Self-eXplainable AI for Medical Image Analysis: A Survey and New Outlooks

Junlin Hou, Sicen Liu, Yequan Bie, Hongmei Wang, Andong Tan, Luyang Luo, Hao Chen

***Abstract*— The increasing demand for transparent and reliable models, particularly in high-stakes decision-making areas such as medical image analysis, has led to the emergence of eXplainable Artificial Intelligence (XAI). Post-hoc XAI techniques, which aim to explain black-box models after training, have been controversial in recent works concerning their fidelity to the models' predictions. In contrast, Self-eXplainable AI (S-XAI) offers a compelling alternative by incorporating explainability directly into the training process of deep learning models. This approach allows models to generate inherent explanations that are closely aligned with their internal decision-making processes. Such enhanced transparency significantly supports the trustworthiness, robustness, and accountability of AI systems in real-world medical applications. To facilitate the development of S-XAI methods for medical image analysis, this survey presents an comprehensive review across various image modalities and clinical applications. It covers more than 200 papers from three key perspectives: 1) input explainability through the integration of explainable feature engineering and knowledge graph, 2) model explainability via attention-based learning, concept-based learning, and prototype-based learning, and 3) output explainability by providing counterfactual explanation and textual explanation. Additionally, this paper outlines the desired characteristics of explainability and existing evaluation methods for assessing explanation quality. Finally, it discusses the major challenges and future research directions in developing S-XAI for medical image analysis.**

***Index Terms*— Self-eXplainable Artificial Intelligence (S-XAI), Medical Image Analysis, Input Explainability, Model Explainability, Output Explainability, S-XAI Evaluation**

I. Introduction

Artificial intelligence (AI), particularly deep learning, has driven significant advancements in medical image analysis, including applications in disease diagnosis, lesion segmentation, medical report generation (MRG), and visual question answering (VQA). Deep neural networks (DNNs) automatically learn features from input data and produce optimal outputs. However, the inherent complexity nature of DNNs hinder our understanding of the decision-making processes behind these models. Consequently, DNNs are often considered as black-box models, which has raised concerns about their transparency, interpretability, and accountability for their successful deployment in real-world clinical applications [1].

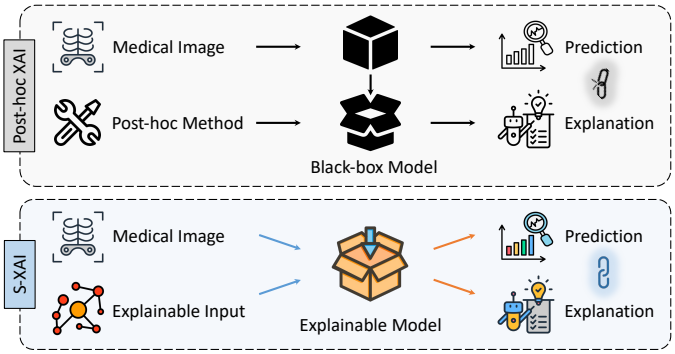

Fig. 1. Illustration of post-hoc XAI versus Self-eXplainable AI (S-XAI).

To tackle the challenge of developing more trustworthy AI systems, research efforts are increasingly focusing on various eXplainable AI (XAI) methods, enhancing transparency [2], fairness [3], and robustness [4]. However, most XAI methods aim to generate explanations for the outputs of black-box AI models after they have been trained, a category known as post-hoc XAI, as illustrated in Fig. 1 top. These methods utilize additional explanation models or algorithms to provide insights into the decision-making process of the primary AI model. In the field of medical image analysis, commonly used post-hoc XAI techniques include feature attribution methods, such as gradient-based approaches (e.g., LRP [5], CAM [6]) and perturbation-based approaches (e.g., LIME [7], Kernel SHAP [8]). Additionally, some methods explored concept attributions, learning human-defined concepts from the internal activations of DNNs (e.g., TCAV [9], CAR [10]). Post-hoc XAI techniques are often model-agnostic, indicating that they can be flexibly applied to a variety of already-trained black-box AI models.

Since post-hoc explanations are generated separately from the primary AI model, several valid concerns have been raised: 1) these explanations may not always be faithful to the actual decision-making process of black-box models [11], [12]; 2) they may lack sufficient detail to fully elucidate the model's functioning [13]. These limitations of post-hoc XAI approaches are particularly problematic in high-stakes

This work was supported by the Hong Kong Innovation and Technology Fund (Project No. MHP/002/22), HKUST (Project No. FS111) and Research Grants Council of the Hong Kong (No. R6003-22 and T45-401/22-N).

J. Hou, Y. Bie, H. Wang, and A. Tan are with the Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Hong Kong, China (email: csejlhou@ust.hk)

S. Liu is with the Department of Engineering, Shenzhen MSU-BIT University, Shenzhen, China (email: liusicen@smbu.edu.cn)

L. Luo is with the Department of Biomedical Informatics, Harvard University, Cambridge, USA (email: luyang_luo@hms.harvard.edu)

H. Chen is with the Department of Computer Science and Engineering, Department of Chemical and Biological Engineering and Division of Life Science, Hong Kong University of Science and Technology, Hong Kong, China; HKUST Shenzhen-Hong Kong Collaborative Innovation Research Institute, Futian, Shenzhen, China. (email: jhc@cse.ust.hk)

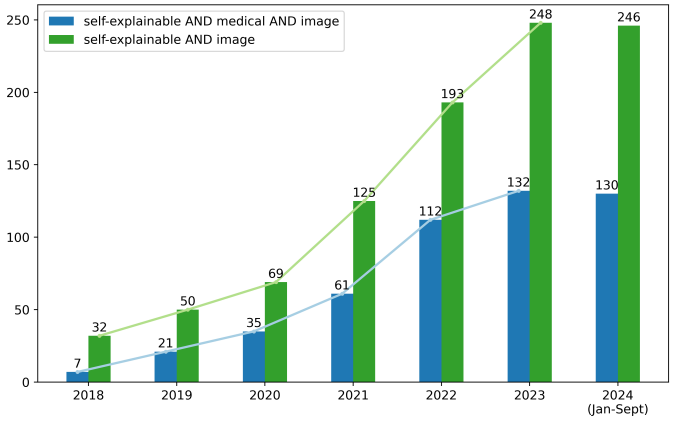

Fig. 2. The upward trend of the total number of S-XAI research papers from 2018 to 2024 (Jan-Sept), where medical articles account for half.

domains like medical image analysis, where clinicians require a deep and trustworthy understanding of how an AI model arrives at its predictions. The issues about the faithfulness and sufficiency of post-hoc explanations highlight the importance of exploring self-explainable AI models as a potentially more reliable and transparent alternative.

Self-eXplainable AI (S-XAI) is a category of XAI methods designed to be interpretable by nature, as illustrated in Fig. 1 bottom. These methods incorporate explainability as an integral part of the model during the training process, rather than generating explanations after the model has been trained. Conventional inherently interpretable methods include various white-box machine learning models, such as decision trees [14], generalized additive models [15], and rule-based systems [16]. In this survey, we focus primarily on DNNs and extend the characteristics of self-explainability across the entire pipeline, from model input to architecture to output, enabling direct inspection and understanding of the reasoning behind the model's predictions without reliance on external explanation methods. In contrast to post-hoc XAI approaches, S-XAI methods aim to provide explanations that are inherent, transparent, and faithful, aligning directly with the model's internal decision-making mechanisms. Such explanations are essential for the effective adoption and clinical integration of AI-powered decision support systems. Furthermore, S-XAI facilitates collaborative decision-making between clinicians and AI systems, fostering better-informed and more accountable medical diagnoses and interventions.

This paper presents the first systematic review of S-XAI for medical image analysis, covering methodology, medical applications, and evaluation metrics, while also offering an in-depth discussion on challenges and future directions. Although there is a wealth of literature on medical XAI surveys [2], [17]–[22] that deliver valuable insights, none have focused specifically on a comprehensive review of S-XAI methods applied to medical image analysis. We analyze more than 200 papers published from 2018 to September 2024, sourced from the proceedings of NeurIPS, ICLR, ICML, AAAI, CVPR, ICCV, ECCV, and MICCAI as well as top-tier journals in the field, including Nature Medicine, Transactions on Pattern Analysis and Machine Intelligence, Transactions on Medical Imaging, Medical Image Analysis, or those cited in related works. The statistics of research articles using keywords *self-explainable*, *medical*, *image* on Google Scholar are presented in Fig. 2, which reveal two key observations: 1) there has been a significant and consistent increase in research papers focused on self-explainable AI over the years, indicating growing interest and emphasis within the research community; 2) nearly half of the total research papers (green bars) are dedicated to applying S-XAI techniques in medical imaging (blue bars), highlighting the vital importance of S-XAI in the medical field.

To summarize, this survey presents insights into S-XAI for medical image analysis, with our contributions outlined below:

1) **Novel Scope of XAI Survey:** As an emerging XAI method that actively offers explainability from the model itself, S-XAI is attracting growing attention from the research community. This work represents the first comprehensive survey on this topic.
2) **Systematic Review of Methods:** We present a novel taxonomy of relevant papers and review them based on input explainability, model explainability, and output explainability. This offers insights into potential technical innovations for S-XAI methods.
3) **Thorough Overview of Applications:** We overview various applications across different anatomical locations and modalities in current S-XAI research. This illustrates the ongoing development of S-XAI technologies in medical image practices, serving as a reference for future applications in diverse scenarios.
4) **Comprehensive Survey of Evaluations:** We analyze a range of desired characteristics and evaluation methods to assess the quality of explainability. This provides guidelines for developing clinically explainable AI systems that are trustworthy and meaningful for end-users.
5) **In-depth Discussion of Challenges and Future Work:** We discuss the key challenges and look forward to the promising future directions. This highlights current shortcomings and identifies new opportunities for researchers to drive further advancements.

II. S-XAI in Medical Image Analysis

Transparency and trustworthiness are essential for deep learning models deployed in real-world applications of medical image analysis. To address this need, the research community has explored various XAI methods and proposed several XAI taxonomy. According to existing literature [17], [19], [23], XAI methods can be categorized by the following criteria.

1) Intrinsic versus Post-hoc: This criteria differentiates whether interpretability is inherent to the model's architecture (intrinsic) or achieved after the model training (post-hoc).

2) Model-specific versus Model-agnostic: Model-specific methods are restricted to particular model classes, whereas model-agnostic methods can be applied to explain any model;

3) Local versus Global: The scope of an explanation distinguishes between those for an individual prediction (local) or those for the entire model behavior (global).

4) Explanation Modality: The common types of explanation include visual explanation, textual explanation, concept explanation, sample explanation, etc.

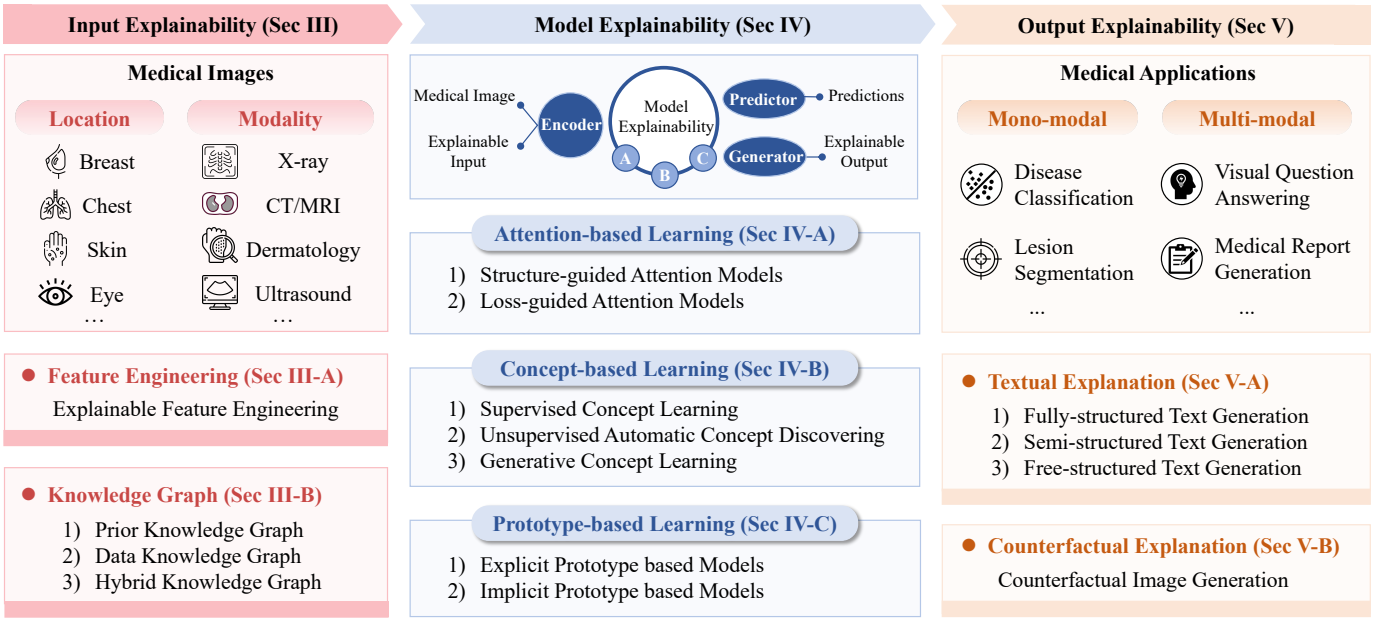

Fig. 3. Overview of Self-eXplainable AI (S-XAI) frameworks including input explainability, model explainability, and output explainability.

This survey concentrates on S-XAI methods for medical image analysis that allow models to inherently explain their own decision-making. As depicted in Fig. 3, we introduce a new taxonomy of S-XAI based on the three key components of DNNs.

1) Input Explainability (Sec. III): Input explainability focuses on integrating additional explainable inputs with deep features of medical images obtained from various anatomical locations and modalities to produce final predictions. By incorporating external knowledge and context-specific information, the accuracy and reliability of these predictions can be significantly improved.

2) Model Explainability (Sec. IV): Model explainability aims to design inherently intepretable model architectures of DNNs. Instead of explaining a black-box model, transforming the model into an interpretable format enhances understanding of how it processes information.

3) Output Explainability (Sec. V): Output explainability refers to the model's ability to generate not just predictions for various medical image tasks but also accompanying explanations through an explanation generator. This capability aids in understanding the rationale behind the model's predictions, facilitating informed medical decision-making.

The following sections summarize and categorize the most relevant works on S-XAI methods applied to medical image analysis. Comprehensive lists of the reviewed S-XAI methods are provided, detailing the employed S-XAI techniques, publication year, anatomical location, image modality, medical application, and the datasets used.

III. Input Explainability

In this section, we will explore input explainability by integrating external domain knowledge, focusing on two key approaches, i.e., a) explainable feature engineering (Sec. III-A) and b) knowledge graph (Sec. III-B). As shown in Fig. 4, these explainable inputs will interact with the deep features of image inputs and be combined to support final predictions.

A. Explainable Feature Engineering

Feature engineering focuses on transforming raw images into a more useful set of human-interpretable features. This process is crucial for traditional machine learning methods to achieve accurate predictions, but it can be time-consuming and demands significant domain expertise. In contrast, deep learning models automatically extract features from raw images, simplifying the manual crafting process but often resulting in reduced interpretability. A promising approach to enhance input explainability is to incorporate explainable feature engineering into deep learning, which injects domain knowledge into the model, as shown in Fig. 4(a). This integration enhances the model's interpretability by ensuring that the learned features are relevant and meaningful for clinical applications. Ultimately, this method improves model performance and offers valuable insights into the decision-making process.

A common strategy in explainable feature engineering is to combine both handcrafted and deep features from an input image to make final predictions [24], [25]. For example, Kapse *et al.* [24] introduce a self-interpretable multiple instance learning (SI-MIL) framework that simultaneously learns from deep image features and handcrafted morphometric and spatial descriptors. They assess the local and global interpretability of SI-MIL through statistical analysis, a user study, and key interpretability criteria. Another line of approach involves incorporating interpretable clinical variables as additional inputs alongside the images, often utilizing multimodal learning techniques [26], [27]. For instance, Xiang *et al.* [26] introduce OvcaFinder, an interpretable model that combines deep learning predictions from ultrasound images with Ovarian–Adnexal Reporting and Data System scores provided by radiologists, as well as routine clinical variables for diagnosing ovarian cancer. This approach enhances diagnostic accuracy and explains the impact of key features and regions on the prediction outcomes.

Discussion: Although explainable feature engineering can be time-consuming, it brings valuable prior knowledge and en-

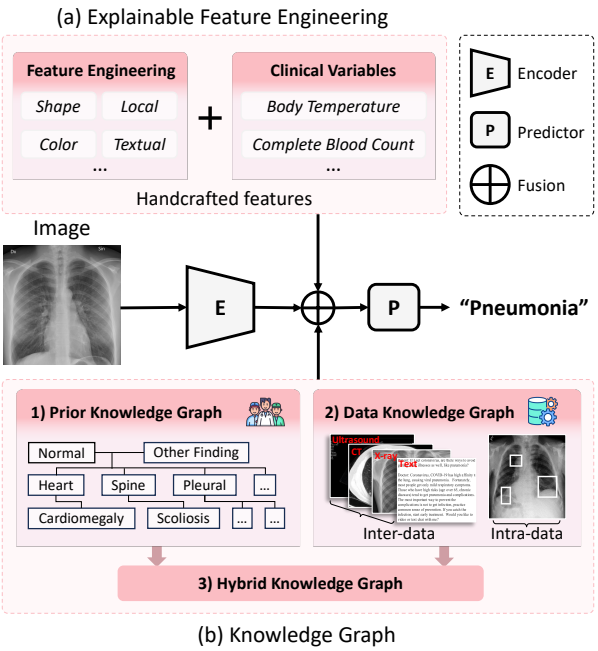

Fig. 4. Input explainability that incorporates (a) explainable feature engineering (b) knowledge graph as additional inputs.

hances the interpretability of deep learning models concerning input features. Despite the increasing research in this area, most studies prioritize accuracy improvements, with limited analysis given to the explainability. Additionally, effective information fusion and interaction poses a key challenge.

B. Knowledge Graph

A knowledge graph (KG) is a structured representation of factual knowledge that captures relationships between entities in a specific area. It provides a way to organize and represent knowledge in a semantically rich and interconnected manner and plays a crucial role in enhancing the interpretability of S-XAI models. Recently, integrating structured domain knowledge into downstream tasks has attracted significant attention of both industry and academia [28]–[30]. This growing interest stems from the recognition that leveraging domain knowledge can greatly improve the performance and effectiveness of various applications. As shown in Fig 4(b), regarding medical imaging analysis, the utilization of KG can be broadly categorized into three categories: 1) prior KG, which serves as a foundational resource that gathers existing domain expertise and established medical knowledge; 2) data KG, which is derived from the analysis of large-scale medical imaging datasets; and 3) hybrid KG, which combines the strengths of both prior and data KGs for medical image analysis.

1) Prior Knowledge Graph: A prior KG in the medical domain is a specialized KG that captures and organizes facts information about medical concepts and their relationships. It can be constructed from a multi-sources, including medical literature, electronic health records, medical ontologies, clinical guidelines, and expert opinions. This graph serves as a comprehensive repository of medical knowledge, encompassing details about diseases, symptoms, treatments, medications, anatomical structures, and more. It provides a vital foundation for medical decision-making, clinical research, and healthcare analytics [31]–[33]. By harnessing the medical prior knowledge encoded in the graph, AI models can gain valuable insights, identify patterns, predict patient outcomes, assist in diagnosis, recommend personalized treatments, and ultimately improve patient care and outcomes [34]–[38]. For example, Liu *et al.* [36] and Huang *et al.* [37] develop KGs based on the professional perspective related to medical images to enhance image understanding. Another way to utilize prior knowledge is by collecting a large number of relationship triples to create a domain-knowledge-enhanced medical VQA dataset. For instance, Liu *et al.* [38] extract a set of 52.6K triplets in the format $< head, relation, tail >$ containing medical knowledge from OwnThink (https://www.ownthink.com). They then use this external information to create SLAKE, a large-scale, semantically annotated, and knowledge-enhanced bilingual dataset for training and testing Med-VQA systems.

Prior KGs enhance S-XAI models by integrating expert-derived knowledge and medical facts, enabling these models to better understand key medical concepts and make more informed predictions. However, the creation of these KGs largely depends on specialized expertise, making the process labor-intensive. Furthermore, these KGs often lack the adaptability required for analyzing dynamic clinical datasets.

2) Data Knowledge Graph: A data KG differs from a prior KG in its construction methodology. While a prior KG relies on expert insights and established medical facts, a data KG is built directly from the dataset itself. This means that instead of relying solely on pre-existing knowledge, the data knowledge graph leverages the inherent information contained within the dataset. This approach allows the data KG to provide a unique perspective and the potential to discover previously unknown relationships and correlations within the data [64]–[67]. There are two primary approaches to leveraging data knowledge for enhancing the explainability of AI models: 1) extracting knowledge directly from the dataset [43], [45], [46], [49], [51], [57], [61]. Liu *et al.* [49] employ a bipartite graph convolutional network to model the intrinsic geometric and semantic relation of ipsilateral views, and an inception graph convolutional network to model the structural similarities of bilateral views. Huang *et al.* [61] develop a medical KG based on the types of diseases and questions concerned by patients during their treatment process. 2) Transferring knowledge from pre-trained models. For example, Qi *et al.* [52] use a pre-trained U-Net to segment lung lobes and then model both the intra-image and inter-image relationships of these lobes and in-batch images through their respective graphs. Elbatel *et al.* [53] distill knowledge from pre-training models to small models for disease classification.

Overall, constructing a data KG involves leveraging the inherent characteristics of the dataset itself to build a graph structure to assist S-XAI models. However, it is important to note that these methods often harbor inherent biases that can vary significantly across different datasets.

3) Hybrid Knowledge Graph: A hybrid KG integrates both the prior KG and the data KG, representing an interactive approach. The prior KG provides a static foundation of established medical facts, while the data KG utilizes dataset charac-

TABLE I
INPUT EXPLAINABILITY METHODS BASED ON KNOWLEDGE GRAPH (KG). THE ABBREVIATIONS HERE ARE CLS: CLASSIFICATION, DET: DETECTION, MRG: MEDICAL REPORT GENERATION, VQA: VISUAL QUESTION ANSWERING.

Method	Year	Location	Modality	Task	Dataset	KG Type
Naseem *et al.* [34]	2023	Multiple	Pathology	VQA	[39]	Prior KG
Zhang *et al.* [35]	2020	Chest	X-ray	MRG	[40]	Prior KG
Liu *et al.* [36]	2021	Chest	X-ray	MRG	[41], [42]	Prior KG
Huang *et al.* [37]	2023	Chest	X-ray	MRG	[41], [42]	Prior KG
Liu *et al.* [38]	2021	Multiple	X-ray, CT, MRI	VQA	[38]	Prior KG
Chen *et al.* [43]	2020	Chest	X-ray	CLS	[40], [44]	Data KG
Zheng *et al.* [45]	2021	Chest	X-ray, CT, US, text	CLS	private	Data KG
Hou *et al.* [46]	2021	Chest	X-ray	CLS	[41], [42]	Data KG
Zhou *et al.* [47]	2021	Chest	X-ray	CLS	[40], [44]	Hybrid KG
Wu *et al.* [48]	2023	Chest	X-ray	CLS	[42]	Hybrid KG
Liu *et al.* [49]	2021	Breast	Mammogram	DET	[50]	Data KG
Zhao *et al.* [51]	2021	Chest	X-ray	DET	[40]	Data KG
Qi *et al.* [52]	2022	Chest	X-ray	DET	[40]	Data KG
Elbatel *et al.* [53]	2023	Chest	Dermatology, Endoscopy	DET	[54], [55]	Data KG
Li *et al.* [56]	2019	Chest	X-ray	MRG	[41]	Hybrid KG
Liu *et al.* [57]	2021	Chest	X-ray	MRG	[41], [42]	Data KG
Li *et al.* [58]	2023	Chest	X-ray	MRG	[41], [42]	Hybrid KG
Kale *et al.* [59]	2023	Chest	X-ray	MRG	[41]	Hybrid KG
Guo *et al.* [60]	2022	Multiple	X-ray, CT, MRI	VQA	[38]	Prior KG
Huang *et al.* [61]	2023	Multiple	X-ray, CT, MRI, US	VQA	[38]	Data KG
Hu *et al.* [62]	2023	Chest	X-ray	VQA	[62]	Hybrid KG
Hu *et al.* [63]	2024	Chest	X-ray	VQA	[62]	Hybrid KG

teristics to dynamically update and enhance this foundational knowledge. By incorporating data-specific insights discovered from the dataset, the hybrid KG allows for the integration of new information and the refinement of existing knowledge. This dynamic updating process ensures that the KG remains up-to-date and relevant. Consequently, the hybrid KG combines the strengths of both the prior and data KG, offering a more comprehensive and adaptable knowledge representation for S-XAI models in the medical field [35], [47], [48], [56], [58], [59], [62], [63]. For instance, Wu *et al.* [48] implement a triplet extraction module to extract medical information from reports, combining entity descriptions with visual signals at the image patch level for medical diagnosis. For the medical report generation tasks, Li *et al.* [56] decompose medical report generation into explicit medical abnormality graph learning and subsequent natural language modeling. Each node in the abnormality graph represents a possible clinical abnormality based on prior medical knowledge, with the correlations among these nodes encoded as edge weights to inform clinical diagnostic decisions. Hu *et al.* [63] utilize large language models to extract labels and build a large-scale medical VQA dataset, Medical-CXR-VQA. They then leverage graph neural networks to learn logical reasoning paths based on this dataset for medical visual question answering task.

In summary, the construction of a hybrid KG relies on prior knowledge and involves automatically adjusting the nodes or edges based on the data characteristics. This process ensures that the KG remains aligned with the specific domain knowledge and captures the most relevant, data-specific information. It provides a comprehensive representation of both data-specific knowledge and prior knowledge, enhancing the interpretability of S-XAI models.

Discussion: The utilization of medical KGs in medical image analysis poses both challenges and promising opportunities. First, integrating diverse prior medical knowledge into a graph format is labor-intensive and costly, requiring constant updates and refinements to incorporate the latest research findings, clinical guidelines, and emerging medical data to maintain up-to-date prior medical knowledge. Another challenge lies in the heterogeneity of medical image data. With the continuous growth of medical image data, the variety of image modalities expands, complicating their effective integration within KGs. Developing robust algorithms to extract meaningful features from medical images and link them with relevant medical KGs remains an ongoing research endeavor.

IV. MODEL EXPLAINABILITY

In this section, we present model explainability by designing interpretable model architectures, such as attention-based learning (Sec. IV-A), concept-based learning (Sec. IV-B), and prototype-based learning (Sec. IV-C).

A. Attention-based Learning

Attention-based learning aims to capture specific areas in an image that are relevant to the prediction task while suppressing irrelevant regions based on feature maps. Therefore, it can be naturally combined with S-XAI methods to provide visual explanations that enhance model decision-making [2], [19], [68]. We categorize attention-based S-XAI models into 1) structure-guided attention models and 2) loss-guided attention models. As illustrated in Fig. 5, the former specifically designs the attention structure and obtains model predictions directly from the attention map, while the latter constrains the attention map using a loss function to ensure attention map align with an ideal interpretable distribution.

1) Structure-guided Attention Model: As shown in the top branch of Fig. 5, structure-guided attention models are characterized by the association between the attention structure in the model and the components directly influencing the

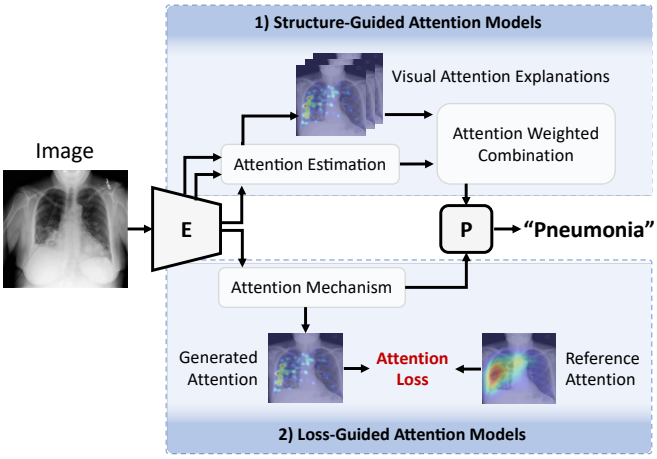

Fig. 5. Attention-based learning, including 1) structure-guided and 2) loss-guided attention models. X-ray images borrowed from [69].

model's predictions. This allows the generated attention map to effectively explain the model's predictions. Jetley *et al.* [70] is the first to introduce attention learning for XAI. They propose an attention estimator which calculates feature compatibility scores to weight feature maps as feature activation scores, which are then directly used as input for a linear classifier. This approach guides the model's attention toward areas that are more relevant to its decision-making while suppressing irrelevant regions. Fukui *et al.* [71] present an Attention Branch Network (ABN), which replaces the fully connection layer of Class Activation Mapping (CAM) [6] using a convolution layer to output class probabilities. There is also a perception branch to apply a classifier to the combination of attention maps from the original features. Furthermore, Li *et al.* [72] propose a slot attention-based method in which the attention output of each slot are directly processed and summed up by a main block named SCOUTER to support a specific category, eliminating the need for a linear classifier and further improving the model's transparency. They also use the output from the slot attention mechanism to represent the model's final confidence for each category. Notably, positive and negative interpretations can be controlled through the parameters in the loss function. This method demonstrates improved interpretability in the glaucoma diagnosis task.

Numerous studies have incorporated multiple attention mechanisms for medical image classification and segmentation tasks [73]–[75]. For example, Schempler *et al.* [73] extend the attention estimator by extracting local information from coarse-scale feature maps for attention gates, facilitating more fine-grained visual interpretation for lesion segmentation or ultrasound diagnosis. Similarly, Gu *et al.* [74] develop a comprehensive attention module that enhances model interpretability through spatial, channel, and scale attention. Their segmentation experiments on skin lesions and fetal organs demonstrate improved performance and better interpretability of target area positioning and scale. Beyond 2D data, how to use attention to explain more complex 3D medical image diagnosis is more challenging. Lozupone *et al.* [76] present an attention module that fuses attention weights from sagittal, coronal, and axial slices to diagnose Alzheimer's disease on 3D MRI brain scans. By integrating these attention scores from three directions, the 3D attention map can be visualized to explain the model's decision-making process. For the fast MRI reconstruction task, Huang *et al.* [77] propose a shifted windows deformable attention mechanism which uses reference points to impose spatial constraints on attention and directly combines the outputs from the attention modules of different windows to produce the model's reconstruction results.

Although structure-guided attention maps can provide explanations for model predictions, they are still difficult to align with clear human-understandable decision-making basis.

2) Loss-guided Attention Model: As shown in the bottom branch of Fig. 5, loss-guided attention models use interpretable labels (i.e., reference attention maps) to construct loss functions that directly constrain the generated attention maps. This method encourages the model to focus on areas that are understandable and beneficial for making predictions. Benefiting from lesion area annotations and professional analyses by doctors, which provide clear references for model decisions, loss-guided attention learning techniques are commonly used in medical image analysis.

Using ground-truth masks of regions of interest (RoIs) to guide the generation of attention maps is a widely adopted approach in medical image classification [80], [83], [87]. For instance, Yang *et al.* [80] directly optimize the attention maps by a Dice loss, which encourages the model to focus on target areas that are highly relevant to the classification of breast cancer microscopy images. To alleviate the challenge of obtaining pixel-level annotations, Yin *et al.* [87] pre-train a histological feature extractor to identify significant clinically relevant feature masks, which are then used to guide and regularize the attention maps. By considering the varying contributions of histological features for classification, the model can selectively focus on different features based on the distribution of nuclei in each instance. In medical image segmentation, labels corresponding to edges and shapes of specific regions are often reused to guide attention in learning semantic information [102], [105], [109]. Sun *et al.* [102] combine spatial attention with the attention estimator in U-Net decoders, enabling the model to interpret learned features at each resolution. They also introduce a gated shape stream alongside the texture stream, where the resulting shape attention maps are aligned with actual edges through binary cross-entropy loss, enhancing the cardiac MRI segmentation.

Compared with lesion masks, eye tracking data provides a more accurate depiction of expert focus, as it captures the way doctors visually process information during diagnosis. Bhattacharya *et al.* [69] leverage the captured doctors' attention to guide model training. They employ a teacher-student network to replicate the visual cognitive behavior of doctors when diagnosing diseases on chest radiographs. The teacher model is trained based on the visual search patterns of radiologists, and the student model utilizes an attention loss to predict attention from the teacher network using eye tracking data.

Discussion: Attention-based S-XAI methods guide model predictions by focusing on critical areas of images, thereby providing effective attention explanations. Structure-guided attention models typically utilize the attention-weighted output as input for the predictor, reflecting the model's decision-

TABLE II
MODEL EXPLAINABILITY METHODS BASED ON ATTENTION-BASED LEARNING. THE ABBREVIATIONS HERE ARE CLS: CLASSIFICATION, SEG: SEGMENTATION, IRE: IMAGE RECONSTRUCTION, REG: REGRESSION

Method	Year	Location	Modality	Task	Dataset	Attention Type
Wang *et al.* [78]	2018	Breast	X-ray	CLS	[79]	Structure-Guided
Yang *et al.* [80]	2019	Breast	Histopathology	CLS	[81]	Loss-Guided
Li *et al.* [72]	2021	Eye	Retinal images	CLS	[82]	Structure-Guided
Yan *et al.* [83]	2019	Skin	Dermatology	CLS	[84], [85]	Loss-Guided
Barata *et al.* [75]	2021	Skin	Dermatology	CLS	[85], [86]	Loss-Guided
Yin *et al.* [87]	2021	Liver	Histopathology	CLS	[88]	Loss-Guided
Bhattacharya *et al.* [69]	2022	Chest	X-ray	CLS	[89]–[98]	Loss-Guided
Lozupone *et al.* [76]	2024	Brain	MRI	CLS	[99]	Structure-Guided
Schempler *et al.* [73]	2019	Abdominal, Fetal	CT, US	SEG + DET	[100], [101]	Structure-Guided
Sun *et al.* [102]	2020	Cardiac	MRI	SEG	[103], [104]	Loss-Guided
Gu *et al.* [74]	2020	Skin, Fetal	Dermatology, MRI	SEG	[86]	Structure-Guided
Karri *et al.* [105]	2022	Skin, Brain, Abdominal	Dermatology, MRI, CT	SEG	[106]–[108]	Loss-Guided
Li *et al.* [109]	2023	Skin	Dermatology	SEG	[54], [85], [86], [106]	Loss-Guided
Huang *et al.* [77]	2022	Head	MRI	IRE	[110]	Structure-Guided
Lian *et al.* [111]	2019	Brain	MRI	REG	[99], [112]	Structure-Guided

making basis. However, these attention explanations often lack clear semantic information and can be subjectively interpreted. In contrast, loss-guided attention models generally provide attention explanations with explicit semantic details. However, since the attention output does not directly influence the model's decisions, evaluating how well these attention maps explain the decision-making process remains challenging. Overall, while attention-based S-XAI methods enhance model transparency and offer insights into decision-making, the understandability of attention maps and their relevance to the decisions still require further investigation.

B. Concept-based Learning

Concept-based S-XAI methods provide explanations in terms of high-level, human-interpretable attributes rather than low-level, non-interpretable features. This approach reveals the inner workings of deep learning models using easily understandable concepts, enabling users to gain deeper insights into underlying reasoning. It also helps in identifying model biases and allows for adjustments to enhance performance and trustworthiness. Most concept-based S-XAI methods focus on making decisions based on a set of concepts while also detailing the contribution of each concept to the final prediction [113]–[116]. These methods introduce concept learning into the training pipeline of the models, instead of simply analyzing explainability after training a black-box model (i.e., post-hoc XAI methods) [117]–[119]. We propose to categorize concept-based S-XAI methods into three types: 1) supervised concept learning, 2) unsupervised automatic concept discovering, and 3) generative concept learning, as shown in Fig. 6.

The term *Concept* has been defined in different ways, which commonly represents high-level attributes [117], [120], [121]. In this paper, we suggest adopting a straightforward and easily understandable categorization: *Textual Concepts* and *Visual Concepts*. *Textual Concepts* refer to textual descriptions of attributes associated with the classes. For example, in Fig. 6, the textual concepts for the classes (i.e., "pneumonia" and "normal") include terms like *Opacity, Effusion, Infiltration*, etc. *Visual Concepts*, on the other hand, consist of semantically meaningful features within the image that may not be

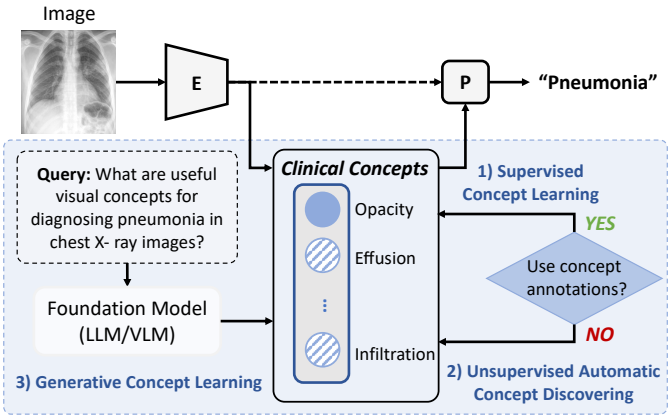

Fig. 6. Concept-based learning, including 1) supervised concept learning, 2) unsupervised automatic concept discovering, and 3) generative concept learning.

explicitly described in natural language. For example, Sun *et al.* [122] consider the instances segmented by SAM [123] as the concepts of a given image.

1) Supervised Concept Learning: Supervised concept learning methods train deep learning models using annotations of textual concepts, particularly by supervising an intermediate layer to represent these concepts. A notable example is Concept Bottleneck Model (CBM) [113], which is an inherently interpretable deep learning architecture. It first maps latent image features to a concept bottleneck layer, where the number of neurons corresponds to the number of human-defined concepts, and then predicts final results based on the concept scores from this layer. By enforcing the neurons in the concept bottleneck layer to learn concept representations supervised by concept labels, CBMs can directly show each concept's contribution to the final prediction (i.e., class-concept relation) using the neuron values of the last layer. Specifically, the authors of CBM conduct experiments on the knee X-ray dataset OAI [124] to explore the importance of concepts such as bone spurs and calcification in determining arthritis grading. Additionally, CBMs allow model editing. When domain experts find certain predicted concept importance unreasonable, they can easily adjust the model's predictions by intervening in the weights

TABLE III
MODEL EXPLAINABILITY METHODS BASED ON CONCEPT-BASED LEARNING. THE ABBREVIATIONS HERE ARE CLS: CLASSIFICATION

Method	Year	Location	Modality	Task	Dataset	Concept
Koh *et al.* [113]	2020	Knee	X-ray	CLS	[124]	Supervised (CBM)
Chauhan *et al.* [125]	2023	Knee, Chest	X-ray	CLS	[124], [44]	Supervised (CBM)
Patricio *et al.* [126]	2023	Skin	Dermatology	CLS	[127], [128]	Supervised (CBM)
Yan *et al.* [129]	2023	Skin	Dermatology	CLS	Private	Supervised (CBM)
Bie *et al.* [130]	2024	Skin	Dermatology	CLS	[127], [128], [131]	Supervised (CBM)
Kim *et al.* [132]	2024	Skin	Dermatology	CLS	[54], [127], [131], [133], [134]	Supervised (CBM)
Lucieri *et al.* [135]	2022	Skin	Dermatology	CLS	[127], [128], [54]	Supervised
Jalaboi *et al.* [136]	2023	Skin	Dermatology	CLS	[137], [138]	Supervised
Hou *et al.* [139]	2024	Skin	Dermatology	CLS	[127], [131]	Supervised
Kim *et al.* [140]	2023	Skin	Dermatology	CLS	[106]	Generated concept
Patricio *et al.* [141]	2024	Skin	Dermatology	CLS	[127], [128], [86]	Generated concept
Pang *et al.* [142]	2024	Skin, Blood cell	Dermatology, Microscopy	CLS	[131], [143]	Supervised (CBM)
Marcinkevivcs *et al.* [144]	2024	Appendix	US	CLS	Private	Supervised (CBM)
Zhao *et al.* [145]	2021	Chest	CT	CLS	[146]	Supervised
Fang *et al.* [147]	2020	Eye	Slit lamp microscopy	CLS	[148]	Concept discovery
Wen *et al.* [149]	2024	Eye	Retinal images	CLS	[150], [151]	Supervised
Kong *et al.* [152]	2022	Thyroid	US	CLS	Private	Concept discovery
Liu *et al.* [153]	2023	Multiple	X-ray, CT	CLS	[90], [154], [155]	Generated concept
Gao *et al.* [156]	2024	Multiple	Dermatology, Pathology, US, X-ray	CLS	[42], [106], [155], [157], [158]	Supervised

of the concept bottleneck layer (test-time intervention). The CBM architecture has inspired many researchers to develop inherently interpretable methods, resulting in a series of CBM-like methods. For example, Concept Embedding Models (CEMs) [159] utilize a group of neurons (concept embeddings) instead of a single neuron to represent a concept, which effectively improves the performance of the original CBM while preserving its interpretability. Different from CBMs, Concept Whitening [139], [160] aims to whiten the latent space of neural networks and aligns the axes of the latent space with known concepts of interest. Zhao *et al.* [145] introduce a hybrid neuro-probabilistic reasoning algorithm for verifiable concept-based medical image diagnosis, which combines clinical concepts with a Bayesian network.

The self-explainable nature of concept-based learning models has led to its application in medical image analysis. Chauhan *et al.* [125] propose Interactive CBMs, which can request labels for certain concepts from a human collaborator. This method is evaluated on chest and knee X-ray datasets. Yan *et al.* [129] discover and eliminate confounding concepts within datasets using spectral relevance analysis [161], and conduct experiments on skin image datasets. Marcinkevics *et al.* [144] adapt CBM for prediction tasks with multiple views of ultrasonography and incomplete concept sets. Kim *et al.* [132] present a medical concept retriever, which connects medical images with text and densely scores images on concept presence. This enables important tasks in medical AI development and deployment, such as data auditing, model auditing, and model interpretation, using a CBM architecture to develop an inherently interpretable model.

However, a significant challenge in supervised concept learning is the scarcity of concept annotations, which require labor-intensive efforts from human experts. Therefore, some researchers prefer unsupervised automatic concept discovering, as it eliminates the need for extra annotations.

2) Unsupervised Automatic Concept Discovering: Models that perform unsupervised concept discovery modify their internal representations to identify concepts within image features without relying on explicit annotations. These discovered concepts may not be associated with human-specified textual concepts. However, these methods can still provide concept-based explanations by visualizing the unsupervised concepts and detailing their contributions to the final predictions. For instance, Ghorbani *et al.* [162] propose Automatic Concept-based Explanations (ACE), which automatically extract visual concepts that are meaningful to humans and important for the network's predictions. Self-Explaining Neural Networks (SENN) [4] first utilize a concept encoder to extract clusters of image representations corresponding to different visual concepts, and also adopt a relevance parametrizer to calculate the relevance scores of concepts. The final prediction is determined by the combination of discovered concepts and the corresponding relevance scores. Inspired by SENN, Sarkar *et al.* [163] propose an ante-hoc explainable framework that includes both a concept encoder and a concept decoder, which map images into concept space and use the concepts to reconstruct the original images, respectively. Yeh *et al.* [164] argue that the discovered concepts may not be sufficient to explain model predictions, so they define a completeness score to evaluate whether the concepts adequately support model predictions and propose a framework for complete concept-based explanations.

Since medical concept annotations are costly and require experts' efforts, unsupervised automatic concept discovering is usually adopted to offer concept-based explanations in medical image analysis without expert-annotated labels. For example, Fang *et al.* [147] address the practical issue of classifying infections by proposing a visual concept mining (VCM) method to explain fine-grained infectious keratitis images. Specifically, they first use a saliency map based potential concept generator to discover visual concepts, and then propose a visual concept-enhanced framework that combines both image-level representations and the discovered concept features for classification. Moreover, Kong *et al.* [152] develop a novel Attribute-Aware Interpretation Learning (AAIL) model to discover clinical concepts, and then adopt a fusion module to integrate these concepts with global features for thyroid nodule diagnosis from ultrasound images.

Although unsupervised automatic concept discovering can offer concept-based explanations, these explanations are abstract and usually cannot be directly described in natural language. To alleviate this issue while also addressing the lack of concept annotations, generative concept learning has become a promising research direction.

3) Generative Concept Learning: Leveraging foundation models, such as Large Language Models (LLMs) and Vision-Language Models (VLMs), can assist in generating and labeling textual concepts. A notable generative concept learning method, namely Language Guided Bottlenecks (LaBo) [165], employs an LLM (GPT-3 [166]) to generate textual concepts for each image category, which are filtered to form the concept bottleneck layer. LaBo then uses a pre-trained VLM (CLIP [167]) to calculate the similarity between input images and the generated concepts to obtain concept scores. The final prediction is based on the multiplication of a weight matrix and these concept scores. Label-free CBM [168] employs a similar pipeline, but trains an independent network that includes a concept bottleneck layer. In the medical domain, Kim *et al.* [140] enhance LaBo [165] by incorporating a more fine-grained concept filtering mechanism and conducted explainability analysis on dermoscopic images, achieving performance improvements compared to the baseline. Similarly, Liu *et al.* [153] employ ChatGPT and CLIP for explainable zero-shot disease diagnosis on X-ray and CT. Bie *et al.* [169] propose an explainable prompt learning framework that leverages medical knowledge by aligning the semantics of images, learnable prompts, and clinical concept-driven prompts at multiple granularities, where the category-wise clinical concepts are obtained by eliciting knowledge from LLMs.

Discussion: Methods that provide concept-based explanations hold significant importance in medical research and applications, particularly in advancing evidence-based medicine. By offering human-understandable explanations, these methods have the capability to help doctors and patients better understand AI-assisted diagnosis, hence holding the potential to make AI technologies effectively supported and disseminated in healthcare. The lack of fine-grained label annotations and the performance-explainability trade-off are the limitations of concept-based methods. Thanks to the development of LLMs, researchers are exploring new ways to alleviate these issues, e.g., generative concept learning [140], [153], [165]. In addition, as there are more and more medical foundation models being developed, incorporating the knowledge of the models and medical experts to efficiently annotate concept labels for datasets will be a promising and meaningful direction. Besides the most popular classification task, other medical applications of concept-based approaches should be further explored.

C. Prototype-based Learning

Prototype-based S-XAI models aims to provide a decision-making process where a model reasons through comparisons with a set of interpretable example prototypes [171]. This reasoning aligns with human recognition patterns, as humans often identify objects by comparing them to example components [172]. These models first extract features from a given

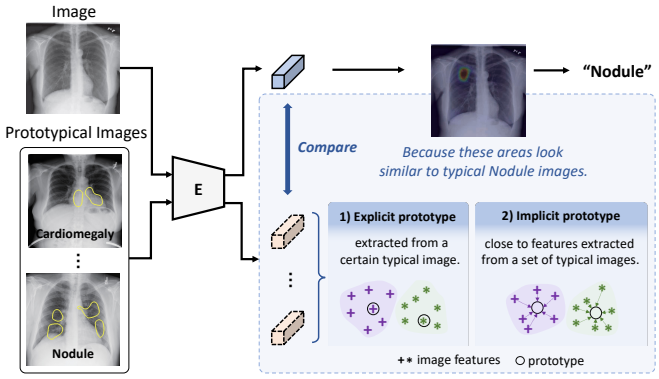

Fig. 7. Prototype-based learning, including 1) explicit prototype and 2) implicit prototype. X-ray images borrowed from [170].

image and then compare the feature maps with the prototypes to calculate similarities. Ultimately, these similarities are combined for the final decision making. This process is considered interpretable because the decision making can be clearly attributed to the contribution of each interpretable prototype (e.g., by the similarity scores). According to how the prototypes are obtained, we define and categorize them to two types: 1) explicit prototypes and 2) implicit prototypes, as presented in Fig. 7. Explicit prototypes are specific high-dimensional feature representations extracted from certain training images, whereas implicit prototypes are latent high-dimensional representations that are close to a set of typical images' representations. All existing prototype-based S-XAI models do not require supervision at the prototype level and aim to automatically find meaningful prototypes to facilitate interpretable decision making.

1) Explicit prototype based models: The first model of this type is ProtopNet [171], which introduces a three-stage training scheme that is widely adopted by subsequent research: 1) Feature extractor training: in this step, the final layer is frozen, and only the feature extraction backbone is trained. 2) Prototype replacement: this step replaces the learned representations in the prototype layer with the nearest feature patch from the training set. 3) Final layer fine-tuning: in this stage, the feature extractor remains fixed while the parameters of the final layer are fine-tuned. Later works closely follow this training scheme while addressing different limitations of this initial framework. For example, ProtoShare [191] proposes to share prototypes across different classes to reduce the overall number of prototypes and enhance model efficiency. A similar idea is explored in ProtoPool [192], where prototypes are assigned to various classes in a differentiable manner. To address the limitation of prior models that use spatially rigid prototypes, ProtoDeform [193] proposes to additionally learn an offset to obtain prototypes that are more spatially flexible. TesNet [194] leverages the Grassman manifold to construct a transparent embedding space, achieving competitive accuracy. Inspired by the theory of support vector machines, ST-protopnet [195] aims to further improve the accuracy of prototype-based models by separating prototypes into support and trivial prototypes, where support prototypes are located near the decision boundary in feature space, while trivial ones lie far from it. To investigate the hierarchical relationships

TABLE IV
MODEL EXPLAINABILITY METHODS BASED ON PROTOTYPE-BASED LEARNING. THE ABBREVIATIONS HERE ARE CLS: CLASSIFICATION, REG: REGRESSION.

Method	Year	Location	Modality	Task	Dataset	Prototype Type
Kim *et al.* [170]	2020	Chest	X-ray	CLS	[40]	Explicit
Singh *et al.* [173]	2021	Chest	X-ray	CLS	[174]	Explicit
Mohammadjafari *et al.* [175]	2021	Brain	MRI	CLS	[176]	Explicit
Barnett *et al.* [177]	2021	Breast	Mammogram	CLS	Private	Explicit
Carloni *et al.* [178]	2022	Breast	Mammogram	CLS	[179]	Explicit
Wang *et al.* [180]	2022	Breast	Mammogram	CLS	[181]	Explicit
Wei *et al.* [182]	2024	Brain	MRI	CLS	[183]	Explicit
Hesse *et al.* [184]	2022	Eye	Retinal images	REG	[185]	Explicit
Santos *et al.* [186]	2024	Eye	Retinal images	CLS	[187]	Explicit
Hesse *et al.* [188]	2024	Brain	MRI, US	REG	[189], [190]	Explicit

between classes, Hase *et al.* [196] propose hierarchical prototypes to offer explanations according to class taxonomy. As prototype-based models are mostly based on linear classifiers, ProtoKnn [197] explores the usage of k nearest neighbors as a classifier and offers counterfactual explanations within the prototype-based framework. Recognizing the importance of interpretability methods for debugging models, ProtoDebug [198] proposes an approach where a human supervisor can provide feedback to the discovered prototypes and learn confounder-free prototypes.

Adopting prototype-based S-XAI models in the medical domain presents additional challenges. Unlike natural images where the representative prototype occupies an area with a relatively stable size, medical image features such as disease regions in chest X-ray images can vary significantly in size. To address this, XProtoNet [170] proposes to predict an occurrence map and summing the similarity scores within those areas, rather than relying solely on the maximum similarity score as done in ProtopNet. Similarly, [173] introduces prototypes with square and rectangular spatial dimensions for COVID-19 detection in chest X-rays. In evaluations of ProtopNet, Mohammadjafari *et al.* [175] observe a performance drop for Alzheimer's disease detection using MRI, whereas Carlon *et al.* [178] report a high-level of interpretability satisfaction from radiologists in breast mass classification using mammograms. In mammogram based breast cancer diagnosis, Wang *et al.* [180] propose to leverage knowledge distillation to improve model performance. To overcome the confounding issue in mammogram based mass lesion classification, Barnett *et al.* [177] employ a multi-stage framework that identifies the mass margin features for malignancy prediction, skipping image patches that have already been used in previous prototypes during the prototype projection step to improve prototype diversity. In brain tumor classification, MProtoNet [182] introduces a new attention module with soft masking and online-CAM loss applied in 3D multi-parametric MRI. To predict the brain age based on MR and ultrasound images, Hesse *et al.* [188] utilize the weighted mean of prototype labels. Additionally, INSightR-Net [184] formulates the diabetic retinopathy grading as a regression task and apply the prototype based framework, while ProtoAL [186] explores an active learning setting for prototype-based models in diabetic retinopathy.

Although these models offer interpretability in a one-to-one mapping to the input image, they can also make it difficult for users to identify which specific property is important in the corresponding image patch (e.g., is it the color or texture that matters in this prototypical area?). This issue can be partially mitigated using implicit prototype based models.

2) Implicit prototype based models: This type of model follows a similar training scheme as the models based on explicit prototypes, with the major difference in avoiding the prototype replacement step, or only projecting the prototype to the training images' feature patches for visualizations. This scheme is simpler than one that includes prototype replacement step and has different interpretability benefits. Li *et al.* [199] propose the earliest work using latent prototypes, which leverages a decoder to visualize the meanings of the learned prototypes. Protoeval [200] designs a set of loss functions to encourage the learned latent prototypes to be more stable and consistent across different images. To address the issue of the same prototype potentially representing different concepts in the real world, Nauta *et al.* [201] introduce PIP-net which learns prototypes by encouraging the augmented two views of the same image patch to be assigned to the same prototype. To help users identify the specific properties in an image that contribute to the final classification (e.g., color or texture), instead of allowing users to observe only one example image patch per prototype, Ma *et al.* [202] propose to illuminate prototypical concepts via multiple visualizations. Due to the interpretability benefits of decision trees, ProtoTree [203] explores the incorporation of decision trees into prototype-based models, using latent prototypes as the nodes throughout the decision-making process. Recently, to address the concern that prototype-based models often underperform their black box counterparts, Tan *et al.* [204] develop an automatic prototype discovery and refinement strategy to decompose the parameters of the trained classification head and thus guarantees the performance.

Discussion: In terms of performance, implicit prototype based models generally outperform explicit ones, probably due to the greater flexibility in prototype learning. Regarding interpretability, both types of models offer unique advantages. For example, explicit prototypes can be intuitively explained through one-to-one mappings to the input image, while implicit prototypes can be explained using a diverse set of images with similar activations. However, in medical image analysis, current prototype-based S-XAI models primarily utilize explicit prototypes. Therefore, investigating the use of implicit

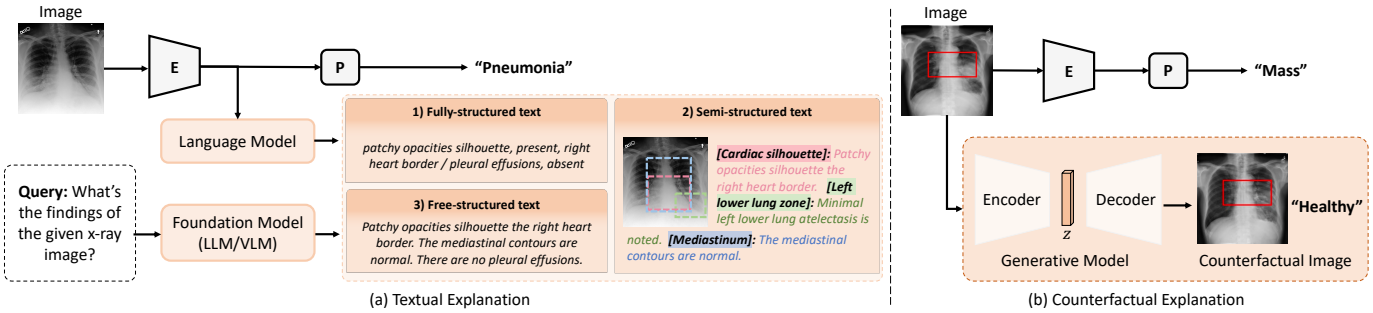

Fig. 8. Output explainability that provides (a) textual explanations, including fully-structured, semi-structured, and free-structured text; and (b) counterfactual explanations. The difference between the generated counterfactual image and raw image (red box) indicates the explanation. X-ray images borrowed from [205].

prototypes in the medical domain could be a promising avenue for future research.

V. Output Explainability

This section discusses output explainability by generating explanations alongside model predictions, including textual (Sec. V-A) and counterfactual (Sec. V-B) explanations.

A. Textual Explanation

Textual explanations in S-XAI involve generating human-readable descriptions that accompany model predictions as part of outputs, similar to image captioning. These methods use natural language to clarify model decisions and typically require textual descriptions for supervision. Some studies explore the integration of textual explanations with visual ones. We categorize these methods into three types based on the structure of textual explanations: 1) fully-structured, 2) semi-structured, and 3) free-structured text, as shown in Fig. 8(a).

1) Fully-structured text generation: To address the challenges posed by complex unstructured medical reports, early efforts transformed target texts into fully structured formats, such as descriptive tags, attributes, or fixed templates, rather than natural language. For example, Pino *et al.* [206] propose CNN-TRG, which detects abnormalities through multilabel classification and generates reports based on pre-defined templates. Some works utilize controlled vocabulary terms (e.g., Medical Subject Headings (MeSH) [207]) to describe image content instead of relying on unstructured reports. Both Shin *et al.* [208] and Gasimova *et al.* [209] employ CNN-RNN frameworks to identify diseases and generate corresponding MeSH sequences, detailing location, severity, and affected organs in chest X-ray images. In addition, Rodin *et al.* [210] present a multitask and multimodal model to produce a short textual summary structured as "[pathology], [present/absent], [(optional) location], [(optional) severity]". However, complete descriptions in natural language are more human-understandable than a set of simple tags, leading several studies to focus on generating reports in a semi-structured format.

2) Semi-structured text generation: Generating semi-structured text involves a partially structured format with predefined topics and constraints in the medical report generation process. For instance, pathology report generation methods [211]–[213] produce reports that focus on describing certain types of cell attributes along with a concluding statement. Additionally, Wang *et al.* [214] introduce a hierarchical framework for medical image explanation, which first predicts semantically related topics and then incorporates these topics as constraints for the language generation model. In the context of hip fracture detection from pelvic X-rays, Gale *et al.* [215] utilize a visual attention mechanism to create terms related to location and characteristics, which are then used to generate sentences structured as: "There is a [degree of displacement], [+/- comminuted][+/- impacted] fracture of the [location] neck of femur [+/- with an avulsed fragment]." More recently, some studies have focused on generating individual sentences based on anatomical regions [216]–[219]. For example, Tanida *et al.* [218] introduce a Region-Guided Radiology Report Generation (RGRG) method that identifies unique anatomical regions in the chest and generates specific descriptions for the most salient areas, ensuring each sentence in the report is linked to a particular anatomical region. Overall, semi-structured approaches effectively balance the rigidity of fully structured reports with the inconsistency of completely free-text reports.

3) Free-structured text generation: With the advancement of language models, reports generated for a given input image are no longer limited to structured formats; instead, they now focus on more open, free-structured text descriptions. These approaches typically involve combining an image encoder to extract visual features with a language model to produce coherent sentences [221]. Several research efforts provide comprehensive explanations that include both textual and visual justifications for diagnostic decisions [222], [224]–[226]. For instance, Spinks and Moens [222] propose a holistic system that delivers diagnosis results along with generated textual captions and a realistic medical image representing the closest alternative diagnosis as visual evidence. Additionally, Wang *et al.* [226] explore a multi-expert Transformer to generate reports and attention-mapping visualization of key medical terms and expert tokens.

In addition to directly generating medical reports, some research studies have incorporated the classification of pathological terms or tags in two distinct ways. The first approach utilizes a "classification-report generation" pipeline, integrating a classifier within the report generation network to enhance feature representations [227], [228]. For example, Yuan *et al.* [227] further employ a sentence-level attention mechanism

TABLE V
OUTPUT EXPLAINABILITY METHODS THAT PROVIDE TEXTUAL EXPLANATIONS. THE ABBREVIATIONS HERE ARE MRG: MEDICAL REPORT GENERATION, CLS: CLASSIFICATION, LOC: LOCATION, SEG: SEGMENTATION, VQA: VISUAL QUESTION ANSWERING, VIS: VISUAL EXPLANATION.

Method	Year	Location	Modality	Task	Dataset	Text Type	Vis.
Pino *et al.* [206]	2021	Chest	X-ray	MRG	[41], [42]	Fully-structured	✓
Shin *et al.* [208]	2016	Chest	X-ray	CLS + MRG	[220]	Fully-structured	-
Gasimova *et al.* [209]	2019	Chest	X-ray	CLS + MRG	[220]	Fully-structured	-
Rodin *et al.* [210]	2019	Chest	X-ray	MRG	[42]	Fully-structured	✓
Zhang *et al.* [211]	2017	Bladder	Pathology	MRG	Private	Semi-structured	✓
Zhang *et al.* [212]	2017	Bladder	Pathology	MRG	Private	Semi-structured	✓
Ma *et al.* [213]	2018	Cervix	Pathology	MRG	Private	Semi-structured	✓
Wang *et al.* [214]	2019	Chest	X-ray	CLS + MRG	[41]	Semi-structured	✓
Gale *et al.* [215]	2019	Pelvic	X-ray	MRG	Private	Semi-structured	✓
Tanida *et al.* [218]	2023	Chest	X-ray	LOC + MRG	[216]	Semi-structured	✓
Wang *et al.* [219]	2022	Chest	X-ray	CLS + MRG	[41], [42]	Semi-structured	-
Singh *et al.* [221]	2019	Chest	X-ray	MRG	[41]	Free-structured	-
Spinks *et al.* [222]	2019	Chest	X-ray	MRG	[41], [223]	Free-structured	✓
Liu *et al.* [224]	2019	Chest	X-ray	MRG	[41], [42]	Free-structured	✓
Chen *et al.* [225]	2020	Chest	X-ray	MRG	[41], [42]	Free-structured	✓
Wang *et al.* [226]	2023	Chest	X-ray	MRG	[41], [42]	Free-structured	✓
Yuan *et al.* [227]	2019	Chest	X-ray	CLS + MRG	[41], [44]	Free-structured	✓
Lee *et al.* [228]	2019	Breast	Mammogram	CLS + MRG	[50]	Free-structured	✓
Zhang *et al.* [229]	2019	Bladder	Pathology	CLS + MRG	[230]	Free-structured	✓
Wang *et al.* [231]	2018	Chest	X-ray	LOC + MRG	[40], [220]	Free-structured	✓
Jing *et al.* [232]	2018	Multiple	X-ray, Pathology	LOC + MRG	[41], [233]	Free-structured	✓
Zeng *et al.* [234]	2020	Multiple	US, X-ray	LOC + MRG	[220]	Free-structured	✓
Tian *et al.* [235]	2018	Abdomen	CT	SEG + MRG	[236]	Free-structured	✓
Thawkar *et al.* [237]	2023	Chest	X-ray	VQA	[42]	Free-structured	-
Zhou *et al.* [238]	2024	Skin	Dermatology	VQA	[131], [239]	Free-structured	-
Moor *et al.* [240]	2023	Multiple	Multiple	VQA	[241]	Free-structured	-
Li *et al.* [242]	2024	Multiple	Multiple	VQA	[38], [39], [243]	Free-structured	-
He *et al.* [244]	2024	Multiple	Multiple	VQA	[39], [42], [243]	Free-structured	-
Chent *et al.* [245]	2024	Multiple	Multiple	VQA	[38], [39], [243], [246]	Free-structured	-
Kang *et al.* [247]	2024	Chest	X-ray	VQA	[41], [42], [248]	Free-structured	-
Chen *et al.* [249]	2024	Chest	X-ray	VQA	[249]	Free-structured	-

alongside a word-level attention model to analyze multi-view chest X-rays, using predicted medical concepts to improve the accuracy of medical reports. Conversely, the second approach employs a "report generation-classification" pipeline, leveraging interpretable region-of-interest (ROI) characterization for final diagnoses. For instance, Zhang *et al.* [229] construct a pathologist-level interpretable diagnostic framework that first detects tumour regions in whole slide images (WSIs), then generates natural language descriptions of microscopic findings with feature-aware visual attention, and finally establishes a diagnostic conclusion. Moreover, integrating region localization and lesion segmentation can enhance the quality of textual explanations [231], [232], [234], [235]. For instance, Wang *et al.* [231] develop a Text-Image Embedding network (TieNet) that incorporates multi-level attention to highlight meaningful text words and X-ray image regions for disease detection and reporting. Leveraging fine-grained annotations of segmentation masks or bounding boxes for lesions, Tian *et al.* [235] combine a segmentation model with a language model, creating a multimodal framework with a semi-supervised attention mechanism for CT report generation.

Compared to traditional report generation approaches, the utilization of LLMs offers a more interactive and comprehensible method for generating textual explanations. Recent medical VLMs applied to various medical images, such as chest X-rays (e.g., XrayGPT [237]), skin images (e.g., SkinGPT [238]), and general medical images (e.g., Med-flamingo [240], LLaMa-Med [242], MedDr [244], HuatuoGPT-Vision [245]), can analyze and respond to open-ended questions about the input images, thanks to their pretraining on extensive datasets of image-report pairs. For instance, XrayGPT [237] demonstrates the alignment of a medical visual encoder (MedClip) with a fine-tuned LLM (Vicuna) using a linear transformation. Given an input image, this combined model can address open-ended questions, such as "What are the main findings and impressions from the given X-ray?". These models not only excel in medical image captioning but also demonstrate exceptional capability in delivering comprehensive explanations for a wide range of medical inquiries. By leveraging their extensive knowledge and understanding, they contribute to the generation of detailed and informative textual explanations within the medical field.

Discussion: Textual explanations have demonstrated significant effectiveness in providing human-interpretable judgments through natural language. This type of S-XAI approach has become especially valuable with the advancement of language models, enabling the generation of lengthy reports and the ability to answer open-ended questions. However, it is crucial to enhance the quality and reliability of these generated textual explanations. Some recent studies utilize techniques such as knowledge decoupling [247] and instruction tuning [249] to address challenges like hallucination, thereby improving the effectiveness and trustworthiness of textual explanations in medical applications.

B. Counterfactual Explanation

Counterfactual explanations describe a causal situation by imagining a hypothetical reality that contradicts the observed facts: *If X had not occurred, Y would not have occurred* [23]. Counterfactual explanations present a contrastive example: for a given original image, it's counterfactual image can alter the model's prediction to a predefined output through the minimal perturbation to observations on the original image, as illustrated in the diagram of Fig. 8(b). Traditional counterfactual explanations are generated through a post-hoc paradigm [205], [250]. Specifically, a classification model is first trained as a black-box model, and then a generative model such as Generative Adversarial Network (GAN) [251] is applied to produce the counterfactual counterpart of the input image.

However, dissociating the model's prediction from its explanation can lead to poor-quality explanations [13]. In particular, counterfactual examples produced in the post-hoc framework are susceptible to issues related to the classifier's robustness and complexity (e.g., overfitting and excessive generalization), resulting in explanations that are inadequate for effective interpretability [252]. To address this, more recent work has explored self-explainable variants of the counterfactual approach. An alternative is to incorporate an explanation generation module directly into the predictor model, such that the model can provide explanations for its own predictions. In general, the predictor and explanation generator are trained jointly, hence the presence of the explanation generator is influencing the training of the predictor. For example, CounterNet [253] integrates the training of the predictive model with the generation of counterfactual explanations into a single, end-to-end process, allowing for joint optimization. Compared to post-hoc approaches, it is able to produce counterfactuals with higher validity. Following a similar pipeline, Guyomard *et al.* [254] propose VCNet, in which the counterfactual generator is based on conditional variational autoencoder (cVAE) whose latent space can be controlled and tweaked to generate more realistic counterfactuals.

In the medical domain, Wilms *et al.* [255] propose an invertible, self-explainable generative model based on efficient normalizing flow for brain age regression and brain sex classification on 3D neuroimaging data. The invertible model can generate predictions during the forward process and produce explanations including voxel-level attribution maps and counterfactual images to clarify its decision-making in the reverse process.

Discussion: S-XAI models that generate counterfactual explanations alongside predictions show greater promise than post-hoc methods. However, their application in medical image analysis remains largely underexplored.

VI. Evaluation

As discussed in the previous sections, various efforts are being made to investigate S-XAI methods in medical image analysis. However, assessing explainability presents significant challenges. In this section, we will outline the desired characteristics and evaluation methods for explainability.

A. Desired characteristics of explainability

It is important for S-XAI models to possess certain desirable qualities when providing explanations. In this regard, Robnik *et al.* [260] enumerated a set of desirable characteristics for high-quality explanations generated by XAI methods. In terms of explanations and explainability methods, Table VI presents expected traits based on our literature review. These characteristics can be used to evaluate and compare different S-XAI approaches.

For medical applications, the characteristics of high-quality explanations should align closely with the real-world necessities of clinical practice. Van *et al.* [19] and Adadi *et al.* [261] summarized the characteristics of XAI methods for medical image analysis and healthcare, respectively. Jin *et al.* [258] proposed clinical XAI guidelines that consist of five criteria for optimizing clinical XAI. These guidelines suggest selecting an explanation form based on understandability and clinical relevance. For the chosen format, the specific XAI technique should be optimized for truthfulness, informative plausibility, and computational efficiency. Usability is another factor that enhances a model's credibility [262]. Individuals are more likely to trust a model that provides insights into how it accomplished its task. In this regard, an interactive and dynamic explanation is preferred over a static one.

B. Evaluation Methods

Doshi-Velez and Kim [263] propose three distinct categories for evaluating XAI methods. 1) Application-grounded evaluations engage experts specific to a field, such as doctors for diagnostic purposes. 2) Human-grounded evaluations involve non-experts assessing the overall quality of explanations. 3) Functionality-grounded evaluations use proxy tasks instead of human input to evaluate explanation quality, which are desirable for interpretability due to constraints related to time and cost. In the medical field, it is crucial to involve end-users, such as junior and senior doctors, in the evaluation process, ideally in contexts that utilize real tasks and data [2].

1) Human-centered evaluation: It is essential to conduct human-centered evaluations in collaboration with medical experts to assess whether end users are satisfied with the explanations provided by S-XAI models. In human-centered evaluations, the quality of the explanations can be assessed using qualitative metrics and quantitative metrics.

Qualitative metrics: Qualitative metrics include evaluating the usefulness, satisfaction, confidence, and trust in provided explanations through interviews or questionnaires [264]. For instance, System Causability Scale [265] measures the quality of interpretability methods applicable in the medical field. Gale *et al.* [215] assess experts' acceptance of explanations by scoring each type on a 10-level Likert scale. Their findings indicate that doctors prefer human-style text explanations over saliency maps and favor a combination of both saliency maps and generated text rather than using either one alone.

Quantitative metrics: Quantitative metrics focus on measuring task performance of human-machine collaboration with factors such as accuracy, response time, likelihood of deviation, ability to detect errors, and even physiological responses

TABLE VI
DESIRABLE QUALITIES OF EXPLANATION METHODS, INDIVIDUAL EXPLANATIONS, AND HUMAN-FRIENDLY EXPLANATIONS.

Type	Qualities	Description	Ref.
Explanations	Faithfulness, Fidelity, Truthfulness	Explanations should truthfully reflect the AI model decision process.	[256]–[258]
	Consistency, Invariance, Robustness	For a fixed model, explanation of similar data points (with similar prediction outputs) should be similar.	[7], [258]
	Understandability, Comprehensibility	Explanations should be easily understandable by clinical users without requiring technical knowledge.	[256], [258]
	Clinical Relevance	Explanation should be relevant to physicians' clinical decision-making pattern, and can support their clinical reasoning process.	[258]
	Plausibility, Factuality, Persuasiveness	Users' judgment on explanation plausibility (i.e., how convincing the explanations are to humans) may inform users about AI decision quality, including potential flaws or biases.	[7], [256], [258]
Explainability Methods	Computational Complexity	The computational complexity of explanation algorithms.	[258], [259]
	Generalizability, Portability	To increase the utility because of the diversity of model architectures.	[259]

[264]. For example, Sayres *et al.* [266] investigate the impact of a deep learning model on doctors' performance in predicting diabetic retinopathy (DR) severity. Ten ophthalmologists with varying levels of experience read images under three conditions: unassisted, predicted grades only, and predicted grades with heatmaps. The results indicate that AI assistance improves diagnostic accuracy, subjective confidence, and time spent. However, in most cases, the combination of grades and heatmaps is only as effective as using grades alone, and actually decreased accuracy for patients without DR.

Overall, human-centered evaluations offer the significant advantage of providing direct and compelling evidence of the effectiveness of explanations [263]. However, these evaluations tend to be costly and time-consuming due to the need to recruit expert participants and obtain necessary approvals, as well as the additional time required for conducting the experiments. Most importantly, these evaluations are inherently subjective.

2) Functionality-grounded evaluation: This category of evaluation, which do not involve human-subject investigations, can be employed to assess the fidelity of explanations. The accuracy of S-XAI methods in generating genuine explanations is referred to as the fidelity of an explainer. In this section, we will present a variety of functionality-grounded evaluation methods for different types of explanations.

Attention-based explanation: In the absence of references, attention-based explanations can be assessed through a causal framework. For example, Petsiuk *et al.* [267] introduce two causal metrics, i.e., deletion and insertion. Following this, Hooker *et al.* [268] propose RemOve And Retrain (ROAR), a method that evaluates how the accuracy of a retrained model decreases when essential features in specific regions are removed. With the manually annotated ground truth data, such as object bounding boxes or semantic masks, the accuracy of attention-based explanations can be evaluated by comparing with these references. Yan *et al.* [83] and Hou *et al.* [269] calculate the Jaccard index value and the AUC score to measure the effectiveness of attention maps, respectively. Additionally, Barnett *et al.* [270] introduce the Activation Precision metric, which quantifies the proportion of relevant information from the relevant region used to classify the mass margin based on radiologist annotations. Furthermore, human expert eye fixation is an emerging data modality that can provide key diagnostic features by tracking the gaze patterns and visual attention of clinicians, which is also utilized as the ground truth of attention maps [271].

Concept-based explanation: To evaluate concept-based explanations, researchers mainly focus on metrics such as Concept Error [113], [163], T-CAV score [117], Completeness Score [164], and Concept Relevance [4], [120]. Additionally, other evaluation methods exist. For example, Zarlenga *et al.* [159] propose Concept Alignment Score (CAS) and Mutual Information to evaluate concept-based explainability. Wang *et al.* [195] adopt Concept Purity to assess the model's capability to discover concepts that only cover a single shape.

Example-based explanation: In the evaluation of example-based explanations, Nguyen and Martinez [272] establish two quantitative metrics: 1) non-representativeness, which evaluates how well the examples represent the explanations, thereby measuring the fidelity of the explanation; and 2) diversity, which gauges the degree of integration within the explanation. Additionally, Huang et al. [200] developed two metrics: 1) a consistency score to determine whether the prototype consistently highlights the same semantically meaningful areas across different images, and 2) a stability score to assess whether it reliably identifies the same area after the image is exposed to noise.

Textual explanation: The common assessment of textual explanations involves using metrics such as BLEU [273], ROUGE-L [274], and CIDEr [275] to compare generated natural language descriptions against ground truth reference sentences provided by experts. Patricio *et al.* [17] conduct a benchmark study of interpretable medical imaging approaches, specifically evaluating the quality of textual explanations for chest X-ray images.

Counterfactual explanation: Singla *et al.* [276] employ three metrics to evaluate counterfactual explanations for chest X-ray classification: 1) Frechet Inception Distance (FID) to assess visual quality, 2) Counterfactual Validity (CV) to determine if the counterfactual aligns with classifier's predictions, and 3) Foreign Object Preservation (FOP), which examines whether patient-specific information is retained. Additionally, they use clinical metrics, including the cardiothoracic ratio and a score

for detecting normal costophrenic recess, to illustrate the clinical utility of the explanations.

VII. Challenges and Future Directions

Despite the rapid advancements in S-XAI for medical image analysis, several significant challenges remain unresolved. In this section, we will analyze the existing challenges and discuss potential future directions to enhance the effectiveness and reliability of S-XAI in the medical domain.

A. S-XAI Benchmark Construction

Establishing benchmarks for S-XAI in medical image analysis. is essential. These benchmarks will standardize evaluations, enable fair comparisons between different methods, and ultimately enhance the reliability of medical AI applications.

1) Dataset construction: One of the main challenges in collecting medical data is the limited availability of doctors to annotate large datasets. This challenge is even more significant in S-XAI, where additional fine-grained annotations, such as concepts and textual descriptions, are necessary. As a result, medical datasets that meet interpretability standards often have a limited volume of data, reducing the generalizability and applicability of S-XAI methods in real-world contexts.

2) Evaluation metrics: Automated evaluation of explanations generated by S-XAI methods poses another significant challenge. In the medical field, human-centered evaluations often rely on the expertise of clinicians. However, the variability in expert opinions can lead to biased and subjective assessments [277]. Meanwhile, existing functionality-grounded evaluations still depend on manual annotations. Thus, developing objective metrics to evaluate the quality of model explanations is likely to become an important research focus.

To tackle these challenges, future directions include leveraging semi-automated annotation tools to assist clinicians in the annotation process, thereby easing their workload. Additionally, developing objective metrics and standardized protocols to assess the quality of model explanations will be a critical research trend in S-XAI.

B. S-XAI in the Era of Foundation Models

Foundation models, including large language models (LLMs) and vision-language models (VLMs), have transformed the AI landscape, finding applications across diverse fields such as natural language processing, computer vision, and multimodal understanding. Notably, medical LLMs [278]–[280] and medical VLMs [237], [240], [281] are designed to encode rich domain-specific knowledge. The intersection of medical foundation models and S-XAI presents significant opportunities for the future of medical AI systems [262].

1) S-XAI benefits foundation models: Foundation models are typically large models with an extremely huge number of parameters, trained on vast datasets. The complexity of these models make it challenging to explore their decision-making processes, which may result in potential biases and a lack of transparency. Apart from leveraging post-hoc XAI techniques (e.g., attribution maps [282], [283]) to interpret the decision-making processes of foundation models, S-XAI methods can enhance the input explainability through explainable prompts [284] and knowledge-enhanced prompts [285], ultimately improving model performance.

2) Foundation models advance S-XAI: Foundation models learn generally useful representations from the clinical knowledge embedded in medical corpora [286]. By harnessing the sophisticated capabilities of foundation models, S-XAI methods can produce user-friendly explanations [287] and support more flexible generative concept-based learning [288], [289]. Moreover, foundation models can facilitate the evaluation of S-XAI methods that emulate human cognitive processes [290].

C. S-XAI with Human-in-the-Loop

Integrating Human-in-the-Loop (HITL) processes is crucial for effectively implementing S-XAI in the medical field. This approach not only enhances the overall performance of AI systems but also fosters trust among medical experts.

1) Enhancing prediction accuracy through human intervention: A HITL framework allows for the identification and removal of potential confounding factors, such as artifacts or biases in datasets, during the training phase. For instance, clinicians can adjust the outputs of predicted concepts, leading to a more accurate concept bottleneck model [129]. This collaborative approach can significantly enhance the model's accuracy by incorporating expert insights.

2) Improving explainability through human feedback: To ensure continuous improvement, a versioning or feedback evaluation system should be established, enabling the final system to build trust during hospital evaluations. Achieving this requires fostering collaboration between S-XAI researchers and clinical practitioners, ensuring that feedback is systematically gathered and used to refine the models.

However, one challenge in integrating HITL processes is the variability in clinician expertise and availability, which can affect the consistency and quality of human feedback. Ensuring that human knowledge is effectively integrated into the AI training process without introducing additional biases or errors is a complex task.

D. Trade-off between Performance and Interpretability

It is widely believed that as model complexity increases to enhance performance, the model's interpretability tends to decline [291], [292]. Conversely, more interpretable models may sacrifice some predictive accuracy. However, it is important to note that some researches contend that *there is no scientific evidence for a general trade-off between accuracy and interpretability* [293]. In fact, recent advancements in concept-based models [132], [139], [142] have demonstrated performance on par with black-box models in medical image applications. This achievement depends on the researcher's ability to identify patterns in an interpretable manner while maintaining the flexibility to accurately fit the data [13]. Future S-XAI methods are expected to aim for an optimization of both performance and interpretability, potentially providing theoretical foundations for this balance.

E. Other Explainability of S-XAI

1) Multi-modal explainability: Embracing multi-modal explainability is a promising direction for S-XAI in the medical field. Medical data often exists in various forms, including images, text, and omics. By integrating these modalities, multi-modal S-XAI approaches can offer more comprehensive and intuitive explanations that align with clinicians processes. Additionally, S-XAI methods can identify correlations during data fusion [294], offering significant potential for discovering new biomarkers. For example, exploring correlations between radiological and pathological images could help discover non-invasive biomarkers as alternatives for tumor diagnosis.

2) Causal explainability: Another direction for S-XAI involves causality, which defines the cause-and-effect relationship and can be mathematically modeled [295]. Traditional deep learning methods in medical imaging often confuse correlation with causation, leading to potentially harmful errors. For instance, DeGrave *et al.* [296] use XAI techniques to audit COVID-19 diagnosis methods from chest X-rays find that these methods are primarily identifying spurious correlations. To address dataset biases, Castro *et al.* [297] emphasize the role of causal reasoning in detecting biases. Luo *et al.* [298] develop debiased models based on biased training data generated from causal assumptions for diagnosing chest X-rays. Incorporating such analyses into the explainability of medical images analysis could be highly beneficial.

VIII. Conclusion

This survey reviews recent advancements in self-explainable artificial intelligence (S-XAI) for medical image analysis. Contrary to previous surveys that primarily focus on post-hoc XAI techniques, this paper emphasizes inherently interpretable S-XAI models, which are gaining traction in research. This survey introduce S-XAI from three key perspectives, i.e., input explainability, model explainability, and output explainability. Additionally, this survey explore the desired characteristics of explainability and various evaluation methods for assessing explanation quality. While significant progress has been made, it also highlights key challenges that need to be tackled and provides insights for future research on trustworthy AI systems in clinical practice. Overall, this survey serves as a valuable reference for the XAI community, particularly within the medical imaging field, and lays the groundwork for future advancements that will improve the transparency and trustworthiness of AI tools in healthcare.

Appendix

A. Public datasets used in S-XAI

We provide an overview of more than 70 datasets currently available for S-XAI in the medical image domain. Table VII presents the key characteristics of these datasets, including modality, scale, and task. In this section, we will introduce the relevant datasets categorized by image modalities, highlighting their contributions to the development of S-XAI. For more detailed information about these datasets, we direct readers to the relevant publications and sources.

1) Radiology: Radiological images generally include modalities such as X-ray, MRI, CT, mammography, and ultrasound. Among these, X-ray is one of the most commonly used modalities in S-XAI for medical image analysis. For instance, the SLAKE [38] dataset collects knowledge triplets from the open source knowledge graph to assist the model in achieving a better understanding of X-ray images. Medical-CXR-VQA [63] focuses on the five types of questions (i.e., *abnormality, presence, view, location* and *type*) and aligns them with the key information of the X-ray image, resulting more reliable answers. VQA-RAD [243] is manually constructed based on clinicians asking naturally occurring questions about radiology images and providing reference answers, resulting in a dataset rich in quality expertise and knowledge to capture the details of radiology images. OAI [124] provides knee X-rays for knee osteoarthritis grading and offers clinical concepts (e.g., joint space narrowing, bone spurs, calcification), making it suitable for concept-based learning. Moreover, datasets such as IU X-ray [41], MIMIC-CXR [42], OpenI [220], and Chest ImaGenome [299] provide a large number of chest X-rays along with corresponding free-text reports, which can facilitate the generation of textual explanations. Finally, MIMIC-CXR-VQA [248], CheXbench [249], SLAKE [38], and MIMIC-Diff-VQA [62] are constructed for the VQA task, providing support for interactive explanations of S-XAI models.

Regarding MRI, SUN09 [103] and AC17 [104] provide cardiac segmentation masks, while BraTS 2020 [107] offers brain masks. However, most MRI datasets have a limited number of samples, which may affect the generalization ability of models. For CT images, datasets like CT-150 [100], NIH-TCIA CT-82 [101], and LiTS [236] can be used for segmentation tasks, while LIDC-IDRI [146] provides lung cancer annotations with grade of eight attributes, benefiting concept-based learning. Additionally, disease diagnoses on mammography datasets [50], [179], [300] and ultrasound datasets [158], [190] also serve as applications for S-XAI methods.

2) Dermatology: In the scope of dermatology, datasets with fine-grained concept annotations are commonly used in concept-based learning for S-XAI. Derm7pt [127] is a dermoscopic image dataset containing 1,011 images with clinical concepts for melanoma skin lesions according to the 7-point checklist criteria [301]. PH^2 dataset [128] includes 200 dermoscopic images of melanocytic lesions with segmentation masks and several clinical concepts. SkinCon [131] is a skin disease dataset containing 3,230 images with 48 clinical concepts densely annotated by dermatologists for fine-grained model debugging and analysis, where the images are selected from the Fitzpatrick 17k [133] and DDI [134] skin image datasets. Other datasets such as ISIC [54], [85], [86] and HAM10000 [106] are also broadly used datasets but without explicit fine-grained concept annotations. Dermoscopic image datasets significantly facilitate the development of S-XAI, however, annotating concept labels requires the efforts of human experts and is labor-intensive. Hence there are currently only a few datasets with a limited number of samples that have fine-grained concept labels. This poses a significant challenge

for concept-based S-XAI, especially in supervised concept learning.

3) Pathology: With regard to pathological image datasets, PathVQA [39] is the first dataset focused on pathology VQA, featuring over 32K open-ended questions derived from 4,998 pathology images. PEIR Gross [233] contains 7,442 image-caption pairs across 21 different sub-categories, with each caption consisting of a single sentence. The Cancer Genome Atlas (TCGA) [230] provides multimodal data for over 20,000 tumor and normal samples, offers multimodal data for more than 20K tumor and normal samples, encompassing clinical data, DNA, and various imaging types (diagnostic images, tissue images, and radiological images). Additionally, datasets such as BACH [81], Biopsy4Grading [88], and NCT [157] are available for classifying breast cancer, non-alcoholic fatty liver disease, and colorectal cancer, respectively. WBCAtt [143] provides 113K microscopic images of white blood cells, annotated with 11 morphological attributes categorized into four main groups: overall cell, nucleus, cytoplasm, and granules. Since pathological images are the "gold standard" for cancer diagnosis, the development of S-XAI models in pathology is highly significant.

4) Retinal images: Regarding retinal disease classification, EyePACS [185] is a large-scale dataset for grading diabetic retinopathy, containing more than 88K images. ACRIMA [82] offers 705 images for glaucoma assessment. In addition to classification labels, FGADR [150], DDR [151], and IDRID [155] also provide fine-grained masks for segmenting various types of lesions.

5) Others: Other medical datasets utilized in S-XAI include the Hyperkvasir endoscopy dataset [55] and the Infectious Keratitis slit lamp microscopy dataset [148]. Additionally, recent datasets like PMC-OA [241] and PMC-VQA [246] collect data from open-source medical literature corpus. These datasets encompass a wealth of multimodal data, which significantly facilitate the development of medical foundation models.

TABLE VII: Public datasets used in the reviewed S-XAI methods.

Dataset	Modality	Scale	Task
SLAKE [38]	X-ray	642 images, 14,028 QA pairs	VQA, SEG, DET
IU X-ray [41]	X-ray	8,121 images, 3,996 texts	MRG
MIMIC-Diff-VQA [62]	X-ray	700K QA pairs, 164K images	VQA
OAI [124]	X-ray	26,626,000 images	CLS
CheXbench [249]	X-ray	6.1M QA pairs	VQA
ChestXray [40]	X-ray	108,948 images	CLS, MRG, DET, LOC
CheXpert [44]	X-ray	224K images	CLS
MIMIC-CXR [42]	X-ray	377K images, 227K texts	CLS, MRG
BCDR-F03 [79]	X-ray	736 images	CLS
RSNA [89]	X-ray	30,000 images	CLS, DET
ZhangLabData [90]	OCT, X-ray	108,312 OCT images, 5,232 X-ray images	CLS
SIIM-FISABIO-RSNA [91]	X-ray	10,178 images	CLS, DET
Public Radiography [92]	X-ray	3,487 images	CLS
COVQU [93]	X-ray	18,479 images	CLS, SEG
COVID-19-NY-SBU [94]	X-ray, CT, MRI	1,384 cases	CLS
MIDRC-RICORD-1C [95]	X-ray	361 cases	CLS
NIH [97]	X-ray	108,948 images	CLS
VinBigData [98]	X-ray	18,000 images	CLS, DET
Montgomery [154]	X-ray	138 images	CLS
COVID-19 [174]	X-ray	761 images	CLS
OpenI [220]	X-ray	7,470 images, 3,955 texts	CLS, MRG
Chest ImaGenome [216]	X-ray	242K images, 217K texts	CLS, MRG, DET
MIMIC-CXR-VQA [248]	X-ray	377K images with QA pairs	VQA
ADNI-1 [99]	MRI	818 subjects	CLS, REG
ADNI-2 [112]	MRI	599 subjects	CLS, REG
SUN09 [103]	MRI	395 slices	SEG
AC17 [104]	MRI	200 volumes	SEG
BraTS 2020 [107]	MRI	494 cases	SEG, CLS
Calgary Campinas MRI [110]	MRI	359 subjects	SEG
OASIS [176]	MRI	416 cases	CLS
BraTS 2014 [183]	MRI	65 scans	SEG
IXI [189]	MRI	600 cases	REG
RICODR [96]	CT, X-ray	240 CT scans, 1,000 X-ray images	CLS
CT-150 [100]	CT	150 scans	SEG
Pancreas-CT [101]	CT	82 scans	SEG
CHAOS [108]	CT, MRI	40 CT scans, 120 MRI scans	SEG
LIDC-IDRI [146]	CT	1,018 scans	CLS, SEG
LiTS [236]	CT	201 scans	SEG
VQA-RAD [243]	CT, MRI, X-ray	3,515 QA pairs, 315 images	VQA
DDSM [50]	Mammogram	2,620 cases	CLS, MRG, SEG
CBIS-DDSM [179]	Mammogram	1,644 cases	CLS, SEG
CMMD [181]	Mammogram	1,775 cases	CLS
BUSI [158]	Ultrasound	780 images	CLS
FGLS [190]	Ultrasound	4,290 volumes	REG
HAM10000 [106]	Dermatology	10,015 images	CLS
ISIC 2016 [84]	Dermatology	1,279 images	CLS, SEG
ISIC 2017 [85]	Dermatology	2,750 images	CLS, SEG
ISIC 2018 [86]	Dermatology	15,121 images	CLS, SEG
ISIC 2019 [54]	Dermatology	33,569 images	CLS
Derm7pt [127]	Dermatology	1,011 images	CLS
PH^2 [128]	Dermatology	200 images	CLS, SEG
SkinCon [131]	Dermatology	3,230 images	CLS
Fitzpatrick 17k [133]	Dermatology	16,577 images	CLS
DDI [134]	Dermatology	656 images	CLS
DermNetNZ [137]	Dermatology	25K images	CLS
SD-260 [138]	Dermatology	6,584 images	CLS
Dermnet [239]	Dermatology	18,856 images	CLS, VQA
PathVQA [39]	Pathology	32K QA pairs, 4,998 images	VQA
BACH [81]	Histopathology	500 images	CLS
Biopsy4Grading [88]	Histopathology	351 images	CLS
WBCAtt [143]	Microscopy	113,278 images	CLS
NCT [157]	Histopathology	100K images	CLS
TCGA [230]	Pathology	20K studies	CLS, MRG
PEIR Gross [233]	Pathology	7,442 image-text pairs	MRG
ACRIMA [82]	Retinal images	705 images	CLS
FGADR [150]	Retinal images	2,842 images	CLS, SEG
DDR [151]	Retinal images	13,673 images	CLS, SEG, DET
IDRID [155]	Retinal images	516 images	CLS, SEG, LOC

Continued on next page

– continued from previous page

Dataset	Modality	Scale	Task
EyePACS [185]	Retinal images	88,702 images	CLS
Messidor [187]	Retinal images	1,200 images	CLS
Hyperkvasir [55]	Endoscopy	110,079 images, 374 videos	SEG, DET
Infectious Keratitis [148]	Slit lamp microscopy	115,408 images	CLS
PMC-OA [241]	Multiple	1.65M image-text pairs	VQA
PMC-VQA [246]	Multiple	227K QA pairs, 149K images	VQA

REFERENCES

[1] X. Jia, L. Ren, and J. Cai, "Clinical implementation of ai technologies will require interpretable ai models," *Medical physics*, no. 1, pp. 1–4, 2020.

[2] Z. Salahuddin *et al.*, "Transparency of deep neural networks for medical image analysis: A review of interpretability methods," *Computers in biology and medicine*, vol. 140, p. 105111, 2022.

[3] L. Luo *et al.*, "Rethinking annotation granularity for overcoming shortcuts in deep learning–based radiograph diagnosis: A multicenter study," *Radiology: Artificial Intelligence*, vol. 4, no. 5, p. e210299, 2022.

[4] D. Alvarez Melis and T. Jaakkola, "Towards robust interpretability with self-explaining neural networks," *Advances in neural information processing systems*, vol. 31, 2018.

[5] S. Bach *et al.*, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PloS one*, vol. 10, no. 7, p. e0130140, 2015.

[6] B. Zhou *et al.*, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929.

[7] M. T. Ribeiro, S. Singh, and C. Guestrin, "" 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*, 2016, pp. 1135–1144.

[8] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[9] B. Kim *et al.*, "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)," in *International conference on machine learning*. PMLR, 2018, pp. 2668–2677.

[10] J. Crabbé and M. van der Schaar, "Concept activation regions: A generalized framework for concept-based explanations," *Advances in Neural Information Processing Systems*, vol. 35, pp. 2590–2607, 2022.

[11] B. Mittelstadt, C. Russell, and S. Wachter, "Explaining explanations in ai," in *Proceedings of the conference on fairness, accountability, and transparency*, 2019, pp. 279–288.

[12] J. Zhang *et al.*, "Overlooked trustworthiness of saliency maps," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2022, pp. 451–461.

[13] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature machine intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

[14] J. R. Quinlan, "Induction of decision trees," *Machine learning*, vol. 1, pp. 81–106, 1986.

[15] T. J. Hastie, "Generalized additive models," in *Statistical models in S*. Routledge, 2017, pp. 249–307.

[16] C. Grosan and A. Abraham, *Intelligent systems*. Springer, vol. 17.

[17] C. Patrício, J. C. Neves, and L. F. Teixeira, "Explainable deep learning methods in medical image classification: A survey," *ACM Computing Surveys*, vol. 56, no. 4, pp. 1–41, 2023.

[18] G. Yang, Q. Ye, and J. Xia, "Unbox the black-box for the medical explainable ai via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond," *Information Fusion*, vol. 77, pp. 29–52, 2022.

[19] B. H. Van der Velden *et al.*, "Explainable artificial intelligence (xai) in deep learning-based medical image analysis," *Medical Image Analysis*, vol. 79, p. 102470, 2022.

[20] R.-K. Sheu and M. S. Pardeshi, "A survey on medical explainable ai (xai): recent progress, explainability approach, human interaction and scoring system," *Sensors*, vol. 22, no. 20, p. 8068, 2022.

[21] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (xai): Toward medical xai," *IEEE transactions on neural networks and learning systems*, vol. 32, no. 11, pp. 4793–4813, 2020.

[22] A. Singh, S. Sengupta, and V. Lakshminarayanan, "Explainable deep learning models in medical image analysis," *Journal of imaging*, vol. 6, no. 6, p. 52, 2020.

[23] M. Christoph, *Interpretable machine learning: A guide for making black box models explainable*. Leanpub, 2020.

[24] S. Kapse *et al.*, "Si-mil: Taming deep mil for self-interpretability in gigapixel histopathology," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 11 226–11 237.

[25] U. Sajid *et al.*, "Breast cancer classification using deep learned features boosted with handcrafted features," *Biomedical Signal Processing and Control*, vol. 86, p. 105353, 2023.

[26] H. Xiang *et al.*, "Development and validation of an interpretable model integrating multimodal information for improving ovarian cancer diagnosis," *Nature Communications*, vol. 15, no. 1, p. 2681, 2024.

[27] N. Lassau *et al.*, "Integrating deep learning ct-scan model, biological and clinical variables to predict severity of covid-19 patients," *Nature communications*, vol. 12, no. 1, pp. 1–11, 2021.

[28] S. Ji *et al.*, "A survey on knowledge graphs: Representation, acquisition, and applications," *IEEE transactions on neural networks and learning systems*, vol. 33, no. 2, pp. 494–514, 2021.

[29] S. Pan *et al.*, "Unifying large language models and knowledge graphs: A roadmap," *IEEE Transactions on Knowledge and Data Engineering*, 2024.

[30] C. Peng *et al.*, "Knowledge graphs: Opportunities and challenges," *Artificial Intelligence Review*, vol. 56, no. 11, pp. 13 071–13 102, 2023.

[31] X. Xie *et al.*, "A survey on incorporating domain knowledge into deep learning for medical image analysis," *Medical Image Analysis*, vol. 69, p. 101985, 2021.

[32] S. Liu and H. Chen, "Knowledge injected multimodal irregular ehrs model for medical prediction," in *International Workshop on Trustworthy Artificial Intelligence for Healthcare*. Springer, 2024, pp. 25–39.

[33] S. Liu *et al.*, "Shape: A sample-adaptive hierarchical prediction network for medication recommendation," *IEEE Journal of Biomedical and Health Informatics*, 2023.

[34] U. Naseem *et al.*, "K-pathvqa: Knowledge-aware multimodal representation for pathology visual question answering," *IEEE Journal of Biomedical and Health Informatics*, 2023.

[35] Y. Zhang *et al.*, "When radiology report generation meets knowledge graph," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 07, 2020, pp. 12 910–12 917.

[36] F. Liu *et al.*, "Exploring and distilling posterior and prior knowledge for radiology report generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 13 753–13 762.

[37] Z. Huang, X. Zhang, and S. Zhang, "Kiut: Knowledge-injected u-transformer for radiology report generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19 809–19 818.

[38] B. Liu *et al.*, "Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering," in *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2021, pp. 1650–1654.

[39] X. He *et al.*, "Pathvqa: 30000+ questions for medical visual question answering," *arXiv preprint arXiv:2003.10286*, 2020.

[40] X. Wang *et al.*, "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2097–2106.

[41] D. Demner-Fushman *et al.*, "Preparing a collection of radiology examinations for distribution and retrieval," *Journal of the American Medical Informatics Association*, vol. 23, no. 2, pp. 304–310, 2016.

[42] A. E. Johnson *et al.*, "Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports," *Scientific data*, vol. 6, no. 1, p. 317, 2019.

[43] B. Chen *et al.*, "Label co-occurrence learning with graph convolutional networks for multi-label chest x-ray image classification," *IEEE journal of biomedical and health informatics*, vol. 24, no. 8, pp. 2292–2302, 2020.

[44] J. Irvin *et al.*, "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 590–597.

[45] W. Zheng *et al.*, "Pay attention to doctor–patient dialogues: Multi-modal knowledge graph attention image-text embedding for covid-19 diagnosis," *Information Fusion*, vol. 75, pp. 168–185, 2021.

[46] D. Hou, Z. Zhao, and S. Hu, "Multi-label learning with visual-semantic embedded knowledge graph for diagnosis of radiology imaging," *IEEE Access*, vol. 9, pp. 15 720–15 730, 2021.

[47] Y. Zhou *et al.*, "Contrast-attentive thoracic disease recognition with dual-weighting graph reasoning," *IEEE Transactions on Medical Imaging*, vol. 40, no. 4, pp. 1196–1206, 2021.

[48] C. Wu *et al.*, "Medklip: Medical knowledge enhanced language-image pre-training for x-ray diagnosis," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 21 372–21 383.

[49] Y. Liu *et al.*, "Act like a radiologist: towards reliable multi-view correspondence reasoning for mammogram mass detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, pp. 5947–5961, 2021.

[50] M. Heath *et al.*, "Current status of the digital database for screening mammography," in *Digital Mammography: Nijmegen, 1998*. Springer, 1998, pp. 457–460.

[51] G. Zhao, "Cross chest graph for disease diagnosis with structural relational reasoning," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 612–620.

[52] B. Qi *et al.*, "Gren: graph-regularized embedding network for weakly-supervised disease localization in x-ray images," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 10, pp. 5142–5153, 2022.

[53] M. Elbatel, R. Martí, and X. Li, "Fopro-kd: fourier prompted effective knowledge distillation for long-tailed medical image recognition," *IEEE Transactions on Medical Imaging*, 2023.

[54] M. Combalia *et al.*, "Bcn20000: Dermoscopic lesions in the wild," *arXiv preprint arXiv:1908.02288*, 2019.

[55] H. Borgli *et al.*, "Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy," *Scientific data*, vol. 7, no. 1, p. 283, 2020.

[56] C. Y. Li *et al.*, "Knowledge-driven encode, retrieve, paraphrase for medical image report generation," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 6666–6673.

[57] F. Liu *et al.*, "Auto-encoding knowledge graph for unsupervised medical report generation," *Advances in Neural Information Processing Systems*, vol. 34, pp. 16 266–16 279, 2021.

[58] M. Li *et al.*, "Dynamic graph enhanced contrastive learning for chest x-ray report generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 3334–3343.

[59] K. Kale *et al.*, "Kgvl-bart: Knowledge graph augmented visual language bart for radiology report generation," in *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 2023, pp. 3401–3411.

[60] H. Guo *et al.*, "Medical visual question answering via targeted choice contrast and multimodal entity matching," in *International Conference on Neural Information Processing*. Springer, 2022, pp. 343–354.

[61] J. Huang *et al.*, "Medical knowledge-based network for patient-oriented visual question answering," *Information Processing & Management*, vol. 60, no. 2, p. 103241, 2023.

[62] X. Hu *et al.*, "Expert knowledge-aware image difference graph representation learning for difference-aware medical visual question answering," in *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2023, pp. 4156–4165.

[63] X. Hu *et al.*, "Interpretable medical image visual question answering via multi-modal relationship graph learning," *Medical Image Analysis*, vol. 97, p. 103279, 2024.

[64] J. Li *et al.*, "Align before fuse: Vision and language representation learning with momentum distillation," *Advances in neural information processing systems*, vol. 34, pp. 9694–9705, 2021.

[65] S. Liu *et al.*, "A hybrid method of recurrent neural network and graph neural network for next-period prescription prediction," *International Journal of Machine Learning and Cybernetics*, vol. 11, pp. 2849–2856, 2020.

[66] S. Liu *et al.*, "Multimodal data matters: language model pre-training over structured and unstructured electronic health records," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 1, pp. 504–514, 2022.

[67] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[68] Z. Niu, G. Zhong, and H. Yu, "A review on the attention mechanism of deep learning," *Neurocomputing*, vol. 452, pp. 48–62, 2021.

[69] M. Bhattacharya, S. Jain, and P. Prasanna, "Radiotransformer: a cascaded global-focal transformer for visual attention–guided disease classification," in *European Conference on Computer Vision*. Springer, 2022, pp. 679–698.

[70] S. Jetley *et al.*, "Learn to pay attention," *arXiv preprint arXiv:1804.02391*, 2018.

[71] H. Fukui *et al.*, "Attention branch network: Learning of attention mechanism for visual explanation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 10 705–10 714.

[72] L. Li *et al.*, "Scouter: Slot attention-based classifier for explainable image recognition," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 1046–1055.

[73] J. Schlemper *et al.*, "Attention gated networks: Learning to leverage salient regions in medical images," *Medical image analysis*, vol. 53, pp. 197–207, 2019.

[74] R. Gu *et al.*, "Ca-net: Comprehensive attention convolutional neural networks for explainable medical image segmentation," *IEEE transactions on medical imaging*, vol. 40, no. 2, pp. 699–711, 2020.

[75] C. Barata, M. E. Celebi, and J. S. Marques, "Explainable skin lesion diagnosis using taxonomies," *Pattern Recognition*, vol. 110, p. 107413, 2021.

[76] G. Lozupone *et al.*, "Axial: Attention-based explainability for interpretable alzheimer's localized diagnosis using 2d cnns on 3d mri brain scans," *arXiv preprint arXiv:2407.02418*, 2024.

[77] J. Huang *et al.*, "Swin deformable attention u-net transformer (sdaut) for explainable fast mri," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2022, pp. 538–548.

[78] H. Wang *et al.*, "Breast mass classification via deeply integrating the contextual information from multi-view data," *Pattern Recognition*, vol. 80, pp. 42–52, 2018.

[79] J. Arevalo *et al.*, "Representation learning for mammography mass lesion classification with convolutional neural networks," *Computer Methods and Programs in Biomedicine*, vol. 127, pp. 248–257, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169260715300110

[80] H. Yang *et al.*, "Guided soft attention network for classification of breast cancer histopathology images," *IEEE transactions on medical imaging*, vol. 39, no. 5, pp. 1306–1315, 2019.

[81] G. Aresta *et al.*, "Bach: Grand challenge on breast cancer histology images," *Medical image analysis*, vol. 56, pp. 122–139, 2019.

[82] A. Diaz-Pinto *et al.*, "Cnns for automatic glaucoma assessment using fundus images: an extensive validation," *Biomedical engineering online*, vol. 18, pp. 1–19, 2019.

[83] Y. Yan, J. Kawahara, and G. Hamarneh, "Melanoma recognition via visual attention," in *Information Processing in Medical Imaging: 26th International Conference, IPMI 2019, Hong Kong, China, June 2–7, 2019, Proceedings 26*. Springer, 2019, pp. 793–804.

[84] D. Gutman *et al.*, "Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (isbi) 2016, hosted by the international skin imaging collaboration (isic)," *arXiv preprint arXiv:1605.01397*, 2016.

[85] N. C. Codella *et al.*, "Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic)," in *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*. IEEE, 2018, pp. 168–172.

[86] N. Codella *et al.*, "Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic)," *arXiv preprint arXiv:1902.03368*, 2019.

[87] C. Yin *et al.*, "Focusing on clinically interpretable features: selective attention regularization for liver biopsy image classification," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24*. Springer, 2021, pp. 153–162.

[88] F. Heinemann, G. Birk, and B. Stierstorfer, "Deep learning enables pathologist-like scoring of nash models," *Scientific reports*, vol. 9, no. 1, p. 18454, 2019.

[89] G. Shih *et al.*, "Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia," *Radiology: Artificial Intelligence*, vol. 1, no. 1, p. e180041, 2019.

[90] D. S. Kermany *et al.*, "Identifying medical diagnoses and treatable diseases by image-based deep learning," *cell*, vol. 172, no. 5, pp. 1122–1131, 2018.

[91] P. Lakhani *et al.*, "The 2021 siim-fisabio-rsna machine learning covid-19 challenge: Annotation and standard exam classification of covid-19 chest radiographs," *Journal of Digital Imaging*, vol. 36, no. 1, pp. 365–372, 2023.

[92] M. E. Chowdhury *et al.*, "Can ai help in screening viral and covid-19 pneumonia?" *Ieee Access*, vol. 8, pp. 132 665–132 676, 2020.

[93] T. Rahman *et al.*, "Exploring the effect of image enhancement techniques on covid-19 detection using chest x-ray images," *Computers in biology and medicine*, vol. 132, p. 104319, 2021.

[94] J. Saltz *et al.*, "Stony brook university covid-19 positive cases," *the cancer imaging archive*, vol. 4, 2021.

[95] E. Tsai *et al.*, "Data from medical imaging data resource center (midrc)-rsna international covid radiology database (ricord) release 1c-chest x-ray, covid+(midrc-ricord-1c)," *The Cancer Imaging Archive*, vol. 10, 2021.

[96] E. B. Tsai *et al.*, "The rsna international covid-19 open radiology database (ricord)," *Radiology*, vol. 299, no. 1, pp. E204–E213, 2021.

[97] X. Wang *et al.*, "Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *IEEE CVPR*, vol. 7. sn, 2017, p. 46.

[98] H. Q. Nguyen *et al.*, "Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations," *Scientific Data*, vol. 9, no. 1, p. 429, 2022.

[99] B. T. Wyman *et al.*, "Standardization of analysis sets for reporting results from adni mri data," *Alzheimer's & Dementia*, vol. 9, no. 3, pp. 332–337, 2013.

[100] H. R. Roth *et al.*, "Hierarchical 3d fully convolutional networks for multi-organ segmentation," *arXiv preprint arXiv:1704.06382*, 2017.

[101] H. R. Roth *et al.*, "Data from pancreas-ct. the cancer imaging archive," *IEEE Transactions on Image Processing*, vol. 10, p. K9, 2016.

[102] J. Sun *et al.*, "Saunet: Shape attentive u-net for interpretable medical image segmentation," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part IV 23*. Springer, 2020, pp. 797–806.

[103] P. Radau *et al.*, "Evaluation framework for algorithms segmenting short axis cardiac mri." *The MIDAS Journal*, 2009.

[104] O. Bernard *et al.*, "Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved?" *IEEE transactions on medical imaging*, vol. 37, no. 11, pp. 2514–2525, 2018.

[105] M. Karri, C. S. R. Annavarapu, and U. R. Acharya, "Explainable multi-module semantic guided attention based network for medical image segmentation," *Computers in Biology and Medicine*, vol. 151, p. 106231, 2022.

[106] P. Tschandl, C. Rosendahl, and H. Kittler, "The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. scientific data. 2018; 5: 180161," *Search in*, vol. 2, 2018.

[107] C. for Biomedical Image Computing and Analytics, "Multimodal brain tumor segmentation challenge 2020: Data," *MICCAI 2020 BraTs*, 2020. [Online]. Available: https://www.med.upenn.edu/cbica/brats2020/data.html

[108] A. E. Kavur *et al.*, "Chaos challenge - combined (ct-mr) healthy abdominal organ segmentation," *Medical Image Analysis*, vol. 69, p. 101950, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361841520303145

[109] H. Li *et al.*, "Pmjaf-net: Pyramidal multi-scale joint attention and adaptive fusion network for explainable skin lesion segmentation," *Computers in Biology and Medicine*, p. 107454, 2023.

[110] R. Souza *et al.*, "An open, multi-vendor, multi-field-strength brain mr dataset and analysis of publicly available skull stripping methods agreement," *NeuroImage*, vol. 170, pp. 482–494, 2018.

[111] C. Lian *et al.*, "End-to-end dementia status prediction from brain mri using multi-task weakly-supervised attention network," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part IV 22*. Springer, 2019, pp. 158–167.

[112] C. R. Jack Jr *et al.*, "Update on the magnetic resonance imaging core of the alzheimer's disease neuroimaging initiative," *Alzheimer's & Dementia*, vol. 6, no. 3, pp. 212–220, 2010.

[113] P. W. Koh *et al.*, "Concept bottleneck models," in *International conference on machine learning*. PMLR, 2020, pp. 5338–5348.

[114] M. Yuksekgonul, M. Wang, and J. Zou, "Post-hoc concept bottleneck models," *arXiv preprint arXiv:2205.15480*, 2022.

[115] R. Jain *et al.*, "Extending logic explained networks to text classification," *arXiv preprint arXiv:2211.09732*, 2022.

[116] A. Tan, F. Zhou, and H. Chen, "Explain via any concept: Concept bottleneck model with open vocabulary concepts," *arXiv preprint arXiv:2408.02265*, 2024.

[117] B. Kim *et al.*, "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)," in *International conference on machine learning*. PMLR, 2018, pp. 2668–2677.

[118] J. R. Clough *et al.*, "Global and local interpretability for cardiac mri classification," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 656–664.

[119] M. Graziani *et al.*, "Concept attribution: Explaining cnn decisions to physicians," *Computers in biology and medicine*, vol. 123, p. 103865, 2020.

[120] R. Achtibat *et al.*, "From attribution maps to human-understandable explanations through concept relevance propagation," *Nature Machine Intelligence*, vol. 5, no. 9, pp. 1006–1019, 2023.

[121] E. Poeta *et al.*, "Concept-based explainable artificial intelligence: A survey," *arXiv preprint arXiv:2312.12936*, 2023.

[122] A. Sun *et al.*, "Explain any concept: Segment anything meets concept-based explanation," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[123] A. Kirillov *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4015–4026.

[124] M. Nevitt, D. Felson, and G. Lester, "The osteoarthritis initiative," *Protocol for the cohort study*, vol. 1, p. 2, 2006.

[125] K. Chauhan *et al.*, "Interactive concept bottleneck models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 5, 2023, pp. 5948–5955.

[126] C. Patrício, J. C. Neves, and L. F. Teixeira, "Coherent concept-based explanations in medical image and its application to skin lesion diagnosis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 3799–3808.

[127] J. Kawahara *et al.*, "Seven-point checklist and skin lesion classification using multitask multimodal neural nets," *IEEE journal of biomedical and health informatics*, vol. 23, no. 2, pp. 538–546, 2018.

[128] T. Mendonça *et al.*, "Ph2: A public database for the analysis of dermoscopic images," *Dermoscopy image analysis*, vol. 2, 2015.

[129] S. Yan *et al.*, "Towards trustable skin cancer diagnosis via rewriting model's decision," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 11 568–11 577.

[130] Y. Bie, L. Luo, and H. Chen, "Mica: Towards explainable skin lesion diagnosis via multi-level image-concept alignment," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 2, 2024, pp. 837–845.

[131] R. Daneshjou *et al.*, "Skincon: A skin disease dataset densely annotated by domain experts for fine-grained debugging and analysis," *Advances in Neural Information Processing Systems*, vol. 35, pp. 18 157–18 167, 2022.

[132] C. Kim *et al.*, "Transparent medical image ai via an image–text foundation model grounded in medical literature," *Nature Medicine*, pp. 1–12, 2024.

[133] M. Groh *et al.*, "Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1820–1828.

[134] R. Daneshjou *et al.*, "Disparities in dermatology ai performance on a diverse, curated clinical image set," *Science advances*, vol. 8, no. 31, p. eabq6147, 2022.

[135] A. Lucieri *et al.*, "Exaid: A multimodal explanation framework for computer-aided diagnosis of skin lesions," *Computer Methods and Programs in Biomedicine*, vol. 215, p. 106620, 2022.

[136] R. Jalaboi *et al.*, "Dermx: An end-to-end framework for explainable automated dermatological diagnosis," *Medical Image Analysis*, vol. 83, p. 102647, 2023.

[137] N. Z. D. Society, "Dermatology images." [Online]. Available: https://dermnetnz.org/

[138] X. Sun *et al.*, "A benchmark for automatic visual classification of clinical skin disease images," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14*. Springer, 2016, pp. 206–222.

[139] J. Hou, J. Xu, and H. Chen, "Concept-attention whitening for interpretable skin lesion diagnosis," *arXiv preprint arXiv:2404.05997*, 2024.

[140] I. Kim *et al.*, "Concept bottleneck with visual concept filtering for explainable medical image classification," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2023, pp. 225–233.

[141] C. Patrício, L. F. Teixeira, and J. C. Neves, "Towards concept-based interpretability of skin lesion diagnosis using vision-language models," in *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2024, pp. 1–5.

[142] W. Pang *et al.*, "Integrating clinical knowledge into concept bottleneck models," in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2024.

[143] S. Tsutsui, W. Pang, and B. Wen, "Wbcatt: a white blood cell dataset annotated with detailed morphological attributes," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[144] R. Marcinkevičs *et al.*, "Interpretable and intervenable ultrasonography-based machine learning models for pediatric appendicitis," *Medical Image Analysis*, vol. 91, p. 103042, 2024.

[145] G. Zhao *et al.*, "Diagnose like a radiologist: Hybrid neuro-probabilistic reasoning for attribute-based medical image diagnosis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 11, pp. 7400–7416, 2021.

[146] S. G. Armato III *et al.*, "The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans," *Medical physics*, vol. 38, no. 2, pp. 915–931, 2011.

[147] Z. Fang *et al.*, "Concept-based explanation for fine-grained images and its application in infectious keratitis classification," in *Proceedings of the 28th ACM international conference on Multimedia*, 2020, pp. 700–708.

[148] Y. Xu *et al.*, "Deep sequential feature learning in clinical image classification of infectious keratitis," *Engineering*, vol. 7, no. 7, pp. 1002–1010, 2021.

[149] C. Wen *et al.*, "Concept-based lesion aware transformer for interpretable retinal disease diagnosis," *IEEE Transactions on Medical Imaging*, 2024.

[150] Y. Zhou *et al.*, "A benchmark for studying diabetic retinopathy: segmentation, grading, and transferability," *IEEE Transactions on Medical Imaging*, vol. 40, no. 3, pp. 818–828, 2020.

[151] T. Li *et al.*, "Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening," *Information Sciences*, vol. 501, pp. 511–522, 2019.

[152] M. Kong *et al.*, "Attribute-aware interpretation learning for thyroid ultrasound diagnosis," *Artificial Intelligence in Medicine*, vol. 131, p. 102344, 2022.

[153] J. Liu *et al.*, "A chatgpt aided explainable framework for zero-shot medical image diagnosis," *arXiv preprint arXiv:2307.01981*, 2023.

[154] S. Jaeger *et al.*, "Two public chest x-ray datasets for computer-aided screening of pulmonary diseases," *Quantitative imaging in medicine and surgery*, vol. 4, no. 6, p. 475, 2014.

[155] P. Porwal *et al.*, "Indian diabetic retinopathy image dataset (idrid): a database for diabetic retinopathy screening research," *Data*, vol. 3, no. 3, p. 25, 2018.

[156] Y. Gao *et al.*, "Aligning human knowledge with visual concepts towards explainable medical image classification," *arXiv preprint arXiv:2406.05596*, 2024.

[157] J. N. Kather, N. Halama, and A. Marx, "100,000 histological images of human colorectal cancer and healthy tissue," *Zenodo10*, vol. 5281, no. 9, 2018.

[158] W. Al-Dhabyani *et al.*, "Deep learning approaches for data augmentation and classification of breast masses using ultrasound images," *Int. J. Adv. Comput. Sci. Appl*, vol. 10, no. 5, pp. 1–11, 2019.

[159] M. Espinosa Zarlenga *et al.*, "Concept embedding models: Beyond the accuracy-explainability trade-off," *Advances in Neural Information Processing Systems*, vol. 35, pp. 21 400–21 413, 2022.

[160] Z. Chen, Y. Bei, and C. Rudin, "Concept whitening for interpretable image recognition," *Nature Machine Intelligence*, vol. 2, no. 12, pp. 772–782, 2020.

[161] S. Lapuschkin *et al.*, "Unmasking clever hans predictors and assessing what machines really learn," *Nature communications*, vol. 10, no. 1, p. 1096, 2019.

[162] A. Ghorbani *et al.*, "Towards automatic concept-based explanations," *Advances in neural information processing systems*, vol. 32, 2019.

[163] A. Sarkar *et al.*, "A framework for learning ante-hoc explainable models via concepts," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 286–10 295.

[164] C.-K. Yeh *et al.*, "On completeness-aware concept-based explanations in deep neural networks," *Advances in neural information processing systems*, vol. 33, pp. 20 554–20 565, 2020.

[165] Y. Yang *et al.*, "Language in a bottle: Language model guided concept bottlenecks for interpretable image classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19 187–19 197.

[166] T. Brown *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[167] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[168] T. Oikarinen *et al.*, "Label-free concept bottleneck models," *arXiv preprint arXiv:2304.06129*, 2023.

[169] Y. Bie *et al.*, "Xcoop: Explainable prompt learning for computer-aided diagnosis via concept-guided context optimization," *arXiv preprint arXiv:2403.09410*, 2024.

[170] E. Kim *et al.*, "Xprotonet: diagnosis in chest radiography with global and local explanations," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 15 719–15 728.

[171] C. Chen *et al.*, "This looks like that: deep learning for interpretable image recognition," *Advances in neural information processing systems*, vol. 32, 2019.

[172] I. Biederman, "Recognition-by-components: a theory of human image understanding." *Psychological review*, vol. 94, no. 2, p. 115, 1987.

[173] G. Singh and K.-C. Yow, "An interpretable deep learning model for covid-19 detection with chest x-ray images," *Ieee Access*, vol. 9, pp. 85 198–85 208, 2021.

[174] J. P. Cohen, P. Morrison, and L. Dao, "Covid-19 image data collection," *arXiv preprint arXiv:2003.11597*, 2020.

[175] S. Mohammadjafari *et al.*, "Using protopnet for interpretable alzheimer's disease classification." in *Canadian Conference on AI*, 2021.

[176] D. S. Marcus *et al.*, "Open access series of imaging studies (oasis): cross-sectional mri data in young, middle aged, nondemented, and demented older adults," *Journal of cognitive neuroscience*, vol. 19, no. 9, pp. 1498–1507, 2007.

[177] A. J. Barnett *et al.*, "A case-based interpretable deep learning model for classification of mass lesions in digital mammography," *Nature Machine Intelligence*, vol. 3, no. 12, pp. 1061–1070, 2021.

[178] G. Carloni *et al.*, "On the applicability of prototypical part learning in medical images: breast masses classification using protopnet," in *International Conference on Pattern Recognition*. Springer, 2022, pp. 539–557.

[179] R. S. Lee *et al.*, "A curated mammography data set for use in computer-aided detection and diagnosis research," *Scientific data*, vol. 4, no. 1, pp. 1–9, 2017.

[180] C. Wang *et al.*, "Knowledge distillation to ensemble global and interpretable prototype-based mammogram classification models," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2022, pp. 14–24.

[181] C. Cui *et al.*, "The chinese mammography database (cmmd): An online mammography database with biopsy confirmed types for machine diagnosis of breast," *The Cancer Imaging Archive*, vol. 1, 2021.

[182] Y. Wei, R. Tam, and X. Tang, "Mprotonet: A case-based interpretable model for brain tumor classification with 3d multi-parametric magnetic resonance imaging," in *Medical Imaging with Deep Learning*. PMLR, 2024, pp. 1798–1812.

[183] B. H. Menze *et al.*, "The multimodal brain tumor image segmentation benchmark (brats)," *IEEE transactions on medical imaging*, vol. 34, no. 10, pp. 1993–2024, 2014.

[184] L. S. Hesse and A. I. Namburete, "Insightr-net: interpretable neural network for regression using similarity-based comparisons to prototypical examples," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2022, pp. 502–511.

[185] C. H. Foundation, "Eyepacs," 2015. [Online]. Available: https://www.kaggle.com/c/diabetic-retinopathy-detection/data

[186] I. B. d. A. Santos and A. C. de Carvalho, "Protoal: Interpretable deep active learning with prototypes for medical imaging," *arXiv preprint arXiv:2404.04736*, 2024.

[187] E. Decencière *et al.*, "Feedback on a publicly distributed image database: The messidor database. image anal & stereology 33: 231–234," 2014.

[188] L. S. Hesse, N. K. Dinsdale, and A. I. L. Namburete, "Prototype learning for explainable brain age prediction," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, January 2024, pp. 7903–7913.

[189] https://brain-development.org/ixi-dataset/.

[190] A. T. Papageorghiou *et al.*, "International standards for fetal growth based on serial ultrasound measurements: the fetal growth longitudinal study of the intergrowth-21st project," *The Lancet*, vol. 384, no. 9946, pp. 869–879, 2014.

[191] D. Rymarczyk *et al.*, "Protopshare: Prototypical parts sharing for similarity discovery in interpretable image classification," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2021, pp. 1420–1430.

[192] D. Rymarczyk *et al.*, "Interpretable image classification with differentiable prototypes assignment," in *European Conference on Computer Vision*. Springer, 2022, pp. 351–368.

[193] J. Donnelly, A. J. Barnett, and C. Chen, "Deformable protopnet: An interpretable image classifier using deformable prototypes," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 265–10 275.

[194] J. Wang *et al.*, "Interpretable image recognition by constructing transparent embedding space," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 895–904.

[195] B. Wang *et al.*, "Learning bottleneck concepts in image classification," in *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, 2023, pp. 10 962–10 971.

[196] P. Hase *et al.*, "Interpretable image recognition with hierarchical prototypes," in *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, vol. 7, 2019, pp. 32–40.

[197] Y. Ukai *et al.*, "This looks like it rather than that: Protoknn for similarity-based classifiers," in *The Eleventh International Conference on Learning Representations*, 2022.

[198] A. Bontempelli *et al.*, "Concept-level debugging of part-prototype networks," in *The Eleventh International Conference on Learning Representations*, 2023.

[199] O. Li *et al.*, "Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[200] Q. Huang *et al.*, "Evaluation and improvement of interpretability for self-explainable part-prototype networks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 2011–2020.

[201] M. Nauta *et al.*, "Pip-net: Patch-based intuitive prototypes for interpretable image classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 2744–2753.

[202] C. Ma *et al.*, "This looks like those: Illuminating prototypical concepts using multiple visualizations," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[203] M. Nauta, R. Van Bree, and C. Seifert, "Neural prototype trees for interpretable fine-grained image recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 14 933–14 943.

[204] A. Tan, Z. Fengtao, and H. Chen, "Post-hoc part-prototype networks," in *Forty-first International Conference on Machine Learning*.

[205] J. Kim, M. Kim, and Y. M. Ro, "Interpretation of lesional detection via counterfactual generation," in *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2021, pp. 96–100.

[206] P. Pino *et al.*, "Clinically correct report generation from chest x-rays using templates," in *Machine Learning in Medical Imaging: 12th International Workshop, MLMI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings 12*. Springer, 2021, pp. 654–663.

[207] C. E. Lipscomb, "Medical subject headings (mesh)," *Bulletin of the Medical Library Association*, vol. 88, no. 3, p. 265, 2000.

[208] H.-C. Shin *et al.*, "Learning to read chest x-rays: Recurrent neural cascade model for automated image annotation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2497–2506.

[209] A. Gasimova, "Automated enriched medical concept generation for chest x-ray images," in *Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support: Second International Workshop, iMIMIC 2019, and 9th International Workshop, ML-CDS 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 17, 2019, Proceedings 9*. Springer, 2019, pp. 83–92.

[210] I. Rodin *et al.*, "Multitask and multimodal neural network model for interpretable analysis of x-ray images," in *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2019, pp. 1601–1604.

[211] Z. Zhang *et al.*, "Mdnet: A semantically and visually interpretable medical image diagnosis network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 6428–6436.

[212] Z. Zhang *et al.*, "Tandemnet: Distilling knowledge from medical images using diagnostic reports as optional semantic references," in *Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part III 20*. Springer, 2017, pp. 320–328.

[213] K. Ma *et al.*, "A pathology image diagnosis network with visual interpretability and structured diagnostic report," in *Neural Information Processing: 25th International Conference, ICONIP 2018, Siem Reap, Cambodia, December 13–16, 2018, Proceedings, Part VI 25*. Springer, 2018, pp. 282–293.

[214] X. Wang *et al.*, "A computational framework towards medical image explanation," in *Artificial Intelligence in Medicine: Knowledge Representation and Transparent and Explainable Systems: AIME 2019 International Workshops, KR4HC/ProHealth and TEAAM, Poznan, Poland, June 26–29, 2019, Revised Selected Papers*. Springer, 2019, pp. 120–131.

[215] W. Gale *et al.*, "Producing radiologist-quality reports for interpretable deep learning," in *2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019)*. IEEE, 2019, pp. 1275–1279.

[216] J. T. Wu *et al.*, "Chest imagenome dataset for clinical reasoning," in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

[217] Q. Li *et al.*, "Anatomical structure-guided medical vision-language pre-training," *arXiv preprint arXiv:2403.09294*, 2024.

[218] T. Tanida *et al.*, "Interactive and explainable region-guided radiology report generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 7433–7442.

[219] L. Wang *et al.*, "An inclusive task-aware framework for radiology report generation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2022, pp. 568–577.

[220] D. Demner-Fushman *et al.*, "Design and development of a multimodal biomedical information retrieval system," *Journal of Computing Science and Engineering*, vol. 6, no. 2, pp. 168–177, 2012.

[221] S. Singh *et al.*, "From chest x-rays to radiology reports: a multimodal machine learning approach," in *2019 Digital Image Computing: Techniques and Applications (DICTA)*. IEEE, 2019, pp. 1–8.

[222] G. Spinks and M.-F. Moens, "Justifying diagnosis decisions by deep neural networks," *Journal of biomedical informatics*, vol. 96, p. 103248, 2019.

[223] Y. Kim *et al.*, "Adversarially regularized autoencoders for generating discrete structures," *arXiv preprint arXiv:1706.04223*, vol. 2, p. 12, 2017.

[224] G. Liu *et al.*, "Clinically accurate chest x-ray report generation," in *Machine Learning for Healthcare Conference*. PMLR, 2019, pp. 249–269.

[225] Z. Chen *et al.*, "Generating radiology reports via memory-driven transformer," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 1439–1449.

[226] Z. Wang *et al.*, "Metransformer: Radiology report generation by transformer with multiple learnable expert tokens," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 11 558–11 567.

[227] J. Yuan *et al.*, "Automatic radiology report generation based on multi-view image fusion and medical concept enrichment," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22*. Springer, 2019, pp. 721–729.

[228] H. Lee, S. Kim, and Y. Ro, "Generation of multimodal justification using visual word constraint model for explainable computer-aided diagnosis," in *Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support*. Springer, 2019.

[229] Z. Zhang *et al.*, "Pathologist-level interpretable whole-slide cancer diagnosis with deep learning," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 236–245, 2019.

[230] N. C. Institute, "The cancer genome atlas program," 2006. [Online]. Available: https://www.cancer.gov/tcga

[231] X. Wang *et al.*, "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*, 2018, pp. 9049–9058.

[232] B. Jing, P. Xie, and E. Xing, "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)*, 2018, pp. 2577–2586.

[233] K. N. Jones *et al.*, "Peir digital library: Online resources and authoring system," in *Proceedings of the AMIA Symposium*. American Medical Informatics Association, 2001, p. 1075.

[234] X. Zeng *et al.*, "Generating diagnostic report for medical image by high-middle-level visual information incorporation on double deep learning models," *Computer methods and programs in biomedicine*, vol. 197, p. 105700, 2020.

[235] J. Tian *et al.*, "A diagnostic report generator from ct volumes on liver tumor with semi-supervised attention mechanism," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II 11*. Springer, 2018, pp. 702–710.

[236] P. Bilic *et al.*, "The liver tumor segmentation benchmark (lits)," *Medical Image Analysis*, vol. 84, p. 102680, 2023.

[237] O. Thawkar *et al.*, "Xraygpt: Chest radiographs summarization using medical vision-language models," *arXiv preprint arXiv:2306.07971*, 2023.

[238] J. Zhou *et al.*, "Pre-trained multimodal large language model enhances dermatological diagnosis using skingpt-4," *Nature Communications*, vol. 15, no. 1, p. 5649, 2024.

[239] kaggle, "Dermnet." [Online]. Available: https://www.kaggle.com/datasets/shubhamgoel27/dermnet

[240] M. Moor *et al.*, "Med-flamingo: a multimodal medical few-shot learner," in *Machine Learning for Health (ML4H)*. PMLR, 2023, pp. 353–367.

[241] W. Lin *et al.*, "Pmc-clip: Contrastive language-image pre-training using biomedical documents," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2023, pp. 525–536.

[242] C. Li *et al.*, "Llava-med: Training a large language-and-vision assistant for biomedicine in one day," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[243] J. J. Lau *et al.*, "A dataset of clinically generated visual questions and answers about radiology images," *Scientific data*, vol. 5, no. 1, pp. 1–10, 2018.

[244] S. He *et al.*, "Meddr: Diagnosis-guided bootstrapping for large-scale medical vision-language learning," *arXiv preprint arXiv:2404.15127*, 2024.

[245] J. Chen *et al.*, "Huatuogpt-vision, towards injecting medical visual knowledge into multimodal llms at scale," *arXiv preprint arXiv:2406.19280*, 2024.

[246] X. Zhang *et al.*, "Pmc-vqa: Visual instruction tuning for medical visual question answering," *arXiv preprint arXiv:2305.10415*, 2023.

[247] S. Kang *et al.*, "Wolf: Large language model framework for cxr understanding," *arXiv preprint arXiv:2403.15456*, 2024.

[248] S. Bae *et al.*, "Ehrxqa: A multi-modal question answering dataset for electronic health records with chest x-ray images," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[249] Z. Chen *et al.*, "Chexagent: Towards a foundation model for chest x-ray interpretation," *arXiv preprint arXiv:2401.12208*, 2024.

[250] K. Schutte *et al.*, "Using stylegan for visual interpretability of deep learning models on medical images," *arXiv preprint arXiv:2101.07563*, 2021.

[251] I. Goodfellow *et al.*, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[252] T. Laugel *et al.*, "Issues with post-hoc counterfactual explanations: a discussion," *arXiv preprint arXiv:1906.04774*, 2019.

[253] H. Guo, T. H. Nguyen, and A. Yadav, "Counternet: End-to-end training of prediction aware counterfactual explanations," in *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2023, pp. 577–589.

[254] V. Guyomard *et al.*, "Vcnet: A self-explaining model for realistic counterfactual generation," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2022, pp. 437–453.

[255] M. Wilms *et al.*, "Towards self-explainable classifiers and regressors in neuroimaging with normalizing flows," in *International Workshop on Machine Learning in Clinical Neuroimaging*, 2021, pp. 23–33.

[256] U. Johansson, R. König, and L. Niklasson, "The truth is in there-rule extraction from opaque models using genetic programming." in *FLAIRS*, 2004, pp. 658–663.

[257] H. Lakkaraju *et al.*, "Faithful and customizable explanations of black box models," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 2019, pp. 131–138.

[258] W. Jin *et al.*, "Guidelines and evaluation of clinical explainable ai in medical image analysis," *Medical Image Analysis*, vol. 84, p. 102684, 2023.

[259] E. Lughofer *et al.*, "Explaining classifier decisions linguistically for stimulating and improving operators labeling behavior," *Information Sciences*, vol. 420, pp. 16–36, 2017.

[260] M. Robnik-Šikonja and M. Bohanec, "Perturbation-based explanations of prediction models," *Human and Machine Learning: Visible, Explainable, Trustworthy and Transparent*, pp. 159–175, 2018.

[261] A. Adadi and M. Berrada, "Explainable ai for healthcare: from black box to interpretable models," in *Embedded systems and artificial intelligence: proceedings of ESAI 2019, Fez, Morocco*. Springer, 2020, pp. 327–337.

[262] X. Wu *et al.*, "Usable xai: 10 strategies towards exploiting explainability in the llm era," *arXiv preprint arXiv:2403.08946*, 2024.

[263] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:1702.08608*, 2017.

[264] J. Zhou *et al.*, "Evaluating the quality of machine learning explanations: A survey on methods and metrics," *Electronics*, vol. 10, no. 5, p. 593, 2021.

[265] A. Holzinger, A. Carrington, and H. Müller, "Measuring the quality of explanations: the system causability scale (scs) comparing human and machine explanations," *KI-Künstliche Intelligenz*, vol. 34, no. 2, pp. 193–198, 2020.

[266] R. Sayres *et al.*, "Using a deep learning algorithm and integrated gradients explanation to assist grading for diabetic retinopathy," *Ophthalmology*, vol. 126, no. 4, pp. 552–564, 2019.

[267] V. Petsiuk, A. Das, and K. Saenko, "Rise: Randomized input sampling for explanation of black-box models," *arXiv preprint arXiv:1806.07421*, 2018.

[268] S. Hooker *et al.*, "A benchmark for interpretability methods in deep neural networks," *Advances in neural information processing systems*, vol. 32, 2019.

[269] J. Hou *et al.*, "Diabetic retinopathy grading with weakly-supervised lesion priors," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[270] A. J. Barnett *et al.*, "Interpretable mammographic image classification using case-based reasoning and deep learning," *arXiv preprint arXiv:2107.05605*, 2021.

[271] S. M. Muddamsetty, M. N. Jahromi, and T. B. Moeslund, "Expert level evaluations for explainable ai (xai) methods in the medical domain," in *International Conference on Pattern Recognition*. Springer, 2021, pp. 35–46.

[272] A.-p. Nguyen and M. R. Martínez, "On quantitative aspects of model interpretability," *arXiv preprint arXiv:2007.07584*, 2020.

[273] K. Papineni *et al.*, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[274] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.

[275] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4566–4575.

[276] S. Singla *et al.*, "Explaining the black-box smoothly—a counterfactual approach," *Medical Image Analysis*, vol. 84, p. 102721, 2023.

[277] S. Tonekaboni *et al.*, "What clinicians want: contextualizing explainable machine learning for clinical end use," in *Machine learning for healthcare conference*. PMLR, 2019, pp. 359–380.

[278] K. Singhal *et al.*, "Towards expert-level medical question answering with large language models," *arXiv preprint arXiv:2305.09617*, 2023.

[279] C. Wu *et al.*, "Pmc-llama: toward building open-source language models for medicine," *Journal of the American Medical Informatics Association*, p. ocae045, 2024.

[280] Y. Gu *et al.*, "Domain-specific language model pretraining for biomedical natural language processing," *ACM Transactions on Computing for Healthcare (HEALTH)*, vol. 3, no. 1, pp. 1–23, 2021.

[281] M. Moor *et al.*, "Foundation models for generalist medical artificial intelligence," *Nature*, vol. 616, no. 7956, pp. 259–265, 2023.

[282] X. Ye and G. Durrett, "Can explanations be useful for calibrating black box models?" in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 6199–6212.

[283] X. Wu *et al.*, "From language modeling to instruction following: Understanding the behavior shift in llms after instruction tuning," in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 2024, pp. 2341–2369.

[284] J. Wei *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24 824–24 837, 2022.

[285] Y. Shi *et al.*, "Mededit: Model editing for medical question answering with external knowledge bases," *arXiv preprint arXiv:2309.16035*, 2023.

[286] K. Singhal *et al.*, "Large language models encode clinical knowledge," *Nature*, vol. 620, no. 7972, pp. 172–180, 2023.

[287] C. Zhao *et al.*, "Automated natural language explanation of deep visual neurons with large models," *arXiv preprint arXiv:2310.10708*, 2023.

[288] Y. Yang *et al.*, "Language in a bottle: Language model guided concept bottlenecks for interpretable image classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19 187–19 197.

[289] C. Singh *et al.*, "Augmenting interpretable models with large language models during training," *Nature Communications*, vol. 14, no. 1, p. 7913, 2023.

[290] S. Bills *et al.*, "Language models can explain neurons in language models," *URL https://openaipublic. blob. core. windows. net/neuron-explainer/paper/index. html.(Date accessed: 14.05. 2023)*, vol. 2, 2023.

[291] D. Gunning and D. Aha, "Darpa's explainable artificial intelligence (xai) program," *AI magazine*, vol. 40, no. 2, pp. 44–58, 2019.
[292] D. Minh *et al.*, "Explainable artificial intelligence: a comprehensive review," *Artificial Intelligence Review*, pp. 1–66, 2022.
[293] C. Rudin *et al.*, "Interpretable machine learning: Fundamental principles and 10 grand challenges," *Statistic Surveys*, vol. 16, pp. 1–85, 2022.
[294] Y. Xu and H. Chen, "Multimodal optimal transport-based co-attention transformer with global structure consistency for survival prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 21 241–21 251.
[295] J. Pearl, *Causality*. Cambridge university press, 2009.
[296] A. J. DeGrave, J. D. Janizek, and S.-I. Lee, "Ai for radiographic covid-19 detection selects shortcuts over signal," *Nature Machine Intelligence*, vol. 3, no. 7, pp. 610–619, 2021.
[297] D. C. Castro, I. Walker, and B. Glocker, "Causality matters in medical imaging," *Nature Communications*, vol. 11, no. 1, p. 3673, 2020.
[298] L. Luo *et al.*, "Pseudo bias-balanced learning for debiased chest x-ray classification," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2022, pp. 621–631.
[299] J. Wu *et al.*, "Chest imagenome dataset," *Physio Net*, 2021.
[300] H. Cai *et al.*, "Breast microcalcification diagnosis using deep convolutional neural network from digital mammograms," *Computational and mathematical methods in medicine*, vol. 2019, no. 1, p. 2717454, 2019.
[301] G. Argenziano *et al.*, "Epiluminescence microscopy for the diagnosis of doubtful melanocytic skin lesions: comparison of the abcd rule of dermatoscopy and a new 7-point checklist based on pattern analysis," *Archives of dermatology*, vol. 134, no. 12, pp. 1563–1570, 1998.