Transparency of deep neural networks for medical image analysis

A review of interpretability methods

Data Science for Health

20.5.2025

Iris Vukovic
Júlia Virgili Mateu

Summary of deck contents

DISCLAIMER: For this work, we do **not differentiate** between interpretability and explainability.

Interpretability

- an attempt to explain the decision-making process of a deep learning model in a way that is understandable

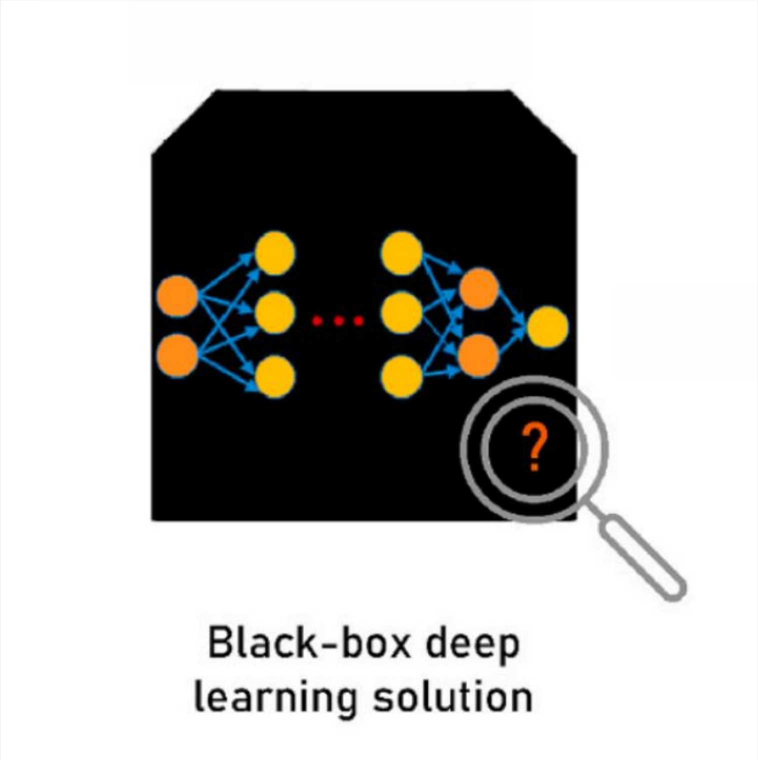
- any technique that attempts to answer the question “Why is the model making this prediction?” for the medical image analysis tasks

Motivation

Need for more **efficient** and **accurate** clinical analysis methods to support clinicians.

Growing amount of available images and technological advancements → **Deep Learning**

Synergy between clinicians and DL lacks **trust** → **Interpretability** as a fundamental approach in method development and deployment

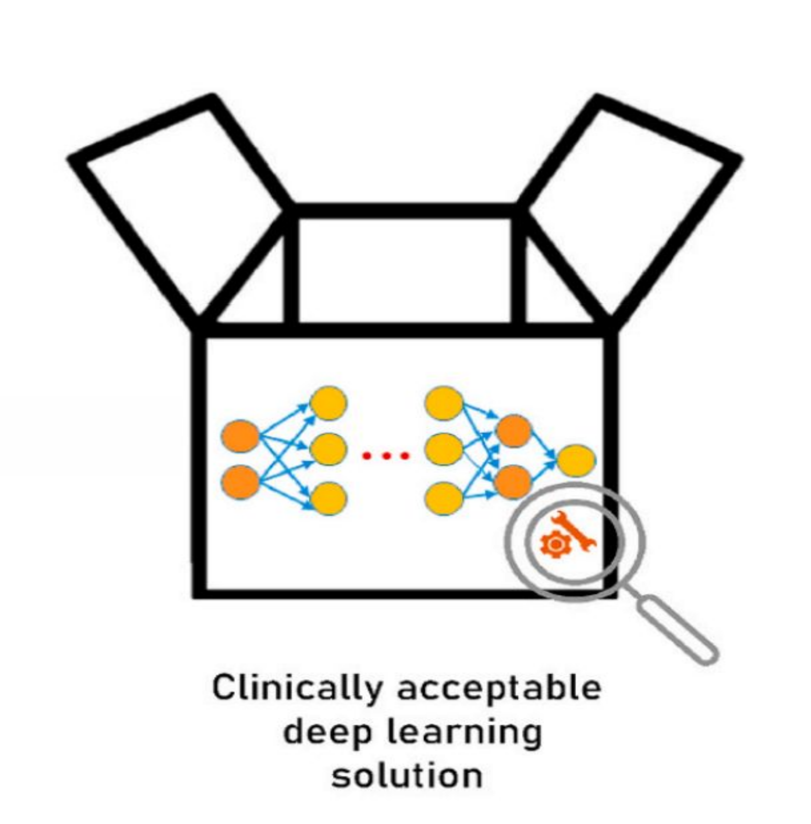


Motivation

Need for more **efficient** and **accurate** clinical analysis methods to support clinicians.

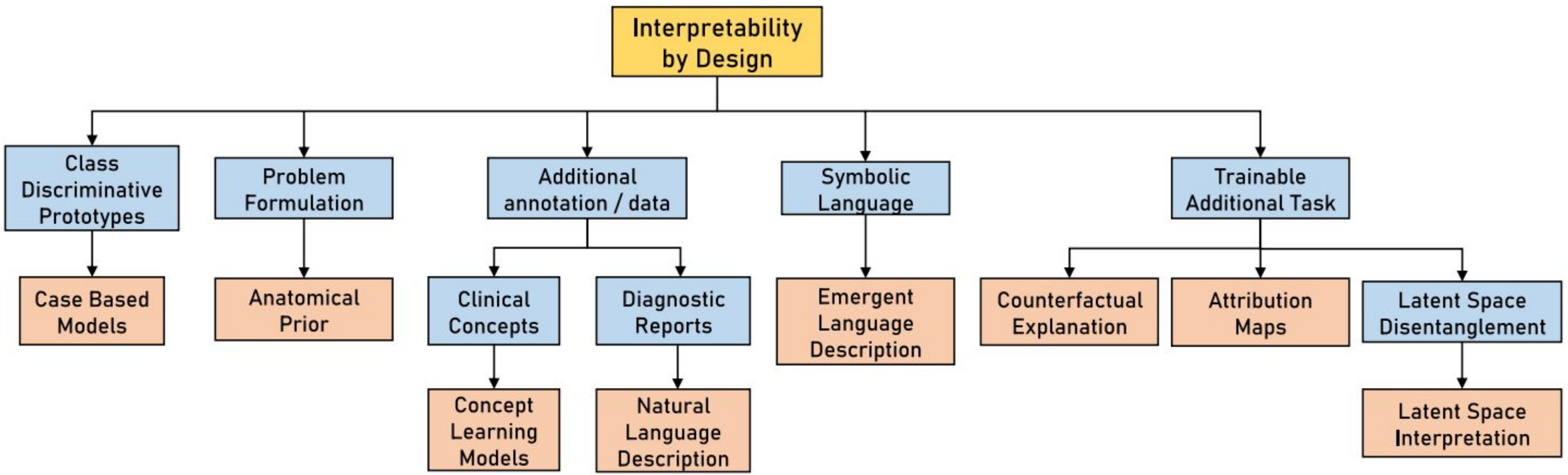
Growing amount of available images and technological advancements → **Deep Learning**

Synergy between clinicians and DL lacks **trust** → **Interpretability** as a fundamental approach in method development and deployment



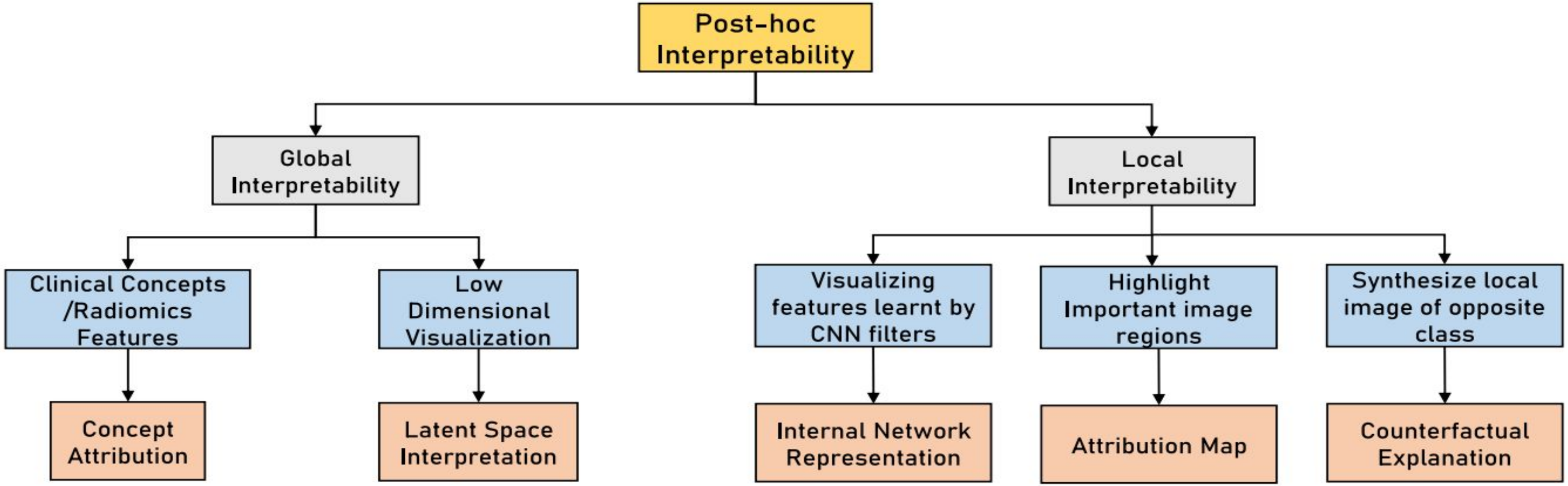
Overview of Methods

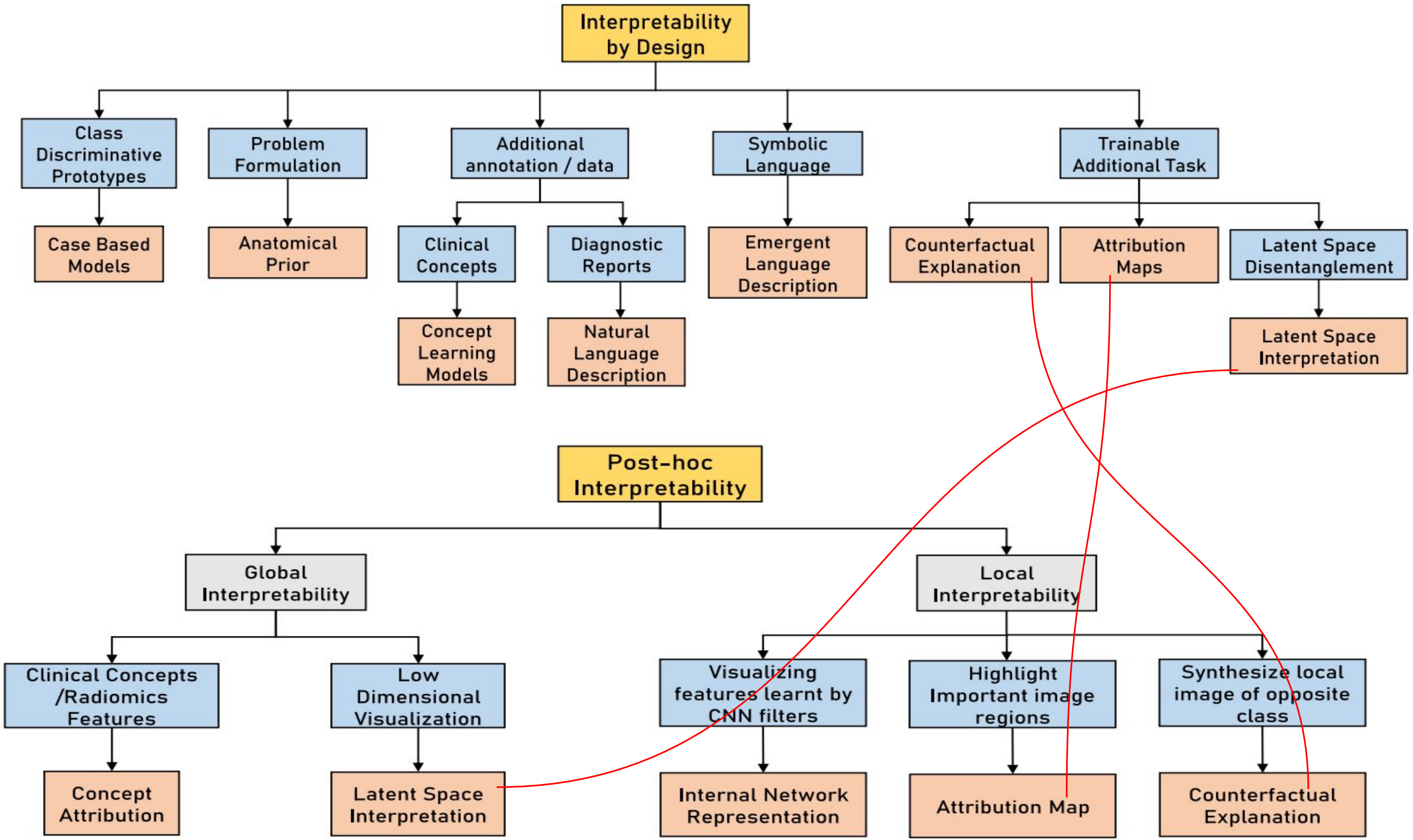
Interpretability can be applied by design in the DL method or as a post-hoc procedure.



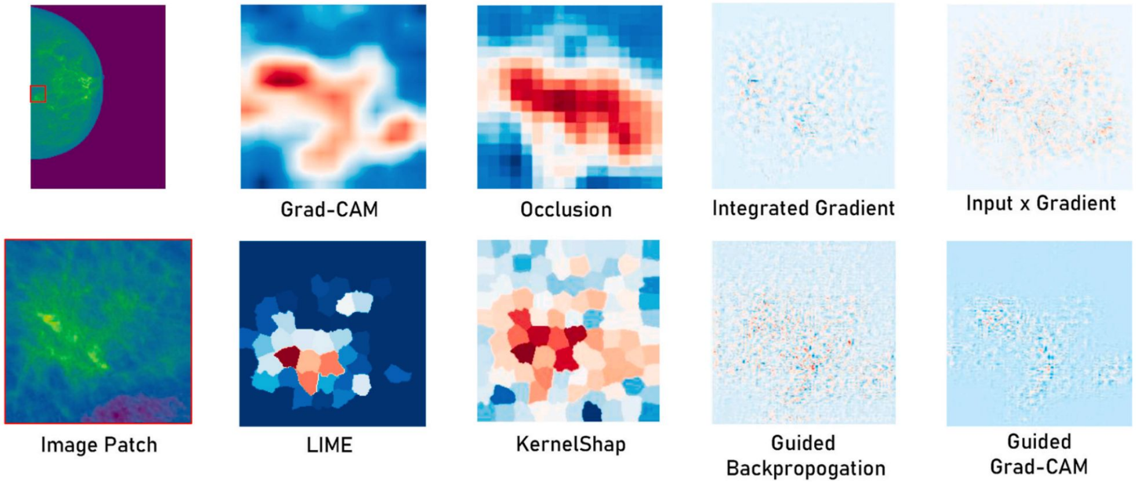
Overview of Methods

Interpretability can be applied by design in the DL method or as a post-hoc procedure.




Interpretability by Design
Class Discriminative Prototypes
Problem Formulation
Additional annotation / data
Symbolic Language
Trainable Additional Task
Case Based Models
Anatomical Prior
Clinical Concepts
Diagnostic Reports
Emergent Language Description
Counterfactual Explanation
Attribution Maps
Latent Space Disentanglement
Concept Learning Models
Natural Language Description
Latent Space Interpretation
Post-hoc Interpretability
Global Interpretability
Local Interpretability
Clinical Concepts /Radiomics Features
Low Dimensional Visualization
Visualizing features learnt by CNN filters
Highlight Important image regions
Synthesize local image of opposite class
Concept Attribution
Latent Space Interpretation
Internal Network Representation
Attribution Map
Counterfactual Explanation

Method 1: Attribution map



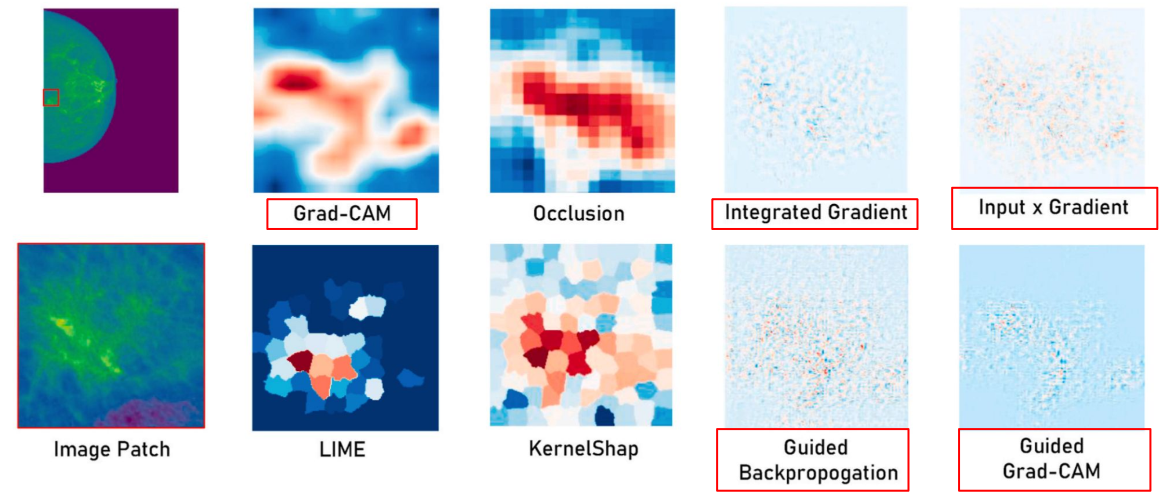
Attribution maps generated by different interpretability methods for explaining a DL model that detects breast mass in mammogram

Highlights regions of the input that are relevant for the prediction

Does not offer information on how these regions contribute to prediction

Many methods related to medical image analysis use attribution maps for interpretability

Method 1: Attribution map

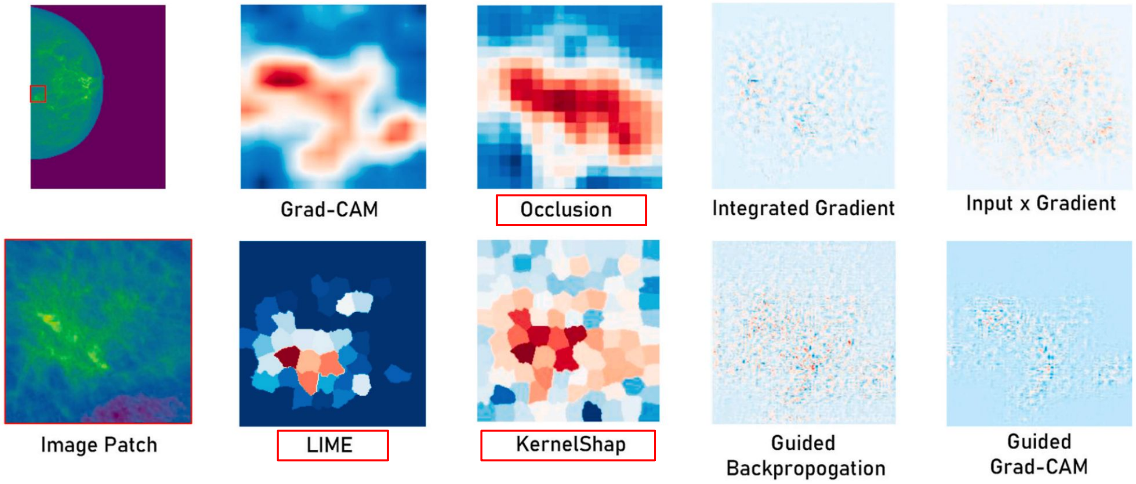


Attribution maps generated by different interpretability methods for explaining a DL model that detects breast mass in mammogram

Gradient-based: generate post-hoc attribution maps by utilizing gradients to identify important parts of input image

Gradient-Class Activation Map (Grad-CAM): localize class-specific image region that are important to model for prediction → gradient of class score w.r.t. each feature map says how sensitive class score is to activations in feature map

Method 1: Attribution map



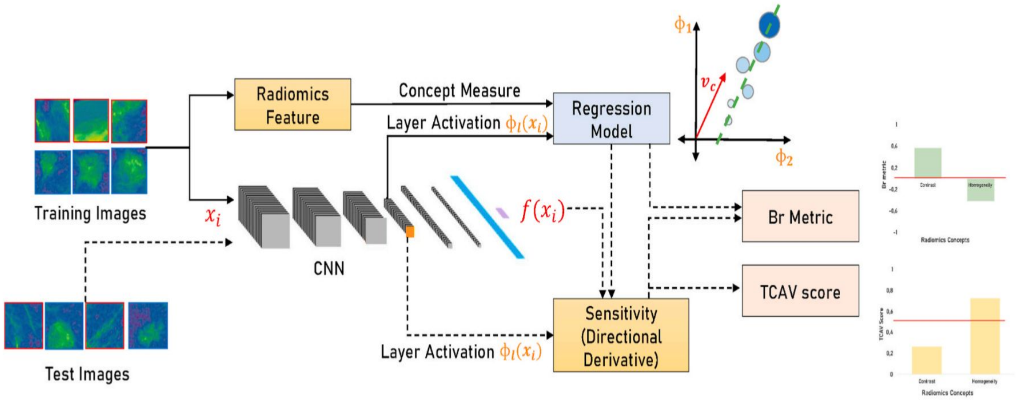
Attribution maps generated by different interpretability methods for explaining a DL model that detects breast mass in mammogram

Perturbation-based: investigate effect of altering different parts of input image on model's prediction

Occlusion: alters image in systematic way to observe the effect on output → altering important parts of the image has a strong effect on the output

Computationally expensive to generate occlusion maps if only small parts of image are perturbed

Method 2: Concept attribution

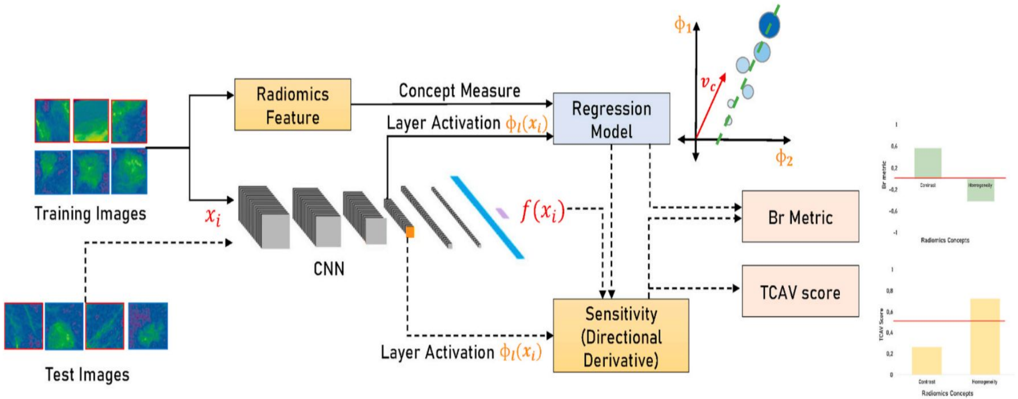


Provides global explanations for DL network in terms of high-level image concepts

Challenging to create a labeled dataset for different concepts

Method 2: Concept attribution: Definitions

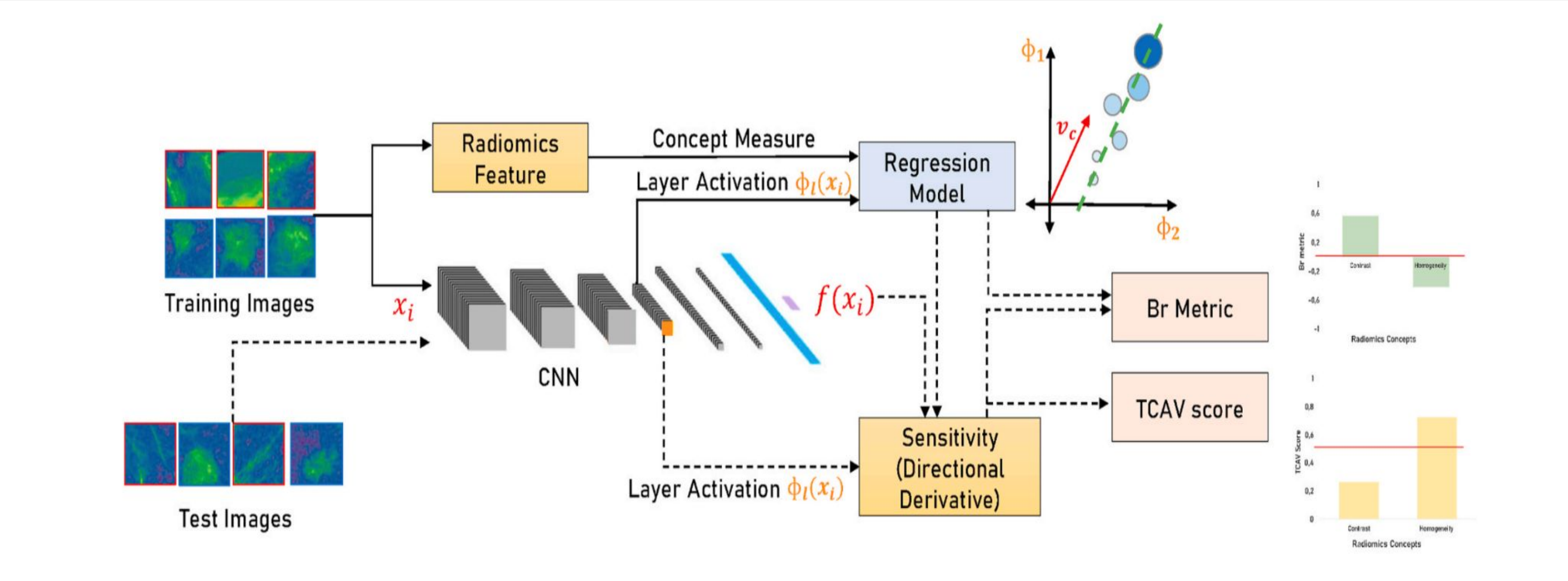
Radiomics: converts medical images into quantitative features which can be used as concepts



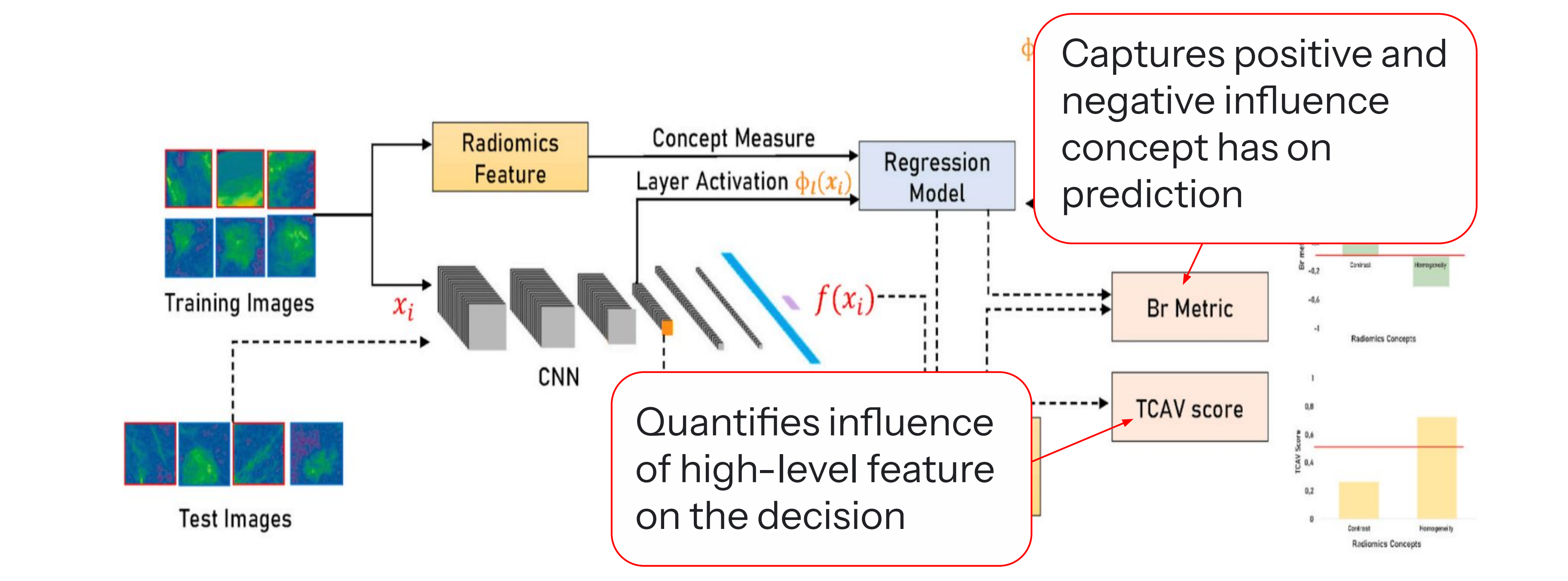
Testing with Concept Activation Vectors (TCAV): linear classifier trained to differentiate between examples contained concept and not → resulting weight vector is a CAV that points in the direction of increasing presence of a concept

Regression Concept Vector: generalization of TCAV that handles continuous value concepts using regression

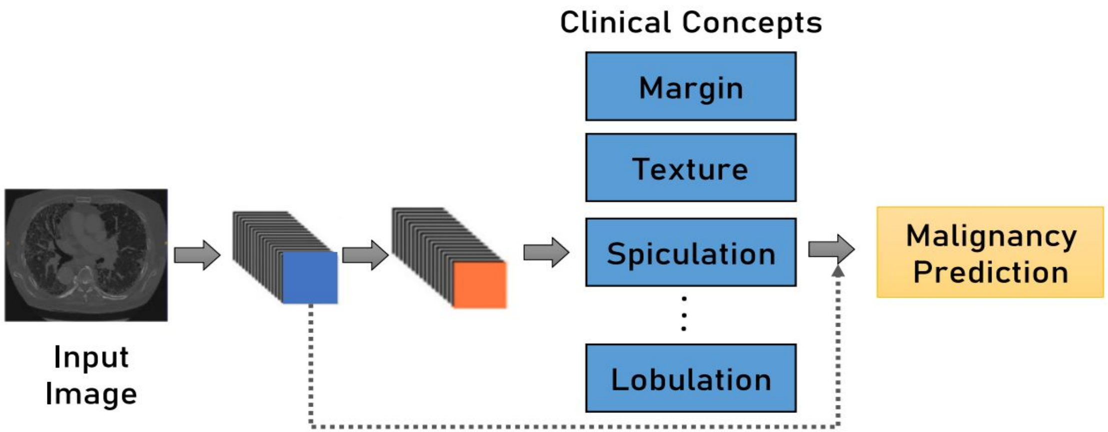
A linear regression model is trained to estimated radiomics features. The influence of radiomics features on the decision of a class is obtained by calculating the directional derivation in the direction of increase of radiomics concepts during testing. The effect of radiomics features is quantified in terms of Bidirectional relevance (Br) metric and Testing with Concept Activation Vectors (TCAV) score.



A linear regression model is trained to estimated radiomics features. The influence of radiomics features on the decision of a class is obtained by calculating the directional derivation in the direction of increase of radiomics concepts during testing. The effect of radiomics features is quantified in terms of Bidirectional relevance (Br) metric and Testing with Concept Activation Vectors (TCAV) score.



Method 3: Concept learning



First predict high-level concepts from input image and then use these concepts for prediction

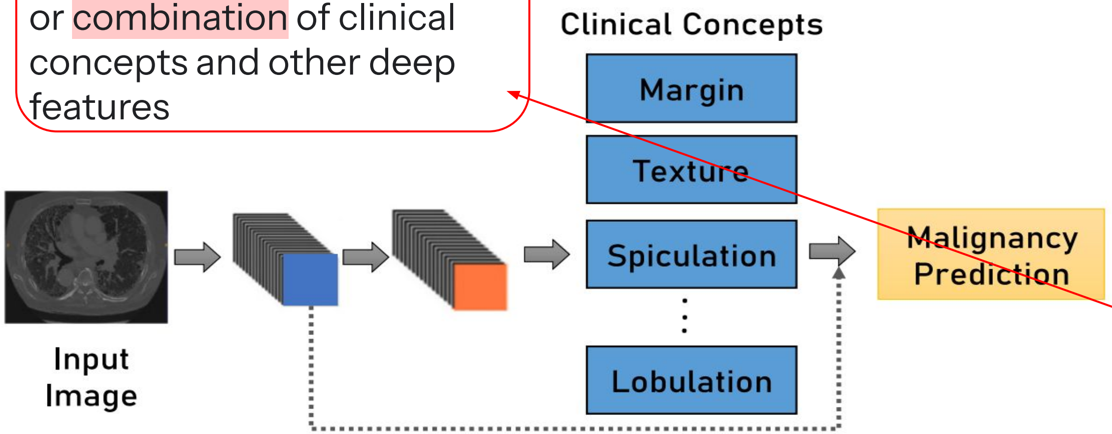
Clinicians can intervene at test time to change predicted value of clinical concept to observe effect on prediction

Annotation of clinical concepts is time-consuming

Can give a false sense of interpretability

Method 3: Concept learning

Final prediction is based on either only clinical concepts or combination of clinical concepts and other deep features



First predict high-level concepts from input image and then use these concepts for prediction

Clinicians can intervene at test time to change predicted value of clinical concept to observe effect on prediction

Annotation of clinical concepts is time-consuming

Can give a false sense of interpretability

Method 4: Counterfactual explanations

$$z = E(X)$$

Encoder

Generate synthetic image from **perturbed latent *z***

$$X' = G(z + \lambda \cdot \alpha)$$

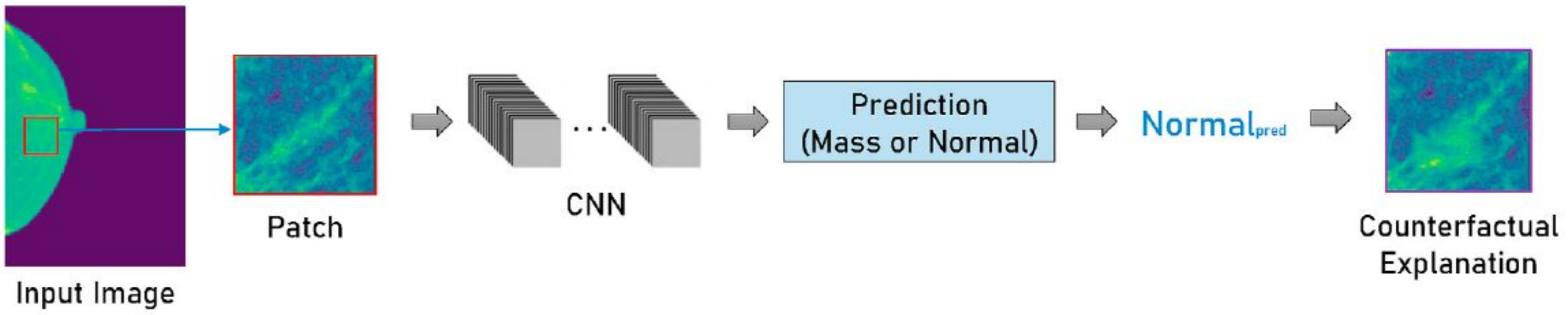
f' : f classifier for maximum change in class prediction

To apply the minimum perturbation to the input image such that we get the maximum change in output: prediction class switch.

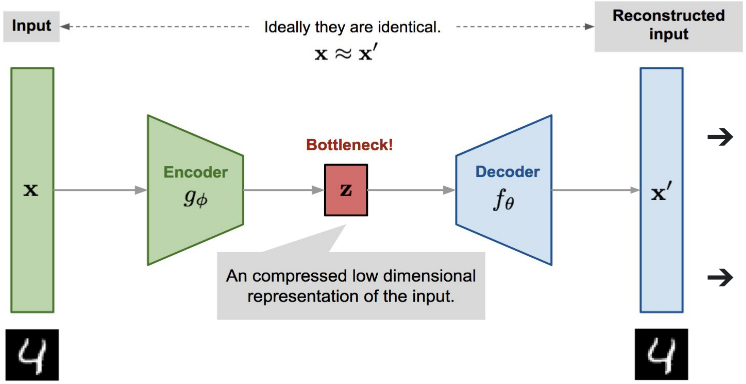
Counterfactual images synthesized via:

- GANs' generator **or**
- Autoencoder's

latent space perturbation

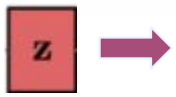


Method 5:
Latent space interpretation

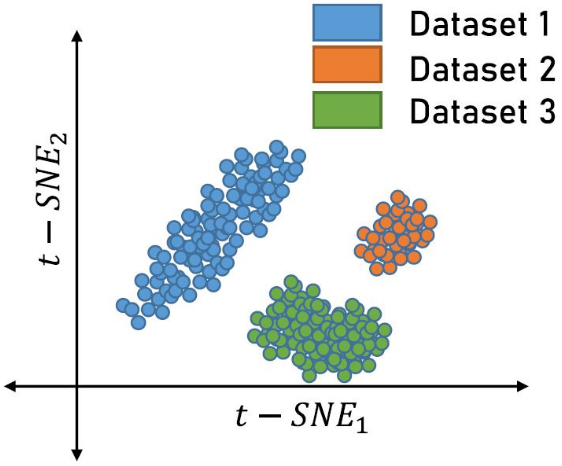


➔ Classification through fully-connected layers, or

➔ Segmentation (CNN)



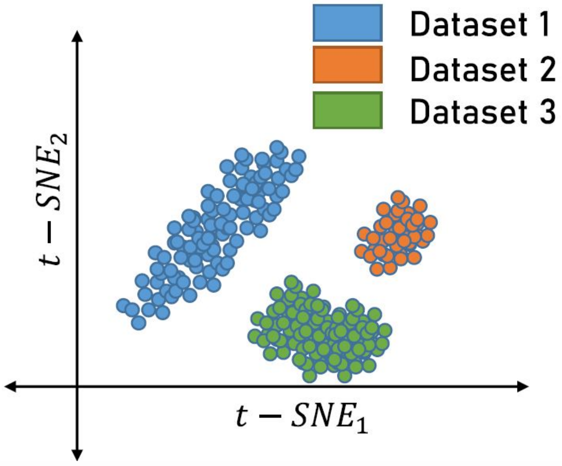
Regular VAE latent space versus **interpretable latent space** ?



Dimensionality reduction methods for visualization:

- **Linear**: PCA
- **Non-linear**: t-distributed Stochastic Neighbor Embedding or t-SNE

Example: dataset bias in the DL model due to difference in acquisition



Method 5: Latent space interpretation

Disentanglement of latent space:

Explicitly aims to **separate out distinct, independent salient factors of variation** in the data.

Each latent dimension ideally corresponds to a **specific, meaningful concept** (e.g., tumor size, organ shape, imaging modality).

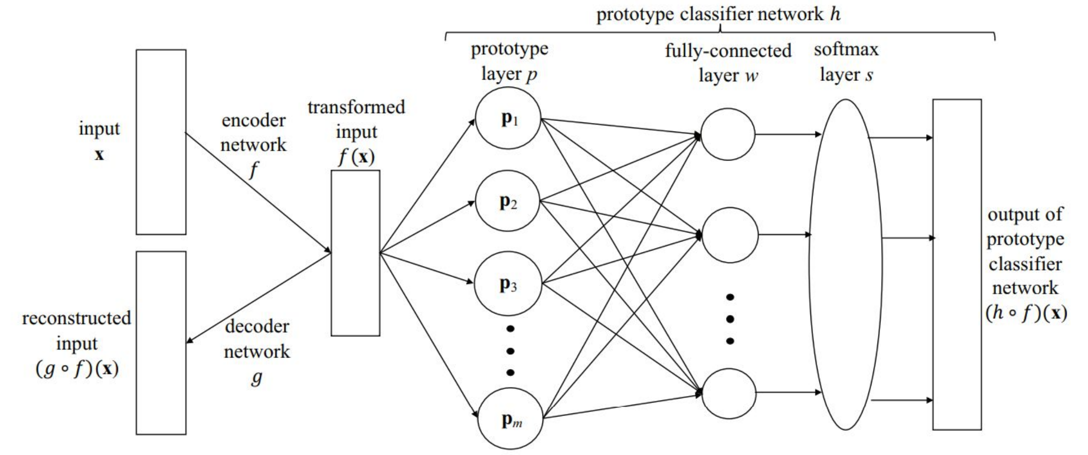
How: Training objectives penalize feature correlation or encourage axis alignment with known factors (predefined).

Dimensionality reduction methods for visualization:

- **Linear**: PCA
- **Non-linear**: t-distributed Stochastic Neighbor Embedding or t-SNE

Method 6: Case-Based Models

Inherently interpretable because the final predictions are made by taking a weighted sum of similarity scores between features extracted from input and class-discriminative prototypes.



m different prototypes

Probability distribution across all classes

Step 1: Learn feature maps that represent **prototypes** from training data.

Step 2: Extract feature maps from input and compute **similarity scores** for each prototype.

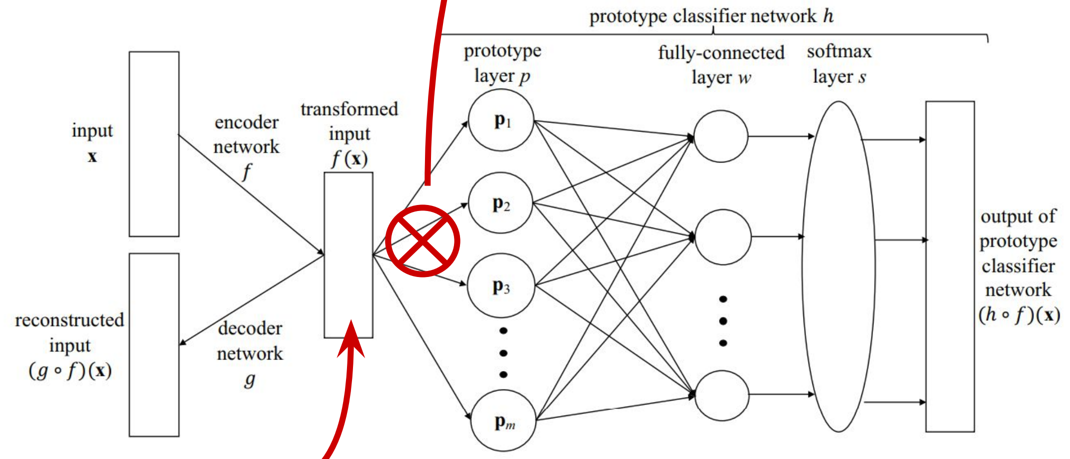
Step 3: Compute a **weighted sum** of the aforementioned scores through a fully connected layer.

Step 4: Make the final **prediction**.

Compute similarity between feature maps **within the latent space**

Method 6: Case-Based Models

Inherently interpretable because the final predictions are made by taking a weighted sum of similarity scores between features extracted from input and class-discriminative prototypes.



feature vector for input **x**

m different prototypes

Probability distribution across all classes

Step 1: Learn feature maps that represent **prototypes** from training data.

Step 2: Extract feature maps from input and compute **similarity scores** for each prototype.

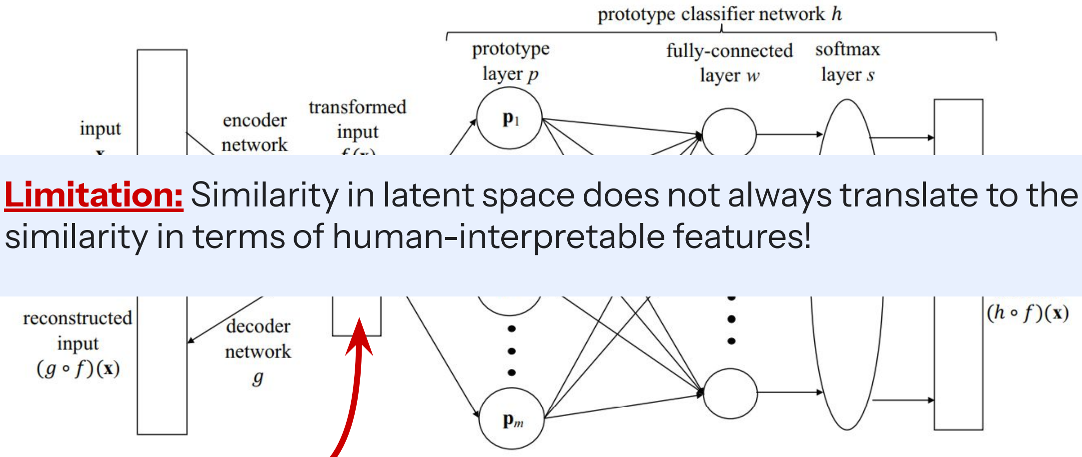
Step 3: Compute a **weighted sum** of the aforementioned scores through a fully connected layer.

Step 4: Make the final **prediction** and **output reasoning weights**.

Loss incorporates 4 terms: classification accuracy, decoder reconstruction and prototype quality terms R1 and R2

Method 6: Case-Based Models

Inherently interpretable because the final predictions are made by taking a weighted sum of similarity scores between features extracted from input and class-discriminative prototypes.



Limitation: Similarity in latent space does not always translate to the similarity in terms of human-interpretable features!

feature vector for input **x**

m different prototypes

Probability distribution across all classes

Step 1: Learn feature maps that represent **prototypes** from training data.

Step 2: Extract feature maps from input and compute **similarity scores** for each prototype.

Step 3: Compute a **weighted sum** of the aforementioned scores through a fully connected layer.

Step 4: Make the final **prediction**.

Evaluation methods

How to quantify explanation quality?

01

Application-grounded

Involve experts for a specific application (e.g.medical diagnosis - doctor)

02

Human-grounded

Human test for the general quality of explanations.

03

Functionality-grounded

Proxy tasks instead of humans to get evaluation metrics that do not involve human interaction. (Desirable due to time and cost constraints.)

Evaluation methods

How to quantify explanation quality?

Difficult to evaluate because there is no ground truth for explanations

01

Application-grounded

Involve experts for a specific application (e.g.medical diagnosis - doctor)

02

Human-grounded

Human test for the general quality of explanations.

03

Functionality-grounded

Proxy tasks instead of humans to get evaluation metrics that do not involve human interaction. (Desirable due to time and cost constraints.)

Conclusions and Takeaways

Incorporation of deep neural networks in the clinical workflow for medical image analysis is held back by the vague understanding of their decision-making process

Both quantitative and qualitative evaluations are necessary to ensure trustworthy explanations

Post-hoc interpretability methods should be utilized carefully as they approximate model behavior and can instill a false sense of confidence

Future directions

→ Case-based models and concept learning models are interpretable by design and have performance similar to black-box CNNs

→ Sanity checks for attribution maps to ensure robustness

→ Multimodal data (images + text + genomics) can increase performance and enhance interpretability

→ Combine human-centered evaluations and quantitative functionality-based evaluations

Thank you