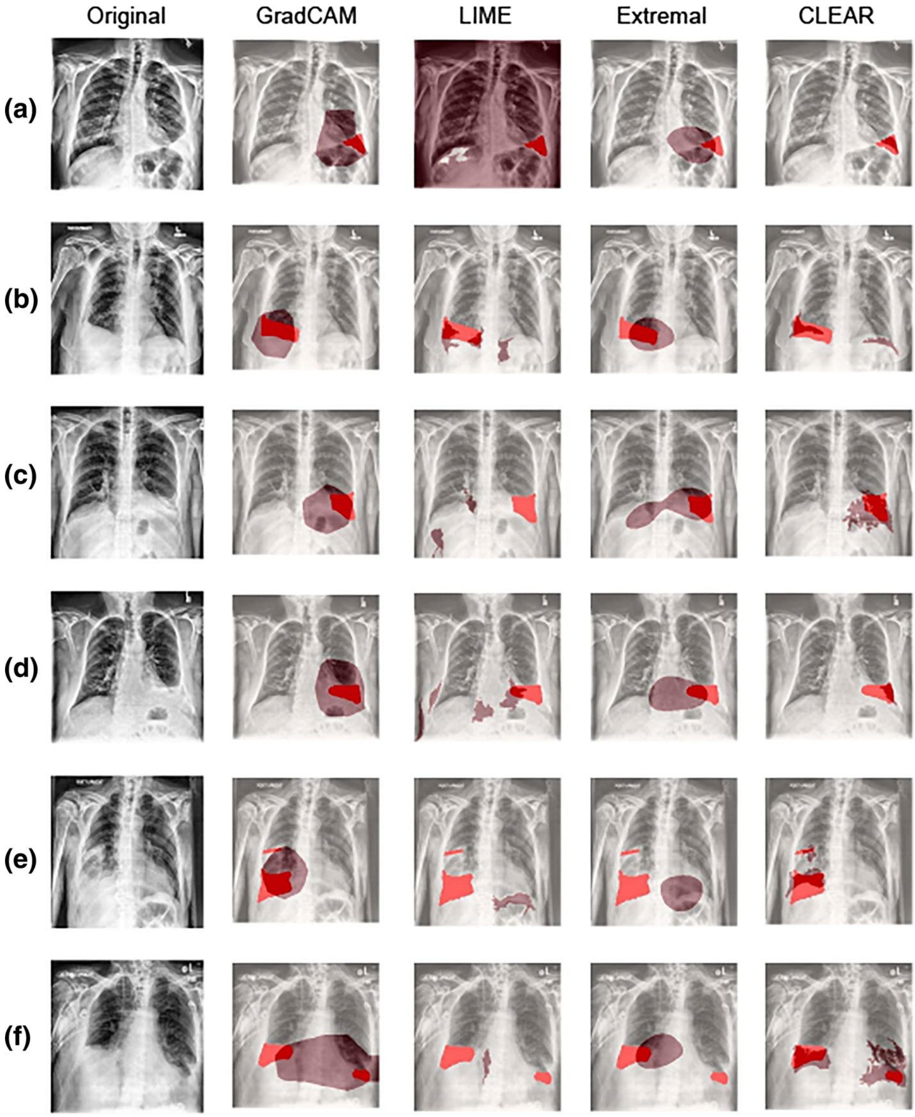


Fig. 12 Representative comparative examples of the identified important segments of a DenseNet-based image classification model (Val Acc: 98.8%) for pleural effusion using (i) CLEAR *Image*, (ii) Grad-CAM, (iii) Extremal Perturbation and (iv) LIME

maps that are above 70 percent of the maximum intensity (i.e. the segments used to calculate the IoU scores). This threshold was determined empirically to yield high IoU score across all the XAI methods evaluated (see Fig. 14).

Figure 12 shows the additional results for the DenseNet model while Fig. 13 presents the results for the VGG16 model. These results have demonstrated higher precision using CLEAR Image in identifying significant segment matching against the annotated ground truth in comparison to other explanation methods. These two figures provide a qualitative

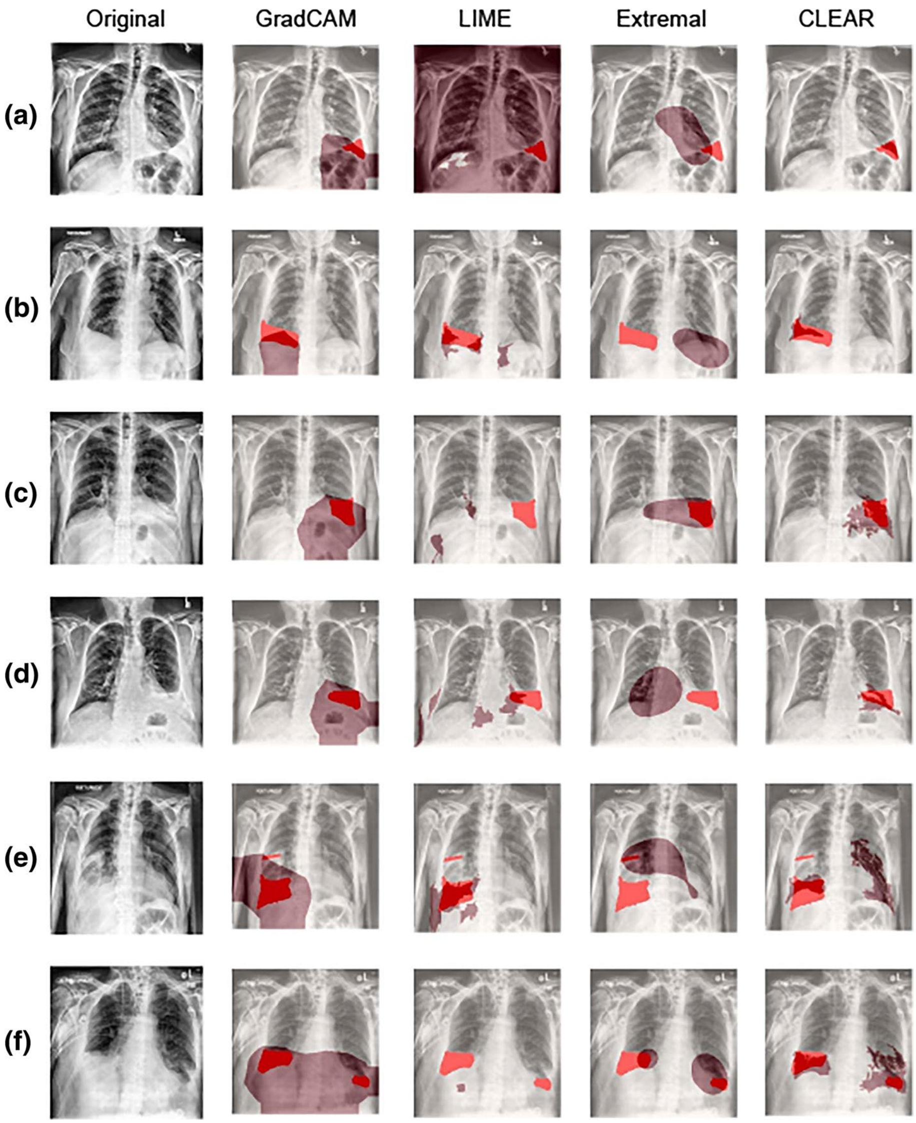
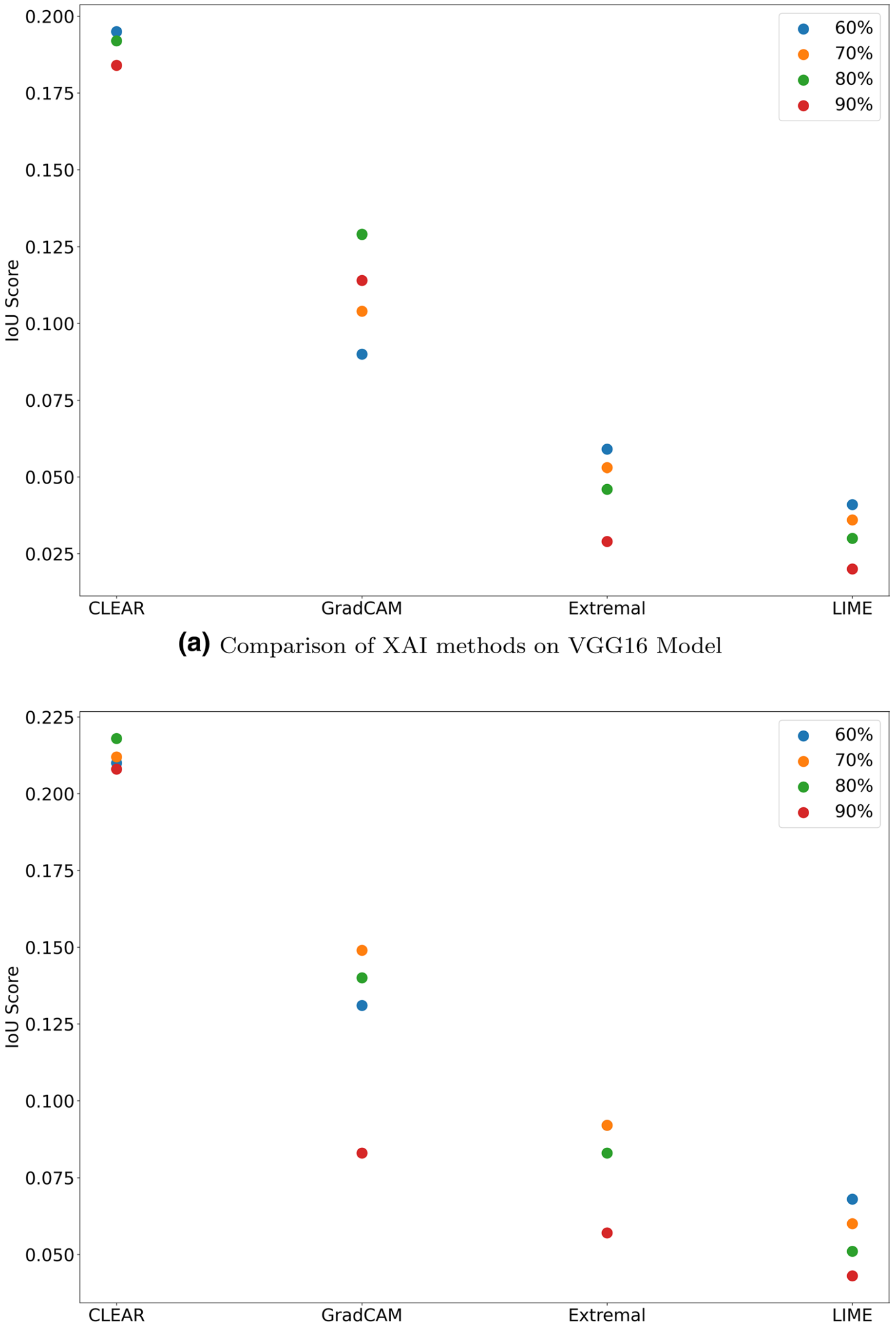


Fig. 13 Representative comparative examples of the identified important segments of a VGG16-based image classification model (Val Acc: 97.5%) for pleural effusion using (i) CLEAR *Image*, (ii) Grad-CAM, (iii) Extremal Perturbation and (iv) LIME

comparison to supplement the results presented in Fig. 7 where CLEAR *Image* outperforms other XAI methods.

Data pre-processing

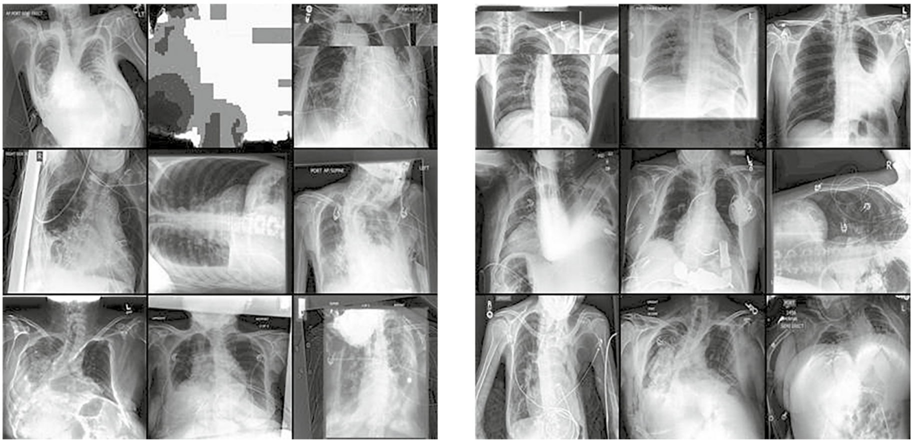
CheXpert has a total of 14 pathological classes including 'No Finding', and these are labelled through an automated rule-based labeller from text radiology reports. For each



(a) Comparison of XAI methods on VGG16 Model

(b) Comparison of XAI methods on DenseNet Model

Fig. 14 Comparison of IoU score against four XAI methods, (1) CLEAR *Image*, (2) GradCAM, (3) Extremal and (4) LIME to determine the threshold of intensity at 10% intervals. CLEAR *Image* outperforms the other XAI methods for each of the 4 intensity thresholds



(a) Diseased **(b)** Healthy

Fig. 15 Representative examples of poorly curated images including image distortion, mis-orientation, obstruction by limbs and support devices as well as significant spine deformation

observation, the Stanford team has classified each radiograph as either negative (0), uncertain (-1) or positive (1). Other metadata includes gender, age, X-ray image projection and presence of supporting devices.

In this study, this dataset (v1.0) was applied for the model development of a binary classification task to demonstrate the capability of CLEAR *Image* as an XAI framework. An initial filtering process of the metadata was applied for the two classes used in the study - (1) Diseased with Pleural Effusion and (2) Healthy (this was assumed to be X-ray images with no findings and no positive observations in any of the pathological conditions). To minimise potential complications with other pathological conditions, X-ray images with only positive in pleural effusion were used with the other pathological categories either as negative/blank.

A review of the filtered images also identified that the dataset was curated with some images having significant artefacts that can hamper model training performance. Figure 15 presents some of these images in both diseased and healthy categories. Many of these consisted of artefacts from image capturing and processing (e.g. image distortion, orientation, low resolutions or miscalibration). Some images were also significantly obstructed by limbs or support devices. Some healthy images were also wrongly labelled according to a hospital doctor, who assisted in our project. A secondary manual filtering was conducted to remove any identified images with artefacts.

The 2440 selected images were split approximately 80/10/10 for the training/validation/testing. The images were also resized to 256 x 256 as the input into the classification model and generative adversarial network (GAN) as described in Sect. 5. Figure 16 presents some typical images in the final dataset for both diseased and healthy categories.

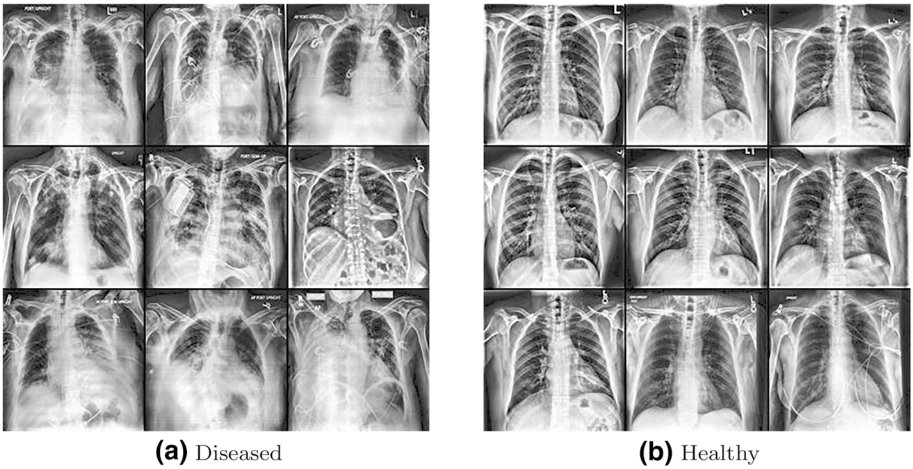


Fig. 16 Representative examples of final images for **a** diseased with identifiable regions of pathology and **b** healthy images with clear air space. All images have minimal obstructions from support devices

Appendix B: Model parameters

DeScarGAN and parameters

The DeScarGAN architecture was adopted for the synthetic dataset in Section 5.1. 80% of the dataset (4000 images) was used for GAN training and 20% of the dataset (1000 images) was used for validation. A total of 2,500 epochs was run and the best epoch was selected on visual quality. Additional 100 images were generated as an out-of-sample test dataset. Adam optimizer was used with $\beta_1 = 0.5$, β_2=0.999. An initial learning rate of 10^{-4} was used and stepped down to a final learning rate of 10^{-6}. Default hyperparameters for loss functions were used to mimic a similar investigation from the original author as shown below in Table 1:

Table 1 Default loss function hyperparameters used in DeScarGAN

Loss term	Weight value
Adversarial loss	$\lambda_{adv,g} = 1$ (generator)
	$\lambda_{adv,d} = 20$ (discriminator)
Gradient penalty loss	$\lambda_{gp} = 10$
Identity loss	$\lambda_{id} = 50$
Reconstruction loss	$\lambda_{rec} = 50$
Classification loss	$\lambda_{cls,g} = 1$ (generator)
	$\lambda_{cls,d} = 5$ (Discriminator)

StarGAN-V2 and parameters

StarGAN-V2 (Choi et al., 2020) has been adopted in this work as a state-of-art GAN network for image translation. The GAN provided the necessary contrastive images for the CheXpert dataset. Default hyperparameters were maintained while notable loss weights are highlighted in Table 2. Adam optimizer was used with $\beta_1 = 0$, β_2=0.99. A total of 50,000 epochs were run for the CheXpert dataset. The style encoding was referenced to the input image for the translation to the targeted class. This aided in maintaining the general features of the images compared to the original. As StarGAN-V2 (Choi et al., 2020) did not constrain its generation to a localised region (e.g. lungs), post-processing of segmentation and blending was implemented for the CheXpert dataset. Segmentation of the lung region was based on a pre-trained model with a U-Net architecture. The segmentation mask was subsequently used to guide the replacement of pixels within the lung region from the GAN generated healthy image onto the original diseased image. Gaussian Blur was applied to minimise the edge effect during the blending process. This post-processing step aided in restricting the feature identification space within the lungs and reducing the computational cost for locating the counterfactuals.

An evaluation of similarity to real healthy images was performed using the Fréchet inception distance (FID) (Heusel et al., 2017) benchmarking against the set of healthy images in the model training dataset. Four image sets were compared: (1) real healthy images in the validation set, set of images with pleural effusion processed as described in Fig. 2 with replacement of lung segments using (2) corresponding GAN-generated healthy images, (3) Gaussian blurred version of the original images and (4) constant value of zero (i.e. black). This FID score indicated how close each of the four compared image sets to the benchmark images in the training set. A low score indicated similarity between the two datasets.

As observed in Fig. 17, the processed images with replacement using GAN generated healthy lung segments resemble more similar to actual healthy images than blurred or black segments. As such, GAN generated processed images as described in Fig. 2 were selected as the choice of synthetic healthy images for this work.

Table 2 Default Loss Function Hyperparameters used in StarGANv2 (Choi et al., 2020)

Loss term	Weight value
Style reconstruction loss	$\lambda_{sty} = 1$
Style diversification loss	$\lambda_{ds} = 1$
Cyclic loss	$\lambda_{cyc} = 1$

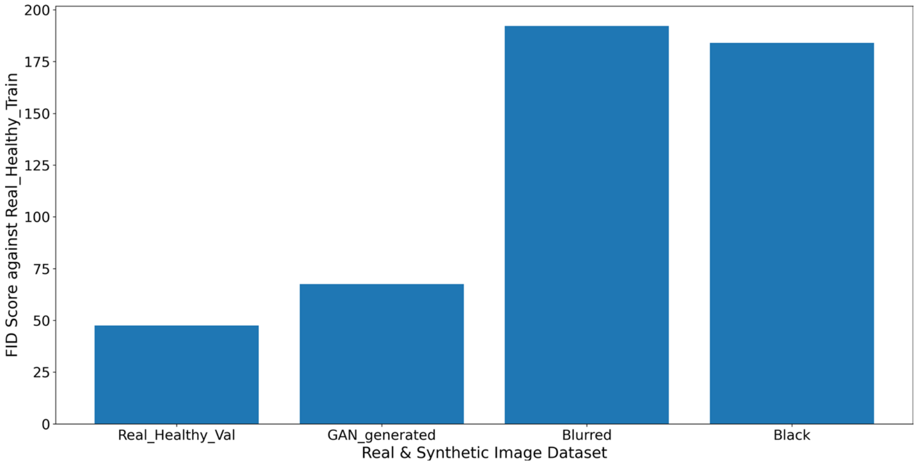


Fig. 17 Comparison of Fréchet inception distance (FID) against the training healthy image dataset with (1) a set of real healthy images in the validation set, set of images with pleural effusion processed as described in Fig. 2 with replacement of lung segments using (2) corresponding GAN-generated healthy images, (3) Gaussian blurred version of the original images and (4) constant value of zero (i.e. black)

Clear *Image* parameters

The default parameters used for the Chest X-ray experiments were:

```
case_study =′ Medical′
max_predictors = 6
num_samples = 1000
regression_type =′ logistic′
logistic_regularise = False
score_type =′ aic′
apply_counterfactual_weights = True
counterfactual_weight = 200
binary_decision_boundary = 0.5
no_polynomimals_no_interactions = True
interactions_only = True
no_intercept = False
centering = True
include_features = False
include_features_list = []
sufficiency_threshold = 0.99
image_infill =′ GAN′
image_all_segments = False
threshold_method =′ manual′
image_use_old_synthetic = False
image_counterfactual_interactions = False
image_segment_type =′ Augmented_GAN′
max_segments_in_counterfactual = 4
min_segs_created_for_Augmented_GAN = 4
min_seg_size = 250
min_seg_increment = 25
image_classes = [′normal′,′ effusion′]
```

For the CLEAR *Image* configuration experiments the parameter ‘image_infill’ had values [‘GAN’, ‘black’] and the parameter image_segment_type had values [‘Augmented_GAN’, ‘Felzenszwalb’]

The same parameter values were used for the synthetic case study except:

```
case_study =′ Synthetic′
image_segment_type =′ Thresholding′
```

Author contributions All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Adam White and Kwun Ho Ngan, with medical advice from James Phelan. The first draft of the manuscript was written by Adam White and Kwun Ho Ngan and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding This research received no external funding

Data availability The code for generating the synthetic dataset is available (see below) and the CheXpert dataset is available at https://stanfordmlgroup.github.io/competitions/chexpert/

Code availability The code (including data generation) is available at https://github.com/ClearExplanationsAI/CLEAR_Image.

Declarations

Conflict of interest The authors declare no competing interest.

References

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). *Sanity checks for saliency maps*. arXiv preprint arXiv:1810.03292.

Arora, S., Risteski, A., & Zhang, Y. (2022). *Do GANs learn the distribution? Some theory and empirics*.

Baumgartner, M. (2009). Inferring causal complexity. *Sociological Methods & Research, 38*(1), 71–101.

Cao, J., Ma, C., Yao, T., Chen, S., Ding, S., & Yang, X. (2022). End-to-end reconstruction-classification learning for face forgery detection. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*. IEEE.

Chang, C.-H., Creager, E., Goldenberg, A., & Duvenaud, D. (2018a). *Explaining image classifiers by counterfactual generation*.

Chang, C.-H., Creager, E., Goldenberg, A., & Duvenaud, D. (2018b). *Explaining image classifiers by counterfactual generation*. arXiv preprint arXiv:1807.08024.

Chattopadhay, A., Sarkar, A., Howlader, P., & Balasubramanian, V. N. (2018). Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)* (pp. 839–847). IEEE.

Choi, Y., Choi, M., Kim, M., Ha, J.-W., Kim, S., & Choo, J. (2018). Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 8789–8797).

Choi, Y., Uh, Y., Yoo, J., & Ha, J.-W. (2020). Stargan v2: Diverse image synthesis for multiple domains. In *2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, (pp. 8185–8194). https://doi.org/10.1109/CVPR42600.2020.00821

Felzenszwalb, P. F., & Huttenlocher, D. P. (2004). Efficient graph-based image segmentation. *International Journal of Computer Vision, 59*(2), 167–181.

Fong, R., Patrick, M., & Vedaldi, A. (2019). Understanding deep networks via extremal perturbations and smooth masks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 2950–2958).

Garfinkel, A. (1982). *Forms of explanation: Rethinking the questions in social theory*.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems* (pp. 2672–2680).

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. In I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information*

Processing Systems. (Vol. 30). Curran Associates Inc. https://proceedings.neurips.cc/paper/2017/file/8a1d694707eb0fefe65871369074926d-Paper.pdf.

Hilton, D. J. (1990). Conversational processes and causal explanation. *Psychological Bulletin, 107*(1), 65.

Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., Seekins, J., Mong, D. A., Halabi, S. S., Sandberg, J. K., Jones, R., Larson, D. B., Langlotz, C. P., Patel, B. N., Lungren, M. P., & Ng, A. Y. (2019). *Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison*.

Karimi, A.-H., Barthe, G., Schölkopf, B., & Valera, I. (2020). *A survey of algorithmic recourse: definitions, formulations, solutions, and prospects*. arXiv preprint arXiv:2010.04050.

Karimi, A.-H., Schölkopf, B., & Valera, I. (2021). Algorithmic recourse: from counterfactual explanations to interventions. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 353–362).

Kazeminia, S., Baur, C., Kuijper, A., van Ginneken, B., Navab, N., Albarqouni, S., & Mukhopadhyay, A. (2020). GANs for medical image analysis. *Artificial Intelligence in Medicine, 109*, 101938.

Kumar, D., Wong, A., & Taylor, G. W. (2017). Explaining the unexplained: A class-enhanced attentive response (clear) approach to understanding deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 36–44).

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* (pp. 4765–4774).

Mescheder, L., Geiger, A., & Nowozin, S. (2018). *Which training methods for GANs do actually converge?* (pp. 3481–3490).

Miller, T. (2018). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence, 267*, 1.

Nie, W., Zhang, Y., & Patel, A. (2018). A theoretical explanation for perplexing behaviors of backpropagation-based visualizations. In *International conference on machine learning* (pp. 3809–3818). PMLR.

Osuala, R., Skorupko, G., Lazrak, N., Garrucho, L., García, E., Joshi, S., Jouide, S., Rutherford, M., Prior, F., Kushibar, K., Diaz, O., & Lekadir, K. (2022). *medigan: A python library of pretrained generative models for enriched data access in medical imaging*.

Paul, L. A. (2009). Counterfactual theories. In *The Oxford handbook of causation*.

Pearl, J. (2000). *Causality: Models, reasoning and inference* (1st ed.). New York, NY, USA: Cambridge University Press.

Pölsterl, S., Wolf, T. N., & Wachinger, C. (2021). Combining 3d image and tabular data via the dynamic affine feature map transform. In *International conference on medical image computing and computer-assisted intervention* (pp. 688–698). Springer.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should i trust you? explaining the predictions of any classifier. In *Proceedings of ACM SIGKDD 2016, KDD '16* (pp. 1135–1144). ACM. ISBN 978-1-4503-4232-2. https://doi.org/10.1145/2939672.2939778.

Ronneberger, O., Fischer, P., & Brox, T. (2015). *U-net: Convolutional networks for biomedical image segmentation*.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., & Chen, X. (2016). *Improved techniques for training GANs*.

Schaffer, J. (2003). Overdetermining causes. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition, 114*(1/2), 23–45.

Schaffer, J. (2004). Trumping preemption. *The Journal of Philosophy, 97*(4), 165–181.

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. In *Proceedings of the IEEE international conference on computer vision* (pp. 618–626).

Shih, S.-M., Tien, P.-J., & Karnin, Z. (2020). GANMEX: One-vs-One attributions guided by GAN-based counterfactual explanation baselines.

Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). *Deep inside convolutional networks: Visualising image classification models and saliency maps*.

Singh, N. K., & Raza, K. (2021). Medical image generation using generative adversarial networks: A review. In R. Patgiri, A. Biswas, & P. Roy (Eds.), *Health informatics: A computational perspective in healthcare* (pp. 77–96). Springer.

Springenberg, J. T., Dosovitskiy, A., Brox, T., & Riedmiller, M. (2014). *Striving for simplicity: The all convolutional net*. arXiv preprint arXiv:1412.6806.

Van Fraassen, B. C., et al. (1980). *The scientific image*. Oxford University Press.

Vedaldi, A., & Soatto, S. (2008). Quick shift and kernel methods for mode seeking. In *European conference on computer vision* (pp. 705–718). Springer.

Verma, S., Dickerson, J., & Hines, K. (2020). Counterfactual explanations for machine learning: A review. arXiv preprint arXiv:2010.10596.

Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech., 31*, 841.

White, A., & Garcez, A. A. (2020). Measurable counterfactual local explanations for any classifier. In *ECAI 2020* (pp. 2529–2535). IOS Press.

Wolleb, J., Sandkühler, R., & Cattin, P. C. (2020). Descargan: Disease-specific anomaly detection with weak supervision. In *International conference on medical image computing and computer-assisted intervention* (pp. 14–24). Springer.

Woodward, J. (2003). *Making things happen: A theory of causal explanation*. Oxford University Press. ISBN 9780195189537.

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In *European conference on computer vision* (pp. 818–833). Springer.

Zhang, J., Bargal, S. A., Lin, Z., Brandt, J., Shen, X., & Sclaroff, S. (2018). Top-down neural attention by excitation backprop. *International Journal of Computer Vision, 126*(10), 1084–1102.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2016). Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2921–2929).

Zhu, J.-Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision* (pp. 2223–2232).