Explainable Deep Learning Methods in Medical Image Classification: A Survey

CRISTIANO PATRÍCIO and JOÃO C. NEVES, University of Beira Interior and NOVA LINCS, Portugal
LUÍS F. TEIXEIRA, University of Porto and INESC TEC, Portugal

The remarkable success of deep learning has prompted interest in its application to medical imaging diagnosis. Even though state-of-the-art deep learning models have achieved human-level accuracy on the classification of different types of medical data, these models are hardly adopted in clinical workflows, mainly due to their lack of interpretability. The black-box nature of deep learning models has raised the need for devising strategies to explain the decision process of these models, leading to the creation of the topic of eXplainable Artificial Intelligence (XAI). In this context, we provide a thorough survey of XAI applied to medical imaging diagnosis, including visual, textual, example-based and concept-based explanation methods. Moreover, this work reviews the existing medical imaging datasets and the existing metrics for evaluating the quality of the explanations. In addition, we include a performance comparison among a set of report generation–based methods. Finally, the major challenges in applying XAI to medical imaging and the future research directions on the topic are discussed.

CCS Concepts: • **Applied computing → Health care information systems;**

Additional Key Words and Phrases: Explainable AI, explainability, interpretability, deep learning, medical image analysis

ACM Reference format:
Cristiano Patrício, João C. Neves, and Luís F. Teixeira. 2023. Explainable Deep Learning Methods in Medical Image Classification: A Survey. *ACM Comput. Surv.* 56, 4, Article 85 (October 2023), 41 pages.
https://doi.org/10.1145/3625287

1 INTRODUCTION

The progress made in the last decade in the field of artificial intelligence (AI) has supported a dramatic increase in the accuracy of most computer vision applications. Medical image analysis is one of the applications where the progress made ensured human-level accuracy on the classification of different types of medical data (e.g., chest X-rays [90] and corneal images [166]). However, in spite of these advances, automated medical imaging is seldom adopted in clinical practice. According to Zachary Lipton [77], the explanation for this apparent paradox is straightforward: doctors

This work was funded by the Portuguese Foundation for Science and Technology (FCT) under the PhD grant no. 2022.11566.BD and supported by NOVA LINCS (UIDB/04516/2020) with the financial support of FCT.IP.
Authors' addresses: C. Patrício and J. C. Neves, University of Beira Interior and NOVA LINCS, Covilhã, Portugal, 6201-001; e-mails: cristiano.patricio@ubi.pt, jcneves@di.ubi.pt; L. F. Teixeira, University of Porto and INESC TEC, Porto, Portugal, 4200-465; e-mail: luisft@fe.up.pt.

0360-0300/2023/10-ART85
https://doi.org/10.1145/3625287

will never trust the decision of an algorithm without understanding its decision process. This fact has raised the need for producing strategies capable of explaining the decision process of AI algorithms, leading to the creation of a novel research topic termed *eXplainable Artificial Intelligence (XAI)*. According to the Defense Advanced Research Projects Agency (DARPA) [46], XAI aims to "*produce more explainable models, while maintaining a high level of learning performance (prediction accuracy); and enable human users to understand, appropriately, trust, and effectively manage the emerging generation of artificially intelligent partners*". In spite of its general applicability, XAI is particularly important in high-stake decisions, such as clinical workflow, where the consequences of a wrong decision could lead to human deaths. This is also evidenced by the European Union's General Data Protection Regulation (GDPR) law, which requires an explanation of the decision-making process of algorithms, thus improving their transparency before they can be used for patient care [41].

It is therefore of utmost importance to invest in the research of novel strategies to improve the interpretability of deep learning methods before deploying them in clinical practice. Previously, the research on this topic focused primarily on devising methods for indirectly analysing the decision process of pre-built models. These methods operate either by analysing the impact of specific regions of the input images on the final prediction (perturbation-based methods [87, 112] and occlusion-based methods [173]) or inspecting the network activations (saliency methods [127, 176]). The fact that these methods can be applied to arbitrary network architectures without requiring an additional customization of the model has supported their popularity in the early days of XAI. However, it has been recently shown that post-hoc strategies suffer from several drawbacks regarding the significance of the explanations [2, 117]. As a consequence, researchers have focused their attention on the design of models/architectures capable of explaining their decision process *per se*. Inherently interpretable models are believed to be particularly useful in medical imaging [117], justifying the recent growth in the number of medical imaging works focusing on this paradigm rather than the traditional post-hoc strategies [60, 162]. In spite of the recent popularity of inherently interpretable models, the existing surveys on the interpretability of deep learning applied to medical imaging have not comprehensively reviewed the progress made in this novel research trend. Also, the significant increase in the number of works focused on the interpretation of the decision process of deep learning applied to medical imaging justifies the need for an updated review of the most recent methods not covered by the last surveys on the topic. Moreover, the particular challenges of medical imaging analysis, including image complexity (anatomical structures, organs, and artifacts are often harder to identify compared with general images), data availability, and the misclassification risk, emphasize the need for a dedicated survey on interpretability applied to medical imaging.

To address these concerns, we comprehensively review the recent advances on explainable deep learning applied to medical diagnosis. In particular, this survey provides the following contributions.

- A review of the recent surveys on the topic of interpretable deep learning in medical imaging, including the major conclusions derived from each work, as well as a comparative analysis to our survey
- An exhaustive list of the datasets commonly used in the study of interpretability of deep learning approaches applied to medical imaging
- A comprehensive review of the state-of-the-art interpretable medical imaging approaches, covering both post-hoc and inherently interpretable models
- A complete description of the metrics commonly used for benchmarking interpretability methods either for visual or textual explanations

- A benchmark of interpretable medical imaging approaches regarding the quality of the textual explanations
- The future research directions on the topic of interpretable deep learning in medical imaging

2 RELATED SURVEYS

Explaining the decisions of deep learning models has been an active area of research, with various methods and algorithms proposed in the literature over the years. The rapid pace of development of XAI has raised the need for comprehensive overviews of the advances in the state-of-the-art and, in most cases, the analysis of specific domains, due to the vast number of works published over the years. Accordingly, in this section, we provide a critical analysis of the existing surveys in deep learning applied to medical imaging (Section 2.1), with a particular focus on explainable approaches (Section 2.2), and we compare the surveys analyzed with our survey (Section 2.3).

2.1 Deep Learning in Medical Image Analysis

The advent of deep learning significantly changed the field of Computer Vision, in which handcrafted feature extraction was replaced by end-to-end learning using Convolutional Neural Networks (CNNs). This new paradigm emerged in 2012 through the seminal work of Krizhevsky et al. [65], but it was not immediately incorporated by all the applications of Computer Vision. Litjens et al. [79] were the first to review the advances in medical image analysis fostered by the advent of deep learning, in which it is clear that the use of CNNs in medical imaging research became the standard approach in 2017, being clearly preferred over traditional handcrafted feature extraction for most of the anatomical regions. Based on the works reviewed, the authors concluded that data preprocessing and data augmentation techniques were essential to obtain superior results and that the combination of medical images with text data (medical reports) could improve the image classification accuracy [124]. Despite the relevance of this survey at the time, the rapid advances in the field of deep learning occurring in the last 5 years have made this work outdated, since the major conclusions of the survey are currently common sense, and novel deep learning models are currently being used in the medical imaging domain.

Considering this, Rehman et al. [111] provided an updated overview of the advances in deep learning applied to medical image analysis. The survey was divided over different pattern recognition tasks (image classification, segmentation, and image registration). Regarding the image classification task, the authors suggested using generative models to perform data augmentation to improve the results. The survey also gave future research directions to overcome the most common challenges identified by Litjens et al. [79]. The use of techniques such as transfer learning and synthetic data generation were suggested to address these challenges while improving the generalization capability of the developed strategies. Nevertheless, the authors concluded that the non-availability of large-scale annotated datasets remains one of the major challenges in medical imaging, which is impacting the performance of the deep learning models due to data overfitting and bias issues.

2.2 Interpretable Deep Learning in Medical Imaging

The works of Litjens et al. [79] and Rehman et al. [111] show that over the years the use of deep learning has greatly improved the performance of medical imaging analysis algorithms, also allowing also the creation of a myriad of approaches for the different image modalities and recognition tasks. Nevertheless, this contrasts with the adoption of these algorithms by clinicians who refuse to rely on decisions that they do not understand [77]. In fact, as foreseen by Litjens et al. [79], the importance of designing interpretable models for medical imaging has been growing over time and is now one of the major challenges in medical imaging. The following describes the different

surveys focused on reviewing the recent advances on interpretable deep learning applied to medical imaging. The major conclusions derived from each work and a brief comparison to our survey are also outlined.

2.2.1 General Reviews. Tjoa and Guan [145] provide a general overview of machine learning and deep learning interpretability methods with an emphasis on their application to the medical field. The authors consider two types of interpretability: (i) perceptive interpretability, which include the saliency methods, and (ii) interpretability by mathematical structures, which include mathematical formulations that can analyze the patterns in data. Although a significant number of works per image modality are covered, the survey lacks a comparison between the reviewed works. Also, the survey of Tjoa and Guan is more suitable for technically oriented readers, which is corroborated by a poor intuitive categorization for the medical community. Most works were discussed based on their mathematical foundation instead of describing the rationale behind the proposed method. As major conclusions, Tjoa and Guan state that combining visual and textual explanations is the most promising modality for conveying the explanations of medical imaging analysis algorithms and that these algorithms should always be considered as a complementary aid/support to clinicians, who should be responsible for the final decision.

Singh et al. [133] propose a review of the works related to the explainability of deep learning models in the context of medical imaging. The methods are broadly divided in two major categories (attribution-based and non-attribution-based) with respect to their capability of determining the contribution of an input feature to the target output. Both categories are reviewed by describing works applied to the different image modalities of medical data. Nevertheless, the survey focuses primarily on the attribution-based category, providing a superficial discussion of existing methods in the different categories of non-attribution methods, including inherently interpretable approaches. Based on the works reviewed, the authors conclude that leveraging patient record data and images can be an exciting research direction to push forward the performance of deep learning in medical imaging. When compared with our survey, the work of Singh et al. lacks a comprehensive analysis of inherently interpretable approaches, an analysis of the available medical imaging datasets, and the benchmarking of the most prominent reviewed methods.

Recently, Salahuddin et al. [119] reviewed a set of interpretability methods which are grouped into nine different categories based on the type of explanations generated. They also discussed the problem of evaluating explanations and described a set of evaluation strategies adopted to quantitatively and qualitatively measure the quality of the explanations. Similar to the other surveys, the authors also emphasized the importance of involving clinicians in designing and validating interpretability models to ensure the utility of the generated explanations. For future perspectives, Salahuddin et al. claimed that case-based and concept-learning models are promising interpretability models for being inherently interpretable and achieving similar performance to black-box CNNs. Despite being one of the most complete surveys on the topic, it lacks the description of most relevant datasets of the field as well as the benchmarking of most prominent approaches reviewed.

2.2.2 Specific Image Modality Reviews. In contrast to the above-mentioned works, the work of Pocevičiūtė et al. [108] is focused on a particular image modality. The XAI techniques devised for digital pathology are reviewed with respect to three criteria: (1) what is going to be explained (e.g., model predictions, predictions uncertainty); (2) explanation modality; and (3) how the explanations are derived (e.g., perturbation-based strategies, interpretable network design). The authors point out the importance of developing a toolbox for objectively measuring the quality of explanations, as the lack of an evaluation framework remains an open problem in the XAI field. Additionally, the authors state that the use of counterfactual examples can enhance the interpretability of the

methods. When compared with our survey, this work has disregarded the textual explanation modality, focusing solely on visual explanations, either by visual examples or saliency maps.

Gulum et al. [45] review the visual explainability techniques applied to cancer detection from Magnetic Resonance Imaging (MRI) scans. Contrary to the work of Pocevičiūtė et al. [108], Gulum et al. discuss the strategies used for measuring the quality of explanations, but they only consider one metric to quantitatively evaluate the explanations. They also emphasize that there is a lack of studies that assess the explanation methods based on human evaluation. As future directions, Gulum et al. highlight the need for developing inherently interpretable approaches as opposed to the traditional post-hoc strategy. Finally, the authors proposed the use of uncertainty estimation associated with model predictions to perceive how a model is confident in making a prediction. Despite its relevance, the work of Gulum et al. is specific to a particular image modality (MRI) and target disease (cancer).

2.2.3 Specific Explanation Modality Reviews. While visual explanations are usually the primary option for explaining the model decisions, these strategies can be unreliable since they often highlight regions regardless of the class of interest [117]. This has fostered the research on textual explanations, and in the particular case of medical imaging led researchers to devise approaches capable of producing different types of textual explanations: (1) textual concepts and (2) textual reports. The recent developments on this topic have been covered in the surveys of Messina et al. [95] and Ayesha et al. [8].

In [95], a thorough overview of the current state-of-the-art on automatic report generation from medical images is provided. The authors review 40 papers with respect to four dimensions: datasets used, model design, explainability, and evaluation metrics. Furthermore, a benchmark of most relevant approaches is provided with respect to the performance in terms of NLP metrics on the IU Chest X-ray dataset. Based on the works analyzed, the authors identify the following challenges and future research directions: (1) the validation of the obtained explanations by clinicians is impractical, being necessary to create automatic metrics positively correlated with the clinicians' opinion; (2) most research has concentrated on chest X-rays, mainly due to the availability of public data; and (3) supervised learning may not be the most adequate strategy for medical report generation learning, whereas reinforcement learning seems a more reasonable training paradigm to explore.

Similar to the work of Messina et al. [95], Ayesha et al. [8] present a detailed survey of the existing automatic caption generation methods for medical images. The most used datasets and the evaluation metrics are also discussed. An extensive study is done around the most significant works under the various deep learning–based medical imaging caption generation methods: encoder-decoder based, retrieval based, attention based, concepts detection based, and patient's metadata based. Additionally, a comparative analysis of the performance of the methods reviewed is provided. Finally, Ayesha et al. suggest some future research directions to deal with the main open issues in medical imaging. They point out the lack of large-scale annotated datasets as the major limitation in the medical imaging field, where the data is scarce and often mislabelled. Also, they claim that the lack of a suitable evaluation metric to assess the generated caption remains an open problem since the evaluation of the generated text is still based on standard NLP metrics, such as BLEU score, ROUGE, METEOR, and CIDEr. As a final remark, Ayesha et al. also indicate the importance of having a model capable of detecting multiple diseases simultaneously.

2.3 Discussion

Despite the significant contributions of each reviewed survey, few have described the most important datasets for medical imaging. Moreover, most surveys focused on particular aspects of

Table 1. Comparative Analysis between the Surveys on the Topic of Explainable Deep Learning Applied to Medical Imaging

Survey	Year	Explanations		Model Type		Medical Imaging Datasets	Benchmarking Performance
		Visual	*Textual*	*Post-hoc*	*In-model*		
Pocevičiūtė et al. [108]	2020	✓	✗	✓	✓	✗	✗
Tjoa and Guan [145]	2020	✓	✓	✓	✓	✗	✗
Singh et al. [133]	2020	✓	✓	✓	✓	✗	✗
Gulum et al. [45]	2021	✓	✗	✓	✓	✗	✗
Ayesha et al. [8]	2021	✗	✓	✗	✓	✓	✓
Salahuddin et al. [119]	2022	✓	✓	✓	✓	✗	✗
Messina et al. [95]	2022	✗	✓	✗	✓	✓	✓
This survey	2023	✓	✓	✓	✓	✓	✓

Our survey is the first to comprehensively review the advances on the topic regarding the different explanation modalities and the explanation processes. Also, it analyses the most relevant datasets on the the field, as well as their use for the development of explainable approaches.

interpretability, such as visual or textual approaches, and few works have comprehensively reviewed inherently interpretable models devised for medical image analysis. Another problem was the lack of a performance comparison among the reviewed methods. Accordingly, this survey covers these limitations by providing a broader overview of the current state-of-the-art XAI applied to medical diagnosis, including uni- and multimodal approaches, followed by the most important medical imaging datasets and a comparative analysis of the models' performance using standard evaluation metrics. In addition, this survey explores a contemporary trend and an under-exploited category of inherently interpretable models, specifically concept-based learning approaches. As delineated in subsequent sections, these approaches are advantageous for medical diagnosis as they provide explanations in the context of high-level concepts that align with the knowledge of the physicians and promote the interaction between physicians and AI through model intervention. Table 1 summarizes the major differences between the reviewed surveys and this survey.

3 BACKGROUND IN XAI

From a historical perspective, the problem of explaining expert systems has its origin in the mid-1980s [98], but the term *XAI* was only introduced in 2004 by Van Lent et al. [152]. Nevertheless, XAI did not gain prominence until deep learning dominated AI; the first sign of this interest was demonstrated by the launching, in 2015, of the Explainable AI program by DARPA, whose primary goal was producing more explainable models to increase their understanding and transparency, leading to greater trust by users. Later, the European Union (EU) [41] introduced legislation about the "right to algorithmic explanation", which provided citizens with the right to receive an explanation for algorithmic decisions obtained from personal data. Considering this, researchers shifted their efforts towards the creation of interpretable models rather than simply focusing on accuracy, leading to an exponential increase in the popularity and interest in XAI, whose number of works on the topic has rapidly increased over the last few years.

This section provides a general overview of the taxonomy of XAI methods, the description of seminal XAI methods, and the existing frameworks providing implementations of these methods (Table 4 in Appendix A.1).

3.1 XAI Methods Taxonomy

Based on the reviewed literature, XAI methods can be categorized according to three criteria: (i) Model-Agnostic versus Model-Specific; (ii) Global Interpretability versus Local Interpretability; and (iii) Post-hoc versus Intrinsic. Figure 1 illustrates the general taxonomy of the XAI methods, and each category is detailed in the following.

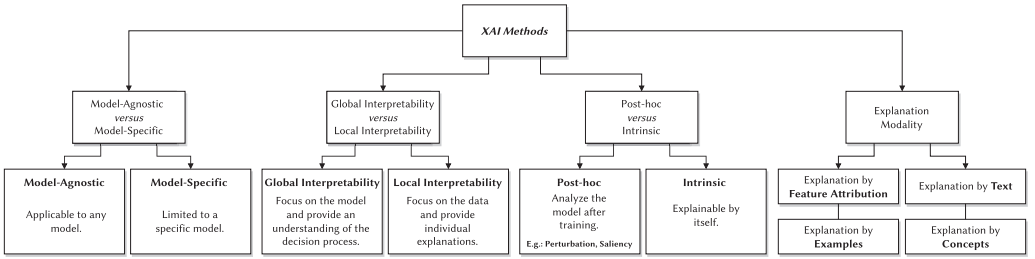


Fig. 1. Proposed categorization of the XAI methods taxonomy. The proposed categorization was inspired by the various taxonomies presented in the reviewed papers.

Model-Agnostic versus Model-Specific. A distinguishing factor between interpretability approaches is their comprehensiveness regarding the models they can be applied to. Model-agnostic methods can be used to explain arbitrary models, not being limited to a specific model architecture. Conversely, model-specific methods are restricted to a specific model architecture, meaning that these methods require access to the model's internal information.

Global Interpretability versus Local Interpretability. The type of explanations provided by XAI methods can be broadly divided into global and local according to whether the explanations provide insights about the model functioning for the general data distribution or for a specific data sample, respectively. Global interpretability methods explain which patterns in the data, i.e., class features, contributed the most to the model's prediction. These explanations can reveal critical reasoning about what the model is learning. On the other hand, local interpretability methods seek to explain why a model performs a specific prediction for a single input.

Post-hoc versus Intrinsic. This criterion distinguishes the methods with respect to whether the explanation mechanism lies in the internal architecture of the model (intrinsic) or if it is applied after the learning/development of the model (post-hoc). Post-hoc methods usually operate by perturbing parts of the data so that they can understand the contribution of different features in the model prediction or by analytically determining the contribution of different features to the model prediction. On the other hand, intrinsic models, also known as *in-model approaches* or *inherently interpretable models*, are self-explainable since they are designed to produce human-understandable representations from the internal model features.

Explanation Modality. Explanation modality refers to the type of explanation provided by each interpretability method. Among the reviewed methods, the explanation can be provided in the form of saliency maps (Explanation by Feature Attribution), semantic descriptions (Explanation by Text), similar examples (Explanation by Examples), or using high-level concepts (Explanation by Concepts). In Section 5, we used this categorization to discuss the reviewed methods.

3.2 Classical XAI Methods

The first attempts to explain deep learning models relied on the post-hoc analysis of the models. In spite of the criticism that post-hoc approaches have been recently subjected to [117], they are still being used in many domains of medical imaging, and their understanding is important to explain the advances in interpretable deep learning. As such, the following sections briefly describe the most popular XAI algorithms according to the two major categories of post-hoc analysis.

3.2.1 Perturbation-Based Methods. The rationale behind perturbation-based methods is to perceive how a perturbation in the input affects the model's prediction. Examples of perturbation-based methods are LIME [112] and SHAP [87].

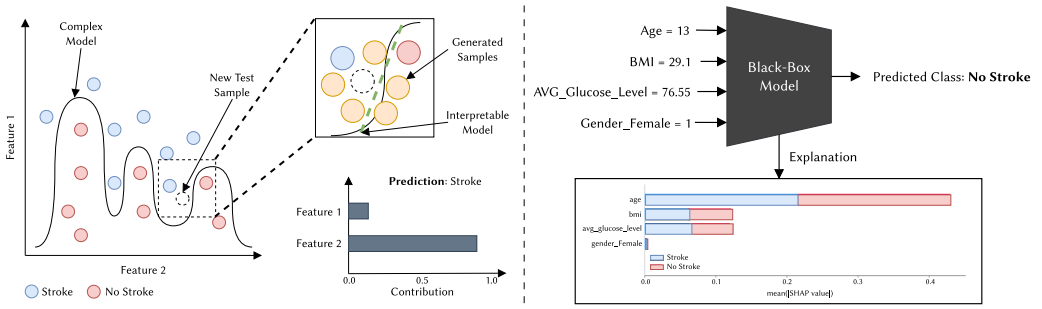
LIME. LIME [112] stands for Local Interpretable Model-agnostic Explanations. As the name suggests, it can explain any black-box model, and according to the XAI taxonomy is a post-hoc, model-agnostic method providing local explanations. The intuition behind LIME is to approximate the complex model (black-box model) locally with an interpretable model, usually denoted as a local surrogate model. Thus, an individual instance is explained locally using a simple interpretable model around the prediction, such as linear models or decision trees. Figure 2(a)) provides an intuitive illustration of the overall functioning of LIME.

In order to approximate the model prediction locally, a new dataset consisting of perturbed samples conditioned on their proximity to the instance being explained is used to fit the interpretable model. The labels for those perturbed samples are obtained through the complex model. In the case of tabular data, the perturbed instances are sampled around the instance being explained by randomly changing the feature values in order to obtain samples both in the vicinity and far away from the instance being explained. Analogously, when LIME is applied to the image classification problem, the image being explained is first segmented into superpixels, which are groups of pixels in the image sharing common characteristics, such as colour and intensity. Then, the perturbed versions of the original data are obtained by randomly masking out a subset of superpixels, resulting in an image with occluded patches. The new dataset used to fit the interpretable model consists of perturbed versions of the image being explained, and the superpixels with the highest positive coefficients in the interpretable model suggest that they largely contributed to the prediction. Thus, they will be selected as part of the interpretable representation that is simply a binary vector indicating the presence or absence of those superpixels.

SHAP. SHAP [87] was inspired by the Shapley values from the cooperative game theory [129] and operates by determining the average contribution of a feature value to the model prediction using all combinations of the features powerset. As an example, given the task of predicting the risk of stroke based on age, gender, and Body Mass Index (BMI), the SHAP explanations for a particular prediction are given in terms of the contribution of each feature. This contribution is determined from the change observed in model prediction when using the 2^n combinations from the features powerset, where the missing features are replaced by random values. Figure 2(b)) illustrates the above-described example. Similar to LIME, SHAP is a local model-agnostic interpretation method that can be applied to both tabular and image data. In the case of tabular data, the explanation is given in the form of importance values to each feature. In the case of image data, it follows a similar procedure to the LIME by calculating the Shapley values for all possible combinations between superpixels. Several variations of SHAP method were proposed to approximate Shapley values in a more efficient way: KernelSHAP, DeepSHAP, and TreeSHAP [86].

3.2.2 Saliency. Saliency maps are one of the most popular techniques to explain the decisions of a model. Saliency methods produce visual explanation maps representing the importance of image pixels to the model classification.

Class Activation Mapping (CAM) [176] is a seminal saliency method that allows the generation of a saliency map using a linear combination of the output of the last Global Average Pooling (GAP) layer of the network. Despite being a seminal contribution, CAM can only be applied to architectures following a specific pattern. To address this problem, Selvaraju et al. [127] proposed the Gradient-weighted Class Activation Mapping (Grad-CAM) [127] that uses the gradient



(a) (b)

Fig. 2. (a) **LIME**. The black curved line represents a decision boundary learned by the complex black-box model. LIME explains a new test sample (dashed circle), by fitting an interpretable model (represented by a green dashed line) to the variations of the test sample (orange circles), which are generated by randomly perturbing the test sample features. The fitted model allows one to perceive the contribution of each feature for classifying that specific test sample. (b) **SHAP**. The predicted risk of stroke of a classification model for a 13-year-old female with a body mass index of 29.1 and an average glucose level of 76.55 was "No Stroke". As evidenced by the bar plot, which provides the Shapley values for each feature, "age" was the feature with a higher impact on the prediction of "No Stroke", followed by the BMI and average glucose level features.

information of the target class with respect to the input image to produce a class-discriminative localization map that acts as a visual explanation for the model's prediction. Grad-CAM is, therefore, a generalization of CAM. Alternatively, SmoothGrad [138] is another gradient-based explanation method whose core idea is to attenuate the noise of the explanations provided by gradient-based techniques. The rationale behind SmoothGrad is to sample multiple images from the input image by adding noise to it. Then, the sensitivity maps are computed for each sampled image. The final map is the average of the sensitivity maps.

The Integrated Gradients (IG) [143] is an attribution method that relies on generating a set of images between the baseline and the original image using linear interpolation. These interpolated images are minor changes in the feature space between the baseline and input image and consistently increase with each interpolated image's intensity. Calculating the gradients per feature (pixels) makes it possible to measure the correlation between changes to a feature and changes in the model's predictions. The pixels with a high score are the ones that contributed the most to the prediction. The Layer-wise Relevance Propagation (LRP) [9] is an alternative solution to the use of gradients, where the decision function is decomposed into the relevance score of each neuron in the network. The output is propagated backwards through the model to determine the relevance score of the input, allowing production of an important heatmap of image pixels.

The main advantage of saliency methods is the reduced computational cost when compared with perturbation-based methods. However, different studies argue that the explanations provided by gradient-based methods can be ambiguous and unreliable as well as sensitive to adversarial perturbations [2, 39].

4 DATASETS

Among the reviewed literature, a set of 25 publicly available medical imaging datasets were considered based on the reviewed papers to provide a thorough overview of the existing medical imaging databases. Table 3 presents the main characteristics of the selected datasets, grouped by image type.

4.1 Chest X-ray

With respect to X-ray imaging modality, IU Chest X-ray [31], Chest X-ray14 [158], CheXpert [53], MIMIC-CXR [55], PadChest [18], COVIDx [155], and VinDr-CXR [101] datasets pertain to the chest anatomical region. IU Chest X-Ray, MIMIC-CXR, and PadChest datasets include free-text radiology reports. The reports are written in English except for the PadChest dataset, whose reports are written in Spanish. MIMIC-CXR and CheXpert are the largest databases, composed of 377, 110, and 224, 316 radiographs, respectively. CheXpert does not provide the raw free-text reports, but it provides an automated rule-based labeller for extracting keywords from medical reports conforming to the Fleischner Society's recommended glossary [48]. This tool was also used in the MIMIC-CXR dataset to extract the labels from the radiology reports. UI Chest X-Ray [31] comprises 7, 470 chest X-ray images jointly with 3, 955 free-text reports, being the most used dataset in the literature. The VineDr-CXR consists of 18, 000 chest X-ray images annotated with 22 findings (local labels) and 6 diagnosis (global labels). The local labels are inferred from the "findings" section in the radiology reports. In contrast, the global labels come from the "impressions" section and indicate suspected diseases, such as "Pneumonia", "Tuberculosis", "Lung Tumor", "Chronic obstructive pulmonary disease", "Other diseases", and "No Findings". Additionally, each "finding" is annotated on the X-ray image with a bounding box. Finally, the COVIDx [155] dataset comprises 13,975 chest X-ray images across 13, 870 patient cases, categorized as "Normal", "Pneumonia" and "COVID-19" cases.

It is worth noting that the majority of the chest X-ray dataset's labels were extracted using an automatic rule-based labeler, such as the CheXpert NLP tool [53]. However, relying on these automated tools can pose several issues concerning the quality of the labels. Consequently, the authors of the VinDr-CXR [101] dataset provided only radiologist-level annotations in both training and test sets.

4.2 Other X-ray Modalities

In the same segment of the X-ray imaging, and similar to the VinDr-CXR [101] dataset approach, VinDr-SpinalXR [102] is a recent dataset comprising 10, 469 spine X-ray images manually annotated by an experienced radiologist with bounding boxes around abnormal findings in 13 categories. The Knee Osteoarthritis [100] dataset contains 8,894 knee X-rays for both knee joint anomaly detection and knee Kellgren and Lawrence grading [58], whose value ranges from 0 to 4, according to the level of severity. Regarding the mammography datasets, Inbreast [99] consists of 410 mammography X-rays along with 115 radiology reports written in Portuguese. Similarly, the CBIS-DDSM [72] dataset is composed of 2, 620 mammography scans categorized as "Normal", "Benign", and "Malignant" cases.

4.3 Dermatoscopy

In the scope of dermatology, the ISIC 2020 [116] dataset consists of 33, 126 skin lesion images of different categorizations (malignant, melanoma, keratosis, etc.). It is part of the International Skin Imaging Collaboration, which promotes annual challenges to enhance the diagnosis of malignant skin lesions in dermoscopy images. The HAM10000 [148] dataset consists of 10, 015 dermatoscopic images categorized as pigmented lesions. The PH^2 [94] dataset comprises 200 dermoscopic images of melanocytic lesions, including common nevi, atypical nevi, and melanoma. Moreover, the PH^2 database includes medical segmentation of the lesions, clinical and histological diagnosis, and the assessment of several dermoscopic criteria. Derm7pt [56] is another dermoscopic image dataset with over 2, 000 images annotated following the 7-point melanoma checklist criteria. Finally, the SKINL2 [30] database consists of a total of 376 light fields of skin lesions manually annotated

under 8 categories based on the type of skin lesion and using the correspondent International Classification of Diseases (ICD) code. SkinCon [28] includes 3,230 images from the Fitzpatrick 17k skin disease dataset [44] annotated with 48 clinical concepts.

4.4 Microscopy

With regard to datasets composed of microscopy images, the BreakHis [140] dataset comprises 9, 109 microscopic images of breast tumor tissue distributed in various magnifying factors (40x, 100x, 200x, and 400x) and categorized into "benign" and "malignant" tumors. The Camelyon 17 [78] dataset consists of 1, 000 annotated whole-slide images (WSI) of lymph nodes. It is part of a challenge whose primary goal is the classification of breast cancer metastases in WSI of histological lymph node sections. Similarly, the Databiox [16] dataset comprises 922 histopathological microscopy images with 4 levels of magnification (4x, 10x, 20x, and 40x) for the task of Invasive Ductal Carcinoma (IDC) grading using 3 grades of IDC.

4.5 Others

For blindness detection purposes, the APTOS [139] dataset provides 5, 590 images of the retina taken using fundus photography. A clinician annotated each image according to the severity of diabetic retinopathy on a scale of 0 (Diabetic Retinopathy) to 4 (Proliferative Diabetic Retinopathy). In the scope of COVID-19, COVID-19 CT Reports (COV-CTR) [74] is a dataset composed of 728 CT scans paired with Chinese and English reports.

Regarding databases with multiple imaging modalities, the Pathology Education Informational Resource (PEIR) is a multidisciplinary public access image database intended for medical education. The PEIR database consists of 33, 648 images and the respective descriptions from different sub-classes of PEIR albums (PEIR Pathology, PEIR Radiology, and PEIR Slice). Similarly, the Radiology Objects in COntext (ROCO) [106] dataset is a multimodal image dataset consisting of 81, 825 radiology images divided into CT, MRI, X-ray Ultrasound, and Mammography. Each image is accompanied by the corresponding caption and the annotated Unified Medical Language System (UMLS) concepts.

4.6 Discussion

According to Table 2, Chest X-ray is the most popular imaging modality regarding simultaneously the number of datasets and their scale. Large-scale datasets have the advantage that they can be used to train a model from scratch. However, the automatic labeling process of some examples could compromise the reliability of the model since some class labels are not human verified and can be mislabeled. On the other hand, as evidenced by the number of images per dataset, the scarcity of large-scale datasets concerning other imaging modalities is evident, hindering the emergence of specialized task-specific models due to the limited training data. The majority of the medical imaging datasets have been gathered primarily for use in classification or segmentation tasks, where the labeling of class and mask annotations are sufficient for their intended purpose. Nevertheless, when these sets are utilized for interpretability purposes, their annotations may prove inadequate for a comprehensive qualitative evaluation. Nonetheless, some datasets, such as SkinCon and others, have been created explicitly with interpretability in mind and contain appropriate annotations to facilitate such evaluations. Accordingly, it is particularly important that researchers consider interpretability issues when building novel medical datasets, ensuring that appropriate annotations are included to further enable comprehensive and informative analyses of the data.

Table 2. Medical Imaging Datasets

Dataset	Image Type	Year	No. Images	Notes
IU Chest X-Ray [31]	Chest X-ray	2016	7,470	Includes reports
Chest X-Ray14 [158]	Chest X-ray	2017	112,120	Multiple labels
CheXpert [53]	Chest X-ray	2019	224,316	Multiple labels
MIMIC-CXR [55]	Chest X-ray	2019	377,110	Includes reports
PadChest* [18]	Chest X-ray	2020	160,868	Includes reports
VinDr-CXR [101]	Chest X-ray	2020	18,000	Multiple labels
COVIDx [155]	Chest X-ray	2020	13,975	Multiclass
Inbreast** [99]	Mammography X-ray	2012	410	Includes reports
CBIS-DDSM [72]	Mammography X-ray	2017	2,620	Multiclass
VinDr-SpineXR [102]	Spinal X-ray	2021	10,469	Multiple labels/BBox
Knee Osteoarthritis [100]	Knee X-ray	2006	8,894	Multiclass
PH^2 [94]	Dermatoscopic Images	2013	200	Multiple labels/Lesion segment
HAM10000 [148]	Dermatoscopic Images	2018	10,015	Multiclass/Lesion segment
SKINL2 [30]	Dermatoscopic Images	2019	376	Multiclass
Derm7pt [56]	Dermatoscopic Images	2019	2,000	Multiclass
ISIC 2020 [116]	Dermatoscopic Images	2020	33,126	Multiple labels
SkinCon [28]	Dermatoscopic Images	2022	3,230	Concept annotations
BreakHis [140]	Microscopy Images	2015	9,109	Multiclass
Camelyon17 [78]	Microscopy Images	2018	1,000	Multiclass
Databiox [16]	Microscopy Images	2020	922	Multiclass
$BCIDR^{(priv)}$ [175]	Microscopy Images	2017	5,000	Includes reports
APTOS [139]	Retina Images	2019	5,590	Multiclass
LIDC-IDRI [7]	CT scans	2011	1,018	Includes annotations
COV-CTR [74]	CT scans	2023	728	Includes reports
PEIR [54]	Photographs	2017	33,648	Includes reports
ROCO [106]	Multimodal	2018	81,825	Includes reports/UMLS Concepts

The datasets marked with "*" have reports written in Spanish. The datasets marked with "**" have reports written in Portuguese.

5 XAI METHODS IN MEDICAL DIAGNOSIS

As aforementioned, deep learning models must confer transparency and trustworthiness when deployed in real-world scenarios. This requirement becomes particularly relevant in clinical practice, where a less informed decision can put patients' lives at risk. Among the reviewed literature, several methods have been proposed to confer interpretability in the deep learning methods applied to medical diagnosis. The following sections summarize and categorize the most relevant works in the scope of explainable models applied to medical diagnosis. We give particular attention to the inherently interpretable neural networks and their applicability to medical imaging. We categorize the methods according to the explanation modality: (i) Explanation by Feature Attribution, (ii) Explanation by Text, (iii) Explanation by Examples, (iv) Explanation by Concepts, and (v) Other Approaches, inspired by the taxonomy presented in [96]. The list of the reviewed methods categorized by the interpretability method employed, image modality, and the dataset is provided in Table 5 in Appendix A.2.[1]

5.1 Explanation by Feature Attribution

Feature attribution methods indicate how much each input feature contributed to the final model prediction. These methods can work on tabular data or image data by depicting feature importance scores in a bar chart or using a saliency map, respectively. Among the existing feature attribution approaches in the literature, we categorize them into (i) perturbation-based methods and (ii) saliency methods, emphasizing their application to medical image analysis. A schematic diagram illustrating the general pipeline of these methods is shown in Figure 3.

[1] An interactive version of Table 5 is provided at this link

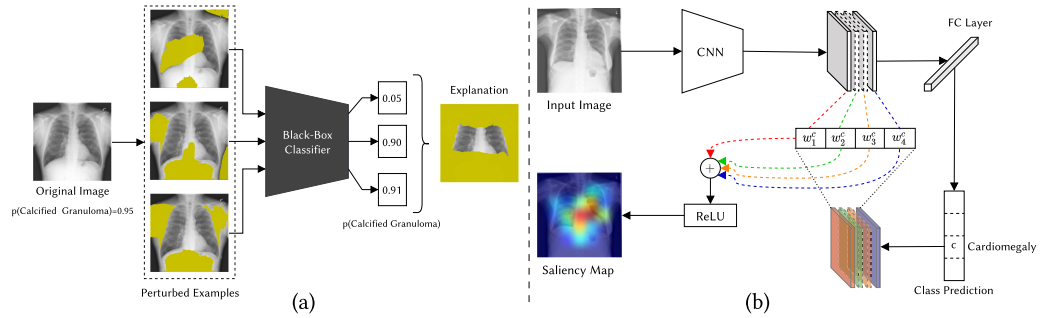


Fig. 3. (a) **Perturbation-based methods**. The input image is randomly perturbed by turning on and off certain pixels, resulting in an image with occluded parts (*Perturbed Examples* in the figure). Then, the perturbed image is fed to the classification model and the prediction confidence is exploited to determine the regions that contributed positively to the class prediction. Those regions will be considered to obtain the final explanation map (*Explanation* in the figure). (b) **Saliency methods**. The input image is fed to the classification model to obtain a class prediction. Then, the gradient is calculated for the score of the class concerning the feature maps of the last convolutional layer. After calculating the importance of the feature map regarding the predicted class, they are weighted with each of respective weight, followed by a ReLU operation to obtain the final saliency map.

5.1.1 Perturbation-Based Methods. As previously described in Section 3.2.1, perturbation-based methods aim to perform a modification in the input data to perceive how it affects the model's prediction. Popular examples of perturbation-based methods are LIME [112] and SHAP [87]. Regarding the use of perturbation methods to explain the prediction of medical diagnosis algorithms, Malhi et al. [91] applied the LIME method to explain the decisions of a classifier to detect bleeding in gastral endoscopy images. Similarly, Punn et al. [109] applied the LIME technique to explain the predictions of various state-of-the-art deep learning models used to classify pulmonary diseases in chest X-ray images. Magesh et al. [89] also used LIME to justify the decisions of a Parkinson's disease classifier model. In the context of melanoma detection, Young et al. [170] used the Kernel SHAP [87] interpretability method to investigate its reliability in providing explanations for a melanoma classifier from dermoscopy images. They concluded that the interpretability strategy highlighted features irrelevant to the final prediction. The authors also conducted sanity checks on the interpretability methods, which confirmed that these methods often produce different explanations for models with similar performance. This can be explained by the fact that the model can learn some spurious correlations, causing interpretability methods to give exaggerated importance to those spurious regions highlighted on the produced saliency maps. In the same context, Wang et al. [157] proposed a multimodal CNN for skin lesion diagnosis, which considered patient metadata and the skin lesion images. To analyze the contribution of each feature regarding the patient metadata, they adopted SHAP. Similarly, Eitel et al. [34] relied on an occlusion-based interpretability method [173] and investigated its robustness for the task of Alzheimer's disease classification.

RISE was proposed by Petsiuk et al. [107], consisting also of a post-hoc model-agnostic method for explaining the predictions of a black-box model through the generation of a saliency map indicating the important pixels to the model's prediction. The core idea is to probe the model with a set of perturbed images from the input image via random masking to perceive the model's response as important regions of the image are randomly occluded. The final saliency map is generated as a linear combination of the generated masks weighted with the output probabilities predicted by the model. For evaluation, the authors proposed two novel metrics, deletion and insertion, based on

the removal and insertion, respectively, of important pixels in the image to perceive the increase or the decrease of the model's performance.

5.1.2 Saliency Methods. Saliency methods allow production of a saliency map in which each pixel is assigned a value that represents its relevance to the prediction of a certain class. Popular techniques are CAM [176], Grad-CAM [127], DeepLIFT [130], and Integrated Gradients [143].

Rajpurkar et al. [110] proposed the CheXNeXt model to detect pulmonary pathologies and used CAM to identify the locations on the chest radiograph that contributed most to the final model prediction.

In the context of detecting COVID-19 from chest radiographs, Lin and Lee [76] used Grad-CAM and Guided Grad-CAM to investigate the regions that the model considered more discriminative. The produced heatmaps showed that when no preprocessing is used, the CNN tends to concentrate on non-lung areas (e.g., spine, heart, background) deemed irrelevant for the classification decision. However, when a masking process is used to highlight only the lung's area, the produced heatmaps highlight only the relevant regions since the CNN attention is limited to the critical area for detecting pulmonary diseases (lung area). Following the same procedures, Lopatina et al. [83] used the DeepLIFT attribution algorithm to investigate the decisions of a multiple sclerosis classification model, and Sayres et al. [123] used Integrated Gradients to provide explanations for the task of predicting diabetic retinopathy from retinal fundus images.

In contrast to the previous approaches, Rio-Torto et al. [114] proposed an in-model joint architecture composed of an explainer and a classifier to produce visual explanations for the predicted class labels. The explainer consists of an encoder-decoder network based on U-Net, and the classifier is based on VGG-16. Since the classifier is trained using the explanations provided by the explainer, the classifier focuses only on the relevant regions of the image containing the class. The qualitative assessment of the provided explanations was carried out by using state-of-the-art explainability methods provided by the Captum library [64]. A quantitative analysis was also provided in terms of accuracy, average precision, AUROC, AOPC [120], and the proposed POMPOM metric.

5.1.3 Discussion. Despite the simplicity of the feature attribution methods and their applicability to a wide range of approaches, these methods may often produce ambiguous explanations, making their qualitative evaluation difficult. Furthermore, preprocessing techniques were required for some methods to generate more plausible explanations [76]. Thus, researchers began exploring other modalities, such as textual explanations. It was discovered that textual explanations were indeed valid explanations and, in some cases, preferred over visual explanations since they are inherently understandable by humans [151].

5.2 Explanation by Text

The use of semantic descriptions became another way of explaining the model decisions since most of the clinicians prefer textual explanations compared with visual explanations only, and the combination of textual and visual explanations over either alone [38]. In general, providing textual explanations can be built on three paradigms: (i) image captioning, (ii) image captioning with visual explanation, and (iii) concept attribution [151]. Figure 4 depicts the general scheme adopted by most works to generate a textual description based on the visual features of the input image.

5.2.1 Image Captioning. The task of generating a textual description for explaining a model decision can be viewed as an extension of image captioning, commonly accomplished with Natural Language Processing (NLP). Indeed, the vast majority of works that aim to generate a textual description for a given input image follow the classical strategy of combining a CNN for extracting the visual features with a recurrent neural network (RNN), e.g., long short-term memory (LSTM),

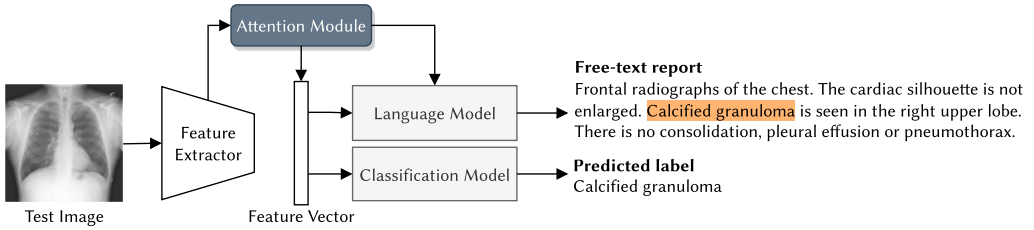


Fig. 4. Explanation by textual descriptions. The typical architecture for obtaining a textual description from image data combines an image embedding model (e.g., CNN) for extracting the features from the input image and a language model (e.g., long short-term memory [LSTM]) for generating the word sentences. The attention module can be inserted between those two models to guide the language model to focus only on relevant regions of the input image to improve the generation of the word sentences.

to generate the word sentences. Based on this paradigm, Sun et al. [142] developed a joint framework to generate sequences of words to provide a textual explanation for the task of diagnosing malignant tumors from breast mammography. Similarly, Singh et al. [136] built on an encoder-decoder framework composed of a CNN and a stacked LSTM for automatically generating radiology reports from chest X-rays. Regarding image captioning with visual explanation, Zhang et al. [175] proposed a multimodal approach, dubbed MDNet, composed of an image-embedding model and a language model, that can generate diagnostic reports, retrieve images by symptom descriptions, and visualize network attention. The MDNet model was evaluated on a dataset (BCIDR) containing histopathological images of bladder cancer. MDNet inspired several approaches. For example, Jing et al. [54] proposed a multi-task learning framework with a co-attention mechanism to guide the generation of text according to the localized regions containing abnormalities. They showed that a hierarchical LSTM model performs better in generating long text reports. Subsequently, Wang et al. [159] proposed TieNet, which makes use of attention modules to extract the most important information from chest X-ray images and use their diagnostic reports in order to guide the model to produce more coherent reports. Similarly, Barata et al. [11] proposed a hierarchical classification model that uses attention modules, including channel and spatial attention, to identify relevant regions in the skin lesions and subsequently guide further the LSTM attending at different locations whilst conferring more transparency to the network. Lee et al. [71] also explained the decisions of a breast masses classifier using both visual and textual explanations based on a CNN-RNN architecture. In the same fashion, Gale et al. [38] proposed a model-agnostic interpretable method based on an RNN to produce textual explanations for the decisions of deep learning classifiers. Furthermore, they developed a visual attention mechanism in charge of highlighting the relevant regions for classifying hip fractures in pelvic X-rays. Yin et al. [169] also used attention mechanisms to attend to the regions at sentence level. They proposed the Hierarchical Recurrent Neural Network (HRNN) model, composed of two-level LSTMs: a word RNN and a sentence RNN. The sentence RNN produces the topic vectors whereas the word RNN receives the output of the sentence RNN and infers the words that constitute the final report. Moreover, they introduced a matching mechanism to map the topic vectors and the sentences into a jointly semantic space that minimizes a contrastive loss. Similar to the work of Yin et al. [169], the model proposed by Liu et al. [82] generates topics from images and then completes sentences from these topics. When compared with [169], this work allows the generation of more coherent report generation due to the use of a fine-tuning process that uses reinforcement learning via CIDEr.

A more disruptive approach was introduced by Li et al. [73] that proposed the Hybrid Retrieval-Generation Reinforced Agent (HRGR-Agent) consisting of a retrieval policy module and a

generation module. The retrieval policy module is responsible for deciding whether sentences are obtained from a generation module or retrieved from the template database, which is composed of a set of template sentences. Moreover, the retrieval policy and generation modules are updated via reinforcement learning, guided by sentence-level and word-level rewards using the CIDEr.

In contrast to the previous approaches, Chen et al. [24] exploited the Transformer [153] architecture, in which they incorporated two memory modules into the decoder. These modules are responsible for memorizing textual patterns and assisting the decoder of the Transformer in generating radiology reports containing relevant information associated with chest X-ray images. The same tendency is reflected in recent works [80, 81, 160, 161, 168], which employ Transformer-based models with additional custom modules to better capture the relevant features of input images, leading to an improved performance in the task of radiological report generation. Recently, Selivanov et al. [126] introduced a novel image captioning architecture that combines two language models, incorporating image-attention (SAT) [165] and text-attention (GPT-3) [17], resulting in an outstanding performance compared with the previous methods. For a more comprehensive analysis of the use of Transformers in medical imaging, we refer the reader to the survey of Shamshad et al. [128].

5.2.2 Concept Attribution. The idea behind concept-based attribution is learning human-defined concepts from the internal activations of a CNN. The use of concepts to provide global explanations was proposed by Kim et al. [59] with the introduction of the Concept Activation Vectors (CAVs), which provide explanations in terms of human-understandable concepts that are typically related to parts of the image. Kim et al. also proposed Testing with CAVs (TCAV), which enables quantification of the importance of a user-defined concept to a classification result. Figure 5 illustrates the typical pipeline of concept-based attribution methods. In the context of medical imaging, the term "microaneurysm" can be viewed as a concept possible of being identified by humans in fundus imaging, and which denotes the presence of diabetic retinopathy [151]. In the same line of research, Graziani et al. [43] proposed a framework for concept-based attribution to generate explanations for CNN decisions of a breast histopathology classifier. They built on TCAV by incorporating Regression Concept Vectors (RCVs), which provide continuous-values measures of a concept instead of solely indicating its presence or absence. This is particularly useful in the medical domain since a value indicating, for example, tumor size is more informative than a binary value indicating its presence or absence. Graziani et al. also concluded that the learning of concepts by an intermediate layer of a CNN could be improved by removing spatial dependencies of the convolutional layers and introducing L2 norm regularization in the regression problem. Recently, Lucieri et al. [84] introduced ExAID, a framework that provides multimodal concept-based explanations for the task of melanoma classification. Their framework relied on CAVs for concept identification and used the TCAV method to estimate the influence of a specific concept on the decision. The authors provided textual explanations in the form of template phrases by only replacing the identified concepts and their importance to the prediction in the phrase structure. In order to localize the regions of the learned concepts in the latent space of the trained classifier, the authors used Concept Localization Maps [85], which use perturbation-based concept localization to generate a saliency map highlighting the relevant regions with respect to the learned concepts, thus providing visual explanations.

5.2.3 Discussion. In summary, except for the work of Chen et al. [24], all the explanation text-based approaches mentioned above rely on RNN architectures to generate text descriptions towards providing a more human-interpretable explanation for a model decision. However, as stated by Pascanu et al. [104], RNN-based approaches, such as LSTM, have some limitations in generating long text reports. On the contrary, concept-based attribution methods provide a more

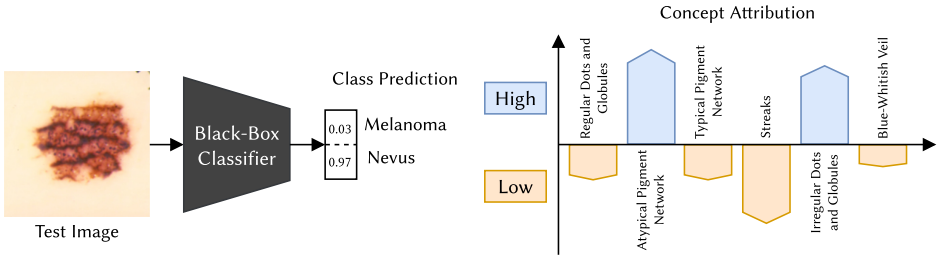


Fig. 5. Explanation by concept attribution. In the first phase, the human-defined concepts (Regular Dots and Globules, Atypical Pigment Network, Typical Pigment Network, Streaks, Irregular Dots and Globules, and Blue-Whitish Veil) are modeled as numeric features, following the CAV technique. Then, after the input image passes through a classifier model, the class prediction is globally explained based on the importance of each concept to the final prediction.

objective and human-understandable way of interpreting classification decisions. However, the main limitation of these methods is the need for manual annotations of the concept examples, which may be impractical for specific medical image modalities, increasing the need for involving clinicians in the annotation tasks.

5.3 Explanation by Examples

The type of methods that explain a model decision by selecting a set of similar examples are dubbed *example-based explanation methods*. Apart from explaining algorithm decisions, this strategy is also commonly used between clinicians to explain the rationale behind their decision process. Below, we categorize example-based explanation methods as follows: (i) Case-Based Reasoning, (ii) Counterfactual Explanations, and (iii) Prototypes.

5.3.1 Case-Based Reasoning. Case-Based Reasoning (CBR) and Content-Based Image Retrieval (CBIR) are example-based explanation methods that aim to search a database for visually similar entries to a specific query image. The general scheme for implementing a CBR system using DNNs is depicted in Figure 6(a). Although the idea of using CBIR systems in clinical settings is not novel [3], there has been renewed interest in CBIR approaches as a way to provide explainability to deep-learning methods for medical diagnosis [5, 25, 47].

Recently, Barnett et al. [13] introduced a novel interpretable AI algorithm (IAIA-BL) for classifying breast masses using CBR. The model provided both a prediction of malignancy and its explanation by using known medical features (mass margins). Given an image region to analyze, the algorithm compared that region with a set of previous similar cases (image patches) using Euclidean distance. The similarity score was then used to provide the mass margin scores for each medical feature. Those scores were then used to predict the malignancy score (benign or malignant). The model was trained using a fine-annotation loss penalizing activations of medically irrelevant regions on the data. The authors also introduced an interpretable evaluation metric, Activation Precision, to quantify the proportion of relevant information from the "relevant region" used to classify the mass margin regarding the radiologist annotations. The experimental results showed that the IAIA-BL achieved comparable performance to black-box models.

A different approach was presented by Tschandl et al. [147], in which they compared the predictions of the ResNet-50 softmax classifier with the diagnostic accuracy obtained by using CBIR. Contrary to Barnett et al. [13], Tschandl et al. measured the cosine similarity between two feature vectors to retrieve the most similar images to the image query. The results showed that the diagnostic accuracy obtained through CBIR is comparable to the performance of a softmax classifier,

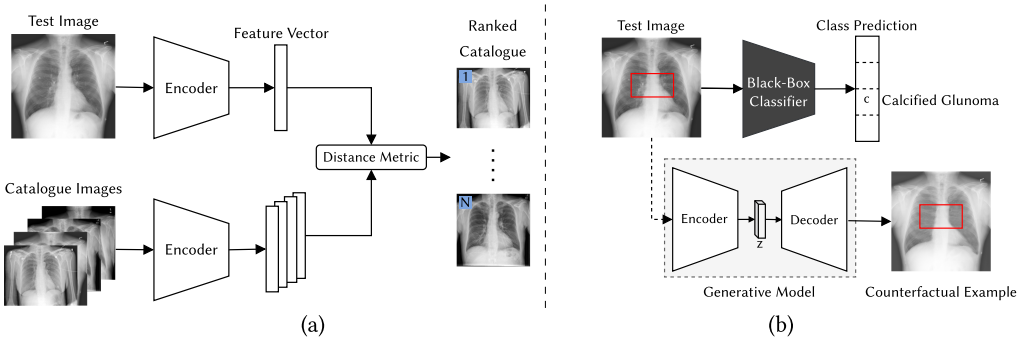


Fig. 6. (a) **Explanation by Case-Based Reasoning**. The feature vector corresponding to the input image is compared against the feature vectors of the images in the catalogue using a distance metric, such as L2 distance. Finally, the images are retrieved from the catalogue ranked by their similarity to the input image. (b) **Explanation by Counterfactual Examples**. To explain the prediction made by a classifier, the input image is modified in a controlled way, typically by using a generative model (e.g., generative adversarial network [GAN]) to shift the original class, i.e., from normal to abnormal or vice-versa. Thus, the counterfactual example intends to explain the prediction by showing that the image was classified as "abnormal" because it is not "normal" as perceived by the absence of tissue inflammation (white spots) in the generated counterfactual example.

leading Tschandl et al. to claim that CBIR can replace traditional softmax classifiers to improve diagnostic interpretability in a clinical workflow.

A more recent approach was introduced by Barata and Santiago [12], in which CBIR was applied to explain the decisions of a CNN model for skin cancer diagnosis. When compared with the work of Tschandl et al., Barata and Santiago implemented an augmented category-cross entropy loss function composed of three regularization losses: the triplet loss, the contrastive loss, and the distillation loss. These losses encourage the model to learn a more structured feature space. The experimental results on the ISIC 2018 [148] dermoscopy dataset confirmed that the combination of different loss functions leads to more structured feature spaces, which improves the performance of the classification model. Lamy et al. [68] proposed an explainable CBR system with a visual interface, combining quantitative and qualitative approaches. In contrast to the above-discussed works, it used numerical data and provided a user study in which clinicians validated their approach.

The recent work of Hu et al. [52] introduced the eXplainable Medical Image Retrieval (X-MIR) approach, which explored the use of similarity-based saliency maps to explain the retrieved images visually. Concretely, they adapted the saliency map generation process for the problem of image retrieval through the use of a similarity-based formulation. These similarity-based saliency methods take as input a retrieval image and a query image for producing a saliency map highlighting the most similar regions of the retrieval image to the query image. To evaluate the quality of the generated saliency maps, the authors adapted two causal metrics, deletion and insertion (refer to Section 6.1 for a detailed description), to measure the decrease or increase, respectively, in image similarity score as the retrieved image is gradually perturbed based on the important regions of its saliency map. The authors evaluated their approach on two medical datasets: COVIDx [155] and ISIC 2017. They found that for both cases, the generated saliency maps focused on relevant regions when retrieved images were correct and observed the contrary when the retrieved images were incorrect. Finally, the authors pointed out for the importance of conducting user studies with clinicians to validate the utility of their approach. Silva et al. [131] also explored the medical image retrieval with the addition of saliency maps to improve the class-consistency of top retrieved

results while enhancing the interpretability of the whole system by accompanying the retrieval with visual explanations.

In order to assess the effectiveness of using a CBIR system as an auxiliary tool for classifying skin lesions through dermatology images, a user-centered study was done by Sadeghi et al. [118]. Sixteen non-expert users were invited to classify skin lesion images among four categories (Nevus, Seborrheic Keratosis, Basal Cell Carcinoma, and Malignant Melanoma) based on two conditions: using CBIR and without using CBIR. The results indicated that CBIR enabled users to make a significantly more accurate classification on a new skin lesion image. These findings suggest that CBIR can indeed help clinicians understand model decisions as well as allow less experienced practitioners to improve their skills. Thus, CBIR systems can have a significant clinical application value as a decision-support tool to accelerate the diagnosis of pathologies.

5.3.2 Counterfactual Explanations. Counterfactual explanations are based on the principle that "an action on the input data will cause an outcome" [96]. The idea is to perturb the input data in a controlled way in order to reverse the final model prediction, being the modified input the counterfactual example, as illustrated in the diagram of Figure 6(b). Furthermore, counterfactual explanations are deemed human-interpretable and post-hoc, meaning that they do not require access to model internals.

The work of Schutte et al. [125] constitutes a new approach in the way of interpreting the predictions made by deep-learning models by using generative models to produce a sequence of images depicting the evolution of a pathology. Through the sequence of images produced, a human can understand which biomarkers triggered the prediction made by the model. Concretely, the proposed method aims to identify the optimal direction in latent space to produce a series of synthetic images with minor modifications leading to different model predictions. By observing these modified synthetic versions of the original image, it is expected that a human can perceive the features that caused the model prediction. Experimental results on two medical image datasets showed that the proposed approach allows visualizing where the most relevant features are localized and how they contributed to the model prediction by analyzing the generated images. Moreover, this generative approach may be helpful for the identification of new biomarkers.

Kim et al. [61] proposed the Counterfactual Generative Network (CGN), which is able to generate counterfactual images to explain the predictions of a pneumonia classifier from chest X-ray images. To guide the CGN towards the generation of contrastive images from query images, the prediction of the classification network was manipulated to shift the original class to the negative class. The subtraction of the counterfactual image from the input image allows the generation of attribution maps evidencing the most relevant regions to the prediction.

In the same fashion, Singla et al. [137] used a conditional Generative Adversarial Network (cGAN) to produce a set of counterfactual images with changed posterior probability to explain the class predictions of a chest X-ray classifier. Additionally, the context from semantic segmentation and object detection was incorporated into the loss function to preserve subtle information about the medical images in the generated counterfactual images. The validity of the generated counterfactual explanations was assessed through the use of three evaluation metrics: (1) the Fréchet Instance Distance score to evaluate the visual quality of the counterfactual images, (2) the Counterfactual Validity score to quantify the class floppiness of the counterfactual images, and (3) the Foreign Object Preservation score to assess the presence of unique properties of patients in the generated explanations. Clinical measurements, cardiothoracic ratio and costophrenic recess, were adopted to demonstrate the utility of the explanations in terms of the clinical context.

Given the recent advances in the scope of image synthesis, the use of generative diffusion probabilistic models [49] to produce counterfactual explanations is an interesting future research

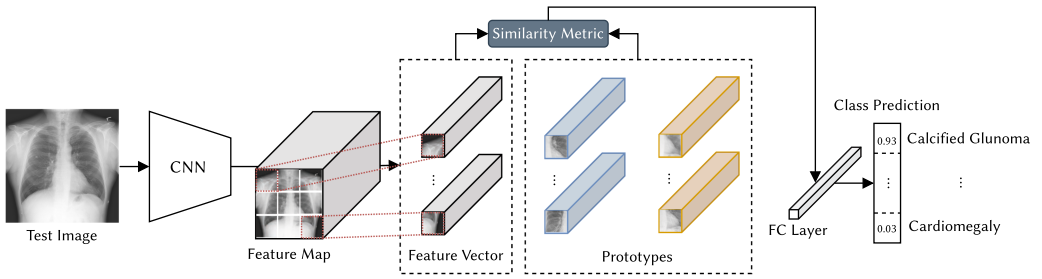


Fig. 7. Explanation by Prototypes. During the model training, a set of prototypes are learned to visually represent a certain class (represented with blue and orange colors in the figure). In the test phase, the features extracted from the test image are compared with a set of prototypes using a similarity metric, such as cosine similarity. Then, the final class prediction is based on the similarity scores computed between the prototypes and the different parts of the input image.

direction as it is under-explored in medical imaging. The benefit of using these models to generate counterfactual explanations is related to their ability to handle missing data and their robustness to distributional shifts [57]. A prominent example is the work of Sanchez et al. [121], which relied on conditional diffusion models for synthesizing healthy counterfactual examples of brain images, allowing segmentation of the lesion through the difference between the observed image and the healthy counterfactual.

5.3.3 Prototypes. While most research on interpretability is still oriented towards the use of post-hoc approaches, some authors have advocated the need for devising inherently interpretable models [117] to obtain explanations that are indeed interpretable by humans. The learning of prototypes during the training phase of the model is a common strategy in the development of inherently interpretable models. This idea was initially explored in [21], in which the authors incorporate a prototype layer at the end of the network [13, 21], named ProtoPNet, for bird species recognition. The rationale behind this approach is that different parts of the image act as class-representative prototypes during training. When a new image needs to be evaluated in the testing phase, the network finds the most similar prototypes to the parts of the test image. The final class prediction is based on a score computed with the similarities between the prototypes. Figure 7 illustrates the general pipeline for deriving a class prediction from the similarity score between different parts of the input image and a set of learned prototypes.

Based on ProtoPNet, Donnelly et al. [32] introduced the Deformable ProtoPNet. This prototypical case-based interpretable neural network provided spatially flexible deformable prototypes, i.e., prototypes that can change their relative position to detect semantically similar parts of an input image.

Despite the significance of ProtoPNet, Hoffmann et al. [50] investigated its shortcomings and proved that ProtoPNet could be susceptible to adversarial and compression noise and, thus, compromise the inner interpretability of the model. Although these limitations were not significant in the bird recognition problem, the picture changes in high-stake applications, such as the healthcare domain, in which the lack of robustness can have severe implications.

Regarding the applicability of these inherently interpretable networks to medical imaging, Kim et al. [60] proposed an interpretable diagnosis framework, dubbed XProtoNet, for chest radiography to learn disease representative features within a dynamic area using an occurrence map. Contrary to ProtoPNet, in XProtoNet the prototypes are class-representative and completely dynamic in terms of area, which is particularly important for accommodating the high variability

in size of discriminative regions of medical images. To produce an appropriate occurrence map, the authors introduced two regularization terms. The L1 loss forces the occurrence area to be small enough to avoid covering irrelevant regions, and the transformation loss approximates each occurrence map with a transformed version via an affine transformation that did not change the relative location of a disease pattern. The experiments on the NIH Chest X-ray dataset [158] confirmed that XProtoNet surpasses the state-of-the-art models in diagnosing chest diseases from X-ray images. Later, Singh et al. [134] introduced an interpretable deep learning model, named Generalized Prototypical Part Network (Gen-ProtoPNet), for detecting COVID-19 from X-ray images. Gen-ProtoPNet was inspired by the original ProtoPNet [21] and the NP-ProtoPNet [135]. Unlike ProtoPNet and NP-ProtoPNet, which use L2 distance to calculate the similarity between prototypes, Gen-ProtoPNet used a generalized version of the L2 distance, allowing the use of prototypical parts of any dimension, i.e., squared and rectangular spatial dimensions. Furthermore, the experiments on two COVID-19 chest X-ray datasets [27, 156] confirmed that using prototypical parts of spatial dimensions bigger than 1 x 1 improves performance of the model, specifically when using the VGG-16 model as the feature extractor.

5.3.4 Discussion. Relying on example-based strategies may be a more trustworthy option since the retrieved examples are plausible and tend to contain similar findings to the input query image. However, the performance of the example-based systems can be compromised if a significant number of examples per class is unavailable. This assumption is also valid in the case of prototype-based approaches, in which performance depends directly on the diversity and amount of class-representative prototypes. Regarding the counterfactual explanations, it is desirable to discover credible causal structures to create ground-truth explanations to improve further the modelling of the interventions made over the images [137].

5.4 Explanation by Concepts

The rationale behind concept-based learning approaches is using human-specified concepts as an intermediate step to derive the final predictions. This idea was used in the works of Kumar et al. [66] and Lampert et al. [67], with specific applications in few-shot learning approaches. In [67], the proposed model first estimated a set of attributes which were subsequently used to infer the final predictions. These types of model architectures were dubbed Concept Bottleneck Models (CBMs) in the work of Koh et al. [63]. In simple terms, these models relied on an encoder-decoder paradigm, in which the encoder is responsible for predicting the concepts given the raw input image, and the decoder leverages the predicted concepts by the encoder to make the final predictions. The encoder is typically a CNN model with a bottleneck layer inserted after the last convolutional layer, whereas the decoder can be a multi-layer perceptron to map the concepts to the final predictions. This pipeline is illustrated in Figure 8. The idea of CBMs can be applied to any model just by inserting the bottleneck layer after the final convolutional layer. However, the main disadvantage of these methods is that annotated concepts are required. Koh et al. [63] provided a systematic study on different ways to learn CBMs. Among the considered setup models, they concluded that the joint training is the preferred approach, which minimizes the weighted sum considering the classification loss and concept loss. The authors also stated that it is possible to intervene in the concept predictions to change the final output, which raises the question of to what extent it is feasible to revisit, for example, 100 concepts and perceive which concept would be modified to make the correct prediction. Later, Yuksekgonul et al. [172] introduced Post-hoc CBM (PCBM) to address some limitations of CBM, specifically the need for concept-level annotations. The authors claim that PCBM can convert any pre-trained model into a concept bottleneck model. When concept annotations are unavailable, PCBM can leverage concept examples from

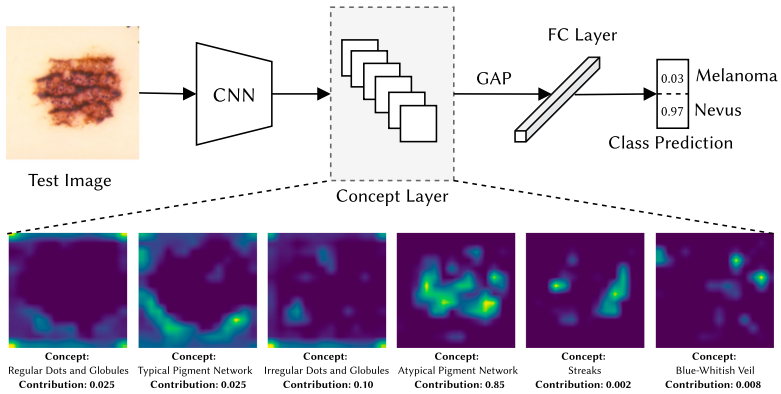


Fig. 8. Explanation by Concepts. In the first phase, the concept layer is trained to predict the concepts associated with the input image. Then, given a test image, the model first predicts the concepts presented in the image, which are subsequently processed by a fully connected layer to infer the final predictions. Simultaneously, it is possible to produce visualizations of the filters of the concept layer that highlight the relevant regions for each concept. Additionally, the contribution of each concept to the final prediction is obtained to perceive which concepts had more influence on the final decision.

other datasets and train linear binary classifiers to distinguish between examples of a single concept and negative examples. Yuksekgonul et al. [172] identified that the problem with CBMs is that they require concept-level annotations per image, which is expensive and difficult to obtain, particularly in the context of skin lesions. CAVs [59] can be adopted to mitigate this by automatically predicting the presence or absence of single concepts on unseen images.

A different approach was introduced by Chen et al. [23] in which they proposed the Concept Whitening (CW) module, which is inserted into a neural network and can replace the Batch Normalization layer so that each point in the latent space has an interpretation in terms of known concepts. In a similar line of research, the decision process was decomposed in a set of human-interpretable concepts, along with a visual interpretation of the spatial localization where these concepts are present in the image [105, 162]. In particular, Patrício et al. [105] proposed an approach for enforcing the visual coherence of concept activations by using a hard attention mechanism to guide the activations of concept filters towards the locations to which the concept is visually related. This strategy has been shown to improve the visual explanation of concept-based approaches for skin lesion diagnosis.

Ghorbani et al. [40] introduced ACE, which can automatically identify a set of high-level concepts in an unsupervised way. In a first phase, each image is segmented in multiple resolutions using the SLIC [1] algorithm. Then, the segments are clustered by similarity according to the Euclidean distance, which is measured in the latent space. Each group of segments represents a different concept, labelled as pseudo-concepts. Lastly, to retain only the important concepts in each group, the TCAV [59] importance score was computed. The work of Fang et al. [35] built on the rationale of the ACE. The authors proposed the Visual Concept Mining (VCM) method to explain the decisions of an infectious keratitis classifier based on human-interpretable concepts. VCM encompasses two stages: (i) the proposed Potential Concept Generator module is responsible for identifying relevant concepts based on the segmentation of image patches according to the relevant regions highlighted on the produced saliency maps; (ii) the visual concept extractor module learns the similarity and diversity among the segmented image parts and groups them according to the DeepCluster [20] algorithm. The authors of VCM claimed that the concepts learned by their

method were coherent with the medical annotations whilst being more diverse for the different classes, contrary to the ACE, which provides too broad concepts.

A disruptive approach was proposed by Sarkar et al. [122] that introduced an ante-hoc explainable model. A concept encoder on top of a backbone classification architecture is used for learning a set of human interpretable concepts, providing an explanation for the classifier predictions. Additionally, the output of the concept encoder is passed to a decoder that reconstructs the input image, encouraging the model to capture the semantic features of the input image. Despite the method only reporting results for generic datasets, it would be interesting to extend this work for medical imaging datasets. Recently, following the philosophy of CBM, Yan et al. [167] proposed a method to improve the trustworthiness of skin cancer diagnosis by allowing doctors to intervene in the decisions of the trained models based on their knowledge and expertise. This human-in-the-loop framework allows for discovering and removing potential confounding behaviors of the model (e.g., artifacts or bias) within the dataset during the training phase. They concluded that modifying the output of the predicted concepts leads to a more accurate model.

5.4.1 Discussion. Although concept-based explanation methods are under-explored in medical imaging, they constitute a promising way of providing human-understandable explanations. The main advantage of these methods is that they are interpretable by design since the final predictions are derived from the learned concepts. However, the dependency on manual annotation of these concepts is the major limitation in concept-learning approaches. Recently, to overcome the annotation dependency of the concepts, proposed methods built on unsupervised techniques to discover a set of pseudo-concepts related to the input image [35, 40].

5.5 Other Approaches

In contrast to the above-discussed approaches, some authors have investigated alternative strategies to confer interpretability to the models, including Bayesian Neural Networks (BNN) to quantify the uncertainty associated with the model prediction or using adversarial training to improve the quality of the generated explanations.

5.5.1 Bayesian Approaches. Despite the success of CNN architectures, it is infeasible to quantify their uncertainty given the deterministic nature associated with the internal parameters. Furthermore, CNNs are likely to overestimate the data when it is biased. In order to address these problems, Thiagarajan et al. [144] proposed the use of Bayesian CNNs (BCNNs), which allow for the uncertainty estimation associated with the predictions. In particular, the uncertainty associated with the predictions of an IDC classifier on breast histopathology images was quantified. The examples characterized by a high value of uncertainty were projected into a lower-dimensional space using the t-SNE [150] technique to facilitate data visualization and interpretation of the test data. The uncertainty allowed selecting the examples requiring human evaluation, which constitutes an interesting approach in the case of problems in the medical imaging domain. Similarly, Billah and Javed [15] relied on BCNNs to quantify the uncertainty associated to the predictions of a classifier model for the diagnosis of blood cancer. Recently, Gour and Jain [42] proposed the UA-ConvNet, an uncertainty-aware CNN to detect COVID-19 from chest X-ray images and provide an estimation of the model uncertainty. For this, they used Monte Carlo dropout [37] to obtain a probability distribution of the model prediction.

5.5.2 Adversarial Training. In adversarial training, examples of the training set are augmented with adversarial perturbations at each training loop, allowing the increase of the robustness of the model when provided with potential malicious examples [10].

The first attempt to use adversarial training to improve interpretability in a medical imaging diagnosis task was made by Margeloiu et al. [92]. They explored the use of adversarial training to improve the interpretability of CNNs, mainly when applied to diagnosing skin cancer. Specifically, the trained model was retrained from scratch using adversarial training with the Projected Gradient Descent (PGD) adversarial attack [88]. The experiments on the dermatology dataset HAM10000 [148] showed that saliency maps of the robust model are significantly sharper and visually more coherent than those obtained from the standard trained model.

However, further research is needed since adversarial training is under-explored in medical imaging interpretability. Specifically, applying the above-referred findings to other datasets and network architectures becomes necessary to perceive the generalization capability of the methods. Furthermore, as stated by the authors in [92], the proposed method is not ready to be deployed in real-world scenarios due to the sensitivity of saliency methods to training noise, which can cause those methods to assign importance to artifacts available in the image (e.g., dark regions and irrelevant medical regions). Therefore, it is crucial to understand the limitations of adversarial training to improve interpretability in medical imaging diagnosis tasks.

5.5.3 Discussion. The uncertainty estimation associated with a classifier's predictions is helpful in the clinical workflow since clinicians can support their decisions based on the uncertainty value. Additionally, the use of adversarial training can be viewed as a method to improve the robustness of the model to adversarial attacks. As also demonstrated by Margeloiu et al. [92], the produced explanations by adversarially trained models seem to be more coherent and sharpening. Despite the under-exploitation of these strategies in medical imaging, these preliminary findings may encourage the emergence of methods adopting these alternative strategies as an additional layer to improve the models' reliability and robustness.

6 EVALUATION METRICS

Depending on the type of explanation modality (visual or textual), there are different ways to assess the quality of the generated explanations. We divide the evaluation metrics used in the literature into two categories: (i) evaluation metrics to assess the quality of visual explanations and (ii) evaluation metrics for measuring the quality of textual explanations.

6.1 Evaluating the Quality of Visual Explanations

Evaluating the quality of model explanations remains an active area of research. A common approach for evaluating the model interpretability, specifically when applied to the medical domain, is to request an expert opinion from the clinicians and radiologists. However, this evaluation approach is time-consuming and depends on the level of experience of the clinicians [38, 149].

Therefore, there have been attempts to propose evaluation metrics capable of objectively assessing the quality of the explanations. Samek et al. [120] were precursors in contributing to the question of how to objectively evaluate the quality of heatmaps by introducing the area over the Most Relevant First (MoRF) perturbation curve (AOPC) measure. This measure is based on a region perturbation strategy that iteratively removes information from some regions of the input image according to its relevance to the class, allowing perception of the performance decay of the model. The conducted experiments showed that a large AOPC value denotes high model sensitivity to the perturbations, indicating that the heatmap is actually informative.

Inspired by the work in [120], Petsiuk et al. [107] proposed two causal metrics, namely deletion and insertion, to evaluate the produced explanations for a black-box model. The deletion metric measures the degradation of the class probability, as important pixels of the image, derived from

the saliency map, are removed. On the other hand, the insertion metric intends to measure the increase of the class probability, as pixels are inserted based on the generated saliency map.

Later, Hooker et al. [51] argued that the modification-based metrics introduced by Petsiuk et al. [107] might not capture the actual reasoning behind the model's degradation since this degradation could be due to artefacts introduced by the values used to replace the removed pixels. Thus, the authors proposed RemOve And Retrain (ROAR) to evaluate interpretability methods by verifying how the accuracy of a retrained model degrades as important features are removed. The most important features are removed in certain regions of the image with a fixed uninformative value for each interpretability method. The main drawback of this metric is the required retraining of the model, which is computationally expensive.

Recently, Rio-Torto et al. [114] proposed the POMPOM (Percentage of Meaningful Pixels Outside the Mask) metric, which determines the number of meaningful pixels outside the region of interest in relation to the total number of pixels, to evaluate the quality of a given explanation. Similarly, Barnett et al. [13] introduced the Activation Precision evaluation metric to quantify the proportion of relevant information from the "relevant region" used to classify the mass margin regarding the radiologist annotations. Despite the relevance of both metrics, they require manual annotations of the masks, which is time-consuming and may be difficult to obtain for some medical image datasets.

In spite of the valuable contribution of these proposed evaluation metrics, we believe that a new evaluation method for model interpretability can be developed with the aid of the BNN characteristics. Considering that the weights of the BNN follow a probability distribution, we can sample different "models" from the posterior distribution and generate an arbitrary number of explanations for a given example [19]. Then, using intersection or union operations over those generated explanations seems to be an interesting direction to estimate whether most of the explanations highlight the same Region of Interest (ROI).

6.2 Evaluating the Quality of Textual Explanations

In this category, the metrics are used to measure the quality of the generated text and originate from generic NLP tasks. The most used metrics in the reviewed papers were BLEU [103], ROUGE-L [75], METEOR [70], and CIDEr [154].

BLEU (Bilingual Evaluation Understudy) score is the most common evaluation metric in NLP. In simple terms, it compares n-gram matches between the generated sentence (also known as *candidate sentence*) and the ground-truth sentence (also known as *reference sentence*), expressed in modified precision[2] for each n-gram. The BLEU metric has N variations (BLEU-N), typically $N \in \{1, 2, 3, 4\}$, with respect to the considered n-grams. The value of BLEU score ranges from 0 to 1, which means that the closer it is to 1, the better the translation. Formally, BLEU score is calculated according to the following formulation:

$$BLEU = BP. \exp \sum_{n=1}^{N} w_n \log p_n, \tag{1}$$

where p_n is the modified precision for n-gram, w_n is a weight, ranging from 0 and 1, and $\sum_{n=1}^{N} w_n = 1$, i.e., if $N = 4$, $w_n = 1/N$, and BP is the brevity penalty that penalizes short generated sentences,

[2]Takes into consideration the maximum frequency of each n-gram in the reference sentence (clipped count). The modified precision is then calculated by summing the clipped counts of the candidate sentence divided by the total number of candidate n-grams.

denoted as

$$BP = \begin{cases} 1 & \text{if } c > r \\ \exp(1 - \frac{1}{c}) & \text{if } c \leq r \end{cases}, \tag{2}$$

where c is the length of candidate translation and r is the reference corpus length.

The ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metric is actually a set of metrics. We focus on the ROUGE-L variant as it was the prevalent metric in most of the reviewed methods. ROUGE-L measures the Longest Common Subsequence (LCS) between the generated sentence and the ground-truth sentence both in terms of precision and recall. This means that if both sentences share a long sub-sentence, the similarity between the two sentences is expected to be high. The final value of ROUGE-L is given in F1 score, as formally described in Equation (3):

$$ROUGE - L_{F1} = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}, \tag{3}$$

where $Precision = \frac{LCS(c,r)}{m}$ and $Recall = \frac{LCS(c,r)}{n}$, with $LCS(c, r)$ denoting the Longest Common Subsequence between the candidate sentence (c) and the reference sentence (r), m is the number of n-grams in the reference sentence, and n the number of n-grams in the candidate sentence.

In contrast to the two above-discussed metrics, METEOR (Metric for Evaluation of Translation with Explicit Ordering) gives attention to the position of the words in the sentence by including a chunk penalty that weights the final score.

$$METEOR = F_{mean} \cdot (1 - Penalty), \tag{4}$$

where $Penalty = 0.5 \cdot (\frac{m}{n})$, where m is the number of chunks and n is the total number of unigram matches, and $F_{mean} = \frac{10PR}{R+9P}$.

Alternatively, CIDEr (Consensus-Based Image Description Evaluation) is an evaluation metric that uses the Term Frequency Inverse Document Frequency (TF-IDF) [115] for weighting each n-gram. The intuition behind CIDEr is that n-grams that frequently appear in the reference sentences are less likely to be informative; hence, they have a lower weight using the IDF term. CIDEr_n, $n = \{1, 2, 3, 4\}$, is the average cosine similarity between the candidate sentence and the reference sentences, considering both precision and recall.

TF-IDF weighting $g_k(s_{ij})$ for each n-gram w_k is computed as follows:

$$g_k(s_{ij}) = \frac{h_k(s_{ij})}{\sum_{w_l \in \omega} h_l(s_{ij})} \log\left(\frac{|I|}{\sum_{I_p \in I} \min(1, \sum_q h_k(s_{pq}))}\right), \tag{5}$$

where $h_k(s_{ij})$ is the number of times an n-gram occurs in a reference sentence s_{ij}, and $h_k(c_i)$ for the candidate sentence c_i, ω is the vocabulary of all n-grams, and I is the set of all images.

$$CIDEr_n(c_i, S_i) = \frac{1}{m} \sum_j \frac{g^n(c_i) \cdot g^n(s_{ij})}{\|g^n(c_i)\| \|g^n(s_{ij})\|}. \tag{6}$$

The final CIDEr score is the weighted average of CIDEr_n:

$$CIDEr(c_i, S_i) = \sum_{n=1}^{N} w_k CIDEr_n(c_i, S_i), \tag{7}$$

where $w_n = 1/N, N = 4$.

Table 3. Performance of the Selected Methods that Generate Textual Explanations for Interpreting the Decision of a Classifier

	Model	Year	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR	CIDEr
	IU Chest X-Ray								
Find. + Impress.	Jing et al. [54]†	2017	0.517	0.386	**0.306**	**0.247**	0.447	**0.217**	0.327
	Singh et al. [136]°	2019	0.374	0.224	0.153	0.110	0.308	0.164	0.360
	HRNN [169]†	2019	0.445	0.292	0.201	0.154	0.344	0.175	0.342
	Selivanov et al. [126]†	2023	**0.520**	**0.390**	0.296	0.235	**0.450**	-	**0.701**
Findings	HRGR-Agent [73]⋆	2018	0.438	0.298	0.208	0.151	0.322	-	0.343
	TieNet[1] [159]⋆	2018	0.330	0.194	0.124	0.081	0.311	-	1.334
	Liu et al. [82]⋆	2019	0.369	0.246	0.171	0.115	0.359	-	**1.490**
	R2Gen [24]⋆	2020	0.470	0.304	0.219	0.165	0.371	0.187	-
	PPKED [80]⋆	2021	0.483	0.315	0.224	0.168	0.376	-	0.351
	CA [81]⋆	2021	0.492	0.314	0.222	0.169	0.381	0.193	-
	ICT [174]⋆	2023	0.503	**0.341**	**0.246**	0.186	0.390	**0.208**	-
	METransformer [160]⋆	2023	0.483	0.322	0.228	0.172	0.380	0.192	0.435
	VLCI [22]⋆	2023	**0.505**	0.334	0.245	**0.189**	0.397	0.204	0.456
	Yang et al. [168]⋆	2023	0.497	0.319	0.230	0.174	**0.399**	-	0.407
	MIMIC-CXR								
	TieNet[1] [159]‡	2018	0.332	0.212	0.142	0.095	0.296	-	1.004
	Liu et al. [82]‡	2019	0.352	0.223	0.153	0.104	0.307	-	1.153
	R2Gen [24]	2020	0.353	0.218	0.145	0.103	0.277	0.142	-
	CA [81]	2021	0.350	0.219	0.152	0.109	0.283	0.151	-
	PPKED [80]	2021	0.360	0.224	0.149	0.106	0.284	0.149	-
	MSAT [161]	2022	0.373	0.235	0.162	0.120	0.282	0.143	0.299
	ICT [174]	2023	0.376	0.233	0.157	0.113	0.276	0.144	-
	METransformer [160]	2023	0.386	0.250	0.169	0.124	0.291	**0.152**	0.362
	Yang et al. [168]	2023	0.386	0.237	0.157	0.111	0.274	-	0.111
	VLCI [22]	2023	0.400	0.245	0.165	0.119	0.280	0.150	0.190
	Selivanov et al. [126]	2023	**0.725**	**0.626**	**0.505**	**0.418**	**0.480**	-	**1.989**

Spaces marked with "-" mean no value is available for the respective metric. [1] Results were taken from the work of Liu et al. [82]. The methods evaluated on the IU Chest X-ray dataset are marked with a symbol (⋆, †, °, ‡), meaning that the methods with each symbol used the same training-validation-test split.

7 PERFORMANCE COMPARISON

In the previous sections, we reviewed several works focused on providing explanations for the output of automated medical diagnosis. At the end of the review, an important question arises: "*What is the best approach?*". Unfortunately, in many cases, there is no trivial answer since most methods adopt different evaluation metrics, making performance comparison with other competing methods unfeasible. Leading up to the question, we compare the performance of some of the methods reviewed. In order to find common ground for a fair comparison of the methods, only those methods that considered the same dataset for evaluation purposes were selected. As presented in Table 3, IU Chest X-ray [31] is the most used dataset among the reviewed methods. As such, we verified whether the methods that were evaluated on the IU Chest X-ray [31] used the same evaluation metrics. Additionally, and since the MIMIC-CXR [55] dataset provides an official training-test partition, we include some methods that report results on this dataset under the same evaluation metrics. In this way, a comparison of the performance between these methods was carried out. Table 3 conveys the results in terms of a set of NLP evaluation metrics (BLEU score, ROUGE, METEOR, and CIDEr) for each of the selected methods.

7.1 Results Discussion

It is worth noting that all the results in Table 3 were taken from the original paper of each method except for the TieNet [159] method, whose results were taken from the work of Liu et al. [82] since in the original paper the authors only provide results in ChestX-ray14 using BLEU, METEOR, and ROUGE-L.

Regarding the methods that considered the "findings+impression" section from the radiology report to generate the free-text report, and as evidenced by the results in Table 3, the model proposed by Selivanov et al. [126] proved to be more accurate in terms of BLUE-1, BLEU-2, and ROUGE-L. The preprocessing and squeezing approaches used for clinical records jointly with the combination of two large language models (Show-Attend-Tell (SAT) [165] and Generative Pretrained Transformer (GPT-3) [17]) can explain the performance improvement. Moreover, generated reports are accompanied by 2D heatmaps that localize each pathology on the input scans. Conversely, the method of Jing et al. [54] demonstrates superior performance in terms of BLEU-3, BLEU-4, and METEOR, which can be explained by the co-attention mechanism adopted by the authors. As the method of Singh et al. [136] does not follow the same data partition, it is not strictly comparable to other methods.

Among the selected methods that only consider the "findings" section, ICT [174] demonstrates superior performance on BLEU-2, BLEU-3, and METEOR metrics. Their Transformer-based model incorporates two modules responsible for capturing inter-intra features of medical reports as auxiliary information and subsequently calibrating the report generation process by integrating that information. This combination results in a performance boost in the report generation model and improves the quality of medical diagnosis. In contrast, with regard to the BLEU-1 score, VLCI [22] exhibits a superior performance, comparable to ICT [174], which could be justified by the cross-modal causal intervention strategy employed by the authors to mitigate spurious correlations from visual and linguistic confounders. Furthermore, the model proposed by Liu et al. [82] adopted a fine-tuning procedure that uses reinforcement learning via CIDEr to ensure more coherent report generation, which could justify the performance in the CIDEr metric.

Regarding the methods that report results for the MIMIC-CXR dataset, it is worth noting that only TieNet [159] and the approach by Liu et al. [82] have followed a distinct data partition strategy. Consequently, these methods are included here primarily for reference, as their divergent data partitioning makes direct comparisons with other methods infeasible. In contrast, the model proposed by Selivanov et al. [126] distinguishes itself from the others by exhibiting superior performance across all metrics except for METEOR (which was not reported by the authors). Their approach leveraged the capabilities of two large language models, SAT and GPT-3, trained on large text corpora. This fusion of language models significantly improved standard text generation scores, as shown in Table 3.

Overall, it is noticeable in Table 3 that the use of Transformer-based models [80, 81, 160, 161, 168] with additional mechanisms to capture complex and relevant features proved to be effective in improving the performance of the generated reports, as observed in the results obtained in the MIMIC-CXR dataset.

8 GENERAL DISCUSSION

Although XAI is a relatively recent research field, its constant growth is undeniable, with applications in many areas, particularly in the medical domain. However, despite the advances and the efforts made toward developing interpretable deep learning–based models for medical imaging, there are open issues that require more research and advances in this growing field. This section identifies open challenges in the literature and potential research paths to further improve the trustworthiness of provided explanations and foster the adoption of deep learning–based systems into clinical routine.

Based on reviewed literature, it can be concluded that the go-to method for model interpretation in medical imaging is producing saliency maps using classical techniques such as Grad-CAM, Integrated Gradients, or LRP. However, as evidenced by some authors, saliency maps can be unreliable and fragile [2, 117], as they often highlight irrelevant regions in the images. In addition, very

similar explanations are frequently given for different classes and often none of them are useful explanations [117]. Thus, the development of inherently interpretable models has been a line of research with promising results in the medical imaging domain. Although these methods remain largely unexplored in medical imaging, future research will undoubtedly be devoted to developing inherently interpretable models. These models have the primary benefit of providing their own explanations, which contributes to their transparency and fidelity, increasing the chances of being adopted into the clinical routine.

On the other hand, different end-users could have different backgrounds and preferences at interpreting the explanation, which can generate some contradictory opinions. This fostered the use of textual explanations, which are preferred over visual explanations by some authors [38] since they are inherently understandable by humans [151]. Since then, methods that generate textual descriptions for explaining a prediction and multimodal methods that combine visual and textual explanations have emerged. However, generating free-text reports is deemed a challenging task since the radiologist reports are technically structured, and the most used language models based on RNNs have some limitations in generating long texts, as stated by Pascanu et al. [104].

As an alternative to text-based explanations, the use of example-based explanations was proposed, since this explanation modality is directly linked to how humans try to explain something to other humans. This way, some example-based approaches have emerged with promising results that were even comparable to the performance of standard classifiers. These example-based methods include CBR approaches and prototype-based and concept-based strategies. Recently, Schutte et al. [125] introduced a disruptive approach that strives to generate synthetic examples to explain a model decision, as discussed in Section 5.5. The possible limitations of using example-based methods are related to the availability of a considerable amount of data covering all the classes in a balanced way, without forgetting the sensibility of these methods to adversarial attacks, even though this can be prevented by using adversarial training [92].

Alternatively, other methods have emerged as candidates for explaining the decision of a model. The adoption of Bayesian Neural Networks to estimate or quantify the uncertainty regarding the model predictions might be an interesting option, although few works attempted to prove its effectiveness.

Future research in medical image interpretability can also include the use of vision Transformers (ViTs) [33]. According to [93], vision Transformers proved to be comparable with CNNs in terms of performance (accuracy) in medical classification tasks. Furthermore, ViTs have the benefit of providing a type of built-in saliency map that is used to better understand the model's decisions.

Regarding medical image datasets, existing publicly available datasets for medical image captioning are limited in number and there is a need to generate more large-size datasets. Moreover, most of the existing datasets have focused only on a few anatomical parts of body, such as the chest, while ignoring other important parts such as the breast and brain [8].

8.1 Challenges and Future Research Trends

Despite the rapid pace of advances in medical imaging and deep learning, there are problems that remain without a definitive solution.

- **Small datasets**: The collection of medical data depends on multiple entities and background bureaucracies. Nevertheless, the main issue is related to the availability of physicians in annotating a vast amount of data, which is time-consuming and costly. This is even more critical in the XAI field, where additional annotations are required (e.g., concepts, textual descriptions). For this reason, interpretability-compliant medical datasets have a lower representation of classes, resulting in poor generalizability and applicability of the developed

methods to real-world scenarios. To overcome these constraints, distinct data augmentation techniques have emerged as an alternative for collecting new data. Recently, Wickramanayake et al. [163] proposed the BRACE framework to augment the dataset based on concept-based explanations from model decisions, which can help to discover the samples in the under-represented regions in the training set. Furthermore, Wickramanayake et al. introduced a utility function to select the images in the under-representation regions and concepts that caused the misclassification. The images with a high utility score are selected to incorporate the training set. On the other hand, the use of generative approaches to perform data augmentation in a controlled way might be an interesting research direction.

- **Insufficient labelled data**: Although most works rely on the supervised learning paradigm, it is often not the best choice when working in the medical domain since the process of label annotation is time-consuming and costly for large-scale datasets, especially in domains such as digital pathology in which the manual annotation is subject to inter- and intra-observer variability [171]. Transfer learning was adopted in most works to address these issues, but this technique is not completely effective in the medical domain since the original models were trained in images belonging to standard object detection datasets (e.g., ImageNet), which do not share the same patterns of medical imagery. Self-Supervised Learning (SSL) has emerged to tackle these challenges, allowing the network to learn visual meaningful feature representations without the need for annotated data [26]. Besides its effectiveness in dealing with scarce labelled data, it confers robustness to the model, rendering it more resistant to adversarial attacks. In medical imaging, the use of SSL seems to be a promising research direction due to the characteristics of medical datasets. Furthermore, contrastive learning approaches have achieved impressive results due to the contrastive loss that encourages the network to learn high-level features that occur in images across multiple views, which are created through the use of geometric transformation such as random cropping, color distortion, and Gaussian blur. For a comprehensive overview of the state-of-the-art of SSL with a particular focus on the medical domain, we refer the reader to [26]. The inherently interpretable models, specifically the concept bottleneck models, require the annotation of concepts for each class or image. The datasets in Table 2 show that the majority does not include this type of annotation, hampering the rapid employment of the concept bottleneck models. Furthermore, despite the significance of the existing methods that emerged to surpass these issues (CAV [59]), annotations regarding clinical concepts remain necessary. This could be solved with closer cooperation between clinicians and the AI community. As discussed in Section 4.6, appropriate annotations in medical imaging datasets are needed to ensure a quick development of interpretability methods for medical diagnosis.
- **Qualitative assessment of the explanations**: The automated evaluation of the explanations provided by interpretability methods remains an open challenge. As previously discussed in Section 6.1, the most adopted method for evaluating the explanations in the context of the medical domain is to resort to the clinician's expertise. However, considering variability in experts' opinions [146], this strategy is particularly biased and subjective. On the other hand, the existing strategies for objectively measuring the quality of visual explanations are still dependent on manual annotations of relevant regions [114] or iterative model retraining (ROAR [51]). For these reasons, we believe that the design of objective metrics for assessing the quality of the model explanations will be one of the important research trends on the topic of XAI.
- **Report generation in medical imaging**: Text-based explanations are usually obtained using RNN-based approaches through the generation of words forming a sentence. Nevertheless, RNN-based approaches have some limitations in generating long text reports [104].

Consequently, the use of Transformers for the automatic generation of radiology reports was adopted in an attempt to overcome the limitations of traditional RNNs, namely, the problem of vanishing gradients. The self-attention mechanism of the Transformer architecture allows for the learning of contextual relationships between the words that constitute the sequence. In addition, Transformer-based networks can be trained faster than traditional RNNs as they allow for simultaneous processing of sequential data. With the rise of foundation models, such as Generative Pretrained Transformers [126], the limitations encountered in previous methods have been alleviated, specifically the coherency of the generated texts. On the other hand, using concept-based approaches as a transition bridge between free-text report generation and concept-based explanations may be an exciting future research direction. Instead of trying to generate free-text reports, which is challenging, having a set of concepts that are sufficient to describe the phenomenon depicted in the image can support clinicians in writing a complete report.

- **Deployment in clinical practice**: The implementation of XAI methods in clinical practice requires rigorous validation to ensure their safety, effectiveness, and reliability, which can be challenging due to the complex and dynamic nature of clinical environments. Additionally, the field of medical imaging is subject to rigorous regulations, and the development and deployment of XAI methods must comply with regulatory and legal requirements [41], such as U.S. Food and Drug Administration approvals [14], data privacy regulations, and liability concerns.

9 CONCLUSIONS

This article reviewed the advances in explainable deep learning applied to medical imaging diagnosis. First, we introduced a comparative analysis between the existing surveys on the topic, in which the major conclusions and weaknesses of each were highlighted. Then, the most prominent XAI methods were briefly described to provide the readers with fundamental concepts of the field, necessary to the discussion of the recent advances in the medical imaging domain. Additionally, several frameworks that implement XAI methods were presented and a brief discussion of the existing medical imaging datasets was provided. We then comprehensively reviewed the works focused on explaining the decision process of deep learning applied to medical imaging. The works were grouped according to the explanation modality, comprising explanations by feature attribution, explanations by text, explanations by examples, and explanations by concepts. Contrary to other surveys on the topic, we focused this review on inherently interpretable models over post-hoc approaches, which has been recently considered a future research direction on deep learning interpretability. A discussion of the adopted evaluation metrics used in the literature was also presented, in which we described the existing metrics to assess the quality of visual explanations and the commonly NLP metrics to evaluate the quality of the generated textual explanations. A comparison of the performance of a set of prominent XAI methods was performed based on the dataset used and the evaluation metrics adopted. Finally, a discussion and future outlook for XAI in medical diagnosis were addressed in which we identified open challenges in the literature and potential research avenues to improve the trustworthiness of provided explanations and to foment the adoption of deep learning–based systems into clinical routine. To conclude, we believe that this survey will be helpful to the XAI community, particularly to the medical imaging field, as an entry point to guide the research and future advances in the topic of XAI.

A APPENDIX

A.1 Intepretability Frameworks

The increasing interest in the interpretability of machine learning fostered the development of frameworks implementing classical XAI methods. The LRP Toolbox [69] was launched in 2016 and provides the implementation of the LRP [9] algorithm for artificial neural networks supporting MATLAB and Python. Additionally, the LRP toolbox released an extension to be compatible with the Caffe Deep Learning framework. DeepExplain [6] is a framework that implements perturbation and gradient-based attribution methods, including Saliency Maps [132], Gradient * Input [130], Integrated Gradients (IG) [143], DeepLIFT [130], ϵ-LRP [9], and DeConvNet [173]. It also supports Tensorflow as well as Keras with Tensorflow backend. Alternatively, iNNvestigate [4] is a more complete toolbox that provides implementations for SmoothGrad [138], DeConvNet [173], Guided-BackProp [141], PatternNet [62], Input * Gradients [130], DeepTaylor [97], PatternAttribution [62], LRP [9], and Integrated Gradients (IG) [143]. It also supports Tensorflow and Keras.

With regard to PyTorch frameworks, TorchRay [36] implements several visualization methods, namely, Gradient [132], Guided-BackProp [141], Grad-CAM [127], and RISE [107]. TorchRay is considered research oriented, since it provides code for reproducing results that appear in several papers. Captum [64] is a PyTorch library that provides state-of-the-art algorithms for model interpretability and understanding. It contains general purpose implementations of Integrated Gradients [143], SmoothGrad [138], VarGrad [113], and others for PyTorch models. Table 4 summarizes the aforementioned frameworks alongside the supported XAI methods. Additionally, we refer the reader to the work of Darias et al. [29], in which some other model-agnostic XAI libraries were approached.

Table 4. Publicly Available Interpretability Frameworks

Framework	Year	Methods	Supported DL Libraries
LRP Toolbox [69]	2016	LRP	Caffe
DeepExplain [6]	2017	Saliency maps, Grad * Input, ϵ-LRP, DeepLIFT, DeConvNet	Tensorflow, Keras
iNNvestigate [4]	2019	SmoothGrad, DeConvNet, Guided-BackProp, PatternNet, LRP Input * Gradients, DeepTaylor, PatternAttribution, IG	Tensorflow, Keras
TorchRay [36]	2019	Gradient, Guided-BackProp, Grad-CAM, RISE	PyTorch
Captum [64]	2019	SmoothGrad, DeConvNet, Guided-BackProp, PatternNet, LRP Input * Gradients, DeepLIFT, DeepTaylor, LIME, Kernel SHAP, IG	PyTorch

A.2 Methods

Table 5. Summary of the XAI Methods Categorized by Interpretability Method Employed, Image Modality, and Dataset

Method	Year	Interpretability Method	Modality	Dataset
Zhang et al. [175]	2017	Attention-based	Microscopic Images	BCIDR
Jing et al. [54]	2017	Attention-based	X-ray	IU Chest X-ray, PEIR Gross
Rajpurkar et al. [110]	2018	CAM	X-ray	ChestX-ray8
Wang et al. [159]	2018	Saliency maps	X-ray	IU Chest X-ray, ChestX-ray14
Gale et al. [38]	2018	SmoothGrad	X-ray	Pelvic X-ray
Li et al. [73]	2018	Text-based	X-ray	IU Chest X-ray, CX-CHR (private)
Malhi et al. [91]	2019	LIME	Endoscopy	Red Lesion Endoscopy
Young et al. [170]	2019	KernelSHAP	Dermoscopy	HAM10000
Eitel et al. [34]	2019	Occlusion	MRI	ADNI Initiative
Sayres et al. [123]	2019	Integrated Gradients	Fundus Images	Private
Barata et al. [11]	2019	CAM	Dermoscopy	ISIC 2017, ISIC 2018
Tschandl et al. [147]	2019	CBIR	Dermoscopy	EDRA, ISIC 2017, Private Dataset
Sun et al. [142]	2019	Text-based	Mammography	Inbreast
Lee et al. [71]	2019	Saliency maps	Mammography	CBIS-DDSM
Lamy et al. [68]	2019	CBR	Mammography	BCW, Mammographic Mass, Breast Cancer
Singh et al. [136]	2019	Stacked LSTM	X-ray	IU Chest X-ray
Yin et al. [169]	2019	t-SNE	X-ray	IU Chest X-ray
Liu et al. [82]	2019	Attention maps	X-ray	IU Chest X-ray, MIMIC-CXR
Windish et al. [164]	2020	Grad-CAM	MRI	IXI, Glioblastoma
Magesh et al. [89]	2020	LIME	DaTscan	PPMI
Lin et al. [76]	2020	Guided Grad-CAM	X-ray	COVIDx
Lopatina et al. [83]	2020	DeepLIFT	MRI	Private
Graziani et al. [43]	2020	TCAV	Microscopic Images	Camelyon16, Camelyon17
Margeloiu et al. [92]	2020	Adversarial Training	Dermoscopy	HAM10000
Rio-Torto et al. [114]	2020	In-model	Microscopic Images	NIH-NCI Cervical Cancer
Fang et al. [35]	2020	Concept-based	Microscopic Images	Infectious Keratitis Dataset
Chen et al. [24]	2020	Multi-Head Attention	X-ray	IU Chest X-ray, MIMIC-CXR
Silva et al. [131]	2020	CBIR, Saliency Map	X-ray	CheXpert
Punn et al. [109]	2021	LIME	X-ray	COVID-19 Dataset
Wang et al. [157]	2021	SHAP	Dermoscopy	HAM10000
Billah and Javed [15]	2021	BCNN	Microscopy Images	ALL-IDB
Barata et al. [12]	2021	CBIR	Dermoscopy	ISIC 2018
Thiagarajan et al. [144]	2021	t-SNE	Microscopic Images	Breast Histopathology
Barnett et al. [13]	2021	Prototype Activation Map	Mammography	Mammography Private Dataset
Kim et al. [61]	2021	Counterfactual Examples	X-ray	Chest X-ray14, VinDr-CXR
Schutte et al. [125]	2021	Grad-CAM	X-ray/Microscopy Images	Osteoarthritis X-ray, Camelyon16
Kim et al. [60]	2021	Saliency maps	X-ray	NIH chest X-ray14
Singh et al. [134]	2021	Prototype Activation Maps	X-ray	CORD-19
Liu et al. [80]	2021	Text-based	X-ray	IU Chest X-ray, MIMIC-CXR
Liu et al. [81]	2021	Text-based	X-ray	IU Chest X-ray, MIMIC-CXR
Lucieri et al. [84]	2022	TCAV	Dermoscopy	ISIC 2019, Derm7pt, PH2, SkinL2
Hu et al. [52]	2022	CBIR	X-ray/Dermoscopy	COVIDx, ISIC 2019
Gour and Jain [42]	2022	Uncertainty-based	X-ray	COVID19CXr
Yuksekgonul et al. [172]	2022	CBM	Dermoscopy	HAM 10000, ISIC 2020
Wang et al. [161]	2022	Text-based	X-ray	MIMIC-CXR
Singla et al. [137]	2023	Counterfactual Examples	X-ray	MIMIC-CXR
Yan et al. [167]	2023	CBM	Dermoscopy	ISIC 2016–2020
Selivanov et al. [126]	2023	Text-based	X-ray	IU Chest X-ray, MIMIC-CXR
Zhang et al. [174]	2023	Text-based	X-ray	IU Chest X-ray, MIMIC-CXR, COV-CTR
Wang et al. [160]	2023	Text-based	X-ray	IU Chest X-ray, MIMIC-CXR
Chen et al. [22]	2023	Text-based	X-ray	IU Chest X-ray, MIMIC-CXR
Yang et al. [168]	2023	Text-based	X-ray	IU Chest X-ray, MIMIC-CXR
Patrício et al. [105]	2023	CBM	Dermoscopy	PH2, Derm7pt

The spaces marked with "-" mean that explanation is only provided through text sentences.

REFERENCES

[1] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. 2012. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34, 11 (2012), 2274–2282.

[2] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. 2018. Sanity checks for saliency maps. *Advances in Neural Information Processing Systems* 31 (2018).

[3] Ceyhun Burak Akgül, Daniel L. Rubin, Sandy Napel, Christopher F. Beaulieu, Hayit Greenspan, and Burak Acar. 2011. Content-based image retrieval in radiology: Current status and future directions. *Journal of Digital Imaging* 24, 2 (2011), 208–222.

[4] Maximilian Alber, Sebastian Lapuschkin, Philipp Seegerer, Miriam Hägele, Kristof T. Schütt, Grégoire Montavon, Wojciech Samek, Klaus-Robert Müller, Sven Dähne, and Pieter-Jan Kindermans. 2019. iNNvestigate neural networks! *Journal of Machine Learning Research* 20, 93 (2019), 1–8.

[5] Stefano Allegretti, Federico Bolelli, Federico Pollastri, Sabrina Longhitano, Giovanni Pellacani, and Costantino Grana. 2021. Supporting skin lesion diagnosis with content-based image retrieval. In *Proceedings of the IEEE International Conference on Pattern Recognition (ICPR'21)*. 8053–8060.

[6] Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. 2018. Towards better understanding of gradient-based attribution methods for deep neural networks. In *International Conference on Learning Representations (ICLR'18)*.

[7] Samuel G. Armato III, Geoffrey McLennan, Luc Bidaut, Michael F. McNitt-Gray, Charles Meyer, Anthony P. Reeves, Binsheng Zhao, Denise R. Aberle, Claudia I. Henschke, Eric A. Hoffman, et al. 2011. The lung image database consortium (LIDC) and image database resource initiative (IDRI): A completed reference database of lung nodules on CT scans. *Medical Physics* 38, 2 (2011), 915–931.

[8] Hareem Ayesha, Sajid Iqbal, Mehreen Tariq, Muhammad Abrar, Muhammad Sanaullah, Ishaq Abbas, Amjad Rehman, Muhammad Farooq Khan Niazi, and Shafiq Hussain. 2021. Automatic medical image interpretation: State of the art and future directions. *Pattern Recognition* 114 (2021), 107856.

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

[10] Tao Bai, Jinqi Luo, Jun Zhao, Bihan Wen, and Qian Wang. 2021. Recent advances in adversarial training for adversarial robustness. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI'21)*. 4312–4321.

[11] Catarina Barata, Jorge S. Marques, and M. Emre Celebi. 2019. Deep attention model for the hierarchical diagnosis of skin lesions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW'19)*. 2757–2765.

[12] Catarina Barata and Carlos Santiago. 2021. Improving the explainability of skin cancer diagnosis using CBIR. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI'21)*. 550–559.

[13] Alina Jade Barnett, Fides Regina Schwartz, Chaofan Tao, Chaofan Chen, Yinhao Ren, Joseph Y. Lo, and Cynthia Rudin. 2021. Interpretable mammographic image classification using case-based reasoning and deep learning. In *Workshop on Deep Learning, Case-Based Reasoning, and AutoML: Present and Future Synergies - IJCAI*.

[14] Stan Benjamens, Pranavsingh Dhunnoo, and Bertalan Meskó. 2020. The state of artificial intelligence-based FDA-approved medical devices and algorithms: An online database. *NPJ Digital Medicine* 3, 1 (2020), 118.

[15] Mohammad Ehtasham Billah and Farrukh Javed. 2022. Bayesian convolutional neural network-based models for diagnosis of blood cancer. *Applied Artificial Intelligence* (2022), 1–22.

[16] Hamidreza Bolhasani, Elham Amjadi, Maryam Tabatabaeian, and Somayyeh Jafarali Jassbi. 2020. A histopathological image dataset for grading breast invasive ductal carcinomas. *Informatics in Medicine Unlocked* 19 (2020), 100341.

[17] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems (NeurIPS)* 33 (2020), 1877–1901.

[18] Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria de la Iglesia-Vayá. 2020. PadChest: A large chest x-ray image dataset with multi-label annotated reports. *Medical Image Analysis* 66 (2020), 101797.

[19] Kirill Bykov, Marina M.-C. Höhne, Adelaida Creosteanu, Klaus-Robert Müller, Frederick Klauschen, Shinichi Nakajima, and Marius Kloft. 2021. Explaining Bayesian neural networks. *arXiv preprint arXiv:2108.10346* (2021).

[20] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. 2018. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision (ECCV'18)*. 132–149.

[21] Chaofan Chen, Oscar Li, Chaofan Tao, Alina Jade Barnett, Jonathan Su, and Cynthia Rudin. 2019. This looks like that: Deep learning for interpretable image recognition. In *Proceedings of the International Conference of Neural Information Processing Systems (NIPS'19)*.

[22] Weixing Chen, Yang Liu, Ce Wang, Guanbin Li, Jiarui Zhu, and Liang Lin. 2023. Visual-linguistic causal intervention for radiology report generation. *arXiv preprint arXiv:2303.09117* (2023).

[23] Zhi Chen, Yijie Bei, and Cynthia Rudin. 2020. Concept whitening for interpretable image recognition. *Nature Machine Intelligence* 2, 12 (2020), 772–782.

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

[25] Deepak Roy Chittajallu, Bo Dong, Paul Tunison, Roddy Collins, Katerina Wells, James Fleshman, Ganesh Sankaranarayanan, Steven Schwaitzberg, Lora Cavuoto, and Andinet Enquobahrie. 2019. XAI-CBIR: Explainable AI system for content based retrieval of video frames from minimally invasive surgery videos. In *Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI'19)*. 66–69.

[26] Alexander Chowdhury, Jacob Rosenthal, Jonathan Waring, and Renato Umeton. 2021. Applying self-supervised learning to medicine: Review of the state of the art and medical implementations. *Informatics* 8, 3 (2021), 59.

[27] Joseph Paul Cohen, Paul Morrison, and Lan Dao. 2020. COVID-19 image data collection. *arXiv 2003.11597* (2020). https://github.com/ieee8023/covid-chestxray-dataset

[28] Roxana Daneshjou, Mert Yuksekgonul, Zhuo Ran Cai, Roberto Novoa, and James Y. Zou. 2022. SkinCon: A skin disease dataset densely annotated by domain experts for fine-grained debugging and analysis. *Advances in Neural Information Processing Systems* 35 (2022), 18157–18167.

[29] Jesus M. Darias, Belén Dıaz-Agudo, and Juan A. Recio-Garcia. 2021. A systematic review on model-agnostic XAI libraries. (2021).

[30] Sergio de Faria, Jose Filipe, Pedro Pereira, Luis Tavora, Pedro Assuncao, Miguel Santos, Rui Fonseca-Pinto, Felicidade Santiago, Victoria Dominguez, and Martinha Henrique. 2019. Light field image dataset of skin lesions. In *Proceedings of the International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC'19)*. 3905–3908.

[31] Dina Demner-Fushman, Marc D. Kohli, Marc B. Rosenman, Sonya E. Shooshan, Laritza Rodriguez, Sameer Antani, George R. Thoma, and Clement J. McDonald. 2016. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association* 23, 2 (2016), 304–310.

[32] Jon Donnelly, Alina Jade Barnett, and Chaofan Chen. 2021. Deformable ProtoPNet: An interpretable image classifier using deformable prototypes. *arXiv preprint arXiv:2111.15000* (2021).

[33] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations (ICLR'21)*.

[34] Fabian Eitel and Kerstin Ritter for the Alzheimer's Disease Neuroimaging Initiative (ADNI). 2019. Testing the robustness of attribution methods for convolutional neural networks in MRI-based Alzheimer's disease classification. In *Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support (IMIMIC'19)*. 3–11.

[35] Zhengqing Fang, Kun Kuang, Yuxiao Lin, Fei Wu, and Yu-Feng Yao. 2020. Concept-based explanation for fine-grained images and its application in infectious keratitis classification. In *Proceedings of the ACM International Conference on Multimedia*. 700–708.

[36] Ruth Fong, Mandela Patrick, and Andrea Vedaldi. 2019. Understanding deep networks via extremal perturbations and smooth masks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV'19)*. 2950–2958.

[37] Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the International Conference on Machine Learning (ICML'16)*. 1050–1059.

[38] William Gale, Luke Oakden-Rayner, Gustavo Carneiro, Andrew P. Bradley, and Lyle J. Palmer. 2019. Producing radiologist-quality reports for interpretable deep learning. In *Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI'19)*. 1275–1279.

[39] Amirata Ghorbani, Abubakar Abid, and James Zou. 2019. Interpretation of neural networks is fragile. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 3681–3688.

[40] Amirata Ghorbani, James Wexler, James Y. Zou, and Been Kim. 2019. Towards automatic concept-based explanations. In *Advances in Neural Information Processing Systems*, Vol. 32.

[41] Bryce Goodman and Seth Flaxman. 2017. European Union regulations on algorithmic decision-making and a "Right to Explanation". *AI Magazine* 38, 3 (2017), 50–57.

[42] Mahesh Gour and Sweta Jain. 2022. Uncertainty-aware convolutional neural network for COVID-19 X-ray images classification. *Computers in Biology and Medicine* 140 (2022), 105047.

[43] Mara Graziani, Vincent Andrearczyk, Stéphane Marchand-Maillet, and Henning Müller. 2020. Concept attribution: Explaining CNN decisions to physicians. *Computers in Biology and Medicine* 123 (2020), 103865.

[44] Matthew Groh, Caleb Harris, Luis Soenksen, Felix Lau, Rachel Han, Aerin Kim, Arash Koochek, and Omar Badri. 2021. 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 (CVPR'21)*. 1820–1828.

[45] Mehmet A. Gulum, Christopher M. Trombley, and Mehmed Kantardzic. 2021. A review of explainable deep learning cancer detection models in medical imaging. *Applied Sciences* 11, 10 (2021), 4573.

[46] David Gunning and David Aha. 2019. DARPA's explainable artificial intelligence (XAI) program. *AI Magazine* 40, 2 (2019), 44–58.

[47] Tarun Gupta, Libin Kutty, Ritu Gahir, Nnamdi Ukwu, Sayantan Polley, and Marcus Thiel. 2021. IRTEX: Image retrieval with textual explanations. In *Proceedings of the IEEE International Conference on Human-Machine Systems (ICHMS'21)*. 1–4.

[48] David M. Hansell, Alexander A. Bankier, Heber MacMahon, Theresa C. McLoud, Nestor L. Müller, and Jacques Remy. 2008. Fleischner Society: Glossary of terms for thoracic imaging. *Radiology* 246, 3 (2008), 697–722.

[49] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems* 33 (2020), 6840–6851.

[50] Adrian Hoffmann, Claudio Fanconi, Rahul Rade, and Jonas Kohler. 2021. This looks like that... does it? Shortcomings of latent space prototype interpretability in deep networks. In *ICML Workshop on Theoretic Foundation, Criticism, and Application Trend of Explainable AI*.

[51] Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. 2019. A benchmark for interpretability methods in deep neural networks. In *Advances in Neural Information Processing Systems*, Vol. 32. 9737–9748.

[52] Brian Hu, Bhavan Vasu, and Anthony Hoogs. 2022. X-MIR: EXplainable medical image retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV'22)*. 440–450.

[53] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. 2019. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 590–597.

[54] Baoyu Jing, Pengtao Xie, and Eric Xing. 2018. On the automatic generation of medical imaging reports. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL'18)*.

[55] Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. 2019. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data* 6, 1 (2019), 1–8.

[56] Jeremy Kawahara, Sara Daneshvar, Giuseppe Argenziano, and Ghassan Hamarneh. 2019. Seven-point checklist and skin lesion classification using multitask multimodal neural nets. *IEEE Journal of Biomedical and Health Informatics* 23, 2 (2019), 538–546.

[57] Amirhossein Kazerouni, Ehsan Khodapanah Aghdam, Moein Heidari, Reza Azad, Mohsen Fayyaz, Ilker Hacihaliloglu, and Dorit Merhof. 2022. Diffusion models for medical image analysis: A comprehensive survey. *arXiv preprint arXiv:2211.07804* (2022).

[58] Oisin J. F. Keenan, George Holland, Julian F. Maempel, John F. Keating, and Chloe E. H. Scott. 2020. Correlations between radiological classification systems and confirmed cartilage loss in severe knee osteoarthritis. *The Bone & Joint Journal* 102-B, 3 (2020), 301–309.

[59] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie J. Cai, James Wexler, Fernanda B. Viégas, and Rory Sayres. 2018. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *Proceedings of the International Conference on Machine Learning (ICML'18)*. 2668–2677.

[60] Eunji Kim, Siwon Kim, Minji Seo, and Sungroh Yoon. 2021. XProtoNet: Diagnosis in chest radiography with global and local explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'21)*. 15719–15728.

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

[62] Pieter-Jan Kindermans, Kristof T. Schütt, Maximilian Alber, Klaus-Robert Müller, Dumitru Erhan, Been Kim, and Sven Dähne. 2018. Learning how to explain neural networks: PatternNet and PatternAttribution. In *International Conference on Learning Representations (ICML'18)*.

[63] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. 2020. Concept bottleneck models. In *Proceedings of the International Conference on Machine Learning (ICML'20)*. 5338–5348.

[64] Kokhlikyanet. 2019. PyTorch Captum. [Online]. Accessed January, 21 2021 from https://github.com/pytorch/captum

[65] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems (NIPS)* 25 (2012).

[66] Neeraj Kumar, Alexander C. Berg, Peter N. Belhumeur, and Shree K. Nayar. 2009. Attribute and simile classifiers for face verification. In *Proceedings of the IEEE International Conference on Computer Vision (CVPR'09)*. 365–372.

[67] Christoph H. Lampert, Hannes Nickisch, and Stefan Harmeling. 2009. Learning to detect unseen object classes by between-class attribute transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'09)*. 951–958.

[68] Jean-Baptiste Lamy, Boomadevi Sekar, Gilles Guezennec, Jacques Bouaud, and Brigitte Séroussi. 2019. Explainable artificial intelligence for breast cancer: A visual case-based reasoning approach. *Artificial Intelligence in Medicine* 94 (2019), 42–53.

[69] Sebastian Lapuschkin, Alexander Binder, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. 2016. The LRP toolbox for artificial neural networks. *Journal of Machine Learning Research* 17, 114 (2016), 1–5.

[70] Alon Lavie and Abhaya Agarwal. 2007. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Workshop on Statistical Machine Translation*. 228–231.
[71] Hyebin Lee, Seong Tae Kim, and Yong Man Ro. 2019. Generation of multimodal justification using visual word constraint model for explainable computer-aided diagnosis. In *Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support*. 21–29.
[72] Rebecca Sawyer Lee, Francisco Gimenez, Assaf Hoogi, Kanae Kawai Miyake, Mia Gorovoy, and Daniel L. Rubin. 2017. A curated mammography data set for use in computer-aided detection and diagnosis research. *Scientific Data* 4, 1 (2017), 1–9.
[73] Christy Y. Li, Xiaodan Liang, Zhiting Hu, and Eric P. Xing. 2018. Hybrid retrieval-generation reinforced agent for medical image report generation. In *Proceedings of the International Conference on Neural Information Processing Systems (NIPS'18)*. 1537–1547.
[74] Mingjie Li, Rui Liu, Fuyu Wang, Xiaojun Chang, and Xiaodan Liang. 2023. Auxiliary signal-guided knowledge encoder-decoder for medical report generation. *World Wide Web* 26, 1 (2023), 253–270.
[75] Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*. 74–81.
[76] Tsung-Chieh Lin and Hsi-Chieh Lee. 2020. Covid-19 chest radiography images analysis based on integration of image preprocess, guided grad-CAM, machine learning and risk management. In *Proceedings of the International Conference on Medical and Health Informatics (ICMHI'20)*. 281–288.
[77] Zachary C. Lipton. 2017. The doctor just won't accept that! *arXiv preprint arXiv:1711.08037* (2017).
[78] Geert Litjens, Peter Bandi, Babak Ehteshami Bejnordi, Oscar Geessink, Maschenka Balkenhol, et al. 2018. 1399 H&E-stained sentinel lymph node sections of breast cancer patients: the CAMELYON dataset. *GigaScience* 7, 6 (2018).
[79] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I. Sánchez. 2017. A survey on deep learning in medical image analysis. *Medical Image Analysis* 42 (2017), 60–88.
[80] Fenglin Liu, Xian Wu, Shen Ge, Wei Fan, and Yuexian Zou. 2021. Exploring and distilling posterior and prior knowledge for radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13753–13762.
[81] Fenglin Liu, Changchang Yin, Xian Wu, Shen Ge, Ping Zhang, and Xu Sun. 2021. Contrastive attention for automatic chest X-ray report generation. In *Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. 269–280.
[82] Guanxiong Liu, Tzu-Ming Harry Hsu, Matthew McDermott, Willie Boag, Wei-Hung Weng, Peter Szolovits, and Marzyeh Ghassemi. 2019. Clinically accurate chest X-ray report generation. In *Machine Learning for Healthcare Conference*. 249–269.
[83] Alina Lopatina, Stefan Ropele, Renat Sibgatulin, Jürgen R. Reichenbach, and Daniel Güllmar. 2020. Investigation of deep-learning-driven identification of multiple sclerosis patients based on susceptibility-weighted images using relevance analysis. *Frontiers in Neuroscience* (2020), 1356.
[84] Adriano Lucieri, Muhammad Naseer Bajwa, Stephan Alexander Braun, Muhammad Imran Malik, Andreas Dengel, and Sheraz Ahmed. 2022. ExAID: A multimodal explanation framework for computer-aided diagnosis of skin lesions. *Computer Methods and Programs in Biomedicine* (2022), 106620.
[85] Adriano Lucieri, Muhammad Naseer Bajwa, Andreas Dengel, and Sheraz Ahmed. 2020. Explaining AI-based decision support systems using concept localization maps. In *Proceedings of the International Conference on Neural Information Processing*. 185–193.
[86] Scott M. Lundberg, Gabriel G. Erion, and Su-In Lee. 2018. Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:1802.03888* (2018).
[87] Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the International Conference on Neural Information Processing Systems (NIPS'17)*. 4768–4777.
[88] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2018. Towards deep learning models resistant to adversarial attacks. In *Proceedings of the International Conference on Learning Representations (ICLR'18)*.
[89] Pavan Rajkumar Magesh, Richard Delwin Myloth, and Rijo Jackson Tom. 2020. An explainable machine learning model for early detection of Parkinson's disease using LIME on DaTscan imagery. *Computers in Biology and Medicine* 126 (2020), 104041.
[90] Anna Majkowska, Sid Mittal, David F. Steiner, Joshua J. Reicher, Scott Mayer McKinney, Gavin E. Duggan, Krish Eswaran, Po-Hsuan Cameron Chen, Yun Liu, Sreenivasa Raju Kalidindi, et al. 2020. Chest radiograph interpretation with deep learning models: Assessment with radiologist-adjudicated reference standards and population-adjusted evaluation. *Radiology* 294, 2 (2020), 421–431.

[91] Avleen Malhi, Timotheus Kampik, Husanbir Pannu, Manik Madhikermi, and Kary Främling. 2019. Explaining machine learning-based classifications of in-vivo gastral images. In *Digital Image Computing: Techniques and Applications (DICTA'19)*. 1–7.

[92] Andrei Margeloiu, Nikola Simidjievski, Mateja Jamnik, and Adrian Weller. 2020. Improving interpretability in medical imaging diagnosis using adversarial training. In *Medical Imaging Meets NeurIPS Workshop*.

[93] Christos Matsoukas, Johan Fredin Haslum, Magnus Söderberg, and Kevin Smith. 2021. Is it time to replace CNNs with Transformers for medical images? In *ICCV 2021: Workshop on Computer Vision for Automated Medical Diagnosis (CVAMD'21)*.

[94] Teresa Mendonça, Pedro Ferreira, Jorge Marques, André Marcal, and Jorge Rozeira. 2013. PH 2-A dermoscopic image database for research and benchmarking. In *Proceedings of the International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS'13)*. 5437–5440.

[95] Pablo Messina, Pablo Pino, Denis Parra, Alvaro Soto, Cecilia Besa, Sergio Uribe, Marcelo Andía, Cristian Tejos, Claudia Prieto, and Daniel Capurro. 2022. A survey on deep learning and explainability for automatic report generation from medical images. *ACM Computing Surveys (CSUR)* 54, 10s (2022), 1–40.

[96] Christoph Molnar. 2022. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. Available from https://christophm.github.io/interpretable-ml-book

[97] Grégoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller. 2017. Explaining nonlinear classification decisions with deep Taylor decomposition. *Pattern Recognition* 65 (2017), 211–222.

[98] Johanna D. Moore and William R. Swartout. 1988. Explanation in expert systems: A survey. *University of Southern California* (1988).

[99] Inês C. Moreira, Igor Amaral, Inês Domingues, António Cardoso, Maria Joao Cardoso, and Jaime S. Cardoso. 2012. INbreast: Toward a full-field digital mammographic database. *Academic Radiology* 19, 2 (2012), 236–248.

[100] M. Nevitt, D. Felson, and Gayle Lester. 2006. The osteoarthritis initiative. *Protocol for the Cohort Study* 1 (2006).

[101] Ha Nguyen, Khanh Lam, Linh Le, Hieu Pham, Dat Tran, et al. 2020. VinDr-CXR: An open dataset of chest X-rays with radiologist's annotations. *arXiv preprint arXiv:2012.15029* (2020).

[102] Hieu T. Nguyen, Hieu H. Pham, Nghia T. Nguyen, Ha Q. Nguyen, Thang Q. Huynh, et al. 2021. VinDr-SpineXR: A deep learning framework for spinal lesions detection and classification from radiographs. In *Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI'21)*. 291–301.

[103] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL'02)*. 311–318.

[104] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. 2013. On the difficulty of training recurrent neural networks. In *Proceedings of the International Conference on Machine Learning (ICML'13)*. 1310–1318.

[105] Cristiano Patrício, João C. Neves, and Luis F. Teixeira. 2023. 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 Workshops (CVPRW'23)*. 3799–3808.

[106] Obioma Pelka, Sven Koitka, Johannes Rückert, Felix Nensa, et al. 2018. Radiology objects in COntext (ROCO): A multimodal image dataset. In *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*. 180–189.

[107] Vitali Petsiuk, Abir Das, and Kate Saenko. 2018. RISE: Randomized input sampling for explanation of black-box models. In *British Machine Vision Conference (BMVC'18)*.

[108] Milda Pocevičiūtė, Gabriel Eilertsen, and Claes Lundström. 2020. Survey of XAI in digital pathology. In *Artificial Intelligence and Machine Learning for Digital Pathology: State-of-the-Art and Future Challenges*. 56–88.

[109] Narinder Singh Punn and Sonali Agarwal. 2021. Automated diagnosis of COVID-19 with limited posteroanterior chest X-ray images using fine-tuned deep neural networks. *Applied Intelligence* 51, 5 (2021), 2689–2702.

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

[111] Aasia Rehman, Muheet Ahmed Butt, and Majid Zaman. 2021. A survey of medical image analysis using deep learning approaches. In *Proceedings of the IEEE International Conference on Computing Methodologies and Communication (ICCMC'21)*. 1334–1342.

[112] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the predictions of any classifier. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1135–1144.

[113] Lorenz Richter, Ayman Boustati, Nikolas Nüsken, Francisco Ruiz, and Omer Deniz Akyildiz. 2020. VarGrad: A low-variance gradient estimator for variational inference. *Advances in Neural Information Processing Systems* 33 (2020), 13481–13492.

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

[115] Stephen Robertson. 2004. Understanding inverse document frequency: On theoretical arguments for IDF. *Journal of Documentation* 60 (2004), 503–520.

[116] Veronica Rotemberg, Nicholas Kurtansky, Brigid Betz-Stablein, Liam Caffery, Emmanouil Chousakos, Noel Codella, Marc Combalia, Stephen Dusza, Pascale Guitera, David Gutman, et al. 2021. A patient-centric dataset of images and metadata for identifying melanomas using clinical context. *Scientific Data* 8, 1 (2021), 1–8.

[117] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1, 5 (2019), 206–215.

[118] Mahya Sadeghi, Parmit K. Chilana, and M. Stella Atkins. 2018. How users perceive content-based image retrieval for identifying skin images. In *Understanding and Interpreting Machine Learning in Medical Image Computing Applications*. 141–148.

[119] Zohaib Salahuddin, Henry C. Woodruff, Avishek Chatterjee, and Philippe Lambin. 2022. Transparency of deep neural networks for medical image analysis: A review of interpretability methods. *Computers in Biology and Medicine* 140 (2022), 105111.

[120] Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. 2016. Evaluating the visualization of what a deep neural network has learned. *IEEE Transactions on Neural Networks and Learning Systems* 28, 11 (2016), 2660–2673.

[121] Pedro Sanchez, Antanas Kascenas, Xiao Liu, Alison Q. O'Neil, and Sotirios A. Tsaftaris. 2022. What is healthy? Generative counterfactual diffusion for lesion localization. In *Deep Generative Models: Second MICCAI Workshop, DGM4MICCAI 2022, Held in Conjunction with MICCAI 2022, Singapore, September 22, 2022, Proceedings*. 34–44.

[122] Anirban Sarkar, Deepak Vijaykeerthy, Anindya Sarkar, and Vineeth N. Balasubramanian. 2022. A framework for learning ante-hoc explainable models via concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'22)*. 10286–10295.

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

[124] Thomas Schlegl, Sebastian M. Waldstein, Wolf-Dieter Vogl, Ursula Schmidt-Erfurth, and Georg Langs. 2015. Predicting semantic descriptions from medical images with convolutional neural networks. In *Proceedings of the International Conference on Information Processing in Medical Imaging (IPMI'15)*. 437–448.

[125] Kathryn Schutte, Olivier Moindrot, Paul Hérent, Jean-Baptiste Schiratti, and Simon Jégou. 2021. Using StyleGAN for visual interpretability of deep learning models on medical images. In *Medical Imaging Meets NeurIPS Workshop*.

[126] Alexander Selivanov, Oleg Y. Rogov, Daniil Chesakov, Artem Shelmanov, Irina Fedulova, and Dmitry V. Dylov. 2023. Medical image captioning via generative pretrained Transformers. *Scientific Reports* 13, 1 (2023), 4171.

[127] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV'17)*. 618–626.

[128] Fahad Shamshad, Salman Khan, Syed Waqas Zamir, Muhammad Haris Khan, Munawar Hayat, Fahad Shahbaz Khan, and Huazhu Fu. 2023. Transformers in medical imaging: A survey. *Medical Image Analysis* (2023), 102802.

[129] L. S. Shapley. 2016. A value for n-person games. In *Contributions to the Theory of Games (AM-28)*, Harold William Kuhn and Albert William Tucker (Eds.). Vol. II. 307–318.

[130] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning important features through propagating activation differences. In *Proceedings of the International Conference on Machine Learning (ICML'17)*, Vol. 70. 3145–3153.

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

[132] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *Workshop at International Conference on Learning Representations (ICLR'14)*.

[133] Amitojdeep Singh, Sourya Sengupta, and Vasudevan Lakshminarayanan. 2020. Explainable deep learning models in medical image analysis. *Journal of Imaging* 6, 6 (2020), 52.

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

[135] Gurmail Singh and Kin-Choong Yow. 2021. These do not look like those: An interpretable deep learning model for image recognition. *IEEE Access* 9 (2021), 41482–41493.

[136] Sonit Singh, Sarvnaz Karimi, Kevin Ho-Shon, and Len Hamey. 2019. From chest X-rays to radiology reports: A multimodal machine learning approach. In *Digital Image Computing: Techniques and Applications (DICTA'19)*. 1–8.
[137] Sumedha Singla, Motahhare Eslami, Brian Pollack, Stephen Wallace, and Kayhan Batmanghelich. 2023. Explaining the black-box smoothly —a counterfactual approach. *Medical Image Analysis* 84 (2023), 102721.
[138] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. 2017. SmoothGrad: Removing noise by adding noise. In *ICML Workshop on Visualization for Deep Learning*.
[139] Asia Pacific Tele-Ophthalmology Society. 2019. APTOS Diabetic Retinopathy Dataset. [Online]. Accessed November, 23 2021 from https://www.kaggle.com/c/aptos2019-blindness-detection/data
[140] Fabio A. Spanhol, Luiz S. Oliveira, Caroline Petitjean, and Laurent Heutte. 2015. A dataset for breast cancer histopathological image classification. *IEEE Transactions on Biomedical Engineering* 63, 7 (2015), 1455–1462.
[141] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. 2014. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806* (2014).
[142] Li Sun, Weipeng Wang, Jiyun Li, and Jingsheng Lin. 2019. Study on medical image report generation based on improved encoding-decoding method. In *Intelligent Computing Theories and Application*. 686–696.
[143] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the International Conference on Machine Learning (ICML'17)*. 3319–3328.
[144] Ponkrshnan Thiagarajan, Pushkar Khairnar, and Susanta Ghosh. 2021. Explanation and use of uncertainty obtained by Bayesian neural network classifiers for breast histopathology images. *IEEE Transactions on Medical Imaging* 41, 4 (2021), 815–825.
[145] Erico Tjoa and Cuntai Guan. 2020. A survey on explainable artificial intelligence (XAI): Towards medical XAI. *IEEE Transactions on Neural Networks and Learning Systems* 32, 11 (2020), 4793–4813.
[146] Sana Tonekaboni, Shalmali Joshi, Melissa D. McCradden, and Anna Goldenberg. 2019. What clinicians want: Contextualizing explainable machine learning for clinical end use. In *Proceedings of the Machine Learning for Healthcare Conference*. 359–380.
[147] Philipp Tschandl, Giuseppe Argenziano, Majid Razmara, and Jordan Yap. 2019. Diagnostic accuracy of content based dermatoscopic image retrieval with deep classification features. *British Journal of Dermatology* 181, 1 (2019), 155–165.
[148] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. 2018. The HAM10000 dataset: A large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data* 5, 1 (2018), 1–9.
[149] Nikos Tsiknakis, Eleftherios Trivizakis, Evangelia E. Vassalou, Georgios Z. Papadakis, Demetrios A. Spandidos, Aristidis Tsatsakis, Jose Sánchez-García, Rafael López-González, Nikolaos Papanikolaou, Apostolos H. Karantanas, et al. 2020. Interpretable artificial intelligence framework for COVID-19 screening on chest X-rays. *Experimental and Therapeutic Medicine* 20, 2 (2020), 727–735.
[150] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* 9, 11 (2008).
[151] Bas H. M. van der Velden, Hugo J. Kuijf, Kenneth G. A. Gilhuijs, and Max A. Viergever. 2022. Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. *Medical Image Analysis* 79 (2022), 102470.
[152] Michael Van Lent, William Fisher, and Michael Mancuso. 2004. An explainable artificial intelligence system for small-unit tactical behavior. In *Proceedings of the National Conference on Artificial Intelligence (AAAI'04)*. 900–907.
[153] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*. 5998–6008.
[154] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'15)*. 4566–4575.
[155] Linda Wang, Zhong Qiu Lin, and Alexander Wong. 2020. COVID-Net: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images. *Scientific Reports* 10, 1 (2020), 1–12.
[156] Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, Kathryn Funk, Rodney Kinney, Ziyang Liu, William Merrill, et al. 2020. CORD-19: The COVID-19 open research dataset. *ArXiv* (2020).
[157] Sutong Wang, Yunqiang Yin, Dujuan Wang, Yanzhang Wang, and Yaochu Jin. 2021. Interpretability-based multimodal convolutional neural networks for skin lesion diagnosis. *IEEE Transactions on Cybernetics* (2021).
[158] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald Summers. 2017. 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 (CVPR'17)*. 2097–2106.
[159] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, and Ronald Summers. 2018. TieNet: Text-image embedding network for common thorax disease classification and reporting in chest X-rays. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'18)*. 9049–9058.
[160] Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. 2023. METransformer: Radiology report generation by Transformer with multiple learnable expert tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'23)*. 11558–11567.

[161] Zhanyu Wang, Mingkang Tang, Lei Wang, Xiu Li, and Luping Zhou. 2022. A medical semantic-assisted Transformer for radiographic report generation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part III*. 655–664.

[162] Sandareka Wickramanayake, Wynne Hsu, and Mong Li Lee. 2021. Comprehensible convolutional neural networks via guided concept learning. In *Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN'21)*. 1–8.

[163] Sandareka Wickramanayake, Wynne Hsu, and Mong Li Lee. 2021. Explanation-based data augmentation for image classification. In *Advances in Neural Information Processing Systems*, Vol. 34.

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

[165] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning (ICML'15)*. 2048–2057.

[166] Yesheng Xu, Ming Kong, Wenjia Xie, Runping Duan, Zhengqing Fang, Yuxiao Lin, Qiang Zhu, Siliang Tang, Fei Wu, and Yu-Feng Yao. 2021. Deep sequential feature learning in clinical image classification of infectious keratitis. *Engineering* 7, 7 (2021), 1002–1010.

[167] Siyuan Yan, Zhen Yu, Xuelin Zhang, Dwarikanath Mahapatra, Shekhar S. Chandra, Monika Janda, Peter Soyer, and Zongyuan Ge. 2023. Towards trustable skin cancer diagnosis via rewriting model's decision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'23)*. 11568–11577.

[168] Shuxin Yang, Xian Wu, Shen Ge, Zhuozhao Zheng, S. Kevin Zhou, and Li Xiao. 2023. Radiology report generation with a learned knowledge base and multi-modal alignment. *Medical Image Analysis* (2023), 102798.

[169] Changchang Yin, Buyue Qian, Jishang Wei, Xiaoyu Li, Xianli Zhang, Yang Li, and Qinghua Zheng. 2019. Automatic generation of medical imaging diagnostic report with hierarchical recurrent neural network. In *Proceedings of the IEEE International Conference on Data Mining (ICDM'19)*. 728–737.

[170] Kyle Young, Gareth Booth, Becks Simpson, Reuben Dutton, and Sally Shrapnel. 2019. Deep neural network or dermatologist? In *Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support*. 48–55.

[171] Hang Yu, Laurence T. Yang, Qingchen Zhang, David Armstrong, and M. Jamal Deen. 2021. Convolutional neural networks for medical image analysis. *Neurocomputing* 444 (2021), 92–110.

[172] Mert Yuksekgonul, Maggie Wang, and James Zou. 2022. Post-hoc concept bottleneck models. *arXiv preprint arXiv:2205.15480* (2022).

[173] Matthew D. Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *Proceedings of the European Conference on Computer Vision (ECCV'14)*. 818–833.

[174] Junsan Zhang, Xiuxuan Shen, Shaohua Wan, Sotirios K. Goudos, Jie Wu, Ming Cheng, and Weishan Zhang. 2023. A novel deep learning model for medical report generation by inter-intra information calibration. *IEEE Journal of Biomedical and Health Informatics* (2023).

[175] Zizhao Zhang, Yuanpu Xie, Fuyong Xing, Mason McGough, and Lin Yang. 2017. MDNet: A semantically and visually interpretable medical image diagnosis network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'17)*. 6428–6436.

[176] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. 2016. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'16)*. 2921–2929.

Received 9 June 2022; revised 25 August 2023; accepted 18 September 2023