Deep Learning for Medical Image Analysis

COMP5423

Hao CHEN

Dept. of CSE,CBE&LIFS, HKUST

jhc@cse.ust.hk

THE DEPARTMENT OF
COMPUTER SCIENCE & ENGINEERING
計算機科學及工程學系

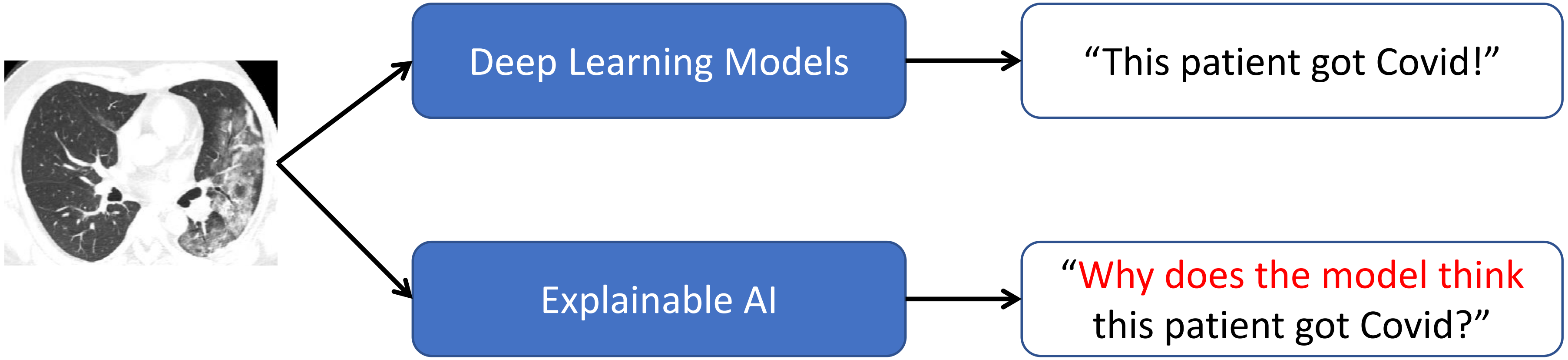
香港科技大學
THE HONG KONG UNIVERSITY OF
SCIENCE AND TECHNOLOGY

Explainability in MIA

- Introduction
- Categories of Explainable AI
- Perturbation-based Methods
- Backpropagation-based Methods
- Concept-based Methods
- Challenges and Future Directions

Why we need explainability in MIA?

- A medical image diagnosis system needs to be transparent, understandable, and explainable to gain the trust of physicians, regulators as well as patients.
- Traceability of the decisions is now a requirement.



Deep Learning Models

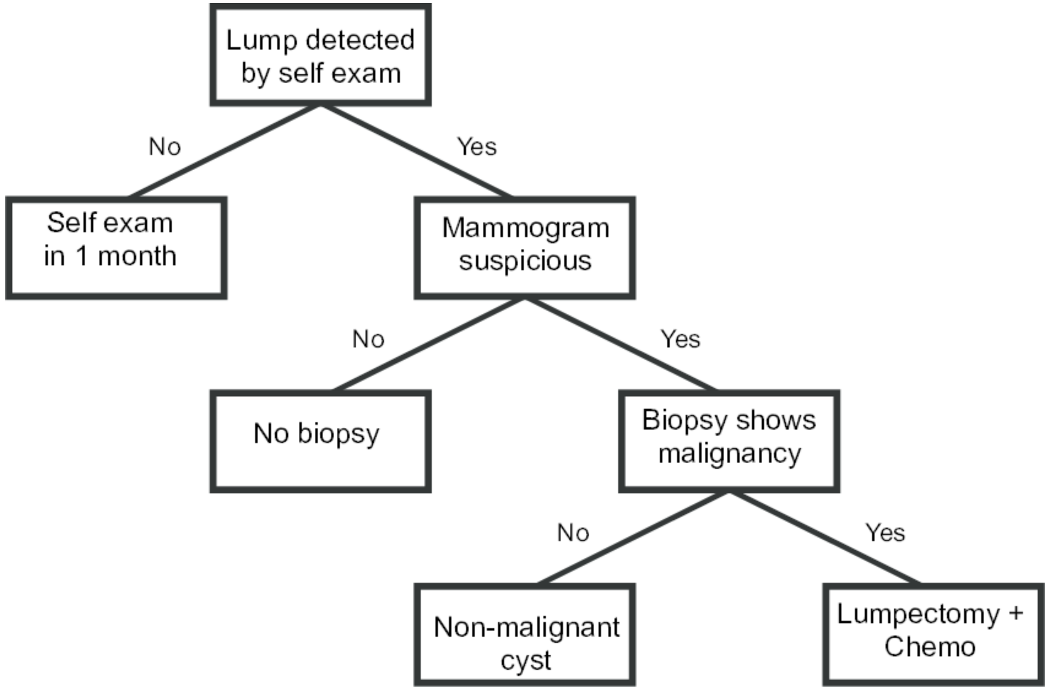
“This patient got Covid!”

Explainable AI

“Why does the model think this patient got Covid?”

Explainable AI

- Some simple AI methods are self-explanatory, e.g., linear regression, decision trees.



An example of decision tree in breast cancer diagnosis and treatment.

Applications of Machine Learning in Cancer Prediction and Prognosis, Cancer Informatics, 2007.

Explainable AI

- Some simple AI methods are self-explanatory, e.g., linear regression, decision trees.
- However, these models lack the complexity required for tasks such as medical image analysis.
- Deep learning with many-layers transformation enables highly non-convex complexity, however, its **black-box** nature affects the applicability in many domains, e.g., finance, autonomous driving, medical diagnosis.
- Explainable AI (XAI) is a promising direction for enhancing the transparency and interpretability of deep learning.

Explainability in MIA

- Introduction
- Categories of Explainable AI
- Perturbation-based Methods
- Backpropagation-based Methods
- Concept-based Methods
- Challenges and Future Directions

Categories of Explainable AI

- **Model Specific vs. Model Agnostic**

Model-specific interpretation methods are based on the parameters of the individual models.

Model Agnostic methods are mainly applicable in post-hoc analysis and not limited to specified model architecture. These methods do not have direct access to the internal model weights or structural parameters.

Explainable Deep Learning Models in Medical Image Analysis. Journal of Imaging, 2020

Categories of Explainable AI

- **Global Methods vs. Local Methods**

Global methods concentrate on the inside of a model by exploiting the overall knowledge about the model, training, and associated data. It tries to explain the behavior of the model in general, e.g., feature importance.

Local interpretable methods are applicable to a single outcome of the model. This can be done by designing methods that can explain the reason for a particular prediction or outcome.

Explainable Deep Learning Models in Medical Image Analysis. Journal of Imaging, 2020

Categories of Explainable AI

- **Pre-Model vs. In-Model vs. Post-Model**

Pre-model methods are independent and do not depend on a particular model architecture, e.g., PCA, t-SNE.

In-model methods are integrated into the model itself.

Post-model methods are implemented after building a model.

Explainable Deep Learning Models in Medical Image Analysis. Journal of Imaging, 2020.
van et al. Visualizing Data Using t-SNE. JMLR 2008

Categories of Explainable AI

- **Surrogate Methods vs. Visualization Methods**

Surrogate methods consist of different models as an ensemble which are used to analyze other black-box models. The black box models can be understood better by interpreting the surrogate model's decisions.

The visualization methods are not a different model, but it helps to explain some parts of the models by visual understanding like visualizing the activation maps.

Explainable Deep Learning Models in Medical Image Analysis. Journal of Imaging, 2020

Categories of Explainable AI

- Model Specific vs. Model Agnostic

Can it explain a particular model or many models?

- Global Methods vs. Local Methods

Does it explain a particular sample or entire model?

- Pre-Model vs. In-Model vs. Post-Model

When does it occur?

- Surrogate Methods vs. Visualization Methods

Does it work separately from the model, or does it visualize the model?

The categories are non-exclusive. There is no universally accepted taxonomy of XAI techniques!

Explainable Deep Learning Models in Medical Image Analysis. Journal of Imaging, 2020

Explainability in MIA

- Introduction
- Categories of Explainable AI
- Perturbation-based Methods
- Backpropagation-based Methods
- Concept-based Methods
- Challenges and Future Directions

Perturbation-based Methods

- This can be implemented by *removing*, *masking*, or *modifying* certain input features, and running the forward pass (output computation), and measuring the **difference** from the original output.

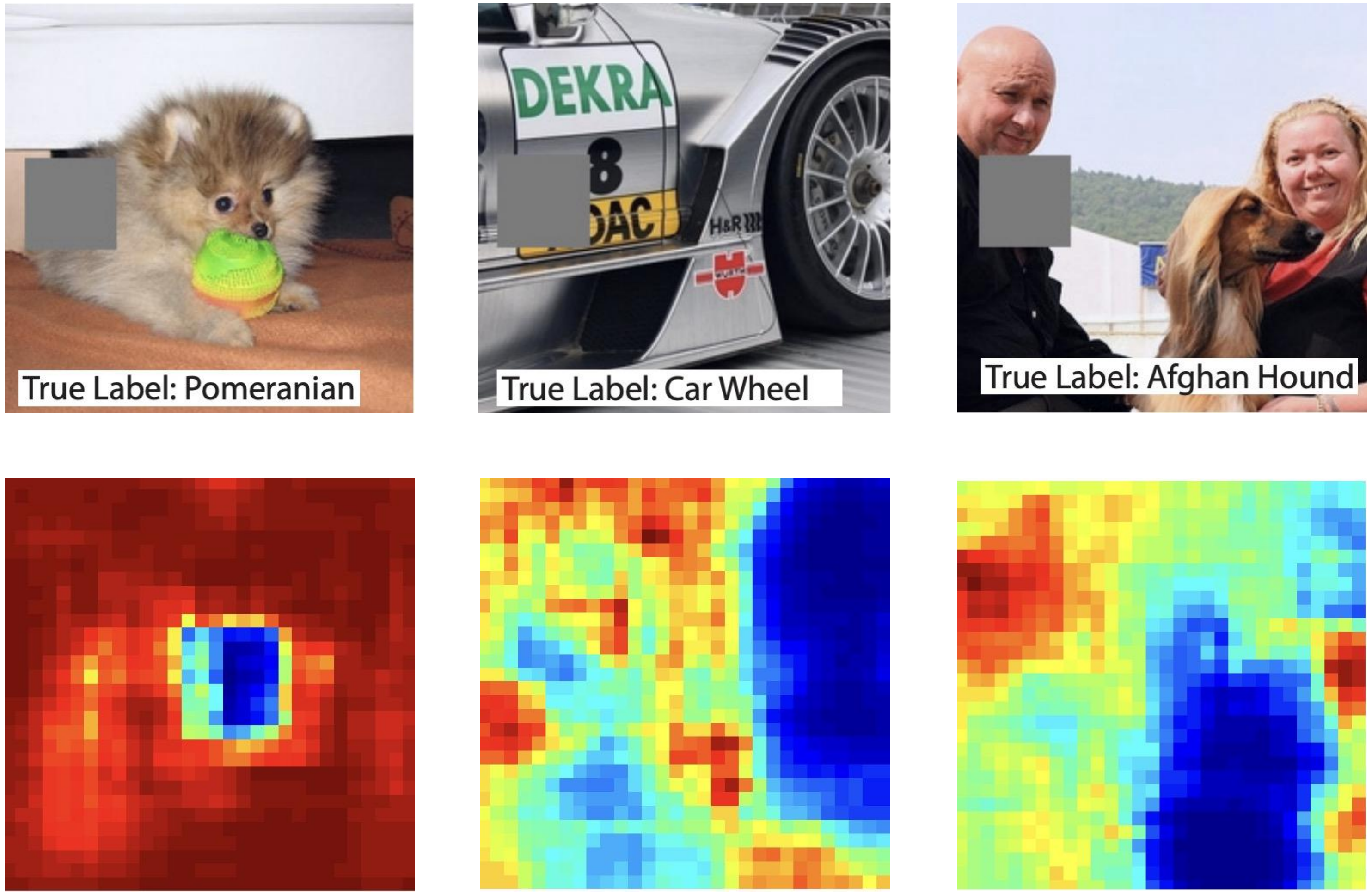
- It can reveal if a model is **overfitting** and learning **irrelevant features**

- It is computationally **expensive** as a forward pass needs to be run after perturbing each group of features of the input.

Occlusion Sensitivity

- Occluding is a good way to investigate the sensitive regions.

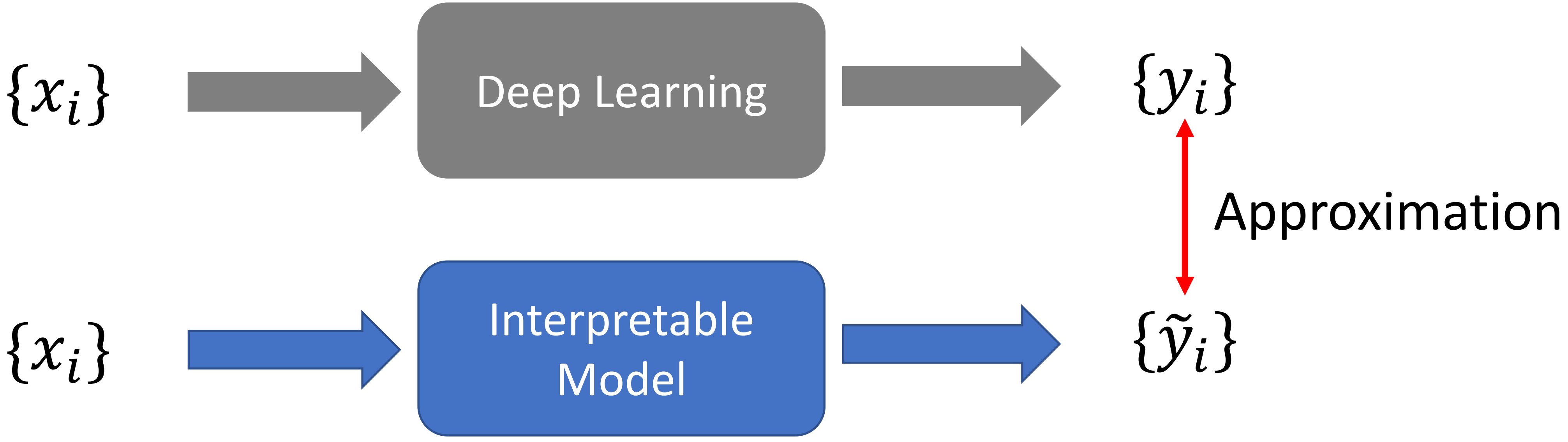
The first row: input images;
The second row: a map of correct class probability, as a function of the position of the gray square.



Zeiler, et al. Visualizing and understanding convolutional networks. ECCV 2014.

LIME (Local Interpretable Model-agnostic Explanations)

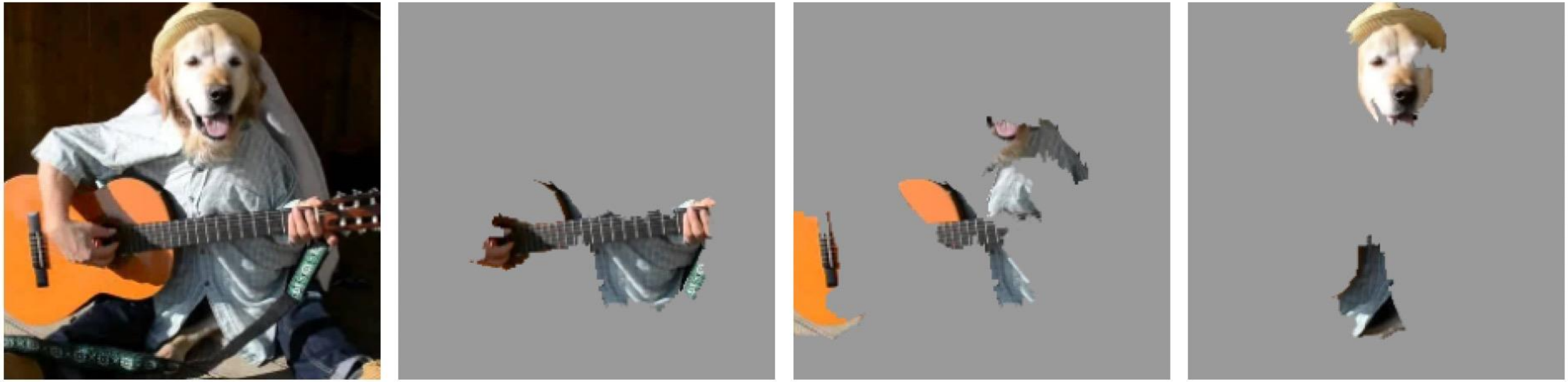
- The LIME approximates the behaviour of a deep neural network using a simpler, more interpretable model locally around the prediction. Interpreting the decisions of this simpler model provides insight into the decisions of the deep neural network.



Ribeiro et al. " Why should I trust you?" Explaining the predictions of any classifier. ACM KDDM 2016.

LIME (Local Interpretable Model-agnostic Explanations)

- The simple model is used to determine the importance of features of the input data, as a proxy for the importance of the features to the deep neural network.



(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Ribeiro et al. “ Why should I trust you?” Explaining the predictions of any classifier. ACM KDDM 2016.

Explainability in MIA

- Introduction
- Categories of Explainable AI
- Perturbation-based Methods
- Backpropagation-based Methods
- Concept-based Methods
- Challenges and Future Directions

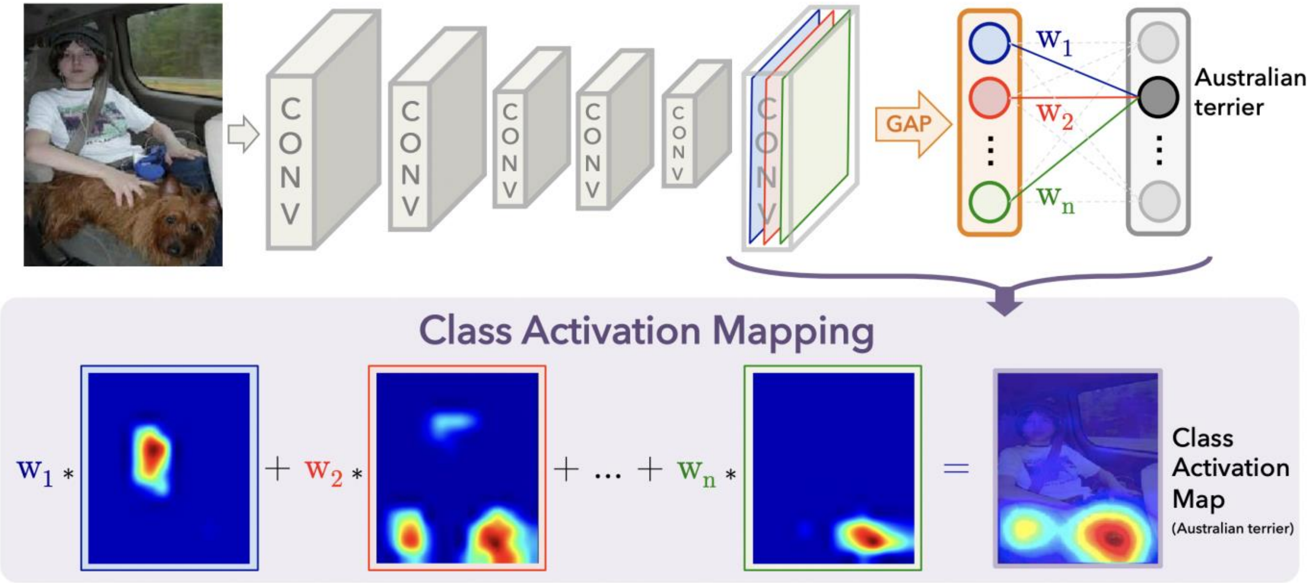
Backpropagation-based Methods

- Backpropagation-based methods compute the attribution for all the input features with a forward and backward pass through the network. The computation cost is usually much lower than perturbation-based methods.

- The ratings of different attributions by domain experts are potentially useful to develop explainable models which are more likely to be trusted by the end users and hence should be a critical part of the development of an XAI system。

Class Activation Map (CAM)

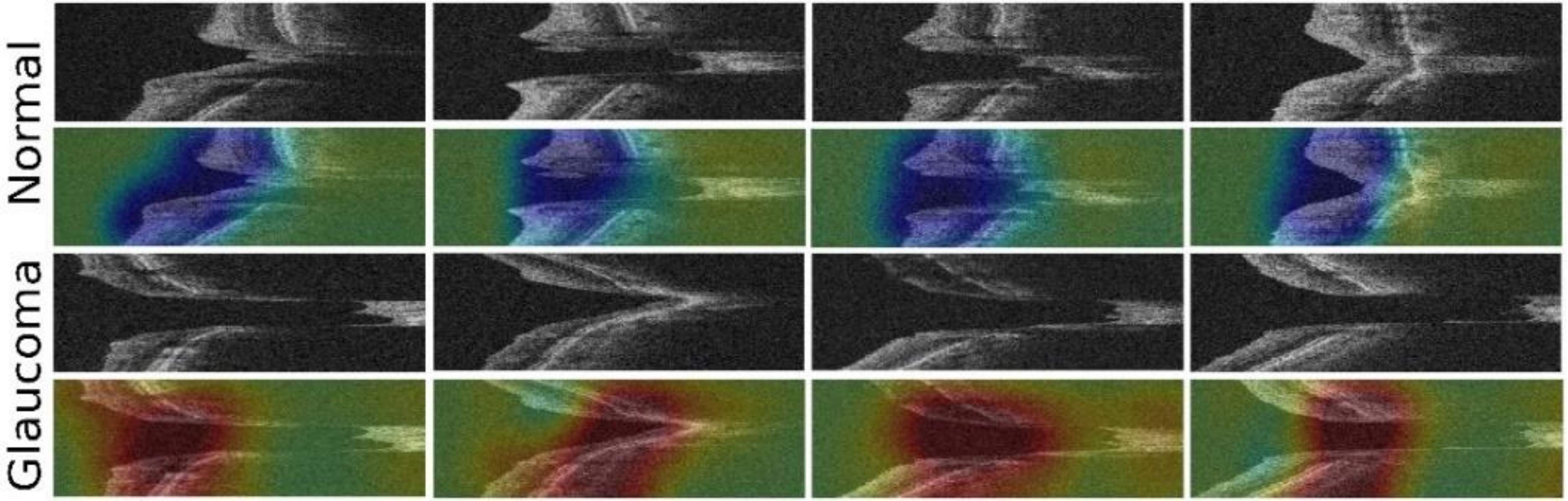
- Global average pooling outputs the spatial average of the feature map of each unit at the last convolutional layer. A weighted sum of these values is used to generate the final output.
- Similarly, we compute a weighted sum of the feature maps of the last convolutional layer to obtain the class activation maps.



Zhou et al. Learning deep features for discriminative localization. CVPR 2016.

Class Activation Map (CAM)

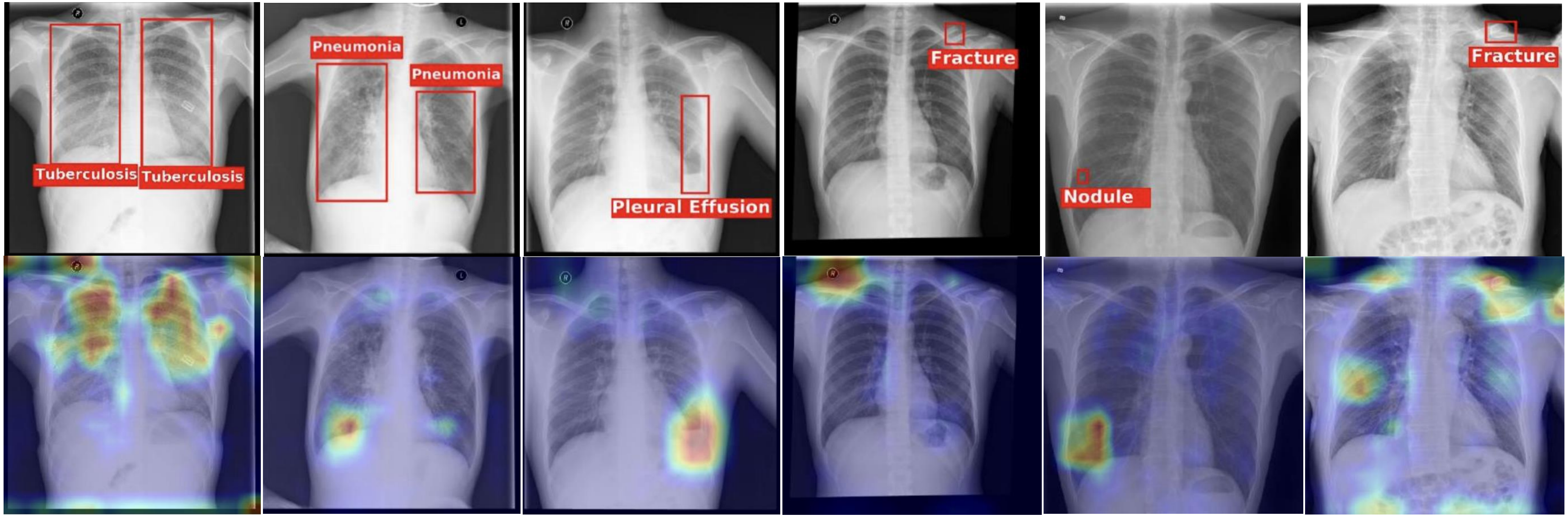
- Glaucoma Diagnosis from OCT Images



Wang et al. Towards multi-center glaucoma OCT image screening with semi-supervised joint structure and function multi-task learning. MIA 2020.

Class Activation Map (CAM)

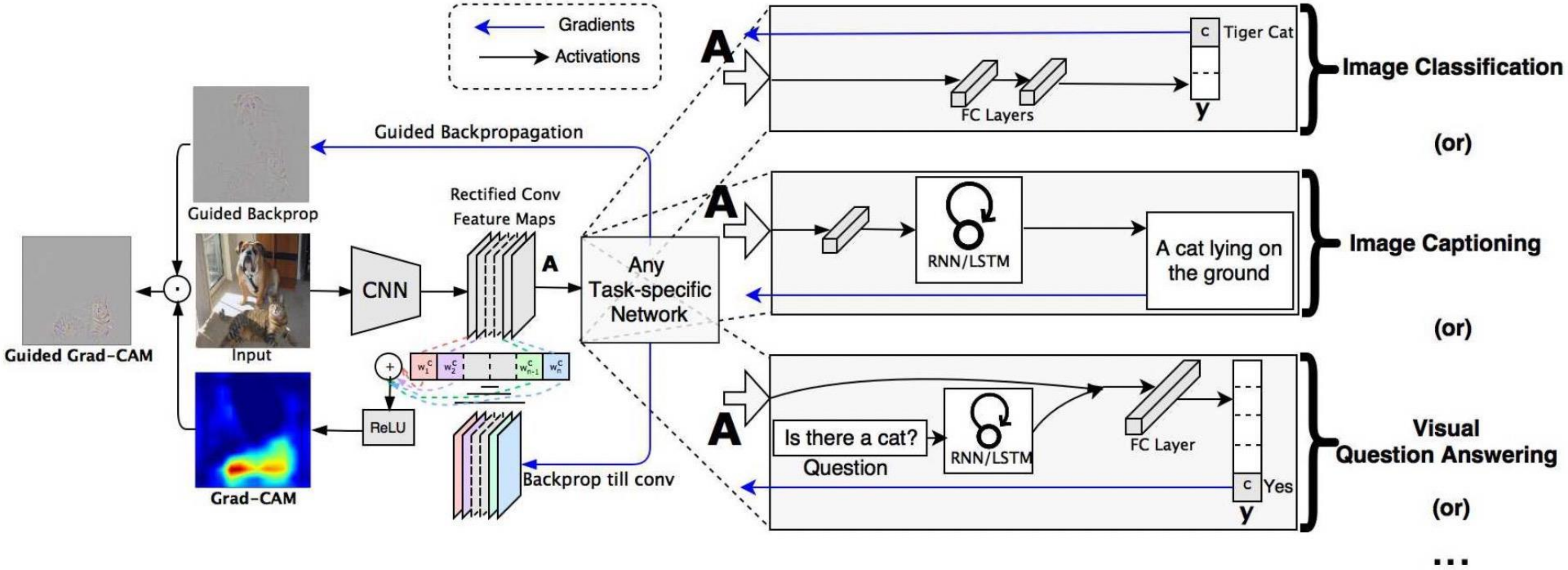
- Disease Diagnosis from Chest X-rays



Luo et al. OXnet: Deep Omni-Supervised Thoracic Disease Detection from Chest X-Rays. MICCAI 2021.

Grad-CAM

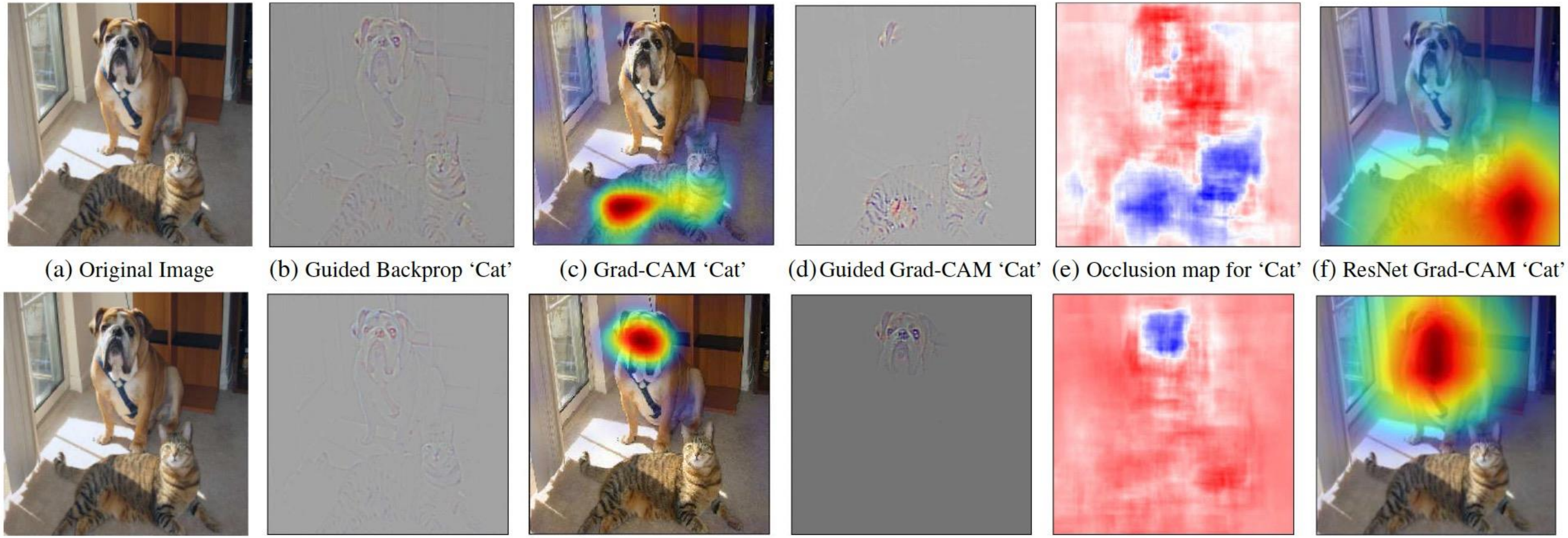
- Gradient-weighted class activation mapping (Grad-CAM) uses the gradients of any target concept flowing into the final convolutional layer to produce a coarse localization map highlighting the important regions in the image for predicting the concept.



Selvaraju et al. Grad-CAM: Visual explanations from deep networks via gradient-based localization. ICCV 2017.

Grad-CAM

- Grad-CAM is a generalization of the CAM method that is applicable to a significantly broader range of CNN model families.



(a) Original Image (b) Guided Backprop 'Cat' (c) Grad-CAM 'Cat' (d) Guided Grad-CAM 'Cat' (e) Occlusion map for 'Cat' (f) ResNet Grad-CAM 'Cat'

(g) Original Image (h) Guided Backprop 'Dog' (i) Grad-CAM 'Dog' (j) Guided Grad-CAM 'Dog' (k) Occlusion map for 'Dog' (l) ResNet Grad-CAM 'Dog'

Selvaraju et al. Grad-CAM: Visual explanations from deep networks via gradient-based localization. ICCV 2017.

Explainability in MIA

- Introduction
- Categories of Explainable AI
- Perturbation-based Methods
- Backpropagation-based Methods
- Concept-based Methods
- Challenges and Future Directions

Concept-based Methods

- Concept-based methods compute the explainable concepts in the intermediate layers of the model and use them for the final prediction task.

- This type of methods typically offers the explainability in an “in-model” style, compared to previous methods that typically occur in a “post-model” style.

Self-explaining Neural Networks (SENN)

- Explainable concepts should fulfill the following principles:

Fidelity: Concepts should preserve relevant information from the original data.

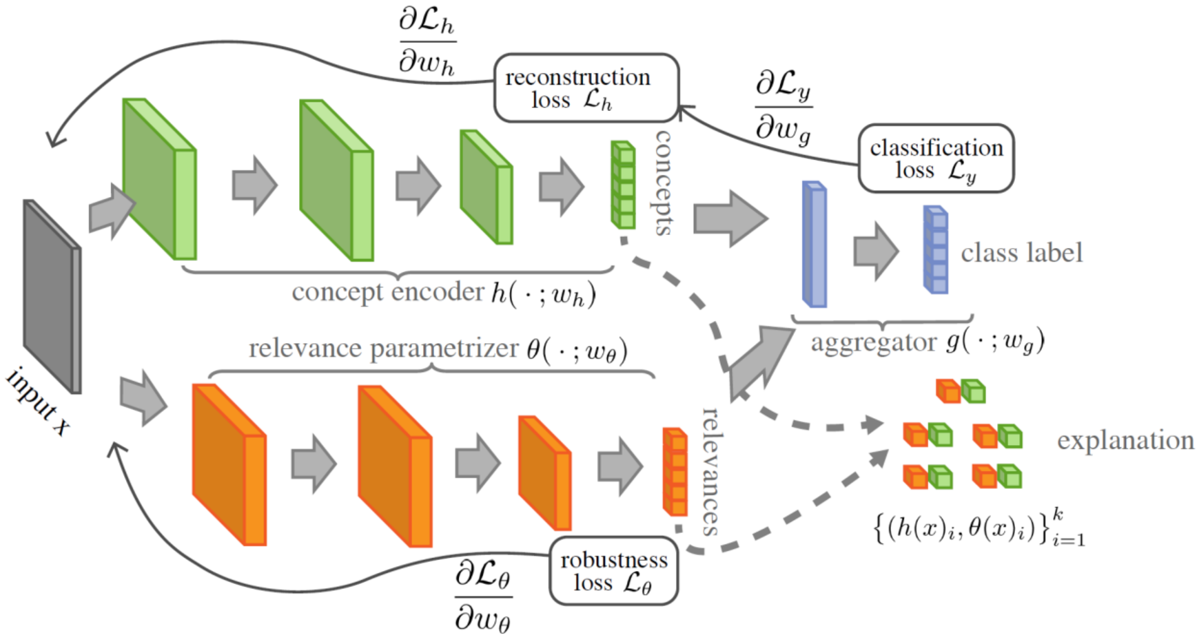
Diversity: Original data should be represented with a few non-overlapping concepts.

Grounding: Concepts should be understood by humans immediately.

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

SENN

- Concept encoder transforms the input into a small set of interpretable basis features.
- Input dependent parametrizer generates relevance scores.
- Aggregation function combines to produce the prediction.

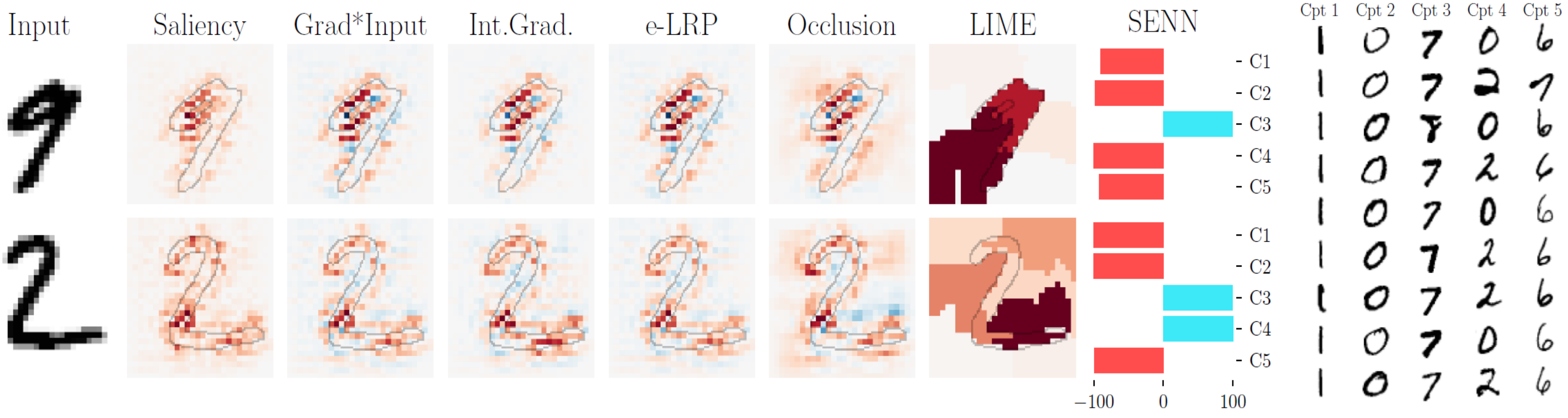


$\frac{\partial \mathcal{L}_h}{\partial w_h}$
reconstruction loss $\mathcal{L}_h$
$\frac{\partial \mathcal{L}_y}{\partial w_g}$
classification loss $\mathcal{L}_y$
concepts
class label
concept encoder $h(\cdot\,;w_h)$
input x
relevance parametrizer $\theta(\cdot\,;w_\theta)$
aggregator $g(\cdot\,;w_g)$
explanation
relevances
$\{(h(x)_i, \theta(x)_i)\}_{i=1}^k$
robustness loss $\mathcal{L}_\theta$
$\frac{\partial \mathcal{L}_\theta}{\partial w_\theta}$

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

SENN

- The explanation includes a characterization of concepts in terms of defining prototypes

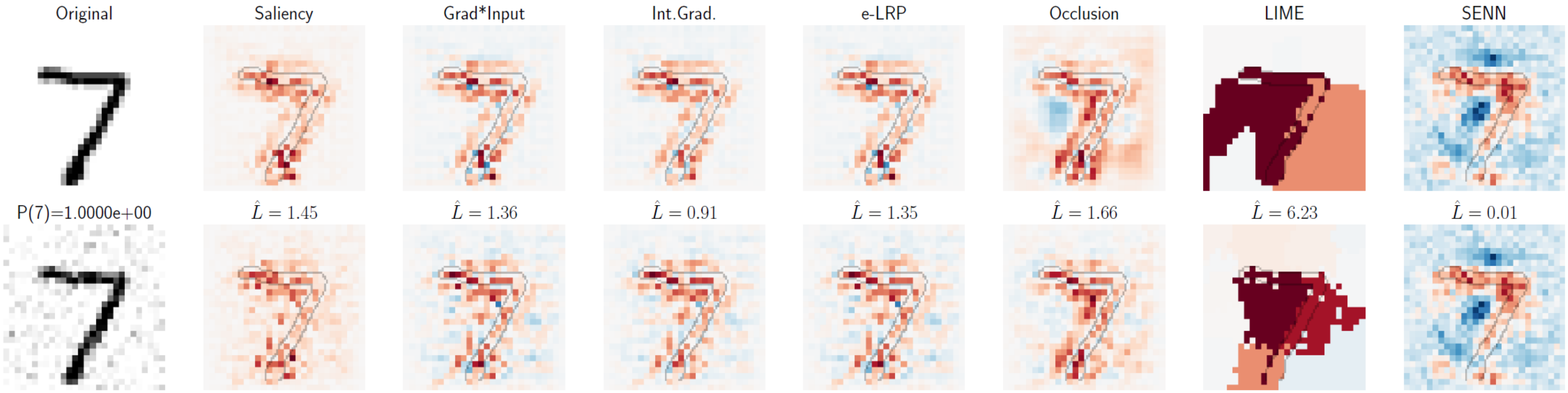


Alvarez Melis, David, and Tommi Jaakkola. "Towards robust interpretability with self-explaining neural networks." Advances in neural information processing systems 31 (2018).

SENN

- MNIST digit *vs* perturbed version with gaussian noise:

The explanations for “post-model” methods vary considerably.

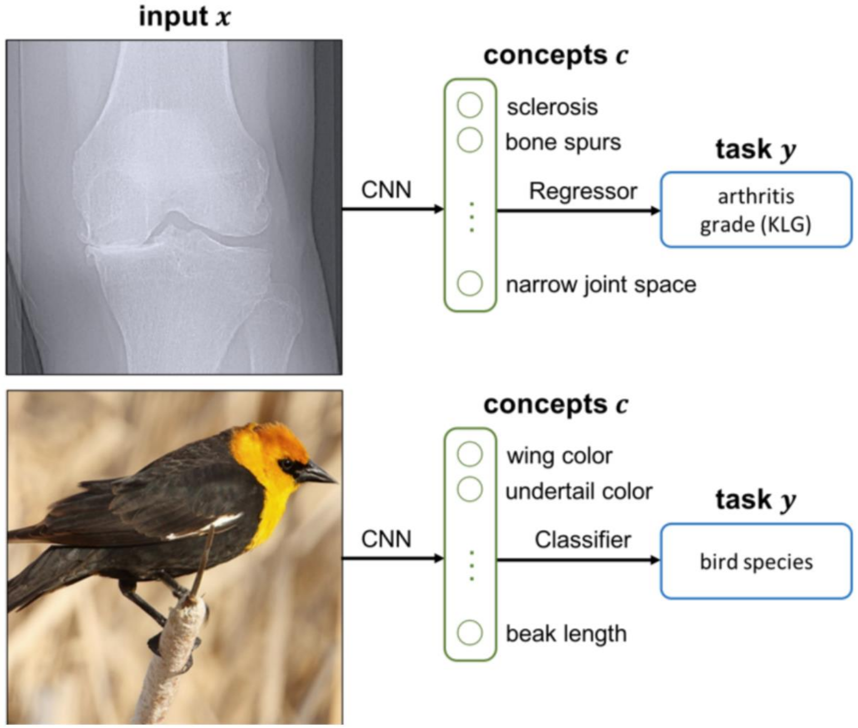


Alvarez Melis, David, and Tommi Jaakkola. "Towards robust interpretability with self-explaining neural networks." Advances in neural information processing systems 31 (2018).

Concept Bottleneck Models

- First predict an intermediate set of human-specified concepts **c**, then use **c** to predict the final output *y*.

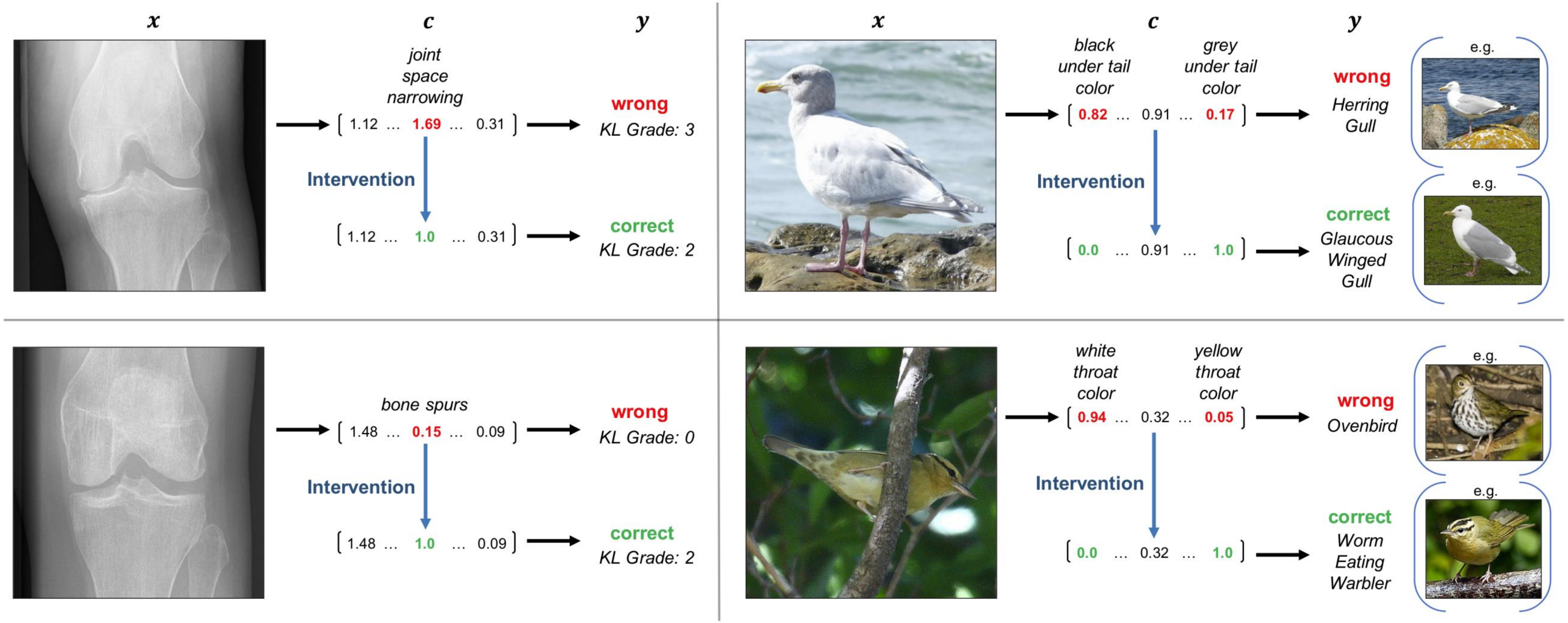
Examples of knee x-ray grading and bird classification.



Koh, Pang Wei, et al. "Concept bottleneck models." International Conference on Machine Learning. PMLR, 2020.

Concept Bottleneck Models

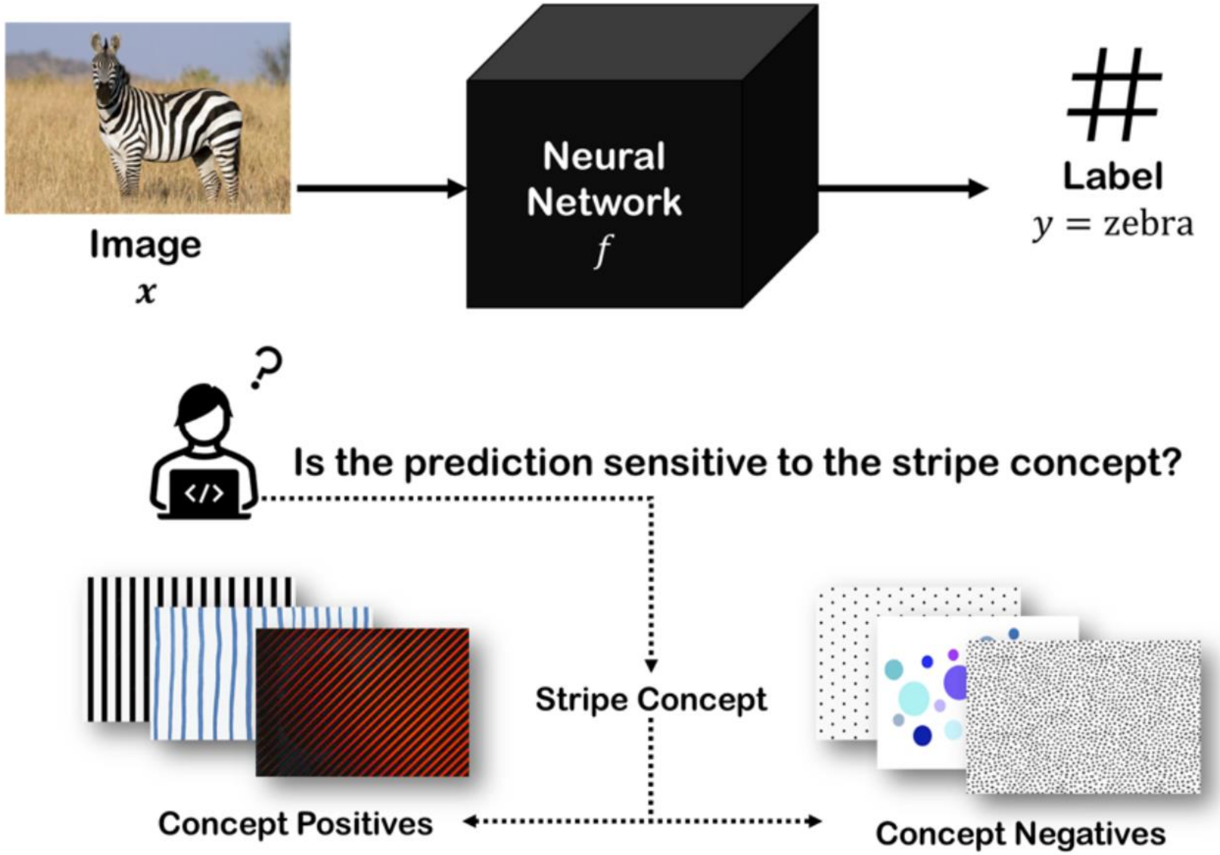
- Intervene the model by editing the predicted concept value.



Koh, Pang Wei, et al. "Concept bottleneck models." International Conference on Machine Learning. PMLR, 2020.

Concept Activation Regions

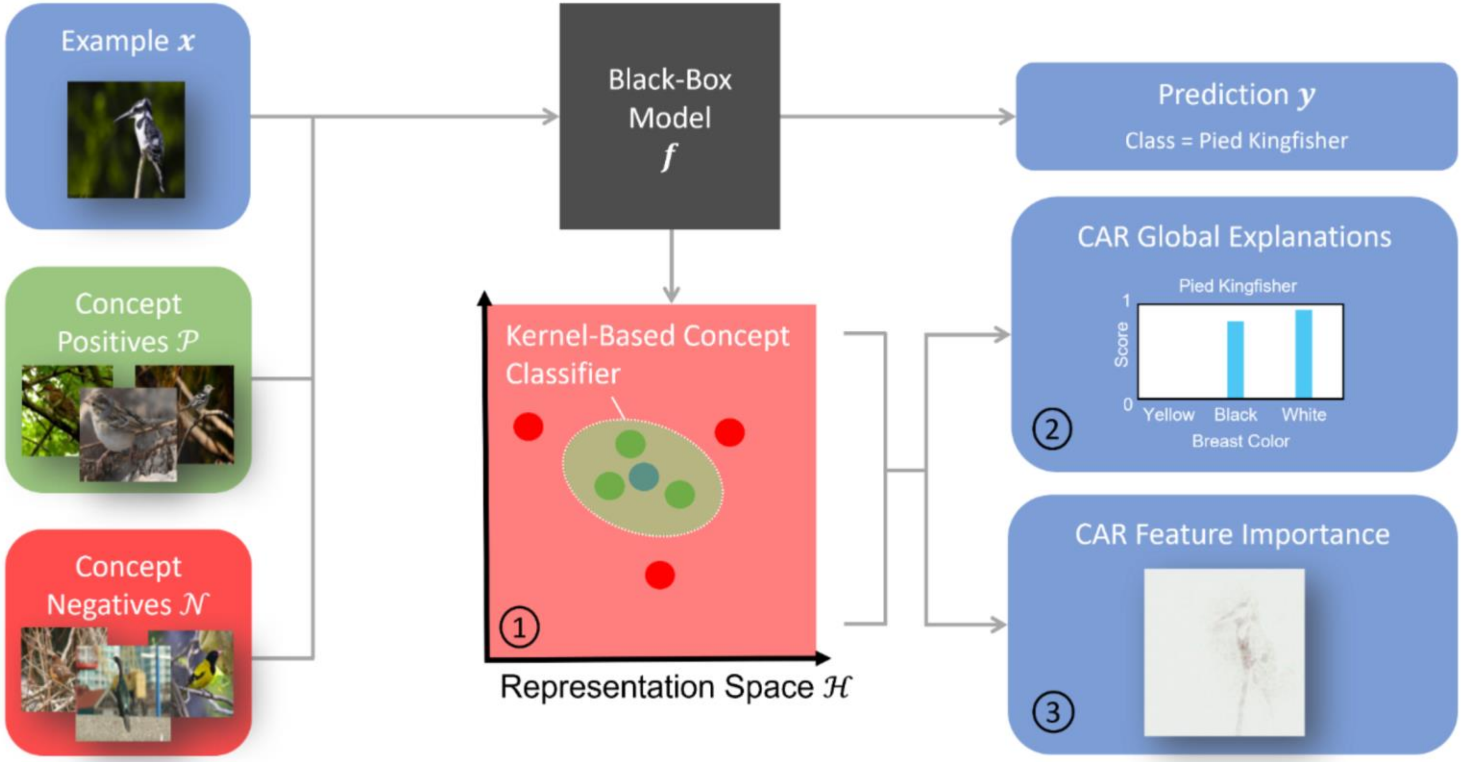
- Each concept is defined by providing positives and negatives.



Crabbé, Jonathan, and Mihaela van der Schaar. Concept Activation Regions: A Generalized Framework For Concept-Based Explanations. NeuIPS2022.

Concept Activation Regions

- Concept activation region discovered by fitting sparse kernel classifiers to distinguish positives and negative examples.



Crabbé, Jonathan, and Mihaela van der Schaar. Concept Activation Regions: A Generalized Framework For Concept-Based Explanations. NeuIPS2022.

Explainability in MIA

- Introduction
- Categories of Explainable AI
- Perturbation-based Methods
- Backpropagation-based Methods
- Concept-based Methods
- Challenges and Future Directions

Challenges and Future Directions

- Most of existing methods target local explainability, i.e., explaining the decisions for a single example. A global explainability of the decisions is needed to analyze the black-box model which may make the right decision due to the wrong reason.
- More quantitative evaluation of explainable AI are yet to be developed.
- Studies have extended the existing explainable methods to fit the requirements in the medical domain. In the future, expert feedback or medical knowledge can be incorporated into the design of such explainability methods, i.e., human-in-the-loop (HITL).
- There is a still long way to go to meet the expectations of end-users, regulators, and the general public.