Using Causal Analysis for Conceptual Deep Learning Explanation

Sumedha Singla[1], **Stephen Wallace**[2], **Sofia Triantafillou**[3], **Kayhan Batmanghelich**[3]

[1]Computer Science Department, University of Pittsburgh, USA

[2]University of Pittsburgh School of Medicine, University of Pittsburgh, USA

[3]Department of Biomedical Informatics, University of Pittsburgh, USA

Abstract

Model explainability is essential for the creation of trustworthy Machine Learning models in healthcare. An ideal explanation resembles the decision-making process of a domain expert and is expressed using concepts or terminology that is meaningful to the clinicians. To provide such explanation, we first associate the hidden units of the classifier to clinically relevant concepts. We take advantage of radiology reports accompanying the chest X-ray images to define concepts. We discover sparse associations between concepts and hidden units using a linear sparse logistic regression. To ensure that the identified units truly influence the classifier's outcome, we adopt tools from Causal Inference literature and, more specifically, mediation analysis through counterfactual interventions. Finally, we construct a low-depth decision tree to translate all the discovered concepts into a straightforward decision rule, expressed to the radiologist. We evaluated our approach on a large chest x-ray dataset, where our model produces a global explanation consistent with clinical knowledge.

1 Introduction

Machine Learning, specifically, Deep Learning (DL) methods are increasingly adopted in healthcare applications. Model explainability is essential to build trust in the AI system [5] and to receive clinicians' feedback. Standard explanation methods for image classification delineates regions in the input image that significantly contribute to the model's outcome [13,17,19]. However, it is challenging to explain *how* and *why* variations in identified regions are relevant to the model's decision. Ideally, an explanation should resemble the decision-making process of a domain expert. This paper aims to map a DL model's neuron activation patterns to the radiographic features and constructs a simple rule-based model that partially explains the Black-box.

Methods based on feature attribution have been commonly used for explaining DL models for medical imaging [1]. However, an alignment between feature attribution and radiology concepts is difficult to achieve, especially when a single region may correspond to several radiographic concepts. Recently, researchers have focused on providing explanations in the form of human-defined concepts [2,12,23]. In medical imaging, such methods have been adopted to derive an explanation for breast mammograms [22], breast histopathology [6] and cardiac MRIs [4]. A major drawback of the current approach is their dependence on explicit

concept-annotations, either in the form of a representative set of images [12] or semantic segmentation [2], to learn explanations. Such annotations are expensive to acquire, especially in the medical domain. We use weak annotations from radiology reports to derive concept annotations. Furthermore, these methods measure correlations between concept perturbations and classification predictions to quantify the concept's relevance. However, the neural network may not use the discovered concepts to arrive at its decision. We borrow tools from causal analysis literature to address that drawback [21].

In this work, we used radiographic features mentioned in radiology reports to define concepts. Using a National Language Processing (NLP) pipeline, we extract weak annotations from text and classify them based on their positive or negative mention [9]. Next, we use sparse logistic regression to identify sets of hidden-units correlated with the presence of a concept. To quantify the causal influence of the discovered concept-units on the model's outcome, we view concept-units as a *mediator* in the treatment-mediator-outcome framework [8]. Using measures from mediation analysis, we provide an effective ranking of the concepts based on their causal relevance to the model's outcome. Finally, we construct a low-depth decision tree to express discovered concepts in simple decision rules, providing the global explanation for the model. The rule-based nature of the decision tree resembles many decision-making procedures by clinicians.

2   Method

We consider a pre-trained *black-box* classifier $f: \mathbf{x} \rightarrow \mathbf{y}$ that takes an image $\mathbf{x}$ as input and process it using a sequence of hidden layers to produce a final output $y \in \mathbb{R}^D$. Without loss of generality, we decompose function f as $\Phi_2 \circ \Phi_1(\mathbf{x})$, where $\Phi_1(\mathbf{x}) \in \mathbb{R}^L$ is the output of the initial few layers of the network and Φ_2 denotes the rest of the network. We assume access to a dataset $\mathcal{X} = \left\{ \left(\mathbf{x}_n, \mathbf{y}_n, \mathbf{c}_n \right) \right\}^N$, where $\mathbf{x}_n$ is input image, $\mathbf{y}_n$ is a d-dimensional one-hot encoding of the class labels and $\mathbf{c}_n \in \mathbb{R}^K$ is a k-dimensional concept-label vector. We define concepts as the radiographic observations mentioned in radiology reports to describe and provide reasoning for a diagnosis. We used a NLP pipeline [9] to extract concept annotations. The NLP pipeline follows a rule-based approach to extract and classify observations from the free-text radiology report. The extracted k^{th} concept-label $\mathbf{c}_n[k]$ is either 0 (negative-mention), 1(positive-mention) or −1 (uncertain or missing-mention). An overview of our method is shown in Fig. 1. Our method consists of three sequential steps:

1. *Concept associations*: We seek to discover sparse associations between concepts and the hidden-units of $f(\cdot)$. We express k^{th} concept as a sparse vector $\beta_k \in \mathbb{R}^L$ that represents a linear direction in the intermediate space $\Phi_1(\cdot)$.

2. *Causal concept ranking*: Using tools from causal inference, we find an effective ranking of the concepts based on their relevance to the classification decision. Specifically, we consider each concept as a mediator in the causal path between the input and the outcome. We measure concept relevance as the effect of a counterfactual intervention on the outcome that passes indirectly through the concept-mediator.

3. *Surrogate explanation function*: We learn an easy-to-interpret function $g(\cdot)$ that mimics function $f(\cdot)$ in its decision. Using $g(\cdot)$, we seek to learn a global explanation for $f(\cdot)$ in terms of the concepts.

2.1 Concept associations

We discover concept associations with intermediate representation $\Phi_1(\cdot)$ by learning a binary classifier that maps $\Phi_1(\mathbf{x})$ to the concept-labels [12]. We treat each concept as a separate binary classification problem and extract a representative set of images $\mathcal{x}^k$, in which concept $c_n[k]$ is present and a random negative set. We define concept vector (β_k) as the solution to the logistic regression model $c_n[k] = \sigma\left(\beta_k^T \, \text{vec}\left(\Phi_1(\mathbf{x}_n)\right)\right) + \epsilon$, where $\sigma(\cdot)$ is the sigmoid function. For a convolutional neural network, $\Phi_1(\mathbf{x}) \in \mathbb{R}^{w \times h \times l}$ is the output activation of a convolutional layer with width w, height h and number of channels l. We experimented with two vectorization for Φ_1. In first, we flatten $\Phi_1(\mathbf{x})$ to be a whl-dimensional vector. In second, we applied a spatial aggregation by max-pooling along the width and height to obtain l-dimensional vector. Unlike TCAV [12] that uses linear regression, we used lasso regression to enable sparse feature selection and minimize the following loss function,

$$\min_{\beta_k} \sum_{\mathbf{x}_n \in \mathcal{x}_k} \ell\left(h_{\beta_k}(\mathbf{x}), c_n[k]\right) + \lambda \|\beta_k\|_1 \qquad (1)$$

where $\ell(\cdot, \cdot)$ is the cross entropy loss, $h_{\beta_k}(\mathbf{x}) = \sigma\left(\beta_k^T \, \text{vec}\left(\Phi_1(\mathbf{x}_n)\right)\right)$ and λ is the regularization parameter. We performed 10-fold nested-cross validation to find λ with least error. The non-zero elements in the concept vector β_k forms the set of hidden units $(\mathcal{V}_k)$ that are most relevant to the kth concept.

2.2 Causal concept ranking

Concept associations identified hidden units that are strongly correlated with a concept. However, the neural network may or may not use the discovered concepts to arrive at its decision. We use tools from causal inference, to quantify what fraction of the outcome is mediated through the discovered concepts.

To enable causal inference, we first define *counterfactual* $\mathbf{x}'$ as a perturbation of the input image $\mathbf{x}$ such that the decision of the classifier is flipped. Following the approach proposed in [20], we used a conditional generative adversarial network (cGAN) to learn the counterfactual perturbation. We conditioned on the output of the classifier, to ensure that cGAN learns a classifier-specific perturbation for the given image $\mathbf{x}$. Next, we used theory from causal mediation analysis to causally relate a concept with the classification outcome. Specifically, we consider concept as a mediator in the causal pathway from the input $\mathbf{x}$ to the outcome $\mathbf{y}$. We specify following effects to quantify the causal effect of the counterfactual perturbation and the role of a mediator in transferring such effect,

1. Average treatment effect (ATE): ATE is the total change in the classification outcome $\mathbf{y}$ as a result of the counterfactual perturbation.

2. Direct effect (DE): DE is the effect of the counterfactual perturbation that comprises of any causal mechanism that *do not* pass through a given mediator. It captures how the perturbation of input image changes classification decision directly, without considering a given concept.

3. Indirect effect (IE): IE is the effect of the counterfactual perturbation which is mediated by a set of mediators. It captures how the perturbation of input image changes classification decision indirectly through a given concept.

Following the potential outcome framework from [18,21], we define the ATE as the proportional difference between the factual and the counterfactual classification outcome,

$$\mathbf{ATE} = \mathscr{E}[\frac{f(\mathbf{x}')}{f(\mathbf{x})} - 1].$$
(2)

To enable causal inference through a mediator, we borrow Pearl's definitions of natural direct and indirect effects [16] (*ref* Fig. 2). We consider set of concept-units $\mathscr{V}_k$ as a mediator, representing the kth concept. We decompose the latent representation $\Phi_1(\mathbf{x})$ as concatenation of response of concept-units $\mathscr{V}_k(\mathbf{x})$ and rest of the hidden units $\bar{\mathscr{V}}_k(\mathbf{x})$ *i.e.*, $\Phi_1(\mathbf{x}) = [\mathscr{V}_k(\mathbf{x}), \bar{\mathscr{V}}_k(\mathbf{x})]$. We can re-write classification outcome as $f(\mathbf{x}) = \Phi_2(\Phi_1(\mathbf{x})) = \Phi_2([\mathscr{V}_k(\mathbf{x}), \bar{\mathscr{V}}_k(\mathbf{x})])$. To disentangle the direct effect from the indirect effect, we use the concept of *do*-operation on the unit level of the learnt network. Specifically, we use $do(\mathscr{V}_k(\mathbf{x}))$ to denote that we set the value of the concept-units to the value obtained by using the original image as input. By intervening on the network and setting the value of the concept units, we can compute the direct effect as the proportional difference between the factual and the counterfactual classification outcome, while holding mediator *i.e.*, $\mathscr{V}_k$ fixed to its value before the perturbation,

$$\mathbf{DE} = \mathscr{E}[\frac{\Phi_2([do(\mathscr{V}_k(\mathbf{x})), \bar{\mathscr{V}}_k(\mathbf{x}')])}{\Phi_2([\mathscr{V}_k(\mathbf{x}), \bar{\mathscr{V}}_k(\mathbf{x})])} - 1].$$
(3)

We compute indirect effect as the expected change in the outcome, if we change the mediator from its original value to its value using counterfactual, while holding everything else fixed to its original value,

$$\mathbf{IE} = \mathscr{E}[\frac{\Phi_2([do(\mathscr{V}_k(\mathbf{x}')), \bar{\mathscr{V}}_k(\mathbf{x})])}{\Phi_2([\mathscr{V}_k(\mathbf{x}), \bar{\mathscr{V}}_k(\mathbf{x})])} - 1].$$
(4)

If the perturbation has no effect on the mediator, then the causal indirect effect will be zero. Finally, we use the indirect effect associated with a concept, as a measure of its relevance to the classification decision.

2.3 Surrogate explanation function

We aim to learn a surrogate function $g(\cdot)$, such that it reproduces the outcome of the function $f(\cdot)$ using an interpretable and straightforward function. We formulated $g(\cdot)$ as a decision tree as many clinical decision-making procedures follow a rule-based pattern. We summarize the internal state of the function $f(\cdot)$ using output of k concept regression functions $h_{\beta_k}(\cdot)$ as follows,

$$\mathbf{w}_n = \left[, \ \text{logit}\left(h_{\beta_1}(\mathbf{x}_n)\right), \text{logit}\left(h_{\beta_2}(\mathbf{x}_n)\right), \cdots \right]. \tag{5}$$

Next, we fit a decision tree function, $g(\cdot)$, to mimic the outcome of the function $f(\cdot)$ as,

$$g^* = \underset{g}{\text{argmin}} \sum_n \mathscr{L}\big(g(\mathbf{w}_n), f(\mathbf{x}_n)\big), \tag{6}$$

where $\mathscr{L}$ is the splitting criterion based on minimizing entropy for highest information gain from every split.

3 Experiments

We first evaluated the concept classification performance and visualized concept-units to demonstrate their effectiveness in localizing a concept. Next, we summarized the indirect effects associated with different concepts across different layers of the classifier. We evaluated a proposing ranking of the concepts based on their causal contribution to the classification decision. Finally, we used the top-ranked concepts to learn a surrogate explanation function in the form of a decision tree. *Data preprocessing:* We perform experiments on the MIMIC-CXR [10] dataset, which is a multi-modal dataset consisting of 473K chest X-ray images and 206K reports. The dataset is labeled for 14 radiographic observations, including 12 pathologies. We used state-of-the-art DenseNet-121 [7] architecture for our classification function [9]. DenseNet-121 architecture is composed of four dense blocks. We experimented with three versions of $\Phi_1(\cdot)$ to represent the network until the second, third, and fourth dense block. For concept annotations, we considered radiographic features that are frequently mentioned in radiology reports in the context of labeled pathologies. Next, we used Stanford CheXpert [9] to extract and classify these observations from free-text radiology reports.

3.1 Evaluation of concept classifiers

The intermediate representations from third dense-block consistently outperformed other layers in concept classification. In Fig. 3, we show the testing-ROC-AUC and recall metric for different concept classifiers. All the concept classifiers achieved high recall, demonstrating a low false-negative (type-2) error.

In Fig. 4, we visualize the activation map of hidden units associated with the concept vector $\mathscr{V}_k$. For each concept, we visualize hidden units that have large logistic regression-coefficient (β_k). To highlight the most activated region for a unit, we threshold activation

map by the top 1% quantile of the distribution of the selected units' activations [2]. Consistent with prior work [3], we observed that several hidden units have emerged as concept detectors, even though concept labels were not used while training *f*. For *cardiac-silhouette*, different hidden units highlight different regions of the heart and its boundary with the lung. For localized concept such as *blunt costophrenic angle*, multiple relevant units were identified that all focused on the lower-lobe regions. Same hidden unit can be relevant for multiple concepts. The top label in Fig. 4. shows the top two important concepts for each hidden unit.

3.2   Evaluating causal concepts using explanation function

We evaluate the success of the counterfactual intervention by measuring ATE. High values for ATE confirms that counterfactual image generated by [20] successfully flips the classification decision. We achieved an ATE of 0.97 for cardiomegaly, 0.89 for pleural effusion and 0.96 for edema. In Fig. 1 (heat-map), we show the distribution of the indirect effect associated with concepts, across different layers. The middle layer demonstrates a large indirect effect across all concepts. This shows that the hidden units in dense-block 3 played a significant role in mediating the effect of counterfactual intervention.

In Fig. 5 (bar-graph), we rank the concepts based on their indirect effect. The top-ranked concepts recovered by our ranking are consistent with the radiographic features that clinicians associates with the examined three diagnoses [11,14,15]. Further, we used the concept sensitivity score from TCAV [12] to rank concepts for each diagnosis. The top-10 concepts identified by our indirect effect and TCAV are the same, while their order is different. The top-3 concepts are also the same, with minor differences in ranking. Both the methods have low importance score for random concept. This confirms that the trend in importance score is unlikely to be caused by chance. For our approach, random concept represents an ablation of the concept-association step. Here, rather than performing lasso regression to identify relevant units, we randomly select units.

To quantitatively demonstrate the effectiveness of our ranking, we iteratively consider x% of top-ranked concepts and retrain the explanation function $g(\mathbf{w})$. In Fig. 5 (bottom-plot), we observe the change in recall metric for the classifier $g(\cdot)$ as we consider more concepts. In the beginning, as we add relevant concepts, the true positive rate increases resulting in a high recall. However, as less relevant concepts are considered, the noise in input features increased, resulting in a lower recall. Fig. 6 visualize the decision tree learned for the best performing model.

4   Conclusion

We proposed a novel framework to derive global explanation for a black-box model. Our explanation is grounded in terms of clinically relevant concepts that are causally influencing the model's decision. As a future direction, we plan to extend our definition of concepts to include a broader set of clinical metrics.

Acknowledgement

References

1. Basu S, Mitra S, Saha N: Deep learning for screening covid-19 using chest x-ray images In: IEEE Symposium Series on Computational Intelligence (SSCI) (2020)

2. Bau D, Zhou B, Khosla A, Oliva A, Torralba A: Network dissection: Quantifying interpretability of deep visual representations. In: IEEE Computer Vision and Pattern Recognition (CVPR) pp. 6541–6549 (2017)

3. Bau D, Zhu JY, Strobelt H, Lapedriza A, Zhou B, Torralba A: Understanding the role of individual units in a deep neural network. National Academy of Sciences 117(48), 30071–30078 (2020)

4. Clough JR, Oksuz I, Puyol-Antón E, Ruijsink B, King AP, Schnabel JA: Global and local interpretability for cardiac mri classification In: Medical Image Computing and Computer-Assisted Intervention (MICCAI). pp. 656–664 (2019)

5. Glass A, McGuinness DL, Wolverton M: Toward establishing trust in adaptive agents In: International Conference on Intelligent User Interfaces (2008)

6. Graziani M, Andrearczyk V, Marchand-Maillet S, Müller H: Concept attribution: Explaining cnn decisions to physicians. Computers in Biology and Medicine 123, 103865 (2020) [PubMed: 32658785]

7. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ: Densely connected convolutional networks In: IEEE Computer Vision and Pattern Recognition (CVPR). pp. 4700–4708 (2017)

8. Imai K, Jo B, Stuart EA: Commentary: Using potential outcomes to understand causal mediation analysis. Multivariate Behavioral Research 46(5) (2011)

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

10. Johnson AE, Pollard TJ, Berkowitz SJ, Greenbaum NR, Lungren MP, Deng CY, Mark RG, Horng S: Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. Scientific data 6(1) (2019)

11. Karkhanis VS, Joshi JM: Pleural effusion: diagnosis, treatment, and management. Open Access Emergency Medicine (OAEM) 4, 31 (2012) [PubMed: 27147861]

12. Kim B, Wattenberg M, Gilmer J, Cai C, Wexler J, Viegas F, et al.: Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav) In: International Conference on Machine Learning (ICML). pp. 2668–2677 (2018)

13. Lundberg SM, Lee SI: A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems 30, 4765–4774 (2017)

14. Milne E, Pistolesi M, Miniati M, Giuntini C: The radiologic distinction of cardiogenic and noncardiogenic edema. American Journal of Roentgenology 144(5), 879–894 (1985) [PubMed: 3872571]

15. Nakamori N, MacMahon H, Sasaki Y, Montner S, et al. : Effect of heart-size parameters computed from digital chest radiographs on detection of cardiomegaly. potential usefulness for computer-aided diagnosis. Investigative radiology 26(6), 546–550 (1991) [PubMed: 1830564]

16. Pearl J: Direct and indirect effects In: Conference on Uncertainty and Artificial Intelligence (UAI). pp. 411–420 (2001)

17. Ribeiro MT, Singh S, Guestrin C: "why should i trust you?" explaining the predictions of any classifier In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1135–1144 (2016)

18. Rubin DB: Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology 66(5), 688 (1974)

19. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D: Grad-cam: Visual explanations from deep networks via gradient-based localization In: International Conference on Computer Vision (ICCV). pp. 618–626 (2017)

20. Singla S, Pollack B, Chen J, Batmanghelich K: Explanation by progressive exaggeration In: International Conference on Learning Representations (ICLR) (2019)

21. Vig J, Gehrmann S, Belinkov Y, Qian S, Nevo D, Singer Y, Shieber S: Investigating gender bias in language models using causal mediation analysis. In: Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H (eds.) Advances in Neural Information Processing Systems vol. 33, pp. 12388–12401 (2020)

22. Yeche H, Harrison J, Berthier T: Ubs: A dimension-agnostic metric for concept vector interpretability applied to radiomics. In: Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support (IMI-MICML-CDS), pp. 12–20. Springer (2019)

23. Zhou B, Sun Y, Bau D, Torralba A: Interpretable basis decomposition for visual explanation In: European Conference on Computer Vision (ECCV). pp. 119–134 (2018)
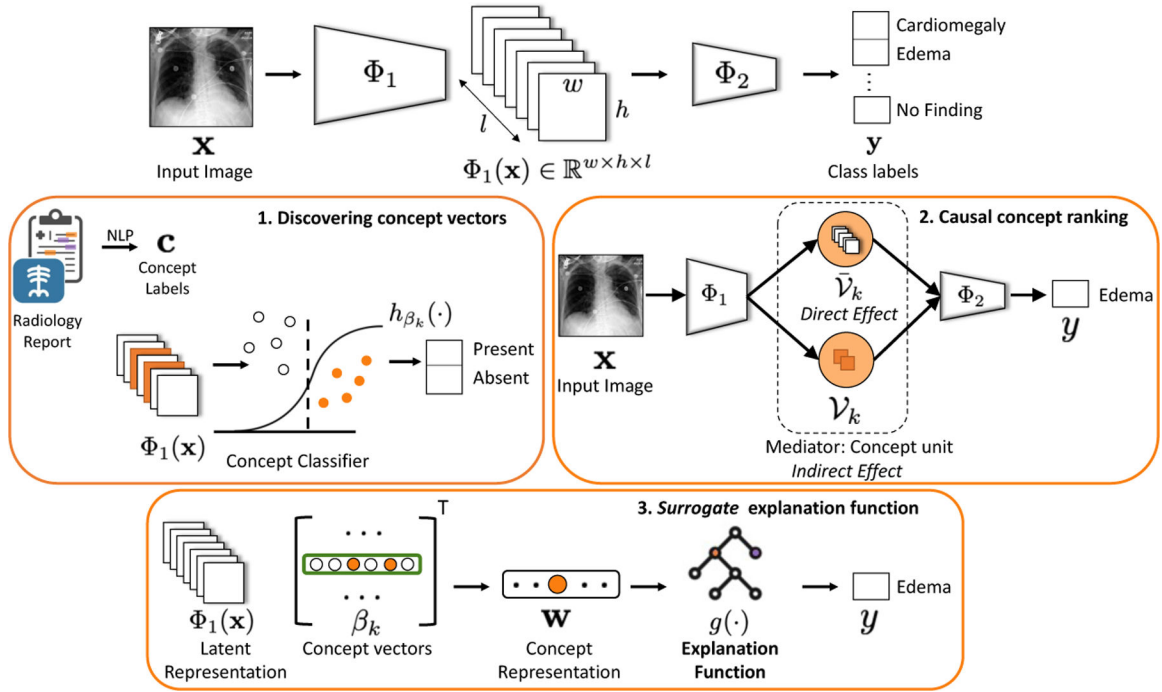
Fig.1.

Method overview: We provide explanation for the black-box function $f(\mathbf{x})$ interms of concepts, that are radiographic observations mentioned in radiology reports. 1) The intermediate representation $\Phi_1(\mathbf{x})$ is used to learn a sparse logistic regression $h_{\beta_k}(\cdot)$ to classify k^{th} concept. 2) The non-zero coefficients of β_k represents a set of concept units $\mathcal{V}_k$ that serves as a mediator in the causal path connecting input $\mathbf{x}$ and outcome y. 3) A decision tree function is learned to map concepts to class labels.
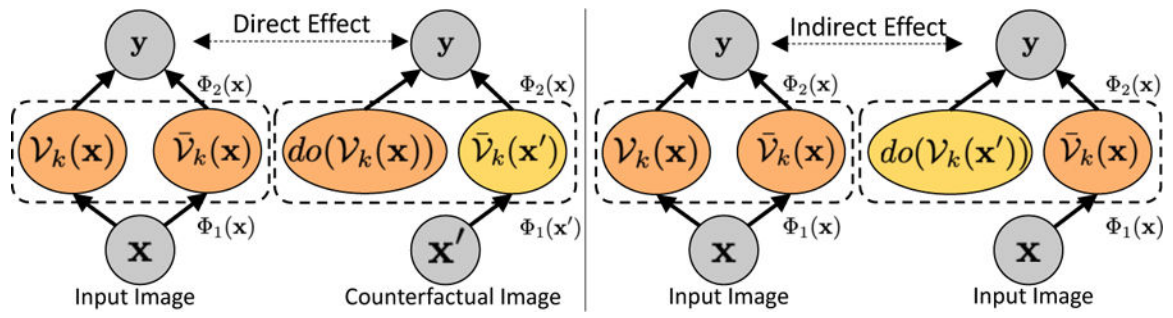
Fig.2.
Illustration of direct and indirect effects in causal mediation analysis.
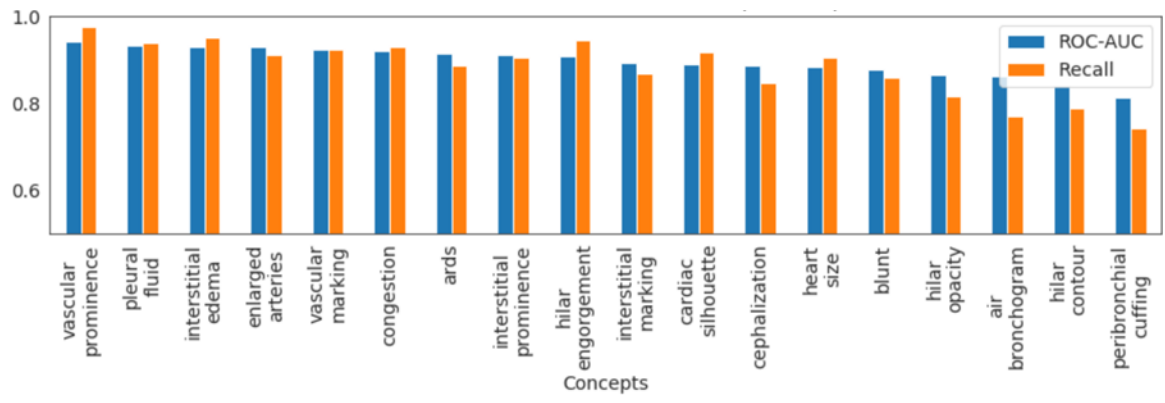
Fig.3.
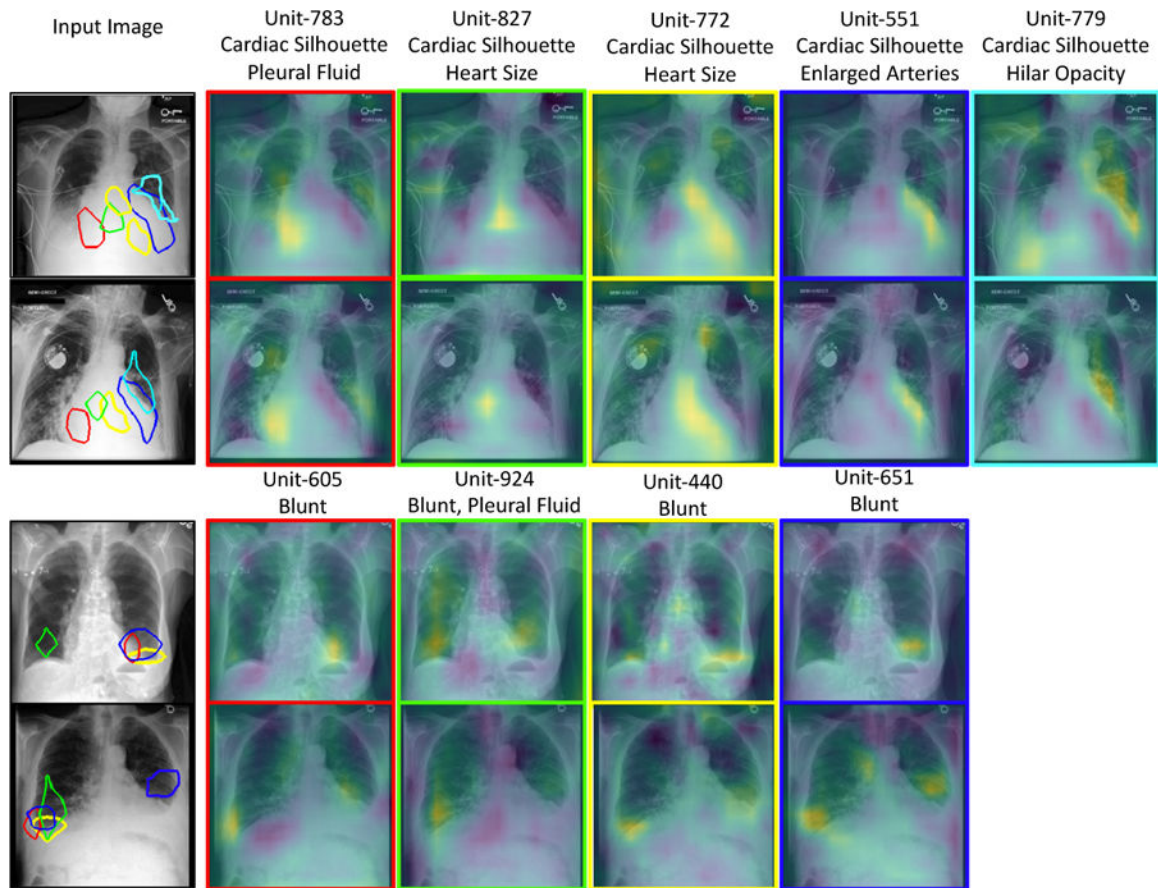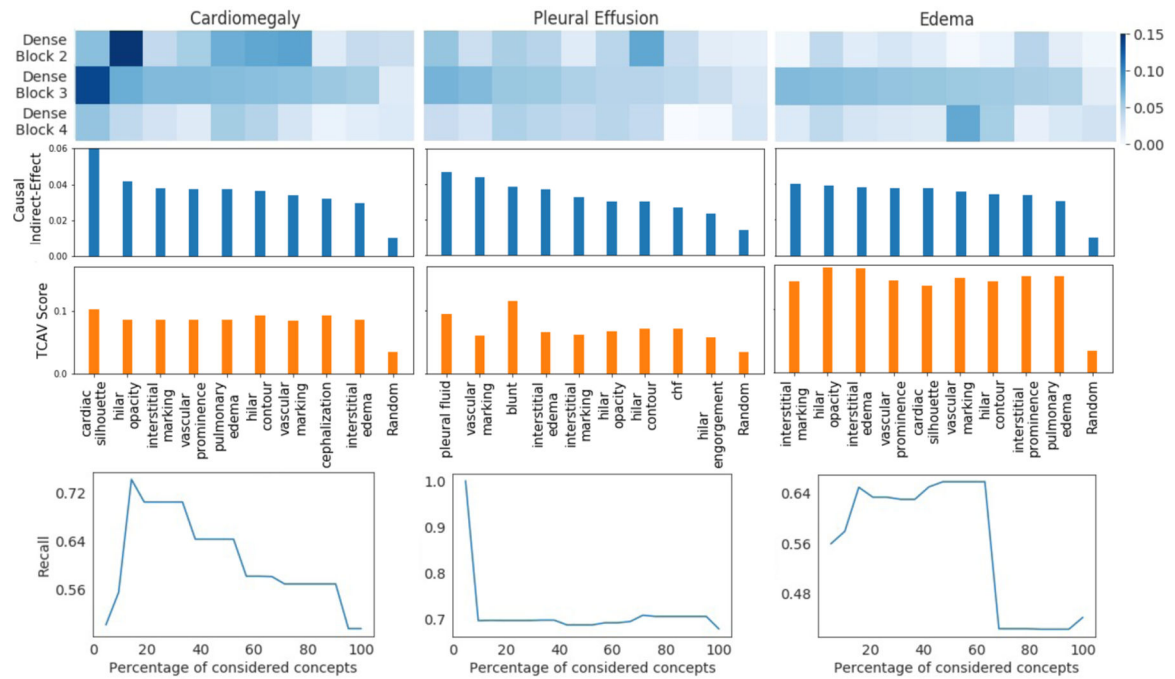AUC-ROC and recall metric for different concept classifiers.

Fig.4.
A qualitative demonstration of the activation maps of the hidden units that act as visual concept detectors. Each column represents one hidden unit identified as part of concept vector $\mathcal{V}_k$. Top two rows show $k = $ *cardiac-silhouette* and bottom rows have $k = $ *blunt costophrenic angle*.

Fig.5.
Indirect effects of the concepts, calculated over different layers of the DenseNet-121 architecture (heat-map). The derived ranking of the concepts based on their causal relevance to the diagnosis (bar-graph). A comparative ranking based on concept sensitivity score from TCAV [12]. The trend of recall metric for the decision tree function $g(\cdot)$, while training using top x% of top-ranked concepts (trend-plot).
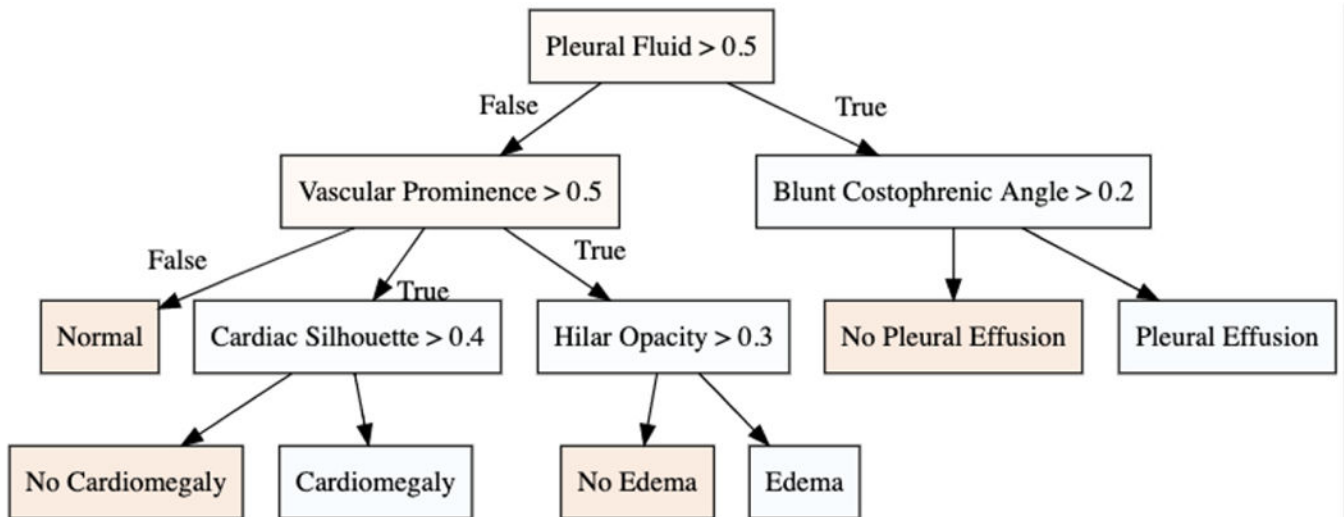
Fig.6.
The decision tree for the three diagnosis with best performance on recall metric.