sensors

Review

Survey of Explainable AI Techniques in Healthcare

Ahmad Chaddad [1,2,*], Jihao Peng [1,†], Jian Xu [1,†] and Ahmed Bouridane [3]

[1] School of Artificial Intelligence, Guilin University of Electronic Technology, Jinji Road, Guilin 541004, China
[2] The Laboratory for Imagery Vision and Artificial Intelligence, Ecole de Technologie Superieure, 1100 Rue Notre Dame O, Montreal, QC H3C 1K3, Canada
[3] Centre for Data Analytics and Cybersecurity, University of Sharjah, Sharjah 27272, United Arab Emirates
* Correspondence: ahmadchaddad@guet.edu.cn
† These authors contributed equally to this work.

Abstract: Artificial intelligence (AI) with deep learning models has been widely applied in numerous domains, including medical imaging and healthcare tasks. In the medical field, any judgment or decision is fraught with risk. A doctor will carefully judge whether a patient is sick before forming a reasonable explanation based on the patient's symptoms and/or an examination. Therefore, to be a viable and accepted tool, AI needs to mimic human judgment and interpretation skills. Specifically, explainable AI (XAI) aims to explain the information behind the black-box model of deep learning that reveals how the decisions are made. This paper provides a survey of the most recent XAI techniques used in healthcare and related medical imaging applications. We summarize and categorize the XAI types, and highlight the algorithms used to increase interpretability in medical imaging topics. In addition, we focus on the challenging XAI problems in medical applications and provide guidelines to develop better interpretations of deep learning models using XAI concepts in medical image and text analysis. Furthermore, this survey provides future directions to guide developers and researchers for future prospective investigations on clinical topics, particularly on applications with medical imaging.

Keywords: explainable AI; medical imaging; deep learning; radiomics

check for updates

Citation: Chaddad, A.; Peng, J.; Xu, J.; Bouridane, A. Survey of Explainable AI Techniques in Healthcare. *Sensors* **2023**, *23*, 634. https://doi.org/10.3390/s23020634

Academic Editors: Samaneh Kouchaki, Xiao-Rong Ding and Saeid Sanei

Received: 7 November 2022
Revised: 14 December 2022
Accepted: 29 December 2022
Published: 5 January 2023

1. Introduction

Currently, artificial intelligence, which is widely applied in several domains, can perform well and quickly. This is the result of the continuous development and optimization of machine learning algorithms to solve many problems, including in the healthcare field, making the use of AI in medical imaging one of the most important scientific interests [1]. However, AI based on deep learning algorithms is not transparent, making clinicians uncertain about the signs of diagnosis. The key question then is how one can provide convincing evidence of the responses. However, there exists a gap between AI models and human understanding, currently known as "black-box" [2] transparency. For this reason, many research works focus on simplifying the AI models for better understanding by clinicians, in order to improve confidence in the use of AI models [3]. For example, the Defense Advanced Research Projects Agency (DARPA) of the United States developed the explainable AI (XAI) model in 2015. Later, in 2021, a trust AI project showed that the XAI can be used in interdisciplinary types of application problems, including psychology, statistics, and computer science, and may provide explanations that increase the trust of users [4].

Typically, XAI is an explainable model providing insights into how the predictions are made to achieve trustworthiness, causality, transferability, confidence, fairness, accessibility, and interactivity [5,6]. For example, as shown in Figure 1, it is strongly recommended to allow the AI model to be understandable for the public when the model outputs a decision. It is noted that the definition of XAI is not clear enough according to [7]. In addition, the two words "explainable" and "interpretable" are associated with XAI terms, by which the black-box models are considered "explainable" when the predictions are considered post hoc methods. An "interpretable" model based on the model itself aims to provide human-understandable outputs as steps [8]. It is also important to note that the definition of explainability depends on the prediction task, as mentioned in [9]. Therefore, the term explainable may be measured based on the target users rather than by uniform standards.

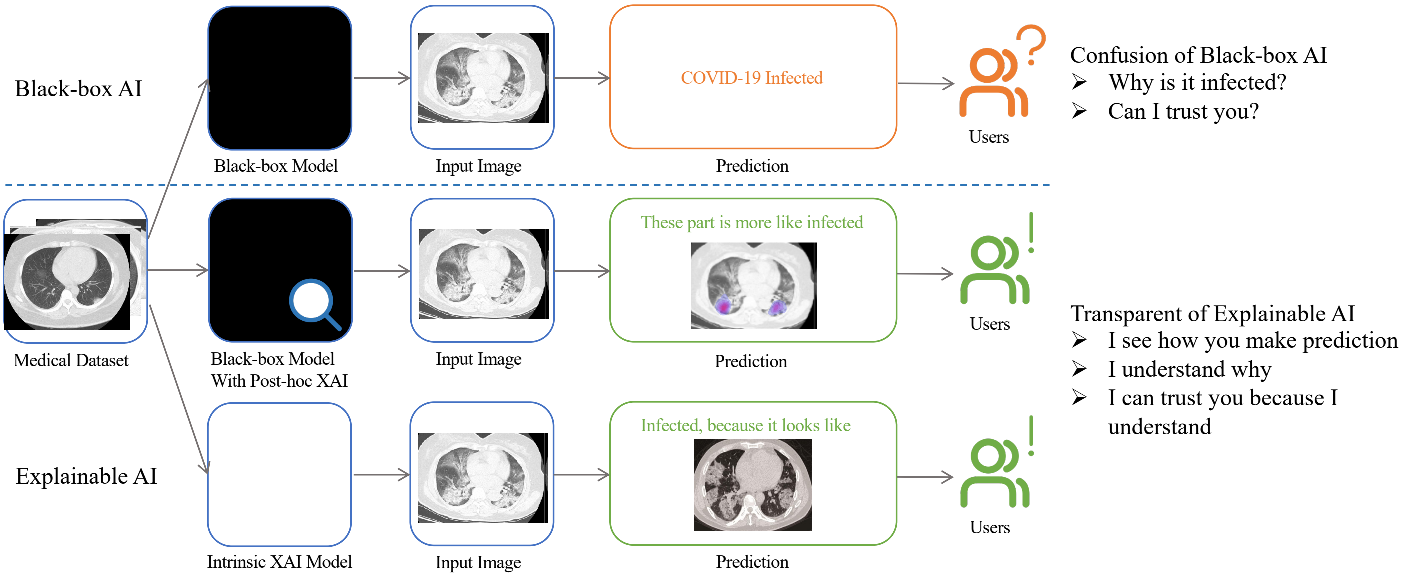

Figure 1. Flowchart of visual comparison between black-box and explainable artificial intelligence, and how the results affect the user. The top branch shows the process of a black-box model. Typically, it provides only results such as classes (e.g., COVID or non-COVID). The other two branches (middle and bottom) represent two XAI methods, referred to in Section 3.1. Specifically, the XAI model (middle) shows the example of saliency map, and the second one (bottom) is the prototype method, as explained in Sections 4.1 and 4.6, respectively. An example of a CT image obtained from the COVID-19 CT scan dataset [10].

In recent years, research outputs of XAI have significantly increased, especially in medical fields, as illustrated in Figure 2. For example, the development of deep learning (DL) models for healthcare has resulted in advanced performance, such as the U-Net model in image segmentation [11]. Despite this progress, DL models still face challenges in clinical practice. Reasons may be related to the inherent high risks in medical decisions. In this context, patients and clinicians are interested to know more about AI-based decisions. In other words, the AI black box is increasingly being used in a variety of high-risk fields where potentially irrational decisions will lead to serious consequences. Therefore, more investigation on XAI is recommended to provide answers to many clinical questions related to accurate and fast diagnosis.

In short, the primary goal of the XAI model is related to people trusting the AI model. For instance, AI users can be divided into two groups: (1) those who have AI "expertise" and (2) those who do not. The first group relates to experts, algorithm developers, and researchers. They focus more on the AI model itself; they develop new methods to monitor the information flow of an algorithm, and explain and optimize the mechanism of an algorithm. The second group of users is generally made up of domain experts, such as radiologists, and the public at large. The expert clinicians require more explanations about AI models to have a technical understanding. Collaborative work between academic and clinical researchers is recommended.

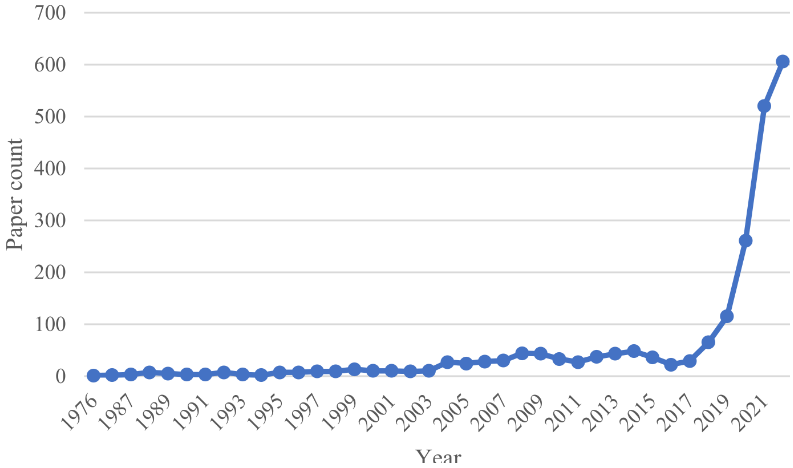

Figure 2. Number of XAI publications added per year from 1976 to 2021. The x axis represents the publishing year, and the y axis shows the number of publications added in a certain year. The number of publications indexed on PubMed (accessed on 1 July 2022: https://pubmed.ncbi.nlm.nih.gov) that matched the search queries related to explainable AI topics in this survey with a term searched of (explainable AI OR explainable artificial intelligence) AND (medicine OR healthcare).

2. XAI Techniques Related to Medical Imaging

To trust AI models, the European Union has proposed seven key requirements, including (1) human agency and oversight; (2) technical robustness and safety; (3) privacy and data governance, (4) transparency; (5) diversity, non-discrimination, and fairness; (6) social and environmental well-being; and (7) accountability [12]. These seven requirements are summarized as follows.

2.1. Confidentiality and Privacy

AI systems require updating data in real-time, which is a systemic risk. The black-box nature of AI can cause many security problems, which can come from internal or external sources. For example, these problems may be related to the algorithm itself or external, such as improper use of users and the creation of false datasets by network attacks [13,14]. In [15], the authors identified three types of security risks related to black-box AI: network attacks, system bias, and mismatch attacks. These threats can have serious consequences for medical systems. For these reasons, several studies have investigated and proposed solutions to XAI models in terms of data security [16,17].

2.2. Ethics and Responsibilities

The medical field has put forward more requirements for the use of AI and how to clarify the ethics and responsibility of AI has become a challenge. For example, irresponsible AI may lead to a loss of medical staff and patients [18]. Furthermore, it involves the ethical issues of data privacy used with AI models. At present, the inspection and accountability of AI are in the early stages. More details about these problems and the interpretable roles of AI can be found in [19,20]. In addition, the concept of responsible AI is discussed and analyzed to develop notions of responsibility for technological domains [21]. Thus, explainability may be an important condition for clarifying these AI responsibilities.

2.3. Bias and Fairness

An AI model is trained by datasets having their own inherent attributes, thereby containing hidden bias. For example, in age and skin-color recognition applications using collected portraits of all ages and races, the AI model showed a preference for light skin and a 45-year-old person [22]. Furthermore, the potential risks in big data algorithms that cannot be ignored allow for increased bias and replication or exacerbate human errors [23]. Notably, the bias may be derived from data, algorithms, and user interactions [24]. This will seriously affect the fairness of AI models, causing different people to get different results. In the case of clinical medicine, different patients may show different symptoms that affect the algorithm. Therefore, it is important to foresee the impact of the AI algorithm's bias in medical healthcare. This should be carefully investigated when deploying AI in personalized medicine.

3. Explainable Artificial Intelligence Techniques

This section provides a brief overview of the categories of XAI that can be used in healthcare. According to the literature published in recent years, there are many criteria used to classify XAI methods [25–27]. Figure 3 shows the criteria for classifying XAI methods and the corresponding categories. Based on these categories, it can be summarized that the most commonly used XAI techniques in medical fields are shown in Table 1. In addition, Table 2 reports the recent papers using the XAI method. For readability purposes, we have divided the table into explainable methods, modalities, and explanations of how explainable methods are applied. We have categorized the XAI techniques and explained in detail the methods as follows.

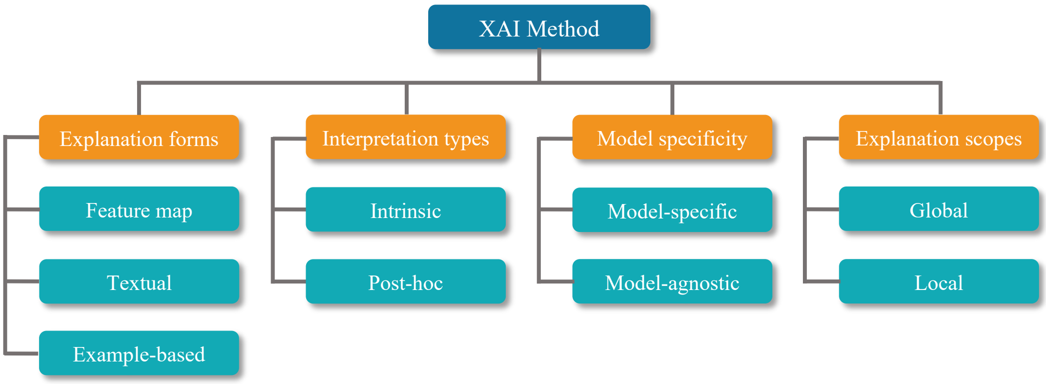

Figure 3. Categorization of explainable AI methods used in this paper. These criteria and categories are summarized from [25–27]. The orange and blue grids represent the criteria and categories, respectively.

Table 1. Summary of explainable AI techniques classified according to Section 3.

Explanation Type	Paper	Technique	Intrinsic	Post Hoc	Global	Local	Model-Specify	Model-Agnostic
Feature	[28]	BP		*		*	*	
	[29]	Guided-BP		*		*	*	
	[30]	Deconv Network		*		*	*	
	[31]	LRP		*		*	*	
	[32]	CAM		*		*	*	
	[33]	Grad-CAM		*		*	*	
	[34]	LIME		*		*		*
	[35]	GraphLIME		*		*		*
	[36]	SHAP		*		*		*
	[37]	Attention	*			*	*	
Example-based	[38]	ProtoPNet	*			*	*	
	[39]	Triplet Network	*		*	*	*	
	[5]	xDNN	*			*	*	
Textual	[40]	TCAV		*	*	*		*
	[41]	Image Captioning	*			*	*	

"*" indicates it belongs to this category, which is defined in Section 3, BP: backpropagation, CAM: class activation map, LRP: layer-wise relevance propagation, LIME: local interpretable model-agnostic explanations, MuSE: model usability evaluation, SHAP: Shapley additive explanations, xDNN: explainable deep neural network, TCAV: testing with concept activation vectors.

Table 2. The following table is a collection of papers that have used interpretable methods in Section 4 to improve the algorithm.

Paper	Organ	XAI	Modality	Contribution
[42]	bone	CAM	X-ray	The model aims to predict the degree of knee damage and pain value through X-ray image.
[43]	lung	CAM	Ultrasound, X-ray	It uses three kinds of lung ultrasound images as datasets, and two networks, VGG-16 and VGG-CAM, to classify three kinds of pneumonia.
[44]	breast	CAM	X-ray	It proposes a globally-aware multiple instance classifier (GMIC) that uses CAM to identify the most informative regions with local and global information.
[45]	lung	CAM	X-ray, CT	The study improves two models, one of them based on MobileNet to classify COVID-19 CXR images, the other one is ResNet for CT image classification.
[46]	lung	CAM	CT	It selects healthy and COVID-19 patient's data for training DRE-Net model.
[47]	lung	Grad-CAM	CT	It proposes a method of deep feature fusion. It achieves better performance than the single use of CNN.
[48]	chest	Grad-CAM	ultrasound	The paper proposes a semi-supervised model based on attention mechanism and disentangled. It then uses Grad-CAM to improve model's explainable.
[49]	lung	Grad-CAM	X-ray	It provides a computer-aided detection, which is composed of the Discrimination-DL and the Localization-DL, and uses Grad-CAM to locate abnormal areas in the image.
[50]	colon	Grad-CAM	colonoscopy	The study proposes DenseNet121 to predict if the patient has ulcerative colitis (UC).
[51]	colon	Grad-CAM	whole-slide images	It investigates the potential of a deep learning-based system for automated MSI prediction.
[52]	lung	Grad-CAM	CT	It shows a classifier based on the Res2Net network. The study uses Activation Mapping to increase the interpretability of the overall Joint Classification and Segmentation system.
[53]	chest	Grad-CAM	CT	It proposes a neighboring aware graph neural network (NAGNN) for COVID-19 detection based on chest CT images.
[54]	lung	Grad-CAM, LIME	X-ray	This work provides a COVID-19 X-ray dataset, and proposes a COVID-CXNet based on CheXNet using transfer learning.
[55]	lung	Grad-CAM, LIME	X-ray, CT	It compares five DL models and uses the visualization method to explain NASNetLarge.
[56]	breast	Attention	X-ray	It provides the triple-attention learning A^3 Net model to diagnose 14 chest diseases.
[57]	bone	Attention	CT	The study introduces a multimodal spatial attention module (MSAM). It uses an attention mechanism to focus on the area of interest.
[58]	colon	Attention	colonoscopy	The proposed Focus U-Net achieves an average DSC and IoU of 87.8% and 80.9%, respectively.
[59]	lung, skin	Saliency	CT, X-ray	The work presents quantitative assessment metrics for saliency XAI. Three different saliency algorithms were evaluated.
[60]	lung	SHAP	EHR	The study introduces a predictive length of stay framework to deal with imbalanced EHR datasets.
[61]	-	SHAP	EHR	The study presents an explainable clinical decision support system (CDSS) to help clinicians identify women at risk for Gestational Diabetes Mellitus (GDM).
[62]	-	SHAP	radiomics	The study proposes a pipeline for interactive medical image analysis via radiomics.
[63]	lung	SHAP	CT	This paper provides a model to predict mutation in patients with non-small cell lung cancer.
[64]	chest	SHAP	EHR	In this paper, it compares the performance of different ML methods (RSFs, SSVMs, and XGB and CPH regression) and uses SHAP value to interpret the models.
[65]	chest	LIME, SHAP	X-ray	The study proposes a unified pipeline to improve explainability for CNN using multiple XAI methods.
[66]	lung	SHAP, LIME, Scoped Rules	EHR	The study provides a comparison among three feature-based XAI techniques on EHR dataset. The results show that the use of these techniques can not replace human experts.
[67]	chest	Image caption	CT	It proposes Medical-VLBERT for COVID-19 CT report generation.

CAM: class activation map, AUC: area under the ROC curve, ROC: receiver operating characteristic curve, LIME: local interpretable model-agnostic explanation, EHR: electronic health record, SHAP: Shapley additive explanations.

3.1. Interpretation Types

Two major types of XAI models are represented by intrinsic and post hoc models. These are two different fields of XAI, which is reflected in the way they operate. Intrinsic is also known as model-based interpretability, where the model itself is interpretable by adjusting the structure and/or components. Post hoc models provide an explanation for a trained model by analyzing the original model and an additional one. Despite the widely used post hoc analysis, its application in the medical field requires more effort to become more practical [68]. The intrinsic and post hoc explanations can be summarized as follows.

3.1.1. Intrinsic Explanation

This type of model is structured to be understandable. In this context, conventional models, such as linear regression models, are relatively simple in structure, though capable of being understandable. For example, a model provides the answer along with the corresponding explanation of the linguistic interpreter [69].

3.1.2. Post Hoc Explanation

Post hoc explanation is related to the interpretable information obtained by external methods when analyzing the model (e.g., a neural network after it has been trained). For example, backpropagation [28], class activation mapping [32], and layer-wise relevance propagation [31] are all post hoc interpretation techniques.

3.2. Model Specificity

XAI models can be grouped as model-specific and model-agnostic, depending on whether the method can be used in multiple types of models.

3.2.1. Model-Specific Explanation

A model-specific approach is applied to a certain scope of application. For example, the method needs to use a particular structure or property in the model. Furthermore, all intrinsic models are model-specific because all of these methods require the use of the structure of the model itself [70].

3.2.2. Model-Agnostic Explanation

A model-agnostic approach has no special requirements for the model. Rather, it is used in most XAI models. For example, in perturbation-based methods such as local interpretable model-agnostic explanations (LIME) [34], several outputs are obtained and interpreted by perturbing the model inputs. This may be classified as post hoc type.

3.3. Explanation Scopes

XAI may be explained over the entire model or for specific inputs and outputs. In this context, two types of explanation can exist: (1) global explanations and (2) local explanations.

3.3.1. Local Explanation

This type considers the model as a black box, focusing on the local variables that contribute to the decision. This leads to the determination of the features which contribute to the decision-making process. Generally, a local explanation focuses on a single input dataset and the characteristic variables associated with it [71].

3.3.2. Global Explanation

This explanation type is interpreted from the model itself. For example, it explains the contribution that relates to the output by getting an understanding of the interaction mechanism of the model variables. This can be formulated as “How does the model predict?” Interpreting the model by a global method depends on the performance of the model. One can generalize the local interpretation to provide an appropriate global interpretation of the model [71].

3.4. Explanation Forms

Explanation forms generated by XAI methods can be divided into three main types: feature-based, textual, and example-based. It is worth noting that some methods can generate multiple types of explanations.

3.4.1. Feature Map

Feature-based explanations present the gradient or hidden feature map values as an approximation of the input importance. They can show which part has the greatest impact on the final output. They are usually represented by the original image with an overlay of a saliency map [28,30].

3.4.2. Textual Explanation

This is a human-comprehensible explanation generated in textual form. Semantic descriptions are used to explain the decision of the model. In an example of image captioning, textual explanations are generated in addition to visual interpretations [72].

3.4.3. Example-Based Explanation

This aims to explain a model by presenting one or more examples similar to the given one. The prototypes consist of the features extracted during network training and are designed as examples [38,73].

4. Introduction of the Explainable AI Method: A Brief Overview

As mentioned previously, XAI is widely used in many fields, in particular, medical imaging. In this section, we focus on the importance of XAI in healthcare applications.

4.1. Saliency

Saliency directly uses the squared value of the gradient as the importance score of different input features [28]. The input can be graph nodes, edges, or node features. It assumes that the higher gradient value is related to the most important features. Although it is simple and efficient, it has several limitations. For example, it can only reflect the sensitivity between the input and output, which cannot express the importance very accurately. In addition, it has a saturation problem. For example, in regions where the performance model reaches saturation, the change in its output relative to any input change is very small, and the gradient can hardly reflect the degree of input contribution.

Guided backpropagation (BP), whose principle is similar to that of the saliency map, modifies the process of backpropagating the gradient [29]. Since the negative gradients are hard to interpret, guided BP only back-propagates the positive gradients and shears the negative gradients to zero. Therefore, guided BP has the same limitations as saliency maps.

One approach to avoid these limitations is to use layer-wise relevance propagation (LRP) [31] and deep Taylor decomposition (DTD) [74]. LRP and DTD are capable of improving a model's interpretability. In DTD, neural networks use complex non-linear functions that are represented by a series of simple functions. In LRP, the relevance of each neuron in the network is propagated backward through the network, thereby allowing it to quantify the contribution of each neuron to the final output. There are several rules designed with a specific type of layer in a neural network [31,74]. To combine LRP and DTD, LRP can be thought of as providing the framework for propagating relevance through a network, whereas DTD provides the means for approximating the complex non-linear functions used by the network. LRP and DTD may lead to overcoming the limitations of saliency maps and provide more accurate explanations [75].

4.2. Class Activation Mapping

Class activation mapping (CAM) is a visualization tool based on convolutional neural networks [32]. It is capable of distinguishing the focus area of the network by obtaining the weight W_k^c of the recognized feature image F_k in the network, where k is the unit of the global average pooling layer and c represents the class. The convolution between these two parameters F_k and W_k^c provides the feature map $M_c(x, y)$, where x and y represent the location (x, y) of the convolutional layer. All types of CAM deformation methods are based on graph activation and weights. However, several methods can be used to obtain the weight value. A good explanation of the CAM model is reported in Algorithm 1. Although this is one of the most commonly used algorithms, it has some challenges. For example, the network's structure requires more flexibility so that the fully connected layer may adapt to the global average pooling layer. For this reason, a new algorithm known as "Gradient-CAM" is proposed to optimize CAM. It uses gradients to compute weight values [33]. First, the network is propagated forward to obtain the feature layer A (e.g., output of the last convolutional layer) and the network predicted value Y (e.g., output value before softmax activation). Then, the weights a are obtained by computing the backpropagation. Finally, the Grad-CAM matrix L may be obtained according to $L^c_{Grad-CAM} = \text{ReLU}\left(\sum_k a^c_k A^k\right)$.

Algorithm 1 Class activation mapping.

```
Require: Image $I^C(H,W)$; Network $N$
Ensure: Replace FC layer with average pooling layer in Network $N$
  procedure CAM(I, N)
      N($I$)                                                        ▷ Input image into network
      $W^c_k \leftarrow (w_1, w_2, w_3, ..., w_k)$                  ▷ Get weights from average polling layer
      $F^c_k \leftarrow (f_1(x,y), f_2(x,y), f_3(x,y), ..., f_k(x,y))$   ▷ Feature map of the last convolution layer
      $M_c(x,y) = \sum_k w^c_k f_k(x,y)$                            ▷ Weighted linear summation
      $M_c(x,y) = \frac{1}{HW} \sum_i^H \sum_j^W M_c(x,y)$          ▷ Normalize and up-sample to Network input size
      $M_c(x,y) = \text{ReLU}(M_c(x,y))$                            ▷ Final image heat map
  end procedure
```

4.3. Occlusion Sensitivity

When training a neural network for image classification, the aim is to know whether this model can locate the position of the main target in the image. By partially occluding the picture, one can observe the situations of the network in the middle layers and the change in the predicted value after inputting the modified image. This leads to an understanding of why the network makes certain decisions. So far, occlusion sensitivity refers to how the probability of a given prediction changes with the occluded part(s) of the image. The higher the output image value, the greater the decrease in the degree of certainty, indicating that the occlusion area is more important in the decision-making process [30].

4.4. Testing with Concept Activation Vectors

Testing with the concept activation vectors (TCAV) is an interpretable method proposed by the Google AI team [40]. Textual concepts are related to an explanation that is simple to understand. In the saliency map, it is not possible to explain the concept of pixels. For this reason, TCAV focuses on capturing high-level concepts in the neural network and attempts to provide a linear transformation from input to concepts using directional derivatives to quantify the importance of user-defined concepts to the classification results. However, this technique requires more investigation to be feasible in medical applications.

4.5. Triplet Networks

The triplet network (TN) concept is an example-based framework [39]. For example, the TN training set consists of three samples: the first is randomly chosen from the "Anchor" training set, while the other two samples are randomly chosen from the training set in the same "Positive" and different "Negative" categories. By adjusting the parameters based on the distance between three inputs, the technique aims to bring the Anchor closer to the Positive and away from the Negative. Since labeling is not necessary, this method can be used for unsupervised learning. The technique is able to provide an explanation through the similarity between samples too.

4.6. Prototypes

A prototype operation can be used as a method to compare the similarity between the target and a typical sample in a category. This sample is known as "prototype" of its class, and it may be easy to understand and compare for users. For example, each unit in the layer stores a weight vector representing an encoded input. The weight can be seen as the feature part of the object. It is interpretable because it makes decisions based on the weighted similarity score extracted from input and prototypes [73]. An example is the "ProtoPNet" CNN model, which has an additional prototype layer used for image classification. This layer takes the output of the previous convolutional layer as input and learns the class prototypes during the training process [38]. The interpretability of ProtoPNet is described as, "this looks like that", in image classification tasks.

The explainable deep neural network (xDNN) is another prototype-based network for image classification [5]. Two out of five layers are prototypes: (1) prototype and (2) MegaClouds layers. The prototype layer can extract the data distribution and then form the linguistic logical rules if … then … for the explanations as follows.

$$\text{if } (I \sim \hat{I}_P) \text{ then } (\textit{class } c) \tag{1}$$

where I is the input image and $\hat{I}_P$ represents the prototype. The prototypes that have the same class label are merged into a MegaCloud M in the MegaClouds layer. The final expression is as follows.

$$\text{if } (x \sim M_1) \text{ or } (x \sim M_2) \text{ or } \dots \text{ or } (x \sim M_M) \text{ then } (\textit{class } c) \tag{2}$$

xDNN was evaluated using multiple datasets, including the COVID-CT dataset [10] and the SARS-CoV-2 CT scan dataset [76]. xDNN achieved slightly lower performance metrics (accuracy, sensitivity, and F1-score) compared to the neighboring aware graph neural network (NAGNN), which uses the post hoc method Grad-CAM to interpret the predictions [53].

4.7. Trainable Attention

Trainable attention is defined as a group of techniques that focus on important information content in digital multimedia data (e.g., image, video, audio, and text). For example, the combination of text and image information in a hidden layer may let clinicians focus on the corresponding information between the ROI in an image and electronic health records (EHR) [77]. In addition, a multilayer visual attention mechanism can be used to interpret medical image analysis. For example, two attention layers, one close to the input and the other close to the output, can be used for explaining the attention mechanism between both input and output [78]. Despite the recent works related to this trainable attention, more work using this technique is needed to provide a niche of visual information to clinicians.

4.8. Shapley Additive Explanations

The Shapley additive explanation (SHAP), which is also a model using Shapley values [36,79], evaluates the importance of an input feature for the final prediction. This model requires much time to calculate the SHAP values. It may combine with other techniques to accelerate the computation of SHAP values [80]. For example, deep explainer (i.e., deep SHAP) is a fast explainability technique that is considered for models with a neural network-based architecture. Generally, SHAP is used to provide explanations and may be used for many clinical topics [61]. However, SHAP applications are still limited to specific problems.

4.9. Local Interpretable Model-Agnostic Explanations

Local interpretable model-agnostic explanation (LIME) is a local model-agnostic method that aims to provide an interpretation of the original model by approximating a new simple model from the predictions of a black-box model locally. The new model is used to interpret the results obtained [34]. This advantage allows LIME to be used with any black-box model to interpret only a single prediction. For example, the approximation model is trained as follows. Given a black-box model and the input data, a perturbation is added to the input data, either by overwriting parts of an image or by removing parts of the words from the text. These new samples are then weighted according to their proximity to the corresponding class, an interpretable model is trained on the obtained dataset, and finally, an explanation is obtained by interpreting the model. The explanation produced by the surrogate model can be expressed through Equation (3).

$$\text{explain}(x) = \text{argmin}_{g \in G} L(f, g, \pi_x) + \Omega(g) \tag{3}$$

where g is a surrogate model in all possible interpretable models and G is used to explain the instance x. L is a fidelity function that measures how close the explanation is to the predictions of the original model. $\Omega(g)$ infers the complexity of the model g, since we want the surrogate model to be interpretable by humans. It minimizes L while keeping model complexity, $\Omega(g)$, low.

For example, using the least absolute shrinkage and selection operator (Lasso) as a factor of an interpretable model g for text classification, LIME can be expressed by K-LASSO with sparse linear explanations, as illustrated in Algorithm 2. Lasso can select features and regularize the weights of features in a linear model. K is the number of features selected via Lasso.

Algorithm 2 Sparse linear explanation using LIME.

```
Require: Classifier $f$; Features number $K$; Instance $x$ to be explained; Similarity kernel $\pi_x$
  $\mathcal{Z} = \{\}$
  $\mathcal{X}_N = \text{SAMPLE\_AROUND}(x')$
  for $x_i$ in $\mathcal{X}_N$ do
      $\mathcal{Z} = \mathcal{Z} \cup (x_i, f(x_i), \pi_x(x_i))$
  end for
  $\omega$ = K-LASSO($\mathcal{Z}, \mathcal{K}$)
  return $\omega$                                  ▷ Explanation for an individual predict
```

Another algorithm called "GraphLIME" can be used for graph neural networks for classification applications. It extends LIME to work in a non-linear environment by sampling N-hop network neighbors and using the Hilbert–Schmidt independence criterion Lasso (HSIC Lasso) as surrogate models [35]. The explanations for "GraphLIME" have the same patterns as "LIME", as expressed in Equation (4)

$$\text{explain}(v) = \text{argmin}_{g \in G} L(f, X_n) \tag{4}$$

The difference is that X_n represents the sampling local information matrix of node v in a graph.

4.10. Image Captioning

Image captioning is an intrinsic explanation method that finally provides interpretations of natural languages [41]. Typically, this approach combines CNNs and long short-term memory networks (LSTM) for encoding the image and text, respectively [81]. These techniques are useful to generate medical reports. For example, "TandemNet" can generate visual interpretations in addition to textual explanation [72]. "TandemNet" is a dual attention model that can effectively combine image and text information, extract useful features, and focus attention for accurate image prediction. Medical Visual-Linguistic BERT, Medical-VLBERT, is another algorithm that has been used to generate medical reports of COVID-19 patients [67]. In this context, a curriculum learning framework, competence-based multimodal curriculum learning (CMCL), was proposed to solve the lack of generation of medical reports [82]. It is worth noting that CMCL mimics the learning process of human doctors from an easy to hard approach by scoring the learning difficulty of the training samples and selecting the appropriate difficulty samples for learning at different training stages of the model.

4.11. Recent XAI Methods

Based on the recent literature related to XAI, as reported in Tables 2 and 3, these XAI models can be summarized as follows.

Table 3. The advantages and disadvantages of the XAI technique. The letters in Code refer to URLs in the footnotes.

Paper	Technique	Simple to Use	Stability	Efficient	Trustworthy	Code	Feature
[33]	Gradient-weighted class activation mapping (Grad-CAM)	+	-	+	-	c1	• Works with any CNN
[34]	Local Interpretable Model-agnostic Explanations (LIME)	+	-	-	+	c2	• Works on text, image, and tabular data • Uses a simple model for an explanation, but complexity must be defined beforehand
[35]	GraphLIME	-	-	-	+	* c3	• Works with GNN
[36]	SHapley Additive exPlanations (SHAP)	+	-	-	na.	c4	• Have a theoretical foundation from Shapley value
[37]	Trainable attention	-	na.	-	+/-	* c5	• Strong anti-noise ability
[5]	xDNN	-	na.	-	+	c6	• Features a prototype and Megacloud layer that can effectively extract prototypes
[40]	Testing with Concept Activation Vectors (TCAV)	+	na.	na.	+	c7	• Use high-level concept for the explanation
[38]	ProtoPNet	+/-	na.	-	-	c8	• Utilized latent space prototypes • Have semantic differences between latent space and input, may cause errors in explanation
[41]	Image Caption	+/-	na.	-	+/-	na.	• Provide a textual explanation • Multiple types of data required

"+" advantage; "-" disadvantage; "*" not official; "na." unavailable; c1 https://github.com/Cloud-CV/Grad-CAM; c2 https://github.com/marcotcr/lime; c3 https://github.com/WilliamCCHuang/GraphLIME; c4 https://github.com/slundberg/shap; c5 https://github.com/SaoYan/LearnToPayAttention; c6 https://github.com/Plamen-Eduardo/xDNN---Python; c7 https://github.com/tensorflow/tcav; c8 https://github.com/cfchen-duke/ProtoPNet. All codes are accessed on 1 September 2022.

Grad-CAM: It can highlight important areas of a saliency map. It can verify the model accuracy by comparing the deviation between important areas and the actual situation [54]. Generally, saliency is used for abnormality localization in medical images, and it is useful when the detection and segmentation problems can be localized in the desired output network [19]. It is widely applied to medical images and has become one of the most popular XAI methods.

LIME: a tool for visual explanation. It can be used to predict the local output. For instance, it is used to provide image explanation [83]. For example, in practical applications, disturbed pixels need to be set according to requirements, and low repeatability limits the interpretability of LIME [84].

SHAP: It is an important tool for analyzing the features. It is usually used to extract features and conduct attribution analysis. Although time-consuming, some features may be relevant to explain the model [64].

Trainable attention: It is a mechanism used for image location and segmentation. For example, interpretability remains to be determined in the dual attention-gated deep neural network [58]. In [85], the attention mechanism weights were seen as at best noisy predictors of the relative importance of specific regions of the input sequence, and they should not be treated as justifications for the model's decisions.

5. Making an Explainable Model through Radiomics

Radiomics is a method to extract features from medical images. These features, known as radiomic features, have the ability to reveal tissue patterns. Radiomic features are used as input into predictive models for clinical classifications [86–88]. Specifically, radiomics can be seen as a multistep process to complete radiomic analysis: (a) image acquisition, (b) image preprocessing, (c) segmentation/labeling leading to identifying regions of interest (ROI), (d) feature extraction and selection, and (e) building predictive models using machine learning [89]. It should be noted that traditional radiomic models consider manual labeling to segment lesions (ROI); this process requires intensive computation and significant effort from radiologists and oncologists to complete the segmentation. With deep learning models, radiomics, also known as deep radiomics, became more practical and was applied in many medical fields, such as pneumonia recognition [54,90]), survival estimation [91–93], and survival prediction [94].

The classification of most medical images is a binary problem (e.g., cancer versus non-cancer) focusing on limited and fixed image features for diagnosis. Thus, it uses saliency maps to highlight the important features. For more complex object classification problems, the network usually requires focusing on more local information. As is known, the detection of disease markers is often expensive, and invasive biopsies require significant analysis and time. Therefore, radiomics is used with a variety of imaging modalities to detect diseases by object detection. In this context, it is desired that the radiomic steps be transparent and explainable. Likewise, DL models for use in medical data analysis should be explainable, as described in [88,90,92,93].

Therefore, there is a need to consider increasing the interpretability of the radiomics diagnosis process. However, radiomics is sensitive to image sampling methods, and different sampling methods affect the sampling characteristics [95]. Improving DL interpretability is critical for the advancement of AI with radiomics. For example, a deep learning predictive model is used for personalized medical treatment [89,92,96]. Despite the wide applications of radiomics and DL models, developing a global explanation model is a massive need for future radiomics with AI.

6. Discussion, Challenges, and Prospects

Most of the recently proposed medical imaging works use post-interpretation rather than model-based interpretation (e.g., CAM and GCAM models are widely used). In fact, these works focus on the application of algorithms, and the interpretable methods are used as a supplement to the algorithms. In the absence of systematic development of XAI, it is a trend to use local interpretation methods to explain the cases studied. In the case of CNNs and their use in medical images, a saliency map is a simple tool for obtaining an explanation of the areas of interest of the network [97]. Eight interpretable methods of saliency maps (+ Grad-CAM, guided backpropagation, and guided Grad-CAM) were evaluated [19]. However, the performance on the testing datasets was not competitive [19]. Therefore, several challenges still face the XAI technique. These are summarized as follows.

6.1. Human-Centered XAI

As AI models involve social decision making, interaction with these models is becoming more important for many users, especially clinicians. The idea is that the interpretation is accepted as an effective tool for communicating with users/persons and AI models.

In a clinical application scenario, XAI provides explanations for doctors and patients. Radiologists also want to know the opinions of others when using AI tools for diagnosis [98]. Obviously, these interpretable methods are not explained to patients. This shows that doctors cannot clearly diagnose diseases in terms of medical treatment. They need strong evidence or more authoritative answers to verify the AI tools. In this case, if the explanations provided by XAI do not meet the doctor's expectations, these explanations will not be taken into account or considered. It would be hard and not acceptable for patients to be informed that they have been diagnosed using a computer-based tool. Providing an incomprehensible explanation to the patient undoubtedly decreases the trustworthiness of AI. To achieve this goal, many researchers are investigating the design of more sophisticated interpretation methods; they may require more time to be partially trusted [99].

At present, the academic community is focused on the human-centered development of interpretable technology [100–102]. Unfortunately, the human–XAI interaction techniques face many challenges that need massive work to be solved [103,104]. In this context, research on human-centered XAI may consider the following: (1) causality (e.g., providing an understandable chain of causal explanations for users), (2) interactivity (e.g., offer explanations from various perspectives, so that users have the option to choose explanations), and (3) counterfactual explanations to enhance human–computer interactions and produces personalized output with the AI model [105].

6.2. AI System Deployment

The use of XAI in clinical decision making gives the models more transparency. However, there are many practical problems related to the speed of operation and implementation [106]. For example, developing, deploying, and applying medical-related AI algorithms involves designers, developers, AI product managers, clinicians, and many other people. This will lead to the development of a new management framework for AI models involving social decision making, and their interaction within the framework to address the desired needs of clinicians and patients alike. The idea is that interpretation is a significant tool for communicating with people and AI models. To achieve this goal, many researchers are looking to design more sophisticated interpretation methods to achieve trustfulness [99]. Currently, the academic community is focused on the human-centered development of interpretable technology [100–102]. Unfortunately, human–XAI interaction techniques also face many challenges. We require significant work to design tools and methods to effectively and appropriately apply AI technology in medical and healthcare care settings [103,104]. Currently, many institutions provide training steps to clinicians to explain the basics of AI models. It is also strongly recommended to generate performance metrics of bias and accuracy for the algorithms to increase the trust level; see Model Cards from Google [107], AI Fact Sheets from IBM [108], and Datasheets for datasets from Microsoft [109]. So far, the most widely required criteria of the XAI model are: (1) easy to use for users, (2) validity, (3) robustness, (4) computational cost, (5) the ability to fine-tune, and (6) open-source development [110].

6.3. Quality of Explanation

When reviewing and investigating papers using the above methods, it was noticed that the XAI functionalities were not as expected when designed. Researchers will face some problems when they use these XAI methods. An XAI model is evaluated by its ability to provide accurate and understandable explanations for its decisions. One can divide the evaluation methods of XAI into two categories: (1) human-centered and (2) computer-centered [111]. In human-centered XAI, the system produces its explanation and is evaluated by human participants. Their feedback is collected and analyzed. However, it requires domain experts such as clinicians to evaluate the explanation performance, making it is highly cost. In computer-centered XAI, the system uses an algorithm to assess the explanation quality of XAI. Among the popular XAI methods, backpropagation has a high rate of success in detecting and recovering from Trojan attacks, particularly for models with large trigger sizes [112]. For example, CAM is effective in detecting the entire trigger region, but may not always provide the accurate localization of the trigger [112]. LIME has a high rate of success in detecting small triggers [112]. In addition, the limited number of examples shows unstable explanations using LIME and SHAP [113]. The pros and cons of the XAI method are summarized in Table 3. The code for these algorithms is also provided. How to evaluate these methods and their explanations is still a popular research challenge [111,114,115], and further research is needed in this context.

6.4. Future Directions of Interpretable Models

Most of the current XAI techniques are post hoc (see Table 2). The most likely reason is their ease of use for target users (clinicians, researchers, etc.), who can adapt their methods to post hoc techniques such as Grad-CAM and LIME. However, it is suggested to develop an interpretable model for any high-risk medical situation [8]. In this context, more clinicians are accepting LIME's chart interpretation, and their satisfaction rate has reached 78%, but most of the clinicians with low satisfaction gave low scores despite their recognition of the explanations provided by LIME [99]. To obtain explanations from the model itself, a general method aims to calculate the loss for the output feature map of each filter in the convolutional layer [116]. This method leads to recognizing which of the filters are activated during the network prediction. In [117], the authors aimed to obtain an abstract structure of a causal model by training a neural network. In [118], the presented method is self-explaining by its similarity to the prototypes, such as comparing specific cases in the network. It is a new form of exploration to explain a GNN by prototype learning. So far, global explainability is desirable in clinical tasks to achieve trust. More particularly, it is necessary for the actual XAI application to take into account users without the necessary AI training. Manual instructions and other further clarifications are strongly recommended. In addition, as reported in [119], an XAI method may be combined with other techniques such as domain adaptation (DA) and federated learning (FL) to achieve better results.

7. Conclusions

This paper has presented several popular XAI methods in terms of their principles and deployment in medical image applications, including their performances. The various algorithms were first classified into numerous distinct categories. In the case of the AI currently applied in the medical imaging fields, a popular XAI related to medical image classifications was discussed. A summary of the applications of the recently proposed XAI approaches to increase the interpretability of their proposed models was also detailed. Furthermore, the need for explainable models for radiomic analysis was also explained and discussed. To conclude, discussion and analysis of the medical requirements of XAI, including its prospects and challenges for further investigation, were also given.

Author Contributions: Conceptualization, A.C.; methodology, A.C.; software, A.C., J.P. and J.X.; validation, A.C. and A.B.; formal analysis, A.C.; investigation, A.C., J.P. and J.X.; resources, A.C.; data curation, A.C., J.P. and J.X.; writing—original draft preparation, A.C., J.P. and J.X.; writing—review and editing, A.C., J.P., J.X. and A.B.; visualization, A.C., J.P. and J.X.; supervision, A.C.; project administration, A.C.; funding acquisition, A.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by NATIONAL NATURAL SCIENCE FOUNDATION grant number 82260360 and the FOREIGN YOUNG TALENTS PROGRAM QN2021033002L.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study.

References

1. Nazar, M.; Alam, M.M.; Yafi, E.; Mazliham, M. A systematic review of human-computer interaction and explainable artificial intelligence in healthcare with artificial intelligence techniques. *IEEE Access* **2021**, *9*, 153316–153348. [CrossRef]
2. von Eschenbach, W.J. Transparency and the black box problem: Why we do not trust AI. *Philos. Technol.* **2021**, *34*, 1607–1622. [CrossRef]

3. Devlin, J.; Chang, M.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2018**, arXiv:1810.04805.
4. Gunning, D.; Aha, D. DARPA's explainable artificial intelligence (XAI) program. *AI Mag.* **2019**, *40*, 44–58.
5. Angelov, P.; Soares, E. Towards explainable deep neural networks (xDNN). *Neural Netw.* **2020**, *130*, 185–194. [CrossRef]
6. Arrieta, A.B.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; García, S.; Gil-López, S.; Molina, D.; Benjamins, R.; et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* **2020**, *58*, 82–115. [CrossRef]
7. Tjoa, E.; Guan, C. A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *32*, 4793–4813. [CrossRef]
8. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **2019**, *1*, 206–215. [CrossRef]
9. Gunning, D.; Stefik, M.; Choi, J.; Miller, T.; Stumpf, S.; Yang, G.Z. XAI—Explainable artificial intelligence. *Sci. Robot.* **2019**, *4*, eaay7120. [CrossRef]
10. Yang, X.; He, X.; Zhao, J.; Zhang, Y.; Zhang, S.; Xie, P. COVID-CT-dataset: A CT scan dataset about COVID-19. *arXiv* **2020**, arXiv:2003.13865.
11. Falk, T.; Mai, D.; Bensch, R.; Çiçek, Ö.; Abdulkadir, A.; Marrakchi, Y.; Böhm, A.; Deubner, J.; Jäckel, Z.; Seiwald, K.; et al. U-Net: Deep learning for cell counting, detection, and morphometry. *Nat. Methods* **2019**, *16*, 67–70. [CrossRef] [PubMed]
12. Smuha, N.A. The EU approach to ethics guidelines for trustworthy artificial intelligence. *Comput. Law Rev. Int.* **2019**, *20*, 97–106. [CrossRef]
13. Bai, T.; Zhao, J.; Zhu, J.; Han, S.; Chen, J.; Li, B.; Kot, A. Ai-gan: Attack-inspired generation of adversarial examples. In Proceedings of the 2021 IEEE International Conference on Image Processing (ICIP), Anchorage, AK, USA, 19–22 September 2021; pp. 2543–2547.
14. Papernot, N.; McDaniel, P.; Jha, S.; Fredrikson, M.; Celik, Z.B.; Swami, A. The limitations of deep learning in adversarial settings. In Proceedings of the 2016 IEEE European symposium on security and privacy (EuroS&P), Saarbrücken, Germany, 21–24 March 2016; pp. 372–387.
15. Kiener, M. Artificial intelligence in medicine and the disclosure of risks. *AI Soc.* **2021**, *36*, 705–713. [CrossRef] [PubMed]
16. Vigano, L.; Magazzeni, D. Explainable security. In Proceedings of the 2020 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW), Genoa, Italy, 7–11 September 2020; pp. 293–300.
17. Kuppa, A.; Le-Khac, N.A. Black Box Attacks on Explainable Artificial Intelligence(XAI) methods in Cyber Security. In Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 19–24 July 2020; pp. 1–8. [CrossRef]
18. Trocin, C.; Mikalef, P.; Papamitsiou, Z.; Conboy, K. Responsible AI for digital health: A synthesis and a research agenda. *Inf. Syst. Front.* **2021**, 1–19. [CrossRef]
19. Arun, N.; Gaw, N.; Singh, P.; Chang, K.; Aggarwal, M.; Chen, B.; Hoebel, K.; Gupta, S.; Patel, J.; Gidwani, M.; et al. Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging. *Radiol. Artif. Intell.* **2021**, *3*, e200267. [CrossRef]
20. Smith, H. Clinical AI: Opacity, accountability, responsibility and liability. *AI Soc.* **2021**, *36*, 535–545. [CrossRef]
21. Tigard, D.W. Responsible AI and moral responsibility: A common appreciation. *AI Ethics* **2021**, *1*, 113–117. [CrossRef]
22. Hazirbas, C.; Bitton, J.; Dolhansky, B.; Pan, J.; Gordo, A.; Ferrer, C.C. Casual conversations: A dataset for measuring fairness in ai. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 2289–2293.
23. Castelvecchi, D. Can we open the black box of AI? *Nat. News* **2016**, *538*, 20. [CrossRef]
24. Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; Galstyan, A. A survey on bias and fairness in machine learning. *ACM Comput. Surv. (CSUR)* **2021**, *54*, 1–35. [CrossRef]
25. Lipton, Z.C. Lipton, Z.C. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* **2018**, *16*, 31–57. [CrossRef]
26. Du, M.; Liu, N.; Hu, X. Techniques for Interpretable Machine Learning. *Commun. ACM* **2019**, *63*, 68–77. [CrossRef]
27. Yang, G.; Ye, Q.; Xia, J. Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond. *Inf. Fusion* **2022**, *77*, 29–52. [CrossRef] [PubMed]
28. Simonyan, K.; Vedaldi, A.; Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv* **2013**, arXiv:1312.6034.
29. Springenberg, J.T.; Dosovitskiy, A.; Brox, T.; Riedmiller, M. Striving for simplicity: The all convolutional net. *arXiv* **2014**, arXiv:1412.6806.
30. Zeiler, M.D.; Fergus, R. Visualizing and understanding convolutional networks. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 818–833.
31. Bach, S.; Binder, A.; Montavon, G.; Klauschen, F.; Müller, K.R.; Samek, W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE* **2015**, *10*, e0130140. [CrossRef] [PubMed]
32. Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; Torralba, A. Learning deep features for discriminative localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2921–2929.
33. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy 22–29 October 2017; pp. 618–626.

34. Ribeiro, M.T.; Singh, S.; Guestrin, C. "Why should i trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 1135–1144.
35. Huang, Q.; Yamada, M.; Tian, Y.; Singh, D.; Chang, Y. Graphlime: Local interpretable model explanations for graph neural networks. *IEEE Trans. Knowl. Data Eng.* **2022**. [CrossRef]
36. Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 4768–4777.
37. Jetley, S.; Lord, N.A.; Lee, N.; Torr, P.H.S. Learn To Pay Attention. *arXiv* **2018**, arXiv:1804.02391.
38. Chen, C.; Li, O.; Tao, D.; Barnett, A.; Rudin, C.; Su, J.K. This Looks Like That: Deep Learning for Interpretable Image Recognition. In *Proceedings of the Advances in Neural Information Processing Systems*; Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2019; Volume 32.
39. Hoffer, E.; Ailon, N. Deep metric learning using triplet network. In Proceedings of the International Workshop on Similarity-Based Pattern Recognition, Copenhagen, Denmark, 12–14 October 2015; pp. 84–92.
40. Kim, B.; Wattenberg, M.; Gilmer, J.; Cai, C.; Wexler, J.; Viegas, F.; sayres, R. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). In Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; Dy, J., Krause, A., Eds.; 2018; Volume 80, pp. 2668–2677.
41. Vinyals, O.; Toshev, A.; Bengio, S.; Erhan, D. Show and Tell: A Neural Image Caption Generator. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015.
42. Pierson, E.; Cutler, D.M.; Leskovec, J.; Mullainathan, S.; Obermeyer, Z. An algorithmic approach to reducing unexplained pain disparities in underserved populations. *Nat. Med.* **2021**, *27*, 136–140. [CrossRef]
43. Born, J.; Wiedemann, N.; Cossio, M.; Buhre, C.; Brändle, G.; Leidermann, K.; Goulet, J.; Aujayeb, A.; Moor, M.; Rieck, B.; et al. Accelerating detection of lung pathologies with explainable ultrasound image analysis. *Appl. Sci.* **2021**, *11*, 672. [CrossRef]
44. Shen, Y.; Wu, N.; Phang, J.; Park, J.; Liu, K.; Tyagi, S.; Heacock, L.; Kim, S.G.; Moy, L.; Cho, K.; et al. An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization. *Med. Image Anal.* **2021**, *68*, 101908. [CrossRef] [PubMed]
45. Jia, G.; Lam, H.K.; Xu, Y. Classification of COVID-19 chest X-ray and CT images using a type of dynamic CNN modification method. *Comput. Biol. Med.* **2021**, *134*, 104425. [CrossRef] [PubMed]
46. Song, Y.; Zheng, S.; Li, L.; Zhang, X.; Zhang, X.; Huang, Z.; Chen, J.; Wang, R.; Zhao, H.; Chong, Y.; et al. Deep learning enables accurate diagnosis of novel coronavirus (COVID-19) with CT images. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2021**, *18*, 2775–2780. [CrossRef] [PubMed]
47. Wang, S.H.; Govindaraj, V.V.; Górriz, J.M.; Zhang, X.; Zhang, Y.D. COVID-19 classification by FGCNet with deep feature fusion from graph convolutional network and convolutional neural network. *Inf. Fusion* **2021**, *67*, 208–229. [CrossRef] [PubMed]
48. Fan, Z.; Gong, P.; Tang, S.; Lee, C.U.; Zhang, X.; Song, P.; Chen, S.; Li, H. Joint localization and classification of breast tumors on ultrasound images using a novel auxiliary attention-based framework. *arXiv* **2022**, arXiv:2210.05762.
49. Wang, Z.; Xiao, Y.; Li, Y.; Zhang, J.; Lu, F.; Hou, M.; Liu, X. Automatically discriminating and localizing COVID-19 from community-acquired pneumonia on chest X-rays. *Pattern Recognit.* **2021**, *110*, 107613. [CrossRef]
50. Sutton, R.T.; zaiane, O.R.; Goebel, R.; Baumgart, D.C. Artificial intelligence enabled automated diagnosis and grading of ulcerative colitis endoscopy images. *Sci. Rep.* **2022**, *12*, 1–10. [CrossRef]
51. Yamashita, R.; Long, J.; Longacre, T.; Peng, L.; Berry, G.; Martin, B.; Higgins, J.; Rubin, D.L.; Shen, J. Deep learning model for the prediction of microsatellite instability in colorectal cancer: A diagnostic study. *Lancet Oncol.* **2021**, *22*, 132–141. [CrossRef]
52. Wu, Y.H.; Gao, S.H.; Mei, J.; Xu, J.; Fan, D.P.; Zhang, R.G.; Cheng, M.M. Jcs: An explainable covid-19 diagnosis system by joint classification and segmentation. *IEEE Trans. Image Process.* **2021**, *30*, 3113–3126. [CrossRef]
53. Lu, S.; Zhu, Z.; Gorriz, J.M.; Wang, S.H.; Zhang, Y.D. NAGNN: Classification of COVID-19 based on neighboring aware representation from deep graph neural network. *Int. J. Intell. Syst.* **2022**, *37*, 1572–1598. [CrossRef]
54. Haghanifar, A.; Majdabadi, M.M.; Choi, Y.; Deivalakshmi, S.; Ko, S. COVID-cxnet: Detecting COVID-19 in frontal chest X-ray images using deep learning. *Multimed. Tools Appl.* **2022**, *81*, 30615–30645. [CrossRef] [PubMed]
55. Punn, N.S.; Agarwal, S. Automated diagnosis of COVID-19 with limited posteroanterior chest X-ray images using fine-tuned deep neural networks. *Appl. Intell.* **2021**, *51*, 2689–2702. [CrossRef] [PubMed]
56. Wang, H.; Wang, S.; Qin, Z.; Zhang, Y.; Li, R.; Xia, Y. Triple attention learning for classification of 14 thoracic diseases using chest radiography. *Med. Image Anal.* **2021**, *67*, 101846. [CrossRef] [PubMed]
57. Fu, X.; Bi, L.; Kumar, A.; Fulham, M.; Kim, J. Multimodal spatial attention module for targeting multimodal PET-CT lung tumor segmentation. *IEEE J. Biomed. Health Inf.* **2021**, *25*, 3507–3516. [CrossRef] [PubMed]
58. Yeung, M.; Sala, E.; Schönlieb, C.B.; Rundo, L. Focus U-Net: A novel dual attention-gated CNN for polyp segmentation during colonoscopy. *Comput. Biol. Med.* **2021**, *137*, 104815. [CrossRef] [PubMed]
59. Hu, B.; Vasu, B.; Hoogs, A. X-MIR: EXplainable Medical Image Retrieval. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 4–8 January 2022; pp. 440–450.
60. Alsinglawi, B.; Alshari, O.; Alorjani, M.; Mubin, O.; Alnajjar, F.; Novoa, M.; Darwish, O. An explainable machine learning framework for lung cancer hospital length of stay prediction. *Sci. Rep.* **2022**, *12*, 607. [CrossRef] [PubMed]
61. Du, Y.; Rafferty, A.R.; McAuliffe, F.M.; Wei, L.; Mooney, C. An explainable machine learning-based clinical decision support system for prediction of gestational diabetes mellitus. *Sci. Rep.* **2022**, *12*, 1170. [CrossRef] [PubMed]

62. Severn, C.; Suresh, K.; Görg, C.; Choi, Y.S.; Jain, R.; Ghosh, D. A Pipeline for the Implementation and Visualization of Explainable Machine Learning for Medical Imaging Using Radiomics Features. *Sensors* **2022**, *22*, 5205. [CrossRef]
63. Le, N.Q.K.; Kha, Q.H.; Nguyen, V.H.; Chen, Y.C.; Cheng, S.J.; Chen, C.Y. Machine learning-based radiomics signatures for EGFR and KRAS mutations prediction in non-small-cell lung cancer. *Int. J. Mol. Sci.* **2021**, *22*, 9254. [CrossRef]
64. Moncada-Torres, A.; van Maaren, M.C.; Hendriks, M.P.; Siesling, S.; Geleijnse, G. Explainable machine learning can outperform Cox regression predictions and provide insights in breast cancer survival. *Sci. Rep.* **2021**, *11*, 6968. [CrossRef]
65. Abeyagunasekera, S.H.P.; Perera, Y.; Chamara, K.; Kaushalya, U.; Sumathipala, P.; Senaweera, O. LISA: Enhance the explainability of medical images unifying current XAI techniques. In Proceedings of the 2022 IEEE 7th International Conference for Convergence in Technology (I2CT), Mumbai, India, 7–9 April 2022; pp. 1–9. [CrossRef]
66. Duell, J.; Fan, X.; Burnett, B.; Aarts, G.; Zhou, S.M. A Comparison of Explanations Given by Explainable Artificial Intelligence Methods on Analysing Electronic Health Records. In Proceedings of the 2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI), Athens, Greece, 27–30 July 2021; pp. 1–4. [CrossRef]
67. Liu, G.; Liao, Y.; Wang, F.; Zhang, B.; Zhang, L.; Liang, X.; Wan, X.; Li, S.; Li, Z.; Zhang, S.; et al. Medical-VLBERT: Medical Visual Language BERT for COVID-19 CT Report Generation with Alternate Learning. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *32*, 3786–3797. [CrossRef] [PubMed]
68. Ghassemi, M.; Oakden-Rayner, L.; Beam, A.L. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit. Health* **2021**, *3*, e745–e750. [CrossRef] [PubMed]
69. Wu, J.; Mooney, R.J. Faithful Multimodal Explanation for Visual Question Answering. *arXiv* **2018**, arXiv:1809.02805.
70. Adadi, A.; Berrada, M. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access* **2018**, *6*, 52138–52160. [CrossRef]
71. Das, A.; Rad, P. Opportunities and challenges in explainable artificial intelligence (xai): A survey. *arXiv* **2020**, arXiv:2006.11371.
72. Zhang, Z.; Chen, P.; Sapkota, M.; Yang, L. Tandemnet: Distilling knowledge from medical images using diagnostic reports as optional semantic references. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Quebec City, QC, Canada, 10–14 September 2017; pp. 320–328.
73. Li, O.; Liu, H.; Chen, C.; Rudin, C. Deep Learning for Case-Based Reasoning Through Prototypes: A Neural Network That Explains Its Predictions. *Proc. AAAI Conf. Artif. Intell.* **2018**, *32, 3530–3537*. [CrossRef]
74. Montavon, G.; Lapuschkin, S.; Binder, A.; Samek, W.; Müller, K.R. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recognit.* **2017**, *65*, 211–222. [CrossRef]
75. Mohamed, E.; Sirlantzis, K.; Howells, G. A review of visualisation-as-explanation techniques for convolutional neural networks and their evaluation. *Displays* **2022**, *73*, 102239. [CrossRef]
76. Soares, E.; Angelov, P.; Biaso, S.; Froes, M.H.; Abe, D.K. SARS-CoV-2 CT-scan dataset: A large dataset of real patients CT scans for SARS-CoV-2 identification. *MedRxiv* **2020**. [CrossRef]
77. Jiang, C.; Chen, Y.; Chang, J.; Feng, M.; Wang, R.; Yao, J. Fusion of medical imaging and electronic health records with attention and multi-head machanisms. *arXiv* **2021**, arXiv:2112.11710.
78. Ron, T.; Hazan, T. Dual Decomposition of Convex Optimization Layers for Consistent Attention in Medical Images. In Proceedings of the 39th International Conference on Machine Learning, Baltimore, MD, USA, 17–23 July 2022; Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., Sabato, S., Eds.; 2022; Volume 162, pp. 18754–18769.
79. Shapley, L.S. 17. A value for n-person games. In *Contributions to the Theory of Games (AM-28), Volume II*; Princeton University Press: Princeton, NJ, USA, 2016; pp. 307–318.
80. Shrikumar, A.; Greenside, P.; Kundaje, A. Learning Important Features Through Propagating Activation Differences. *arXiv* **2017**, arXiv:1704.02685.
81. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
82. Liu, F.; Ge, S.; Wu, X. Competence-based multimodal curriculum learning for medical report generation. *arXiv* **2022**, arXiv:2206.14579.
83. Malhi, A.; Kampik, T.; Pannu, H.; Madhikermi, M.; Främling, K. Explaining machine learning-based classifications of in-vivo gastral images. In Proceedings of the 2019 Digital Image Computing: Techniques and Applications (DICTA), Perth, Australia, 2–4 December 2019; pp. 1–7.
84. Ye, Q.; Xia, J.; Yang, G. Explainable AI for COVID-19 CT classifiers: An initial comparison study. In Proceedings of the 2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS), Aveiro, Portugal, 7–9 June 2021; pp. 521–526.
85. Serrano, S.; Smith, N.A. Is attention interpretable? *arXiv* **2019**, arXiv:1906.03731.
86. Chaddad, A.; Kucharczyk, M.J.; Daniel, P.; Sabri, S.; Jean-Claude, B.J.; Niazi, T.; Abdulkarim, B. Radiomics in glioblastoma: Current status and challenges facing clinical implementation. *Front. Oncol.* **2019**, *9*, 374. [CrossRef]
87. Chaddad, A.; Kucharczyk, M.J.; Cheddad, A.; Clarke, S.E.; Hassan, L.; Ding, S.; Rathore, S.; Zhang, M.; Katib, Y.; Bahoric, B.; et al. Magnetic resonance imaging based radiomic models of prostate cancer: A narrative review. *Cancers* **2021**, *13*, 552. [CrossRef]
88. Chaddad, A.; Toews, M.; Desrosiers, C.; Niazi, T. Deep radiomic analysis based on modeling information flow in convolutional neural networks. *IEEE Access* **2019**, *7*, 97242–97252. [CrossRef]
89. Singh, G.; Manjila, S.; Sakla, N.; True, A.; Wardeh, A.H.; Beig, N.; Vaysberg, A.; Matthews, J.; Prasanna, P.; Spektor, V. Radiomics and radiogenomics in gliomas: A contemporary update. *Br. J. Cancer* **2021**, *125*, 641–657. [CrossRef] [PubMed]
90. Chaddad, A.; Hassan, L.; Desrosiers, C. Deep radiomic analysis for predicting coronavirus disease 2019 in computerized tomography and X-ray images. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *33*, 3–11. [CrossRef]

91. Gupta, S.; Gupta, M. Deep Learning for Brain Tumor Segmentation using Magnetic Resonance Images. In Proceedings of the 2021 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), Melbourne, Australia, 13–15 October 2021; pp. 1–6. [CrossRef]
92. Chaddad, A.; Daniel, P.; Zhang, M.; Rathore, S.; Sargos, P.; Desrosiers, C.; Niazi, T. Deep radiomic signature with immune cell markers predicts the survival of glioma patients. *Neurocomputing* **2022**, *469*, 366–375. [CrossRef]
93. Chaddad, A.; Zhang, M.; Desrosiers, C.; Niazi, T. Deep radiomic features from MRI scans predict survival outcome of recurrent glioblastoma. In Proceedings of the International Workshop on Radiomics and Radiogenomics in Neuro-Oncology, Shenzhen, China, 13 October 2019; pp. 36–43.
94. Moridian, P.; Ghassemi, N.; Jafari, M.; Salloum-Asfar, S.; Sadeghi, D.; Khodatars, M.; Shoeibi, A.; Khosravi, A.; Ling, S.H.; Subasi, A.; et al. Automatic Autism Spectrum Disorder Detection Using Artificial Intelligence Methods with MRI Neuroimaging: A Review. *arXiv* **2022**, arXiv:2206.11233.
95. Scapicchio, C.; Gabelloni, M.; Barucci, A.; Cioni, D.; Saba, L.; Neri, E. A deep look into radiomics. *Radiol. Med.* **2021**, *126*, 1296–1311. [CrossRef] [PubMed]
96. Garin, E.; Tselikas, L.; Guiu, B.; Chalaye, J.; Edeline, J.; de Baere, T.; Assénat, E.; Tacher, V.; Robert, C.; Terroir-Cassou-Mounat, M.; et al. Personalised versus standard dosimetry approach of selective internal radiation therapy in patients with locally advanced hepatocellular carcinoma (DOSISPHERE-01): A randomised, multicentre, open-label phase 2 trial. *Lancet Gastroenterol. Hepatol.* **2021**, *6*, 17–29. [CrossRef] [PubMed]
97. Akula, A.R.; Wang, K.; Liu, C.; Saba-Sadiya, S.; Lu, H.; Todorovic, S.; Chai, J.; Zhu, S.C. CX-ToM: Counterfactual explanations with theory-of-mind for enhancing human trust in image recognition models. *iScience* **2022**, *25*, 103581. [CrossRef] [PubMed]
98. Ehsan, U.; Liao, Q.V.; Muller, M.; Riedl, M.O.; Weisz, J.D. Expanding explainability: Towards social transparency in ai systems. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, Yokohama, Japan, 8–13 May 2021; pp. 1–19.
99. Kumarakulasinghe, N.B.; Blomberg, T.; Liu, J.; Leao, A.S.; Papapetrou, P. Evaluating local interpretable model-agnostic explanations on clinical machine learning classification models. In Proceedings of the 2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS), Rochester, MN, USA, 28–30 July 2020; pp. 7–12.
100. Evans, T.; Retzlaff, C.O.; Geißler, C.; Kargl, M.; Plass, M.; Müller, H.; Kiehl, T.R.; Zerbe, N.; Holzinger, A. The explainability paradox: Challenges for xAI in digital pathology. *Future Gener. Comput. Syst.* **2022**, *133*, 281–296. [CrossRef]
101. Salahuddin, Z.; Woodruff, H.C.; Chatterjee, A.; Lambin, P. Transparency of deep neural networks for medical image analysis: A review of interpretability methods. *Comput. Biol. Med.* **2022**, *140*, 105111. [CrossRef]
102. Gebru, B.; Zeleke, L.; Blankson, D.; Nabil, M.; Nateghi, S.; Homaifar, A.; Tunstel, E. A Review on Human–Machine Trust Evaluation: Human-Centric and Machine-Centric Perspectives. *IEEE Trans. Hum.-Mach. Syst.* **2022**, *52*, 952–962. [CrossRef]
103. Adebayo, J.; Muelly, M.; Abelson, H.; Kim, B. Post hoc explanations may be ineffective for detecting unknown spurious correlation. In Proceedings of the International Conference on Learning Representations, Virtual Event, 3–7 May 2021.
104. Alqaraawi, A.; Schuessler, M.; Weiß, P.; Costanza, E.; Berthouze, N. Evaluating saliency map explanations for convolutional neural networks: A user study. In Proceedings of the 25th International Conference on Intelligent User Interfaces, Cagliari, Italy 17–20 March 2020; pp. 275–285.
105. Stepin, I.; Alonso, J.M.; Catala, A.; Pereira-Fariña, M. A Survey of Contrastive and Counterfactual Explanation Generation Methods for Explainable Artificial Intelligence. *IEEE Access* **2021**, *9*, 11974–12001. [CrossRef]
106. Sutton, R.T.; Pincock, D.; Baumgart, D.C.; Sadowski, D.C.; Fedorak, R.N.; Kroeker, K.I. An overview of clinical decision support systems: Benefits, risks, and strategies for success. *NPJ Digit. Med.* **2020**, *3*, 17. [CrossRef]
107. Mitchell, M.; Wu, S.; Zaldivar, A.; Barnes, P.; Vasserman, L.; Hutchinson, B.; Spitzer, E.; Raji, I.D.; Gebru, T. Model cards for model reporting. In Proceedings of the Conference on Fairness, Accountability, and Transparency, Atlanta, GA, USA, 29–31 January 2019; pp. 220–229.
108. Arnold, M.; Bellamy, R.K.; Hind, M.; Houde, S.; Mehta, S.; Mojsilović, A.; Nair, R.; Ramamurthy, K.N.; Olteanu, A.; Piorkowski, D.; et al. FactSheets: Increasing trust in AI services through supplier's declarations of conformity. *IBM J. Res. Dev.* **2019**, *63*, 6:1–6:13. [CrossRef]
109. Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J.W.; Wallach, H.; Iii, H.D.; Crawford, K. Datasheets for datasets. *Commun. ACM* **2021**, *64*, 86–92. [CrossRef]
110. van der Velden, B.H.; Kuijf, H.J.; Gilhuijs, K.G.; Viergever, M.A. Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. *Med. Image Anal.* **2022**, *79*, 102470. [CrossRef] [PubMed]
111. Lopes, P.; Silva, E.; Braga, C.; Oliveira, T.; Rosado, L. XAI Systems Evaluation: A Review of Human and Computer-Centred Methods. *Appl. Sci.* **2022**, *12*, 9423. [CrossRef]
112. Lin, Y.S.; Lee, W.C.; Celik, Z.B. What do you see? Evaluation of explainable artificial intelligence (XAI) interpretability through neural backdoors. *arXiv* **2020**, arXiv:2009.10639.
113. Nguyen, H.T.T.; Cao, H.Q.; Nguyen, K.V.T.; Pham, N.D.K. Evaluation of Explainable Artificial Intelligence: SHAP, LIME, and CAM. In Proceedings of the FPT AI Conference 2021, Ha Noi, Viet Nam, 6–7 May 2021; pp. 1–6.
114. Nauta, M.; Trienes, J.; Pathak, S.; Nguyen, E.; Peters, M.; Schmitt, Y.; Schlötterer, J.; van Keulen, M.; Seifert, C. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. *arXiv* **2022**, arXiv:2201.08164.

115. Zhang, Y.; Xu, F.; Zou, J.; Petrosian, O.L.; Krinkin, K.V. XAI Evaluation: Evaluating Black-Box Model Explanations for Prediction. In Proceedings of the 2021 II International Conference on Neural Networks and Neurotechnologies (NeuroNT), Saint Petersburg, Russia, 16 June 2021; pp. 13–16.
116. Zhang, Q.; Wu, Y.N.; Zhu, S.C. Interpretable Convolutional Neural Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018.
117. Geiger, A.; Wu, Z.; Lu, H.; Rozner, J.; Kreiss, E.; Icard, T.; Goodman, N.; Potts, C. Inducing causal structure for interpretable neural networks. In Proceedings of the International Conference on Machine Learning, Baltimore, MD, USA, 17–23 July 2022; pp. 7324–7338.
118. Zhang, Z.; Liu, Q.; Wang, H.; Lu, C.; Lee, C. ProtGNN: Towards Self-Explaining Graph Neural Networks. *Proc. AAAI Conf. Artif. Intell.* **2022**, *36*, 9127–9135. [CrossRef]
119. Chaddad, A.; Li, J.; Katib, Y.; Kateb, R.; Tanougast, C.; Bouridane, A.; Abdulkadir, A. Explainable, Domain-Adaptive, and Federated Artificial Intelligence in Medicine. *arXiv* **2022**, arXiv:2211.09317.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.