Universiteit Utrecht
CENTAI

Minor Research Project
MSc in Bioinformatics and Biocomplexity

Fairness and Explainability in Chest X-ray Image Classifiers

Examiner:
Dr. Anastasia Giachanou

Host institute supervisors:
Prof. André Pansisson
Dr. Alan Perotti
Dr. Claudio Borile
Dr. Michele Starnini

Student:
Gemma Bel Bordes (1988018)

July 2023

Abstract

Artificial intelligence (AI) is increasingly being used in healthcare, particularly for interpreting medical images. However, there are growing concerns regarding the presence of biases in these AI models, which raise important fairness considerations. This study investigates biases in artificial intelligence (AI) models for chest X-ray diagnosis and explores the role of Explainable AI (XAI) in understanding model decisions. Biases were observed in model performance across different patient groups and diseases. Various XAI techniques were employed to generate explanations for model decisions, and comparisons were made with explanations provided by doctors. We identified an optimized version of occlusion as the most accurate XAI technique in this case, which also provided a consistent accuracy of the explanations across all patient groups. Indeed, the explanations remained equally accurate regardless of variations in model performance for different subgroups, suggesting the absence of model bias amplification. Evaluating the correctness of XAI explanations posed challenges due to the limited availability of ground truth. In order to increase the power of our analysis, we explored alternative evaluation methods, like deletion or insertion curves, but reported them as unsuitable for chest X-ray images. We have therefore established some recommendations for using XAI on chest X-ray images. Given the reported absence of biases in the explanations, our aim is also to instill confidence in clinical stakeholders regarding XAI techniques.

Keywords: fairness, explainability, chest X-rays, artificial intelligence.

Layman's Summary

Artificial Intelligence (AI) is a powerful technology that mimics human learning and reasoning to solve problems. In healthcare, AI can be used to create models that diagnose diseases based on chest X-ray images, similar to how a doctor would.

However, researchers have found that these AI models can sometimes be biased, just like humans. For example, a model may perform better for males than females, leading to unfair outcomes. To build trust in these models, it is crucial for doctors to understand how they make decisions.

This is where Explainable AI (XAI) comes in. XAI aims to make AI models more understandable by providing explanations for their decisions. In the case of chest X-ray diagnosis, an XAI technique would highlight the important regions in the image that influenced the model's decision. For instance, if the model determines a heart-related disease, the explanation should point out the heart as the crucial area.

In our study, we confirmed the presence of biases in chest X-ray models. We noticed that the performance of the model varied depending on the disease and the group of patients (e.g., males vs. females, younger vs. older). To understand why the model was making these decisions, we used XAI techniques to generate explanations.

However, we observed that different XAI techniques provided different explanations. So, we compared the explanations with those given by doctors and selected the technique that aligned better with the doctors' explanations. With this analysis we were able to find the best explainer for this case. Interestingly, we found that the XAI technique's explanations were equally accurate for all groups of patients. This suggests that even if the model performs differently for different patient's groups, the explanations are not affected by this. The explanations will be equally accurate for all groups. However, we need to conduct further research with more cases to confirm these findings.

Furthermore, we encountered challenges in evaluating the correctness of the XAI explanations. It is difficult to gather enough information about how doctors explain their decisions, making alternative evaluation methods necessary. We tried to evaluate explanations without the doctor's explanations. But unfortunately, the evaluation methods we tried were not suitable for our chest X-ray images.

In summary, addressing biases and enhancing the interpretability of AI models through XAI techniques are vital steps in ensuring fairness and trust in healthcare applications. Further research is needed to validate the findings, but we suggest that XAI methods do not amplify model biases, which should motivate clinicians to trust XAI. Developing suitable evaluation methods will contribute to advancing the field of XAI in chest X-ray diagnosis. We would like to ultimately use these techniques to evaluate why the model is sometimes wrong.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Motivation and contribution

Healthcare is being revolutionized with technological advances and, particularly, with Artificial Intelligence (AI). Research on AI applied to medicine is rapidly expanding, with the number of related publications currently five times greater than it was a decade ago [1]. The use of AI for the interpretation of medical imaging has been of great success [2], and its implementation seems to be particularly promising in the radiology field, as X-ray images follow a universal standard that facilitates the integration with AI [3]. The urgent need for rapid interpretation of chest X-ray images during the COVID-19 pandemic has further accelerated research in this area [4, 5].

Despite the notable advancements, the adoption of these tools in hospitals across many high-income countries remains sluggish [6]. Several studies have highlighted two significant concerns voiced by clinical stakeholders: a lack of transparency, which refers to the challenge of understanding the decision-making process of AI models, and the potential biases embedded within these models [7, 8].

First of all, biases in AI pose a genuine concern that has been relatively recently identified in the medical and other fields [9, 10]. Since then, this issue has been carefully studied [11–14]. These biases can arise either from the training data that fails to represent the demographics of the target population [10] or from inherent biases in the learning process itself [15]. When a model exhibits bias, it demonstrates unequal performance across different subpopulations, often characterized by protected attributes such as sex, age, or race. In other words, biased models are inherently unfair, and in the context of medicine, this poses significant risks and violates bioethical principles [16]. Indeed, these fairness issues have also been reported for models classifying chest X-ray images [17–19], as well as other imaging modalities like MRI [20, 21].

On the other hand, models designed for image interpretation are often complex and regarded as black boxes. And understanding the output of a black box is challenging. Explainable AI (XAI) tackles this problem of transparency, by aiming to elucidate the behavior of such models using various techniques [22]. The form of the explanation will depend on the goal and input of the model. In the case of image classifiers, built with Convolutional Neural Networks (CNN), we expect a visual explanation in the form of a heatmap. This heatmap highlights those features of the input image that were important for the final classification. These are also known as saliency maps or attribution heatmaps. XAI on medical imaging is extensively reviewed in [23, 24].

This project combines both concerns, fairness and explainability in an attempt to audit models. Specifically, we concentrate on classifiers that categorize chest X-ray images for specific conditions or diseases, known as multi-label classifiers. Ideally, we want (1) to once again establish the presence of biases within these classifiers, and (2) to employ XAI techniques to comprehend the reasons behind the frequent misclassification of data points from a particular subpopulation. It is important to note that we attribute biases to the model itself rather than the data. For instance, we may observe the model excessively focusing on the female breast as a factor for predicting a disease that is actually unrelated

to it.

Nevertheless, prior to delving into the understanding of misclassifications using XAI, it is crucial to establish the trustworthiness of the explanations provided. We aim (1) to benchmark a selection of XAI techniques (i.e., explainers) with different metrics, and (2) to assess whether the explanations are equally accurate for all subpopulations. By adding this explainability step, we might be generating new sources of biases or amplifying those embedded in the model.

1.2 Related work

The integration of fairness and explainability to comprehend biases was inspired by the work of FairLens [12]. It begins by identifying the subpopulations for which the model demonstrates poorer performance. Subsequently, it leverages XAI techniques to provide explanations for misclassified instances within those subpopulations. Notably, the model evaluated in FairLens was a multi-label classifier utilizing tabular data. To the best of our knowledge, there is currently no literature available that applies a similar approach specifically to image datasets.

Several datasets with chest X-ray images have been released. The NIH ChestX-ray [25] was initially released with the annotation for 8 chest diseases, but they later added more labels adding up to 14 diseases (NIH ChestX-ray 14). With the release of this dataset, the first efforts were made to employ state-of-the-art algorithms for classifying these images [25, 26]. A few years later, other groups published similar but larger datasets, like CheXpert [27], MIMIC-CXR [28] and PadChest [29], which further stimulated research on new classifiers.

Researchers quickly raised concerns regarding fairness issues in chest X-ray classifiers. It was discovered that gender-imbalanced datasets led to biased classifiers [18]. As such, classifiers trained with a higher proportion of male images, showed a lower performance when tested with female images, and vice versa. Furthermore, Seyyed et al. reported algorithmic underdiagnosis biases affecting traditionally underserved subpopulations, such as black female patients with low income [17]. Discussion about the sources of the biases (i.e., whether they were data-based or algorithmic biases) followed this publication [30, 31].

Regarding algorithmic biases, some proposed techniques for model debiasing involved fine-tuning and pruning methods [32]. However, another study found that aiming for equal performance across the entire population actually resulted in worsened performance for all subpopulations [33]. They concluded that their debiasing strategies did not surpass the effectiveness of simple data balancing techniques.

In a different approach, Luo et al. focused on addressing shortcut learning as a potential source of biases and introduced a novel algorithm to mitigate it [34]. Shortcut learning occurs when a model learns spurious correlations that are irrelevant to the classification task and do not generalize to other datasets. This behavior was observed in COVID image classifiers, where the learned shortcuts were lateral markers indicating some information about the image acquisition process [35]. However, shortcuts can also

manifest as anatomical features that differentiate patient groups based on demographics rather than diseases, such as the female breast. Notably, models have been trained to predict patient demographics (i.e., age, sex, and race) from chest X-ray images [36]. In a recent study, the same trained model used for disease classification achieved remarkable success in classifying sex and race [37]. But based on other analysis, they argued that this could not be solely used to establish the source of the model biased to be anatomical shortcuts that differentiate patients demographically.

In terms of explainability, there is no consensus about the XAI technique to be used with chest X-ray image classifiers [23, 38]. It is common to find publications that use explainers as an additional step (post-hoc) for assessing the model (e.g., [39]). But also, some researchers propose modifications to the CNN (like [40, 41]) to make it more interpretable. Among the post-hoc explainers, those generating Class Attention Maps (CAM) are prevalent [38]. Recent benchmarking studies have compared different explainers used with chest X-ray image classifiers, with GradCAM being reported as superior but still limited compared to human annotations of disease localization [42]. Another novel explainer, PYLON, has also undergone benchmarking with other explainers by using human annotations as the ground truth [43]. However, there is no standardized metric for comparing these explainers, although intersection over union (IOU) is commonly used to evaluate the overlap. Efforts have also been made to evaluate explanations without relying on ground truth annotations, and fidelity metrics have been proposed. These are calculated by perturbing the image regions regarded as important for the classification, and quantifying the change in model's output [44, 45]. However, the debate around these fidelity metrics is ongoing [46–48].

1.3 Theoretical background

1.3.1 DenseNet-121 model

Seyyed et. al. reported biased chest X-ray classifiers [17] using the DenseNet-121 architecture [49]. DenseNet-121 is a densely connected Convolutional Neural Network (CNN) designed to address the problem of information loss that occurs when the number of layers is too high. Particularly, DenseNet-121 is built in such a way that all convolutional layers (120 in total) are connected to each other within four dense blocks (Fig 1). Within a dense block, all feature maps from one layer are concatenated to those produced by the previous layer, which have the same size. Downsampling is performed after each dense block using convolution and pooling operations. At the end of the network, the resulting feature maps from the last block are average pooled and passed through a fully connected layer. For the chest X-ray classifier, the fully connected layer finally converts the feature maps into a linear vector, with each element representing the score for a specific disease. These scores will be then transformed into probabilities ranging from 0 to 1 using a sigmoid layer, assuming a multi-label classification problem (i.e., one image can be classified with more than one disease independently of the probability for the rest of the diseases). The model used pre-trained weights from the ImageNet dataset [50] rather than being trained from scratch.

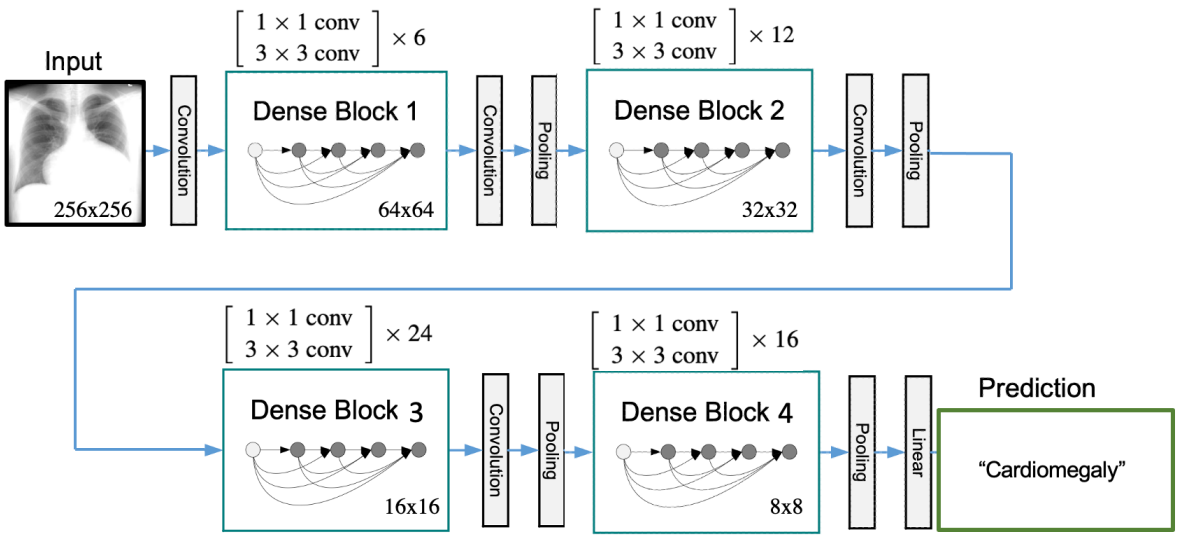

Fig. 1. Densenet-121 architecture. Denset-121 consists of a total of 120 convolutional layers, which are connected to each other within 4 dense blocks (with 6, 12, 24 and 16 convolutional units, respectively). Between blocks, downsampling of the feature maps is done via convolution and pooling (also referred to as transition layers). In case of feeding the network with a 256x256 image, the output size of each dense block are shown at the right-bottom corner. Figure adapted from [49].

1.3.2 Explainability methods

XAI techniques have been categorized under different aspects [51]. First, depending on the complexity of the model being explained, explanations can be generated intrinsically or in a post-hoc way. If generated intrinsically, the model needs to be interpretable on its own, so it must be simple enough to extract explanations or rules during the training of the model. However, image classifiers are normally too complex, making this kind of intrinsic explanation not suitable, though there is ongoing research on interpretable classifiers [52, 53]. Indeed, obtaining a post-hoc explanation is more common in these cases. This is done by analyzing the trained model or attempting to understand the relationships learned between the input and the output. These last options also branch off XAI techniques into model-aware, when they rely on the internal structure of the model, or model-agnostic, when they just assess the relationships between the input and output of the model. And finally, based on the scope of the explanation, we can distinguish between local and global explainers. While local explanations assess single predictions, a global explanation would summarize it to understand the model as a whole. Again, given the complexity of image classifiers, the latter is difficult to achieve.

As previously mentioned, the categorization of XAI techniques for image classifiers mainly focuses on the distinction between model-aware or model-agnostic methods, while fixing the explanation to be local and post-hoc. All these methods will produce an attribution heatmap, but they might follow different approaches for this. In the literature, we can still find another classification based on the approach: gradient- (or backpropagation-) based and perturbation-based techniques. But, for the cases we will present, it translates to the same model-aware vs model-agnostic distinction.

GradCAM [54], Integrated Gradients (IG) [55] and GradientSHAP [56] are three gradient-based (and model-aware) methods recurrently used. Gradients are important for understanding how features influence the score of the model for a given class. The idea of such explainers is to compute the gradients of the output with respect to the input (IG and GradientSHAP) or the extracted features (GradCAM) via backpropagation, using the same weights obtained during the training of the model. These gradients are then used to calculate the attribution of each pixel for the class of interest.

Given a class c, the function targeting the class as F_c and an input x:

- GradCAM [54]: for a given class c, this explainer computes the gradient of the model score $F(x)$ with respect to feature maps A^k by backpropagation until the last convolutional layer ($\frac{\partial F_c(x)}{\partial A^k}$). For each feature map (of size Z), the gradients are average-pooled to compute the neuron importance weights α_c^k as

$$\alpha_c^k = \frac{1}{Z} \sum_i \sum_j \frac{\partial F_c(x)}{\partial A_{ij}^k}$$

  Next, a weighted linear combination of the feature maps is computed to obtain the attribution heatmap, which includes only positive values because of the ReLU function:

$$L_{GradCAM}^c = ReLU(\sum_k \alpha_k^c A^k)$$

  It is important to note that (1) this heatmap will have the same size as the convolutional feature maps (i.e., Z) and (2) $F(x)$ corresponds here to the score given by the model in the form of logits and not the probabilities.

- IG [55]: this method relies on a baseline x', which normally corresponds to a black image. IG generates different image inputs that go from this baseline to the original image, controlled by α. For each of these inputs, it computes the gradients of the output with respect to the original input. The integrated gradient needs to be computed independently for each input dimension i as:

$$IG_i(x) = (x_i - x_i') \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha(x - x'))}{\partial x_i} d\alpha$$

  Normally, to improve the resulting IG heatmaps, SmoothGrad [57] is also implemented. SmoothGrad creates noisy copies of the original image, for which the gradients are computed independently for each case and averaged at the end.

- GradientSHAP: this explainer released by Captum [56], a Pytorch library for model interpretation, follows the ideas behind expected gradients [58] and SHAP values [59]. It intrinsically uses SmoothGrad [57]. Instead of using a single baseline, this is chosen given a baseline distribution D. It also selects random points along the transition from the baseline to the original image (α parameter) and computes the expected gradient of the output with respect to these randomly chosen points. The final SHAP value corresponds to the expected gradient multiplied by the difference

between the input and the baseline value:

$$SHAP_i(x) = (x_i - x_i')\mathbb{E}_{x' \sim D, \alpha \sim U(0,1)} \frac{\partial F(x' + \alpha(x - x'))}{\partial x_i}$$

On the other hand, occlusion [60] is a straightforward example of perturbation-based (model-agnostic) methods. These methods also require a baseline similar to the one used by IG. However, they do not compute the gradients, but simply perturb the original input image and quantify the changes in the model's output.

- Occlusion [60]: by occluding different patches from the original image, the attribution of the occluded patch is computed as the difference between the output given the original image and the perturbed image. The process iteratively hides different image regions, and this results in the final heatmap. The patch size and stride (the pixel distance for patch movement) can be adjusted. When regions are occluded in multiple steps (with a smaller stride than the patch size), the attribution is computed as the average difference in output. It is worth noting that all pixels occluded by one single patch will receive the same attribution value.

2 Results

2.1 The models exhibit reproducible biases and give rise to disease-specific affected subpopulations

We trained and assessed different classifiers to predict the presence or absence of 13 to 14 chest diseases from X-ray images. Following the work done by Seyyed et al. [17], we utilized three datasets with similar data statistics but varying in size (S1): NIH ChestX-ray14 (NIH), CheXpert (CXP), and MIMIC-CXR (MIMIC). To ensure reproducibility, we adopted the same data splits for training, testing, and validation as described in the paper. Additionally, we employed an alternative data split for NIH, where we grouped all the images with a bounding box annotation (i.e., ground truth image localization of a disease) into the test set (S2). This alternative version of NIH is referred to as 'alternative NIH' in the rest of the report. We reduced the size of all the images to reduce the storage space.

Since we aim to reproduce the models from the original work, we needed to first achieve a similar performance. The original study presented underdiagnosis and overdiagnosis rates for the "no finding" label. Hence, we also computed the false positive rate (FPR) and false negative rate (FNR) specifically for the "no finding" label for our models, that had been trained with an upgraded Pytorch [61] version. These rates were calculated for different subpopulations based on the division of patients by sex and age. We obtained very similar results to those in the original work (Fig S1), as well as close accuracy of the models given by the area under the curve (AUC) (Table S3). By ensuring consistency in these performance metrics, we can validate the reproducibility of the models, even with the image resizing and Pytorch upgrade.

With the alternative NIH dataset, we observed changes in the model's performance with respect to the original NIH case. Still, significant biases persisted. While the original study by Seyyed et al. [17] primarily focused on the "no finding" label to detect underdiagnosis and overdiagnosis, our main focus lies in evaluating the model's performance when classifying specific diseases. Specifically, we were interested in assessing the sensitivity of the model in predicting each disease. To accomplish this, we computed the true positive rate (TPR) for each disease and subpopulation (2). Our analysis revealed that while some diseases exhibited lower performance for female and younger patients (such as atelectasis, effusion, mass, nodule, and edema), the detection of other diseases demonstrated inferior results for males (e.g., pneumothorax) and older patients (e.g., cardiomegaly). Hence, the subpopulations affected by lower performance were inconsistent and varied depending on the specific disease being evaluated.

2.2 Explanations with occlusion show the strongest agreement with the ground truth-bounding boxes

In order to understand the reasons behind the model's disease predictions, we wanted to explain the regions of interest that contribute to classifying a label as positive. But before this, we want to ensure the quality and reliability of these explanations. We

Fig. 2. Sensitivity biases in the model. Model performance calculated as the true positive rate or sensitivity for the different subpopulations (by sex: males (M) and females (F); and by age divided into ventiles as 0-20, 20-40, 40-60, 60-80 and 80 or older). The model is trained and tested with the alternative NIH.

analyzed a selection of four explainers for image classification: GradCAM [54], integrated gradients (IG) [55], GradientSHAP [56, 59], and occlusion [60].

To assess which method produced the most accurate explanations, we focused on the bounding boxes that were annotated for some images of the NIH dataset (984 in total). These bounding boxes indicate the true localization of diseases in the images. As these annotated images are exclusively available in the alternative NIH test set, our analysis in this section will solely refer to this dataset. Specifically, we consider the pairs of X-ray images and diseases that were correctly classified as positive (TP) and possess bounding box annotations (574 in total). For each of these pairs, we employed the four explainers to generate attribution heatmaps (i.e., explanations). It is worth noting that we were working with a reduced set of eight disease labels: atelectasis, cardiomegaly, effusion, infiltration, mass, nodule, pneumonia, and pneumothorax. This is because the bounding box coordinates were only provided for the previous version of NIH (NIH ChestX-ray8 [25]).

We conducted a disease-specific evaluation of the attribution heatmaps using three metrics: Intersection over Union (IoU), Point Localization Accuracy (PLA), and the Area Under the Curve (AUC) for the Receiver Operating Characteristic (ROC) curve generated by our attributions (as the prediction) and the bounding boxes (as the true values).

When quantitatively evaluating the attributions with the 3 metrics mentioned above, we detected that GradCAM and occlusion clearly outperformed the other two explainers (Table 1). The accuracy of the attribution heatmaps varies significantly depending on the disease being explained. Cardiomegaly consistently achieved the highest values for IoU, PLA, and AUC, while explanations for atelectasis and nodules tended to have lower values overall. Upon visual inspection of the attribution heatmaps, it was evident that GradCAM and occlusion produced lower-resolution heatmaps compared to IG and GradientSHAP (see Figure S2). This was totally expected given how these explainers work. Additionally, the heatmaps generated by IG and GradientSHAP exhibited high levels of noise and sparsity, confirming their lower quality as assessed quantitatively.

Table 1. Evaluation of the explanations with bounding boxes. Evaluation of the performance of each explainer (GradCAM, occlusion, Integrated Gradients (IG) and GradientSHAP) for different diseases. Intersection over Union (IoU), Point Localization Accuracy (PLA) and the Area Under the Curve (AUC) are computed with the bounding box annotations (i.e., ground truth localization). The dataset (and model) used here is the alternative NIH. Only true positive cases with bounding box annotation are considered.

	GradCAM			Occlusion			IG			GradientSHAP		
	IoU	PLA	AUC	IoU	PLA	AUC	IoU	PLA	AUC	IoU	PLA	AUC
Atelectasis	0.048 (0.058)	0.094	0.544 (0.274)	0.149 (0.097)	0.248	0.787 (0.209)	0.049 (0.041)	0.060	0.533 (0.028)	0.074 (0.044)	0.308	0.573 (0.031)
Cardiomegaly	0.564 (0.103)	0.977	0.949 (0.036)	0.487 (0.104)	0.977	0.901 (0.039)	0.155 (0.018)	0.616	0.548 (0.010)	0.116 (0.035)	0.826	0.576 (0.020)
Effusion	0.154 (0.133)	0.258	0.736 (0.219)	0.174 (0.125)	0.292	0.783 (0.176)	0.065 (0.036)	0.142	0.531 (0.019)	0.064 (0.042)	0.333	0.563 (0.032)
Infiltration	0.242 (0.192)	0.485	0.756 (0.237)	0.264 (0.154)	0.606	0.800 (0.161)	0.082 (0.041)	0.161	0.528 (0.016)	0.089 (0.052)	0.475	0.575 (0.029)
Mass	0.160 (0.156)	0.317	0.860 (0.203)	0.222 (0.149)	0.683	0.855 (0.157)	0.033 (0.037)	0.024	0.507 (0.031)	0.089 (0.053)	0.463	0.580 (0.033)
Nodule	0.022 (0.018)	0.045	0.731 (0.245)	0.055 (0.056)	182	0.874 (0.195)	0.008 (0.005)	0.114	0.520 (0.032)	0.060 (0.046)	0.296	0.577 (0.035)
Pneumonia	0.302 (0.193)	0.444	0.891 (0.123)	0.276 (0.156)	0.833	0.802 (0.148)	0.093 (0.053)	0.222	0.537 (0.016)	0.096 (0.045)	0.390	0.579 (0.022)
Pneumothorax	0.089 (0.118)	0.102	0.710 (0.173)	0.151 (0.143)	0.245	0.769 (0.139)	0.050 (0.032)	0.122	0.517 (0.015)	0.031 (0.030)	0.041	0.525 (0.031)

Since occlusion (1) generally showed higher accuracy than GradCAM and (2) it could be tuned with some hyperparameters choice (i.e., the occlusion patch size and stride), we decided to explore it further. The first explanations with occlusion were created using a patch of size 32x32 and stride 32. We conducted the same evaluation experiments but for four other cases, using smaller patches and strides to aim for higher-resolution heatmaps. A higher resolution translated to a larger computational time to obtain the heatmaps Table S4. Even though we expected higher-resolution heatmaps to perform better, it was hard to conclude it from the results in Fig 3. While IoU and PLA remained relatively stable across different hyperparameter choices, the differences were not consistent among diseases. Conversely, changes in the AUC consistently followed the same trends across diseases. High-resolution heatmaps (i.e., small patch and stride) achieved lower AUC values, but this happened especially when the patch size and stride have the same size. Based on these AUC results and considering computational time, occlusion with a patch size of 32x32 and stride of 16 seems to be a reasonable choice.

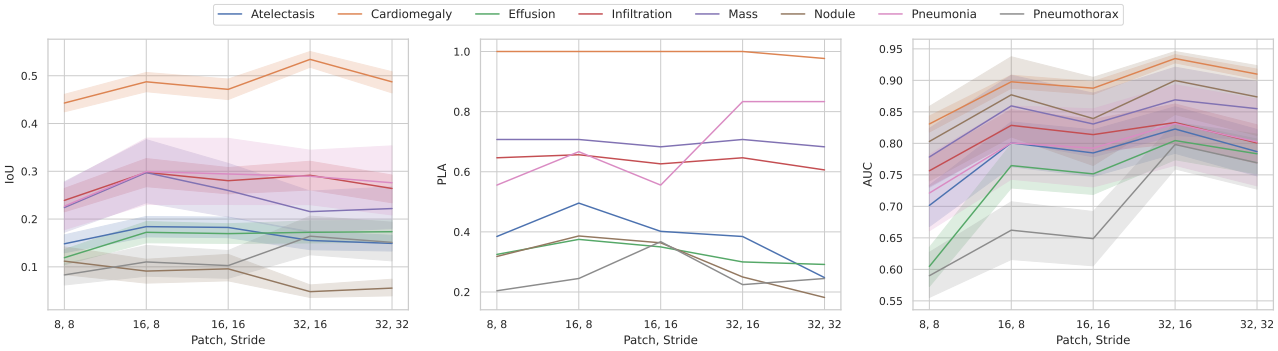

Fig. 3. Occlusion performance with different hyperparameters. Evaluation of the occlusion explanations for different hyperparameters choices. Different window and stride sizes are evaluated. The x-axis go from smaller to larger patches and strides, which correspond to high-resolution to low-resolution heatmaps. Intersection over Union (IoU), Point Localization Accuracy (PLA) and the Area Under the Curve (AUC) are computed with the bounding box annotations (i.e., ground truth localization). AUC and IoU are shown with the average value and the 95% confidence interval. The dataset (and model) used here is the alternative NIH. Only true positive cases with bounding box annotation are considered.

We observed variations in the behavior of the metrics IoU, PLA, and AUC when evaluating the various occlusion hyperparameter choices. This rose the question of which metric was most suitable for evaluating explanations. Analyzing the results obtained with a patch size of 32x32 and a stride size of 16 in occlusion, we examined the correlations between these metrics (Fig 4A-C). IoU and PLA showed a stronger correlation with each other than with AUC. However, PLA can not be used to evaluate individual explanations and it is instead calculated for a group of explanations, in this case for each disease (see section 3.5.1). Furthermore, we explored the correlation between IoU and AUC for individual explanations (Fig 4D) and found numerous cases with high AUC but low IoU scores. We visually explored some cases (Fig 4E). In general, explanations with a high IoU and AUC were those that were inside the bounding box, and for which this bounding box was large. Cases with low IoU and low AUC were those where the explanation highlights a completely different region. But in cases of disagreement, where AUC was high but IoU was very low, the correct region was highlighted by the explanation, but the small size of

the bounding box penalized the IoU score. It is worth mentioning that the computation of the IoU required the binarization of the heatmap with a threshold that directly affects the size of the highlighted region. Therefore, we believe AUC is a more reliable choice for evaluating this type of explanation. It seems to be more flexible, but still coherent, with the comparison between explanations and bounding boxes.

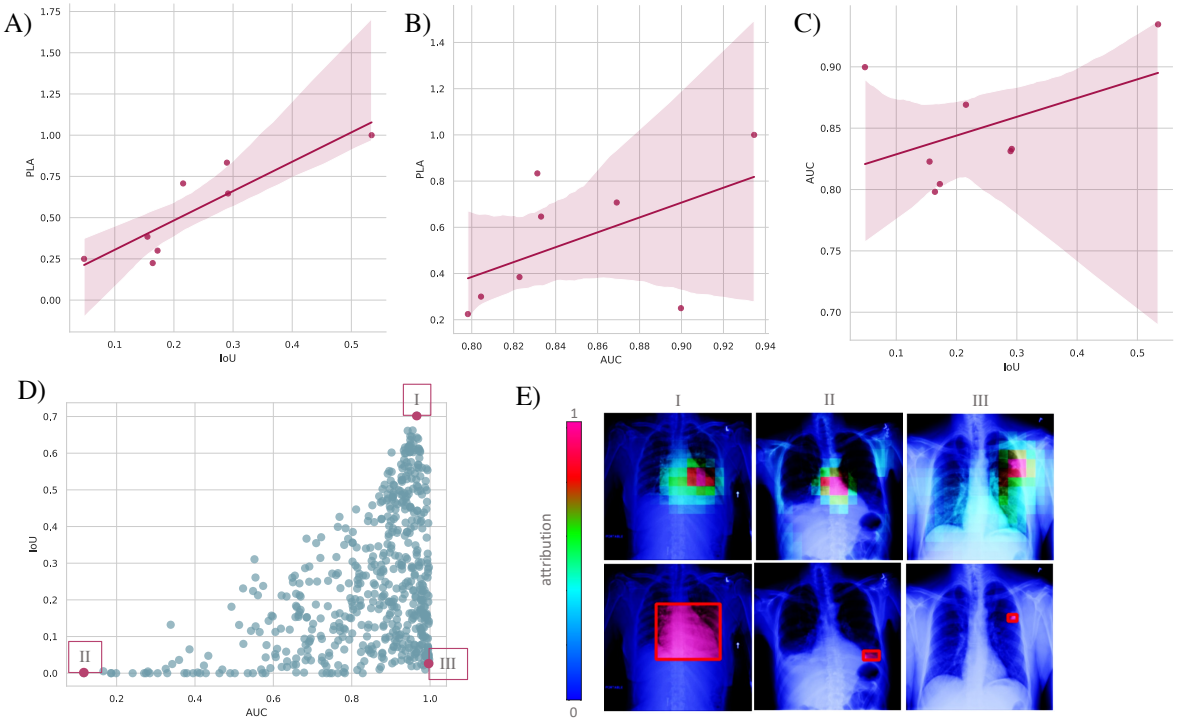

Fig. 4. Agreement among the metrics used to evaluate the explanations with bounding boxes. (A-C) depict the correlation between Intersection over Union (IoU), Point Localization Accuracy (PLA) and the Area Under the Curve (AUC) metrics that have been computed for each disease (for IoU and AUC, the average is presented). Regression lines with 95% confidence interval are plotted. (D) shows the correlation between IoU and AUC individually for each case, without its aggregation into diseases. Three individual cases are annotated (I-III), which are shown in panel (E). For each case, explanations with occlusion are shown at the top, and the bounding box annotation is shown at the bottom. The attributions are normalized between [0, 1] for visualization purposes (E). The dataset (and model) used for this analysis is the alternative NIH. Only true positive cases with bounding box annotation are considered. For (E), the attributions are normalized between [0, 1] for visualization purposes.

2.3 The explanations of the decisions of the models are equally accurate among different subpopulations

Once we have selected the best explainer and evaluation metrics, we wanted to know whether there were subpopulations for which the explanations were worse. Since we had previously detected model biases, could we find similar biases in the explanations?

We assessed the explanations from occlusion with the AUC metric. Instead of

computing the average AUC of all the explanations for a specific disease, we computed it for the disease but also for each subpopulation (Fig 5). Our analysis showed that, overall, the accuracy of the explanations was similar among the subpopulations. This similarity was evident for diseases like cardiomegaly, while diseases like atelectasis did exhibit some differences. However, the mean AUC values for atelectasis still fell within the error bars of the other subpopulations. Based on these findings, we argue that although the model performs differently on various subpopulations, the explainer did not amplify these biases. The explanations remained equally accurate once obtained, regardless of the subpopulation.

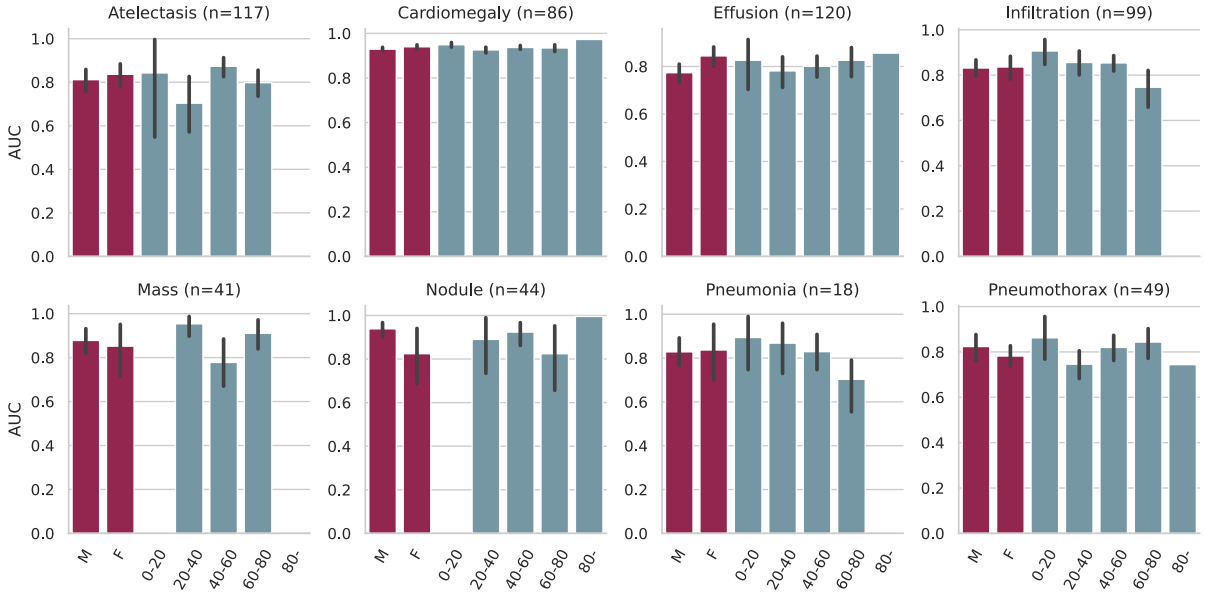

Fig. 5. Performance of the explainer on different subpopulations. Area Under the Curve (AUC) computed with the bounding box annotations (i.e., ground truth localization) and the attributions given by occlusion, given for each subpopulation (by sex: males (M) and females (F); and by age divided into ventiles as 0-20, 20-40, 40-60, 60-80 and 80 or older). The dataset (and model) used here is the alternative NIH. All test images that were correctly classified for the presence of a disease (i.e., true positives) and that had the bounding box annotation were considered for this analysis. 95% confidence intervals are shown. For each disease, the number of cases is shown.

However, our analysis might have limited statistical power. We only focused on TP cases with bounding box annotations, resulting in a total of 574 cases, which further decreases when analyzing them by disease and by subpopulation.

2.4 The evaluation of the explanations without bounding boxes seems to be unsuitable for chest X-ray images

We tried to expand our analysis by incorporating two metrics that do not rely on bounding boxes, allowing us to overcome the limitation of only working with TP cases. This would enable us to extend our analysis to larger datasets, including CXP and MIMIC, and even the full NIH dataset by considering all cases.

First, we calculated the faithfulness correlation [44] for the same explanations evaluated in the previous section. This metric computes the correlation between the change in model output and the attributions of subsets of pixels that are iteratively masked from the original image (see section 3.5.2). We compared this metric with the AUC score, but we could not detect any correlation (Fig S3). Therefore, we were reluctant to use this metric to extend the bias analysis to other data.

Second, we wanted to evaluate the explanations with deletion or insertion curves [45]. These curves are obtained with the model probabilities that are given for the same image but with different parts of it masked. The image is iteratively masked based on the most important regions, given by the heatmap attributions. From these curves, we then compute the Area Under the Deletion Curve and Insertion Curve (AUDC and AUIC). For the following analysis, we used the normal NIH dataset and the GradCAM explainer, although we argue that it can be perfectly translated to any other dataset and explainer choice.

When plotting some of these curves as a sanity check, we detected an unusual behavior in some cases (Fig S4). When inserting the important regions, the probability did not significantly rise until adding approximately 75% of the original image. In contrast, for the deletion curves, we could see they followed the expected behavior: the probability went quickly down when masking important regions of the image. Surprisingly, when creating a deletion curve by randomly masking regions without following the importance order, it appeared very similar to the curves generated using the ordered deletion procedure. Since the mask we used for the process was a black image, we wondered whether there was a problem with the model. Has the model learned to predict certain diseases based on black patterns? Is the model "modeling" the black instead of anatomical features?

To answer this, we used another masking technique. Instead of masking the image with black regions, we also mask it with the same regions but coming from another image, which corresponded to the mean image of the whole test set. We hypothesized a larger disparity between curves created randomly and by importance order when using the mean image mask compared to the black mask.

We compared the AUDC for curves obtained by masking regions randomly and by importance order, both masking with black regions or with the mean image. We computed the difference between AUDC obtained randomly and by importance order (AUDCdiff) for the two mask cases. We have finally compared the normalized AUDCdiff obtained with the two masks for each image associated with each disease (Fig 6). We can see there was no clear difference between using the mean image instead of a black image when masking. Since we were no longer restricted to the cases with bounding box annotation, we performed this analysis also for the rest of the diseases (Fig S5) and observed the same results. It is also worth mentioning that there was a great number of cases for which the AUDCdiff was negative, meaning that a random deletion would obtain a better score (lower AUDC). Based on these findings, we rejected our initial hypothesis about the model and argue that the issue lies with the evaluation metric itself. For chest X-ray images, the evaluation of explanations using deletion or insertion curves appears to be unsuitable.

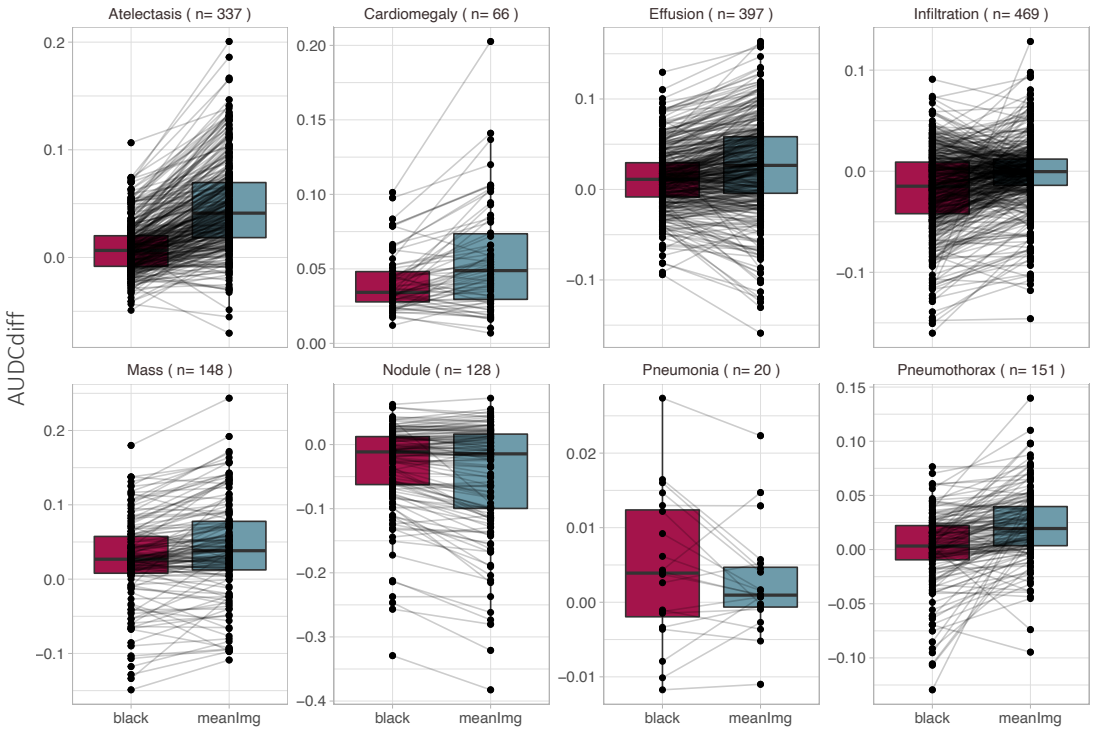

Fig. 6. Comparison between masks used for the deletion curves. The difference between the area under the curve of deletion curves produced with (1) a descending order of the pixels by importance and (2) a random order of the pixels, are shown as the AUDCdiff. The AUDCdiff is shown for two masks that were used to generate the deletion curves, the black image and the mean image of all test images (meanImg). Paired lines are depicted to show individual comparisons. The analysis was performed for all true positive cases of the test set with the NIH dataset and the GradCAM explainer.

3 Methods

3.1 Chest X-Ray datasets

Based on Seyyed's work [17], we utilized three datasets: NIH ChestXRay 14 (NIH) [25], CheXpert (CXP) [27], and MIMIC-CXR (MIMIC) [28]. These datasets consist of multiple images per patient, belonging to either a single or different studies (time points). Images were annotated manually or using natural language processing, identifying 15 labels (NIH) or 14 labels (CXP and MIMIC), which include chest diseases and an additional label for "no finding." In CXP and MIMIC, we treated unknown or uncertain labels as negative cases. To ensure label consistency, we considered "no finding" as positive only when no diseases were present and negative if at least one positive disease was identified. Patient sex and age were annotated for all images. In the case of MIMIC, we merged MIMIC-CXR and MIMIC-IV [62] to obtain patient demographics. Additional technical and demographic information about the datasets can be found in [STable1]. Access to these datasets required a data use agreement and completion of the credentialing process, including a small course through Physionet [63] for MIMIC.

To reduce resource requirements, we stored smaller versions of the images by fixing the image height at 512 pixels while maintaining the original proportions.

Initially, we split the data into training-validation-test sets using Seyyed et al.'s partitioning approach [17], resulting in approximately 80-10-10 splits without any patient overlap [STable1]. However, our MIMIC splits slightly differed due to unmatched patients resulting from merging MIMIC-CXR and MIMIC-IV. In addition, we created a different data split for NIH to obtain a test set containing patients with at least one image annotated with bounding boxes, which provide ground truth explanations of disease localization in the images. This alternative NIH split was necessary because the overlap between the images in the original test set and those with bounding boxes was insufficient (49 images instead of 984). The alternative NIH split followed a proportion of 70-15-15.

3.2 Chest X-Ray classifiers

We used the code to build the models (classifiers) from the work by Seyyed et al. [17], with slight modifications for GPU reproducibility. These models, implemented in PyTorch, adopt the DenseNet-121 architecture [49] and utilize ImageNet [50] pre-trained weights. The initial learning rate was 5e-4 and halved if no validation loss improvement was observed for 3 epochs. The training concluded after 10 epochs without improvement. Data augmentation involved random flips and rotations, with images resized to 256x256 pixels and normalized using ImageNet statistics. As a multi-label classifier, a sigmoid function was applied to obtain label probabilities. These probabilities were then transformed into binary predictions using a threshold learned during the validation process.

We trained the models for each dataset (NIH, alternative NIH, CXP, and MIMIC) with an updated PyTorch version (v1.2.0) that is compatible with the explainability library Captum [56].

3.3 Model performance metrics

To evaluate the model's performance and compare it to the original results, we calculated the average AUC. We also computed the false positive rate (FPR) and false negative rate (FNR) for the "no finding" label. Additionally, we assessed the model's sensitivity by calculating the TPR for each disease. These rates were computed for specific subpopulations based on patient stratification by sex and age, divided into ventiles ranging from 0-20 to 80 and older.

3.4 Explanation of the model decision

To interpret the model's decisions for image and disease label pairs, we utilized the Captum library [56], which provides model interpretability functionalities in PyTorch [61]. We selected four explainers from Captum: GradCAM [54], Integrated Gradients (IG) [55], GradientShap [56, 59], and occlusion [60]. These explainers generate pixel-level attributions, visualized as heatmaps, indicating the importance of each pixel for the model's prediction.

For GradCAM, we obtained explanations from the final convolutional layer, resulting in an 8x8 resolution heatmap. For IG, we applied SmoothGrad [57] (Noise Tunnel in Captum), which involves creating multiple images with random noise and averaging the attributions obtained for this set of images. Since GradientShap already incorporates SmoothGrad, we used it with default parameters. As for occlusion, we experimented with different hyperparameter choices to refine the explanations. We started with a window and stride size of 32x32 (inspired by [42]), gradually reducing the parameters by two until reaching a window and stride size of 8. Except for GradCAM, all explainers required a baseline, which we set as a black image. For all the explainers, we only take positive attributions into account.

We generated explanations solely from the test set, excluding the "no finding" label and misclassified images.

3.5 Metrics for evaluating the explanation

We assessed the explanations using different approaches based on the availability of bounding boxes.

3.5.1 Metrics based on the bounding box annotation

For those images that have the bounding box annotation, we used the following metrics:

- Intersection over union (IoU): it computes the ratio of the overlap area to the union area between the binarized attributions heatmap and the bounding box. To binarize the heatmap, we applied Otsu's thresholding method (as in [42]).
- Point Localization Accuracy (PLA) [proposed in [43]]: it measures the accuracy of localizing the most important pixel, defined as the pixel with the highest attribution,

within the bounding box. The PLA is calculated by dividing the number of hits (where the most important pixel is inside the bounding box) by the total number of cases considered. This metric is computed independently for each disease. For low-resolution heatmaps such as GradCAM and occlusion, we determine the most important pixel as the pixel located in the middle of the most important region of pixels.

- Area Under the Curve (AUC): it considers the explanation as the prediction and the bounding box as the ground truth. Then the AUC is computed as it is normally done for model performance assessment. To compute the AUC, we first convert the 2D matrices of the heatmap and the bounding box into 1D arrays.

3.5.2 Metrics not based on the bounding box annotation

Alternatively, we used two other evaluation metrics that do not rely on the bounding boxes but only on the model output and the attributions:

- Faithfulness correlation (FC) [44]: it is computed by masking a random subset of image pixels with black in an iterative manner. The Pearson correlation is then calculated between the difference in model output and the sum of the attributions assigned to those masked pixels. In each iteration, we compute (1) the difference between the model output with the original image and the model output with the partially masked image, and (2) the sum of attribution values corresponding to the masked pixels. We perform 200 iterations, masking a subset of 1024 pixels each time.
- Area Under the Deletion Curve (AUDC) [45]: it is computed by iteratively masking regions of the image in a descending order based on the attributions heatmap. Starting from the original image, we mask regions one by one until the entire image is masked. At each step, we recompute the model output and construct the deletion curve. The area under the deletion curve is then computed. The smaller the AUDC, the better the explanation.
- Area Under the Insertion Curve (AUIC) [45]: starting with a completely masked image and gradually revealing regions by inserting them one by one, from most to least important regions according to the heatmap. At each step, we recompute the model output and construct the deletion curve. The area under the deletion curve is then computed. The larger the AUIC, the better the explanation.

For both AUDC and AUIC, we worked with regions of size 32x32, which translated to 64 iterations (images that are fed to the model are 256x256). Before computing the AUDC, we normalized the x-axis with the number of iterations to [0, 1]. The normalized axis corresponds to the fraction of perturbed pixels. We employed two types of masks: a black mask and the mean image of all the images in the test set.

Additionally, we computed the AUDC for deletion curves created by randomly masking regions of the image, without sorting based on attribution values. We then evaluated the difference between the AUDC of the descending-deletion curve and the random-deletion curve, which we refer to as AUDCdiff. To facilitate a meaningful

comparison of this difference between the black mask and mean image mask (which have different endpoints, which correspond to the probability of the fully masked image), we normalized the AUDCdiff for the mean image mask by fixing the curve range and using the AUDCdiff obtained with the black mask as a reference. We do so with this formula:

$$\text{AUDCdiff (mean image)} = \text{AUDCdiff (black)} \frac{\text{deletion curve range (black)}}{\text{deletion curve range (mean image)}}$$

where the deletion curve range is the difference between the endpoint (i.e., probability given the completely masked image: either black or mean image) and the start point (i.e., probability given the original image: same for both masking strategies).

4 Discussion

The final purpose of this work was to use XAI to understand fairness issues on chest X-ray image classifiers. The starting point was to reproduce the models used by Seyyed et. al., for which they reported underdiagnosis biases for underserved subpopulations [17]. Therefore, we have analyzed the performance disparities for the classification of individual diseases among different subpopulations (stratified by sex and age). Then, we aimed to use XAI techniques, the explainers, to understand the decisions given by the biased models. In order to ensure a good quality of the explanations, we have evaluated four popular explainers by quantitatively comparing them with (ground truth) bounding boxes. Occlusion outperformed the other explainers. With the previous evaluation, we also compared the accuracy of the explanations for each subpopulation, and this suggested new biases. Since the number of cases used in this analysis was low, we also assessed the explanations with metrics that do not rely on the bounding boxes, known as fidelity metrics. However, these metrics - especially, deletion curves - did not work well with chest X-ray images.

In terms of reproducing the biased models, we have found the same underdiagnosis biases they reported for the three datasets: NIH, CXP and MIMIC. Even when using an updated Pytorch version, compatible with the XAI techniques. It is important to note that NIH shows a different underdiagnosed group (males) than CXP and MIMIC (young, females), also in the original results. Very slight differences in FPR and FNR exist because (1) we ran the model once whereas they run it 5 times, and (2) for CXP and MIMIC we corrected the annotations to be consistent with the "no finding" label. Also, for the MIMIC case, we were not able to obtain the exact original dataset due to unmatched patients between the images and the demographics from MIMIC-IV. Alternatively, for NIH we have used a customized data split, for which all the images published with bounding boxes were present in the test set. We have also built a model for this and detected major changes with respect to the original results for NIH, with a generally smaller FPR and greater FNR for the "no finding" label. But still, we report male patients being underdiagnosed while female patients being overdiagnosed. This alternative NIH model is used for the rest of the study, if not specified otherwise.

While they basically focused on the underdiagnosis rate as the FPR for the "no finding" label, we were interested in the precision of the model for each disease that can be diagnosed. We have therefore assessed it as the TPR of each disease label. We have seen that, depending on the disease label of interest, this rate changes among subpopulations, exhibiting different biases (i.e., the negatively affected subgroup for a certain disease is not the same as for another disease). Out of the 14 diseases, the model was notably more precise for males in 6 cases and for females in 5 cases. A general increase of the TPR for males was also reported by Seyyed et. al. in a previous publication [19]. In terms of age biases, there are also a few diseases for which the precision increases with the age of the patient. For example, effusion is a clear example of a disease that favors male and older patients. However, it is worth mentioning that we did not perform further analysis with intersectional subpopulations, making these findings to be taken independently for sex and age differences.

On the other hand, using explainability to understand the model decisions was not as straightforward as expected. Given there is no common pipeline to do so, we first encountered two important choices to be made: the explainer and the evaluation of its explanations. For the evaluation/benchmark of the explainer, we have chosen 4 post-hoc explainers that are commonly used in the literature and/or easily to be used with Captum. These were GradCAM, occlusion, IG and GradientSHAP. Following other benchmark works [42, 43], we have used the IoU and PLA to evaluate the explanations (i.e., attribution heatmaps). On top of these two, we have also evaluated them with the AUC that results from considering the bounding box as the ground truth and the heatmap as the prediction. We have evaluated all TP cases independently for each of the 8 diseases with bounding box annotations. With a similar trend for all three metrics, occlusion (a model-agnostic explainer) outperformed the rest. Following occlusion, GradCAM also performed much better than IG and GradientSHAP, which produced highly noisy heatmaps. Similar performance of IG and GradCAM was reported in [42] using the CXP dataset. And the publication of a novel (but not post-hoc) explainer, PYLON, also reported similar GradCAM results with NIH [43]. As shown in our results, they also obtained a much higher accuracy of cardiomegaly explanations, compared to other diseases like pneumothorax.

Since occlusion can be fine-tuned with two hyperparameters, the patch size and stride, we analyzed different combinations to get the best choice. These hyperparameters control the occlusion process of the image and, therefore, the resolution of the explanation. Even though we expected higher resolution explanations to obtain better results, we could see that the results for PLA and IoU did not change much. However, the AUC was higher for lower-resolution heatmaps, but it always showed higher values when the stride was smaller than the patch size. This makes two consecutive occlusion steps hide a shared region of the image, and compute the average of the two attributions. This result seems plausible, as the region to be shared is occluded with different neighboring pixels, and this could improve the attribution heatmap. We have chosen a patch size of 32x32 and a stride of 16 (note that the images are 256x256). It would be interesting to see what would happen for cases with the same patch size but a much smaller stride. Probably it is not the high resolution of the patch that gave us worse performance, but the small occlusion patch.

IoU, PLA and AUC behaved differently in the previous analysis. When assessing their respective correlations, we have seen a better correlation between PLA and IoU than these two with AUC. PLA and IoU are more rigid than AUC, so it makes sense they give more similar results. The problem with PLA is that it can not be calculated for single images (for a single image we can just say whether the most important pixel is inside the bounding box or not), but PLA is the ratio between hits and the total number of cases we are comparing. IoU needs a binarized attribution heatmap, but the threshold choice is not trivial and it is a source of debate in object detection [64]. Also, IoU was reported to be low when comparing different human annotations on the chest X-ray images [42], due to the variable sizes of the bounding boxes the doctors annotated between each other. We have seen that there are a high number of cases for which the AUC was very high but the IoU was extremely low. By exploring some of these cases, we could see that even if the bounding box was similarly located as the important region of the explanation, if this

region was bigger they were highly penalized by the IoU score. Therefore we propose to use the AUC as a more flexible, but still coherent, metric.

We hypothesized that model biases could be translated into what we call explanation biases. So we have evaluated whether the explanations were more accurate for certain subpopulations, with the AUC. We do see that their accuracy change among different subpopulations. However, we can not detect a consistent relationship between these and the model biases defined for underdiagnosis and overdiagnosis. In the case of the detection of nodules, we saw they were underdiagnosed in males and overdiagnosed in females. With this analysis, we see that the explanations for females are worse. In the case of pneumothorax, overdiagnosed in males and underdiagnosed in females, but worse explanations are still produced for females. We might not have enough statistical power to reach a strong conclusion due to the low number of cases we are considering (only TP cases: 574 in total, but to be analyzed independently for each disease). We could not find literature doing this kind of analysis to compare it with. Still, we do think this finding can help us motivating clinical stakeholders to trust XAI since model biases were not found in the explanations.

Finally, we wanted to find a way to assess explanations without relying on the bounding boxes. We have tried two fidelity (or faithfulness) metrics, which should capture how well the explanation captures the true behavior of the model. The first one, the faithfulness correlation, showed no agreement with the previous (and bounding box-based) AUC metric. The second approach with deletion or insertion curves does not seem to be suitable for chest X-ray images. We have detected that a random perturbation of the image produced similar deletion curves to those created when the perturbations followed the importance of the pixels. Also, for the insertion curves we have detected that the model needs a great amount of the original pixel to be able to detect the signal. Since we were perturbing these pixels with a black mask, these behaviors made us think of the following. Maybe the model was focusing on black features instead of anatomical features. We then compared the curves produced by masking the images (1) with a black mask and (2) with a mask that maintained the anatomical structure of the chest. For the latter, we changed image regions by the same regions coming from the mean image of all chest X-rays. However, we have shown that there are no big differences, so random perturbations were close to importance-ordered perturbations for both masks. This means it is not a problem of the model, but it also means that deletion curves are not suitable for our case. Moreover, going back to the explanations we obtained with XAI, where normally the highlighted regions were anatomical features, we can indeed corroborate that the model did not focus on black features.

The fact that deletion curves do not work for this kind of images, but occlusion outperformed other explainers might be surprising. Occlusion builds the explanation also by perturbing the image with black pixels. However, occlusion uncovers the previously occluded region at each step (so in every step there is only one region hidden), while when we create a deletion curve we leave the previously hidden region covered (until we obtain a completely occluded image). We could therefore think that deletion curves remove any correlation between regions of the image that could be important for the model. That is why we also chose to compute the faithfulness correlation, which occludes random pixels of

the image, which are different every iteration. But since this metric was not in agreement with the bounding box metric, we can not reach any conclusion on this.

One work reported that this last deletion process generates out-of-distribution images, and this could be the source of a strange performance of the model since it did not learn such images [65]. They proposed to retrain the model with these partially occluded images with ROAR. While we understand the proposed issue when perturbing with a black value (or any constant value), we argue that we would avoid it when masking with the mean X-ray. This type of mask follows the concept of meaningful perturbation, introduced in [66], and is similar to the approach proposed by Lenis et. al. to inpaint the image with regions from a healthy sample [67], although they use it for explaining the image, not for evaluation. These should not produce such out-of-distribution samples. But instead, we still found that deletion curves with the mean image mask were also problematic.

5 Conclusions

In conclusion, we once again highlight the presence of inherent biases in classifiers for chest X-ray images. While this applies to different datasets, we specifically focused on the model trained using the NIH dataset. Our findings reveal a significant variation in the model's precision among different subpopulations, focusing on patients' sex and age. The subpopulation adversely affected varies depending on the predicted disease, although there is a slight trend towards higher precision for male and older patients.

To address the disparate model performance, we introduced the explainability step, which helps in understanding these variations. Additionally, we recommend an optimized version of occlusion as the XAI technique that provides more accurate explanations compared to doctors' annotations. Moreover, we propose evaluating visual explanations by computing the AUC between the explanations and the ground truth annotations. Our study demonstrates that this metric offers greater flexibility while remaining coherent, unlike commonly used metrics such as IoU.

Despite the presence of precision biases in the model, our evaluation of the explanations provided by occlusion reveals a different picture. Interestingly, we have not observed significant differences in the accuracy of these explanations among the various subpopulations. The explainer seems to not amplify model biases. Given the reported absence of biases in the explanations, our aim is also to instill confidence in clinical stakeholders regarding XAI techniques.

We strongly discourage researchers from using deletion or insertion curves to evaluate visual explanations for chest X-ray images. It is essential to individually examine the behavior of these curves, especially when dealing with other types of images.

Our study contributes to advancing fairness and explainability in chest X-ray image interpretation. Further research will provide valuable recommendations for improving these classifiers before being deployed in hospitals. We have established some initial guidelines to evaluate the classifiers with XAI, but the next steps should be focused on the knowledge we can gain given the explanations. We aim to explain not only the correct

classifications but the misclassifications for correcting the model or, at least, for warning the clinical stakeholders about these mistakes.

6 References

1. Rajpurkar, P., Chen, E., Banerjee, O. & Topol, E. J. AI in health and medicine. en. *Nature Medicine* **28.** Number: 1 Publisher: Nature Publishing Group, 31–38 (2022).

2. Barragán-Montero, A. *et al.* Artificial intelligence and machine learning for medical imaging: A technology review. en. *Physica Medica* **83,** 242–256 (2021).

3. Brady, A. P. The vanishing radiologist—an unseen danger, and a danger of being unseen. en. *European Radiology* **31,** 5998–6000 (2021).

4. Asif, S., Wenhui, Y., Jin, H. & Jinhai, S. *Classification of COVID-19 from Chest X-ray images using Deep Convolutional Neural Network* in *2020 IEEE 6th International Conference on Computer and Communications (ICCC)* (2020), 426–433.

5. Ibrahim, A. U., Ozsoz, M., Serte, S., Al-Turjman, F. & Yakoi, P. S. Pneumonia Classification Using Deep Learning from Chest X-ray Images During COVID-19. en. *Cognitive Computation* (2021).

6. Jha, S. & Topol, E. J. Upending the model of AI adoption. English. *The Lancet* **401.** Publisher: Elsevier, 1920 (2023).

7. Lambert, S. I. *et al.* An integrative review on the acceptance of artificial intelligence among healthcare professionals in hospitals. en. *npj Digital Medicine* **6.** Number: 1 Publisher: Nature Publishing Group, 1–14 (2023).

8. Wang, W., Chen, L., Xiong, M. & Wang, Y. Accelerating AI Adoption with Responsible AI Signals and Employee Engagement Mechanisms in Health Care. en. *Information Systems Frontiers* (2021).

9. Zou, J. & Schiebinger, L. AI can be sexist and racist - it's time to make it fair. eng. *Nature* **559,** 324–326 (2018).

10. Buolamwini, J. & Gebru, T. *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification* in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency* (eds Friedler, S. A. & Wilson, C.) **81** (PMLR, 2018), 77–91.

11. Obermeyer, Z., Powers, B., Vogeli, C. & Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **366,** 447–453 (2019).

12. Panigutti, C., Perotti, A., Panisson, A., Bajardi, P. & Pedreschi, D. FairLens: Auditing black-box clinical decision support systems. *Information Processing and Management* **58.** arXiv: 2011.04049 Publisher: Elsevier Ltd, 102657 (2021).

13. Obermeyer, Z., Powers, B., Vogeli, C. & Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **366,** 447–453 (2019).

14. Saleiro, P. *et al. Aequitas: A Bias and Fairness Audit Toolkit* en. 2018.

15. Chen, I. Y. *et al.* Ethical Machine Learning in Healthcare. *Annual Review of Biomedical Data Science* **4,** 123–144 (2021).

16. Beauchamp, T. & Childress, J. *Principles of Biomedical Ethics* (Oxford University Press, 1979).

17. Seyyed-Kalantari, L., Zhang, H., McDermott, M. B., Chen, I. Y. & Ghassemi, M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature Medicine* **27.** Publisher: Springer US, 2176–2182 (2021).

18. Larrazabal, A. J., Nieto, N., Peterson, V., Milone, D. H. & Ferrante, E. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences of the United States of America* **117.** ISBN: 1919012117, 12592–12594 (2020).

19. Seyyed-Kalantari, L., Liu, G., McDermott, M., Chen, I. Y. & Ghassemi, M. CheXclusion: Fairness gaps in deep chest X-ray classifiers. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* **26.** arXiv: 2003.00827, 232–243 (2020).

20. Puyol-Antón, E. *et al.* Fairness in Cardiac Magnetic Resonance Imaging: Assessing Sex and Racial Bias in Deep Learning-Based Segmentation. *Frontiers in Cardiovascular Medicine* **9** (2022).

21. Ribeiro, F., Shumovskaia, V., Davies, T. & Ktena, I. *How fair is your graph? Exploring fairness concerns in neuroimaging studies* in *Proceedings of the 7th Machine Learning for Healthcare Conference* (eds Lipton, Z., Ranganath, R., Sendak, M., Sjoding, M. & Yeung, S.) **182** (PMLR, 2022), 459–478.

22. Saeed, W. & Omlin, C. Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities. en. *Knowledge-Based Systems* **263,** 110273 (2023).

23. Van der Velden, B. H. M., Kuijf, H. J., Gilhuijs, K. G. A. & Viergever, M. A. Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. en. *Medical Image Analysis* **79,** 102470 (2022).

24. Nazir, S., Dickson, D. M. & Akram, M. U. Survey of explainable artificial intelligence techniques for biomedical imaging with deep neural networks. en. *Computers in Biology and Medicine* **156,** 106668 (2023).

25. Wang, X. *et al.* ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017* **2017-Janua.** arXiv: 1705.02315 ISBN: 9781538604571, 3462–3471 (2017).

26. Rajpurkar, P. *et al.* CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. arXiv: 1711.05225, 3–9 (2017).

27. Irvin, J. *et al.* CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *33rd AAAI Conference on Artificial Intelligence.* arXiv: 1901.07031 ISBN: 9781577358091, 590–597 (2019).

28. Johnson, A. E. *et al.* MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data* **6.** Publisher: Springer US, 1–8 (2019).

29. Bustos, A., Pertusa, A., Salinas, J. M. & de la Iglesia-Vayá, M. PadChest: A large chest x-ray image dataset with multi-label annotated reports. *Medical Image Analysis* **66.** arXiv: 1901.07441, 1–35 (2020).

30. Bernhardt, M., Jones, C. & Glocker, B. Potential sources of dataset bias complicate investigation of underdiagnosis by machine learning algorithms. *Nature Medicine* **28.** Publisher: Springer US, 1157–1158 (2022).

31. Seyyed-Kalantari, L., Zhang, H., McDermott, M. B., Chen, I. Y. & Ghassemi, M. Reply to: 'Potential sources of dataset bias complicate investigation of underdiagnosis by machine learning algorithms' and 'Confounding factors need to be accounted for in assessing bias by machine learning algorithms'. *Nature Medicine* **28.** Publisher: Springer US, 1161–1163 (2022).

32. Marcinkevi˘cs, R., Ozkan, E. & Vogt, J. E. Debiasing Deep Chest X-Ray Classifiers using Intra- and Post-processing Methods. en.

33. Zhang, H. *et al. Improving the Fairness of Chest X-ray Classifiers* arXiv:2203.12609 [cs, eess]. 2022.

34. Luo, L., Xu, D., Chen, H., Wong, T.-T. & Heng, P.-A. *Pseudo Bias-Balanced Learning for Debiased Chest X-Ray Classification* in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022* (eds Wang, L., Dou, Q., Fletcher, P. T., Speidel, S. & Li, S.) (Springer Nature Switzerland, Cham, 2022), 621–631.

35. DeGrave, A. J., Janizek, J. D. & Lee, S.-I. AI for radiographic COVID-19 detection selects shortcuts over signal. en. *Nature Machine Intelligence* **3,** 610–619 (2021).

36. Adleberg, J. *et al.* Predicting Patient Demographics From Chest Radiographs With Deep Learning. *Journal of the American College of Radiology* **19.** Publisher: American College of Radiology, 1151–1161 (2022).

37. Glocker, B., Jones, C., Bernhardt, M. & Winzeck, S. Algorithmic encoding of protected characteristics in chest X-ray disease detection models. English. *eBioMedicine* **89.** Publisher: Elsevier (2023).

38. Miró-Nicolau, M., Moyà-Alcover, G. & Jaume-i-Capó, A. Evaluating Explainable Artificial Intelligence for X-ray Image Analysis. en. *Applied Sciences* **12,** 4459 (2022).

39. Visuña, L., Yang, D., Garcia-Blas, J. & Carretero, J. Computer-aided diagnostic for classifying chest X-ray images using deep ensemble learning. *BMC Medical Imaging* **22,** 178 (2022).

40. Wang, B. *et al.* Automatic creation of annotations for chest radiographs based on the positional information extracted from radiographic image reports. *Computer Methods and Programs in Biomedicine* **209,** 106331 (2021).

41. Hou, J. & Gao, T. Explainable DCNN based chest X-ray image analysis and classification for COVID-19 pneumonia detection. en. *Scientific Reports* **11.** Number: 1 Publisher: Nature Publishing Group, 1–15 (2021).

42. Saporta, A. *et al.* Benchmarking saliency methods for chest X-ray interpretation. en. *Nature Machine Intelligence* **4.** Number: 10 Publisher: Nature Publishing Group, 867–878 (2022).

43. Preechakul, K., Sriswasdi, S., Kijsirikul, B. & Chuangsuwanich, E. Improved image classification explainability with high-accuracy heatmaps. en. *iScience* **25,** 103933 (2022).

44. Bhatt, U., Weller, A. & Moura, J. M. F. *Evaluating and Aggregating Feature-based Model Explanations* en. in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence* (International Joint Conferences on Artificial Intelligence Organization, Yokohama, Japan, 2020), 3016–3022.

45. Petsiuk, V., Das, A. & Saenko, K. *RISE: Randomized Input Sampling for Explanation of Black-box Models* arXiv:1806.07421 [cs]. 2018.

46. Gomez, T., Fréour, T. & Mouchère, H. *Metrics for saliency map evaluation of deep learning explanation methods* arXiv:2201.13291 [cs]. 2022.

47. Tomsett, R., Harborne, D., Chakraborty, S., Gurram, P. & Preece, A. Sanity Checks for Saliency Metrics. en. *Proceedings of the AAAI Conference on Artificial Intelligence* **34,** 6021–6029 (2020).

48. Bylinskii, Z., Judd, T., Oliva, A., Torralba, A. & Durand, F. *What do different evaluation metrics tell us about saliency models?* arXiv:1604.03605 [cs]. 2017.

49. Huang, G., Liu, Z., van der Maaten, L. & Weinberger, K. Q. *Densely Connected Convolutional Networks* arXiv:1608.06993 [cs]. 2018.

50. Deng, J. *et al. Imagenet: A large-scale hierarchical image database* in *2009 IEEE conference on computer vision and pattern recognition* (2009), 248–255.

51. Adadi, A. & Berrada, M. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access* **6.** Conference Name: IEEE Access, 52138–52160 (2018).

52. Chen, Z., Bei, Y. & Rudin, C. Concept whitening for interpretable image recognition. *Nature Machine Intelligence* **2.** arXiv: 2002.01650 Publisher: Springer US ISBN: 4225602000265, 772–782 (2020).

53. Chen, C. *et al.* This looks like that: Deep learning for interpretable image recognition. *Advances in Neural Information Processing Systems* **32.** arXiv: 1806.10574, 1–12 (2019).

54. Selvaraju, R. R. *et al. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization* en. in *2017 IEEE International Conference on Computer Vision (ICCV)* (IEEE, Venice, 2017), 618–626.

55. Sundararajan, M., Taly, A. & Yan, Q. *Axiomatic Attribution for Deep Networks* en. arXiv:1703.01365 [cs]. 2017.

56. Kokhlikyan, N. *et al. Captum: A unified and generic model interpretability library for PyTorch* 2020.

57. Smilkov, D., Thorat, N., Kim, B., Viégas, F. & Wattenberg, M. *SmoothGrad: removing noise by adding noise* arXiv:1706.03825 [cs, stat]. 2017.

58. Erion, G., Janizek, J. D., Sturmfels, P., Lundberg, S. M. & Lee, S.-I. Improving performance of deep learning models with axiomatic attribution priors and expected gradients. *Nature Machine Intelligence* **3,** 620–631 (2021).

59. Lundberg, S. M. & Lee, S.-I. *A Unified Approach to Interpreting Model Predictions* in *Advances in Neural Information Processing Systems* **30** (Curran Associates, Inc., 2017).

60. Zeiler, M. D. & Fergus, R. en. in *Computer Vision – ECCV 2014* (eds Fleet, D., Pajdla, T., Schiele, B. & Tuytelaars, T.) Series Title: Lecture Notes in Computer Science, 818–833 (Springer International Publishing, Cham, 2014).

61. Paszke, A. *et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library* arXiv:1912.01703 [cs, stat]. 2019.

62. Johnson, A. *et al. MIMIC-IV* 2020.

63. Goldberger, A. L. *et al.* PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. *Circulation* **101.** Circulation Electronic Pages: http://circ.ahajournals.org/content/101/23/e215.full PMID:1085218; doi: 10.1161/01.CIR.101.23.e215, e215–e220 (2000 (June 13)).

64. Padilla, R., Passos, W. L., Dias, T. L. B., Netto, S. L. & da Silva, E. A. B. A Comparative Analysis of Object Detection Metrics with a Companion Open-Source Toolkit. en. *Electronics* **10.** Number: 3 Publisher: Multidisciplinary Digital Publishing Institute, 279 (2021).

65. Hooker, S., Erhan, D., Kindermans, P.-J. & Kim, B. *A Benchmark for Interpretability Methods in Deep Neural Networks* in *Advances in Neural Information Processing Systems* **32** (Curran Associates, Inc., 2019).

66. Fong, R. & Vedaldi, A. *Interpretable Explanations of Black Boxes by Meaningful Perturbation* in *Proceedings of the IEEE International Conference on Computer Vision* (2017), 3429–3437.

67. Lenis, D. *et al. Domain Aware Medical Image Classifier Interpretation by Counterfactual Impact Analysis* en. in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020* (eds Martel, A. L. *et al.*) (Springer International Publishing, Cham, 2020), 315–325.

7 Supplementary data

7.1 Supplementary figures

Fig. S1. Reproduction of the models' biases. Comparison of the False Positive Rate (FPR) and False Negative Rate (FNR) for the "no finding" label across subpopulations (by sex: males (M) and females (F); and by age divided into ventiles as 0-20, 20-40, 40-60, 60-80 and 80 or older). Our results are depicted on the left (A, C, E, G) and the originally reported results on the right (B, D, F) [[31]], which are presented for 5 runs (95% confidence intervals are plotted). A-B correspond to the NIH dataset, C-D to the CXP dataset, and E-F to the MIMIC dataset. G corresponds to the alternative NIH model (with a customized data split). Note our models have been trained only once with a previous image resizing and with an upgraded Pytorch version.

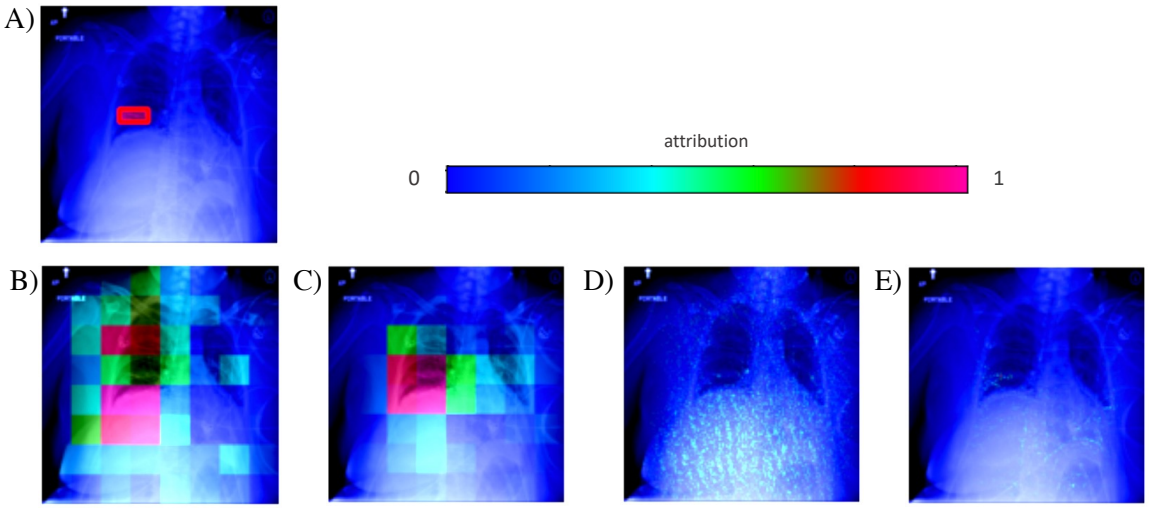

Fig. S2. [Example of the explanations produced by the different explainers. An example of an NIH image with atelectasis is shown. (A) depicts the bounding box annotation (i.e., the ground truth disease localization), while (C-E) show the attribution heatmap produced by GradCAM (B), occlusion (C), integrated gradients (D) and GradientSHAP (E). The attributions are normalized between [0, 1] for visualization purposes.

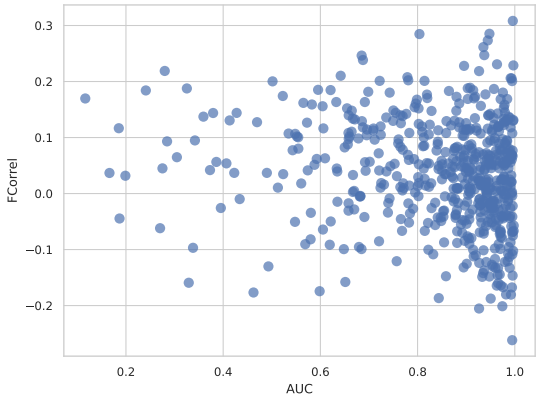

Fig. S3. Correlation between AUC and Faithfulness correlation. Area Under the Curve (AUC) metric using the bounding box annotation vs faithfulness correlation, which does not rely on the bounding box.

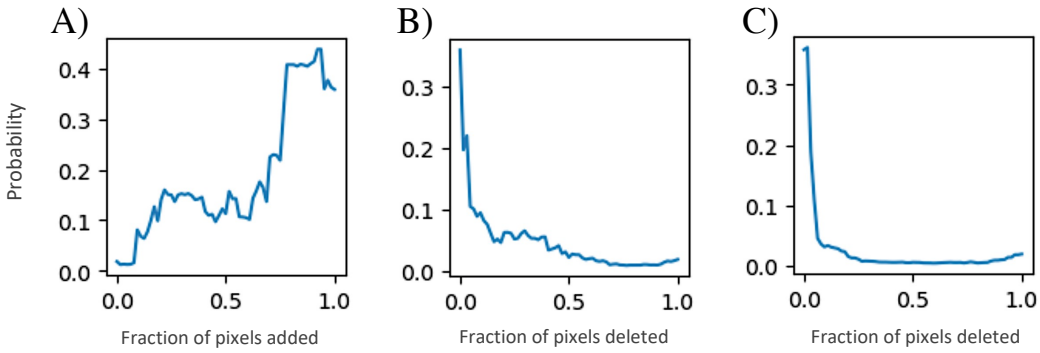

Fig. S4. Example of the strange behavior of insertion and deletion curves. Given the GradCAM explanation for a NIH image with nodule, this depicts the (A) insertion curve, (B) deletion curve, and (C), deletion curve when masking regions with a random order.

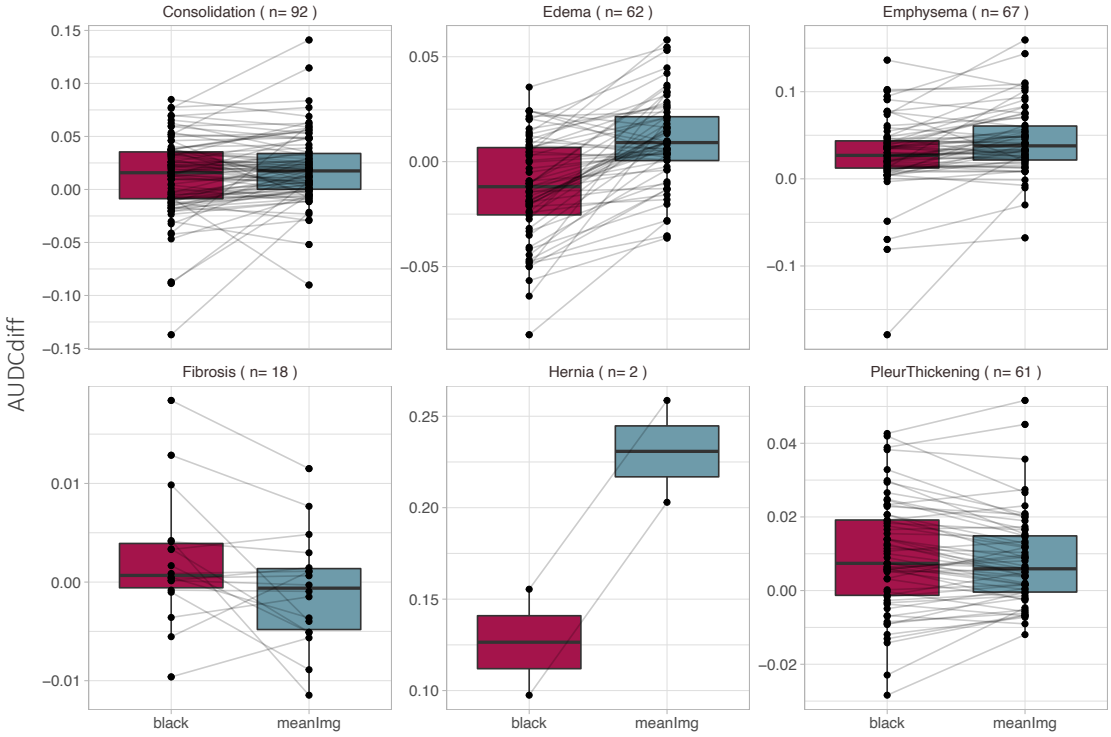

Fig. S5. Comparison between masks used for the deletion curves (for the rest of diseases). The difference between the area under the curve of deletion curves produced with (1) a descending order of the pixels by importance and (2) a random order of the pixels, are shown as the AUDCdiff. The AUDCdiff is shown for two masks that were used to generate the deletion curves, the black image and the mean image of all test images (meanImg). Paired lines are depicted to show the differences individually for each case. The analysis was performed for all true positive cases of the test set with the NIH dataset and the GradCAM explainer.

7.2 Supplementary tables

Table S1. Datasets summary. Original NIH ChestX-ray 14, CheXpert and MIMIC-CXR datasets information.

	NIH ChestX-ray14	**CheXpert**	**MIMIC-CXR**
Size	43 GB	450 GB	550 GB
#Images (#Patients)	112,120 (30,805)	223,648 (64,740)	371,547 (64,967)
Image size (average)	1024x1024	2282x2635	2485x2695
Image view	Frontal	Frontal and lateral	Frontal and lateral
#Labels (diseases + "no finding")	14+1	13+1	13+1
Demographics	Age and sex	Age and sex	Age and sex (also insurance and race)
Bounding box annotation	Yes	No	No

Table S2. Data splits and demographics statistics. Information about the data splits to be used by the classifiers. Statistics about population demographics are shown for the total number of images of each dataset. The splits corresponding to the alternative NIH ChestX-ray14 splits are written in parenthesis, and they correspond to the split where all the images with bounding boxes are included in the test set. The rest of the datasets follow the splits used by Seyyed et. al. [17]

	NIH ChestX-ray14 (alternative)	**CheXpert**	**MIMIC-CXR**
Training set (#images)	98,892 (83,367)	178,352	297,895
Test set (#images)	6,373 (15,938)	22,274	36,386
Validation set (#images)	6,855 (12,815)	23,022	37,266
Total (#images)	112,120	223,648	371,547
%male	56.49	59.36	52.16
%female	43.51	40.64	47.84
%0-20 years	6.28	0.87	1.38
%20-40 years	26.22	13.18	13.58
%40-60 years	43.92	31.01	32.34
%60-80 years	22.67	38.94	39.18
%80- years	0.91	16.01	10.86

Table S3. Reproduction of the original models. Comparison of our models' performance with the originally reported results [17], given by the average Area Under the Curve (AUC) of the model. The original AUC is reported with ±95% confidence interval for the 5 runs. Note our models have been trained for only one run, with a previous image resizing and with an upgraded Pytorch version. Alternative NIH corresponds to another dataset split of NIH, that was not used in the original paper.

	Original AUC	**Our AUC**
NIH	0.835 ± 0.002	0.835
Alternative NIH		0.783
CXP	0.805 ± 0.001	0.799
MIMIC	0.834 ± 0.001	0.830

Table S4. Occlusion hyperparameter choices. Summary of the experiments for optimizing occlusion, based on the patch size and the stride. These parameters affect the resolution of the attribution heatmap and its computing time. Note that the maximum heatmap resolution equals the image resolution, which is 256x256.

Window size	**Stride**	**Resolution**	**Computing time (s)**
32	32	8x8	1.72
32	16	16x16	4.53
16	16	16x16	5.92
16	8	32x32	18.43
8	8	32x32	19.30

7.3 Supplementary materials

All the code used for this project can be found on the following repository: https://github.com/gemmabb/FairMedImages.