Explainable AI

(In medical image processing)

A Comprehensive Taxonomy for Explainable Artificial Intelligence: A Systematic Survey of Surveys on Methods and Concepts

Gesina Schwalbe[1,2*†] and Bettina Finzel[2†]

[1*]Continental AG, Regensburg, Germany.
[2]Cognitive Systems Group, University of Bamberg, Bamberg, Germany.

*Corresponding author(s). E-mail(s): gesina.schwalbe@continental-corporation.com;
Contributing authors: bettina.finzel@uni-bamberg.de;
†These authors contributed equally to this work.

Abstract

In the meantime, a wide variety of terminologies, motivations, approaches, and evaluation criteria have been developed within the research field of explainable artificial intelligence (XAI). With the amount of XAI methods vastly growing, a taxonomy of methods is needed by researchers as well as practitioners: To grasp the breadth of the topic, compare methods, and to select the right XAI method based on traits required by a specific use-case context. Many taxonomies for XAI methods of varying level of detail and depth can be found in the literature. While they often have a different focus, they also exhibit many points of overlap. This paper unifies these efforts and provides a complete taxonomy of XAI methods with respect to notions present in the current state of research. In a structured literature analysis and meta-study, we identified and reviewed more than 50 of the most cited and current surveys on XAI methods, metrics, and method traits. After summarizing them in a

Expert Systems With Applications 213 (2023) 118888

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Quod erat demonstrandum? - Towards a typology of the concept of explanation for the design of explainable AI

Federico Cabitza [a,b], Andrea Campagner [a], Gianclaudio Malgieri [c,d], Chiara Natali [a], David Schneeberger [e], Karl Stoeger [e], Andreas Holzinger [f,*]

[a] DISCo, University of Milano-Bicocca, viale Sarca 336, Milano, 20126, Italy
[b] IRCCS Orthopedic Institute Galeazzi, via Galeazzi, 4, Milano, 20161, Italy
[c] Augmented Law Institute, EDHEC Business School, 24 avenue Gustave Delory, CS 50411, Roubaix Cedex 1, 59057, France
[d] eLaw, Leiden University, Rapenburg 70, Leiden, 2311 EZ, Netherlands
[e] University of Vienna, Schottenbastei 10-16, Vienna, 1010, Austria
[f] University of Natural Resources and Life Sciences Vienna, Peter Jordan Straße 82, Vienna, 1190, Austria

ARTICLE INFO

Keywords:
Explainable AI
XAI
Explanations
Taxonomy
Artificial intelligence
Machine learning

ABSTRACT

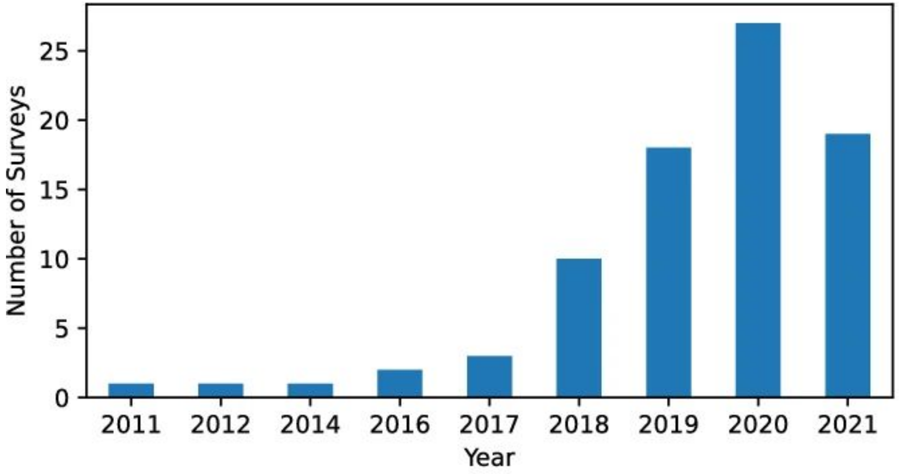
In this paper, we present a fundamental framework for defining different types of explanations of AI systems and the criteria for evaluating their quality. Starting from a structural view of how explanations can be constructed, i.e., in terms of an explanandum (what needs to be explained), multiple explanantia (explanations, clues, or parts of information that explain), and a relationship linking explanandum and explanantia, we propose an explanandum-based typology and point to other possible typologies based on how explanantia are presented and how they relate to explanandia. We also highlight two broad and complementary perspectives for defining possible quality criteria for assessing explainability: epistemological and psychological (cognitive). These definition attempts aim to support the three main functions that we believe should attract the interest and further research of XAI scholars: clear inventories, clear verification criteria, and clear validation methods.

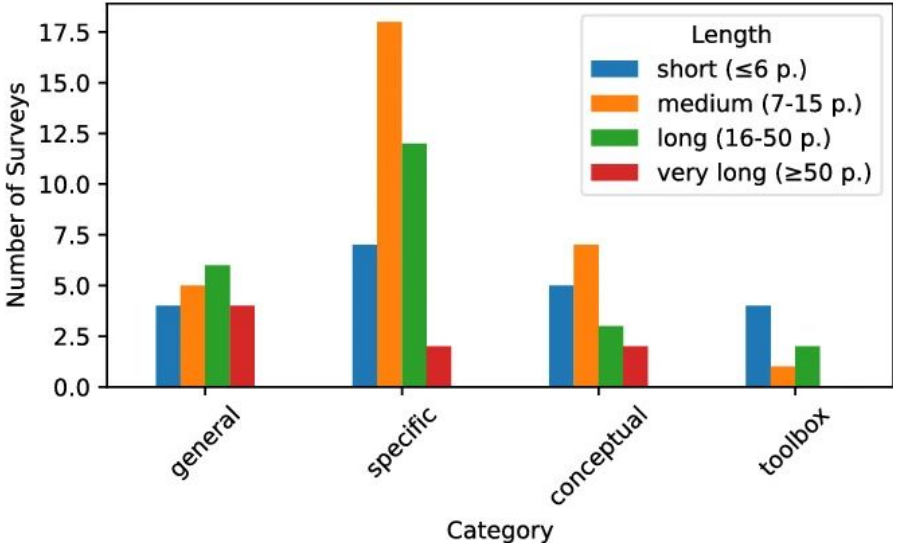
1. Introduction

It is well-known and easily verifiable that the interest in artificial intelligence has grown almost exponentially both in academic research and professional practice (Johnson, Albizri, Harfouche, & Fosso-Wamba, 2022). This is also mirrored by the increasing number of articles that mention this broad expression in the last 10 years. A similar trend can be observed as for the number of articles that talk about a specific feature of AI systems, that is *explainability* (see Fig. 1). This is probably due to the fact that the computational paradigm legal proceedings, where the output of a black box AI system could pose severe risks and consequences for the involved users.

However, in this contribution we will not speculate about the concept of explainability: for the sake of argument, here we simply intend explainability as *the requirement (and related capability) to associate a proper explanation to the output of an AI system*. This requirement can virtually be addressed by anyone, the so called explainer, but it is usually assigned to the AI system itself (or to one of its components, usually called *explanator* Vilone & Longo, 2021), which then becomes

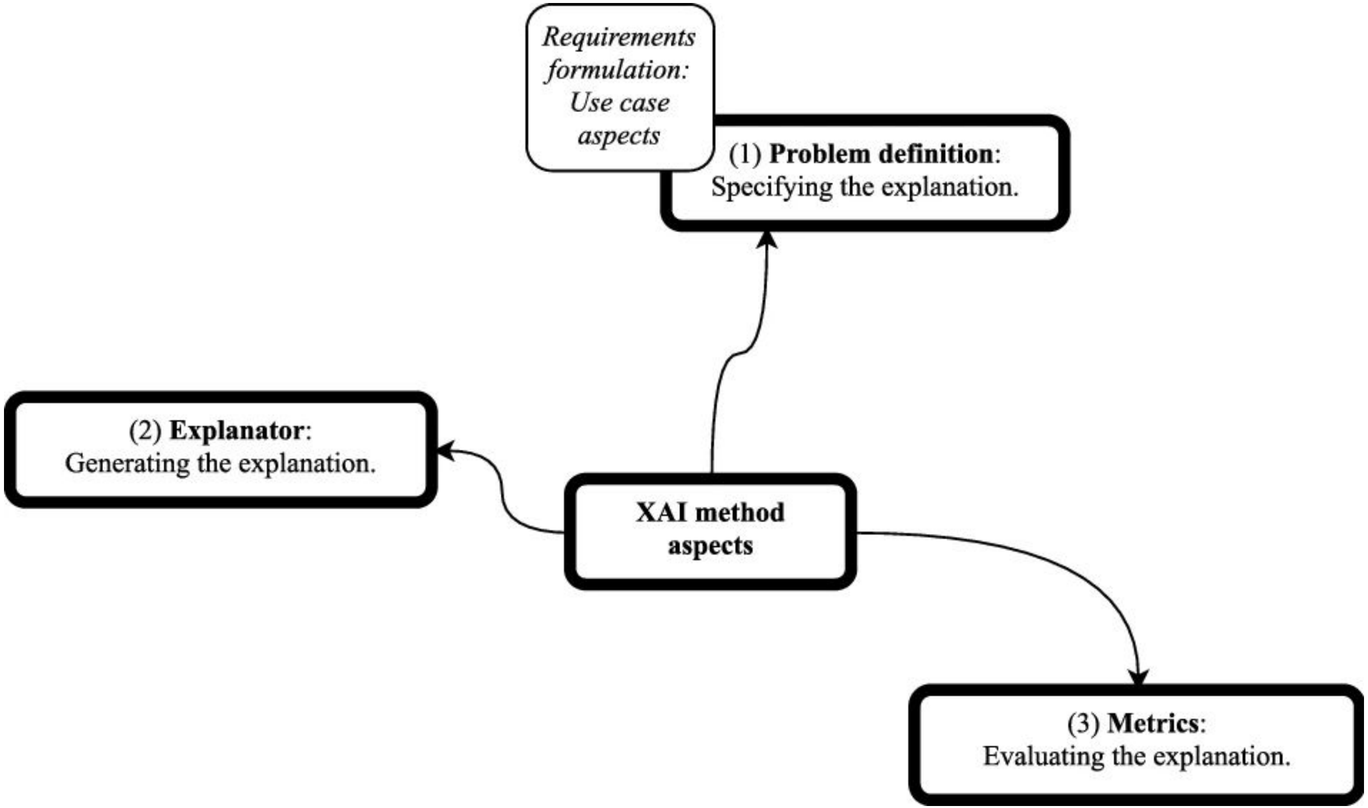
Surveys of XAI

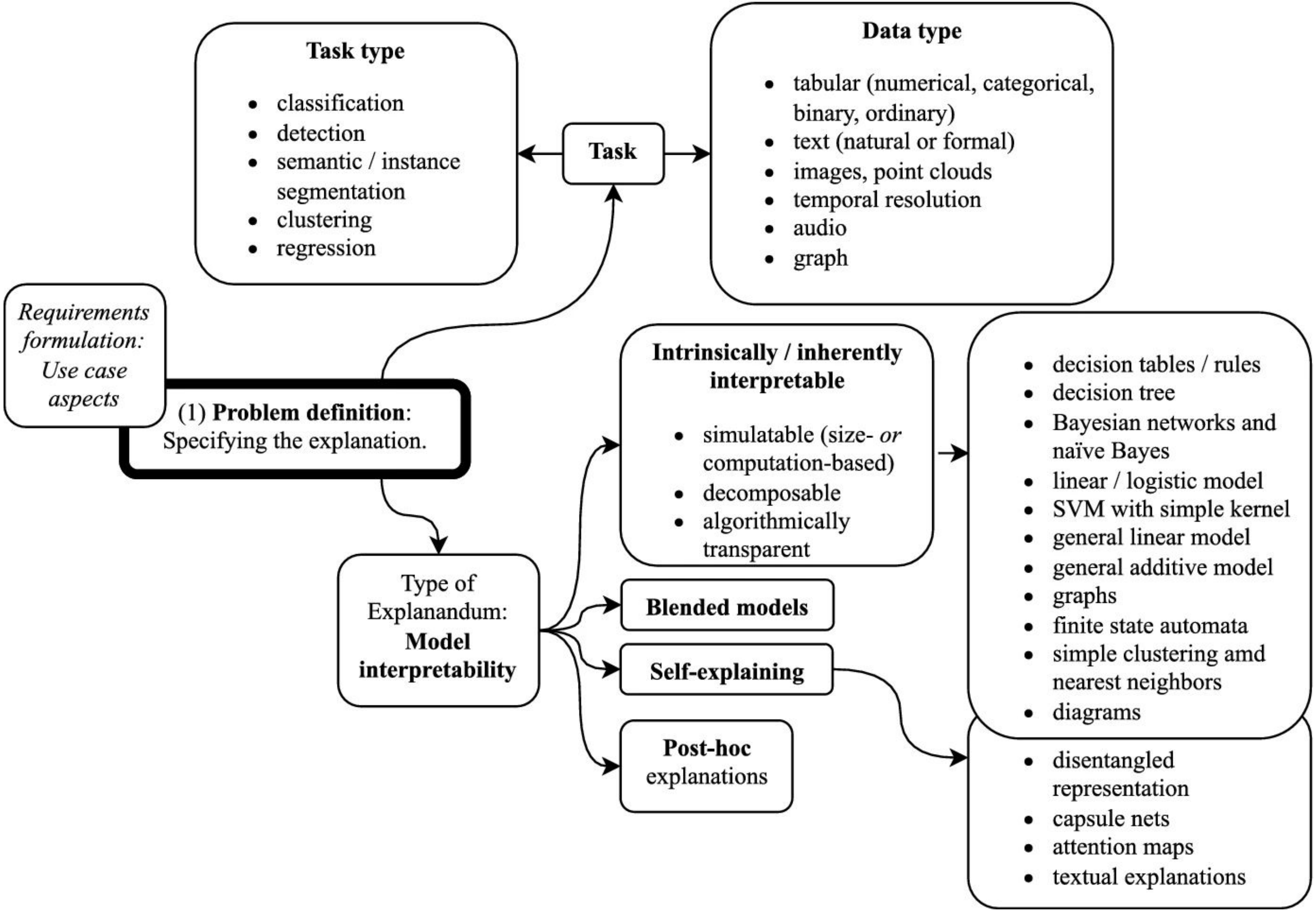




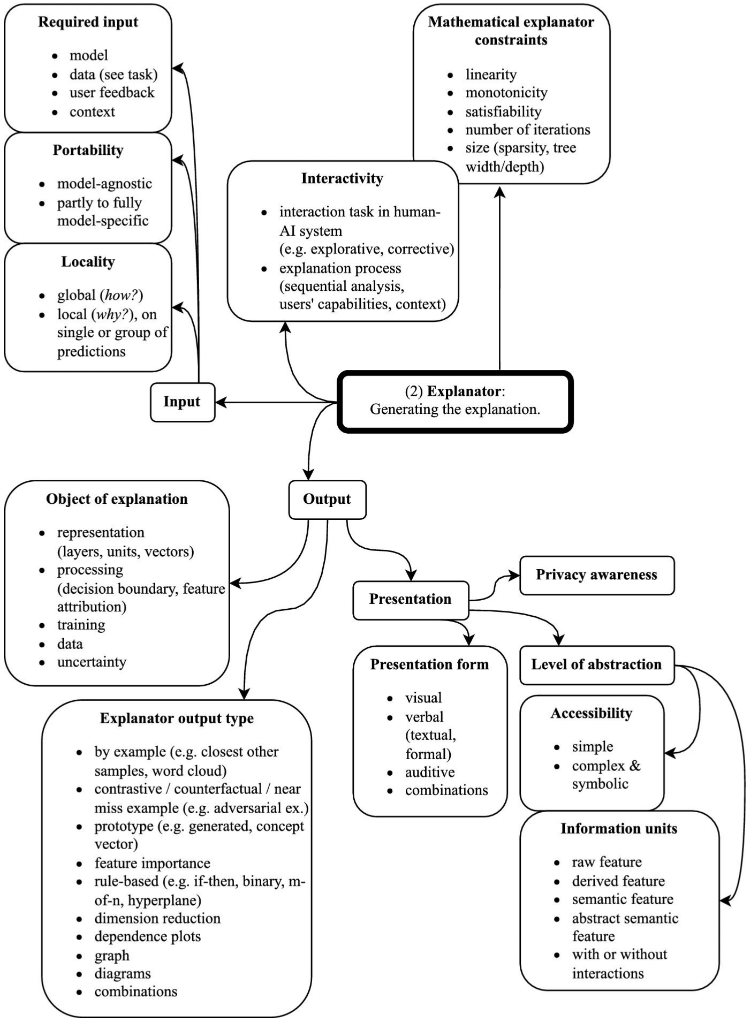
A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts, Schwalbe and Finzel, Data Mining and Knowledge Discovery, 2023

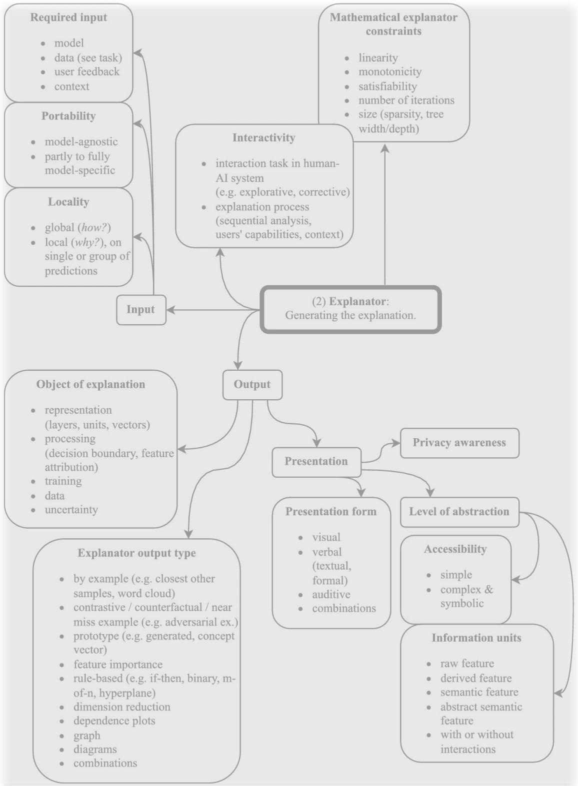
XAI overview

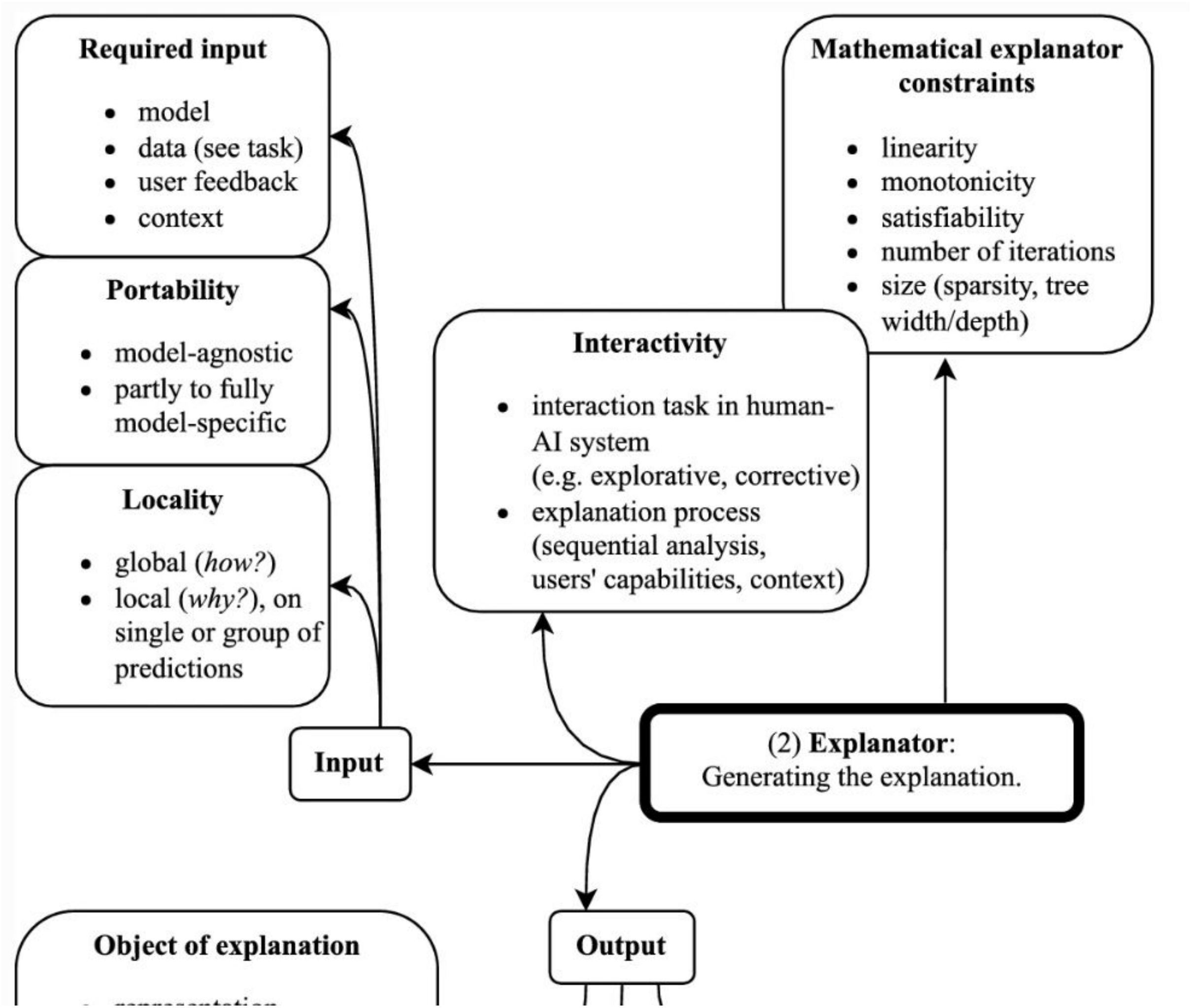


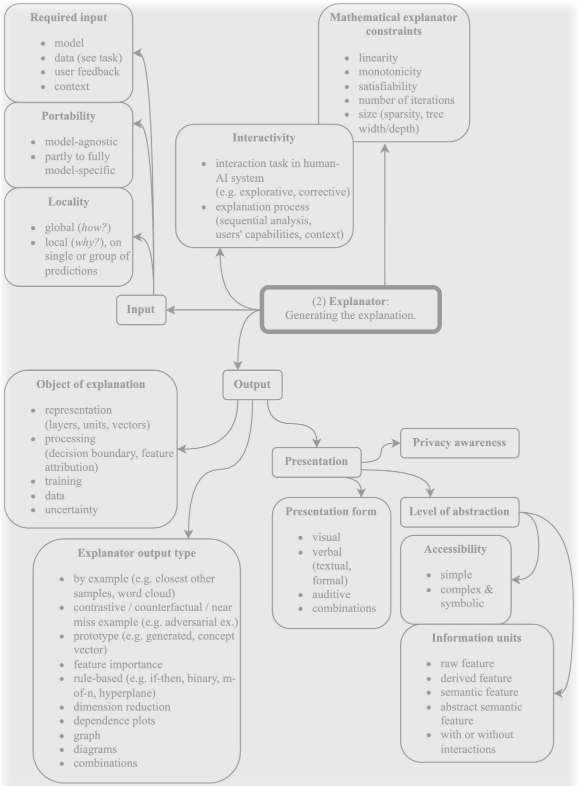

Task type
classification
detection
semantic / instance segmentation
clustering
regression
Task
Data type
tabular (numerical, categorical, binary, ordinary)
text (natural or formal)
images, point clouds
temporal resolution
audio
graph
Requirements formulation: Use case aspects
(1) Problem definition: Specifying the explanation.
Type of Explanandum: Model interpretability
Intrinsically / inherently interpretable
simulatable (size- or computation-based)
decomposable
algorithmically transparent
decision tables / rules
decision tree
Bayesian networks and naïve Bayes
linear / logistic model
SVM with simple kernel
general linear model
general additive model
graphs
finite state automata
simple clustering amd nearest neighbors
diagrams
Blended models
Self-explaining
disentangled representation
capsule nets
attention maps
textual explanations
Post-hoc explanations

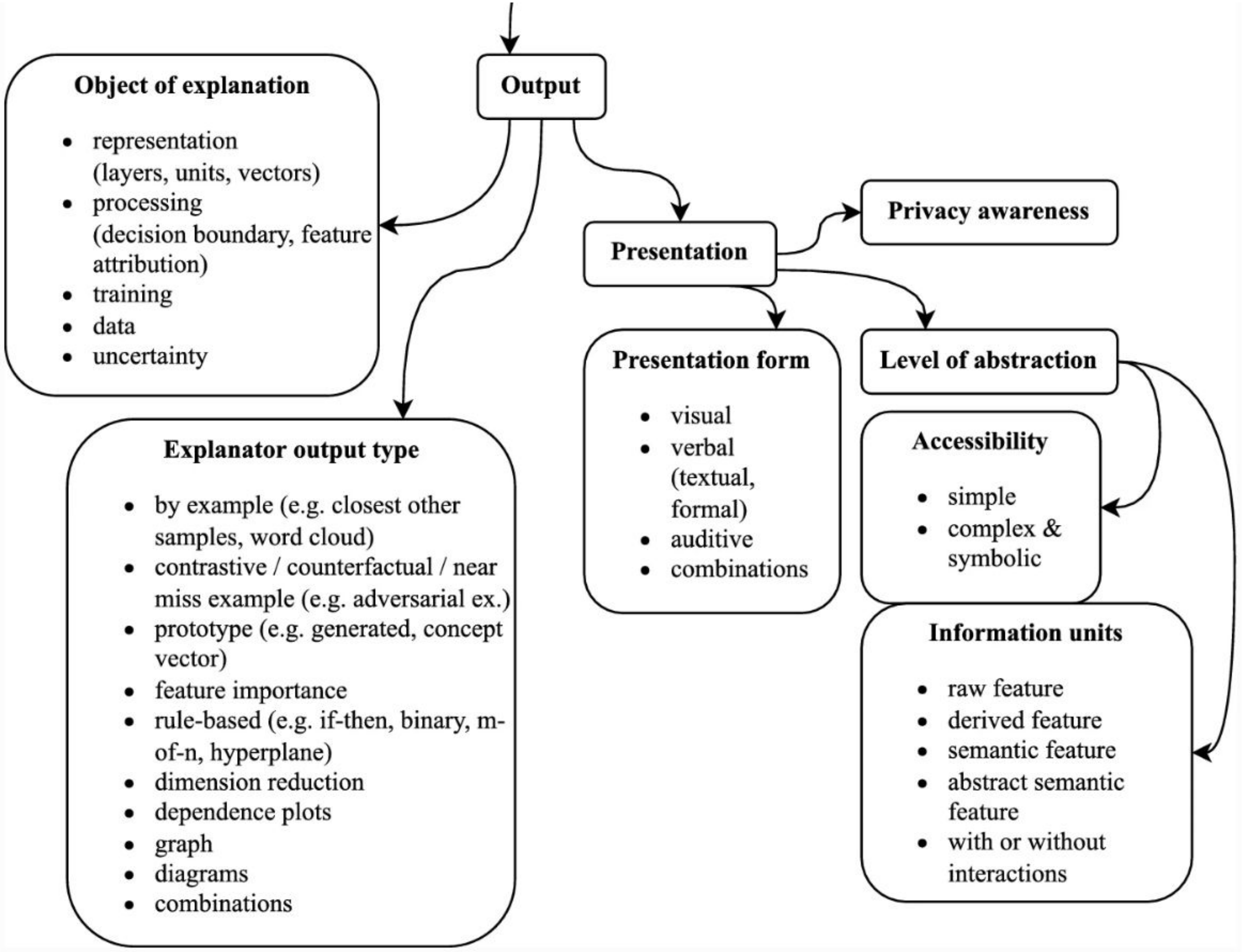
XAI Explanator

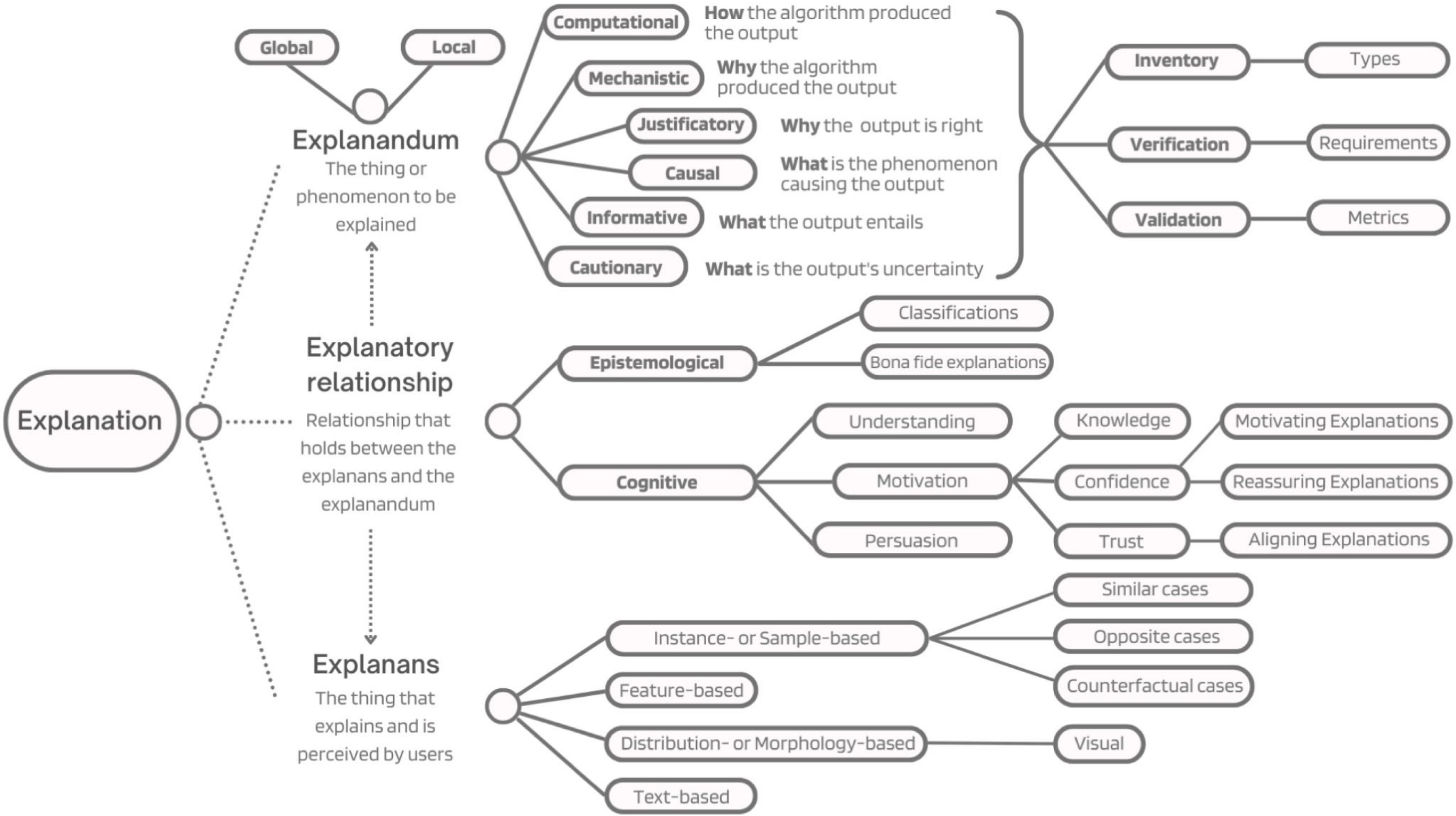



Required input
model
data (see task)
user feedback
context
Mathematical explanator constraints
linearity
monotonicity
satisfiability
number of iterations
size (sparsity, tree width/depth)
Portability
model-agnostic
partly to fully model-specific
Interactivity
interaction task in human-AI system (e.g. explorative, corrective)
explanation process (sequential analysis, users' capabilities, context)
Locality
global (how?)
local (why?), on single or group of predictions
Input
(2) Explanator: Generating the explanation.
Object of explanation
representation (layers, units, vectors)
processing (decision boundary, feature attribution)
training
data
uncertainty
Output
Privacy awareness
Presentation
Presentation form
visual
verbal (textual, formal)
auditive
combinations
Level of abstraction
Accessibility
simple
complex & symbolic
Explanator output type
by example (e.g. closest other samples, word cloud)
contrastive / counterfactual / near miss example (e.g. adversarial ex.)
prototype (e.g. generated, concept vector)
feature importance
rule-based (e.g. if-then, binary, m-of-n, hyperplane)
dimension reduction
dependence plots
graph
diagrams
combinations
Information units
raw feature
derived feature
semantic feature
abstract semantic feature
with or without interactions


Required input
model
data (see task)
user feedback
context
Mathematical explanator constraints
linearity
monotonicity
satisfiability
number of iterations
size (sparsity, tree width/depth)
Portability
model-agnostic
partly to fully model-specific
Interactivity
interaction task in human-AI system (e.g. explorative, corrective)
explanation process (sequential analysis, users' capabilities, context)
Locality
global (how?)
local (why?), on single or group of predictions
Input
(2) Explanator: Generating the explanation.
Object of explanation
Output


Required input
model
data (see task)
user feedback
context
Portability
model-agnostic
partly to fully model-specific
Locality
global (how?)
local (why?), on single or group of predictions
Input
Mathematical explanator constraints
linearity
monotonicity
satisfiability
number of iterations
size (sparsity, tree width/depth)
Interactivity
interaction task in human-AI system (e.g. explorative, corrective)
explanation process (sequential analysis, users' capabilities, context)
(2) Explanator: Generating the explanation.
Output
Object of explanation
representation (layers, units, vectors)
processing (decision boundary, feature attribution)
training
data
uncertainty
Presentation
Privacy awareness
Presentation form
visual
verbal (textual, formal)
auditive
combinations
Level of abstraction
Accessibility
simple
complex & symbolic
Information units
raw feature
derived feature
semantic feature
abstract semantic feature
with or without interactions
Explanator output type
by example (e.g. closest other samples, word cloud)
contrastive / counterfactual / near miss example (e.g. adversarial ex.)
prototype (e.g. generated, concept vector)
feature importance
rule-based (e.g. if-then, binary, m-of-n, hyperplane)
dimension reduction
dependence plots
graph
diagrams
combinations


Output
Object of explanation
representation (layers, units, vectors)
processing (decision boundary, feature attribution)
training
data
uncertainty
Presentation
Privacy awareness
Presentation form
visual
verbal (textual, formal)
auditive
combinations
Level of abstraction
Accessibility
simple
complex & symbolic
Information units
raw feature
derived feature
semantic feature
abstract semantic feature
with or without interactions
Explanator output type
by example (e.g. closest other samples, word cloud)
contrastive / counterfactual / near miss example (e.g. adversarial ex.)
prototype (e.g. generated, concept vector)
feature importance
rule-based (e.g. if-then, binary, m-of-n, hyperplane)
dimension reduction
dependence plots
graph
diagrams
combinations


Explanation
Global
Local
Explanandum
The thing or phenomenon to be explained
Computational
How the algorithm produced the output
Mechanistic
Why the algorithm produced the output
Justificatory
Why the output is right
Causal
What is the phenomenon causing the output
Informative
What the output entails
Cautionary
What is the output's uncertainty
Inventory
Types
Verification
Requirements
Validation
Metrics
Explanatory relationship
Relationship that holds between the explanans and the explanandum
Epistemological
Classifications
Bona fide explanations
Cognitive
Understanding
Motivation
Persuasion
Knowledge
Confidence
Trust
Motivating Explanations
Reassuring Explanations
Aligning Explanations
Explanans
The thing that explains and is perceived by users
Instance- or Sample-based
Similar cases
Opposite cases
Counterfactual cases
Feature-based
Distribution- or Morphology-based
Visual
Text-based

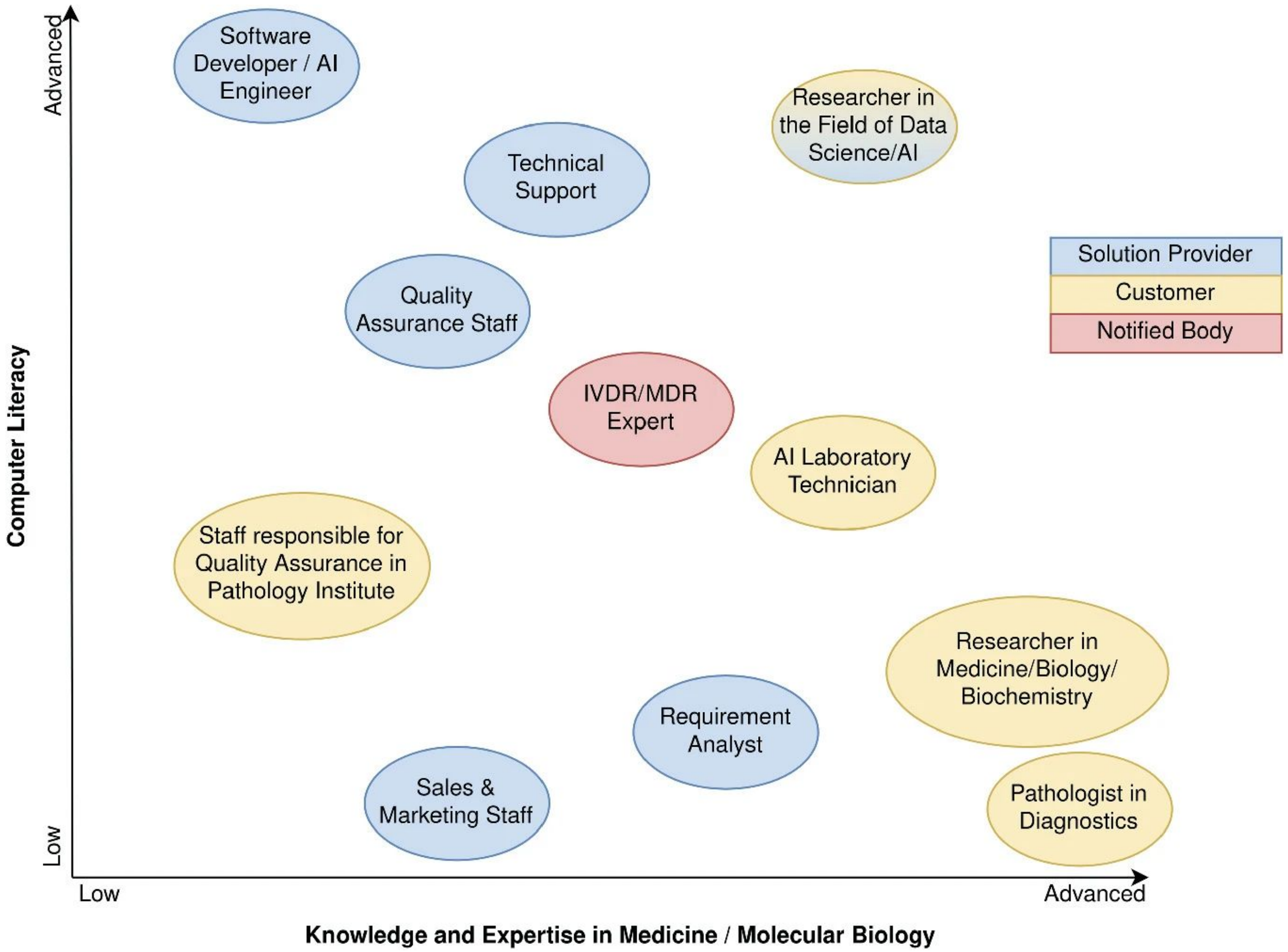
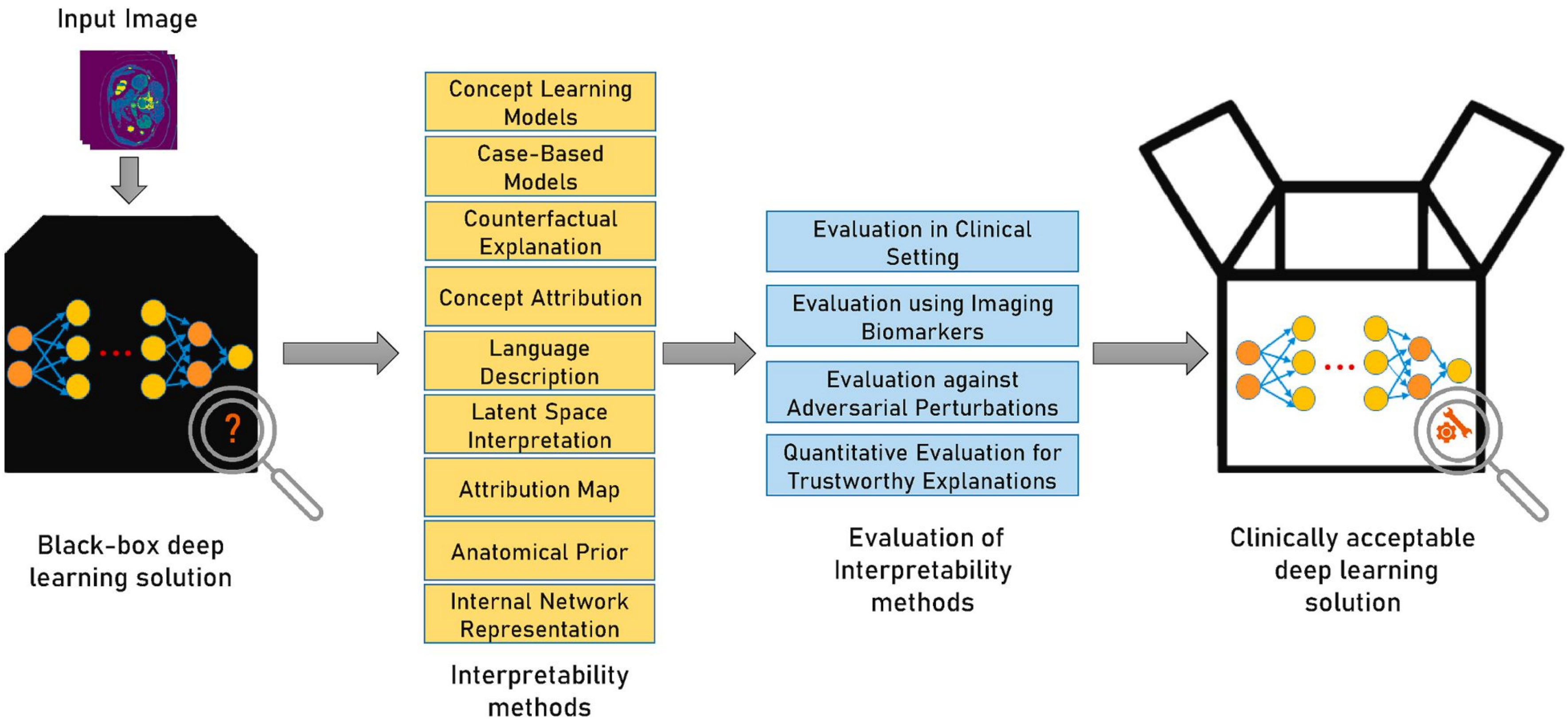

Advanced
Low
Computer Literacy
Software Developer / AI Engineer
Technical Support
Researcher in the Field of Data Science/AI
Solution Provider
Customer
Notified Body
Quality Assurance Staff
IVDR/MDR Expert
AI Laboratory Technician
Staff responsible for Quality Assurance in Pathology Institute
Researcher in Medicine/Biology/ Biochemistry
Requirement Analyst
Sales & Marketing Staff
Pathologist in Diagnostics
Low
Advanced
Knowledge and Expertise in Medicine / Molecular Biology

Image Model Explainability workflow

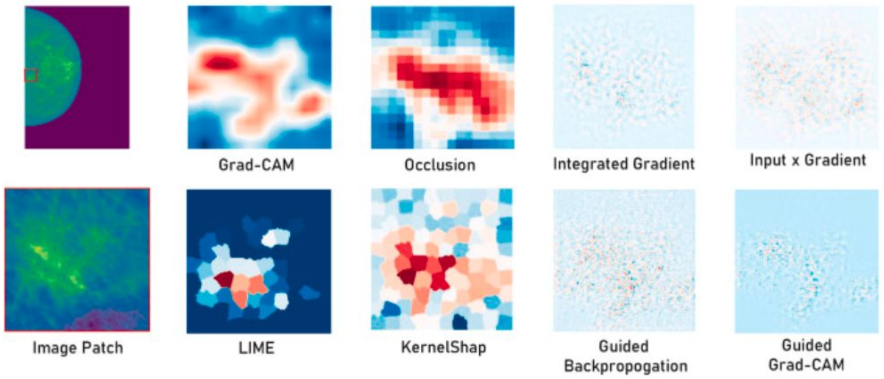


Attribution Map

Post-hoc explanations are provided by highlighting the regions of the input image that the model considers important.

No information is provided on how the relevant regions contribute to the prediction, multiple classes can have the same regions highlighted.

Attribution maps



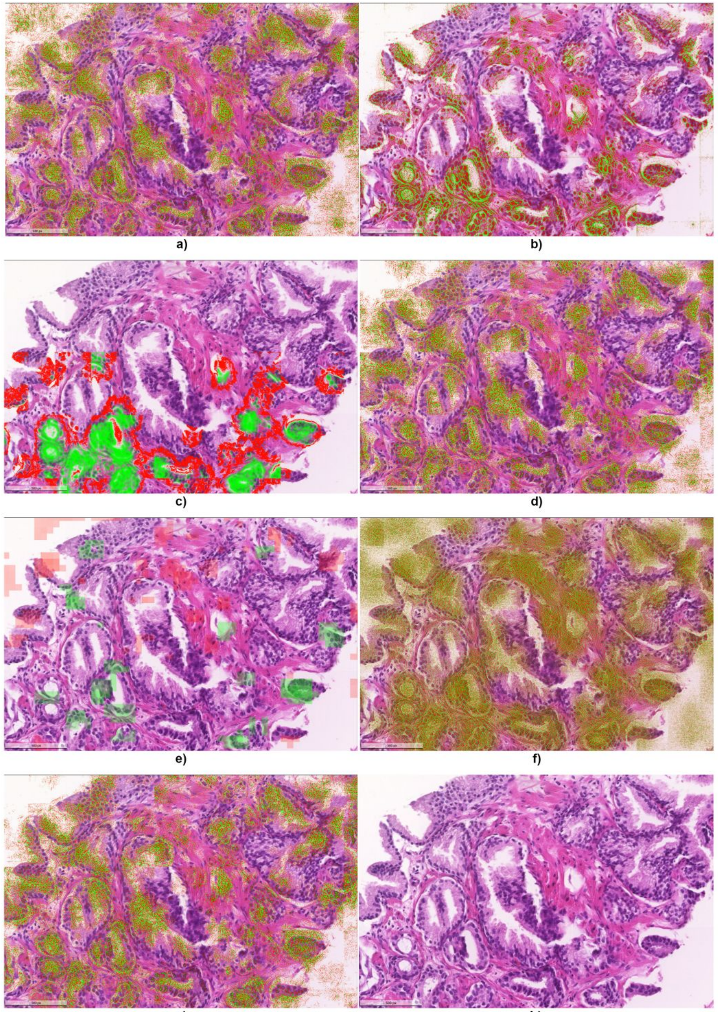
Layerwise Relevance Propagation (LRP)	[15]	Alzheimer's disease classification
	[35]	Multiple Sclerosis diagnosis
Class Activation Maps (CAM)	[16]	Oral cancer classification
Gradient-Class Activation Maps (Grad-CAM)	[104]	Automatic brain tumor grading
	[101]	Detection of COVID-19 from Chest X-ray and CT scans
Integrated Gradient (IG)	[137]	Diabetic Retinopathy (DR) prediction
	[150]	Multiple Sclerosis classification
Occlusion	[64]	Diagnosis of age-related macular degeneration and diabetic macular edema in OCT images
	[139]	Diagnosis of Alzheimer's disease
Local Interpretable Model-agnostic Explanations (LIME)	[93]	Parkinson's disease detection
	[122]	Congestive heart failure prediction
kernel SHAP (Linear LIME+Shapley values)	[161]	Skin cancer detection
	[171]	Lung nodule classification
Trainable Attention	[154]	Melanoma recognition
	[156]	Classification of breast cancer microscopy images
	[59]	Organ segmentation in 3D abdominal CT scans
SmoothGrad	[44]	Identification of cardiac structure, estimation of cardiac function and prediction of systemic phenotypes from Echocardiography
	[102]	Classification of estrogen receptor status from breast MRI
Guided BackPropagation (GBP)	[51]	Lung adenocarcinoma classification
DeepLIFT (Learning Important FeaTures)	[89]	Diagnosis of Multiple Sclerosis
Deep SHAP (DeepLIFT+Shapley values)	[62]	Breast lesion classification, lung lesion classification
	[107]	COVID and Pneumonia classification from chest X-rays
Deep Taylor Decomposition (Deep Taylor)	[127]	Content-based image retrieval for pleural effusion condition in Chest X-Ray images
Multi-Layer Class Activation Maps (MLCAM)	[57]	Feature localization for Confocal Laser Endomicroscopy Glioma images
Expected Gradients	[28]	COVID-19 detection
Contextual Decomposition Explanation Penalization (CDEP)	[109]	Skin Lesion Classification

Attribution maps

a) I*G
b) Guided Backprop
c) Deep Taylor Decomposition
d) LRP
e) Occlusion sensitivity
f) DeconvNet
g) Integrated Gradients
h) Original Image

Computed using

https://github.com/albermax/innvestigate

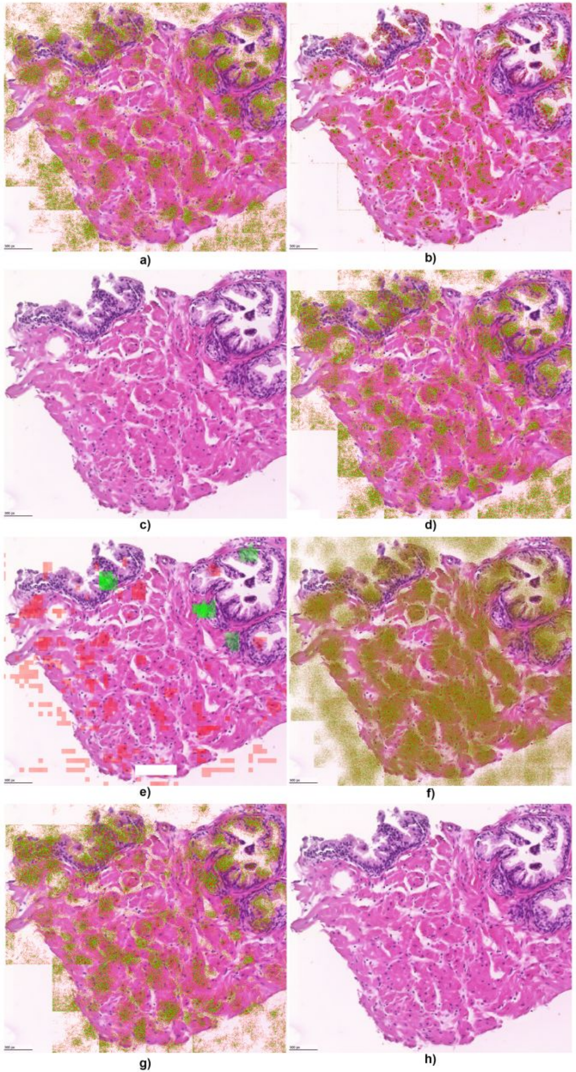


Attribution maps

a) I*G
b) Guided Backprop
c) Deep Taylor Decomposition
d) LRP
e) Occlusion sensitivity
f) DeconvNet
g) Integrated Gradients
h) Original Image

Computed using

https://github.com/albermax/innvestigate

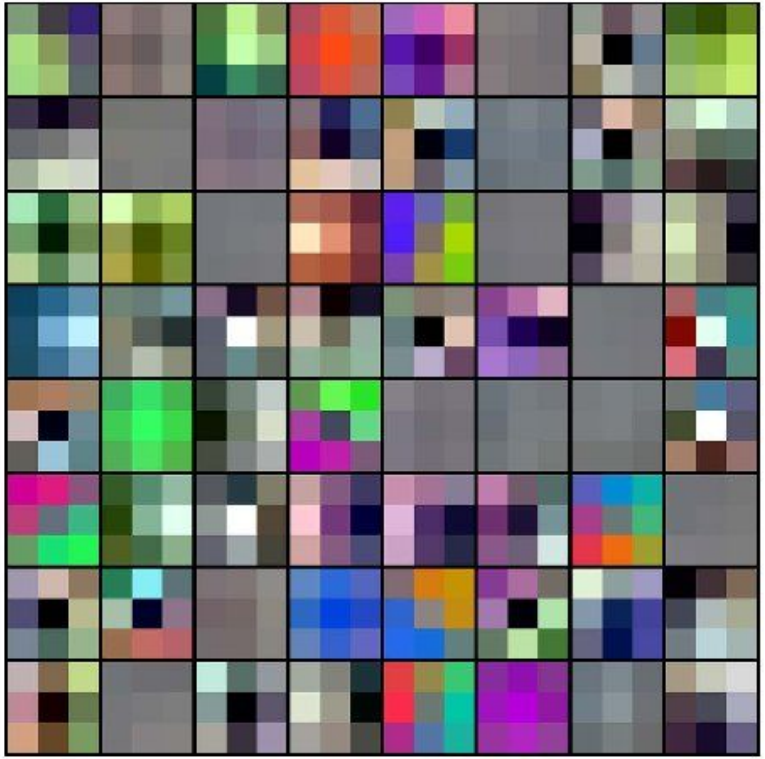


Internal Network Representation

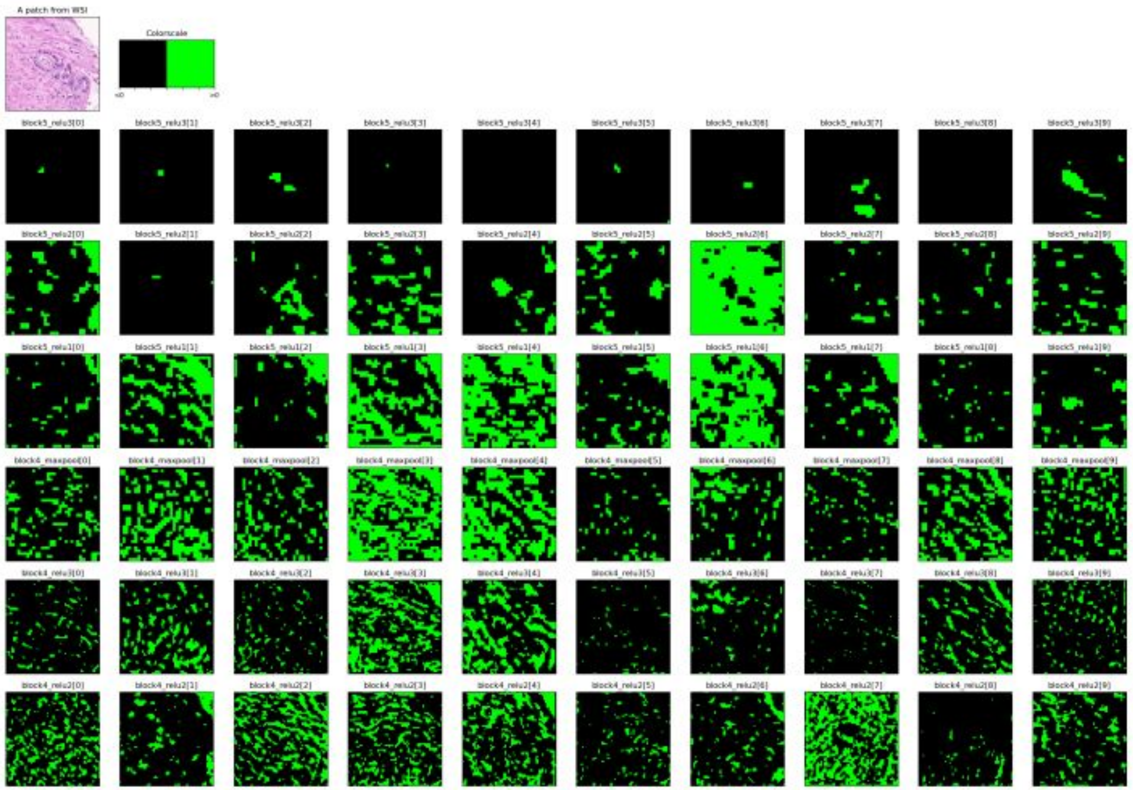
Visualization of features learned by different filters in a CNN.

The structures and patterns that different filters learn to identify are hard to interpret in medical images.

Filter visualization

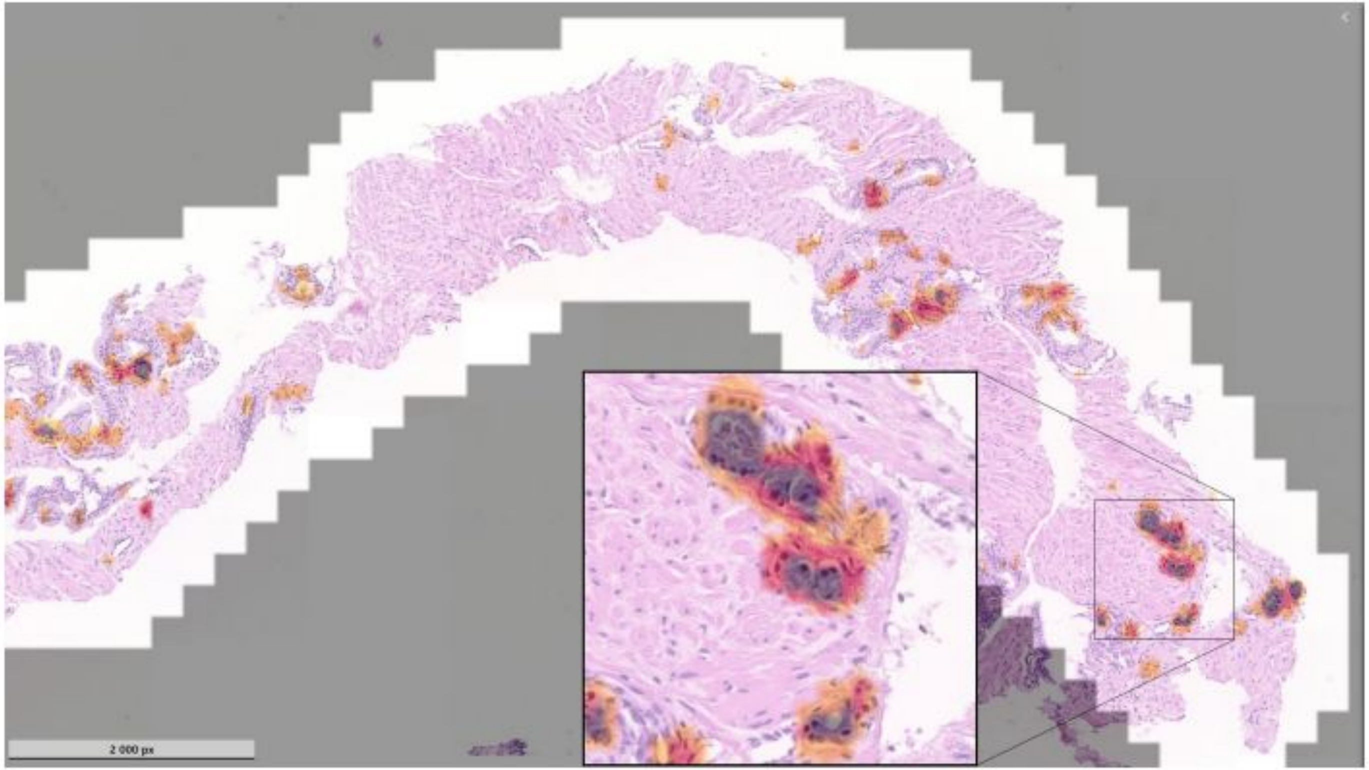


Internal Network Representation



Internal Network Representation - clustering

Cluster pixels according to feature maps activated "above" them.



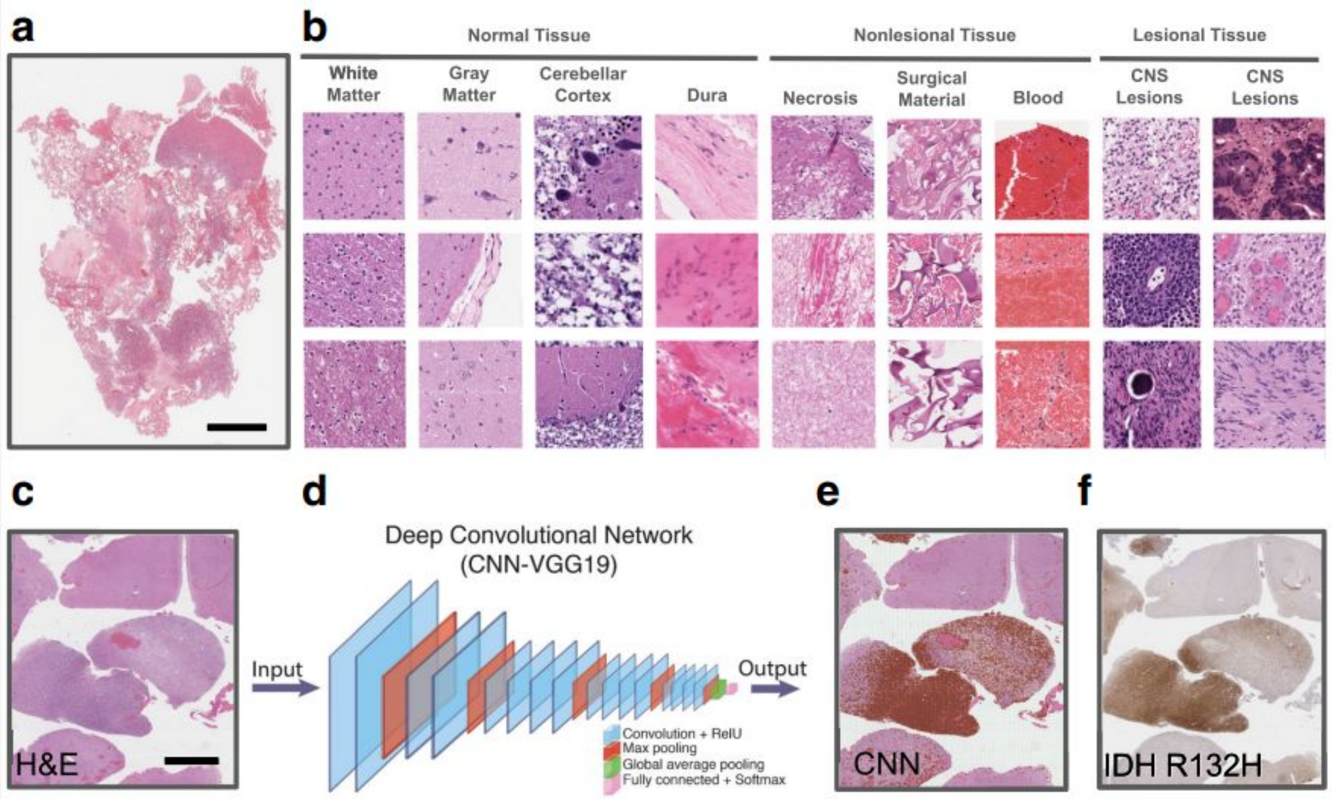
Latent Space Interpretation

The latent space is used to uncover the salient factors of variation learned in the data with respect to the clinical knowledge. Visualization of high-dimensional latent space in two dimensions to identify similarities and outliers.

Loss of information when the high-dimensional feature space is projected to two dimensions. The similarity in latent space does not always translate to the similarity in terms of human-interpretable features.

Latent Space Interpretation - tissue segmentation

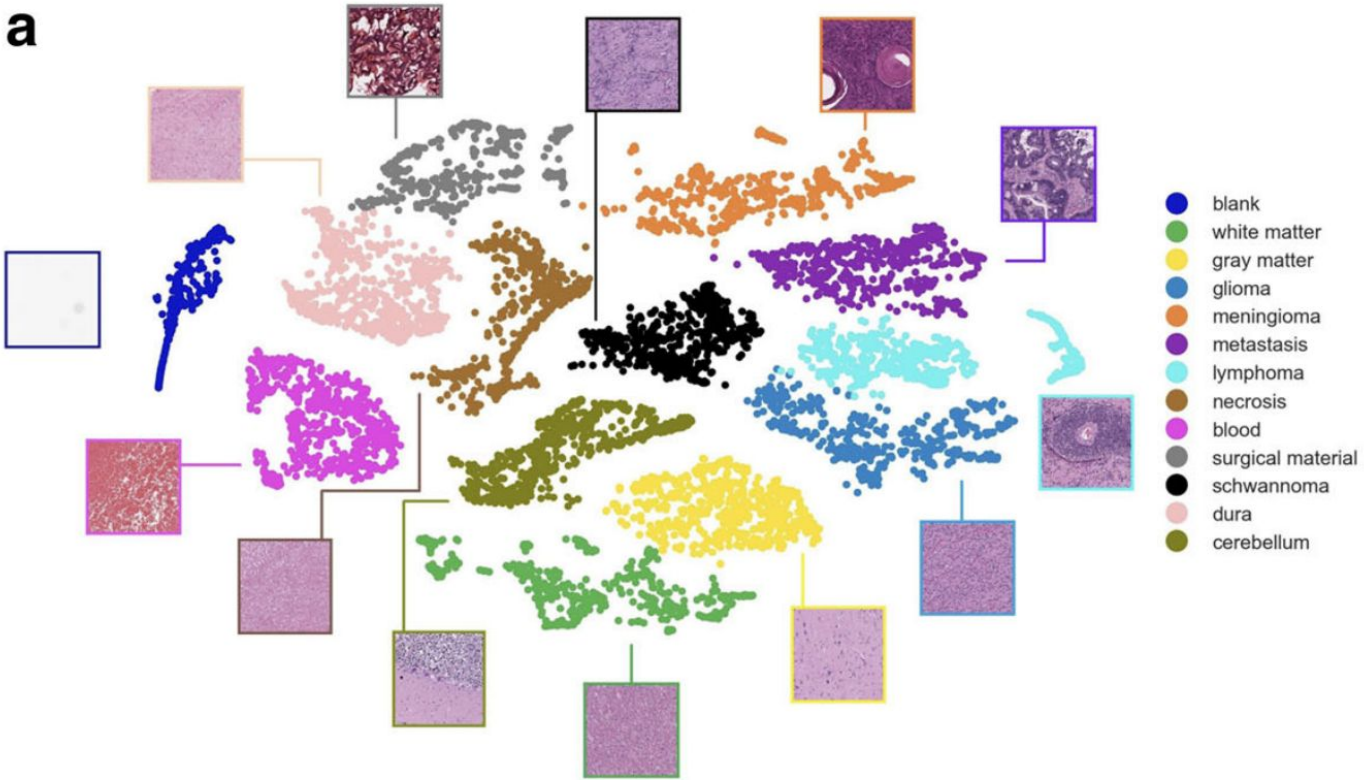
- Patch based tissue segmentation (brain)
- 13-classes (white/grey matter, blioma, etc.)
- Simple multiclass network on 1024x1024 patches



Visualizing histopathologic deep learning classification and anomaly detection using nonlinear feature space dimensionality reduction, Han, BMC Bioinformatics, 2018

Latent Space Interpretation - tissue segmentation

2D t-SNE visualization of the final hidden layer features



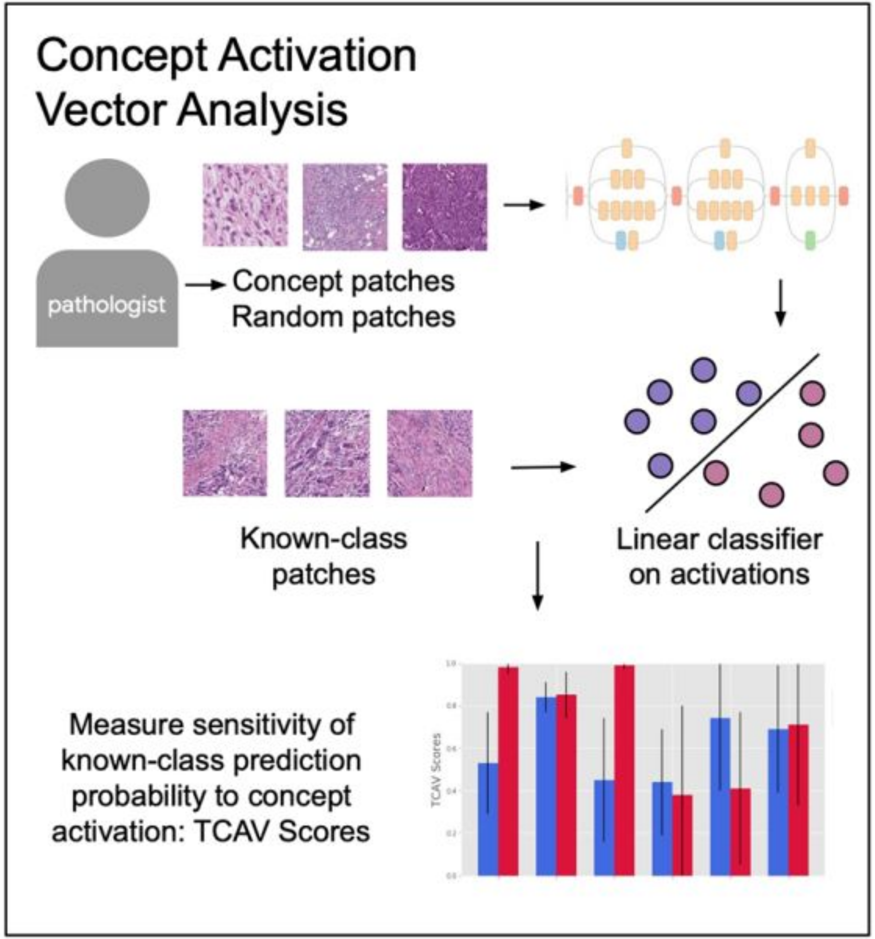
Concept Attribution

Global explanations to quantify the influence of high-level image concepts/features on the model predictions.

Difficult to annotate high-level clinical concepts, features used for interpretability may not be reproducible.

TCAV

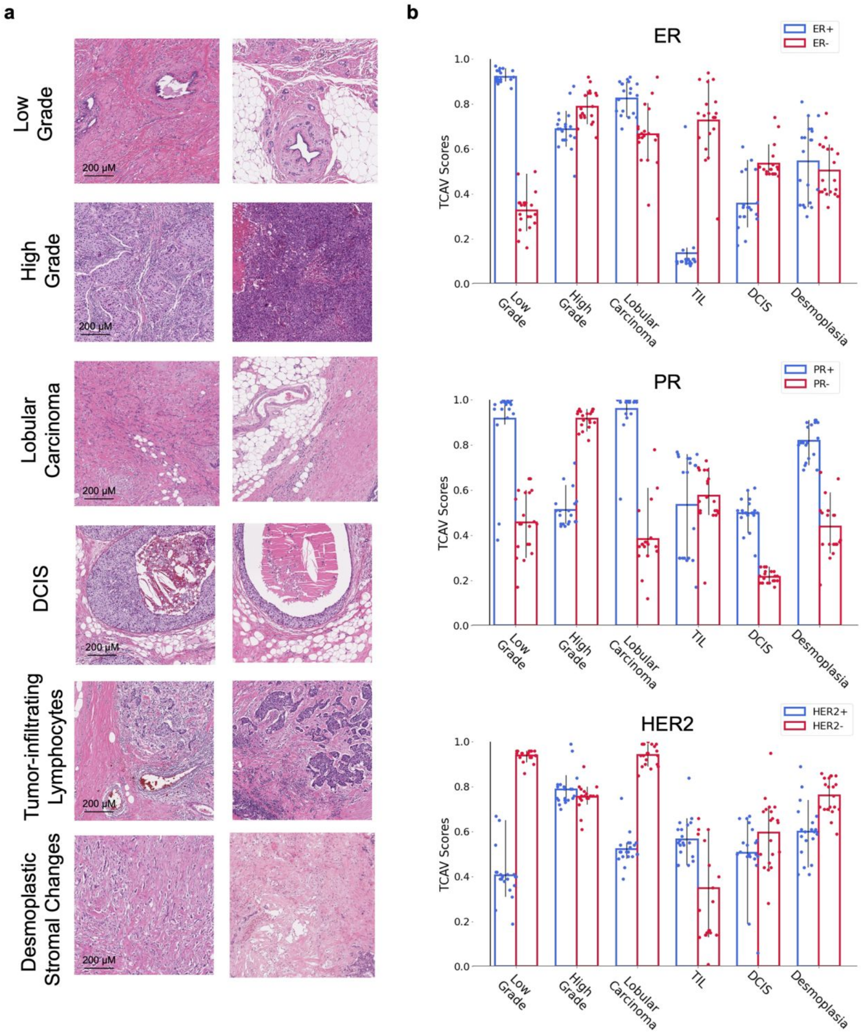
- Linear classifier in the latent space - distinguish latent representation of the concept vs random input
- Derivative of the output w.r.t. the normal vector of the linear classifier



Determining breast cancer biomarker status and associated morphological features using deep learning, Gamble et al, NComms, 2021

Concept Attribution

- Predict mutations of ER, PR, HER2
- Observe cancer-related concepts

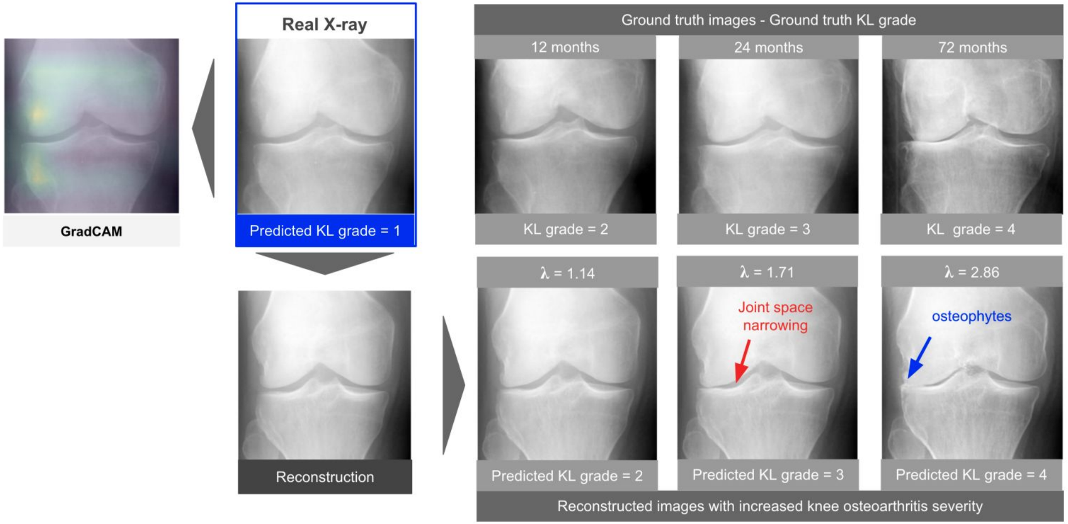


Counterfactual Explanation

Input images are perturbed in a realistic manner to generate the opposite prediction.

Possibility of unrealistic perturbations to the input images, the resolution of the generated counterfactual images is limited.

Counterfactuals - Using Generative Models



- Trained generator from latent **w** to images
- Take latent representation **w'** of a real image
- Consider **w'**+α**D** where **D** is the direction of increasing/decreasing model output

Using Stylegan for Visual Interpretability of Deep Learning Models on Medical Images, Schutte et al, https://arxiv.org/abs/2101.07563, 2021

Language Description

Textual justifications are provided along with the predictions.

Structured diagnostic reports require more annotation efforts, duplication of training sentences during testing.

Language Description

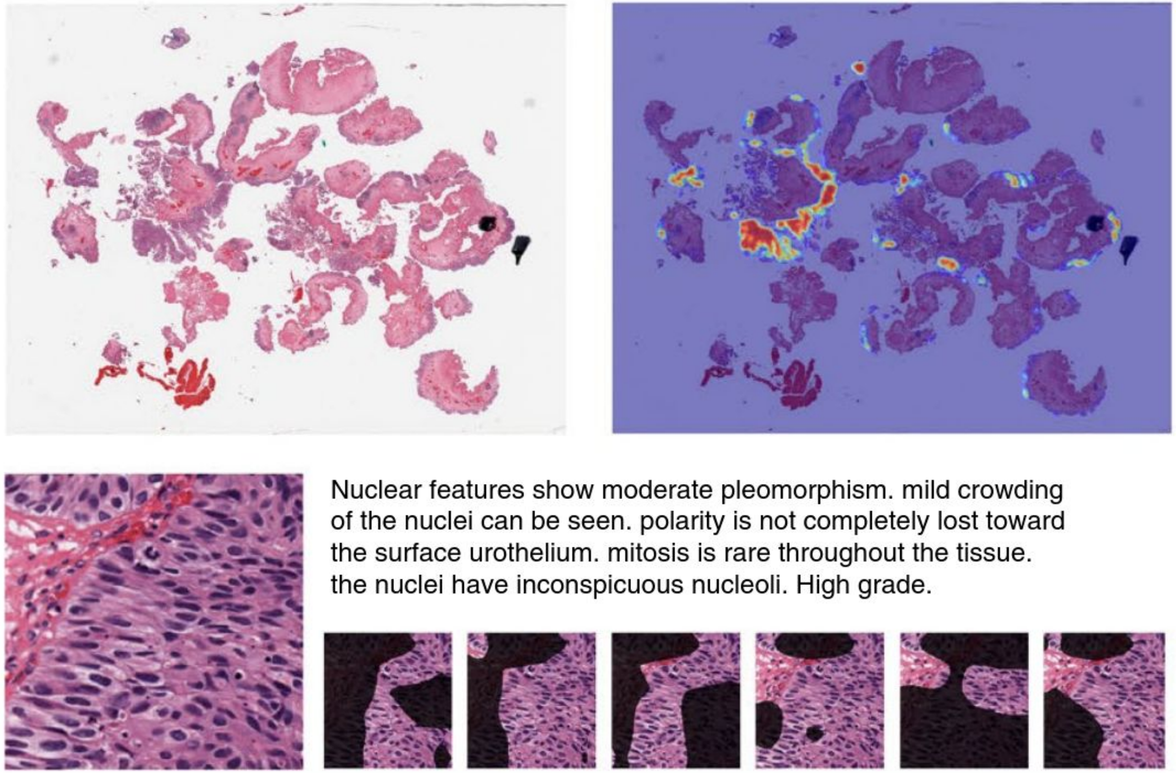
Training data:

Image labels (tumor)

- Patches 1024 x 1024
- Pixel-level labels of tumor

Textual descriptions

- Microscopic findings - 5 types of cellular features
- Vocabulary size 112 (21,265 image-report pairs)
- Feature aware attention (indicates what it sees when generating the text)



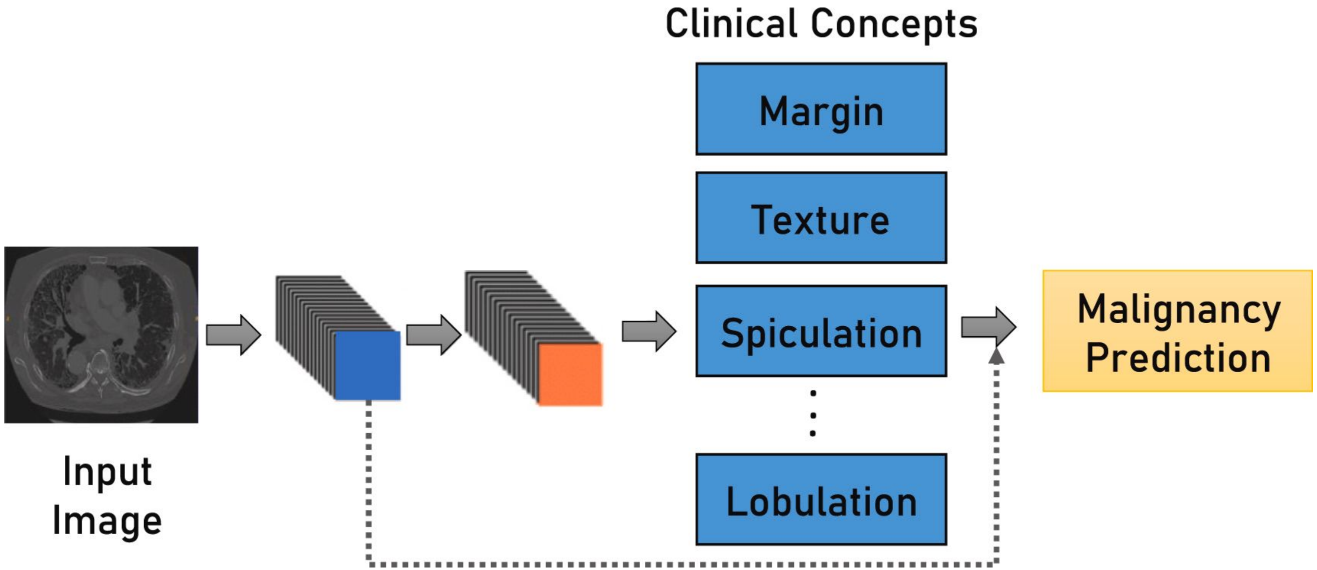
Pathologist-level interpretable whole-slide cancer diagnosis with deep learning, Zhang et al, Nature learning machines, 2019

Concept Learning Models

High-level clinical concepts are first predicted and the final classification is made using these concepts.

Additional annotation cost, learned concepts may encode information beyond the intended clinical concepts due to information leakage.

Concept Learning Models



- Can be misleading as the learned encoding contains information in addition to the concept representation (cheating)

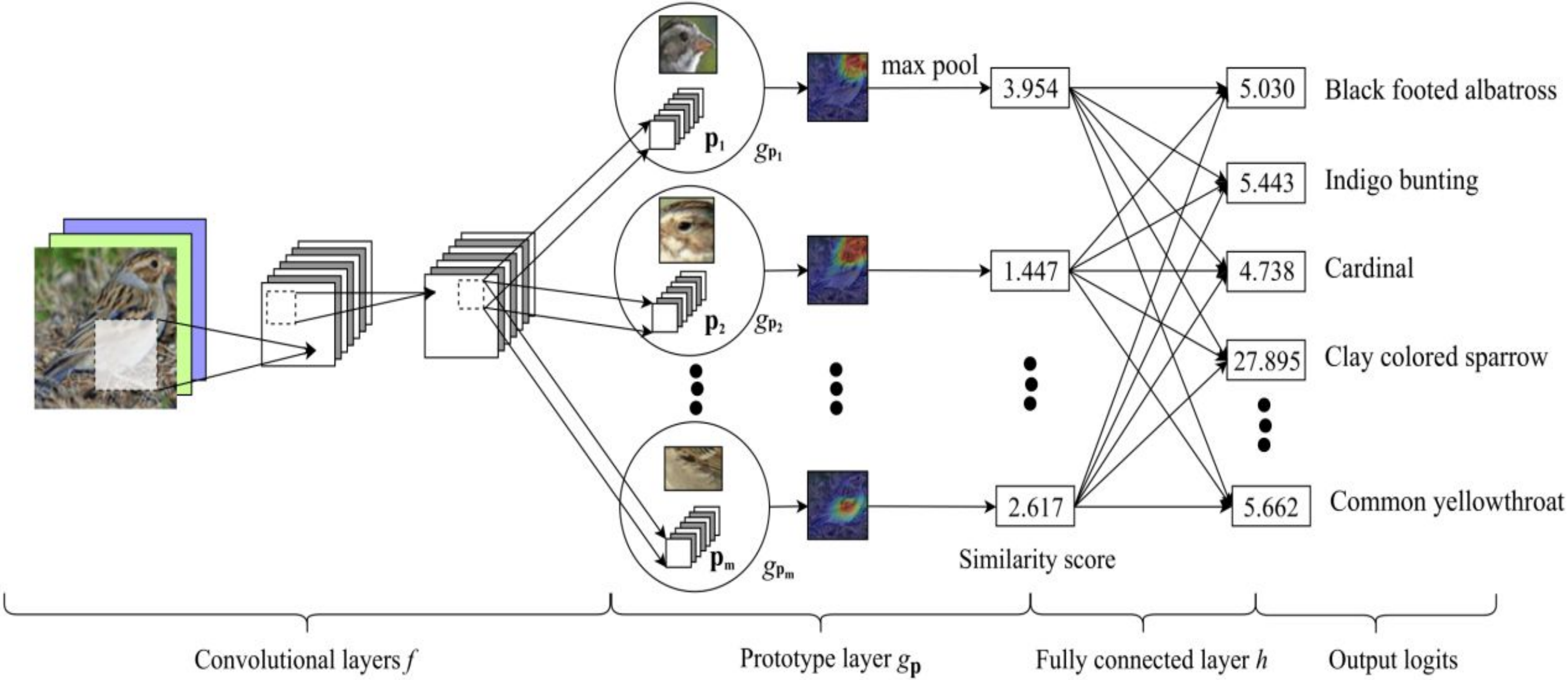
DO CONCEPT BOTTLENECK MODELS LEARN AS INTENDED?, Margeloiu, ICLR 2021

Case-Based Models

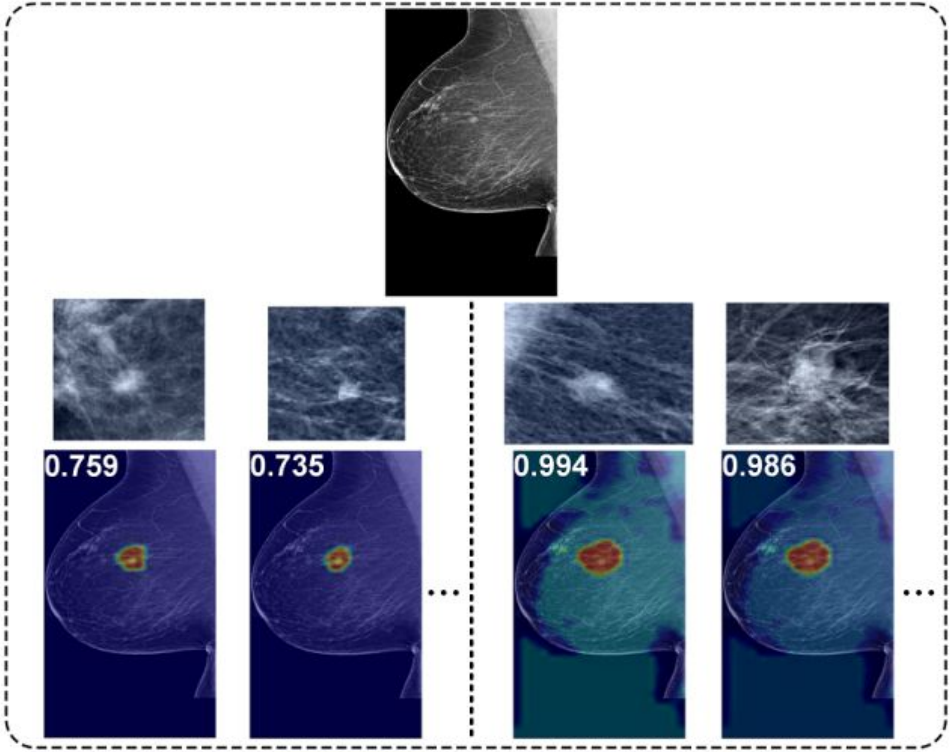
Class discriminative prototypes are learned and the final classification is performed by comparing features extracted from input images with the prototypes.

Susceptibility to corruption by noise and compression artefacts, difficult to train.

Case-Based Models - ProtoPNet



Case-Based Models - ProtoPNet - Mammogram



- Non-cancer and cancer prototypes
- Similarity heatmaps to the prototypes
- prototypes can be corrupted due to the semantic gap between similarity in latent space and in input space

Is ProtoPNet Really Explainable? Evaluating and Improving the Interpretability of Prototypes, Huang et al, https://arxiv.org/abs/2212.05946, 20222

Anatomical Prior

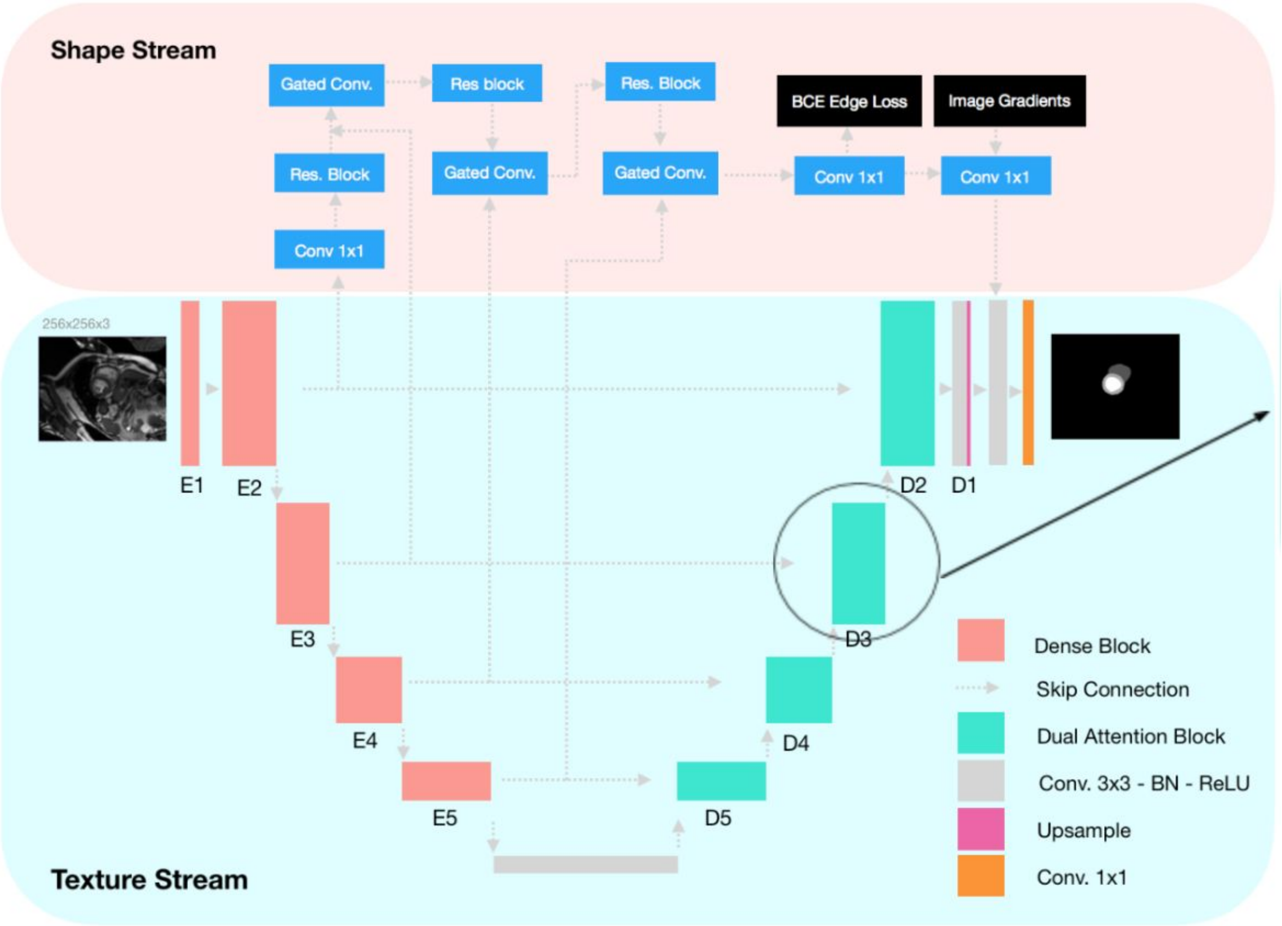
Task-specific structural information is incorporated in the design process of the network.

Specialized clinical knowledge may be required, anatomical prior cannot be utilized for all problems.

Priors

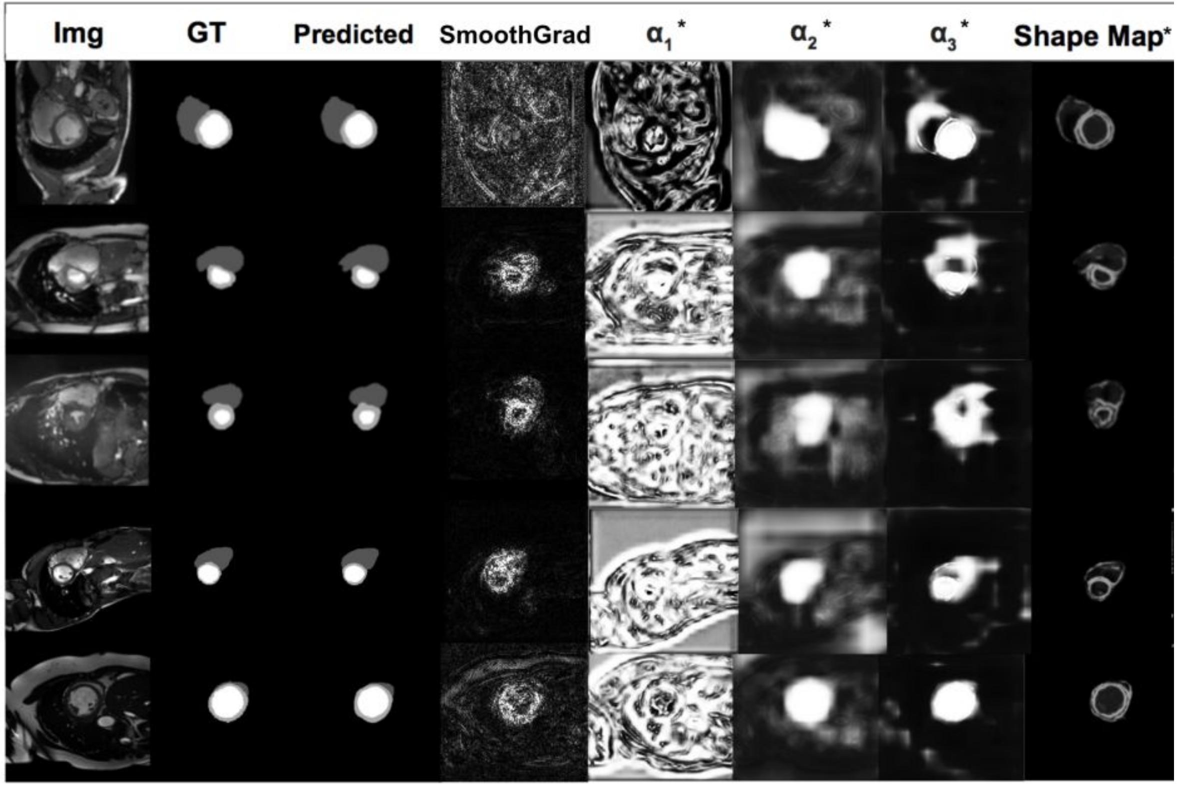
UNet for segmentation

Add shape stream incorporating shape loss (length of boundary, area)



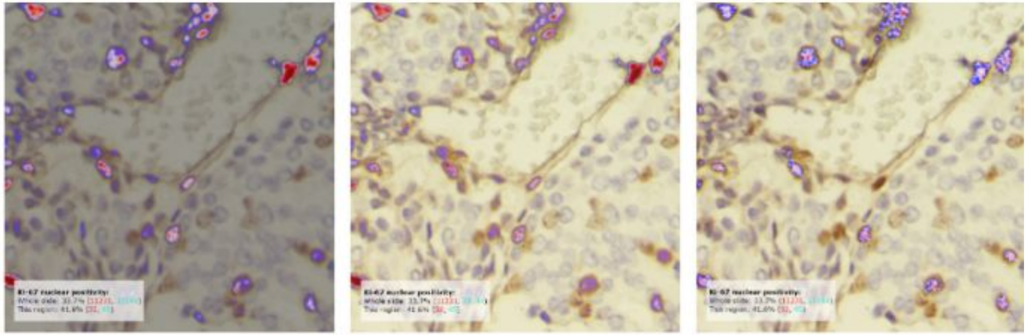
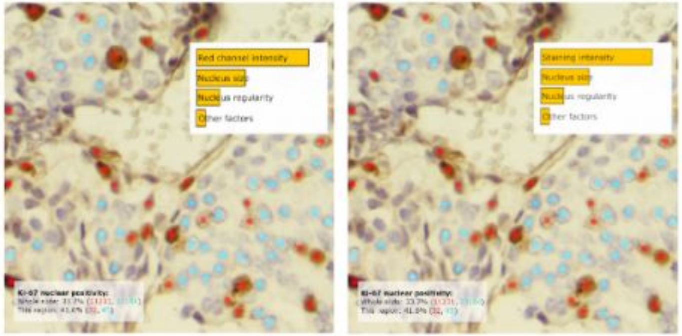
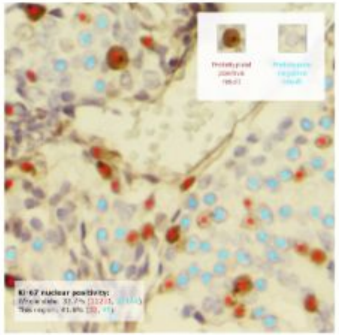
SAUNet: Shape Attentive U-Net for Interpretable Medical Image Segmentation, Sun et al, https://arxiv.org/pdf/2001.07645v3.pdf, 2020

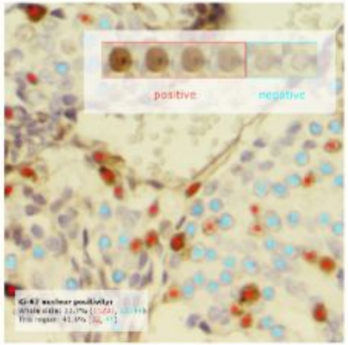
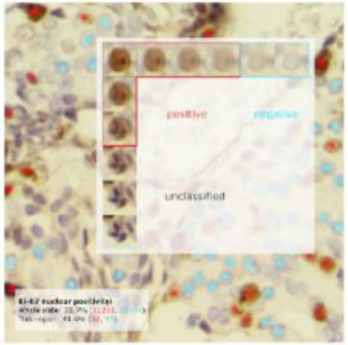
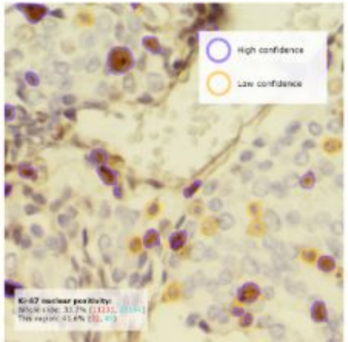
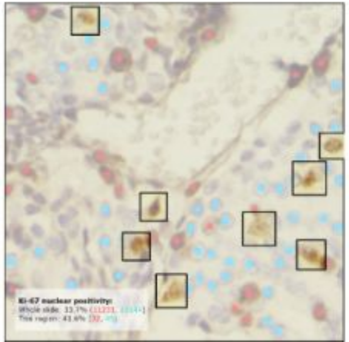
Priors



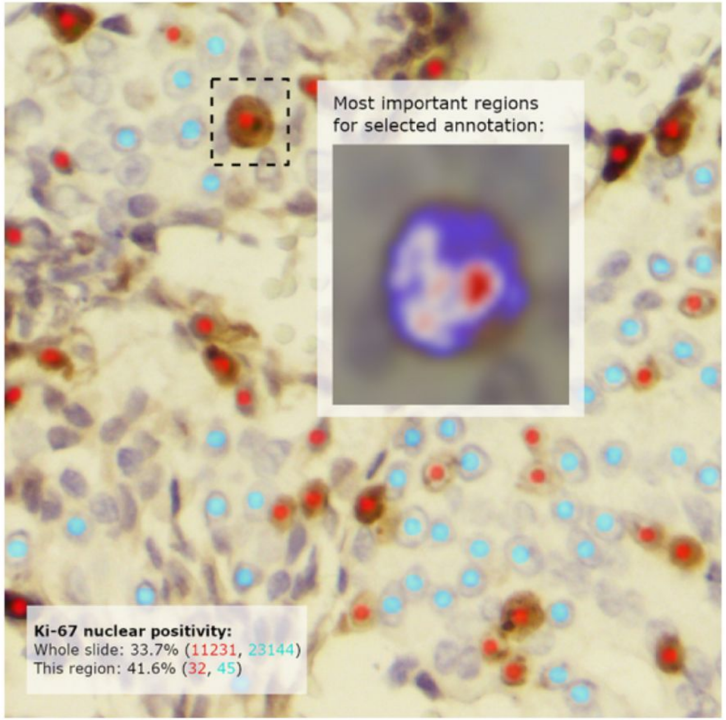
The explainability paradox

- mixed-methods study of user interaction with samples of state-of-the-art AI explainability techniques for digital pathology
- How are state-of-the-art xAI approaches interpreted and evaluated by expert users in a typical diagnostic setting?
- How do these interpretations and evaluations inform principles for the development of safe and effective xAI?
- Evaluation of five explanation generating methods
  Saliency maps, concept attribution, prototype, counterfactual, trust score
- AI-assisted Ki-67 quantification was chosen as a representative task from the slide examination step of the digital pathology workflow

Saliency Map (Global)	Show the most relevant pixels for the positive classifications within this region of interest	
Concept Attribution	Show the most important features attributed to positive classifications	
Prototypes	Show prototypical positively and negatively classified annotations within this region	

Counterfactuals (One-axis)	Show generated examples interpolating between positive and negative examples, showing model classifications for each	
Counterfactuals (Two-axis)	Show generated examples changing in two principal factors of variation, showing model classifications for each	
Trust Scores	Display low-confidence annotations for review	 

empaia



Hide explanation

Local saliency map

Show the most relevant pixels for the classification of a selected annotation

I find the explanation intuitively understandable *

Strongly disagree | 1 | 2 | 3 | 4 | 5 | 6 | 7 | Strongly agree

The explanation helps me to understand factors relevant to the algorithm *

Strongly disagree | 1 | 2 | 3 | 4 | 5 | 6 | 7 | Strongly agree

The explanation helps me to decide whether I can trust the generated annotations *

Strongly disagree | 1 | 2 | 3 | 4 | 5 | 6 | 7 | Strongly agree

The explanation provides me with valuable information for my work *

Strongly disagree | 1 | 2 | 3 | 4 | 5 | 6 | 7 | Strongly agree

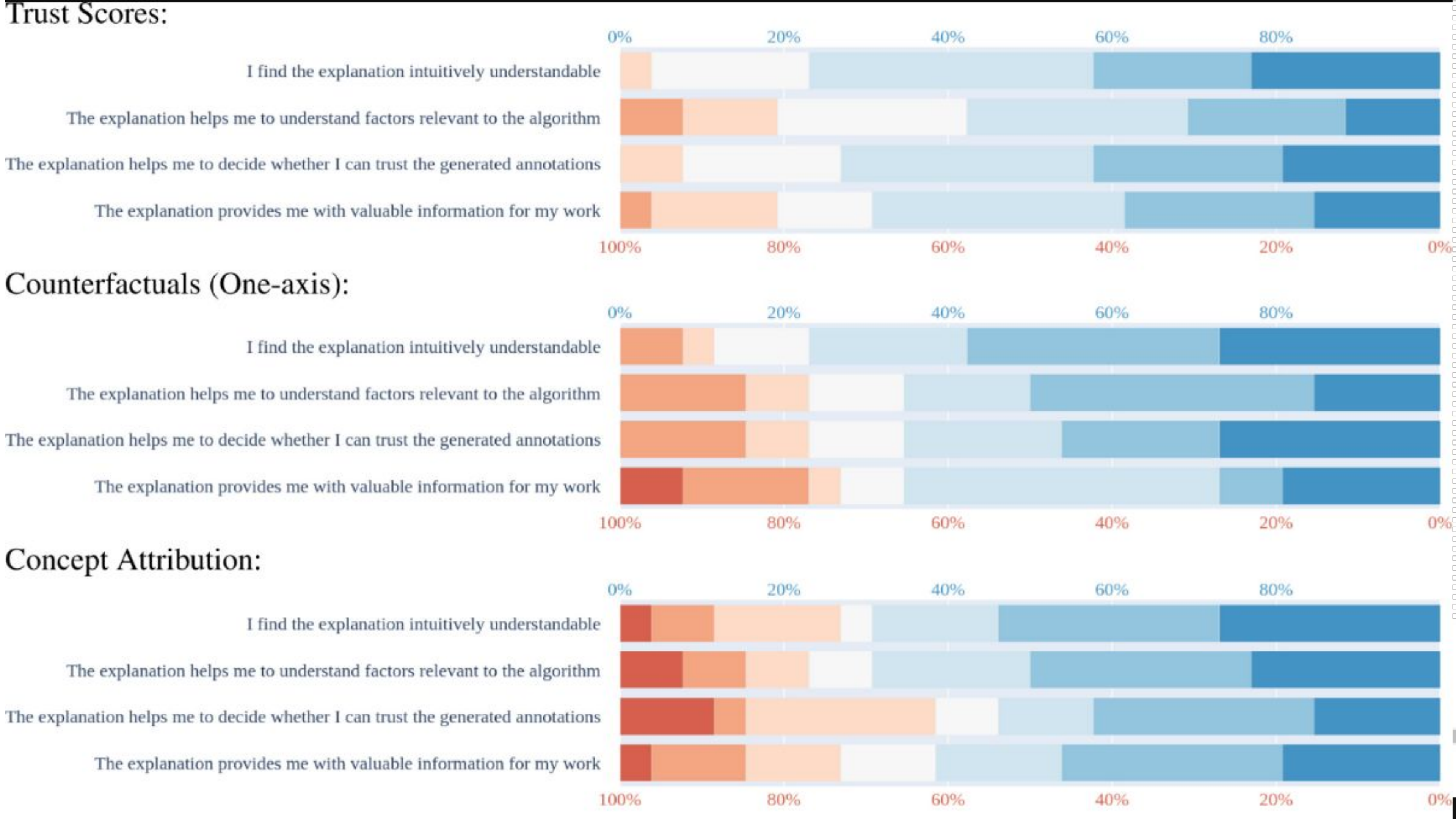
Additional comments

Previous

Next

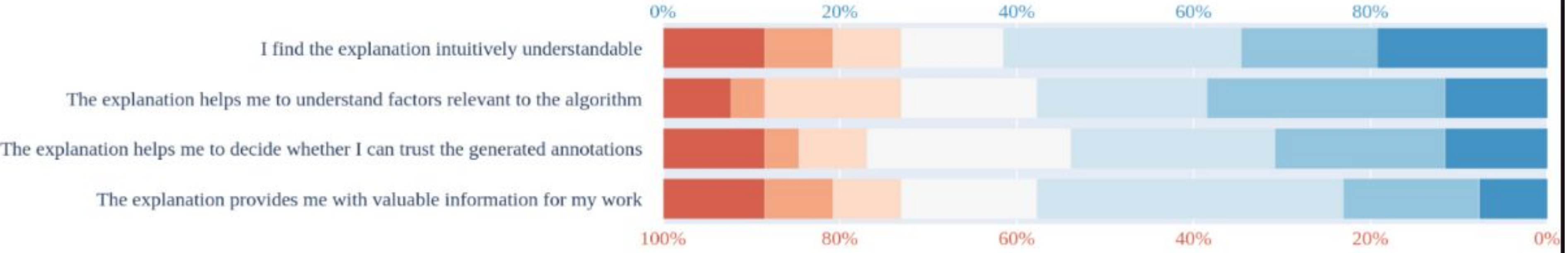
Evaluation

- Questionnaire for 25 respondents
- individuals holding professional roles in pathology or neuropathology
  - consultant (12)
  - researcher (6)
  - pathologist in training (4)
  - technician (3)

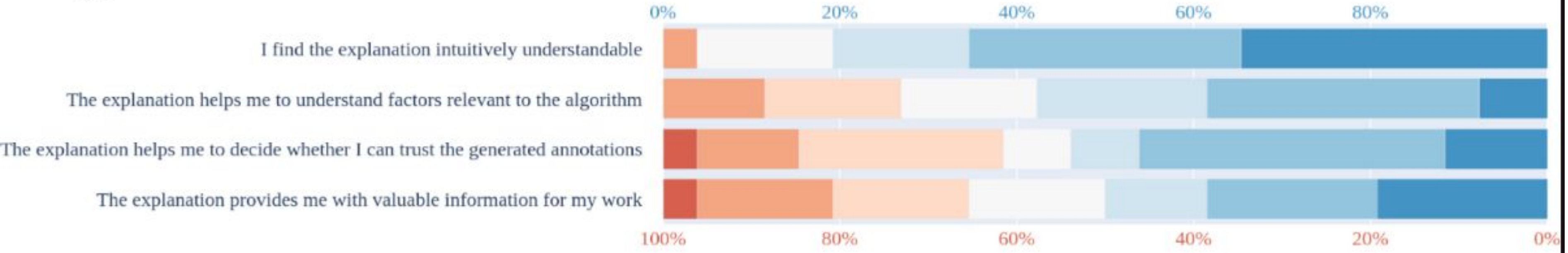


Trust Scores:
0%
20%
40%
60%
80%
I find the explanation intuitively understandable
The explanation helps me to understand factors relevant to the algorithm
The explanation helps me to decide whether I can trust the generated annotations
The explanation provides me with valuable information for my work
100%
80%
60%
40%
20%
0%
Counterfactuals (One-axis):
0%
20%
40%
60%
80%
I find the explanation intuitively understandable
The explanation helps me to understand factors relevant to the algorithm
The explanation helps me to decide whether I can trust the generated annotations
The explanation provides me with valuable information for my work
100%
80%
60%
40%
20%
0%
Concept Attribution:
0%
20%
40%
60%
80%
I find the explanation intuitively understandable
The explanation helps me to understand factors relevant to the algorithm
The explanation helps me to decide whether I can trust the generated annotations
The explanation provides me with valuable information for my work
100%
80%
60%
40%
20%
0%

Counterfactuals (Two-axis):



Prototypes:



Saliency map (Global):