Attri-Net: Inherently Interpretable Multi-Label Classification Using Class-Specific Counterfactuals

Susu Sun[1], Stefano Woerner[1], Andreas Maier[2], Lisa M. Koch[1], Christian F. Baumgartner[1]
[1]University of Tübingen, [2]University of Erlangen-Nuremberg
susu.sun@uni-tuebingen.de

Motivation

Problem:

- Interpretability is important for high-stakes medical applications.
- Deep Neural Networks are difficult to explain.
- Widely used post-hoc explanations have several drawbacks.

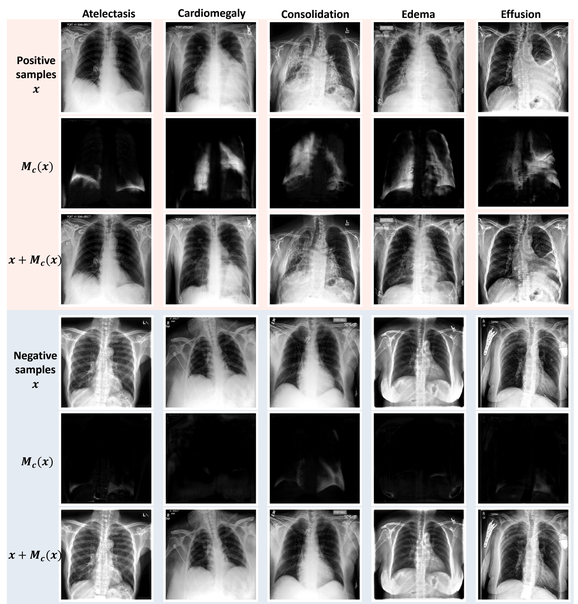
Contributions:

- We designed an **inherently interpretable multi-label** classification model Attri-Net.
- Interpretability is inherent through linear models that use human-understandable attribution maps for classification.

Counterfactual Attribution Map

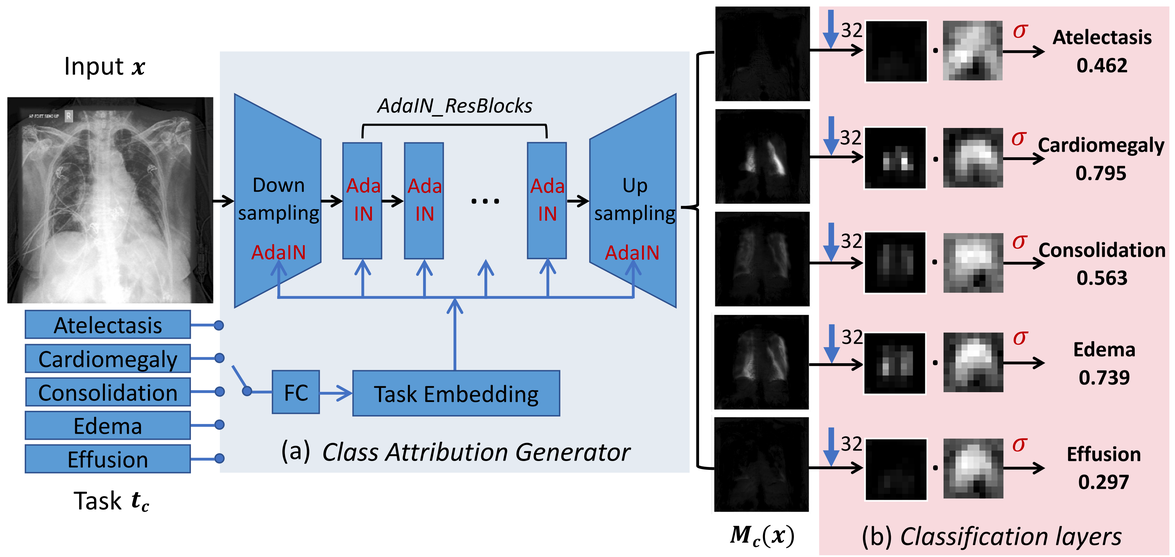
Class-Specific Counterfactual Attribution Map $M_c(\mathbf{x})$

- is an additive map that highlights disease-relevant regions.
- the summed image $\hat{\mathbf{x}} = \mathbf{x} + M_c(\mathbf{x})$ appears to come from the negative class $c = 0$.



Source — Code — Paper

Attri-Net Framework



- **Class Attribution Generator:** generates class-specific counterfactual attribution map $M_c(\mathbf{x})$ to identify disease effects in image $\mathbf{x}$ corresponding to certain disease class c (blue box in figure).
- **Classification Layers:** classify only based on counterfactuals with simple logistic regression classifiers (red box in figure).
- **Task Switching Mechanism:** makes multi-label classification possible by injecting specific diagnostic tasks to the generator through Adaptive Instance Normalization layers (AdaIN modules in figure).

Human interpretable counterfactual attribution map + **interpretable** linear classifier –> **inherent interpretability**

Quantitative Evaluation

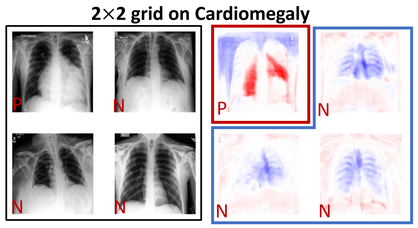
Class sensitivity property requires distinct attribution maps for disease-positive and disease-negative images.

Class Sensitivity Score:

$$s_c = \frac{1}{Z} \sum_{p_{c=1} \in I_{c=1}} p_{c=1}$$

$$\text{with} \quad Z = \sum_k \sum_{p_{c=1} \in I_k} p_{c=1}$$

$p_{c=1}$: attributions show postive disease effects at localization p.



Attri-Net achieves higher class sensitivity scores than other explanation methods on three datasets.

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**

Training

We train Attri-Net end-to-end with four loss terms to meet ensure attribution maps:

- preserve sufficient class relevant information for classification.
- are human-interpretable.

Overall loss for Class Attribution Generator:

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

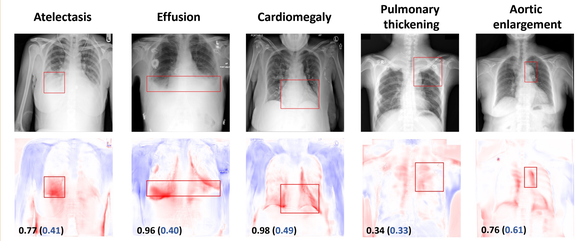
Classification Peformance

Attri-Net achieves comparable classification performance with both black box model and interpretable model on three datasets.

Model	CheXpert	ChestX-ray8	VindrCXR
ResNet50	0.7727	0.7445	0.8986
CoDA-Nets	0.7659	0.7727	0.9322
ours	0.7405	0.7762	0.9405

Qualitative Evaluation

Attri-Net highlights regions associated with specific diseases that are consistent with clinical knowledge.



Attri-Net generates distinct patterns for different diseases, making the attribution maps more human-understandable.