Inherently Interpretable Multi-Label Classification Using Class-Specific Counterfactuals

Author name(s) withheld EMAIL(S) WITHHELD
Address withheld

Editors: Under Review for MIDL 2023

Abstract

Interpretability is essential for machine learning algorithms in high-stakes application fields such as medical image analysis. However, high-performing black-box neural networks do not provide explanations for their predictions, which can lead to mistrust and suboptimal human-ML collaboration. Post-hoc explanation techniques, which are widely used in practice, have been shown to suffer from severe conceptual problems. Furthermore, as we show in this paper, current explanation techniques do not perform adequately in the multi-label scenario, in which multiple medical findings may co-occur in a single image. We propose Attri-Net[1], an inherently interpretable model for multi-label classification. Attri-Net is a powerful classifier that provides transparent, trustworthy, and human-understandable explanations. The model first generates class-specific attribution maps based on counterfactuals to identify which image regions correspond to certain medical findings. Then a simple logistic regression classifier is used to make predictions based solely on these attribution maps. We compare Attri-Net to five post-hoc explanation techniques and one inherently interpretable classifier on three chest X-ray datasets. We find that Attri-Net produces high-quality multi-label explanations consistent with clinical knowledge and has comparable classification performance to state-of-the-art classification models.

1. Introduction

The clinical adoption of machine learning (ML) technology is hindered by the black-box nature of deep learning models. Their inscrutability may lead to a lack of trust (Dietvorst et al., 2015), or blind trust among clinicians (Tschandl et al., 2020; Gaube et al., 2021), and may result in ethical as well as legal problems (Grote and Berens, 2020). Therefore, transparency has been identified as one of the key properties for deploying machine learning technology in high-stakes application areas such as medicine (Rudin, 2019).

The most commonly used category of techniques for understanding the decision mechanisms of ML models are *post-hoc* methods which apply a heuristic to a trained model trying to understand the decision mechanism retrospectively after the prediction is made. Gradient-based techniques such as Guided Backpropagation (Springenberg et al., 2014) perform local function approximation of the black-box model by differentiating the prediction with respect to the input pixels. The faithfulness of such methods to the decision mechanisms has recently been put into question by Adebayo et al. (2018) and Arun et al. (2021) who showed that explanations remain unchanged despite randomisation of network weights. Perturbation-based methods such as LIME (Ribeiro et al., 2016), or SHAP (Lundberg and Lee, 2017) also approximate the local decision function. These methods cannot currently

1. The code for Attri-Net is available at https://github.com/anonymous/Attri-Net

© 2023 Anonymous.

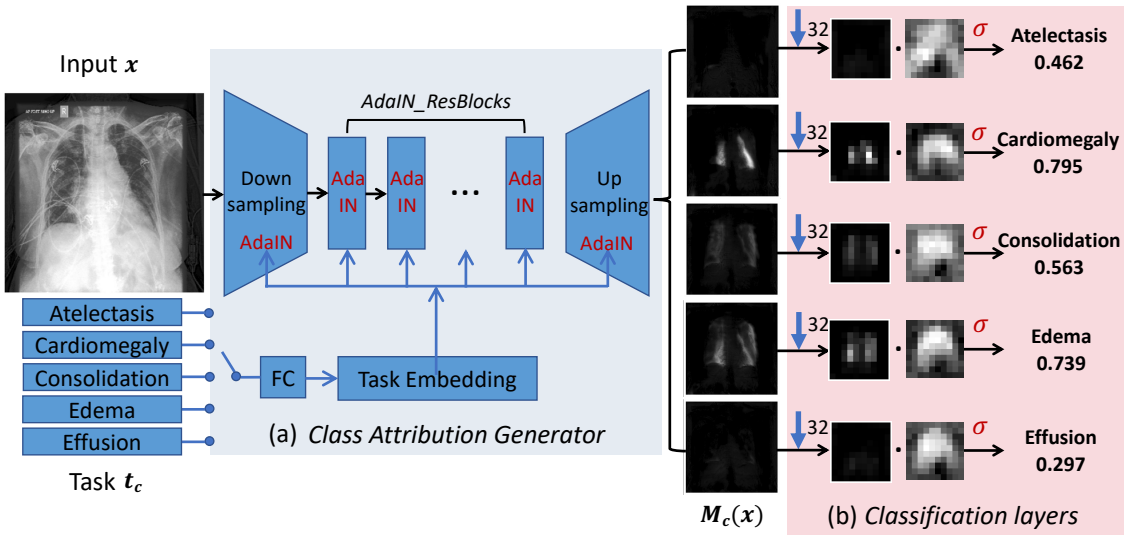

Figure 1: Attri-Net Framework. Given an input image and a task, our visual feature attribution generator (a) produces counterfactual attribution maps. Based on these maps a logistic regression classifier (b) produces the final prediction for each class.

produce explanations at the pixel level and are computationally demanding. Another line of work including Class Activation Mappings (CAM) (Zhou et al., 2016) and GradCAM (Selvaraju et al., 2017) attempts to construct neural network architectures from which the decision mechanism can be directly inferred. However, these techniques are limited by the spatial resolution of their explanations and do not explain the reasoning mechanism on a pixel-level. BagNet (Brendel and Bethge, 2019) addresses this issue by severely restricting the global receptive field of the network which can negatively affect classification performance. Additionally, explanations are partially generated by image patches that are not used for making the prediction. This means that, again, the explanations do not directly correspond to the decision mechanism. Another class of methods obtain post-hoc explanations by generating hypothetical counterfactual images of an opposite class (Schutte et al., 2021; Joshi et al., 2018), or by exaggerating features of the predicted class (Cohen et al., 2021; Singla et al., 2019). A related line of post-hoc methods derives classifications from an intermediate representation of the counterfactual generator (Bass et al., 2020; Cetin et al., 2022). We also note that some techniques attempt to generate counterfactuals without the aim to explain a classifier (Baumgartner et al., 2018; Nemirovsky et al., 2020).

While explanations of post-hoc techniques may appear reasonable there is no guarantee that they explain what the classifier actually does, and there is, in fact, growing evidence that they are not faithful to the actual decision mechanism (Adebayo et al., 2018; Han et al., 2022; White and Garcez, 2019). In contrast, *inherently interpretable* methods use prediction systems for which the decision mechanism is directly revealed to the user. These models are by definition faithful to the decision mechanism because the explanation *is* the decision mechanism. Prior work includes methods in which the final predictions are directly based on human-interpretable concepts (Alvarez Melis and Jaakkola, 2018; Chen et al., 2020; Koh et al., 2020), prototypical representations of classes (Chen et al., 2019; Barnett et al.,

2021), or direct attribution to image patches (Javed et al., 2022). The recently proposed Convolutional Dynamic Alignment Networks (CoDA-Nets) (Bohle et al., 2021) is, to our knowledge, the only existing model providing inherently interpretable visual explanations on a pixel-level. The method expresses network weights as a function of the input image in a way that allows them to formulate the networks decision for a specific input image as a linear classifier. We note that there are to our knowledge no inherently interpretable methods based on counterfactual explanations.

The majority of visual explanation techniques were developed for binary or multi-class problems. Many clinical tasks, however, are multi-label problems, where multiple classes can apply simultaneously. For example, in chest X-ray diagnosis, which we study in this work, an image often contains multiple findings. In our experiments, we found that existing visual explanation techniques are not well-suited to this important type of decision problems. In particular, the explanations are not specific to the class and they tend to highlight similar regions for all classes, in some cases even when the class is not present in the image.

In this paper, we propose Attri-Net, an inherently interpretable visual explanation technique designed specifically for the multi-label scenario. Our model predicts class-specific counterfactual attribution maps as intermediate representations. The attribution maps, which are conceptually based on the visual feature attribution GAN (VA-GAN) approach introduced by Baumgartner et al. (2018), represent residual images that contain all existing evidence of a class in an input image. Attri-Net then uses these class-specific attribution maps as input features in a final linear classification layer. We evaluate Attri-Net on three widely used chest X-ray datasets and demonstrate that the method produces high-quality inherently interpretable explanations with a high class sensitivity while retaining classification performance comparable to state-of-the-art models.

2. Methods

In this paper, we address the multi-label classification scenario with C classes, where each class c with label $y_c \in \{0, 1\}$ can independently occur in an image, i.e. multiple co-existing medical findings are possible. In the following, we first introduce our method for generating counterfactual class attribution maps for each class c (see Fig. 1a). Then, we show how a logistic regression classifier is used to obtain the final predictions based on those attribution maps (see Fig. 1b). Lastly, we explain how these two components are trained end-to-end in our proposed Attri-Net framework.

2.1. Counterfactual class attribution map generation

The core of our method is an image-to-image network $M_c(\mathbf{x}) : \mathbb{R}^{h \times w} \mapsto \mathbb{R}^{h \times w}$ which generates residual counterfactual class attribution maps for an input image $\mathbf{x}$. Intuitively, the output of M_c represents how each pixel in the input should change in order to remove the effect of class c from the image. Like Baumgartner et al. (2018), we learn an additive mapping M_c that makes the output image appear to come from the opposite class, that is

$$\hat{\mathbf{x}} = \mathbf{x} + M_c(\mathbf{x}) \,,$$

such that the generated counterfactual image $\hat{\mathbf{x}}$ is indistinguishable from images sampled from the distribution $p(\mathbf{x}|y_c = 0)$ of real images *not* containing class c. To ensure the

correct behaviour of M_c, we simultaneously train a class-specific discriminator network D_c to distinguish between real and fake images with $y_c = 0$. Specifically, we use the Wasserstein GAN loss (Arjovsky et al., 2017; Baumgartner et al., 2018). Details on the optimisation of D_c are given in Appendix B.1. Given a discriminator function D_c we can write the following adversarial loss term ensuring that $\hat{\mathbf{x}}$ is a realistic counterfactual not containing class c and, by extension, that M_c outputs realistic residual class attribution maps:

$$\mathcal{L}_{\text{adv}}^{(c)} = \mathop{\mathbb{E}}_{\mathbf{x}\sim p(\mathbf{x}|y_c=1)} \left[-D_c(\mathbf{x} + M_c(\mathbf{x})) \right]. \tag{1}$$

Some examples of generated counterfactuals $\hat{\mathbf{x}}$ are shown in Appendix A.1.

To discourage the network from attributing superfluous pixels not belonging to a given class, we additionally encourage the class attribution maps to be sparse using an L_1 regularization term similar to Baumgartner et al. (2018). To further encourage the generator to produce smaller effects when the class is present in an image than when it is not present, we divide the loss into two differently weighted terms with a larger weight α_0 for class-negative, and a smaller weight α_1 class-positive examples, i.e.,

$$\mathcal{L}_{\text{reg}}^{(c)} = \alpha_0 \mathop{\mathbb{E}}_{\mathbf{x}\sim p(\mathbf{x}|y_c=0)} \left[\|M_c(\mathbf{x})\|_1 \right] + \alpha_1 \mathop{\mathbb{E}}_{\mathbf{x}\sim p(\mathbf{x}|y_c=1)} \left[\|M_c(\mathbf{x})\|_1 \right]. \tag{2}$$

We use $\alpha_0 = 2, \alpha_1 = 1$ for all experiments in this paper.

The functions M_c, and D_c are implemented as neural networks building on the StarGAN architecture (Choi et al., 2018) which produced superior results to alternative options we explored such as the original VA-GAN architecture. Although it is feasible to design a network M to produce class attribution maps for all labels as multiple output channels in a single forward pass, preliminary experiments revealed inadequate class attribution in the multi-label scenario. Instead, we build on the recently proposed task switching network (Sun et al., 2021) where adaptive instance norm (AdaIN) layers are used to switch between related tasks. In our work, tasks correspond to the generation of attribution maps for different classes. Each task is represented as a task vector $\mathbf{t}_c$ which is a one-hot encoding upsampled by a factor of 20 as in (Sun et al., 2021). This encoding is then converted into a task embedding via a small fully connected network and fed to AdaIN layers which are placed throughout the network (as shown in Fig. 1a). The AdaIN layers then toggle the behaviour of the network. To combine this paradigm with the StarGAN architecture, we replaced all instance normalization layers of the original generator and discriminator networks with AdaIN layers. The architecture is described in greater detail in Appendix B.2. The mask generator and discriminator can now be expressed as $M_c(\mathbf{x}) = M(\mathbf{x}, \mathbf{t}_c)$, and $D_c(\mathbf{x}) = D(\mathbf{x}, \mathbf{t}_c)$, respectively. The class attribution maps for all labels can be obtained by repeated forward passes through M while iterating through the $\mathbf{t}_c$ vectors of all classes.

2.2. Classification using a logistic regression classifier

Given a class-specific counterfactual attribution map obtained using $M(\mathbf{x}, \mathbf{t}_c)$, we want to predict the presence of class c in an image. To achieve this, the respective attribution map is downsampled and used as input to a logistic regression classifier. That is,

$$p(y_c|\mathbf{x}) = \sigma\Big(\sum_{i,j} w_{ij}^{(c)} \cdot S_\gamma(M(\mathbf{x}, \mathbf{t}_c))_{ij} \Big), \tag{3}$$

where S_γ is a 2D average pooling operator that downsamples by a factor of γ, $w_{ij}^{(c)}$ denotes the weights associated to each pixel of the down-sampled attribution map for class c, and σ is the sigmoid function. In preliminary experiments, we found $\gamma = 32$ to perform robustly and we use this value for all experiments.

The classifier is trained using a standard multi-label classification loss $\mathcal{L}_{\text{cls}}^{(c)}$, which is the sum of the binary cross entropy losses for each class. Note that, since our framework is trained end-to-end, M also receives gradients from that loss and is thereby encouraged to create class attribution maps that are linearly classifiable.

To further encourage the attribution maps to be discriminative for positive and negative examples of each class, we apply the center loss proposed by Wen et al. (2016), which has been shown to lead to more discriminative feature representations. Extending the idea, here, we define class centers $\mathbf{v}_{y_c=0}, \mathbf{v}_{y_c=1} \in \mathbb{R}^{h \times w}$ which are learnable and converge to prototypical representations of attribution maps corresponding to positive and negative instances of each class c. The center loss draws the class attribution maps closer to their respective class centers, resulting in a more clustered feature space where positive and negative samples are better linearly separable. The overall center loss can be written as

$$\mathcal{L}_{\text{ctr}}^{(c)} = \frac{1}{2} \left(\mathop{\mathbb{E}}_{\mathbf{x} \sim p(\mathbf{x}|y_c=0)} \left[\| M(\mathbf{x}, \mathbf{t}_c) - \mathbf{v}_{y_c=0} \|_2^2 \right] + \mathop{\mathbb{E}}_{\mathbf{x} \sim p(\mathbf{x}|y_c=1)} \left[\| M(\mathbf{x}, \mathbf{t}_c) - \mathbf{v}_{y_c=1} \|_2^2 \right] \right) . \quad (4)$$

The class center images are updated for each mini-batch in a separate gradient update interleaved with the updates of the network parameters as described by Wen et al. (2016). The final class center images as well as the logistic regression weights $w_{ij}^{(c)}$ may be used to further interpret the model's behaviour on a global level. Examples of both are shown in Appendix A.2. However, we leave the exploration of global interpretability to future work.

2.3. Training

Our Attri-Net framework can be trained end-to-end with four loss terms enforcing our essential requirements: Firstly, the attribution map should preserve sufficient class relevant information such that a satisfactory classification result can be obtained. Secondly, the attribution maps should be human-interpretable. The overall training objective for the class attribution generator M with weight parameters φ is given by

$$\min_{\varphi} \sum_c \lambda_{\text{cls}} \mathcal{L}_{\text{cls}}^{(c)} + \lambda_{\text{adv}} \mathcal{L}_{\text{adv}}^{(c)} + \lambda_{\text{reg}} \mathcal{L}_{\text{reg}}^{(c)} + \lambda_{\text{ctr}} \mathcal{L}_{\text{ctr}}^{(c)} , \quad (5)$$

where we use the hyperparameters λ_* to balance the losses. We chose $\lambda_{\text{cls}} = 100,\ \lambda_{\text{adv}} = 1,\ \lambda_{\text{reg}} = 100,\ \lambda_{\text{ctr}} = 0.01$ for our experiments. An ablation study on the effect of the different losses can be found in Appendix A.3. During training, we repeatedly iterate through the different classes c and, for each, sample two mini-batches, one containing positive examples of the current class and the other negative examples. Using these two batches we perform gradient updates for all losses before moving on to the next class. We use the ADAM optimizer (Kingma and Ba, 2014) with a learning rate of 10^{-4} and a batch size of 4 to optimize our model. Furthermore, following Wen et al. (2016), we use stochastic gradient descent for updating the center loss parameters. Training converges within 72 hours on an Nvidia V100

Table 1: Classification performance measured by area under the ROC curve (AUC).

Model	CheXpert	ChestX-ray8	VindrCXR
ResNet50 (Azizi et al., 2021)	0.7687	-	-
SimCLR (Azizi et al., 2021)	0.7702	-	-
LSE (Ye et al., 2020)	-	0.7554	-
ChestNet (Ye et al., 2020)	-	0.7896	-
ResNet50	0.7727	0.7445	0.8986
CoDA-Nets	0.7659	0.7727	0.9322
ours	0.7405	0.7762	0.9405

Table 2: Comparison of class sensitivity scores.

Model	CheXpert	ChestX-ray8	VindrCXR
ResNet + GB	0.3183	0.3028	0.1727
ResNet + GCam	0.1434	0.1570	0.1931
ResNet + LIME	0.2347	0.2609	0.2422
ResNet + SHAP	0.4745	0.4122	0.3714
ResNet + Gifsplan.	0.2748	0.5817	0.4396
CoDA-Nets	0.3576	0.4138	0.4464
ours	**0.4880**	**0.6160**	**0.5509**

GPU. After training we select the decision threshold which maximises the Youden-index (sensitivity + specificity - 1) for each class on the validation set. We also perform this step for the baseline methods.

3. Experiments and Results

Data. We evaluated our proposed Attri-Net on the three widely used chest X-ray datasets CheXpert (Irvin et al., 2019), ChestX-ray8 (Wang et al., 2017), and VinDrCXR (Nguyen et al., 2020). Following (Irvin et al., 2019) and (Azizi et al., 2021) for the CheXpert and ChestX-ray8 datasets we used the classes "Atelectasis", "Cardiomegaly", "Consolidation", "Edema", and "Pleural Effusion". For the VinDr-CXR dataset, we selected the five pathologies with the highest number of samples, which were "Aortic enlargement", "Cardiomegaly", "Pulmonary fibrosis", "Pleural thickening", and "Pleural effusion". We split all datasets into a training (80%), testing (10%) and validation (10%) fold. Since the test set of Chexpert was not publicly available and the official validation set was small, we adopted the method used in (Azizi et al., 2021) to split the train set into train, validation, and test sets.

Classification performance. To assess the classification performance, we compared our model with the state-of-the-art inherently interpretable model CoDA-Nets (Bohle et al., 2021) as well as a standard black-box ResNet50 model. We also report the results of Azizi et al. (2021) and Ye et al. (2020) on CheXpert and ChestX-ray8, respectively. Attri-Net overall performed comparable to the state-of-the-art (see Tab. 1), with an area under the ROC curve that was slightly lower on CheXpert, similar to other methods on ChestX-ray8, and slightly better on VindrCXR.

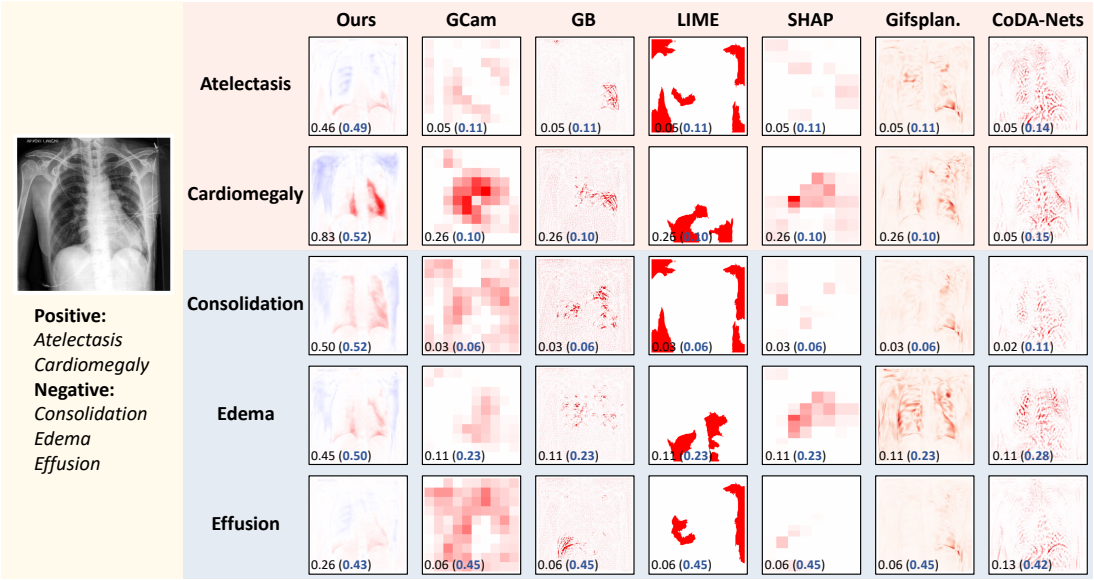

Figure 2: Visual comparison of explanations for an example image from the CheXpert dataset. Predicted class probabilities are indicated in the lower left corner of each attribution map with the respective decision threshold in parentheses.

Interpretability. Khakzar et al. (2021) argue that if different areas of an image are responsible for predicting different classes, then also the explanations should be different. They coin this property "class sensitivity". In the context of multi-label classification, the explanation for an image containing a class should have higher attribution than an image where the class is absent. We measured class sensitivity following Bohle et al. (2021) and created a series of 2×2 grids of explanations, where each grid contained only one positive example of a given class (see Appendix A.4 for example grids). We then represented class sensitivity by the sum of attributions in the positive example divided by the sum of all attributions in the grid. The optimal scenario where only the disease positive map contains any attributions, and disease negative attribution maps are blank, yields a sensitivity of 1. We computed the average sensitivity over 200 grids for each class c.

Our method led to a substantially and consistently higher class sensitivity than the inherently interpretable baseline, CoDA-Nets, across all datasets (see Tab. 2). For the black-box ResNet, we compared five post-hoc explanations techniques, i.e. Guided Backpropagation (Springenberg et al., 2014), GradCAM (Selvaraju et al., 2017), LIME (Ribeiro et al., 2016), SHAP (Lundberg and Lee, 2017) and the recently proposed Gifsplanation (Cohen et al., 2021). The post-hoc methods varied considerably with SHAP and Gifsplanation performing comparably to CoDA-Nets, but substantially worse than our Attri-Net.

Qualitative examination of example explanations supported these results. Our proposed Attri-Net produced class attribution maps that clearly highlight the parts of the underlying anatomy that support the respective classes (see Fig. 2 for a representative example from the CheXpert dataset). Moreover, the attributions for different classes were clearly distinct from each other, each one focusing on different anatomical areas. Examples from the ChestX-ray8 and VinDr-CXR datasets can be found in Appendix A.5. In contrast, the

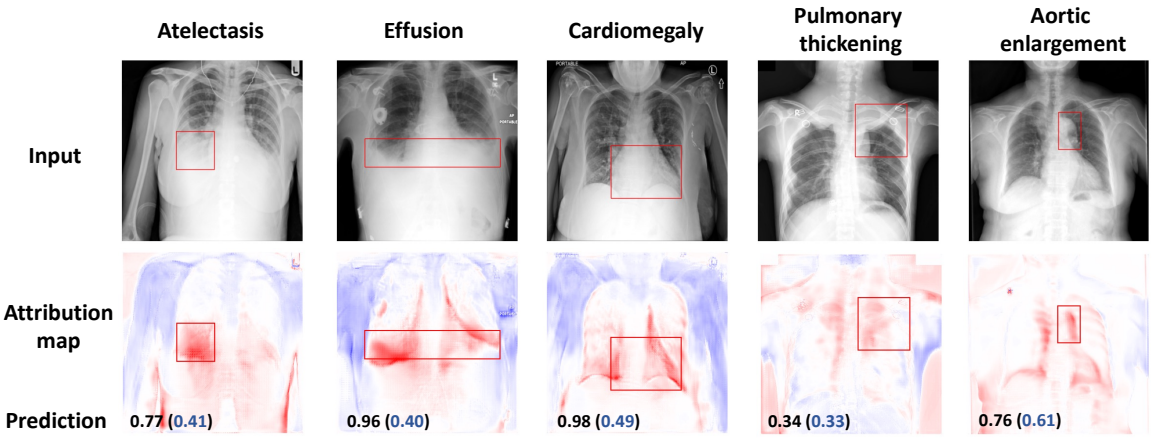

Figure 3: Attribution maps for samples from ChestX-ray8 and VindrCXR with GT bounding boxes. Decision thresholds are given in parentheses.

inherently interpretable baseline, CoDA-Nets, produced visually similar attributions for all classes (rightmost column in Fig. 2). We further observed that the baseline techniques were mostly not useful for identifying which parts of the anatomy contributed to a prediction. While Guided Backpropagation qualitatively provided the most useful explanations of the baselines, its attributions were very noisy as is typical for this technique. We further examined Attri-Net explanations on example images of each class where pathology bounding boxes were available (Fig. 3). Attri-Net generally highlighted regions associated with the respective pathologies, with particularly sensitive attribution maps when the final prediction was highly confident (i.e. the examples with atelectasis, effusion, and cardiomegaly). We also observed some relatively strong attributions in regions outside the bounding boxes. As our class attribution maps were based on counterfactuals that were designed to realistically remove all effects of a pathology, we hypothesise they may have uncovered additional effects correlated with the classes which were not part of the clinical grading protocol.

4. Discussion and Conclusion

We proposed Attri-Net, a novel inherently interpretable multi-label classifier and showed that it produces high-quality explanations substantially outperforming all baselines in terms of class sensitivity while retaining classification performance comparable to state-of-the-art black-box models. Explanations of the black-box model were highly dependent on the post-hoc technique, and fundamentally differed from each other even on the same image. This erodes trust in their capacity to provide necessary transparency in high-stakes applications and shows the need for inherently interpretable models such as ours, where the predictions are formed directly and linearly from visually interpretable class attribution maps.

The qualitative and quantitative assessments in this paper suggest that our method provides useful explanations, but there remain important avenues for future work. As pointed out by Nemirovsky et al. (2020) counterfactual explanations may sometimes be difficult to interpret. Therefore, the evaluation of different explanation types in human-ML collaboration setting will be a next crucial step.

Acknowledgments

Acknowledgments withheld.

References

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. Advances in neural information processing systems, 31, 2018.

David Alvarez Melis and Tommi Jaakkola. Towards robust interpretability with self-explaining neural networks. Advances in neural information processing systems, 31, 2018.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In International conference on machine learning, pages 214–223. PMLR, 2017.

Nishanth Arun, Nathan Gaw, Praveer Singh, Ken Chang, Mehak Aggarwal, Bryan Chen, Katharina Hoebel, Sharut Gupta, Jay Patel, Mishka Gidwani, et al. Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging. Radiology: Artificial Intelligence, 3(6), 2021.

Shekoofeh Azizi, Basil Mustafa, Fiona Ryan, Zachary Beaver, Jan Freyberg, Jonathan Deaton, Aaron Loh, Alan Karthikesalingam, Simon Kornblith, Ting Chen, et al. Big self-supervised models advance medical image classification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 3478–3488, 2021.

Alina Jade Barnett, Fides Regina Schwartz, Chaofan Tao, Chaofan Chen, Yinhao Ren, Joseph Y Lo, and Cynthia Rudin. Interpretable mammographic image classification using case-based reasoning and deep learning. arXiv preprint arXiv:2107.05605, 2021.

Cher Bass, Mariana da Silva, Carole Sudre, Petru-Daniel Tudosiu, Stephen Smith, and Emma Robinson. Icam: Interpretable classification via disentangled representations and feature attribution mapping. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, Advances in Neural Information Processing Systems, volume 33, pages 7697–7709. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/56f9f88906aebf4ad985aaec7fa01313-Paper.pdf.

Christian F. Baumgartner, Lisa M. Koch, Kerem Can Tezcan, Jia Xi Ang, and Ender Konukoglu. Visual feature attribution using wasserstein gans. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018.

Moritz Bohle, Mario Fritz, and Bernt Schiele. Convolutional dynamic alignment networks for interpretable classifications. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10029–10038, 2021.

Wieland Brendel and Matthias Bethge. Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. arXiv preprint arXiv:1904.00760, 2019.

Irem Cetin, Maialen Stephens, Oscar Camara, and Miguel A González Ballester. Attri-vae: Attribute-based interpretable representations of medical images with variational autoencoders. Computerized Medical Imaging and Graphics, page 102158, 2022.

Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This looks like that: deep learning for interpretable image recognition. Advances in neural information processing systems, 32, 2019.

Zhi Chen, Yijie Bei, and Cynthia Rudin. Concept whitening for interpretable image recognition. Nature Machine Intelligence, 2(12):772–782, 2020.

Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 8789–8797, 2018.

Joseph Paul Cohen, Rupert Brooks, Sovann En, Evan Zucker, Anuj Pareek, Matthew P Lungren, and Akshay Chaudhari. Gifsplanation via latent shift: a simple autoencoder approach to counterfactual generation for chest x-rays. In Medical Imaging with Deep Learning, pages 74–104. PMLR, 2021.

Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. Algorithm aversion: people erroneously avoid algorithms after seeing them err. Journal of Experimental Psychology: General, 144(1):114, 2015.

Susanne Gaube, Harini Suresh, Martina Raue, Alexander Merritt, Seth J Berkowitz, Eva Lermer, Joseph F Coughlin, John V Guttag, Errol Colak, and Marzyeh Ghassemi. Do as ai say: susceptibility in deployment of clinical decision-aids. NPJ digital medicine, 4(1): 1–8, 2021.

Thomas Grote and Philipp Berens. On the ethics of algorithmic decision-making in healthcare. Journal of medical ethics, 46(3):205–211, 2020.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. Advances in neural information processing systems, 30, 2017.

Tessa Han, Suraj Srinivas, and Himabindu Lakkaraju. Which explanation should i choose? a function approximation perspective to characterizing post hoc explanations. arXiv preprint arXiv:2206.01254, 2022.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In Proceedings of the AAAI conference on artificial intelligence, volume 33, pages 590–597, 2019.

Syed Ashar Javed, Dinkar Juyal, Harshith Padigela, Amaro Taylor-Weiner, Limin Yu, and Aaditya Prakash. Additive mil: Intrinsic interpretability for pathology. arXiv preprint arXiv:2206.01794, 2022.

Shalmali Joshi, Oluwasanmi Koyejo, Been Kim, and Joydeep Ghosh. xgems: Generating examplars to explain black-box models. arXiv preprint arXiv:1806.08867, 2018.

Ashkan Khakzar, Yang Zhang, Wejdene Mansour, Yuezhi Cai, Yawei Li, Yucheng Zhang, Seong Tae Kim, and Nassir Navab. Explaining covid-19 and thoracic pathology model predictions by identifying informative input features. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 391–401. Springer, 2021.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In International Conference on Machine Learning, pages 5338–5348. PMLR, 2020.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. Advances in neural information processing systems, 30, 2017.

Daniel Nemirovsky, Nicolas Thiebaut, Ye Xu, and Abhishek Gupta. Countergan: Generating realistic counterfactuals with residual generative adversarial nets. arXiv preprint arXiv:2009.05199, 2020.

Ha Q Nguyen, Khanh Lam, Linh T Le, Hieu H Pham, Dat Q Tran, Dung B Nguyen, Dung D Le, Chi M Pham, Hang TT Tong, Diep H Dinh, et al. Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations. arXiv preprint arXiv:2012.15029, 2020.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pages 1135–1144, 2016.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5):206–215, 2019.

Kathryn Schutte, Olivier Moindrot, Paul Hérent, Jean-Baptiste Schiratti, and Simon Jégou. Using stylegan for visual interpretability of deep learning models on medical images. arXiv preprint arXiv:2101.07563, 2021.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE international conference on computer vision, pages 618–626, 2017.

Sumedha Singla, Brian Pollack, Junxiang Chen, and Kayhan Batmanghelich. Explanation by progressive exaggeration. arXiv preprint arXiv:1911.00483, 2019.

Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. arXiv preprint arXiv:1412.6806, 2014.

Guolei Sun, Thomas Probst, Danda Pani Paudel, Nikola Popović, Menelaos Kanakis, Jagruti Patel, Dengxin Dai, and Luc Van Gool. Task switching network for multi-task learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 8291–8300, 2021.

Philipp Tschandl, Christoph Rinner, Zoe Apalla, Giuseppe Argenziano, Noel Codella, Allan Halpern, Monika Janda, Aimilios Lallas, Caterina Longo, Josep Malvehy, et al. Human–computer collaboration for skin cancer recognition. Nature Medicine, 26(8):1229–1234, 2020.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2097–2106, 2017.

Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In European conference on computer vision, pages 499–515. Springer, 2016.

Adam White and Artur d'Avila Garcez. Measurable counterfactual local explanations for any classifier. arXiv preprint arXiv:1908.03020, 2019.

Wenwu Ye, Jin Yao, Hui Xue, and Yi Li. Weakly supervised lesion localization with probabilistic-cam pooling. arXiv preprint arXiv:2005.14480, 2020.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2921–2929, 2016.

Appendix A. Additional evaluations

A.1. Examples of counterfactual generations

Examples of counterfactual images obtained by adding the class-specific visual attribution map to the input image, i.e. $\hat{\mathbf{x}} = \mathbf{x} + M_c(\mathbf{x})$, are shown in Fig. 4.

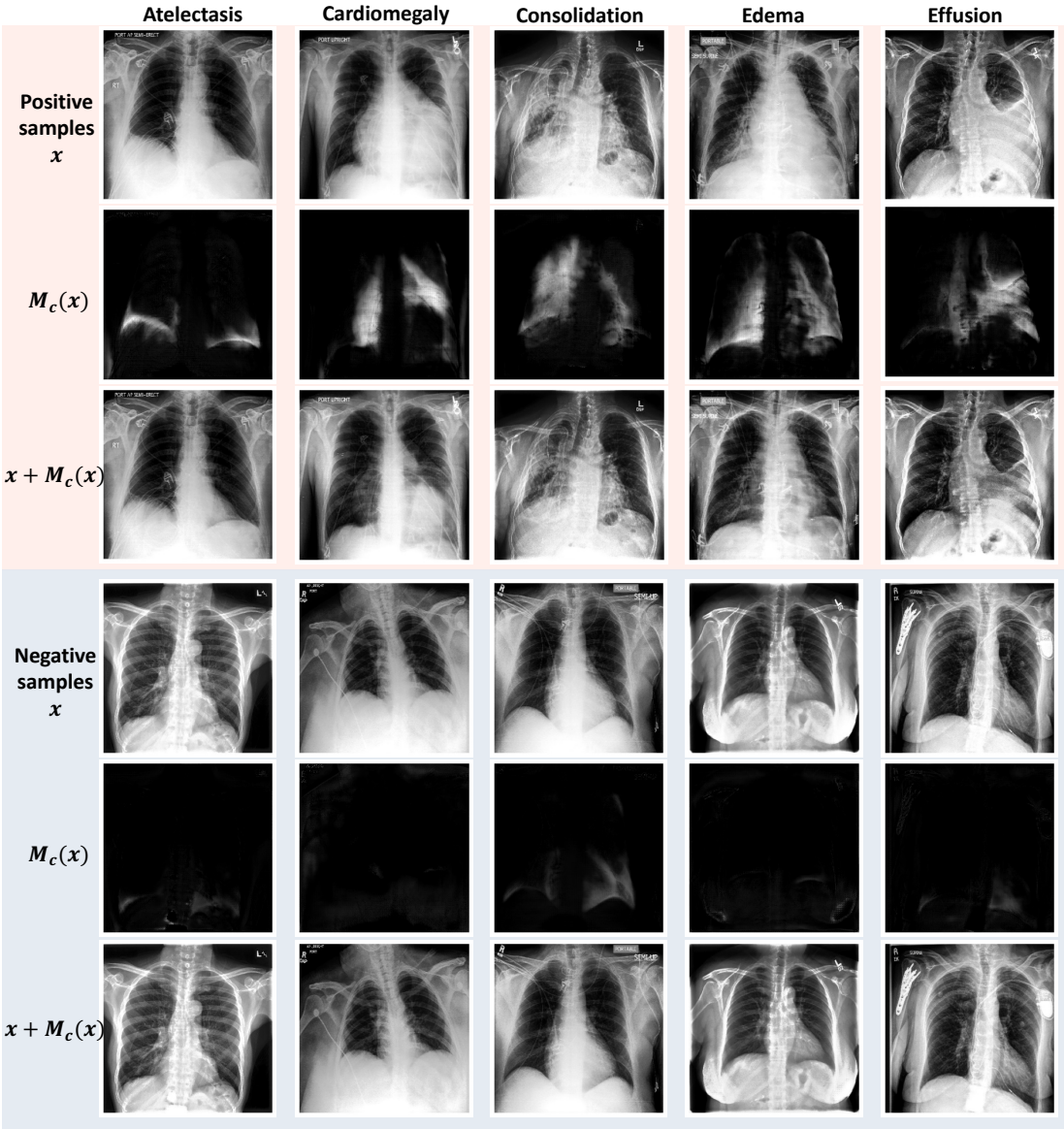

Figure 4: Counterfactual image generation. The examples in the top group of rows show input images containing evidence for different classes c. For those, the evidence for class c is removed by adding $M_c(\mathbf{x})$. The bottom group of rows contains images with no evidence for class c. Those images remain mostly unchanged by adding the output of $M_c(\mathbf{x})$.

A.2. Global interpretability

A distinction is often made between local explanations, which explain the prediction for a specific input image, and *global* explanations, which explain the decision mechanisms of the ML algorithm as a whole (i.e. for all input images). While the primary focus of our paper was on local interpretability, we may gain some global insights about the decision mechanism of the classifier through interpretation of the positive and negative class centers introduced in Section 2.2, as well as the class specific weights of the logistic regression classifier. The class centers capture some prototypical aspects of the respective classes, while the classifier weights can tell us which areas of the images the classifier is paying attention to for each class . Fig. 5 shows the class centres for five diseases and the weights of the corresponding classifiers trained on the ChestX-ray8 dataset.

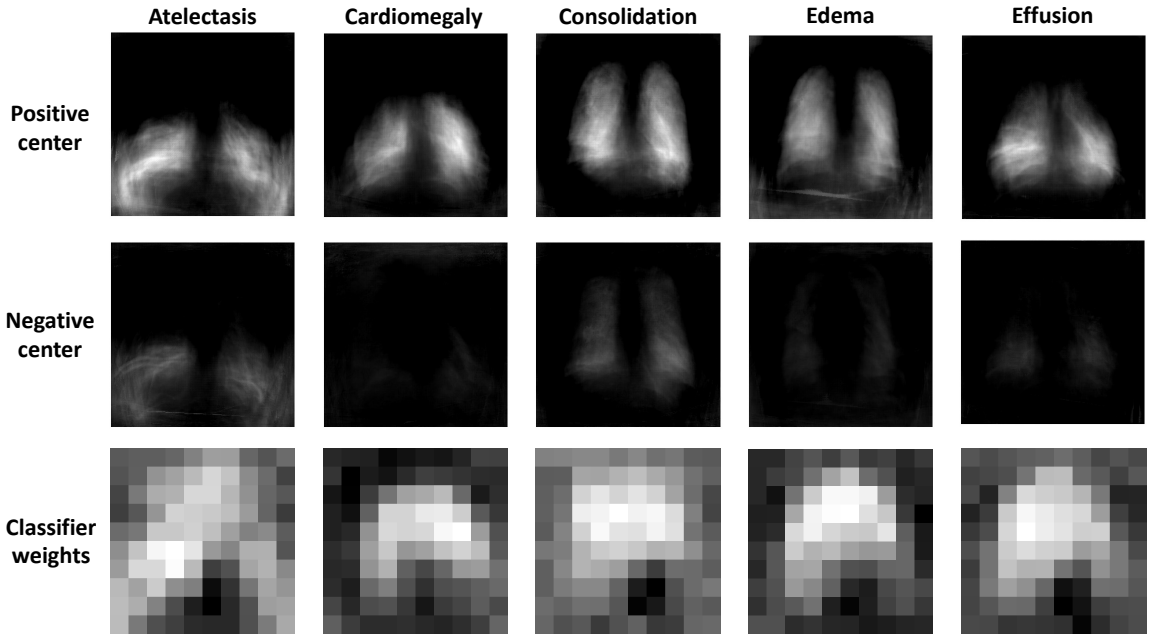

Figure 5: Attribution centers of disease and corresponding classifiers' weights provide a global explanation of Attri-Net.

A.3. Ablation study of the loss terms

An ablation study on the effects of the losses used for training Attri-Net can be found in Tab. 3. Example attributions for all combinations for an image from the ChestX-ray8 dataset are shown in Fig. 6.

Table 3: Ablation study on the four losses. Evaluated on the Vindr-CXR dataset.

Model	Loss terms	Classification AUCs	Class sensitivity
Attri-Net$_{\text{cls}}$	$\mathcal{L}_{\text{cls}}$	0.9339	0.2516
Attri-Net$_{\text{cls_adv}}$	$\mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{adv}}$	**0.9444**	0.1602
Attri-Net$_{\text{cls_adv_reg}}$	$\mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{adv}} + \mathcal{L}_{\text{reg}}$	0.9397	0.5259
Attri-Net$_{\text{all}}$	$\mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{adv}} + \mathcal{L}_{\text{reg}} + \mathcal{L}_{\text{ctr}}$	0.9405	**0.5509**

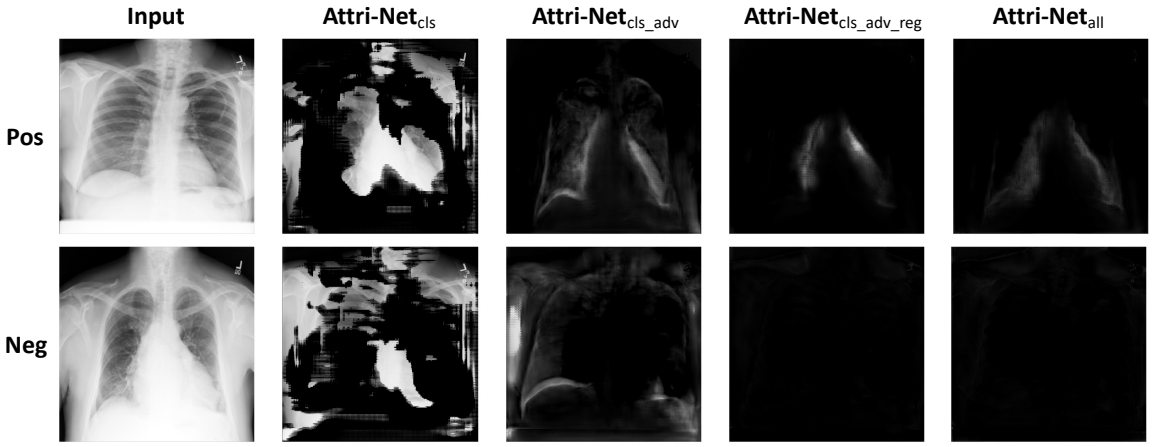

Figure 6: Examples of attribution maps obtained for positive and negative samples of the disease cardiomegaly on ChestX-ray8 for different subsets of our losses.

A.4. Class sensitivity image grids

The class sensitivity evaluations in Section 3 is based on class sensitivity grids as proposed by Bohle et al. (2021). In Fig. 7, Fig. 8, and Fig. 9 we show examples of such grids for all studied classes on the CheXpert dataset, for Attri-Net, CoDA-Nets and Gifsplanation, respectively.

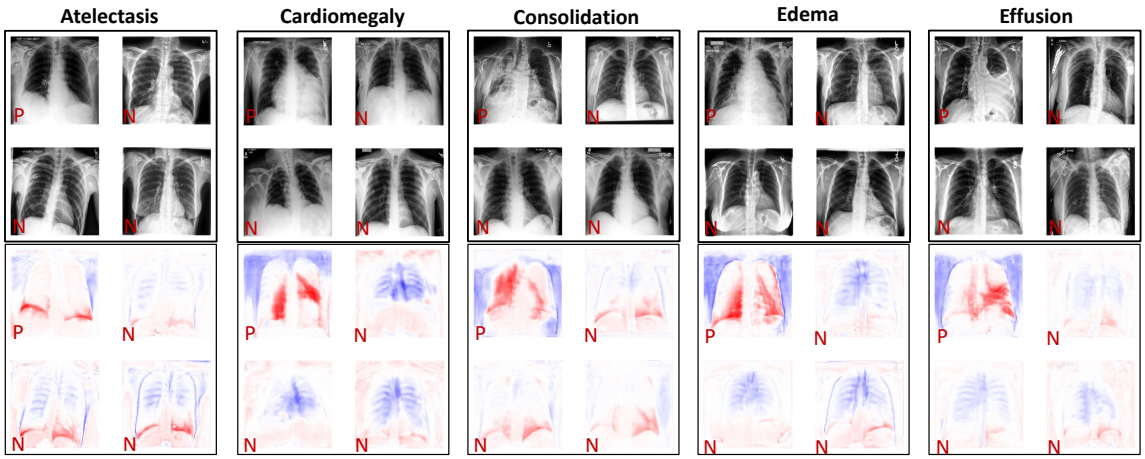

Figure 7: Class sensitivity image grids obtained using Attri-Net. The first row shows image grids, the second row shows the respective attribution maps. P and N denote class-positive and class-negative examples, respectively.

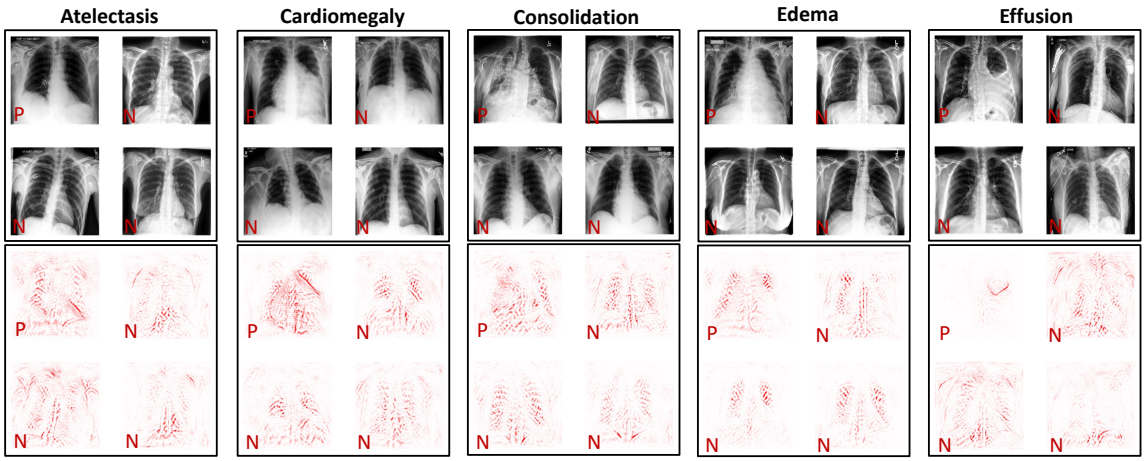

Figure 8: Class sensitivity image grids obtained using CoDA-Nets. The first row shows image grids, the second row shows the respective attribution maps. P and N denote class-positive and class-negative examples, respectively.

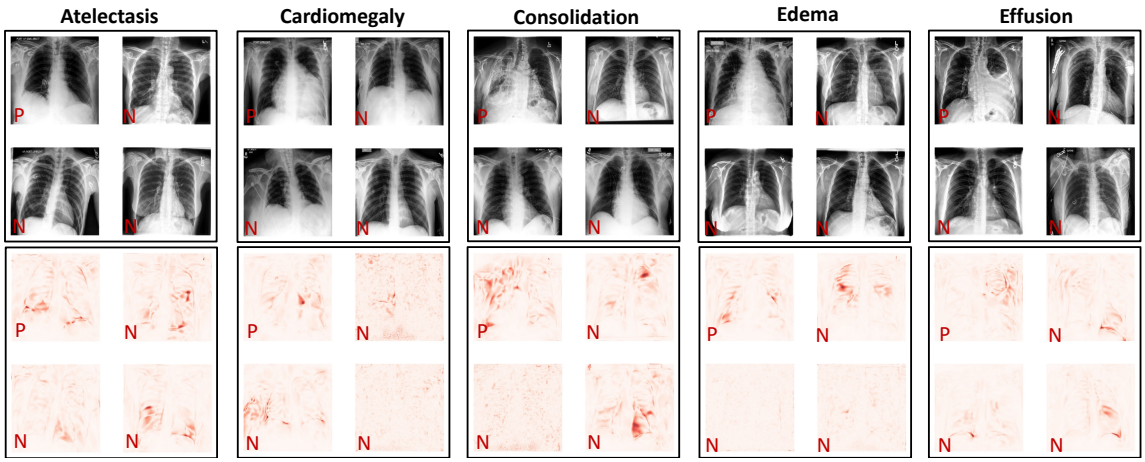

Figure 9: Class sensitivity image grids obtained using Gifsplanation. The first row shows image grids, the second row shows the respective attribution maps. P and N denote class-positive and class-negative examples, respectively.

A.5. Example explanations for ChestX-ray8 and Vindr-CXR

Fig. 10 and Fig. 11 contain additional examples of visual attributions using all compared methods derived from the ChestX-ray8 and Vindr-CXR datasets, respectively.

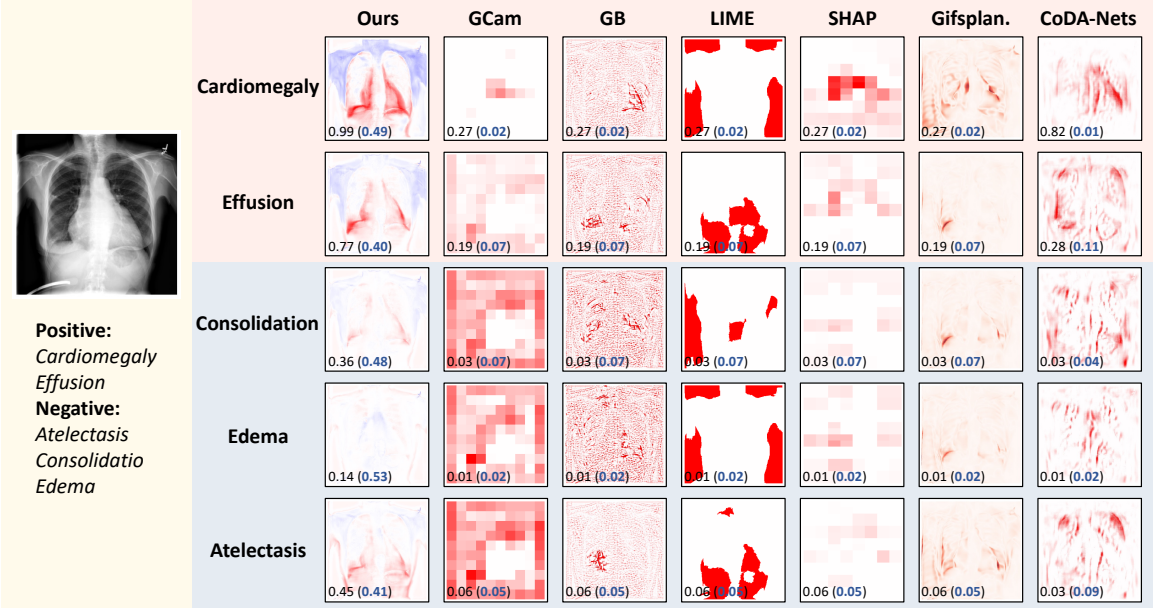

Figure 10: Explanations for an example image from the ChestX-ray8 dataset.

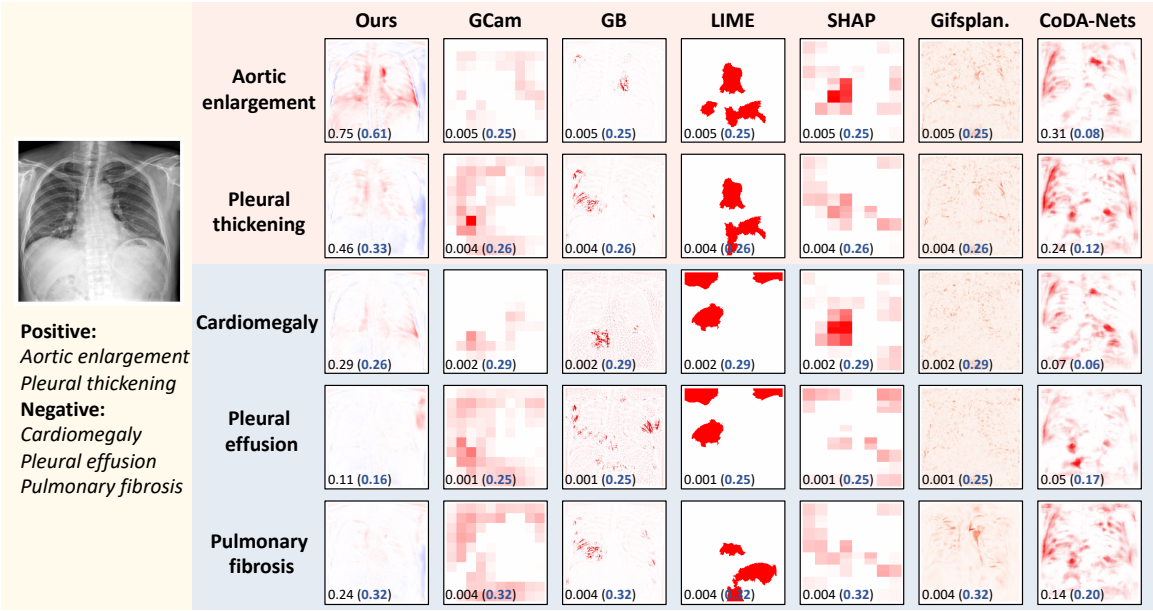

Figure 11: Explanations for an example image from the VindrCXR dataset.

Appendix B. Additional training details

B.1. Discriminator training

The Attri-Net framework requires training a discriminator function D in parallel to the class attribution generator M. The weight parameters θ of the discriminator are computed in separate gradient update steps using the Wasserstein GAN (Arjovsky et al., 2017) objective. The full discriminator optimisation objective is then given by

$$\min_{\theta} \sum_{c} \underset{\mathbf{x}\sim p(\mathbf{x}|y_c=0)}{\mathbb{E}} [D_c(\mathbf{x}|\theta)] + \underset{\mathbf{x}\sim p(\mathbf{x}|y_c=1)}{\mathbb{E}} [D_c(\mathbf{x} + M_c(\mathbf{x})|\theta)] \,,$$

where we omitted the gradient penalty loss which ensures the discriminator fulfills the Lipschitz-1 constraint dictated by the Wasserstein GAN objective (Gulrajani et al., 2017).

B.2. Network architecture

The network architecture of the attribution map generator and the discriminator of the Attri-Net framework are shown in Tab. 4 and Tab. 5, respectively. L refers to the length of input/output features, N are the number of output channels and K the kernel size.

Table 4: Attri-Net class attribution generator network architecture.

Layers	Input → Output	Layer information
Task embedding layer	Task code $\mathbf{t}_c$ → Task embedding $\mathbf{t}'_c$	8 × FC(L100,L100)
Down-sampling	(Input image $\mathbf{x}$, $\mathbf{t}'_c$) → $\mathbf{out}_{\text{down}}$	Ada_Conv: CONV(N64, K7x7), AdaIN, ReLU Ada_Conv: CONV(N128, K4x4), AdaIN, ReLU Ada_Conv: CONV(N256, K4x4), AdaIN, ReLU
Bottlenecks	($\mathbf{out}_{\text{down}}$, $\mathbf{t}'_c$) → $\mathbf{out}_{\text{bn}}$	Ada_ResBlock: CONV(N256, K3x3), AdaIN, ReLU Ada_ResBlock: CONV(N256, K3x3), AdaIN, ReLU Ada_ResBlock: CONV(N256, K3x3,), AdaIN, ReLU Ada_ResBlock: CONV(N256, K3x3), AdaIN, ReLU Ada_ResBlock: CONV(N256, K3x3), AdaIN, ReLU Ada_ResBlock: CONV(N256, K3x3), AdaIN, ReLU
Up-sampling	($\mathbf{out}_{\text{bn}}$, $\mathbf{t}'_c$) → $\mathbf{out}_{\text{up}}$	Ada_DECONV(N128, K4x4), AdaIN, ReLU Ada_DECONV(N64, K4x4), AdaIN, ReLU CONV(N1, K7x7)
Output layer	($\mathbf{x}$, $\mathbf{out}_{\text{up}}$) → $M_c(\mathbf{x})$	$M_c(\mathbf{x}) = \tanh(\mathbf{x} + \mathbf{out}_{\text{up}}) - \mathbf{x}$

Table 5: Attri-Net discriminator network architecture.

Layers	Input → Output	Layer information
Task embedding layer	Task code $\mathbf{t}_c$ → Task embedding $\mathbf{t}'_c$	8 × FC(L100,L100)
Input layer **Hidden layers**	($\mathbf{x}/\hat{\mathbf{x}}$, $\mathbf{t}'_c$) → $\mathbf{out}_{\text{hid}}$	Ada_Conv: CONV(N64, K4x4), AdaIN, ReLU Ada_Conv: CONV(N128, K4x4), AdaIN, ReLU Ada_Conv: CONV(N256, K4x4), AdaIN, ReLU Ada_Conv: CONV(N512, K4x4), AdaIN, ReLU Ada_Conv: CONV(N1024, K4x4), AdaIN, ReLU Ada_Conv: CONV(N2048, K4x4), AdaIN, ReLU
Output layer	$\mathbf{out}_{\text{hid}} \rightarrow \mathcal{L}_{\text{adv}}^{(c)}$	CONV(N1, K3x3)