Explainable AI for Medical Data: Current Methods, Limitations, and Future Directions

MD IMRAN HOSSAIN and GHADA ZAMZMI, University of South Florida, USA
PETER R. MOUTON, SRC Biosciences, USA
MD SIRAJUS SALEKIN, YU SUN, and DMITRY GOLDGOF, University of South Florida, USA

With the power of parallel processing, large datasets, and fast computational resources, deep neural networks (DNNs) have outperformed highly trained and experienced human experts in medical applications. However, the large global community of healthcare professionals, many of whom routinely face potentially life-or-death outcomes with complex medicolegal consequences, have yet to embrace this powerful technology. The major problem is that most current AI solutions function as a metaphorical black-box positioned between input data and output decisions without a rigorous explanation for their internal processes. With the goal of enhancing trust and improving acceptance of artificial intelligence– (AI) based technology in clinical medicine, there is a large and growing effort to address this challenge using eXplainable AI (XAI), a set of techniques, strategies, and algorithms with an explicit focus on explaining the "hows and whys" of DNNs. Here, we provide a comprehensive review of the state-of-the-art XAI techniques concerning healthcare applications and discuss current challenges and future directions. We emphasize the strengths and limitations of each category, including image, tabular, and textual explanations, and explore a range of evaluation metrics for assessing the effectiveness of XAI solutions. Finally, we highlight promising opportunities for XAI research to enhance the acceptance of DNNs by the healthcare community.

CCS Concepts: • **Computing methodologies** → **Machine learning**; • **General and reference** → **Evaluation;**

Additional Key Words and Phrases: Explainability, medical data, responsible AI, deep neural networks, interpretable AI

ACM Reference format:
Md Imran Hossain, Ghada Zamzmi, Peter R. Mouton, Md Sirajus Salekin, Yu Sun, and Dmitry Goldgof. 2025. Explainable AI for Medical Data: Current Methods, Limitations, and Future Directions. *ACM Comput. Surv.* 57, 6, Article 148 (February 2025), 46 pages.
https://doi.org/10.1145/3637487

1 INTRODUCTION

Since the early 2010s, **artificial intelligence (AI)** powered by deep learning has demonstrated remarkable performance in the medical domain with applications for skin disease identification

Authors' addresses: Md I. Hossain, G. Zamzmi, Md S. Salekin, Y. Sun, and D. Goldgof, Department of Computer Science and Engineering, ENB, University of South Florida, 4202 E. Fowler Ave, Tampa, FL 33620, USA; e-mails: {mdimranh, ghadh, salekin, yusun}@usf.edu, goldgof@mail.usf.edu; P. R. Mouton, SRC Biosciences, 1810 W. Kennedy Blvd, Tampa, Florida 33606, USA; e-mail: petermouton@usf.edu.

0360-0300/2025/02-ART148 $15.00
https://doi.org/10.1145/3637487

[172], COVID-19 detection [193], early pain detection in neonates [178], the diagnosis of retinal disease [113], image segmentation [83], and classification of various malignancies from pathological images of breast [11] and brain cancer [93]. Despite these advancements, implementing **deep neural networks (DNNs)** has been slow to gain wide acceptance in various clinical healthcare settings. This reluctance arises due to prioritizing accuracy over making the decision-making processes explainable [142]. Explainability adds a powerful tool for verifying and improving performance by detecting weaknesses, learning patterns in the input data, and identify clinically meaningless features in the millions of input parameters and hundreds of network layers [180]. Most importantly, using XAI to make healthcare algorithms more transparent will help clinicians build greater trust in their decision-making processes.

Existing XAI approaches for image analysis can be classified as attribution based and non-attribution based. Attribution-based approaches highlight the decision-making region of the image by generating a heatmap, while non-attribution-based approaches help analyze model behavior, model debugging, and even explain the prediction decision step by step. Attribution-based approaches have been used with several medical imaging applications, including COVID detection [4], knee pain assessment [24], Alzheimer's disease classification [22], and diagnosing multiple sclerosis (MS) [44]. Examples of non-attribution-based approaches include clinically meaningful concepts selection for explaining cardiac **magnetic resonance imaging (MRI)** segmentation [64], image perturbation for knee osteoarthritis severity [183], and prototypical images for classifying cancer or non-cancer [212].

In addition, the multimodal XAI method is becoming more and more prevalent in healthcare applications due to the variety of medical data sources involved. Instead of concentrating on a single data type, the multimodal XAI explains different imaging modalities like X-ray, CT, and MRI or different data kinds like images, time-series data, sequential data, clinical records, and metadata. XAI clarifies the roles of different data types [222], explicates the relationship and significance of different data types [235], elucidates the critical features [89], and offers insights into the influential data points from different modalities.

The importance of explainability and its role in creating trustworthy AI pushed researchers to conduct comprehensive reviews of current XAI methods. For instance, general XAI concept, taxonomy, various definitions, reviews of both deep and shallow models, programming implementation, research topics associated with explainability, challenges of XAI, and guidelines for responsible AI have been presented in References [8, 36, 42, 79, 123, 219, 232]. Similarly, the authors of Reference [21] categorized the explanations methods, described the methods based on three popular data types (images, tabular data, and text), and additionally reported a comparison to assess several visual XAI methods. For multi-modal data, Joshi et al. [87] reviewed related methods, presented the challenges of XAI in multi-modal datasets, and discussed the significance, challenges, and future trends. Several reviewers have been published in the case of XAI for medical imaging applications. For example, Huff et al. [78] reviewed the visualization methods to highlight the important parts of medical images. In addition, the applications of the XAI methods in medical image analysis according to the anatomical location are presented in Reference [215]; other reviews can be found in References [127, 177, 211] where the authors present XAI methods for understanding the decision of medical imaging tasks in terms of clinical overflow.

Unlike the previous works, this article presents a comprehensive review of XAI techniques for medical data, including image, tabular, textual, and multimodal; provides a summary of evaluation metrics; highlights the strengths, weaknesses, and recommendations of each explanation category; and focuses on future research direction. This review covers various XAI factors that contribute to the principles of responsible AI, which are presented in Figure 1. As shown in the figure, understanding explainability (in terms of data, taxonomy, metrics, etc.) is the first step towards

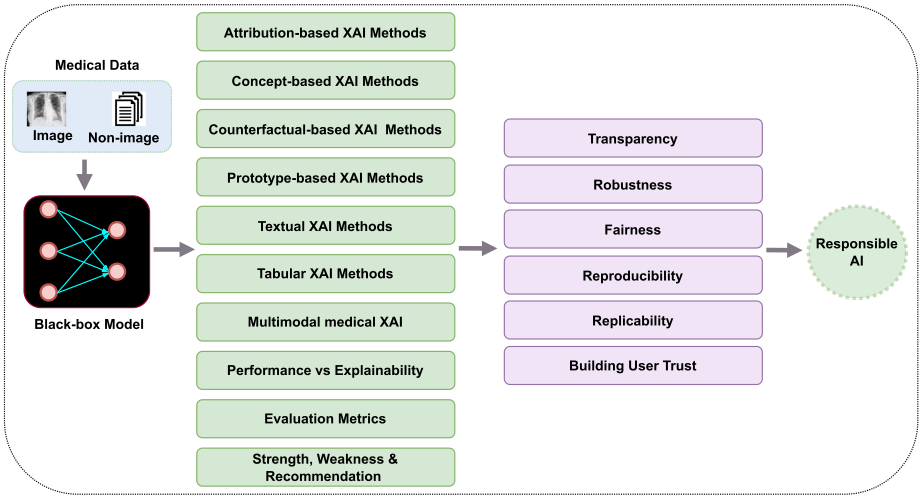

Fig. 1. Key XAI challenges discussed in this review, as well as their impact on the responsible AI principles.

developing transparent, robust, fair, and reproducible AI; i.e., explainability is the central pillar of responsible AI as providing an explanation of why AI models behave in a specific way can lead to better transparency while making it easier to detect wrong and unfair decisions, which is critical to building user's trust.

The main contributions of this review paper are summarized as follows:

— This article comprehensively reviews and analyses current XAI methods applied to various types of medical data, including image, text, tabular, and multimodal. It also introduces several XAI methods not yet utilized in medical data analysis but that have the potential to explain the outcome of DNNs with medical image input.
— It provides a taxonomy of XAI techniques for medical images, which clearly demonstrates the underlying relationships between different techniques and allows systematic analysis of explainable algorithms.
— It summarizes existing evaluation metrics to determine the validity and trustworthiness of different medical data explanation methods.
— It discusses the strengths and limitations of existing attribution-based, non-attribution-based, tabular, and textual XAI techniques to help researchers select the desired XAI method, depending on the application. It also highlights several future research opportunities and directions for developing different interpretability methods.

Before proceeding further, we want to note that we use the terms "interpretability" and "explainability" interchangeably, considering that there is a lack of consensus in defining these terms [8]. Miller [143] says interpretable ML refers to the degree to which a human can understand the cause of a decision (of a model). In Reference [96], interpretability is defined as the degree to which a human can consistently predict the model's result, while in Reference [36], interpretability is defined as a desirable quality or feature of an algorithm that provides enough expressive data to understand how the algorithm works. The Cambridge Dictionary defines it as follows: "If something is interpretable, it is possible to find its meaning or possible to find a particular meaning in it." Hence, explainability or interpretability can be defined as the additional information to reason the decision process, feature relevance, or identifying influential features generated by the ML model itself or another algorithm.

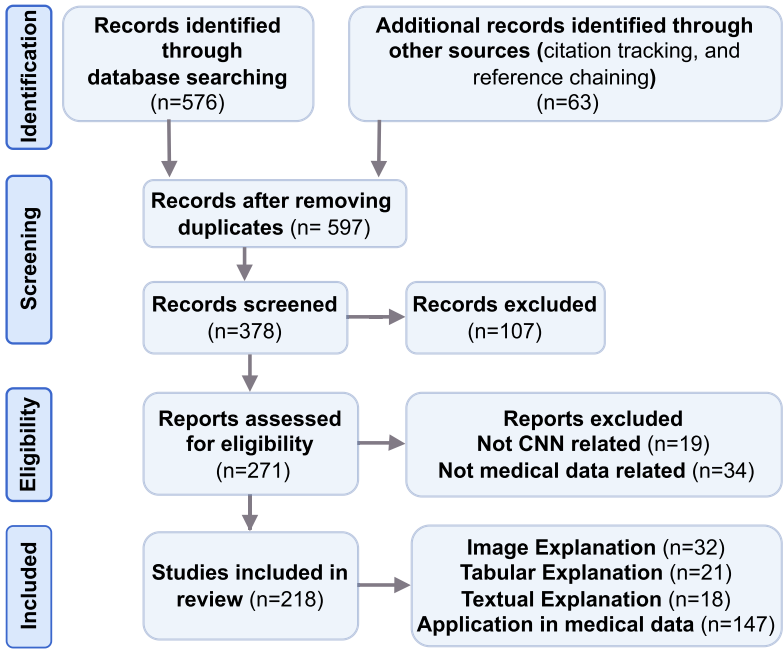

Fig. 2. PRISMA flow diagram of our review. It shows the number of papers identified, screened, assessed, and included in this review.

1.1 Literature Review Design

The search and selection process for this review are presented as a PRISMA [145] flow diagram in Figure 2. This systematic review is based on research articles collected using various search engines such as Google Scholar, IEEE Xplore, ACM Digital Library, Scopus, Pubmed, and CiteSeer. We retrieved relevant journal articles, scientific conferences, technical reports, and online articles published up to August 2023 using keywords such as explainable deep learning, XAI review/ survey, explainability, interpretability, explainable AI in the medical data, XAI in medical data, explainable machine learning, and explainability assessment/evaluation. We identified a total of 639 studies using the search keywords. We then filtered out irrelevant articles and selected relevant articles based on the following criteria: (1) the study is written in English; (2) the publication includes a novel or existing explainable method applicable to medical data; (3) the study contains published journal articles, conference papers, technical reports, open-access articles, or highly cited arXiv papers; (4) the study is related to explainable deep learning in medical data applications; and (5) the XAI method has the potential to be applied to medical data analysis, i.e., the method has been applied in several datasets, provides a visual explanation or explains the decision-making process, and shows significant results with existing evaluation metrics. We screened the identified paper carefully, included a total of 218 studies and excluded the study that failed to fulfill these criteria.

1.2 Organization of the Survey

This review is organized as follows. Section 2 provides a general taxonomy of XAI techniques for medical data analysis. In Section 3, we discuss four categories of XAI approaches for medical images based on the nature of their explanations and present applications of these methods in medical images. Section 4 discusses non-image-based explanation methods with applications in the medical field. We summarize the multimodal data explanations in Section 5. Various evaluation

metrics for XAI techniques are provided in Section 6, and, finally, limitations and future research directions are highlighted in Section 7.

2 TYPES OF XAI

Explainability follows the statement that no single algorithm is enough to solve all the problems better than every single algorithm. A hybrid approach often provides a concrete solution where multiple algorithms can be used. Explainability methods can be categorized into the following four categories.

2.1 Attribution vs. Non-Attribution

An attribution-based method generates a visual explanation by highlighting the image region associated with the model prediction [10, 25, 90, 149, 163, 170, 185, 190, 197, 202, 250]. A localization map can be produced by a gradient-based process [25, 185, 250] or perturbing the input pixel to identify important features such as occlusion [241], **Randomized Input Sampling for Explanation (RISE)** [163], and **Local Interpretable Model-agnostic Explanations (LIME)** [10]. The non-attribution method focuses on how and why a decision has been made and explains the prediction using the concept [55, 56, 97, 237], prototype [27, 102], and altered prediction [29, 37]. Instead of only focusing on the pixels, they also deal with model behavior, analyze model sensitivity and stability, or help in model debugging.

2.2 Intrinsic vs. Post Hoc

Intrinsic methods are usually integrated into the model and inherently interpretable and backed by various models (e.g., rule-based model [114] and decision tree [3]). Intrinsic explainers are directly dependent on the model architecture and are not usually transferable to another neural network without redesigning the XAI algorithm for a new model. However, the Post-Hoc XAI method is independent of the model architecture and can be applied to any trained CNN model. The accuracy of the existing network does not change; however, another algorithm is needed to explain. For example, attribution-based [25, 90, 149, 163, 170, 185, 190, 197, 202], concept learning explanation [55, 56, 97, 237] can explain externally of a trained network without sacrificing the prediction accuracy.

2.3 Local vs. Global

The local explanation method involves working on specific data instances and interpreting the reason for a decision of a classifier based on these data points. This explanation provides insights into specific features that positively or negatively impact the decision. Most of the perturbation-based methods such as **Contrastive Explanations Method (CEM)** [37], **Learning to Explain (L2X)** [29], saliency or attribution map methods, i.e., **Class Activation Mapping (CAM)** [250], GradCAM [185], and RISE [163], are in this category. However, the global explanation interprets the entire behavior of the model. It gives insights into the overall knowledge of the model. For instance, analyzing important features that help to improve the overall model performance is a global explanation method. Various complex networks adapted into the linear, tree-based, or rule-based models are also in this category, as they provide inherently global interpretation. Some of the global explanations methods include **Testing with Concept Activation Vectors (TCAV)** [97] and **Automatic Concept-based Explanations (ACE)** [55].

2.4 Model-specific vs. Model-agnostic

Model-specific methods are based on the internal model architecture and parameters and are appropriate for explaining specific structures (e.g., a specific CNN model). For example,

Table 1. A List of Different XAI Methods in Computer Vision Tasks, Especially for Image Analysis, and Their Categories

Category	XAI method and reference	Year	Remarks
Attribution-based	Layerwise relevance propagation (LRP) [10]	2015	Post hoc visual explanation methods highlight the image region that the DNN model thinks is important. Gradient-based and perturbation-based approaches are very popular for producing heat maps. The visualization-based method is easy to implement but has several limitations.
	CAM [250]	2016	
	LIME [170]	2016	
	SHAP (Shapley additive explanations) [133]	2016	
	DeepLIFT [190]	2017	
	Gradient-weighted CAM (Grad-CAM) [185]	2017	
	Integrated Gradient (IG) [202]	2017	
	SmoothGrad [197]	2017	
	GradCAm++ [25]	2018	
	RISE [163]	2018	
	Anchor [171]	2018	
	XRAI [90]	2019	
	Similarity Difference and Uniqueness [149]	2020	
Concept-based	TCAV [97]	2018	High-level concepts are manually or automatically selected and used to explain the prediction.
	ACE [55]	2019	
	Causal Concept Effect (CACE) [56]	2019	
	CONCEPT SHAP [237]	2020	
Counter-factual-based	CEM [37]	2018	Purtured input images are fed into the model to generate an altered prediction.
	L2X [29]	2018	
	Guided Prototypes [128]	2021	
	ABELE [60]		
Prototype-based	Maximum Mean Discrepancy critic (MMD-critic) [96]	2016	Prototypical part selection and comparison among train and test image
	Influence Function [102]	2017	
	Prototypical Part Network (ProtoPNet) [27]	2019	
Others	Decision Tree [245]	2019	Wavelet-based, tree, or feature correlation-based methods can be applied to medical images for interpretation.
	Rahul et el. [159]	2019	
	Adaptive Wavelet Distillation [66]	2021	
	Human–AI Interfaces [70]	2021	

GNNExplainer [239] is model specific due to its ability to only explain the graph neural network model. Neural additive model [3] is another example, which combines generalized additive models with the expressivity of DNNs to learn the relationships between the input and the output. However, the model agnostic method is independent of the model, can be applied to other domains, and does not deal with the model weights and the parameter directly. Different post hoc [25, 29, 37, 55, 56, 90, 97, 149, 163, 197, 237] XAI methods are in this category, as these methods are independent of the model architecture.

3 IMAGE-BASED XAI METHODS AND APPLICATIONS

XAI for image data can be broadly categorized into two main categories: attribution and non-attribution. Within the attribution-based methods, we further divide them into gradient-based, perturbation-based, and layerwise relevance propagation-based techniques. However, non-attribution methods can be divided into concept-based, counterfactual-based, and prototype-based techniques. Tables 1–3 lists techniques appropriate for medical image analysis.

3.1 Attribution-based Explanations

Saliency map (SM), used in Reference [192], generates heatmaps on the image predicted by a deep learning network. The purpose of an SM is to highlight the prominent region of an image on which a human's eye would focus first to recognize an object. An SM can be produced by either assigning a saliency value to every pixel or segmenting an image into several pixel groups and assigning a saliency value to every group of pixels.

3.1.1 Gradient-based Methods. These types produce attribution maps of input images using the gradient or backpropagation to highlight the important part of the prediction. The explanations are usually model agnostic and post hoc.

CAM [250] is one of the popular visualization methods, which creates an activation map and can localize the decision-making features on an image. CAM uses **global average pooling (GAP)** after the last convolutional layers and before the final **fully connected (FC)** layer. Given $f_k(x, y)$ corresponds to the activation of unit k (convolution filters) in the last layer at location (x, y) for a particular image. The weight corresponding to class c for unit k is w_c^k. Then, the input to the softmax layer for a given class c is computed as $S_c = \sum_{x,y} \sum_k w_c^k f_k(x, y)$ The class activation map M_c is computed as $M_c(x, y) = \sum_k w_C^k f_k(x, y)$, where M_c indicates how important the activation in the spatial location (x, y) is for classifying an image to class c. The class activation map is overlaid on the input image, and a highlighted region is generated for each class. CAM can be applied to the network, which sequentially has a GAP FC layer and softmax.

GradCAM [185] is a local explainer that solves the issues of CAM, i.e., no need for architectural modification or retraining the network. It uses the gradients of the network flowing into the last convolution layer to assign an important score to each neuron and produces a coarse localization map highlighting important pixels of an image to predict the target class. At first, GradCAM computes the gradient for each class score y_c with respect to the features maps activation A^k of the last convolution layer. Next, neuron importance scores w_k^c are calculated by averaging the gradient flowing back over the activation map's size Z. The weights w_k^c for a particular feature map A_k and class c is expressed as $w_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial Y^c}{\partial A_{ij}^k}$. The heatmap $L_{Grad-CAM}^c$ is generated by combining the weights and the forward activation map followed by a **rectified linear unit (ReLU)**, $L_{Grad-CAM}^c = ReLU(\sum_k w_k^c A^k)$. An up-sampling technique is employed via bilinear interpolation of the same size as the input images to produce the class activation map. The paper additionally presents Guided Grad-CAM, where a pointwise multiplication is done among upsampled saliency map L^c with the pixel-space visualization generated by Guided Backpropagation.

GradCAM++ [25] solves the shortcomings of GradCAM, i.e., provides better visual explanation and localizes when multiple objects are present in a single image. GradCAM++ reformulates the structure of weights w_k^c as in Equation (1) so that all relevant regions of the input image get equal priority,

$$w_k^c = \sum_i \sum_j \left[\frac{\frac{\partial^2 Y^c}{(\partial A_{ij}^k)^2}}{2\frac{\partial^2 Y^c}{(\partial A_{ij}^k)^2} + \sum_a \sum_b A_{ab}^k \{\frac{\partial^3 Y^c}{(\partial A_{ij}^k)^3}\}} \right] .relu\left(\frac{\partial Y^c}{\partial A_{ij}^k}\right). \tag{1}$$

Similarly to GradCAM, the saliency map is computed as $L_{ij}^c = ReLU(\sum_k w_k^c . A_{ij}^k)$ using ReLU and a forward activation map.

Integrated Gradient (IG) [202] requires no change of the original network and can evaluate the model performance, understand feature importance, and identify data skew. The method satisfies two axioms, i.e., sensitivity and implementation invariance. To evaluate the sensitivity, a baseline image, which could be a black/white or a random image, is considered as a starting point. If the

baseline image differs from the input image in one feature and if the network predicts a different outcome for the baseline and input image, then a differing feature should have a non-zero relevance score. Implementation invariance is satisfied when two networks whose predictions are identical for all inputs, despite implementation differences, have identical attributes for the same baseline and the input image. Formally, if a deep network F, an input x, a baseline input x', and the gradient $\frac{\partial F(x)}{\partial x_i}$, then the IG along ith dimension can be expressed as

$$IG_i(x) = (x_i - x_i') \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha(x - x'))}{\partial x_i} d\alpha, \tag{2}$$

where α is distributed in the range [0,1]. In practice, the definite integral computation is costly; therefore, Riemman approximation (3) solves the issue and can be written as

$$IG_i^{approx}(x) ::= (x_i - x_i') \sum_{k=1}^{m} \frac{\partial F(x' + \frac{k}{m}(x - x'))}{\partial x_i} \frac{1}{m}, \tag{3}$$

where m is the step size in the approximation of the integral, the weakness of the method is that IG does not give emphasis on global feature importance or does not interpret feature interactions.

Smoothgrad [197] sharpens the gradient-based saliency map by adding noise to the input image and smooths the saliency map when averaging is done on sensitivity maps created from these perturbed images. A saliency map for any input image (x) can be expressed as $M_c(x) = \partial S_c(x)/\partial x$, where $\partial S_c(x)$ is the derivation of the class function S_c for a class c. It signifies the effect of changing a tiny amount in each pixel of x on the classification score. Due to these local variations in $\partial S_c(x)/\partial x$, noise is added to the saliency map, i.e., raw gradient-based SM tends to be noisy. SmoothGrad uses a Gaussian Kernel to smooth the gradient. Hence, the new smoothed version, which takes the average of the sensitivity map can be formulated as follows:

$$\hat{M}_c(x) = \frac{1}{n} \sum_{1}^{n} M_c(x + \mathcal{N}(0, \sigma^2)), \tag{4}$$

where n is the number of samples, $\mathcal{N}(0, \sigma^2)$ is the Gaussian noise. Another advantage is that SmoothGrad can be combined with gradient-based methods like IG or vanilla gradient to improve visualization maps using the above averaging procedure.

XRAI [90] is an INTGRAD-based method that can be applied to any DNN model to determine the most salient image regions. The method is divided into three stages: segmentation, gathering attribution, and choosing appropriate regions. First, image segmentation is performed multiple times using different parameters, and therefore the attribution results do not depend on a specific hyper-parameter or image segmentation method. Second, the method uses IG with a baseline as a black-and-white average for generating an attribution map. Third, for selecting the region, XRAI leverages that the sum of all attributions for input is equal to the difference between input softmax and baseline softmax. Further, among the two regions, the one that sums more positively should get more priority from the classifier. For this consideration, XRAI selects an empty mask and then adds the region having maximum gain in the combined attribution per area. The algorithm runs until the mask is filled on the baseline image and the desired SM is obtained. Additionally, the paper introduces two metrics (softmax Information Curves and Accuracy Information Curves) for evaluating the attribution map quality.

Other earlier gradient-based methods like DeConvNet [241] consist of deconvolution and unpooling layers and generate a saliency map by zeroing the negative gradient during the backward pass. Guided BackPropagation [198] is a further improvement of the DeConvNet for a better visual explanation.

3.1.2 Perturbation-based Methods. These techniques measure the impact of perturbation of the different parts of an input image on the model's prediction. For example, the occlusion method [241] is used to visualize various parts of the image most closely associated with the classification. The authors found that the feature activity map also changed when a specific image region was occluded.

LIME was developed by Ribeiro et al. [170] in which the image is perturbed by changing on and off a certain amount of the super-pixel. The Quickshift segmentation algorithm generates the super-pixel; a uniform solid color is used to make it on and off. Subsequently, the perturbed image is fed into the neural network, and each perturbed image's prediction score is recorded. In the next step, the cosine metric helps to compute the distance between each perturbed image and the original image. A kernel function maps the distance between zero and one(weights). Afterward, a linear regression model is learned on top using the data and results from the perturbation, prediction score, and weights. Each learned coefficient from the linear model belongs to each super-pixel of the segmented image. This coefficient signifies the importance of each super-pixel for predicting a target class.

Anchor [171] is based on creating anchors or conditions by determining a decision rule that provides a stable explanation for a specific instance regardless of perturbation or model variations. The method uses the if-then rule to anchor a prediction to identify the important image segment such that the variation among the remaining segments does not influence the prediction. The image is segmented as LIME to observe variations of the model prediction, and a feature is an anchor feature if it consistently leads to a particular prediction. Anchor searches a set, $D = z|f(z) = f(x), z\epsilon x$, where z is part of the images, x is the input image, and $f(\cdot)$ is a black-box model [45].

RISE [163] is a model agnostic, i.e., it can generate the saliency map without model modification and can be applied directly to any existing network. In RISE, the input image x is perturbed via a random mask, and only a subset of the input image pixels is kept preserved. The masked images work as input to the base model, then the response for each masked image is recorded; more specifically, the confidence score is calculated by the black-box model on the masked image. The final saliency map is a result of combining the random binary mask M and combined weights coming from the output probabilities of the predicted class working on the masked image. When the mask pixel is important, the value $f(x \odot M)$ is high, where f denotes the black-box model, and $\odot$ indicates element-wise multiplication.

Similarity Difference and Uniqueness [149] is another visual explanation method based on the perturbation of the input image. First, masks are generated from the last CNN layer, converting the feature activation maps of class c, i.e., $f^c = [f_1^c, f_2^c \ldots\ldots f_N^c]$, into a binary mask M_i^c. Next, feature activation image mask A_c^i is obtained by a pointwise multiplication of the input image x and an interpolated binary mask M_i^c, which can be expressed as $A_c^i = F(x \odot M_i^c)$. Then probability prediction scores (P_i^c) and (P_{org}^c) for all the feature activation image masks and original input images are computed, respectively, to obtain the similarity difference, $SD_i^c = exp(\frac{-1}{2\sigma^2}||P_{org}^c - P_i^c||)$. The uniqueness (U_i^c) among the prediction score vectors of feature image masks is measured to figure out the high salient regions as $U_i^c = \sum_{j=1}^{N} ||P_i^c - P_j^c||$. Finally, feature importance weight W_i^c is computed for a visual explanation as $W_i^c = SD_i^c \cdot U_i^c$. The feature importance weight determines which feature has more influence in predicting a class.

3.1.3 Layerwise Relevance Propagation. **Layerwise relevance propagation (LRP)** [10] is post hoc and model specific and provides local explanations for image data. LRP uses the decomposition method to explain the neural network's prediction. It redistributes output or prediction $f(x)$ of the network f backward direction to input space with the help of local distribution rules, and in the process, a relevance score R_i is assigned to each pixel. Given i indexes neuron activation at a

particular layer l, pixelwise relevance scores R_j at layer $l+1$ and w_{ij} is the weights learning from the data of neuron i to neuron j, then the simple LRP rule using a local redistribution rule can be expressed as (5)

$$R_i^{(l)} = \sum_j \frac{x_i \cdot w_{ij}}{\sum_{i'} x_{i'} \cdot w_{i'j}} R_j^{(l+1)}. \tag{5}$$

If a neuron gets more activated, then it gets a larger share of redistributed relevance. If the relevance is redistributed from the network output to the input, then it is possible to find the important pixel of the image and generate the heat map or relevance map.

3.1.4 Other Methods. **Deep Learning Important FeaTures (DeepLIFT)** [190] is a backpropagation-based method similar to LRP and uses a reference input like the IG for explanation. It describes how the output changes from the baseline in terms of changes in the input. Formally, the target output t and reference output t_o and $\Delta t = t - t_o$, which is the difference-from-reference for output, $x_1, x_2, x_3 \ldots\ldots x_n$ is some neurons in the intermediate layer, and Δx is the difference of x and x', then the contribution scores $C_{\Delta x_i \Delta t}$ assigned by DeepLIFT for input feature x_i can be expressed as $\Delta t = \sum_{i=1}^{N} C_{\Delta x_i \Delta t}$, where N is the input neuron to compute t. Here, $C_{\Delta x_i \Delta t}$ can be thought of how much influence the difference-from-reference of x_i has on Δt. The contribution score can be computed by some rules like the linear rule, rescale rule, or revealcancel rule, which is elaborated on in the original paper. Additionally, the amount of relevance among input and output differences can be determined by a multiplier $m_{\Delta x \Delta t}$ as: $m_{\Delta x \Delta t} = \frac{C_{\Delta x \Delta t}}{\Delta x}$. It is possible to apply the chain rule for multipliers as written in (6) where $m_{\Delta x_i \Delta t}$ follows the multiplier equation's properties,

$$m_{\Delta x_i \Delta t} = \sum_j m_{\Delta x_i \Delta y_j} m_{\Delta y_j} \Delta t. \tag{6}$$

This chain rule helps to compute relevance scores layer by layer via backpropagation while the saliency map on the image is generated, similarly to the LRP method. Lundberg and Lee [133] first proposed **Shapley additive explanations (SHAP)** values for measuring feature importance and then introduced the Deep SHAP method by combining DeepLIFT and Shapley values for the explanation.

3.2 Attribution-based Explanations for Medical Images

Visual methods have been used in the medical field to aid clinicians by explaining the decision of a deep neural network (Table 2). For instance, in Reference [4], the authors utilized the saliency map produced by CAM to obtain superior performance over baseline results on COVID datasets. SnapMIX [75], an augmentation method, is used by utilizing the heatmap in each box generated on COVID-19 positive and negative images, where the label of the virtual training image is generated by combining the labels of these source images. Then, the paper delineates the link between whether CAM-heatmap is related to the obtained classification. Furthermore, they used the heatmap to validate the stability of the classification outcome by the location of the image most relevant to classification. In Reference [189], the authors produced high-resolution CAM by combining feature maps from multiple CNN layers and accurately localizing brain tumors on MRI data. DenseNet and CAM are used to produce robust visual explanations by improving the resolution of CAM for the brain gender classification [51]. Deep-learning-assisted CAM-based visualization has been used for classifications of knee MRI [20] and in knee pain assessment by generating heat maps of the MR image [24]. Another modification of the CAM, RESpond CAM [248], performed better than GradCAM for biomedical three-dimensional (3D) imaging inputs. Lee et al. [111] generated attention heatmaps of benign and metastatic lymph nodes by CAM and superimposed on the CT image to compare the position of the actual lymph node and the image region highlighted

Table 2. Application of Attribution-based Interpretability Methods to Highlight Salient Pixels in Medical Imaging

Methods	Modality	Reference	Application
CAM [250]	MRI	[20, 24, 51, 189]	CAM has been used for various applications such as COVID-19 detection, cancer identification, tumor classification, or even medical image augmentation purposes; CAM visualizes salient pixels in various locations i,g., Bladder Brain, Breast, Skin, Cardiovascular, Chest, Gastrointestinal, and Thyroid.
	CT	[106, 110–112, 134, 224]	
	Ultrasound	[122, 167, 220]	
	X-ray	[9, 41, 77, 94, 169]	
	Histology	[76, 95, 200]	
	Endoscopy	[221]	
	Dermatoscopy	[118]	
	Fundus Image	[85, 106], [124]	
GradCAM [185]	MRI	[119, 152, 162]	GradCAM visualizes the salient pixels in CNN-based image classifiers such as glaucoma detection from optical coherence tomography, COVID-19 detection, polyps classification, and pathological pattern diagnosis from endocytoscopic images.
	CT	[160, 164]	
	Ultrasound	[158]	
	X-ray	[23, 92, 121, 242]	
	Histology	[64, 84, 101, 104, 206]	
	OCT	[208]	
	Endoscopy	[80]	
	Fundus Image	[100, 139]	
LRP [10]	MRI	[22, 44, 58, 234]	LRP for Alzheimer's disease classification, diagnosing MS, discriminating schizophrenia.
	Histology	[67]	
SHAP [133]	MRI	[214]	Breast density estimation
	Dermatoscopy	[240]	Melanoma detection
Integrated Gradient [202]	MRI	[156]	Estrogen receptor status classification.
	Eye	[182]	DR severity level prediction.
Guided-back propagation [198]	MRI	[82]	Pathological region localization
	Endoscopy	[226]	Colorectal polyps segmentation
DeepLIFT [190]	MRI	[129]	Diagnosis of Multiple Sclerosis
LIME [170]	DaTscan	[137]	Parkinson's disease detection
	Chest Radiograph	[184]	Congestive heart failure visualization
SmoothGrad [197]	MRI	[156]	Classification of estrogen receptor
	Ultrasound	[54]	Interpretation of echocardiograms

Note: MRI = magnetic resonance imaging; CT: computed tomography, OCT: optical coherence tomography.

by the CNN model. Additionally, Rajpurkar et al. [169] developed CheXNeXt to detect different pathologies from the ChestX-ray8 dataset and used CAM to identify which region on the chest radiograph was most responsible for the final prediction.

In Reference [227], GradCAM focuses on the area containing a glioblastoma, vestibular schwannoma, or no tumor of brain MRI slices. The authors used GradCAM to visualize features and significant regions in whole-slide images to classify various types of polyps in Reference [104]. In Reference [208], the Grad-CAM visualization technique was applied to highlight the relevant image region for glaucoma detection from **optical coherence tomography (OCT)** probability map images and to understand the reason for ambiguity in false negatives and false positives. Besides, Itoh et al. [80] used a CNN-based classifier and Grad-CAM to visualize decision-reasoning regions of pathological pattern diagnosis from a large endocytoscopic image dataset. They

generated Grad-CAM visualization from three convolutional layers of the trained CNN to differentiate decision-making regions. Moreover, Grad-CAM and Guided Grad-CAM [121] explain the decision to detect COVID-19 from chest radiography images and identify the most discriminative image region.

Yan et al. [234] utilized LRP to distinguish schizophrenia patients, where LRP helped to identify functional network connectivity patterns in fMRI data. LRP has been used to diagnose multiple sclerosis by visualizing diagnosis-relevant features to make the CNN model transparent [44]. In addition, in Reference [58], the authors used 3D-CNN with LRP for neonatal magnetic resonance imaging. They analyzed the impact of different registration techniques on the image dataset using the generated relevance maps from the LRP. Furthermore, Boehle et al. [22] utilized LRP to produce a heatmap on MRI data in the image region responsible for Alzheimer's disease.

Tang et al. [205] proposed the Discovery CAM to visually explain the skin lesion and chest X-ray datasets. They used calibrated confidence methods to integrate the weights in the final classification layer to ensure that explanations align with doctors' ground truth. A model-agnostic, optimization-based, saliency method was proposed in Reference [138], using mammography and the curated breast imaging data, where a saliency map is produced in four stages: (i) classification score is obtained; (ii) a masked image is produced, inpainted, and classified; (iii) saliency loss is adjusted between the original image and map quality based on score difference; and (iv) optimization of saliency loss is performed several steps to get the resulting map. Further, an in-model explainer [173] was proposed to explain visually predicted class labels for the cervical cancer dataset. A VGG-based classifier was trained using the visual explanations created by an encoder–decoder, which allows the explainer only to use the pertinent parts of the image for classification.

In Reference [182], IG was used to generate explanatory heatmaps for diabetic retinopathy (DR) severity. This technique produces pixel-based maps that calculate each pixel contribution for a DR severity level prediction. Further, in Reference [129], DeepLIFT was used to identify salient features for MS classifications. The method was selected among other visual XAI algorithms depending on the quantitative evaluation of the trained model's perturbation. Further, Table 2 provides the application of visualization-based methods in medical image analysis. In addition, a comprehensive list showing the applications of visual XAI in the medical image according to anatomical location can be found in Reference [215].

3.3 Concept-based Explanations

A deep learning algorithm extracts low-level features from the image, like lines or edges, that are hard to understand for a human, as these features do not construct a meaningful high-level concept. Feature-based explanations applied to a black box model often create non-sensible interpretations [2]. However, concept-based explanations construct the explanation in a way humans can understand the behavior of a DNN model by evaluating concepts' contribution to conclude a particular decision [237]. For instance, stripes can be a concept for predicting a zebra.

TCAVs method is a global explanation method that uses human-friendly concepts to interpret the internal state of a trained CNN model. The concept is determined by the user, which could be input features from the training data or user-provided data. There is great flexibility to define and refine concepts even for non-expert ML model analysts as a hypotheses test is done during analysis. In this method, the concept is represented as a vector, and directional derivatives are applied to see how the concept vector affects the final classification score. The method needs high-level concepts (positive sets) and some random images (negative sets) to construct such a concept vector. To differentiate between activation layers of two sets, i.e., interested concepts and random examples, a binary linear classifier is trained. **Concept Activation Vector (CAV)** v_C^l for concept C in layer l is orthogonal to the decision-making boundaries for this binary linear classifier. CAV signifies

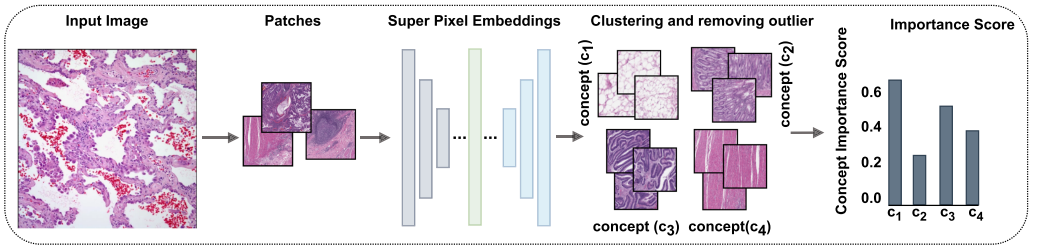

Fig. 3. Concept-based explanation, in which patches are obtained from the input image; human interpretable concepts (e.g., c_1, c_2, c_3, c_4) are selected from these patches; clustering of the concepts is performed, and the TCAV score is calculated, which determines how important these concepts are for a particular prediction. Input images before CNN are taken from the Cancer Genome Atlas (TCGA) database (https://www.cancer.gov/ccg/research/genome-sequencing/tcga).

the direction pointing away from the random examples to the concept of interest. For example, the directional derivative $\frac{dp(z)}{d(v_C^l)}$ of zebra class z determines how much a concept ('strips,' 'dotted line,' 'meshed') has an impact on the classification of zebra. TCAV score analyses the importance of each concept, i.e., if the directional derivatives change positively, then the concept is important; otherwise, not.

ACE [55] automatically extracts random examples and high-level visual concepts from the image presented in the training data. This method selects segments of multiple resolutions from the training images belonging to the same class. All the segments resized as original images are given as input to the CNN model. Similar segments are clustered in the activation space of the final layer using Euclidean distance, and low-similarity cluster segments are removed as an outlier. After that, an importance score like TCAV is computed for the clusters that contributed to the prediction of a class. Besides, the explanation results were evaluated quantitatively by human involvement in two experiments: (i) identifying the intruder concept and (ii) identifying the concept's meaning.

ConceptShap [237] is a post hoc, concept-based method used to enhance the interpretability of TCAV and ACE by answering how much a concept contributes to a prediction. It uses a concept discovery method to find a particular set of complete and interpretable concepts, where complete means the concept can fully explain the model's decision. The difference with the ACE method is that concepts are consistently clustered to certain coherent spatial regions to find the closeness with the concept and its nearest neighbors. If the concept vector with a high completeness score is $C_s = \{c_1, c_2, \ldots . c_n\}$, then the method uses Shapley Value [187] to find the importance of each concept for quantifying the important attributes of an image.

Causal Concept Effect (CACE) [56] can be defined as the causal effect of an understandable human concept on the DNN model's prediction. TCAV method faces difficulties finding concepts if the training data consists of multiple classes, biases in data, and the presence of color variation in the data. However, CACE explains the causal effect of the concept's presence or absence on the classifier's output. Additionally, the authors used **Variational Auto Encoders (VAEs)** for approximating VAE-CaCE, which estimates the causal effect of the generated concept. Experimentation shows that CACE performs better in the presence of biases and correlation in data.

3.4 Concept-based Explanations for Medical Images

Several studies have used concept-based explanations for medical image analysis (Table 3). Figure 3 shows a general overflow of how concept-based explanation can be applied where the final prediction is relevant to selected clinical concepts or deep features. Kim et al. [97] utilized

TCAV to explain the predictions of DR levels. The clinical concepts related to the diagnosis of DR level show a high TCAV score, whereas low scores for non-diagnostic concepts. For example, DR-level diagnosis by doctors depends on microaneurysms or aneurysms; therefore, these diagnostic concepts showed a comparatively high TCAV score. Further, the TCAV method is applied to cardiac MRI image data to provide clinically meaningful explanations for the predictions of a black-box classifier in Reference [83]. Moreover, MRI images are processed by selecting the region of interest or patches and extracting superpixels from the obtained patches. Next, the resized patches are fed into the network (UNET) to get activation from the middle layer. Then, concept clustering is done on the latent space of superpixels; therefore, it helps to get to CAVs defined as a perpendicular vector to the decision boundary between cluster data and random counterparts obtained by training a binary classifier.

In Reference [57], the authors extended TCAV into a regression problem as the diagnostic concepts are usually assessed continuously by calculating **regression concept vectors (RCVs)**. Usually, TCAV finds a discriminator between user-defined concepts and random images using directional derivatives, whereas they measured the direction of the greatest increase for a continuous concept. The method calculates the relevance of a concept using the RCV with the help of bidirectional relevance scores. Further, the authors detected tumor tissue in breast lymph node samples and found that nuclei texture was an important concept. Moreover, correlation and nuclei contrast were relevant for classifying breast tissue patches.

Clough et al. [33] identified the disease of cardiac MR segmentation and used a concept activation vector to interpret the decision concerning clinically standard measurements. The authors used a variational autoencoder to get a latent representation and a decoder to reconstruct the image for local explanation. In the intermediate layers of the network, TCAV was used to find which diagnostically meaningful or clinical biomarkers were most associated with cardiac disease. For example, filling rates and ventricular ejection had a significant score related to the disease; therefore, the network recognized these clinical features as significant. In another extension of TCAV, the author introduced a new metric called Uniform unit Ball surface Sampling [236] to show the concept's importance in CNN layers and explained the radiomics concept using mammographic images. Moreover, TCAV was used for skin lesion classification in Reference [130] to learn human understandable dermoscopic concepts such as pigment networks, streaks, dots and globules, blue-whitish veils, asymmetry, regression structure, and color. In Reference [50], the authors presented TCAV for interpreting the estimation of breast cancer biomarkers using six concepts, e.g., high-grade carcinoma, invasive lobular carcinoma, low-grade carcinoma, tumor-adjacent desmoplastic stromal changes, ductal carcinoma in situ, and tumor-infiltrating lymphocytes.

Fang et al. [46] proposed a **visual concept mining (VCM)** method based on human-understandable concepts to interpret infectious keratitis classification. VCM comprises two main components: (i) the concept generator automatically searches the relevant concepts and (ii) the visual concept extractor learns concepts' similarity and diversity by clustering the concepts of different classes. Moreover, Sauter et al. [181] used the ACE method to extract visual concepts automatically from digital histopathology data. They demonstrated that ACE helped to discover class-correlated bias, sampling bias, measurement bias, and class sampling ratio bias.

3.5 Counterfactual-based Explanations

A counterfactual explanation for images makes a minimal change in the images to alter a predefined output and interpret predictions of an individual instance [218]. VAE and **generative adversarial network (GAN)** are usually used to generate counterfactual explanations by realistic synthesis of the input images [88, 126]. The prediction depends on specific pixel variations or even

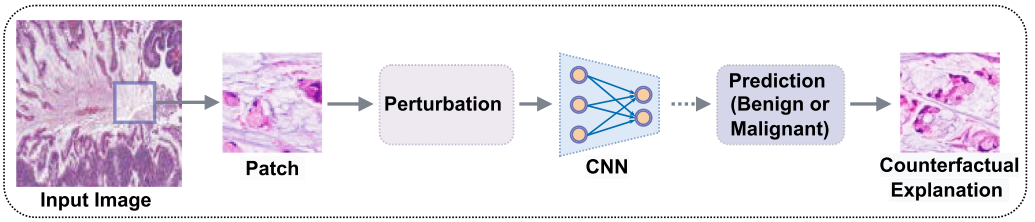

Fig. 4. An example of counterfactual explanation where the original image can be perturbed to alter the prediction. For example, a benign class can be predicted as malignant or vice versa. Input images before CNN are taken from the TCGA database.

on the whole altered image. For instance, predicting 03 from a counterfactual XAI is possible by perturbing the pixels of digit 08.

Guided by Prototypes [128] proposes a model-agnostic method that searches interpretable counterfactual instances to provide powerful insights about model prediction. It analyses the decision process to find the important features of a decision. The method demonstrates the changes in input features that can help alter the prediction of a predefined output. It perturbs the input features from the test data until the desired prediction is obtained. In the case of image perturbation, the pixel is modified until the closest image to the original one is found, and of course, a different classification result for the perturbed image is achieved. To get the meaningful perturbation, the objective loss function can be defined as $L = c \cdot L_{pred} + \beta \cdot L_1 + L_2 + L_{AE} + L_{proto}$, where the $c \cdot L_{pred}$ helps to predict another class using the perturbed images and other terms for regularization. In Figure 4, an application of the method is presented where perturbation of the image can alter the prediction, i.e., if the classifier predicts an image as benign, then the counterfactual explanation would give the output as malignant or vice versa.

CEM [37] is suitable when the classes are close, i.e., changing a few pixels alters class prediction. To explain the classification, the technique finds which pixels should be present or absent in an image. It uses the fact that an image x is classified as class y because the features $f_a \ldots\ldots f_d$ are present, and features $f_m \ldots . fp$ are absent, formally known as **pertinent positives (PP)** or **pertinent negatives (PN)**, respectively. Pertinent positives are the factors (pixel for the image) that lead to the same classification as original instances and pertinent negatives lead to different classes w.r.t original class. If an input x_0, then the perturbation δ, the modified input example $x \epsilon \chi$ is defined as $x = x_0 + \delta$, where χ represents data space, then PN's can be found using the following optimization equation,

$$\min_{\delta \epsilon \chi / x_0} c \cdot f_k^{neg}(x_0, \delta) + \beta ||\delta||_1 + ||\delta)||_2^2 + \gamma ||x_0 + \delta - AE(x)||_2^2), \tag{7}$$

where $f_k^{neg}(x_0, \delta)$ is the loss function that encourages the prediction to be a different class, $\beta ||\delta||_1 + ||\delta)||_2^2$ helps to select efficient features, and $||x_0 + \delta - AE(x)||_2^2$ is the reconstruction error of x when an autoencoder is used. For pertinent positives, features that are easily available in the input data are considered. PPs are also considered as an optimization problem that is as follows:

$$\min_{\delta \epsilon \chi \cap x_0} c \cdot f_k^{pos}(x_0, \delta) + \beta ||\delta||_1 + ||\delta)||_2^2 + \gamma ||\delta - AE(\delta)||_2^2), \tag{8}$$

where $\delta \epsilon \chi \cap x_0$ is the interpretable perturbation.

L2X [29] is a model agnostic approach that employs instance-wise feature selection, i.e., finds the importance score for each feature while predicting an instance. It extracts the most informative features for a given instance via mutual information. The technique computes mutual information

between selected features and response variables using a variational approximation and assigns a score for patches (groups of pixels). If the score for a selected patch is positive, then that patch positively impacts the prediction of an instance, and if the score is negative, then the patch is not the important one.

Adversarial Black box Explainer generating Latent Exemplars (ABELE) [60] is a local, model-agnostic XAI method that provides a set of exemplars (examples classified with the same label) and counter-exemplar images (counterfactuals) and generates the saliency map. The method turns the input image to its latent representation utilizing an adversarial autoencoder and generates a neighborhood in the latent space. A decision tree classifier is trained on the synthetic images found in the local neighborhood to mimic the behavior of the black box locally. The decision rule and counterfactual rules are extracted from the decision tree to generate the exemplars and counter-exemplars. The saliency map is created from the image region that contributes to the classification and the region that shifts to another class.

3.6 Counterfactual-based Explanations for Medical Images

The counterfactual explanation is suitable for the medical image (Table 3) because it synthesizes small, explainable changes to a query image to generate the desired alteration. For instance, the training calibration-based explainers [210] is used to interpret the model prediction for identifying pneumonia-related anomalies in the chest X-ray images. In this work, a normal class x is transformed into its latent representation using an encoder, and counterfactual $\bar{x}$ is learned in the latent space by a calibration-driven optimization such that the classifier's output changes to abnormal from the abnormal class.

Schutte et al. [183] presented an explanation method using StyleGAN [91] to predict knee osteoarthritis severity on X-ray images and tumor probability prediction on histology images based on alteration of the images to generate different outcomes. A StyleGAN is trained on input images to create a mapping between the image and latent vectors. The method finds the optimal direction in the latent representation to generate a fluctuation in model prediction. Afterward, the synthetic images for altering the prediction are generated by shifting the input image latent representation along the optimal direction. The generative method shows where the predictive features are located in the images and how they impact the prediction.

Furthermore, the counterfactual generative network [99] explains lesion prediction by generating different counterfactual examples from the input chest X-ray images. The framework generates counterfactual lesional images from query images using counterfactual manipulation training. The attribution maps of lesional regions are generated by subtracting the counterfactual image from the input images. Similarly, Singla et al. [195] provided counterfactual visual explanations that targeted three labels: cardiomegaly, pleural effusion, and edema for chest X-ray images based on conditional Generative Adversarial Network. The framework generates the counterfactuals from input images while preserving the anatomical shape and small details using specialized reconstruction loss.

Mertes et al. [140] proposed GANterfactual framework to generate counterfactual image explanations for pneumonia detection from X-ray images using an adversarial image-to-image translation algorithm. The authors used CycleGAN [251] with a modified counterfactual loss function to alter the input image such that the classifier predicted the incorrect class. Another counterfactual application using image-to-image translation based on CycleGAN [251] was applied to justify the decision of Diabetic Macular Edema prediction in Reference [151]. Further, in Reference [16], classification with feature attribution was proposed using VAE-GAN to disentangle class relevance features from the background for better interpretability in brain image datasets. Moreover, Gifsplanation [34] was proposed using an autoencoder and gradient update that can change the latent

representation of a particular input image. Multiple images were generated to produce a short video (gif) by enhancing or reducing the features utilized for a prediction to explain the decision of chest X-ray classifiers. Ghandeharioun et al. [52] presented a method named DISSECT that trains jointly a generator, a discriminator, and a concept disentangler to produce concept traversals defined as a series of instances with increasing degrees of concepts that influence a model's decision. The procedure was validated on melanoma classification, and it was found that large lesions, asymmetrical shapes, and jagged borders are responsible for prediction.

In Reference [184], **Generative Visual Rationales (GVRs)** identify the features of congestive heart failure on chest radiographs and detect bias and overfitted models. A neural network and a generative model were trained on the encoded representations of the labeled and unlabeled datasets, respectively, to estimate B-type natriuretic peptides. The GVR was created by comparing the reconstructed healthy radiograph and the diseased radiograph produced by the generative model. Additionally, ABELE [60] was implemented to offer the practitioner with explanation for skin lesion diagnosis, where Progressive Growing Adversarial Autoencoder was trained to reconstruct low-resolution skin lesion images [141]. The method provides medical exemplars and counterexemplars for the classification diagnosis to enhance interpretability.

3.7 Prototype-based Explanations

Prototype explanation helps users to reason the model's prediction by examining cases similar to the original model. Humans often use representative examples for categorization and decision-making [21]. Similarly, prototype-based explanation models employ representative examples to cluster and explain the data. For a particular class, the XAI finds the best-matched prototypical images.

Influence Functions [102] helps determine model behavior, model debugging, and detecting data errors. It analyzes the training data to find the most responsible training point for predicting a class. If the training examples $z_1, z_2 \ldots\ldots z_n$, where $z_i = (x_i, y_i)$, then x_i is the input images, y_i is the labels; if a single training image z is upweighted by a small amount, then the influence function gives an approximation for parameter $\theta \epsilon \Theta$ and the new changed parameter $\hat{\theta}_{\epsilon,z}$ is formulated as follows: $\hat{\theta}_{\epsilon,z} \stackrel{\text{def}}{=} arg\ min_{\theta \epsilon \Theta}(1-\epsilon)\frac{1}{n}\sum_{i=1}^{n} L(z_i,\theta) + \epsilon L(z,\theta)$, where loss $L(z,\theta)$ for a data point. In particular, at a test point z_{test}, the influence of upweighting z on the loss has the following closed-form expression:

$$\begin{aligned} \mathcal{L}_{up,loss}(z, z_{test}) &\stackrel{\text{def}}{=} \frac{dL(z_{test}, \hat{\theta}_{\epsilon,z})}{d\epsilon}|_{\epsilon=0} \\ &= -\nabla_\theta L(z_{test}, \hat{\theta})^T H_{\hat{\theta}}^{-1} \nabla_\theta L(z, \hat{\theta}), \end{aligned} \tag{9}$$

where $H_{\hat{\theta}} \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} \nabla_\theta^2 L(z_i, \hat{\theta})$ is the Hessian. So the influence of a training point would be higher if the upweighting of a training point has more impact on the loss parameter.

Prototypical Part Network (ProtoPNet) [27] is a post hoc explanation and a model-agnostic method that can be used to explain the image data. The method identifies some parts of the test image that look like some prototypical parts of the training image and makes a final classification based on a weighted combination of the similarity score between the learned prototype and parts of the test image. To illustrate, let $H \times W \times D$ be the shape of the convolutional output $f(x)$, for a given input image x, where $H = W = 7$ and D could be 128, 256, 512, using cross-validation. Then two additional 1×1 convolutional layers help to learn the prototype of spatial dimension 1×1 with the depth D of each prototype. Since the channel dimension is the same for both prototype and convolutional output, while the H and W of each prototype are smaller than the whole CNN output, each prototype denotes an activation pattern in a patch of CNN output, which eventually

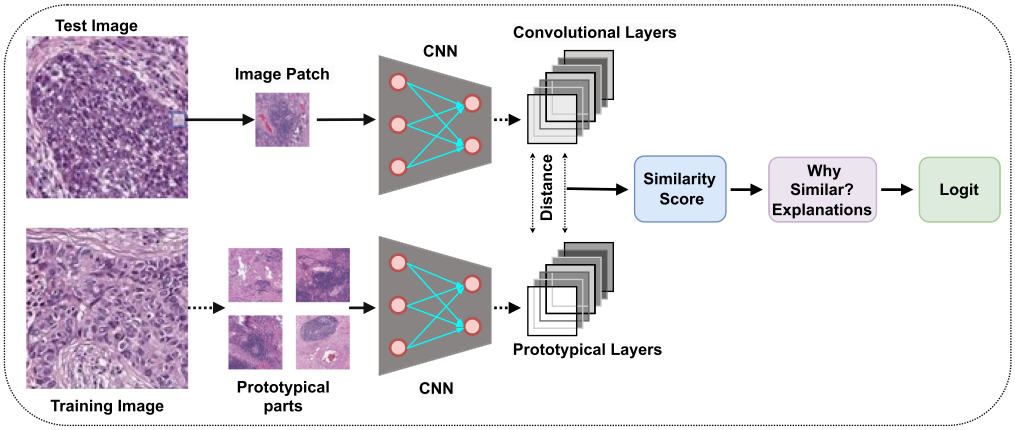

Fig. 5. Visual comparison of the patch from test image and learned prototypical patch from training data. The explanation is based on how these patches are similar in texture, contrast, shape, hue, and saturation. Training and test images before CNN are taken from the TCGA database.

corresponds to some prototypical image patch in the original pixel space. Hence, each prototype is the latent representation of some prototypical part of some training image. Afterward, a similarity score determines how strong a prototypical part is present in the image, while the heatmap highlights that part of the image.

Reference [153] is a modification of ProtoPNet, which explains why the model considers the prototype and image patch similar. The modified method is especially applicable when the similarity between the prototype and patch is not very noticeable. It enhances the explanation of a prototype with additional quantitative information about the visual characteristics of prototypes. Specifically, it figures out the influence of color hue, saturation, shape, texture, and contrast in a prototype, making it possible to understand why the model deems two images similar.

Furthermore, Donnelly et al. [39] introduced Deformable ProtoPNet based on ProtoPNet that can provide spatially adjustable deformable prototypes. A deformable prototype consists of several prototypical parts that alter their relative spatial positions to detect similar parts of an input image in an adaptive manner.

Maximum Mean Discrepancy critic (MMD-critic) [96] can help understand complex or unlabeled data distribution. It uses the prototype and criticism to characterize the dataset to enhance the explainability. Criticism is defined as the data point in the input space that is not recorded by prototypes or does not fit the model. It uses maximum mean discrepancy that measures the difference between the prototype distribution and the total data distribution, and criticisms are selected from parts of the dataset that are underrepresented by the prototypes. For dog classification, the MMD-critic learns reasonable prototypes, whereas the criticism selects dog categories with an unusual type, such as pictures of dogs with costumes, having the movement of dogs. When data distributions need to be well explained, prototypes, together with criticism, are helpful to facilitate human reasoning and understanding.

3.8 Prototype-based Explanations for Medical Images

Figure 5 represents the typical process of the prototypical explanation implementation for the medical image. Several studies used prototypical explanations in the medical image, as presented in Table 3. For example, the influence function was used in Reference [221] to explain lesion classification by identifying relevant features and showing which part of the image contains those features.

Table 3. A Summary of Different Non-attribution-based Methods' Applications for Medical Images

Methods	Modality	Paper	Remarks
Concept-based	MRI	[33]	Coronary artery disease detection
	Fundus image	[97]	DR prediction
	CT	[230]	Predicting the malignancy of lung nodules
	OCT	[207]	Glaucoma detection in optical coherence tomography
	MRI	[83]	Segmentation of left ventricle, right ventricle and myocardium
	Histopathology	[57]	
	MRI	[33]	Relevant concepts detection using Regression Concept Vectors
	Skin	[130]	
	Histology	[50]	Clinical biomarkers identification for cardiac disease
	Eye	[46]	learning human understandable dermoscopic concepts
	Histopathology	[181]	Estimation of breast cancer biomarkers
			Infectious keratitis classification
			Visual concepts extraction automatically for skin cancer
Counter-factual based	X-ray	[210]	Pneumonia-related anomalies identification
	X-ray	[183]	Knee osteoarthritis severity prediction
	X-ray	[138]	Categorization of the sample with masses or healthy
	X-ray	[195]	Lesion prediction
	X-ray	[52]	Detecting biases of a melanoma classifier
	X-ray	[99]	Lesion prediction by generating counterfactuals
	X-ray	[140]	Pneumonia detection
	MRI	[16]	Disentangling class relevance features
	X-ray	[34]	Mass prediction by generating gif
	Eye	[151]	Diabetic macular edema prediction
	X-ray	[184]	Identifying the features of congestive heart failure
	Skin	[141]	Skin lesion diagnosis
Influence function	MRI	[221]	Explanation of lesion classification
	Spectrogram	[73]	Neonatal pain assessment
Prototype-based	Endoscopy	[221]	Ulcer recognition in wireless capsule endoscopy
	Ultrasound	[117]	Thyroid nodule diagnosis
	Histology	[212]	Cancer or non-cancer classification
	Skin	[172]	Skin cancer recognition
	X-ray	[98]	Diagnosis in chest radiography
	X-ray	[193]	COVID detection
	X-ray	[15]	Mass lesions classification
	Skin	[14]	Melanoma classification
	X-ray	[74]	Similar image retrieval from to explain the chest X-ray
	X-ray	[191]	Localization of salient image regions

Radiologist features importance score was determined by the influence function for classifying a specific lesion.

Additionally, the influence function was employed in Reference [73] to explain neonatal pain assessment from the spectrogram image of the neonate's audio signal. The paper identifies the most helpful training examples to justify a decision and the harmful training images to remove bad-quality data. In Reference [212], the author explained the decision of the neural network to classify cancer from the pathological image using a prototype-based interpretation. The paper uses a VAE to encode the image patches and cluster the patches to get the prototype. The VAE decodes

the prototypes to obtain a visual interpretation, allowing humans to understand the component visually. In addition, based on prototype occurrences, a weight is calculated, which helps to ascertain the prototype's contribution to classifying cancer.

Contextual decomposition explanation penalization [172] explains skin cancer by ignoring the spurious examples in the training data. The paper uses an explanation term that compares the user's and model's interpretations. In another application [117], the author applied a prototype to interpret the decision in thyroid nodule diagnosis. The method inputs the feature maps extracted using the CNN into the prototype layer, which utilizes the built-in prototypes to generate similarity scores by comparing them with the feature map. The prototype having the largest similarity score contains the information of the target class.

Kim et al. [98] presented XProtoNet to understand disease-specific features for identifying disease in chest X-ray images. The technique adaptively predicts an occurrence area of disease by comparing it with the learned prototypes of an image. The author used transformation loss to generate a suitable occurrence map and L_1 loss for covering small occurrence areas to avoid irrelevant image regions. Singh et al. [193] introduced Generalized Prototypical Part Network for COVID detection from chest X-ray images that uses a generalized version of distance function L_2 to compute the similarity between the prototypes. Unlike ProtoPNet, the generalized L_2 function utilizes prototype parts of any dimension, such as rectangular spatial dimensions and squared spatial dimensions, for image classification. Furthermore, interpretable AI algorithm for breast lesions [15] compares test mammograms to prototypical images of different mass margin types for classifying mass lesions. The method learns medically relevant prototypes and localizes relevant areas in the images. Distance function L_2 compares each patch of the convolutional feature maps with each learned prototype and helps to generate similarity maps. Similarity scores from the similarity maps feed into the fully connected layers to predict benign or malignant lesions. Moreover, the activation precision metric was introduced to measure the proportion of pertinent data marked by the radiologist-annotator used to classify mass margin.

Another similarity-based method was proposed in Reference [14] for melanoma classification using similar image retrieval to justify the diagnosis. The method uses a loss function with regularization terms to improve the feature space and interpretability in a clinical workflow. While, in Reference [68], a search tool for similar histopathology images was developed based on retrieving similar histologic features, cancer grades, and organ sites. Hu et al. [74] applied similarity-based saliency maps to explain the chest X-ray image retrieval approach. The method takes a retrieval and query image to identify the most similar regions by deletion and insertion metrics. Moreover, in Reference [191], the authors investigated the application of interpretability to localize salient image regions for better feature representations and medical image retrieval. Experiments on the chest X-ray dataset demonstrate that the method supports retrieval with visual explanations, improves the class consistency of the retrieved images, and improves the interpretability of the entire system.

3.9 Other XAI Approaches

This section provides additional XAI methods based on decision trees, wavelet transforms, feature correlation, and human–AI interfaces.

Interpreting CNNs via Decision Trees [245] aims to transform the convoluted features from the different filters inside a neural network into semantically meaningful concepts and evaluate which filter or part of the image contributes how much to a prediction. A prediction is obtained by bridging the middle layer features into the semantic meaningful concepts. The decision tree indicates all decision modes of a CNN, where the root node mostly consists of common decision modes and represents prediction rationales shared by many images. Further, each leaf node represents a

specific decision mode of an individual image, while nodes close to the leaves correspond to modes shared by minority images. Decision modes help to explain how the filters/parts of the image are associated with the final prediction.

Describing deep features through the lens of radiologist features [159], the authors have explained deep features extracted from the last layers before the classification layer of a CNN with respect to semantic features and traditional quantitative features. An experienced radiologist generates semantic features from a CT scan of a lung tumor and includes the common characteristics of a tumor. Quantitative features are extracted from a tumor phenotype by Definiens software and a radiologist, which provides information about the nodule, such as nodule size, histogram-based information, and pixel intensity. Moreover, wrapper feature selection and the random forest were applied to traditional quantitative or semantic features, respectively, to select the best subset of features. The Pearson correlation coefficient was calculated for each semantic feature with the deep features, and the five most correlated features for each semantic feature were selected. Then, each semantic feature is replaced with the correlated deep features and checked whether the same classification accuracy could be achieved. Since the same original classification accuracy is obtained with traditional quantitative features, shape-based quantitative features explain deep features.

The Adaptive Wavelet Distillation [66] method distills knowledge from a trained DNN into a valid wavelet transform. The method employs three types of losses in the formulation of the techniques: The reconstruction loss permits the reconstruction of the initial data and guarantees that the wavelet transform is invertible, the wavelet loss ensures an accurate wavelet transformation, and the interpretation loss enables knowledge distillation into the wavelet model from the pretrained model. During the transformation process, the wavelet coefficients help to explain the model's prediction. An application of the methods is molecular partner prediction using the LSTM model, where the reasonable feature is extracted from the process, like the maximum value of the times series or trace length.

Human–AI Interfaces to Support Explainability Reference [70] highlights the fact that a human expert in the loop can play a crucial role in medical XAI by providing the experience and conceptual knowledge of what an AI system cannot do alone. Figure 6 shows a process of explanation where explanatory statements s are acquired by a machine s_m or a human s_h where s is a function of r, k, c, i.e., $s = f(r, k, c)$. Here r is the representation of an unknown fact u_e associated with an entity, where k represents preexisting knowledge for a machine that is embedded in an algorithm, and for a human, it is made up of implicit, explicit, or/and tacit knowledge, and c is the context, which for a human, the physical environment where the decision is made and for a machine is the runtime environment. The unknown entity u_e represents the ground truth g_t, modeled by human m_h or a machine m_m. The goal is to determine whether a human can understand the explanation given by the machine and to what extent a human can understand it. In an ideal scenario, both the machine and human statements are homogeneous ($m_h = m_m$) and similar to the ground truth. For example, if a pathologist is satisfied with the result of the XAI algorithm, then $m_h = m_m$ is satisfied, meaning a congruence exists between humans and machines.

4 NON-IMAGE-BASED XAI METHODS AND APPLICATIONS

In this section, we provide the explanation methods for non-image-based data such as tabular, textual, time-series, and sequential data.

4.1 Tabular Data Explanation

We briefly present a list of tabular data explanation methods for black-box algorithms that can be implemented for medical data, such as clinical records and patient metadata. We follow the same

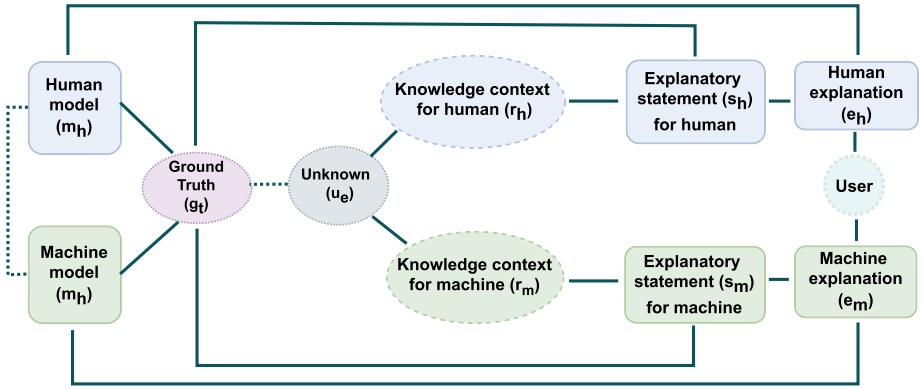

Fig. 6. Explanation from human and model need to be congruent with explanatory statements that are related to ground truth [70].

taxonomy as in Reference [21], where the authors present four types of explanations for tabular data based on (i) Features Importance, (ii) Rule, (iii) Counterfactual, and (iv) Prototype.

Features Importance-based XAI Method assigns an importance score to each feature to understand the features' contribution to a prediction. Higher importance scores imply a feature has more influence on the prediction, whereas lower importance scores imply a feature has less influence. We discuss some commonly used XAI methods for determining feature importance. For example, LIME [170], already described in Section 3.1.2, returns justifications as feature importance score. LIME creates a local interpretable model (decision tree or linear model) trained on a new dataset based on randomly selected data points close to the instance being explained, which approximate interpretable representations of the actual data. Finally, Least Absolute Shrinkage and Selection Operator regularisation is employed to maintain only the most required features, and regression coefficients are utilized as feature importance scores. Another similar approach, MAPLE [165], uses a local linear modeling technique and a random forest for neighborhood selection where the linear model's coefficients determine each feature's local impact.

SHAP [133] based on cooperative game theory is an additive feature attribution technique that linearly combines simplified input features. SHAP follows three desirable properties: (i) *local accuracy* ensures the model's prediction for a specific instance should not deviate mainly from the expected average prediction for simplified inputs; (ii) *missingness* necessitates that features that are absent from the original input have no attributed influence on SHAP values if the simplified inputs serve to indicate feature presence; and (iii) *consistency* states that if a model changes, then the contribution of a simplified input should increase, and the SHAP value should also increase. Moreover, **Neural Additive Models (NAM)** [3] are designed to combine the deep neural networks' predictability with additive models' inherent interpretability. NAMs train multiple neural network architectures in an additive manner to interpret the contributions of each input feature. Additionally, in Reference [32], **Equi-explanation Maps (EEM)** was proposed to summarize the logic of the black-box model to provide a concise representation of global explanations by splitting the desirable explanation features region into subspaces depending on similar logic. Further, LRP [10], discussed in Section 3.1.3, attributes an importance score to each feature depending on its contribution to the final model prediction using local propagation rules backward through the model layers.

The rule-based XAI method provides explanations based on predefined rules to the end user by reasoning about the final model prediction. For example, **Local Rule-Based Explanations**

Table 4. A List of XAI Techniques That Can Be Employed to Explain Medical Tabular Data

Category	XAI method and reference	Data Type	Remarks
Features Importance-based	LRP [10]	Any	The explainer assigns an importance score to each feature for a prediction. A higher importance score indicates that the feature contributes positively to predictions.
	Local Interpretable Model-agnostic Explanations	Any Tabular	
	MAPLE [165]	Any	
	SHAP [133]	Tabular	
	NAMs [3]	Tabular	
	EEM [32]		
Rule-based	LORE [60]	Tabular	Rule-based explainer extracts rules or counterfactual rules for the end-user about the decision process.
	ANCHOR [171]	Any	
	RuleMatrix [144]	Tabular	
	GLObal to loCAL eXplainer [186]	Tabular	
Counterfactual-based	CEM [37]	Tabular	Counterfactual-based explainers modify the input features to produce an altered prediction. It identifies the features' dependency that led to a specific decision.
	Diverse Counterfactual Explanations [147]	Tabular	
	C-CHVAE [161]	Tabular	
	Actionable REcourse Summaries [115]	Tabular	
	Feasible and Actionable Counterfactual Explanations [166]		
Prototype-based	Prototype Selection [19]	Tabular	Prototype-based explainers identify similar data or examples belonging to the targeted class for a prediction.
	MMD-critic [96]	Text	
	PROTODASH [65]	Text	
	Tree Space Prototype [204]	Tabular	

(LORE) [60] uses a genetic algorithm for generating a synthetic neighborhood of the data point being explained. An explanation utilizing the neighborhood includes a decision rule and a set of counterfactual rules. The counterfactual rules illustrate the conditions that can be changed on input data to alter the prediction and enable the user to characterize the instance and its neighbor. ANCHOR [171], discussed in Section 3.1.2, provides rules as explanations using if-then rules such that if some features represent precision conditions or rules for a prediction, then any modifications to other features should not influence model prediction. Another rule-based tabular explanation method named RuleMatrix [144], a matrix-based rule visualizer to verify the rules. The technique extracts and filters a list of rules based on confidence and support thresholds that approximate a trained classifier. Then, a visual interface allows users to analyze and navigate the rules because the model may contain too many rules. Then, a visual interface helps users analyze and navigate the rules since the rule-based method may contain too many rules or a complex rules composition. Furthermore, GLObal to loCAL eXplainer [186] develops global explanation by hierarchically aggregating local explanations from local decision rules.

Counterfactual-based explanations make a small change to the input data to impact a classifier prediction positively from the user's perspective. For example, Diverse Counterfactual Explanations [147] returns a set of diverse counterfactuals by solving an optimization problem to provide feasibility and diversity. The feasibility ensures that the counterfactuals follow user constraints and contexts, while diversity in generated counterfactuals provides various ways of altering the outcome class by adding a regularization term in the loss function to penalize similar counterfactuals. Moreover, CEM [37], described in Section 5, uses PP and PN to generate contrastive explanations. Another model-agnostic tabular data explainer named **Counterfactual Conditional**

Table 5. A List of XAI Techniques That Can Be Employed to Explain Medical Text Data

Category	XAI method and reference	Data Type	Remarks
Sentence Highlighting	LRP [10]	Any	Post hoc explanation methods assign an importance score to every word or group of words and highlight the words in the sentence for a prediction. The image-based attribution approach needs a slight modification to suit the text-based scenario.
	SHAP [133]	Any	
	DeepLIFT [190]	Any	
	IG [202]	Any	
	L2X [29]	Any	
	XRAI [90]	Any	
	Local Interpretation Of Neural nETworkS through penultimate layer decoding [29]	Any	
	LIME [170]	Any	
Attention-based	Vaswani et al. [216]	Text	Attention mechanisms highlight the salient words, and attention weight ranks tokens in a sentence.
	Li et al. [116]	Text	
	exBERT [72]	Text	
Others	DoctorXAI [155]	Any	Textual explanations methods that may not fit in sentence highlighting or attention-based. DoctorXAI, LORE, and ANCHOR are rule-based explanations, while XSPELLS and LASTS find exemplars and counter-exemplars.
	ANCHOR [171]	Any	
	XSPELLS [108]	Text	
	LASTS [62]	Text	
	LORE [61]	Any	
	CAT [26]	Text	
	POLYJUICE [231]	Text	
	QUINT [1]	Text	
	Rajani et al. [168]	Text	

Heterogeneous Autoencoder (C-CHVAE) [161] utilizes an autoencoder on heterogeneous data for counterfactuals generation. A distance function in the real input space is not required for C-CHVAE to generate counterfactuals since measuring meaningful distance in tabular data is difficult. Feasible and Actionable Counterfactual Explanations [166] returns actionable counterfactuals that are coherent with the underlying data distribution and connected via feasible paths. These paths are the shortest path determined via density-weighted metrics. Moreover, Actionable REcourse Summaries [115] provides global counterfactual explanations for a whole reference population.

The prototype-based method identifies a small subset of samples that can view as a condensed representation of whole data. For instance, MMD-critic [96], described in Section 3.7, produces prototypes and criticisms to increase the explanations for complex data distribution using Maximum Mean Discrepancy. The data points that are closer to the data distribution are known as prototypes, and those that are further away are known as criticisms. Prototypes define the dataset's overall behavior, while criticisms are recordings that the prototypes fail to explain adequately. A variation of MMD-critic is PROTODASH [65], which represents the significance of each prototype by returning non-negative weights. Prototype Selection [19] solves a set cover optimization problem and summarizes the dataset using selecting prototypical instances from the dataset. Afterward, a nearest-neighbor rule is applied to the group of prototypes so that the method can be utilized as a classifier. Moreover, Tree Space Prototype [204] selects prototypes adaptively from the dataset to explain tree ensemble methods.

4.2 Textual Data Explanation

Combining image and textual XAI techniques can provide a comprehensive way of explaining the black-box model compared to the image explanation alone. The textual XAI can take the form

of report generation and report generation with a visual explanation for an input image. Most textual explanations employ (i) sentence highlighting that assigns each word or token a predicted importance score and highlights that word or token in the sentences and (ii) CNN for feature extraction, and RNN for generating the word sentences.

Some attribution-based methods can be modified to obtain sentence highlighting for text explanation. LIME [170] can be adopted for highlighting important words in a sentence. Lime generates a neighborhood of sentences from an input sentence by randomly replacing one or more words with blank spaces. Then, a linear model is trained using the perturbed texts that can mimic the original black box model, i.e., both models should predict the identical class labels. The coefficients from the linear model indicate the significance of each word.

Integrated Gradient (IG) [202], discussed in Section 3.1, requires a baseline image to generate explanation, while for text-based networks, the baseline could be a zero embedding vector. The baseline calculates the saliency value of a specific word by evaluating the word's contribution to the model output in the absence of any input information.

Deep Learning Important FeaTures (DeepLIFT) [190], described in Section 3.1, can also be applied for sentence highlighting, employing the same principle of INTGRAD. L2X [29] L2X can be adapted to generate sentence highlighting explanations for text data. In the case of text, the patches are the group of words or phrases. L2X assigns importance scores to each phrase rather than individual words for generating explanations. Local Interpretation Of Neural nETworkS through penultimate layer decoding [146] constructs a local neighborhood of input texts at the penultimate layer and records the decision of the network for the neighborhood and provides an explanation using a linear model like LIME. Furthermore, LRP [10], XRAI [90], and SHAP [133] can be used to explain the black-box model for textual data.

The attention-based method was introduced by Xu et al. [233] to demonstrate which parts of the images are most important in realizing the caption. The attention mechanism directs the model to learn attention weight, which aids in understanding the context of all the information in a sentence. Li et al. [116] employed an attention-based method to generate a visual explanation and error analysis by identifying the impact of erasing words. The weights of the attention layers of RNN provide the importance score for every word in a sentence. A higher score indicates redder highlighting, and the model is more sensitive to the deletion of a specific word. Furthermore, exBERT [72] delivers users to explore the model's reasoning process by providing insights about the attention weights and internal representation.

Further, we provide additional explainers that might not fall under sentence highlighting or attention-based. For example, ANCHOR [171] can be used to provide text explanations at the sentence level by identifying the important word related to a model's decision. In the case of text, a sentence is perturbed by modifying or removing certain words to make an interpretable representation that includes specific tokens (words). Next, the anchor or high-precision rule is established such that specific words from the perturbed sentence help to predict a particular class with high confidence. An example of such words could be in the case of 'This book is not bad.' where mode predicts 'positive' is 'not' and 'bad.'

XSPELLS [108] is a model-agnostic local method consisting of exemplar and counter-exemplar sentences as explanations. The method generates meaningful synthetic texts from the latent space using VAE and creates neighbors of the text to explain. The randomly generated neighbors are used to learn a decision tree that helps to characterize exemplar and counter-exemplar texts. LORE [60] utilizes a local interpretable predictor to explain a specific instance in the form of rules and counterfactual rules. Given an instance x and black-box b where $b(x) = y$, LORE applies a genetic algorithm to produce a balanced set of synthetic neighbors Z of that instance such that for some instances, $b(z) = b(x)$, while for other cases, $b(z) \neq b(x)$. Then, the method constructs a decision

tree using the balanced set Z and derives the decision rule explaining the cause for the decision and a set of counterfactual rules identifying conditions leading to an altered outcome. Additionally, POLYJUICE [231] is a model-agnostic that produces textual counterfactuals for the purpose of the explanation. Other methods, such as LASTS [62], DoctorXAI [155], QUINT [1], and CAT [26] can be adopted to explain the textual data.

4.3 Explanation for Other Data Types

LASTS [62] is a modification of ABELE for time-series classification. The explanation by LASTS consists of (i) a shapelet-based rule and shapelet-based counterfactual rules and (ii) a set with exemplars and counter-exemplar. Like ABELE, LASTS also utilize generated neighborhood in the latent feature space using the autoencoder to learn a latent decision tree that provides a local decision rule and counter-factual rules. Such rules identify exemplars and counter-exemplars and the associated reconstructed time series is used for learning shapelets. Finally, to extract a shapelet-based rule and counterfactual rules, a shapelet-based tree is learned from the reconstructed exemplars and counter-exemplars.

Panigutti et al. [155] introduced DoctorXAI as a model-agnostic and local explainer based on sequential, multi-labeled, and ontology-linked data to predict patients' next visit time. Doctor XAI perturbs the sequential data to create a local synthetic neighborhood of a patient by using the semantic data encoded in the ontology and uses a black-box algorithm for labeling. The health-record data of such semantic patients is then transformed into a format suited for decision tree training. Doctor XAI uses such a trained decision tree to extract a rule-based explanation.

4.4 XAI Applications in Non-image Medical Data

In this section, we present the application of the explainable algorithm in non-image medical data (Table 6). For example, in Reference [249], TRACER was introduced for interpretable risk prediction of acute kidney injury. TRACER relies on an RNN-based model that learns both time-variant and invariant feature importance for each patient using a featurewise transformation mechanism and a self-attention mechanism, respectively. In Reference [107], the authors used an attention mechanism and visualization-based explanation for heart failure disease diagnosis from **electronic medical records (EMRs)** data. In Reference [199], Time-aware and Co-occurrence-aware Network was proposed based on a self-attention mechanism and time-aware gated recurrent unit, which interprets the prediction and a diagnosis graph for the patient. In Reference [154], an interpretable COVID-19 diagnosis framework was developed from cough sounds and symptoms metadata. The framework integrates the attention mechanism and shows that the most recurrent symptoms for infected patients are fever, dizziness, cough, and chest pain. At the same time, the **t-distributed Stochastic Neighbor Embedding (t-SNE)** [213] finds the most discriminating cough sound features for each class. In Reference [188], an RNN-based survival model named DeepAISE was developed for periodical sepsis prediction. Feature important scores were computed as model explanations to find the top contributing factors to individual sepsis risk w.r.t. input features, similar to saliency maps for CNN. Another attention-based explanation [148] was used to detect **atrial fibrillation (AF)** using a single-lead ECG signal. The system learns clinically meaningful components for the detection task from input signals, such as waves and heartbeats.

A SHAP-based explanation for eye state detection was used in Reference [203] using an **electroencephalogram (EEG)**. The results of a gradient-boosted tree model and a DNN model were compared to evaluate the interpretability of each model. Shapley values indicate that both models shared the same top three important features with a slight variation. Moreover, in Reference [38], the paper used **Mel-Frequency Cepstral Coefficient (MFCC)** features from PCG signals for heart anomaly detection. The framework utilizes a pre-trained LSTM model to segment the

Table 6. A List of XAI Applications in Non-image Medical Data

XAI Method	Data Type	Paper	Remarks
Attention-based	EMRs	[107]	Heart failure diagnosis using the RNN model.
Attention-based	Longitudinal EMR	[249]	Acute kidney injury risk prediction.
Attention-based	Electrocardiogram (ECG) signals	[148]	AF detection.
SHAP	EEG	[203]	Eye state detection using EEG signal.
SHAP	Electronic Health Record (EHR)	[28]	Forecasting adverse surgical events.
SHAP	Phonocardiogram (PCG) signals	[38]	Heart anomaly detection using MFCC features.
SHAP	Electronic Health Record (EHR)	[150]	Risk of mortality prediction using GRU.
SHAP	Genomic data	[228]	Interpretable cancer classification.
Attention	Cough sounds, and symptoms	[154]	COVID-19 diagnosis from sounds and metadata.
Saliency Map	Electronic Health Record (EHR)	[188]	Periodical sepsis prediction.
GradCAM	Electromyography (EMG)	[63]	Neuro-robotic systems.
Attention-based	Electronic Health Record (EHR)	[199]	Mortality prediction, disease prediction, readmission prediction, and diagnosis prediction.
LIME	Electrocardiogram (ECG) signals	[81]	AF detection.
LRP	3D motion data	[48]	Freeze of Gait movement detection.

input MFCC and a CNN learns spatial features. Additionally, both SHAP and occlusion maps explain the hidden representations of the model and reveal that a correct PCG signal classification occurs when the model focuses on the features between fundamental heart sound regions (S1 and S2 segments).

LIME has been used to explain the time series for AF detection from single-lead ECG signal [81]. The framework highlights the most relevant segments of the signal that matched with the cardiologists' identified features for AF detection, e.g., the absence of P-wave, electrical activity, and variability of R-R intervals. Further, Reference [48], in LRP, was used to explain the DNN decisions for detecting movement that anticipates **freezing of gait (FOG)** in Parkinson's disease. LRP visualizes the underlying characteristics that CNN considers crucial for modeling the pathology. Moreover, the method found that the most relevant features that characterize the movement preceding FOG are fixed knee extension during the stance period, fixed ankle dorsiflexion, and reduced peak knee flexion during the wing phase.

5 MULTIMODAL XAI APPLICATION IN MEDICAL DATA

Healthcare applications inherently require a multimodal approach due to the interconnection and diversity of data sources involved. The multi-modal system combines various imaging modalities such as CT, MRI, and PET or blending of varied data types such as images, 1D signals, clinical records, and demographic metadata rather than focusing on a single data type. Different unimodal XAI techniques can be extended to the multimodal explanation approach and can be applied in medical fields (Table 7). For example, DIME [135] extends the core idea of LIME to explain a multimodal model. The method determines the dominant factor of a multimodal prediction by disentangling the model into unimodal contributions and multimodal interactions. More specifically,

DIME determines the importance of each modality from the linear model weights similar to the LIME technique and utilizes the weights to generate interpretable visualization.

In this study, we focus on the significance of multimodal XAI in healthcare applications such as report generation from images, medical image captioning with visual explanations, patient metadata or tabular data, or a combination of imaging data (e.g., X-rays, MRIs, and CT scans) with health records. Image captioning can be viewed as creating a textual description for interpreting the model decision, commonly used in Natural Language Processing tasks. Sun et al. [201] employed the joint model (CNN-RNN) to generate medical imaging reports from mammography data where the CNN produces a global feature of the image, and RNN decodes the text matching. Likewise, Singh et al. [194] developed an encoder–decoder framework using CNN as an encoder and a stacked LSTM as a decoder to generate radiology reports automatically from chest X-rays images. Zhang et al. [247] proposed a multimodal approach composed of a language model and image embedding model, MDNET, that takes pathology bladder cancer images as input, produces diagnostic reports, retrieves images based on symptom descriptions, and generates network attention. An auxiliary attention sharpening module improves the attention maps to highlight carcinoma-informative regions. Jing et al. [86] proposed a framework for multi-task learning that includes a co-attention mechanism for generating text according to the regions of radiology and pathology images containing abnormalities. Moreover, the authors develop a hierarchical LSTM network that captures long-range semantics effectively and generates high-quality long reports. Wang et al. [225] proposed Text-Image Embedding network for extracting the most distinctive parts from chest X-ray images and text representations. The authors showed how various parts of the radiological findings correspond to multiple saliency maps in the image. Similarly, Barata et al. [13] proposed an explainable hierarchical model using attention modules that mimic hierarchical decisions made by a dermatologist. The channel and spatial attention can identify relevant features and regions in the dermoscopy images to improve the explainability. Lee et al. [109] provided textual explanation and visual justification to explain the diagnostic decision of a breast masses classifier based on the CNN-RNN model. Similarly, Gale et al. [49] proposed a model agnostic method based on the LSTM model to generate descriptive sentences to explain the decision of a CNN classifier. Furthermore, a visual attention mechanism highlights the relevant region for detecting hip fractures in pelvic X-rays. In Reference [238], the authors presented an attention-based framework for detecting abnormalities and generating medical reports automatically from chest X-ray images. A hierarchical RNN model consists of a sentence RNN, which produces topic vectors, and a word RNN generates an appropriate sentence using the topic vectors. Moreover, a matching mechanism maps the topic vectors and sentences into the same semantic space to create more accurate reports. Similarly, Liu et al. [125] presented a hierarchical generation strategy using CNN-RNN-RNN architecture to produce topics from radiological images and then generate words from topics. The method uses a fine-tuning technique employing reinforcement learning to produce a more coherent report. Chen et al. [31] proposed a memory-driven transformer model to generate radiology reports from chest X-ray images.

Furthermore, Das et al. [35] proposed visual dialog, where a human interacts with a machine about an image's visual content. Specifically, when provided with an image, a conversation history, and a question, the AI agent infers a context and provides a correct response to the inquiry. In [105], the visual dialog was implemented in the public radiology dataset, RadVisDial. The authors described how AI agents could be practically useful and clinically beneficial to chest X-ray images.

Additionally, in Reference [235], the authors implemented attribution-based XAI for image classification for COVID-19 patients and segmentation for hydrocephalus patients using CT and MRI datasets via multi-modal and multi-center data fusion. For explaining classification, a modified CAM highlights the salient pixels and locates the infected area in CT images. Kernel SHAP (as

Table 7. A Brief Summary of XAI Applications in Multimodal Setting for Medical Data

XAI Methods used/based on	Modality	Paper	Remarks
SHAP + grad-CAM	Skin + Metadata	[222]	Skin lesion diagnosis based on patient metadata.
Report generation (CNN-RNN)	X-ray + Text	[201]	Medical reports from mammography data.
Report generation (CNN-RNN)	X-ray + Text	[194]	Generation of radiology reports automatically from chest X-ray images.
Visual Dialog (CNN-RNN)	X-ray + Text	[105]	Generation of visual dialog in radiology.
Concept Activation Vectors	Dermatology + Text	[131]	Visual and textual XAI for melanoma classification.
Kernel SHAP + CAM + t-SNE	CT + MRI	[235]	COVID-19 detection and segmentation.
Diagnostic report generation (CNN-RNN)	Pathology + Text	[247]	Report generation, image retrieval, and visual attention to provide justifications.
Generation of medical reports (CNN-RNN)	X-ray + Text	[86]	Generating text from image regions containing abnormalities and generating long reports.
Multi-level attention (CNN-RNN)	X-ray + Text	[225]	Report generation, highlighting the important words and image regions.
SHAP	Histology + Genomic data	[13]	Morphologic and molecular correlation of prognosis for 14 cancer types.
Attention-based	X-ray + Tabular	[30]	COVID-19 classification.
Visual and textual justification (CNN-RNN)	X-ray + Text	[109]	Explanation of the diagnostic decision for a breast masses classifier.
Radiology report generation	X-ray + Text	[49]	Descriptive sentence generation and an attention mechanism for detecting hip fractures.
Generation of medical reports (CNN-RNN)	X-ray + Text	[238]	Abnormalities detection and medical reports generation from chest X-ray images.
Report generation (CNN-RNN-RNN)	X-ray + Text	[125]	Chest X-ray report generation using reinforcement learning.
Radiology report generation	X-ray + Text	[31]	Radiology reports by memory-driven transformer.
Image Captioning	X-ray + Text	[174]	X-ray captioning and pathology localization.
LIME	MRI + gene data	[89]	Alzheimer's disease (AD) classification.
Image Captioning	Histology + Text	[136]	Report generation with visual interpretation.

described in Reference [133]) assesses the contributions of each super-pixel. Specifically, super-pixels associated with lesion areas influence the positive prediction, while those corresponding to disease-free areas sway the prediction negatively. Further, to explain brain ventricle segmentation, UNET with ResNet was trained on multi-modal MRI data obtained from hydrocephalus patients. The PCA [229] method projects the encoder's lowest bottom features into a 2D latent space for feature space visualization. t-SNE [213] visualizes the learned features and distribution of the targeted image, revealing the method's weaknesses and strengths. In Reference [131], a framework termed ExAID was proposed for a multi-modal explanation consisting of visual maps and textual explanations for melanoma classification from dermoscopic images. ExAID utilized CAVs for dermatological concept identification and the TCAV method for estimating the impact of a particular concept on a model decision. The framework used **Concept Localization Map (CLM)** [132] to

locate the learned concept in the trained classifier's latent space. CLM uses perturbation-based concept localization to highlight the region connected with learned concepts, thus generating visual explanations. Further, ExAID provides textual explanations from concept predictions and directional derivatives utilized in TCAV. Wang et al. [222] proposed IM-CNN for skin lesion diagnosis based on patient metadata and skin lesion images. Metadata comprises the patient's basic information (location, age, sex) and extracted features (globule, pigment network, border irregularity, and asymmetry) from skin images. SHAP analyzes the features of the metadata and assigns an importance score to each feature, whereas grad-CAM generates an interpretable visual output of skin lesion images.

6 EVALUATION OF XAI METHODS

Due to the absence of ground truth for explanations, current explanation methods do not follow a specific set of evaluation metrics. Therefore, the researchers' decision for selecting the evaluation methods depends highly on the application. We present below some of the evaluation techniques that are widely used by the research community.

A Taxonomy of Evaluation: A framework proposed in Reference [40] consists of three evaluation methods: (i) *application grounded* , which requires a human expert for a specific application, e.g., doctors performing diagnoses; (ii) *human grounded*, which is applicable when the task is simple and does not require an expert human while still requires a lay human to evaluate the quality of interpretations, e.g., a normal human evaluating the quality of saliency maps; and (iii) *functionally grounded*, which requires no human—however, it requires proxy tasks for evaluation. Comparing the results of the explanation method with radiologists' manual delineations of tumors could be an example. This evaluation is suitable when the method is not mature yet or human subject involvement is unethical.

Attribution Map Evaluation: Some of the commonly used evaluation metrics for attribution methods are given below: Pointing Game Metric [243] extracts a maximum point from the saliency map. If the point lies on the bounding box of a specified class, then a hit is considered; otherwise, a miss. Then, the accuracy is obtained from the number of miss and hit for each object as $Accuracy = \frac{\#Hits}{\#Misses+\#Hits}$. Deletion and insertion [179] are saliency map metrics where the former finds the degraded class probability score when important pixels are removed, and the latter measures increased class probability when pixels are inserted. Hooker et al. [71] claimed that perturbing the highest-scoring regions cause a distribution shift in the data; hence they introduced RemOve And Retrain, which retrains on the perturbed images and checks whether accuracy drops or increases. Moreover, in Reference [175], the Remove and Debias metric eliminates the need for retraining and reduces computational costs using the mutual information between low-important pixels and the class.

The Attribution Localisation Metric [103] is calculated as the ratio of the sum of positive attributions in the bounding box to the overall attribution. Ghorbani et al. [53] used Top-k Intersection as an evaluation metric, which calculates the intersection size of the most relevant feature of the original and perturbed image. A high score is preferable as the overlap between the two masks should be high. Another similar metric is the concept influence score [209], which provides insights into which semantic visual concept influences prediction and measures the amount of pixelwise intersection between an explanation and segmentation map using top-k features.

Two other metrics are described in Reference [7]: (i) *Relevance Mass Accuracy* is computed as the ratio of the sum of positive or relevance values within the bounding box over the sum of overall positive attributions of the entire image, and (ii) *Relevance Rank Accuracy* determines how much high-intensity relevance is included inside the ground-truth mask. Moreover, Area Under the Curve [47] is another common metric for visual explanations. In Reference [244], **Area over**

Perturbation Curve (AOPC) [179] evaluates the saliency map produced from ultrasound images where a high AOPC value indicates the heatmap is, in fact, relevant. Given an input image x, N is the total number of images, K is perturbation steps, and $f(x)$ specifies the certainty of an object's presence in the image x, then AOPC is the difference between $f(x)$ scores with and without perturbation. Mathematically, AOPC is defined as (10)

$$AOPC = \frac{1}{N}\sum_{n=0}^{N}(f(x_n)^{(0)} - \frac{1}{K}\sum_{n=0}^{K} f(x_n)^{(k)}). \tag{10}$$

Zhang et al. [244] adapted AOPC for the quantitative evaluation of attributions maps for interpretability of a CNN for fetal head circumference estimation. In the case of model parameter randomization, the metrics are Model Parameter Randomization [2] and Random Logit Metric [196]. Furthermore, robustness metrics such as Local Lipschitz Estimate [6] determine the consistency of the explanation for similar instances. To evaluate the faithfulness of the methods, faithfulness correlation [18], Pixel-Flipping [10], and Region Perturbation [179] can be used.

Evaluation of Human-AI Interface Tool: Holzinger et al. [69] introduced the **System Causability Scale (SCS)** to measure the quality of explanation in the case of human–AI interfaces. The technique evaluates the user's perception of an explanation process using the usability combined with causability. SCS was applied to measure the characteristics and quality of the human–AI interface in coronary artery disease estimation.

Evaluation Metrics for Counterfactual Explanation: The metrics utilized in Reference [195] for the counterfactual explanation are **Counterfactual Validity (CV)** [147], **Frechet Inception Distance (FID)**, and **Foreign Object Preservation (FOP)**. CV score determines whether counterfactual explanation is associated with classifier prediction, i.e., if the classifier predicts an image as normal, then counterfactual prediction should be as abnormal by the classifier. FID quantifies the visual quality of the explanations by calculating the feature distance between the input and counterfactual image. FOP metric checks the retention of the individual patient information in the explanations. Looveren et al. [128] used IM1 and IM2 metrics to evaluate the interpretability where IM1 calculates the reconstruction errors ratio between counterfactual instances, and IM2 compares similarities among reconstructed counterfactual instances.

Quantitative Evaluation of Concept Learning Model: Concept-based models such as TCAV may produce meaningless concept activation vectors if the random concept is selected. To avoid spurious results, a statistical significance test for a specific concept provides the stability of a particular concept [97]. A two-sided t-test of the TCAV score can determine the resulting concepts relevant to a class prediction when the null hypothesis is rejected. Furthermore, human involvement in evaluating the concept ensures the correct selection of the concept [55, 237].

Evaluation of Tabular Data Explanation Methods: XAI metric selection is an open issue that depends on the specific application or goal of the evaluation. For example, in Reference [21], the authors used fidelity and stability for tabular data explanation methods. Fidelity [59] evaluates how well the methods can resemble or mimic the decision-making process of the black box model. Stability evaluates the consistency or coherence of explanations when dealing with similar instances. Stability for an input x, and explanation model $f_e(x)$ can be calculated using Lipschitz constant [5] as $L_x = max\frac{||f_e(x)-f_e(x')||}{||x-x'||}, \forall_{x'} \in \mathcal{N}_x$, where $\mathcal{N}_x$ is the neighbourhood of instance x.

Furthermore, distance metrics can be used to compute the quality of explanations. Euclidean distance or pairwise distance can quantify how much the XAI method-generated explanation deviates from the ground-truth explanation. Root Mean Squared Error and Area Under the Curve are two common evaluation metrics that can be employed based on the specific characteristics of the problem [3, 65]. Moreover, most quantitative evaluations measure the model performance variation

by adding or removing the features through the input perturbation process [5]. A robust explanation should be robust to the minor perturbation of input data. Attribution and non-attribution evaluation methods can be adapted according to particular applications for tabular explanations.

Evaluation of Textual Explanation Methods: Commonly used metrics for textual evaluations are **Bilingual Evaluation Understudy (BLEU)** [157], **Metric for Evaluation of Translation with Explicit Ordering (METEOR)** [12], **Recall-Oriented Understudy for Gisting Evaluation (ROUGE)** [120], and **Consensus-Based Image Description Evaluation (CIDEr)** [217].

BLEU is a standard metric for textual explanation evaluation in natural language processing tasks, which calculates the similarity between the generated sentence and the ground-truth sentence. The metric count n-grams (sequences of n words) overlap between the generated and the reference sentence. The BLEU is often calculated using n-grams up to size N, where $N = 1, 2, 3, 4$, and the scores are geometrically averaged. Formally, the BLEU score is determined using the following formula: $BLEU = BP.exp(\sum_{n=1}^{N} w_n log p_n)$, where the brevity penalty (BP) penalizes shorter generated sentences than the reference sentence, w_n is the weight and p_n is the modified n-gram precision. The BLEU score is between 0 and 1; closer to 1 means better translation quality.

The ROUGE package consists of metrics (ROUGE-N, ROUGE-L, ROUGE-W, and ROUGE-S) for automatically evaluating generated sentences compared to reference sentences. ROUGE-L is commonly used, which measures the **Longest Common Subsequence (LCS)** between the generated sentence and the reference sentences regarding recall R and precision P. F1-based ROUGE-L can be calculated between two sentences x of length m and y of length n as $ROUGE-L_{f1} = 2\frac{P_{lcs}.Rlcs}{P_{lcs}+Rlcs}$, where $P_{lcs} = \frac{LCS(x,y)}{m}$ and $Rlcs = \frac{LCS(x,y)}{n}$. The reference and generated sentences should match exactly for the ROUGE scores to be 1, which range from 0 to 1.

Another commonly used metric is METEOR, which utilizes synonyms, stemming, and paraphrases to evaluate the quality of translations. METEOR score is computed as follows: (i) an F-score is computed as $Fmean = \frac{10PR}{(R+9P)}$, (ii) a fragmentation penalty M to ensure the diversity in the translation is given by $M = 0.5\frac{\#chunks}{\#unigram-matched}$, and (iii) finally the score, $METEOR = Fmean(1 - M)$.

Additionally, CIDEr compares the similarity between a human-generated ground-truth sentence and the generated sentence. The CIDE score for n-grams utilizes average cosine similarity between the ground-truth sentence and the generated sentence in terms of both precision and recall. Furthermore, another cosine similarity-based method named BERTSCORE [246] measures the similarity of the two sentences based on embeddings obtained from pre-trained BERT.

7 LIMITATIONS AND FUTURE DIRECTIONS

This article reviews XAI methods with a specific focus on medical data. Although these methods show good to excellent results, the question is whether these methods can be implemented in the clinical setting. Our extensive review of XAI methods uncovers several challenges and considerations associated with these methods that warrant careful evaluation before their integration into medical data applications. We summarize these challenges (Table 8) and propose several directions for future research.

Limited Clinical Applications of Attribution Map: Most existing saliency map-based methods highlight the important pixels of an image. However, these methods failed in various evaluation testing. For example, the saliency map's effectiveness was assessed in the context of medical imaging. To achieve this, eight techniques were utilized, specifically trained on the SIIM-ACR Pneumothorax Segmentation and RSNA Pneumonia Detection datasets. The evaluation was based on four key trustworthiness criteria: utility, sensitivity to weight randomization, repeatability, and reproducibility. The techniques examined were Gradient Explanation, Smoothgrad, IG, Smooth

Table 8. An Overview of the Strengths, Weaknesses, and Recommendations of Different Types of XAI Methods

Category	Strength	Weakness	Recommendation
Attribution-based	Post hoc, and local explanations; easy to understand with visual interpretation; simple implementation; determines specific features' importance.	Inconsistent heatmap generation; repeatability and reproducibility issues; no information about the decision-making process; sensitivity to perturbations.	A non-visual method may be applied simultaneously to examine the congruence with the visual XAI based on the application's context.
Concept-based	High-level concepts explain the model's decision; human involvement ensures the selected concepts are accurate; provide better interpretable insights, better analysis of the model's behavior.	Concept selection is subjective and relies on humans; additional annotation costs; concepts may suffer the completeness issue; in-depth mathematical knowledge is required.	Easy to implement but recommended in a human-in-the-loop system; comparison of different models' behavior using the same concepts.
Counter-factual-based	Provides causal relationships between input features and prediction; highlights image regions where the model is less robust; identifies bias and vulnerability.	Generating counterfactuals is computationally expensive; perturbation of images could be unrealistic; good data quality is required.	Suitable to employ when minimum changes of features lead to altered classification or when identifying salient features is essential.
Prototype-based	Easy to understand; both global and local explanation; able to find bad quality training data; errors and biases identification.	Expensive and difficult to train; susceptible to noise or artifacts; human involvement in validating the correct prototype.	Explaining individual test points; when the end-users are non-experts.
Tabular-based	Most important features identification; behavior analysis of different models; outlier features detection; bias detection.	Tabular XAI might not accurately explain highly correlated features; struggle to identify the most contributing factor for high-dimensional data.	Applicable if tabular data or metadata are available; Multiple XAI methods should be applied to compare explanation results for robustness.
Textual-based	Easily understandable to non-experts; both global and local interpretations; enhance the model transparency through interactive feedback.	Generating coherent explanations can be challenging; limited datasets; require more annotation cost, bias, or explanation variations.	Suitable if text annotation is available; recommended if the end-user is non-expert.

IG, GradCAM, XRAI, Guided-backdrop, and Guided GradCAM. Interestingly, none of these methods satisfied all the criteria set forth. The benchmarks for this evaluation were the area under the precision–recall curve and the structural similarity index. However, a standout observation was that XRAI excelled in terms of localization, utility, repeatability, and reproducibility when crafting a saliency map for radiology data. Further, the attribution-based method failed against the robustness test in several experiments. The robustness of four visualization-based methods, namely guided backpropagation, gradient input, LRP, and Occlusion, was tested for Alzheimer's disease classification [43]. Repeatedly training a CNN with the identical setting showed that these four methods generated an inconsistent heatmap. Moreover, saliency maps are susceptible to adversarial attacks [53], i.e., a small perturbation in the input image may generate a large variation in the saliency map, though the output prediction remains the same. The bias term is also non-negligible towards making the attribution map, which correlates with the output prediction [223]. Further research shows that attribution methods fail in randomization tests. For example, Adebayo et al. [2] showed that Guided Backprop and Guided GradCAM methods could produce the visual explanation without proper training. Hence, visualization-based methods need to be evaluated with caution when applied to medical images, and future research should focus on developing attribution systems that are more robust, effective, reproducible, and capable of producing consistent saliency maps. Another major issue of the gradient-based methods (e.g., integrated gradients)

depends on selecting a reference point, such as a baseline image. Hence, future research should focus on converting the reference point to a hyperparameter.

Limitation of Non-attribution Explanation Methods: The shortcomings of existing non-attribution (e.g., concept learning, counterfactual, or prototype-based) methods are computationally expensive, need a domain expert, or require high annotation cost. For example, the main limitation of concept learning models is that they need additional annotation costs because they require a human to select concept examples. Some other drawbacks are that a misleading explanation is possible due to confounding concepts, and concepts may not causally affect the model's prediction. In some cases, where the methods (e.g., ACE) extract visual concepts automatically from the image, hence, no human in the loop is necessary, while the concepts may suffer the completeness issue [237]. To illustrate, the method may select ten concepts with high selection scores; however, they may not be sufficient to explain the predictions. A limited method (e.g., ConceptSHAP) can solve the above-mentioned issues of selecting concepts [56] by adapting game-theoretic concept, which needs in-depth mathematical knowledge to understand. Hence, the future direction of concept learning should be developing less complex techniques, requiring minimum human involvement, and automatically selecting robust, complete concepts.

The counterfactual explanation limitation is that the targeted image perturbation process may be unrealistic. Besides, an autoencoder is needed to generate counterfactual images; therefore, a good representation is limited due to poor data quality or insufficient data. Hence, importance should be given to developing image perturbation process approaches.

Furthermore, prototype-based methods take a long time to train as they compare the test image with every input in the train set and are susceptible to noise or compression artifacts. In some XAI methods, a human is needed to validate the reason for prediction. In that case, the main limitation is that a specialist's requirement adds additional cost. Therefore, future investigations should emphasize these issues and design improved prototype-based methods capable of achieving maximum performance.

Limitation of Tabular Explanation: The majority of tabular explanations methods provide local explanations, but fail to provide explanations of the global behavior of the black-box model. In addition, due to the difficulty of distinguishing a specific feature's contribution, tabular XAI approaches may not explain accurately if the features are highly correlated. Further, if the tabular data is high-dimensional, i.e., data with high features, then the explanation methods might struggle to identify the most contributing features, or their contributions may not be accurately captured.

Limitation of Textual Explanation: The availability of the dataset plays a vital role in advancing the textual explanation in the medical field or report generation from the images. A domain-specific expert with language specialization can generate the ground-truth sentences, hence the additional annotation cost. Furthermore, generating coherent explanations may be challenging because various word choices or phrasings could result in various explanations. Therefore, in the future, emphasis can be placed on gathering expert-generated ground-truth statements and creating new XAI algorithms for coherent and trustworthy reports.

Lack of Evaluation Metrics: Another future direction of could be developing evaluation techniques. Although several evaluation metrics exist, these are not established, and mostly adopted directly from deep learning or computer vision metrics. One of the reasons for limited quantitative evaluation metrics is the lack of ground truth for the explanation. Hence, an enormous opportunity exists to develop XAI evaluation techniques as the field is still immature.

Interpretability vs. Accuracy Tradeoff: A common myth in deep learning research is that a tradeoff exists between interpretability and accuracy; i.e., a model with high accuracy usually offers less explainability and vice versa. Such belief has confined researchers to producing more explainable models. However, the interpretability vs. accuracy tradeoff is reversed; i.e., the model with

more interpretability results in better accuracy [176]. Hence, future research can be implementing and designing XAI algorithms that have high explainability ability as well as high performance.

Complex Architectures and Multimodal Datasets: Another research direction could focus on assessing the quality of existing XAI methods on more complex deep models or datasets. For instance, while the influence function yields precise outcomes for shallow networks, its results tend to be inaccurate when applied to deeper networks, as discussed in Reference [17]. Further, the performance of an existing method is often examined on a simpler dataset (e.g., MNIST, CIFAR-10), while medical image datasets are often complex and have different characteristics . Hence, current XAI methods should be examined for complex architectures and multimodal datasets.

8 CONCLUSION

Explainable AI for understanding the decision-making process of deep neural networks is crucial in critical applications (e.g., medical data analysis). In this article, we explored existing XAI techniques and summarized several applications of these methods to medical data such as image, tabular, textual, and multimodal. We divided medical image XAI methods into four categories: attribution-based, concept learning-based, counterfactual-based, and prototype-based explanations; at the same time, we presented XAI methods for tabular and text data and their application in the medical field. We further explained the importance of interpretable ML research and clarified different concepts, definitions, and a taxonomy underlying DNN model explainability, which has revolved around medical data, particularly medical images. The paper also summarizes commonly used evaluation metrics for evaluating XAI methods. Further, this article outlines each category's current challenges and limitations to assist researchers in selecting the XAI method carefully according to their problem and dataset. Finally, we highlighted the potential future directions for explainable research in medical data.

In summary, this article provided a comprehensive review of current XAI techniques applied to medical data, addressed the limitations, and provided future research directions to obtain a trustworthy and responsible AI systems.

REFERENCES

[1] Abdalghani Abujabal, Rishiraj Saha Roy, Mohamed Yahya, and Gerhard Weikum. 2017. Quint: Interpretable question answering over knowledge bases. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 61–66.

[2] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. 2018. Sanity checks for saliency maps. *Adv. Neural Inf. Process. Syst.* 31 (2018).

[3] Rishabh Agarwal, Levi Melnick, Nicholas Frosst, Xuezhou Zhang, Ben Lengerich, Rich Caruana, and Geoffrey E Hinton. 2021. Neural additive models: Interpretable machine learning with neural nets. *Adv. Neural Inf. Process. Syst.* 34 (2021).

[4] Ameen Ali, Tal Shaharabany, and Lior Wolf. 2021. Explainability guided multi-site COVID-19 CT Classification. arXiv:2103.13677. Retrieved from https://arxiv.org/abs/2103.13677

[5] David Alvarez Melis and Tommi Jaakkola. 2018. Towards robust interpretability with self-explaining neural networks. *Adv. Neural Inf. Process. Syst.* 31 (2018).

[6] David Alvarez-Melis and Tommi S. Jaakkola. 2018. On the robustness of interpretability methods. arXiv:1806.08049. Retrieved from https://arxiv.org/abs/1806.08049

[7] Leila Arras, Ahmed Osman, and Wojciech Samek. 2020. Ground truth evaluation of neural network explanations with clevr-xai. arXiv:2003.07258. Retrieved from https://arxiv.org/abs/2003.07258

[8] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* 58 (2020), 82–115.

[9] Worawate Ausawalaithong, Arjaree Thirach, Sanparith Marukatat, and Theerawit Wilaiprasitporn. 2018. Automatic lung cancer prediction from chest X-ray images using the deep learning approach. In *Proceedings of the 11th Biomedical Engineering International Conference (BMEiCON'18)*. IEEE, 1–5.

[10] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS One* 10, 7 (2015), e0130140.

[11] Jun Bai, Russell Posner, Tianyu Wang, Clifford Yang, and Sheida Nabavi. 2021. Applying deep learning in digital breast tomosynthesis for automatic breast cancer detection: A review. *Med. Image Anal.* 71 (2021), 102049.

[12] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. 65–72.

[13] Catarina Barata, M. Emre Celebi, and Jorge S. Marques. 2021. Explainable skin lesion diagnosis using taxonomies. *Pattern Recogn.* 110 (2021), 107413.

[14] Catarina Barata and Carlos Santiago. 2021. Improving the explainability of skin cancer diagnosis using CBIR. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 550–559.

[15] Alina Jade Barnett, Fides Regina Schwartz, Chaofan Tao, Chaofan Chen, Yinhao Ren, Joseph Y. Lo, and Cynthia Rudin. 2021. A case-based interpretable deep learning model for classification of mass lesions in digital mammography. *Nat. Mach. Intell.* 3, 12 (2021), 1061–1070.

[16] Cher Bass, Mariana da Silva, Carole Sudre, Petru-Daniel Tudosiu, Stephen Smith, and Emma Robinson. 2020. ICAM: Interpretable classification via disentangled representations and feature attribution mapping. *Adv. Neural Inf. Process. Syst.* 33 (2020), 7697–7709.

[17] Samyadeep Basu, Phil Pope, and Soheil Feizi. 2020. Influence functions in deep learning are fragile. In *International Conference on Learning Representations*.

[18] Umang Bhatt, Adrian Weller, and José M. F. Moura. 2021. Evaluating and aggregating feature-based model explanations. In *Proceedings of the 29th International Conference on International Joint Conferences on Artificial Intelligence*. 3016–3022.

[19] Jacob Bien and Robert Tibshirani. 2011. Prototype selection for interpretable classification.

[20] Nicholas Bien, Pranav Rajpurkar, Robyn L. Ball, Jeremy Irvin, Allison Park, Erik Jones, Michael Bereket, Bhavik N. Patel, Kristen W. Yeom, Katie Shpanskaya, et al. 2018. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of MRNet. *PLoS Med.* 15, 11 (2018), e1002699.

[21] Francesco Bodria, Fosca Giannotti, Riccardo Guidotti, Francesca Naretto, Dino Pedreschi, and Salvatore Rinzivillo. 2023. Benchmarking and survey of explanation methods for black box models. *Data Min. Knowl. Discov.* (2023), 1–60.

[22] Moritz Böhle, Fabian Eitel, Martin Weygandt, and Kerstin Ritter. 2019. Layer-wise relevance propagation for explaining deep neural network decisions in MRI-based Alzheimer's disease classification. *Front. Aging Neurosci.* (2019), 194.

[23] Luca Brunese, Francesco Mercaldo, Alfonso Reginelli, and Antonella Santone. 2020. Explainable deep learning for pulmonary disease and coronavirus COVID-19 detection from X-rays. *Comput. Methods Progr. Biomed.* 196 (2020), 105608.

[24] Gary H. Chang, David T. Felson, Shangran Qiu, Ali Guermazi, Terence D. Capellini, and Vijaya B. Kolachalama. 2020. Assessment of knee pain from MR imaging using a convolutional Siamese network. *Eur. Radiol.* 30, 6 (2020), 3538–3548.

[25] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N. Balasubramanian. 2018. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV'18)*. IEEE, 839–847.

[26] Saneem Chemmengath, Amar Prakash Azad, Ronny Luss, and Amit Dhurandhar. 2022. Let the CAT out of the bag: Contrastive attributed explanations for text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 7190–7206.

[27] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K. Su. 2019. This looks like that: Deep learning for interpretable image recognition. *Adv. Neural Inf. Process. Syst.* 32 (2019).

[28] Hugh Chen, Scott M. Lundberg, Gabriel Erion, Jerry H. Kim, and Su-In Lee. 2021. Forecasting adverse surgical events using self-supervised transfer learning for physiological signals. *NPJ Digit. Med.* 4, 1 (2021), 167.

[29] Jianbo Chen, Le Song, Martin Wainwright, and Michael Jordan. 2018. Learning to explain: An information-theoretic perspective on model interpretation. In *International Conference on Machine Learning*. PMLR, 883–892.

[30] Richard J. Chen, Ming Y. Lu, Drew F. K. Williamson, Tiffany Y. Chen, Jana Lipkova, Zahra Noor, Muhammad Shaban, Maha Shady, Mane Williams, Bumjin Joo, et al. 2022. Pan-cancer integrative histology-genomic analysis via multimodal deep learning. *Cancer Cell* 40, 8 (2022), 865–878.

[31] Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. 2020. Generating radiology reports via memory-driven transformer. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'20)*. 1439–1449.

[32] Tanya Chowdhury, Razieh Rahimi, and James Allan. 2022. Equi-explanation maps: Concise and informative global summary explanations. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*. 464–472.
[33] James R. Clough, Ilkay Oksuz, Esther Puyol-Antón, Bram Ruijsink, Andrew P. King, and Julia A. Schnabel. 2019. Global and local interpretability for cardiac MRI classification. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 656–664.
[34] Joseph Paul Cohen, Rupert Brooks, Sovann En, Evan Zucker, Anuj Pareek, Matthew P. Lungren, and Akshay Chaudhari. 2021. Gifsplanation via latent shift: A simple autoencoder approach to counterfactual generation for chest x-rays. In *Medical Imaging with Deep Learning*. PMLR, 74–104.
[35] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M. F. Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 326–335.
[36] Arun Das and Paul Rad. 2020. Opportunities and challenges in explainable artificial intelligence (xai): A survey. arXiv:2006.11371. Retrieved from https://arxiv.org/abs/2006.11371
[37] Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Paishun Ting, Karthikeyan Shanmugam, and Payel Das. 2018. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. *Adv. Neural Inf. Process. Syst.* 31 (2018).
[38] Theekshana Dissanayake, Tharindu Fernando, Simon Denman, Sridha Sridharan, Houman Ghaemmaghami, and Clinton Fookes. 2020. A robust interpretable deep learning classifier for heart anomaly detection without segmentation. *IEEE J. Biomed. Health Inf.* 25, 6 (2020), 2162–2171.
[39] Jon Donnelly, Alina Jade Barnett, and Chaofan Chen. 2022. Deformable protopnet: An interpretable image classifier using deformable prototypes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10265–10275.
[40] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *STAT* 1050 (2017), 2.
[41] Jared A. Dunnmon, Darvin Yi, Curtis P. Langlotz, Christopher Ré, Daniel L. Rubin, and Matthew P. Lungren. 2019. Assessment of convolutional neural networks for automated classification of chest radiographs. *Radiology* 290, 2 (2019), 537–544.
[42] Rudresh Dwivedi, Devam Dave, Het Naik, Smiti Singhal, Omer Rana, Pankesh Patel, Bin Qian, Zhenyu Wen, Tejal Shah, Graham Morgan, et al. 2022. Explainable AI (XAI): Core ideas, techniques and solutions. *ACM Comput. Surv.* (2022).
[43] Fabian Eitel, Kerstin Ritter, Alzheimer's Disease Neuroimaging Initiative (ADNI, et al. 2019. Testing the robustness of attribution methods for convolutional neural networks in MRI-based Alzheimer's disease classification. In *Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support*. Springer, 3–11.
[44] Fabian Eitel, Emily Soehler, Judith Bellmann-Strobl, Alexander U. Brandt, Klemens Ruprecht, René M Giess, Joseph Kuchling, Susanna Asseyer, Martin Weygandt, John-Dylan Haynes, et al. 2019. Uncovering convolutional neural network decisions for diagnosing multiple sclerosis on conventional MRI using layer-wise relevance propagation. *NeuroImage: Clin.* 24 (2019), 102003.
[45] Feng-Lei Fan, Jinjun Xiong, Mengzhou Li, and Ge Wang. 2021. On interpretability of artificial neural networks: A survey. *IEEE Trans. Radiat. Plasma Med. Sci.* 5, 6 (2021), 741–760.
[46] Zhengqing Fang, Kun Kuang, Yuxiao Lin, Fei Wu, and Yu-Feng Yao. 2020. Concept-based explanation for fine-grained images and its application in infectious keratitis classification. In *Proceedings of the 28th ACM International Conference on Multimedia*. 700–708.
[47] Tom Fawcett. 2006. An introduction to ROC analysis. *Pattern Recogn. Lett.* 27, 8 (2006), 861–874.
[48] Benjamin Filtjens, Pieter Ginis, Alice Nieuwboer, Muhammad Raheel Afzal, Joke Spildooren, Bart Vanrumste, and Peter Slaets. 2021. Modelling and identification of characteristic kinematic features preceding freezing of gait with convolutional neural networks and layer-wise relevance propagation. *BMC Med. Inf. Decis. Making* 21, 1 (2021), 1–11.
[49] William Gale, Luke Oakden-Rayner, Gustavo Carneiro, Lyle J. Palmer, and Andrew P. Bradley. 2019. Producing radiologist-quality reports for interpretable deep learning. In *Proceedings of the IEEE 16th International Symposium on Biomedical Imaging (ISBI'19)*. IEEE, 1275–1279.
[50] Paul Gamble, Ronnachai Jaroensri, Hongwu Wang, Fraser Tan, Melissa Moran, Trissia Brown, Isabelle Flament-Auvigne, Emad A. Rakha, Michael Toss, David J. Dabbs, et al. 2021. Determining breast cancer biomarker status and associated morphological features using deep learning. *Commun. Med.* 1, 1 (2021), 14.
[51] Kai Gao, Hui Shen, Yadong Liu, Lingli Zeng, and Dewen Hu. 2019. Dense-cam: Visualize the gender of brains with mri images. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN'19)*. IEEE, 1–7.
[52] Asma Ghandeharioun, Been Kim, Chun-Liang Li, Brendan Jou, Brian Eoff, and Rosalind Picard. 2021. DISSECT: Disentangled simultaneous explanations via concept traversals. In *International Conference on Learning Representations*.

[53] Amirata Ghorbani, Abubakar Abid, and James Zou. 2019. Interpretation of neural networks is fragile. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 3681–3688.
[54] Amirata Ghorbani, David Ouyang, Abubakar Abid, Bryan He, Jonathan H. Chen, Robert A. Harrington, David H. Liang, Euan A. Ashley, and James Y. Zou. 2020. Deep learning interpretation of echocardiograms. *NPJ Digit. Med.* 3, 1 (2020), 10.
[55] Amirata Ghorbani, James Wexler, James Y. Zou, and Been Kim. 2019. Towards automatic concept-based explanations. *Advances in Neural Information Processing Systems* 32 (2019).
[56] Yash Goyal, Amir Feder, Uri Shalit, and Been Kim. 2019. Explaining classifiers with causal concept effect (cace). arXiv:1907.07165. Retrieved from https://arxiv.org/abs/1907.07165
[57] Mara Graziani, Vincent Andrearczyk, and Henning Müller. 2018. Regression concept vectors for bidirectional explanations in histopathology. In *Understanding and Interpreting Machine Learning in Medical Image Computing Applications*. Springer, 124–132.
[58] Irina Grigorescu, Lucilio Cordero-Grande, A. David Edwards, Joseph V. Hajnal, Marc Modat, and Maria Deprez. 2019. Investigating image registration impact on preterm birth classification: An interpretable deep learning approach. In *International Workshop on Preterm, Perinatal and Paediatric Image Analysis*. Springer, 104–112.
[59] Riccardo Guidotti, Anna Monreale, Fosca Giannotti, Dino Pedreschi, Salvatore Ruggieri, and Franco Turini. 2019. Factual and counterfactual explanations for black box decision making. *IEEE Intell. Syst.* 34, 6 (2019), 14–23.
[60] Riccardo Guidotti, Anna Monreale, Stan Matwin, and Dino Pedreschi. 2020. Black box explanation by learning image exemplars in the latent feature space. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD'19), Part I*. Springer, 189–205.
[61] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Dino Pedreschi, Franco Turini, and Fosca Giannotti. 2018. Local rule-based explanations of black box decision systems. arXiv:1805.10820. Retrieved from https://arxiv.org/abs/1805.10820
[62] Riccardo Guidotti, Anna Monreale, Francesco Spinnato, Dino Pedreschi, and Fosca Giannotti. 2020. Explaining any time series classifier. In *Proceedings of the IEEE 2nd International Conference on Cognitive Machine Intelligence (CogMI'20)*. IEEE, 167–176.
[63] Paras Gulati, Qin Hu, and S Farokh Atashzar. 2021. Toward deep generalization of peripheral emg-based human-robot interfacing: A hybrid explainable solution for neurorobotic systems. *IEEE Robot. Autom. Lett.* 6, 2 (2021), 2650–2657.
[64] Manish Gupta, Chetna Das, Arnab Roy, Prashant Gupta, G. Radhakrishna Pillai, and Kamlakar Patole. 2020. Region of interest identification for cervical cancer images. In *Proceedings of the IEEE 17th International Symposium on Biomedical Imaging (ISBI'20)*. IEEE, 1293–1296.
[65] Karthik S. Gurumoorthy, Amit Dhurandhar, Guillermo Cecchi, and Charu Aggarwal. 2019. Efficient data representation by selecting prototypes with importance weights. In *Proceedings of the IEEE International Conference on Data Mining (ICDM'19)*. IEEE, 260–269.
[66] Wooseok Ha, Chandan Singh, Francois Lanusse, Srigokul Upadhyayula, and Bin Yu. 2021. Adaptive wavelet distillation from neural networks through interpretations. *Adv. Neural Inf. Process. Syst.* 34 (2021).
[67] Miriam Hägele, Philipp Seegerer, Sebastian Lapuschkin, Michael Bockmayr, Wojciech Samek, Frederick Klauschen, Klaus-Robert Müller, and Alexander Binder. 2020. Resolving challenges in deep learning-based analyses of histopathological images using explanation methods. *Sci. Rep.* 10, 1 (2020), 1–12.
[68] Narayan Hegde, Jason D. Hipp, Yun Liu, Michael Emmert-Buck, Emily Reif, Daniel Smilkov, Michael Terry, Carrie J. Cai, Mahul B. Amin, Craig H. Mermel, et al. 2019. Similar image search for histopathology: SMILY. *NPJ Digit. Med.* 2, 1 (2019), 56.
[69] Andreas Holzinger, André Carrington, and Heimo Müller. 2020. Measuring the quality of explanations: The system causability scale (SCS). *Künstl. Intell.* 34, 2 (2020), 193–198.
[70] Andreas T. Holzinger and Heimo Muller. 2021. Toward human–AI interfaces to support explainability and causability in medical AI. *Computer* 54, 10 (2021), 78–86.
[71] Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. 2019. A benchmark for interpretability methods in deep neural networks. *Adv. Neural Inf. Process. Syst.* 32 (2019).
[72] Benjamin Hoover Hendrik Strobelt, and Sebastian Gehrmann. 2020. exBERT: A visual analysis tool to explore learned representations in transformer models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.
[73] Md Imran Hossain, Ghada Zamzmi, Peter Mouton, Yu Sun, and Dmitry Goldgof. 2023. Enhancing neonatal pain assessment transparency via explanatory training examples identification. In *Proceedings of the IEEE 36th International Symposium on Computer-Based Medical Systems (CBMS'23)*. IEEE, 311–316.
[74] Brian Hu, Bhavan Vasu, and Anthony Hoogs. 2022. X-mir: Explainable medical image retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 440–450.

[75] Shaoli Huang, Xinchao Wang, and Dacheng Tao. 2021. Snapmix: Semantically proportional mixing for augmenting fine-grained data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 1628–1636.
[76] Yongxiang Huang and Albert Chung. 2019. Evidence localization for pathology images using weakly supervised learning. In *Proceedings of the International Conference on Medical Image Computing and Computer-assisted Intervention*. Springer, 613–621.
[77] Zhicheng Huang and Dongmei Fu. 2019. Diagnose chest pathology in X-ray images by learning multi-attention convolutional neural network. In *Proceedings of the IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC'19)*. IEEE, 294–299.
[78] Daniel T. Huff, Amy J. Weisman, and Robert Jeraj. 2021. Interpretation and visualization techniques for deep learning models in medical imaging. *Phys. Med. Bio.* 66, 4 (2021), 04TR01.
[79] Rami Ibrahim and M. Omair Shafiq. 2023. Explainable convolutional neural networks: A taxonomy, review, and future directions. *ACM Computing Surveys* 55, 10 (2023), 1–37.
[80] Hayato Itoh, Zhongyang Lu, Yuichi Mori, Masashi Misawa, Masahiro Oda, Shin-ei Kudo, and Kensaku Mori. 2020. Visualising decision-reasoning regions in computer-aided pathological pattern diagnosis of endoscytoscopic images based on CNN weights analysis. In *Medical Imaging 2020: Computer-Aided Diagnosis*, Vol. 11314. SPIE, 761–768.
[81] Praharsh Ivaturi, Matteo Gadaleta, Amitabh C. Pandey, Michael Pazzani, Steven R. Steinhubl, and Giorgio Quer. 2021. A comprehensive explanation framework for biomedical time series classification. *IEEE J. Biomed. Health Inf.* 25, 7 (2021), 2398–2408.
[82] Amir Jamaludin, Timor Kadir, and Andrew Zisserman. 2017. SpineNet: Automated classification and evidence visualization in spinal MRIs. *Med. Image Anal.* 41 (2017), 63–73.
[83] Adrianna Janik, Jonathan Dodd, Georgiana Ifrim, Kris Sankaran, and Kathleen Curran. 2021. Interpretability of a deep learning model in the application of cardiac MRI segmentation with an ACDC challenge dataset. In *Medical Imaging 2021: Image Processing*, Vol. 11596. International Society for Optics and Photonics, 1159636.
[84] Junyi Ji. 2019. Gradient-based interpretation on convolutional neural network for classification of pathological images. In *Proceedings of the International Conference on Information Technology and Computer Application (ITCA'19)*. IEEE, 83–86.
[85] Hongyang Jiang, Kang Yang, Mengdi Gao, Dongdong Zhang, He Ma, and Wei Qian. 2019. An interpretable ensemble deep learning model for diabetic retinopathy disease classification. In *Proceedings of the 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC'19)*. IEEE, 2045–2048.
[86] Baoyu Jing, Pengtao Xie, and Eric Xing. 2018. On the automatic generation of medical imaging reports. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2577–2586.
[87] Gargi Joshi, Rahee Walambe, and Ketan Kotecha. 2021. A review on explainability in multimodal deep neural nets. *IEEE Access* 9 (2021), 59800–59821.
[88] Shalmali Joshi, Oluwasanmi Koyejo, Warut Vijitbenjaronk, Been Kim, and Joydeep Ghosh. 2019. Towards realistic individual recourse and actionable explanations in black-box decision making systems. *J. Environ. Sci. (Chin.)* (2019).
[89] Md Sarwar Kamal, Aden Northcote, Linkon Chowdhury, Nilanjan Dey, Rubén González Crespo, and Enrique Herrera-Viedma. 2021. Alzheimer's patient analysis using image and gene expression data and explainable-AI to present associated genes. *IEEE Trans. Instrum. Meas.* 70 (2021), 1–7.
[90] Andrei Kapishnikov, Tolga Bolukbasi, Fernanda Viégas, and Michael Terry. 2019. Xrai: Better attributions through regions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4948–4957.
[91] Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4401–4410.
[92] Satyananda Kashyap, Alexandros Karargyris, Joy Wu, Yaniv Gur, Arjun Sharma, Ken C. L. Wong, Mehdi Moradi, and Tanveer Syeda-Mahmood. 2020. Looking in the right place for anomalies: Explainable Ai through automatic location learning. In *Proceedings of the IEEE 17th International Symposium on Biomedical Imaging (ISBI'20)*. IEEE, 1125–1129.
[93] Justin Ker, Yeqi Bai, Hwei Yee Lee, Jai Rao, and Lipo Wang. 2019. Automated brain histology classification using machine learning. *J. Clin. Neurosci.* 66 (2019), 239–245.
[94] Ashkan Khakzar, Shadi Albarqouni, and Nassir Navab. 2019. Learning interpretable features via adversarially robust optimization. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 793–800.
[95] Amirhossein Kiani, Bora Uyumazturk, Pranav Rajpurkar, Alex Wang, Rebecca Gao, Erik Jones, Yifan Yu, Curtis P. Langlotz, Robyn L. Ball, Thomas J. Montine, et al. 2020. Impact of a deep learning assistant on the histopathologic classification of liver cancer. *NPJ Digit. Med.* 3, 1 (2020), 1–8.
[96] Been Kim, Rajiv Khanna, and Oluwasanmi O. Koyejo. 2016. Examples are not enough, learn to criticize! criticism for interpretability. *Adv. Neural Inf. Process. Syst.* 29 (2016).
[97] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. 2018. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International Conference on Machine Learning*. PMLR, 2668–2677.

[98] Eunji Kim, Siwon Kim, Minji Seo, and Sungroh Yoon. 2021. XProtoNet: Diagnosis in chest radiography with global and local explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15719–15728.

[99] Junho Kim, Minsu Kim, and Yong Man Ro. 2021. Interpretation of lesional detection via counterfactual generation. In *Proceedings of the IEEE International Conference on Image Processing (ICIP'21)*. IEEE, 96–100.

[100] Mijung Kim, Jong Chul Han, Seung Hyup Hyun, Olivier Janssens, Sofie Van Hoecke, Changwon Kee, and Wesley De Neve. 2019. Medinoid: Computer-aided diagnosis and localization of glaucoma using deep learning. *Appl. Sci.* 9, 15 (2019), 3064.

[101] Kranthi Kiran G. V. and G. Meghana Reddy. 2019. Automatic classification of whole slide pap smear images using CNN with PCA based feature interpretation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 0–0.

[102] Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. In *International Conference on Machine Learning*. PMLR, 1885–1894.

[103] Maximilian Kohlbrenner, Alexander Bauer, Shinichi Nakajima, Alexander Binder, Wojciech Samek, and Sebastian Lapuschkin. 2020. Towards best practice in explaining neural network decisions with LRP. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN'20)*. IEEE, 1–7.

[104] Bruno Korbar, Andrea M. Olofson, Allen P. Miraflor, Catherine M. Nicka, Matthew A. Suriawinata, Lorenzo Torresani, Arief A. Suriawinata, and Saeed Hassanpour. 2017. Looking under the hood: Deep neural network visualization to interpret whole-slide image analysis outcomes for colorectal polyps. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 69–75.

[105] Olga Kovaleva, Chaitanya Shivade, Satyananda Kashyap, Karina Kanjaria, Joy Wu, Deddeh Ballah, Adam Coy, Alexandros Karargyris, Yufan Guo, David Beymer Beymer, et al. 2020. Towards visual dialog for radiology. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*. 60–69.

[106] Devinder Kumar, Graham W. Taylor, and Alexander Wong. 2019. Discovery radiomics with CLEAR-DR: Interpretable computer aided diagnosis of diabetic retinopathy. *IEEE Access* 7 (2019), 25891–25896.

[107] Bum Chul Kwon, Min-Je Choi, Joanne Taery Kim, Edward Choi, Young Bin Kim, Soonwook Kwon, Jimeng Sun, and Jaegul Choo. 2018. Retainvis: Visual analytics with interpretable and interactive recurrent neural networks on electronic medical records. *IEEE Trans. Vis. Comput. Graph.* 25, 1 (2018), 299–309.

[108] Orestis Lampridis, Riccardo Guidotti, and Salvatore Ruggieri. 2020. Explaining sentiment classification with synthetic exemplars and counter-exemplars. In *Proceedings of the 23rd International Conference on Discovery Science (DS'20)*. Springer, 357–373.

[109] Hyebin Lee, Seong Tae Kim, and Yong Man Ro. 2019. Generation of multimodal justification using visual word constraint model for explainable computer-aided diagnosis. In *Proceedings of the 2nd International Workshop on Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support (iMIMIC 2019) and 9th International Workshop (ML-CDS 2019), Held in Conjunction with MICCAI 2019*. Springer, 21–29.

[110] Hyunkwang Lee, Sehyo Yune, Mohammad Mansouri, Myeongchan Kim, Shahein H. Tajmir, Claude E. Guerrier, Sarah A. Ebert, Stuart R. Pomerantz, Javier M. Romero, Shahmir Kamalian, et al. 2019. An explainable deep-learning algorithm for the detection of acute intracranial haemorrhage from small datasets. *Nat. Biomed. Eng.* 3, 3 (2019), 173–182.

[111] Jeong Hoon Lee, Eun Ju Ha, DaYoung Kim, Yong Jun Jung, Subin Heo, Yong-Ho Jang, Sung Hyun An, and Kyungmin Lee. 2020. Application of deep learning to the diagnosis of cervical lymph node metastasis from thyroid cancer with CT: External validation and clinical utility for resident training. *Eur. Radiol.* 30 (2020), 3066–3072.

[112] Yiming Lei, Yukun Tian, Hongming Shan, Junping Zhang, Ge Wang, and Mannudeep K. Kalra. 2020. Shape and margin-aware lung nodule classification in low-dose CT images via soft activation mapping. *Med. Image Anal.* 60 (2020), 101628.

[113] H. A. Leopold, A. Singh, S. Sengupta, J. S. Zelek, and V. Lakshminarayanan. 2020. Recent advances in deep learning applications for retinal diagnosis using OCT. *Tate of the Art in Neural Networks* (2020).

[114] Benjamin Letham, Cynthia Rudin, Tyler H. McCormick, and David Madigan. 2015. Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. *Ann. Appl. Stat.* 9, 3 (2015), 1350–1371.

[115] Dan Ley, Saumitra Mishra, and Daniele Magazzeni. 2022. Global counterfactual explanations: Investigations, implementations and improvements. In *ICLR 2022 Workshop on PAIR@2Struct: Privacy, Accountability, Interpretability, Robustness, Reasoning on Structured Data*.

[116] Jiwei Li, Will Monroe, and Dan Jurafsky. 2016. Understanding neural networks through representation erasure. arXiv:1612.08220. Retrieved from https://arxiv.org/abs/1612.08220

[117] Mengze Li, Kun Kuang, Qiang Zhu, Xiaohong Chen, Qing Guo, and Fei Wu. 2020. IB-M: A flexible framework to align an interpretable model and a black-box model. In *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM'20)*. IEEE, 643–649.

[118] Weipeng Li, Jiaxin Zhuang, Ruixuan Wang, Jianguo Zhang, and Wei-Shi Zheng. 2020. Fusing metadata and dermoscopy images for skin disease diagnosis. In *Proceedings of the IEEE 17th International Symposium on Biomedical Imaging (ISBI'20)*. IEEE, 1996–2000.

[119] WangMin Liao, BeiJi Zou, RongChang Zhao, YuanQiong Chen, ZhiYou He, and MengJie Zhou. 2019. Clinical interpretable deep learning model for glaucoma diagnosis. *IEEE J. Biomed. Health Inf.* 24, 5 (2019), 1405–1412.

[120] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*. 74–81.

[121] Tsung-Chieh Lin and Hsi-Chieh Lee. 2020. Covid-19 chest radiography images analysis based on integration of image preprocess, guided grad-CAM, machine learning and risk management. In *Proceedings of the 4th International Conference on Medical and Health Informatics*. 281–288.

[122] Zehui Lin, Shengli Li, Dong Ni, Yimei Liao, Huaxuan Wen, Jie Du, Siping Chen, Tianfu Wang, and Baiying Lei. 2019. Multi-task learning for quality assessment of fetal head ultrasound images. *Med. Image Anal.* 58 (2019), 101548.

[123] Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. 2020. Explainable ai: A review of machine learning interpretability methods. *Entropy* 23, 1 (2020), 18.

[124] Chi Liu, Xiaotong Han, Zhixi Li, Jason Ha, Guankai Peng, Wei Meng, and Mingguang He. 2019. A self-adaptive deep learning method for automated eye laterality detection based on color fundus photography. *PLoS One* 14, 9 (2019), e0222025.

[125] Guanxiong Liu, Tzu-Ming Harry Hsu, Matthew McDermott, Willie Boag, Wei-Hung Weng, Peter Szolovits, and Marzyeh Ghassemi. 2019. Clinically accurate chest x-ray report generation. In *Machine Learning for Healthcare Conference*. PMLR, 249–269.

[126] S. Liu, B. Kailkhura, D. Loveland, and H. Yong. 2019. *Generative Counterfactual Introspection for Explainable Deep Learning*. Technical Report. Lawrence Livermore National Laboratory, Livermore, CA.

[127] Hui Wen Loh, Chui Ping Ooi, Silvia Seoni, Prabal Datta Barua, Filippo Molinari, and U Rajendra Acharya. 2022. Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022). *Comput. Methods Progr. Biomed.* (2022), 107161.

[128] Arnaud Van Looveren and Janis Klaise. 2021. Interpretable counterfactual explanations guided by prototypes. In *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 650–665.

[129] Alina Lopatina, Stefan Ropele, Renat Sibgatulin, Jürgen R. Reichenbach, and Daniel Güllmar. 2020. Investigation of deep-learning-driven identification of multiple sclerosis patients based on susceptibility-weighted images using relevance analysis. *Front. Neurosci.* 14 (2020), 609468.

[130] Adriano Lucieri, Muhammad Naseer Bajwa, Stephan Alexander Braun, Muhammad Imran Malik, Andreas Dengel, and Sheraz Ahmed. 2020. On interpretability of deep learning based skin lesion classifiers using concept activation vectors. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN'20)*. IEEE, 1–10.

[131] Adriano Lucieri, Muhammad Naseer Bajwa, Stephan Alexander Braun, Muhammad Imran Malik, Andreas Dengel, and Sheraz Ahmed. 2022. ExAID: A multimodal explanation framework for computer-aided diagnosis of skin lesions. *Comput. Methods Progr. Biomed.* 215 (2022), 106620.

[132] Adriano Lucieri, Muhammad Naseer Bajwa, Andreas Dengel, and Sheraz Ahmed. 2020. Explaining ai-based decision support systems using concept localization maps. In *Proceedings of the 27th International Conference on Neural Information Processing (ICONIP'20), Part IV 27*. Springer, 185–193.

[133] Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* 30 (2017).

[134] Luyang Luo, Hao Chen, Xi Wang, Qi Dou, Huangjing Lin, Juan Zhou, Gongjie Li, and Pheng-Ann Heng. 2019. Deep angular embedding and feature correlation attention for breast MRI cancer analysis. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 504–512.

[135] Yiwei Lyu, Paul Pu Liang, Zihao Deng, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2022. Dime: Fine-grained interpretations of multimodal models via disentangled local explanations. arXiv:2203.02013. Retrieved from https://arxiv.org/abs/2203.02013

[136] Kai Ma, Kaijie Wu, Hao Cheng, Chaochen Gu, Rui Xu, and Xinping Guan. 2018. A pathology image diagnosis network with visual interpretability and structured diagnostic report. In *Proceedings of the 25th International Conference on Neural Information Processing (ICONIP'18), Proceedings, Part VI 25*. Springer, 282–293.

[137] Pavan Rajkumar Magesh, Richard Delwin Myloth, and Rijo Jackson Tom. 2020. An explainable machine learning model for early detection of Parkinson's disease using LIME on DaTSCAN imagery. *Comput. Biol. Med.* 126 (2020), 104041.

[138] David Major, Dimitrios Lenis, Maria Wimmer, Gert Sluiter, Astrid Berg, and Katja Bühler. 2020. Interpreting medical image classifiers by optimization based counterfactual impact analysis. In *Proceedings of the IEEE 17th International Symposium on Biomedical Imaging (ISBI'20)*. IEEE, 1096–1100.

[139] Qier Meng, Yohei Hashimoto, and Shin'ichi Satoh. 2020. How to extract more information with less burden: Fundus image classification and retinal disease localization with ophthalmologist intervention. *IEEE J. Biomed. Health Inf.* 24, 12 (2020), 3351–3361.

[140] Silvan Mertes, Tobias Huber, Katharina Weitz, Alexander Heimerl, and Elisabeth André. 2022. Ganterfactual–counterfactual explanations for medical non-experts using generative adversarial learning. *Front. Artif. Intell.* 5 (2022), 825565.

[141] Carlo Metta, Riccardo Guidotti, Yuan Yin, Patrick Gallinari, and Salvatore Rinzivillo. 2021. Exemplars and counterexemplars explanations for image classifiers, targeting skin lesion labeling. In *Proceedings of the IEEE Symposium on Computers and Communications (ISCC'21)*. IEEE, 1–7.

[142] Richard Meyes, Constantin Waubert de Puiseau, Andres Posada-Moreno, and Tobias Meisen. 2020. Under the hood of neural networks: Characterizing learned representations by functional neuron populations and network ablations. arXiv:2004.01254. Retrieved from https://arxiv.org/abs/2004.01254

[143] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.* 267 (2019), 1–38.

[144] Yao Ming, Huamin Qu, and Enrico Bertini. 2018. Rulematrix: Visualizing and understanding classifiers with rules. *IEEE Trans. Vis. Comput. Graph.* 25, 1 (2018), 342–352.

[145] David Moher, Alessandro Liberati, Jennifer Tetzlaff, Douglas G. Altman, and PRISMA Group*. 2009. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *Ann. Intern. Med.* 151, 4 (2009), 264–269.

[146] Ioannis Mollas, Nikolaos Bassiliades, and Grigorios Tsoumakas. 2020. Lionets: Local interpretation of neural networks through penultimate layer decoding. In *Proceedings of the International Workshops of Machine Learning and Knowledge Discovery in Databases (ECML PKDD'19), Part I*. Springer, 265–276.

[147] Ramaravind K. Mothilal, Amit Sharma, and Chenhao Tan. 2020. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 607–617.

[148] Sajad Mousavi, Fatemeh Afghah, and U. Rajendra Acharya. 2020. HAN-ECG: An interpretable atrial fibrillation detection model using hierarchical attention networks. *Comput. Biol. Med.* 127 (2020), 104057.

[149] Satya M. Muddamsetty, N. S. Jahromi Mohammad, and Thomas B. Moeslund. 2020. Sidu: Similarity difference and uniqueness method for explainable ai. In *Proceedings of the IEEE International Conference on Image Processing (ICIP'20)*. IEEE, 3269–3273.

[150] Thanakron Na Pattalung, Thammasin Ingviya, and Sitthichok Chaichulee. 2021. Feature explanations in recurrent neural networks for predicting risk of mortality in intensive care patients. *J. Personal. Med.* 11, 9 (2021), 934.

[151] Arunachalam Narayanaswamy, Subhashini Venugopalan, Dale R. Webster, Lily Peng, Greg S. Corrado, Paisan Ruamviboonsuk, Pinal Bavishi, Michael Brenner, Philip C. Nelson, and Avinash V. Varadarajan. 2020. Scientific discovery by generating counterfactuals using image translation. In *Proceedings of the 23rd International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI'20), Part I 23*. Springer, 273–283.

[152] Parth Natekar, Avinash Kori, and Ganapathy Krishnamurthi. 2020. Demystifying brain tumor segmentation networks: Interpretability and uncertainty analysis. *Front. Comput. Neurosci.* 14 (2020), 6.

[153] Meike Nauta, Annemarie Jutte, Jesper Provoost, and Christin Seifert. 2021. This looks like that, because... explaining prototypes for interpretable image recognition. In *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 441–456.

[154] Ankit Pal and Malaikannan Sankarasubbu. 2021. Pay attention to the cough: Early diagnosis of COVID-19 using interpretable symptoms embeddings with cough sound signal processing. In *Proceedings of the 36th Annual ACM Symposium on Applied Computing*. 620–628.

[155] Cecilia Panigutti, Alan Perotti, and Dino Pedreschi. 2020. Doctor XAI: An ontology-based approach to black-box sequential data classification explanations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 629–639.

[156] Zachary Papanastasopoulos, Ravi K. Samala, Heang-Ping Chan, Lubomir Hadjiiski, Chintana Paramagul, Mark A. Helvie, and Colleen H. Neal. 2020. Explainable AI for medical imaging: deep-learning CNN ensemble for classification of estrogen receptor status from breast MRI. In *Medical Imaging 2020: Computer-aided Diagnosis*, Vol. 11314. International Society for Optics and Photonics, 113140Z.

[157] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. 311–318.

[158] Arijit Patra and J. Alison Noble. 2019. Incremental learning of fetal heart anatomies using interpretable saliency maps. In *Proceedings of the Annual Conference on Medical Image Understanding and Analysis*. Springer, 129–141.

[159] Rahul Paul, Matthew Schabath, Yoganand Balagurunathan, Ying Liu, Qian Li, Robert Gillies, Lawrence O. Hall, and Dmitry B. Goldgof. 2019. Explaining deep features using radiologist-defined semantic features and traditional quantitative features. *Tomography* 5, 1 (2019), 192–200.

[160] Rahul Paul, Matthew Schabath, Robert Gillies, Lawrence Hall, and Dmitry Goldgof. 2020. Convolutional neural network ensembles for accurate lung nodule malignancy prediction 2 years in the future. *Comput. Biol. Med.* 122 (2020), 103882.

[161] Martin Pawelczyk, Klaus Broelemann, and Gjergji Kasneci. 2020. Learning model-agnostic counterfactual explanations for tabular data. In *Proceedings of the Web Conference*. 3126–3132.

[162] Sérgio Pereira, Raphael Meier, Victor Alves, Mauricio Reyes, and Carlos A. Silva. 2018. Automatic brain tumor grading from MRI data using convolutional neural networks and quality assessment. In *Understanding and Interpreting Machine Learning in Medical Image Computing Applications*. Springer, 106–114.

[163] Vitali Petsiuk, Abir Das, and Kate Saenko. 2018. Rise: Randomized input sampling for explanation of black-box models. arXiv:1806.07421. Retrieved from https://arxiv.org/abs/1806.07421

[164] Kenneth A. Philbrick, Kotaro Yoshida, Dai Inoue, Zeynettin Akkus, Timothy L. Kline, Alexander D. Weston, Panagiotis Korfiatis, Naoki Takahashi, and Bradley J. Erickson. 2018. What does deep learning see? Insights from a classifier trained to predict contrast enhancement phase from CT images. *Am. J. Roentgenol.* 211, 6 (2018), 1184–1193.

[165] Gregory Plumb, Denali Molitor, and Ameet S. Talwalkar. 2018. Model agnostic supervised local explanations. *Adv. Neural Inf. Process. Syst.* 31 (2018).

[166] Rafael Poyiadzi, Kacper Sokol, Raul Santos-Rodriguez, Tijl De Bie, and Peter Flach. 2020. FACE: Feasible and actionable counterfactual explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 344–350.

[167] Xiaofeng Qi, Lei Zhang, Yao Chen, Yong Pi, Yi Chen, Qing Lv, and Zhang Yi. 2019. Automated diagnosis of breast ultrasonography images using deep neural networks. *Med. Image Anal.* 52 (2019), 185–198.

[168] Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain Yourself! Leveraging language models for commonsense reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 4932–4942.

[169] Pranav Rajpurkar, Jeremy Irvin, Robyn L. Ball, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis P. Langlotz, et al. 2018. Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLoS Med.* 15, 11 (2018), e1002686.

[170] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "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*. 1135–1144.

[171] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[172] Laura Rieger, Chandan Singh, William Murdoch, and Bin Yu. 2020. Interpretations are useful: Penalizing explanations to align neural networks with prior knowledge. In *International Conference on Machine Learning*. PMLR, 8116–8126.

[173] Isabel Rio-Torto, Kelwin Fernandes, and Luís F. Teixeira. 2020. Understanding the decisions of CNNs: An in-model approach. *Pattern Recogn. Lett.* 133 (2020), 373–380.

[174] Ivan Rodin, Irina Fedulova, Artem Shelmanov, and Dmitry V. Dylov. 2019. Multitask and multimodal neural network model for interpretable analysis of x-ray images. In *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM'19)*. IEEE, 1601–1604.

[175] Yao Rong, Tobias Leemann, Vadim Borisov, Gjergji Kasneci, and Enkelejda Kasneci. 2022. A consistent and efficient evaluation strategy for attribution methods. In *Proceedings of the International Conference on Machine Learning*. PMLR, 18770–18795.

[176] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* 1, 5 (2019), 206–215.

[177] Zohaib Salahuddin, Henry C. Woodruff, Avishek Chatterjee, and Philippe Lambin. 2022. Transparency of deep neural networks for medical image analysis: A review of interpretability methods. *Comput. Biol. Med.* 140 (2022), 105111.

[178] Md Sirajus Salekin, Ghada Zamzmi, Dmitry Goldgof, Rangachar Kasturi, Thao Ho, and Yu Sun. 2021. Multimodal spatio-temporal deep learning approach for neonatal postoperative pain assessment. *Comput. Biol. Med.* 129 (2021), 104150.

[179] Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. 2016. Evaluating the visualization of what a deep neural network has learned. *IEEE Trans. Neural Netw. Learn. Syst.* 28, 11 (2016), 2660–2673.

[180] Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. 2017. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. t arXiv:1708.08296. Retrieved from https://arxiv.org/abs/1708.08296

[181] Daniel Sauter, Georg Lodde, Felix Nensa, Dirk Schadendorf, Elisabeth Livingstone, and Markus Kukuk. 2022. Validating automatic concept-based explanations for AI-Based digital histopathology. *Sensors* 22, 14 (2022), 5346.

[182] Rory Sayres, Ankur Taly, Ehsan Rahimy, Katy Blumer, David Coz, Naama Hammel, Jonathan Krause, Arunachalam Narayanaswamy, Zahra Rastegar, Derek Wu, et al. 2019. Using a deep learning algorithm and integrated gradients explanation to assist grading for diabetic retinopathy. *Ophthalmology* 126, 4 (2019), 552–564.

[183] Kathryn Schutte, Olivier Moindrot, Paul Hérent, Jean-Baptiste Schiratti, and Simon Jégou. 2021. Using stylegan for visual interpretability of deep learning models on medical images. arXiv:2101.07563. Retrieved from https://arxiv.org/abs/2101.07563

[184] Jarrel C. Y. Seah, Jennifer S. N. Tang, Andy Kitchen, Frank Gaillard, and Andrew F. Dixon. 2019. Chest radiographs in congestive heart failure: Visualizing neural network learning. *Radiology* 290, 2 (2019), 514–522.

[185] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*. 618–626.

[186] Mattia Setzu, Riccardo Guidotti, Anna Monreale, Franco Turini, Dino Pedreschi, and Fosca Giannotti. 2021. Glocalx-from local to global explanations of black box ai models. *Artif. Intell.* 294 (2021), 103457.

[187] L. S. Shapley. 1953. A value for n-person games. In *Contributions to the Theory of Games (AM 28)*, Volume II, 307–317.

[188] Supreeth P. Shashikumar, Christopher S. Josef, Ashish Sharma, and Shamim Nemati. 2021. DeepAISE–an interpretable and recurrent neural survival model for early prediction of sepsis. *Artif. Intell. Med.* 113 (2021), 102036.

[189] Sumeet Shinde, Tanay Chougule, Jitender Saini, and Madhura Ingalhalikar. 2019. HR-CAM: Precise localization of pathology using multi-level learning in CNNs. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 298–306.

[190] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning important features through propagating activation differences. In *Proceedings of the International Conference on Machine Learning*. PMLR, 3145–3153.

[191] Wilson Silva, Alexander Poellinger, Jaime S. Cardoso, and Mauricio Reyes. 2020. Interpretability-guided content-based medical image retrieval. In *Proceedings of the 23rd International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI'20), Part I 23*. Springer, 305–314.

[192] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *Proceedings of the Workshop at International Conference on Learning Representations*. Citeseer.

[193] Gurmail Singh and Kin-Choong Yow. 2021. An interpretable deep learning model for COVID-19 detection with chest X-ray images. *IEEE Access* 9 (2021), 85198–85208.

[194] Sonit Singh, Sarvnaz Karimi, Kevin Ho-Shon, and Len Hamey. 2019. From chest x-rays to radiology reports: A multimodal machine learning approach. In *Proceedings of the Digital Image Computing: Techniques and Applications (DICTA'19)*. IEEE, 1–8.

[195] Sumedha Singla, Motahhare Eslami, Brian Pollack, Stephen Wallace, and Kayhan Batmanghelich. 2023. Explaining the black-box smoothly–a counterfactual approach. *Med. Image Anal.* 84 (2023), 102721.

[196] Leon Sixt, Maximilian Granz, and Tim Landgraf. 2020. When explanations lie: Why many modified bp attributions fail. In *Proceedings of the International Conference on Machine Learning*. PMLR, 9046–9057.

[197] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. 2017. Smoothgrad: Removing noise by adding noise. arXiv:1706.03825. Retrieved from https://arxiv.org/abs/1706.03825

[198] J. Springenberg, Alexey Dosovitskiy, Thomas Brox, and M. Riedmiller. 2015. Striving for simplicity: The all convolutional net. In *Proceedings of the ICLR (Workshop Track)*.

[199] Chenxi Sun, Hongna Dui, and Hongyan Li. 2021. Interpretable time-aware and co-occurrence-aware network for medical prediction. *BMC Med. Inf. Decis. Making* 21, 1 (2021), 1–12.

[200] Hao Sun, Xianxu Zeng, Tao Xu, Gang Peng, and Yutao Ma. 2019. Computer-aided diagnosis in histopathological images of the endometrium using a convolutional neural network and attention mechanisms. *IEEE J. Biomed. Health Inf.* 24, 6 (2019), 1664–1676.

[201] Li Sun, Weipeng Wang, Jiyun Li, and Jingsheng Lin. 2019. Study on medical image report generation based on improved encoding-decoding method. In *Proceedings of the 15th International Conference on Intelligent Computing Theories and Application (ICIC'19), Part I 15*. Springer, 686–696.

[202] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the International Conference on Machine Learning*. PMLR, 3319–3328.

[203] Amirhessam Tahmassebi, Jennifer Martin, Anke Meyer-Baese, and Amir H. Gandomi. 2020. An interpretable deep learning framework for health monitoring systems: A case study of eye state detection using eeg signals. In *Proceedings of the IEEE Symposium Series on Computational Intelligence (SSCI'20)*. IEEE, 211–218.

[204] Sarah Tan, Matvey Soloviev, Giles Hooker, and Martin T. Wells. 2020. Tree space prototypes: Another look at making tree ensembles interpretable. In *Proceedings of the ACM-IMS on Foundations of Data Science Conference*. 23–34.

[205] Claire Tang. 2020. Discovering unknown diseases with explainable automated medical imaging. In *Proceedings of the 24th Annual Conference on Medical Image Understanding and Analysis (MIUA'20)*. Springer, 346–358.

[206] Ziqi Tang, Kangway V. Chuang, Charles DeCarli, Lee-Way Jin, Laurel Beckett, Michael J. Keiser, and Brittany N. Dugger. 2019. Interpretable classification of Alzheimer's disease pathologies with a convolutional neural network pipeline. *Nat. Commun.* 10, 1 (2019), 1–14.

[207] Kaveri A. Thakoor, Sharath C. Koorathota, Donald C. Hood, and Paul Sajda. 2020. Robust and interpretable convolutional neural networks to detect glaucoma in optical coherence tomography images. *IEEE Trans. Biomed. Eng.* 68, 8 (2020), 2456–2466.

[208] Kaveri A. Thakoor, Xinhui Li, Emmanouil Tsamis, Paul Sajda, and Donald C. Hood. 2019. Enhancing the accuracy of glaucoma detection from OCT probability maps using convolutional neural networks. In *Proceedings of the 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC'19)*. IEEE, 2036–2040.

[209] Jonas Theiner, Eric Müller-Budack, and Ralph Ewerth. 2022. Interpretable semantic photo geolocation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 750–760.

[210] Jayaraman J. Thiagarajan, Kowshik Thopalli, Deepta Rajan, and Pavan Turaga. 2022. Training calibration-based counterfactual explainers for deep learning models in medical image analysis. *Sci. Rep.* 12, 1 (2022), 1–15.

[211] Erico Tjoa and Cuntai Guan. 2020. A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE Trans. Neural Netw. Learn. Syst.* 32, 11 (2020), 4793–4813.

[212] Kazuki Uehara, Masahiro Murakawa, Hirokazu Nosato, and Hidenori Sakanashi. 2019. Prototype-based interpretation of pathological image analysis by convolutional neural networks. In *Proceedings of the Asian Conference on Pattern Recognition*. Springer, 640–652.

[213] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *J. Mach. Learn. Res.* 9, 11 (2008).

[214] Bas H. M. van der Velden, Markus H. A. Janse, Max A. A. Ragusi, Claudette E. Loo, and Kenneth G. A. Gilhuijs. 2020. Volumetric breast density estimation on MRI using explainable deep learning regression. *Sci. Rep.* 10, 1 (2020), 1–9.

[215] Bas H. M. van der Velden, Hugo J. Kuijf, Kenneth G. A. Gilhuijs, and Max A. Viergever. 2022. Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. *Med. Image Anal.* (2022), 102470.

[216] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Adv. Neural Inf. Process. Syst.* 30 (2017).

[217] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4566–4575.

[218] Sahil Verma, John Dickerson, and Keegan Hines. 2020. Counterfactual explanations for machine learning: A review. arXiv:2010.10596. Retrieved from https://arxiv.org/abs/2010.10596

[219] Giulia Vilone and Luca Longo. 2020. Explainable artificial intelligence: A systematic review. arXiv:2006.00093. Retrieved from https://arxiv.org/abs/2006.00093

[220] Lituan Wang, Lei Zhang, Minjuan Zhu, Xiaofeng Qi, and Zhang Yi. 2020. Automatic diagnosis for thyroid nodules in ultrasound images by deep neural networks. *Med. Image Anal.* 61 (2020), 101665.

[221] Sen Wang, Yuxiang Xing, Li Zhang, Hewei Gao, and Hao Zhang. 2019. Deep convolutional neural network for ulcer recognition in wireless capsule endoscopy: Experimental feasibility and optimization. *Comput. Math. Methods Med.* 2019 (2019).

[222] Sutong Wang, Yunqiang Yin, Dujuan Wang, Yanzhang Wang, and Yaochu Jin. 2021. Interpretability-based multimodal convolutional neural networks for skin lesion diagnosis. *IEEE Trans. Cybernet.* 52, 12 (2021), 12623–12637.

[223] Shengjie Wang, Tianyi Zhou, and Jeff Bilmes. 2019. Bias also matters: Bias attribution for deep neural network explanation. In *Proceedings of the International Conference on Machine Learning*. PMLR, 6659–6667.

[224] Xi Wang, Hao Chen, An-Ran Ran, Luyang Luo, Poemen P. Chan, Clement C. Tham, Robert T. Chang, Suria S. Mannil, Carol Y. Cheung, and Pheng-Ann Heng. 2020. Towards multi-center glaucoma OCT image screening with semi-supervised joint structure and function multi-task learning. *Med. Image Anal.* 63 (2020), 101695.

[225] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, and Ronald M. Summers. 2018. Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 9049–9058.

[226] Kristoffer Wickstrøm, Michael Kampffmeyer, and Robert Jenssen. 2020. Uncertainty and interpretability in convolutional neural networks for semantic segmentation of colorectal polyps. *Med. Image Anal.* 60 (2020), 101619.

[227] Paul Windisch, Pascal Weber, Christoph Fürweger, Felix Ehret, Markus Kufeld, Daniel Zwahlen, and Alexander Muacevic. 2020. Implementation of model explainability for a basic brain tumor detection using convolutional neural networks on MRI slices. *Neuroradiology* 62, 11 (2020), 1515–1518.

[228] Eloise Withnell, Xiaoyu Zhang, Kai Sun, and Yike Guo. 2021. XOmiVAE: An interpretable deep learning model for cancer classification using high-dimensional omics data. *Brief. Bioinf.* 22, 6 (2021), bbab315.

[229] Svante Wold, Kim Esbensen, and Paul Geladi. 1987. Principal component analysis. *Chemometr. Intell. Lab. Syst.* 2, 1-3 (1987), 37–52.

[230] Botong Wu, Zhen Zhou, Jianwei Wang, and Yizhou Wang. 2018. Joint learning for pulmonary nodule segmentation, attributes and malignancy prediction. In *Proceedings of the IEEE 15th International Symposium on Biomedical Imaging (ISBI'18)*. IEEE, 1109–1113.

[231] T. Wu, M. Tulio Ribeiro, J. Heer, and D. Weld. 2021. Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models. In *Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP'21)*.

[232] Gabrielle Ras, Ning Xie, Marcel Van Gerven, and Derek Doran. 2020. Explainable deep learning: A field guide for the uninitiated. *Journal of Artificial Intelligence Research* 73 (2022), 329–396.
[233] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the International Conference on Machine Learning*. PMLR, 2048–2057.
[234] Weizheng Yan, Sergey Plis, Vince D. Calhoun, Shengfeng Liu, Rongtao Jiang, Tian-Zi Jiang, and Jing Sui. 2017. Discriminating schizophrenia from normal controls using resting state functional network connectivity: A deep neural network and layer-wise relevance propagation method. In *Proceedings of the IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP'17)*. IEEE, 1–6.
[235] Guang Yang, Qinghao Ye, and Jun Xia. 2022. Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond. *Inf. Fusion* 77 (2022), 29–52.
[236] Hugo Yeche, Justin Harrison, and Tess Berthier. 2019. UBS: A dimension-agnostic metric for concept vector interpretability applied to radiomics. In *Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support*. Springer, 12–20.
[237] Chih-Kuan Yeh, Been Kim, Sercan Arik, Chun-Liang Li, Tomas Pfister, and Pradeep Ravikumar. 2020. On completeness-aware concept-based explanations in deep neural networks. *Adv. Neural Inf. Process. Syst.* 33 (2020), 20554–20565.
[238] Changchang Yin, Buyue Qian, Jishang Wei, Xiaoyu Li, Xianli Zhang, Yang Li, and Qinghua Zheng. 2019. Automatic generation of medical imaging diagnostic report with hierarchical recurrent neural network. In *Proceedings of the IEEE International Conference on Data Mining (ICDM'19)*. IEEE, 728–737.
[239] Zhitao Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. 2019. Gnnexplainer: Generating explanations for graph neural networks. *Adv. Neural Inf. Process. Syst.* 32 (2019).
[240] Kyle Young, Gareth Booth, Becks Simpson, Reuben Dutton, and Sally Shrapnel. 2019. Deep neural network or dermatologist? In *Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support*. Springer, 48–55.
[241] Matthew D. Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *Proceedings of the European Conference on Computer Vision*. Springer, 818–833.
[242] Bofei Zhang, Jimin Tan, Kyunghyun Cho, Gregory Chang, and Cem M. Deniz. 2020. Attention-based cnn for kl grade classification: Data from the osteoarthritis initiative. In *Proceedings of the IEEE 17th International Symposium on Biomedical Imaging (ISBI'20)*. IEEE, 731–735.
[243] Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. 2018. Top-down neural attention by excitation backprop. *Int. J. Comput. Vis.* 126, 10 (2018), 1084–1102.
[244] Jing Zhang, Caroline Petitjean, Florian Yger, and Samia Ainouz. 2020. Explainability for regression CNN in fetal head circumference estimation from ultrasound images. In *Interpretable and Annotation-Efficient Learning for Medical Image Computing*. Springer, 73–82.
[245] Quanshi Zhang, Yu Yang, Haotian Ma, and Ying Nian Wu. 2019. Interpreting cnns via decision trees. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6261–6270.
[246] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating text generation with BERT. In *Proceedings of theInternational Conference on Learning Representations*.
[247] Zizhao Zhang, Yuanpu Xie, Fuyong Xing, Mason McGough, and Lin Yang. 2017. Mdnet: A semantically and visually interpretable medical image diagnosis network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6428–6436.
[248] Guannan Zhao, Bo Zhou, Kaiwen Wang, Rui Jiang, and Min Xu. 2018. Respond-cam: Analyzing deep models for 3d imaging data by visualizations. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 485–492.
[249] Kaiping Zheng, Shaofeng Cai, Horng Ruey Chua, Wei Wang, Kee Yuan Ngiam, and Beng Chin Ooi. 2020. Tracer: A framework for facilitating accurate and interpretable analytics for high stakes applications. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*. 1747–1763.
[250] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. 2016. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2921–2929.
[251] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*. 2223–2232.

Received 14 October 2022; revised 28 August 2023; accepted 1 December 2023