Review

Explainable Deep Learning Models in Medical Image Analysis

Amitojdeep Singh [1,2,]*[iD], Sourya Sengupta [1,2] and Vasudevan Lakshminarayanan [1,2]

[1] Theoretical and Experimental Epistemology Laboratory, School of Optometry and Vision Science, University of Waterloo, Waterloo, ON N2L 3G1, Canada; sourya.sengupta@uwaterloo.ca (S.S.); vengulak@uwaterloo.ca (V.L.)
[2] Department of Systems Design Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada
*   Correspondence: amitojdeep.singh@uwaterloo.ca

check for
updates

Abstract: Deep learning methods have been very effective for a variety of medical diagnostic tasks and have even outperformed human experts on some of those. However, the black-box nature of the algorithms has restricted their clinical use. Recent explainability studies aim to show the features that influence the decision of a model the most. The majority of literature reviews of this area have focused on taxonomy, ethics, and the need for explanations. A review of the current applications of explainable deep learning for different medical imaging tasks is presented here. The various approaches, challenges for clinical deployment, and the areas requiring further research are discussed here from a practical standpoint of a deep learning researcher designing a system for the clinical end-users.

Keywords: explainability; explainable AI; XAI; deep learning; medical imaging; diagnosis

1. Introduction

Computer-aided diagnostics (CAD) using artificial intelligence (AI) provides a promising way to make the diagnosis process more efficient and available to the masses. Deep learning is the leading artificial intelligence (AI) method for a wide range of tasks including medical imaging problems. It is the state of the art for several computer vision tasks and has been used for medical imaging tasks like the classification of Alzheimer's [1], lung cancer detection [2], retinal disease detection [3,4], etc. Despite achieving remarkable results in the medical domain, AI-based methods have not achieved a significant deployment in the clinics. This is due to the underlying black-box nature of the deep learning algorithms along with other reasons like computational costs. It arises from the fact that, despite having the underlying statistical principles, there is a lack of ability to explicitly represent the knowledge for a given task performed by a deep neural network. Simpler AI methods like linear regression and decision trees are self-explanatory as the decision boundary used for classification can be visualized in a few dimensions using the model parameters. However, these lack the complexity required for tasks such as classification of 3D and most 2D medical images. The lack of tools to inspect the behavior of black-box models affects the use of deep learning in all domains including finance and autonomous driving where explainability and reliability are the key elements for trust by the end-user. A schematic explaining the relationship between deep learning and the need for explanations is shown in Figure 1.

A medical diagnosis system needs to be transparent, understandable, and explainable to gain the trust of physicians, regulators as well as the patients. Ideally, it should be able to explain the complete logic of making a certain decision to all the parties involved. Newer regulations like the European General Data Protection Regulation (GDPR) are making it harder for the use of black-box models in

all businesses including healthcare because retraceability of the decisions is now a requirement [5]. An artificial intelligence (AI) system to complement medical professionals should have a certain amount of explainability and allow the human expert to retrace the decisions and use their judgment. Some researchers also emphasize that even humans are not always able to or even willing to explain their decisions [5]. Explainability is the key to safe, ethical, fair, and trust-able use of artificial intelligence (AI) and a key enabler for its deployment in the real world. Breaking myths about artificial intelligence (AI) by showing what a model looked at while making the decision can inculcate trust among the end-users. It is even more important to show the domain-specific features used in the decision for non-deep learning users like most medical professionals.
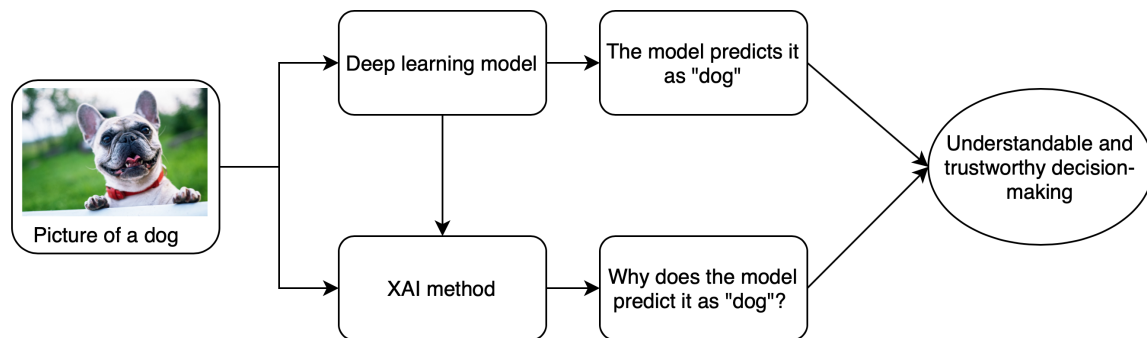

Figure 1. A brief schematic of basics of XAI methods.

The terms explainability and interpretability are often used interchangeably in the literature. A distinction between these was provided in [6] where interpretation was defined as mapping an abstract concept like the output class into a domain example, while explanation was defined as a set of domain features such as pixels of an image the contribute to the output decision of the model. A related term to this concept is the uncertainty associated with the decision of a model. Deep learning classifiers are usually not able to say "I don't know" in situations with ambiguity and instead return the class with the highest probability, even if by a narrow margin, making uncertainty a crucial topic. Lately, uncertainty has been analyzed along with the problem of explainability in many studies to highlight the cases where a model is unsure and in turn, make the models more acceptable to non-deep learning users. There have been studies about the uncertainty of machine learning algorithms which include those for endoscopic videos [7] and tissue parameter estimation [8]. We limit the scope of this paper to explainability methods and discuss uncertainty if a study used it along with explainability. The topic of uncertainty in deep learning models can be itself a subject of a future review. As noted earlier, deep learning models are considered as non-transparent as the weights of the neurons can not be understood as knowledge directly. [9] showed that neither the magnitude or the selectivity of the activations, nor the impact on network decisions is sufficient for deciding the importance of a neuron for a given task. A detailed analysis of the terminologies, concepts and, use cases of explainable artificial intelligence (AI) is provided in [10].

This paper describes the studies related to the explainability of deep learning models in the context of medical imaging. A general taxonomy of explainability approaches is described briefly in the next section and a comparison of various attribution based methods is performed in Section 3. Section 4 reviews various explainability methods applied to different medical imaging modalities. The analysis is broken down into Sections 4.1 and 4.2 depending upon the use of attributions or other methods of explainability. The evolution, current trends, and some future possibilities of the explainable deep learning models in medical image analysis are summarized in Section 5.

2. Taxonomy of Explainability Approaches

Several taxonomies have been proposed in the literature to classify different explainability methods [11,12]. Generally, the classification techniques are not absolute, it can vary widely depending upon the characteristics of the methods and can be classified into many overlapping or non-overlapping classes simultaneously. Different kinds of taxonomies and classification methods are discussed briefly here and a detailed analysis of the taxonomies can be found in [10,11] and a flow chart for them is shown in Figure 2.
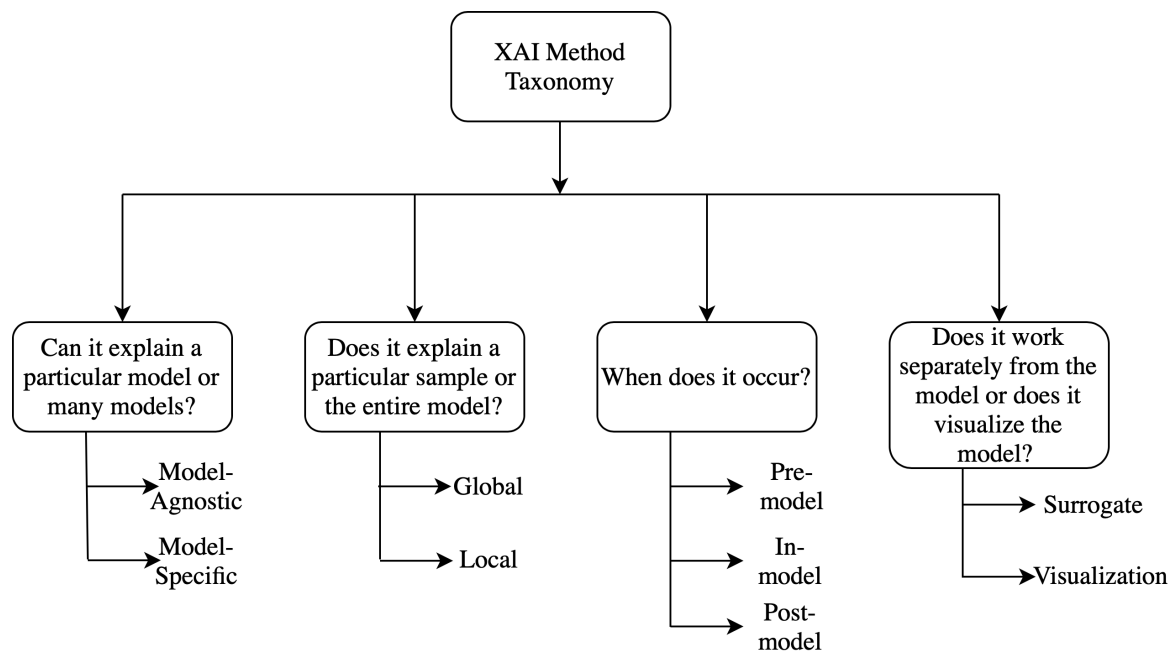

Figure 2. Taxonomy of XAI methods.

2.1. Model Specific vs. Model Agnostic

Model-specific interpretation methods are based on the parameters of the individual models. The graph neural network explainer (GNNExplainer) [13] is a special type of model-specific interpretability where the complexity of data representation needs specifically the graph neural network (GNN). Model Agnostic methods are mainly applicable in post-hoc analysis and not limited to specified model architecture. These methods do not have direct access to the internal model weights or structural parameters.

2.2. Global Methods vs. Local Methods

Local interpretable methods are applicable to a single outcome of the model. This can be done by designing methods that can explain the reason for a particular prediction or outcome. For example, it is interested in specific features and their characteristics. On the contrary, global methods concentrate on the inside of a model by exploiting the overall knowledge about the model, the training, and the associated data. It tries to explain the behavior of the model in general. Feature importance is a good example of this method, which tries to figure out the features which are in general responsible for better performance of the model among all different features.

2.3. Pre-Model vs. In-Model vs. Post-Model

Pre-model methods are independent and do not depend on a particular model architecture to use it on. Principal component analysis (PCA) [14], t-Distributed Stochastic Neighbor Embedding (t-SNE) [15] are some common examples of these methods. Interpretability methods, integrated into the model itself, are called as in-model methods. Some methods are implemented after building a

model and hence these methods are termed as post model and these methods can potentially develop meaningful insights about what exactly a model learned during the training.

2.4. Surrogate Methods vs. Visualization Methods

Surrogate methods consist of different models as an ensemble which are used to analyze other black-box models. The black box models can be understood better by interpreting the surrogate model's decisions by comparing the black-box model's decision and surrogate model's decision. The decision tree [16] is an example of surrogate methods. The visualization methods are not a different model, but it helps to explain some parts of the models by visual understanding like activation maps.

It is to be noted that these classification methods are non-exclusive, these are built upon different logical intuitions and hence have significant overlaps. For example, most of the post-hoc models like attributions can also be seen as model agnostic as these methods are typically not dependent upon the structure of a model. However, some requirements regarding the limitations on model layers or the activation functions do exist for some of the attribution methods. The next section describes the basic concept and subtle difference between various attribution methods to facilitate a comparative discussion of the applications in Section 4.

3. Explainability Methods—Attribution Based

There are broadly two types of approaches to explain the results of deep neural networks (DNN) in medical imaging—those using standard attribution based methods and those using novel, often architecture, or domain-specific techniques. A majority of the papers for explaining deep learning in medical image diagnosis use attribution based methods. Their model agnostic plug and play nature along with readily available open-source implementations make them a convenient solution. The deep learning practitioners can, therefore, focus on designing a model optimal for a given task and use these easy to generate explanations for understanding the model better. Since attribution based studies are a majority (with many of them using multiple attribution based methods), we discuss them beforehand. The applications for those methods are ordered according to the anatomical districts, i.e., organ groups for the diagnosed diseases in Section 4.1. Other methods are used in only a few studies each which typically uses a single method and are hence discussed along with their applications in Section 4.2 which is ordered by the explainability method used.

The problem of assigning an attribution value or contribution or relevance to each input feature of a network led to the development of several attribution methods. The goal of an attribution method is to determine the contribution of an input feature to the target neuron which is usually the output neuron of the correct class for a classification problem. The arrangement of the attributions of all the input features in the shape of the input sample forms heatmaps known as the attribution maps. Some examples of attribution maps for different images are shown in Figure 3. The features with a positive contribution to the activation of the target neuron are typically marked in red while those negatively affecting the activation are marked in blue. These are the features or pixels in case of images providing positive and negative evidence of different magnitudes, respectively.

The commonly used attribution methods are discussed in this section and the applications in the next section. It must be noted that some of the approaches like DeepTaylor [17] provide only positive evidence and can be useful for a certain set of tasks. The attribution methods can be applied on a black box convolutional neural network (CNN) without any modification to the underlying architecture making them a convenient yet powerful XAI tool. An empirical comparison of some of the methods discussed in this section and a unified framework called DeepExplain is available in [18]. Most of the methods discussed here apart from the newer Deep Learning Important FeaTures (LIFT) and Deep SHapley Additive exPlanations (SHAP) are implemented in the iNNvestigate toolbox [19].
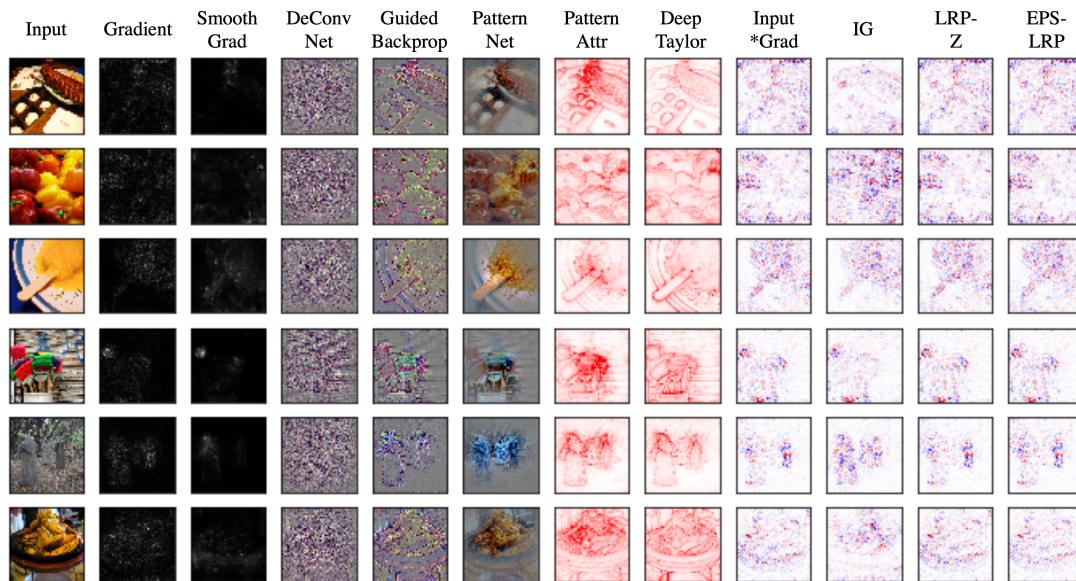
Figure 3. Attributions of VGG-16 with images from Imagenet using the methods implemented in [19].

3.1. Perturbation Based Methods—Occlusion

Perturbation is the simplest way to analyze the effect of changing the input features on the output of an AI model. This can be implemented by removing, masking, or modifying certain input features, and running the forward pass (output computation), and measuring the difference from the original output. This is similar to the sensitivity analysis performed in parametric control system models. The input features affecting the output the most are ranked as the most important. It is computationally expensive as a forward pass needs to be run after perturbing each group of features of the input. In the case of image data the perturbation is performed by covering parts of an image with a grey patch and hence occluding them from the system's view. It can provide both positive and negative evidence by highlighting the responsible features.

This technique was applied by Zeiler and Fergus [20] to the convolutional neural network (CNN) for the image classification task. Occlusion is the benchmark for any attribution study as it is a simple to perform model agnostic approach which reveals the feature importance of a model. It can reveal if a model is overfitting and learning irrelevant features as in the case of adversarial examples [21]. The adversarial examples are the inputs designed to cause the model to make a false decision and are like optical illusions for the models. In that case, the model misclassifies the image (say a cat as a dog) despite the presence of discriminating feature

Occluding all features (pixels) one-by-one and running the forward pass each time can be computationally expensive and can take several hours per image [18]. It is common to use patches of sizes such as 5×5, 10×10, or even larger depending on the size of the target features and computational resources available.

Another perturbation based approach is Shapley value sampling which computes approximate Shapely Values by taking each input feature for a sample number of times. It a method from the coalitional game theory which describes the fair distribution of the gains and losses among the input features. It was originally proposed for the analysis of regression [22]. It is slower than all other approaches as the network has to be run samples × number of features times. As a result it is not a practical method in its original form but has led to the development of game theory-based methods like Deep SHapley Additive exPlanations (SHAP) as discussed in the next subsection.

3.2. Backpropagation Based Methods

These methods compute the attribution for all the input features with a single forward and backward pass through the network. In some of the methods these steps need to be repeated multiple

times but it is independent of the number of input features and much lower than for perturbation-based methods. The faster run-time comes at the expense of a weaker relationship between the outcome and the variation of the output. Various backpropagation based attribution methods are described in Table 1. It must be noted that some of these methods provide only positive evidence while others provide both positive and negative evidence. The methods providing both positive and negative evidence tend to have high-frequency noise which can make the results seem spurious. [18].

Table 1. Backpropagation based attribution methods.

Method	Description	Notes
Gradient	Computes the gradient of the output of the **target neuron** with respect to the input.	The **simplest** approach but is usually not the most effective.
DeConvNet [20]	Applies the **ReLU to the gradient computation instead** of the gradient of a neuron with ReLU activation.	Used to **visualize the features** learned by the layers. **Limited** to CNN models with **ReLU activation**.
Saliency Maps [23]	Takes the **absolute value of the partial derivative** of the target output neuron with respect to the input features to find the features which affect the output the most with least perturbation.	**Can't distinguish between positive and negative** evidence due to absolute values.
Guided backpropagation (GBP) [24]	Applies the **ReLU to the gradient computation in addition** to the gradient of a neuron with ReLU activation.	Like DeConvNet, it is textbflimited to CNN models with **ReLU activation**.
LRP [25]	**Redistributes the prediction score** layer by layer with a backward pass on the network using a particular rule like the ϵ**-rule** while ensuring numerical stability	There are alternative stability rules and **limited** to CNN models with **ReLU activation** when all activations are **ReLU**.
Gradient × input [26]	Initially proposed as a method to **improve sharpness of attribution maps** and is computed by multiplying the signed partial derivative of the output with the input.	It **can approximate occlusion** better than other methods in certain cases like multi layer perceptron (MLP) with Tanh on MNIST data [18] while being instant to compute.
GradCAM [27]	Produces **gradient-weighted class activation maps** using the gradients of the target concept as it flows to the final convolutional layer	Applicable to **only CNN** including those with fully connected layers, structured output (like captions) and reinforcement learning.
IG [28]	Computes the **average gradient** as the input is varied from the **baseline** (often zero) to the actual input value unlike the Gradient × input which uses a single derivative at the input.	It is **highly correlated with the rescale rule of DeepLIFT** discussed below which can act as a good and faster approximation.
DeepTaylor [17]	Finds a rootpoint near each neuron with a value close to the input but with output as 0 and uses it to recursively estimate the attribution of each neuron using **Taylor decomposition**	Provides **sparser explanations**, i.e., focuses on key features but provides **no negative evidence** due to its assumptions of only positive effect.
PatternNet [29]	Estimates the input signal of the output neuron using an **objective function**.	Proposed to counter the incorrect attributions of other methods on **linear systems** and generalized to deep networks.
Pattern Attribution [29]	Applies Deep Taylor decomposition by searching the **rootpoints in the signal direction** for each neuron	Proposed along with **PatternNet** and uses decomposition instead of signal visualization
DeepLIFT [30]	Uses a reference input and computes the reference values of all hidden units using a forward pass and then proceeds backward **like LRP**. It has two variants—**Rescale rule** and the one introduced later called **RevealCancel** which treats positive and negative contributions to a neuron separately.	Rescale is strongly related to and **equivalent in some cases to** ϵ**-LRP** but is **not applicable to models involving multiplicative rules**. **RevealCancel handles such cases** and using RevealCancel for convolutional and Rescale for fully connected layers reduces noise.
SmoothGrad [31]	An improvement on the gradient method which averages the gradient over multiple inputs with additional noise	Designed to visually sharpen the attributions produced by gradient method using class score function.
Deep SHAP [32]	It is a fast **approximation** algorithm to compute the game theory based **SHAP values**. It is connected to DeepLIFT and uses **multiple background samples** instead of one baseline.	Finds attributions for **non neural net models** like trees, support vector machines (SVM) and **ensemble** of those with a neural net using various tools in the the SHAP library.

An important property of attribution methods known as completeness was introduced in the DeepLIFT [30] paper. It states that the attributions for a given input add up to the target output minus the target output at the baseline input. It is satisfied by integrated gradients, DeepTaylor and Deep SHAP but not by DeepLIFT in its rescale rule. A measure generalizing this property is proposed in [18] for a quantitative comparison of various attribution methods. It is called sensitivity-n and involves comparing the sum of the attributions and the variation in the target output in terms of PCC. Occlusion is found to have a higher PCC than other methods as it finds a direct relationship between the variation in the input and that in the output.

The evaluation of attribution methods is complex as it is challenging to discern between the errors of the model and the attribution method explaining it. Measures like sensitivity-n reward the methods designed to reflect the network behavior closely. However, a more practically relevant measure of an attribution method is the similarity of attributions to a human observer's expectation. It needs to be performed with a human expert for a given task and carries an observer bias as the methods closer to the observer expectation can be favored at the cost of those explaining the model behavior. We underscore the argument that the ratings of different attribution methods by experts of a specific domain are potentially useful to develop explainable models which are more likely to be trusted by the end users and hence should be a critical part of the development of an XAI system.

4. Applications

The applications of explainability in medical imaging are reviewed here by categorizing them into two types—those using pre-existing attribution based methods and those using other, often specific methods. The methods are discussed according to the explainability method and the medical imaging application. Table 2 provides a brief overview of the methods.

4.1. Attribution Based

A majority of the medical imaging literature that studied interpretability of deep learning methods used attribution based methods due to their ease of use. Researchers can train a suitable neural network architecture without the added complexity of making it inherently explainable and use a readily available attribution model. This allows the use of either a pre-existing deep learning model or one with a custom architecture for the best performance on the given task. The former makes the implementation easier and allows one to leverage techniques like transfer learning [33,34] while the latter can be used to focus on specific data and avoid overfitting by using fewer parameters. Both approaches are beneficial for medical imaging datasets which tend to be relatively smaller than computer vision benchmarks like ImageNet [35].

Post-model analysis using attributions can reveal if the model is learning relevant features or if it is overfitting to the input by learning spurious features. This allows researchers to adjust the model architecture and hyperparameters to achieve better results on the test data and in turn a potential real-world setting. In this subsection, some recent studies using attribution methods across different medical imaging modalities are reviewed in the order of the anatomical districts from top to bottom of the human body. The reviewed tasks include explanations of deep learning for diagnosing conditions from brain MRI, retinal imaging, breast imaging, CT scans, chest X-ray as well as skin imaging.

4.1.1. Brain Imaging

A study comparing the robustness of various attribution based methods for convolutional neural network (CNN) in Alzheimer's classification using brain MRI [36] performed a quantitative analysis of different methods. Gradient × input, Guided backpropagation (GBP), LRP, and occlusion were the compared methods. The L2 norm between the average attribution maps of multiple runs for the same model to check the repeatability of heatmaps for identically trained models. It was found to be an order of magnitude lower for the first three methods compared to the baseline occlusion since occlusion covers a larger area. LRP performed the best overall indicating the superiority of a completely attribution based method over function and signal-based methods. The similarity between the sum, density, and gain (sum/density) for the top 10 regions of the attributions across the runs was also the highest for LRP. In another study [37] GradCAM and Guided backpropagation (GBP) were used to analyze the clinical coherence of the features learned by a CNN for automated grading of brain tumor from MRI. For the correctly graded cases, both the methods had the most activation in the tumor region while also activating the surrounding ventricles which can indicate malignancy as well. In some cases, this focus on non-tumor regions and some spurious patterns in Guided backpropagation (GBP) maps lead to errors indicating unreliability of the features.

4.1.2. Retinal Imaging

A system producing IG heatmaps along with model predictions was explored as a tool to assist diabetic retinopathy (DR) grading by ophthalmologists [38]. This assistance was found to increase the accuracy of the grading compared to that of an unassisted expert or with the model predictions alone. Initially, the system increased the grading time but with the user's experience, the grading time decreased and the grading confidence increased, especially when both predictions and heatmaps were used. Notably, the accuracy did reduce for patients without DR when model assistance was used and an option to toggle the assistance was provided. An extension of IG called Expressive gradients (EG) was proposed in [39] for weakly supervised segmentation of lesions for Age-related macular degeneration (AMD) diagnosis. A convolutional neural network (CNN) with a compact architecture outperformed larger existing convolutional neural network (CNN)s and Expressive gradients (EG) highlighted the regions of interest better than conventional IG and Guided backpropagation (GBP) methods. Expressive gradients (EG) extends IG by enriching input-level attribution map with high-level attribution maps. A comparative analysis of various explainability models including DeepDeep Learning Important FeaTures (LIFT), DeepSHapley Additive exPlanations (SHAP), IG, etc. was performed for on a model for detection of choroidal neovascularization (CNV), diabetic macular edema (DME), and drusens from optical coherence tomography (OCT) scans [40]. Figure 4 highlights better localization achieved by newer methods (e.g., DeepSHAP) in contrast to noisy results from older methods (e.g., saliency maps).
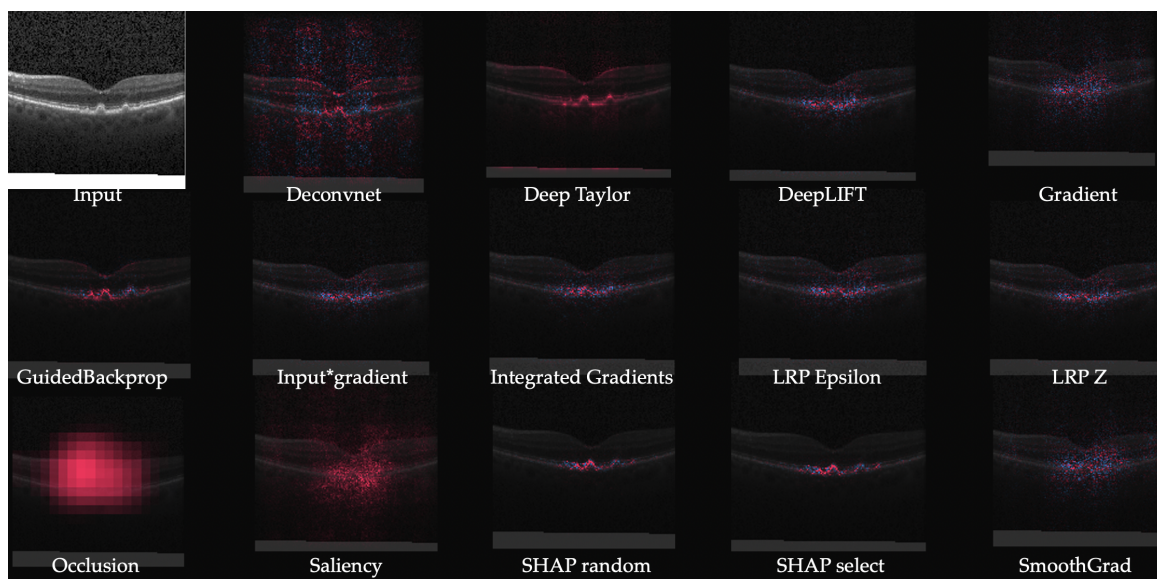

Figure 4. Example of heat maps from a retinal OCT image [40].

4.1.3. Breast Imaging

IG and SmoothGrad were used to visualize the features of a convolutional neural network (CNN) used for classifying estrogen receptor status from breast MRI [41]. The model was observed to have learned relevant features in both spatial and dynamic domains with different contributions from both. The visualizations revealed the learning of certain irrelevant features resulting from pre-processing artifacts. These observations led to changes in the pre-processing and training approaches. An earlier study for breast mass classification from mammograms [42] using two different convolutional neural network (CNN)s—AlexNet [43] and GoogleNet [44]—employed saliency maps to visualize the image features. Both the convolutional neural network (CNN)s were seen to learn the edges of the mass which are the main clinical criteria, while also being sensitive to the context.

4.1.4. CT Imaging

A DeepDreams [45] inspired attribution method was presented in [46] for explaining the segmentation of tumor from liver CT images. This novel method formulated using the concepts of DeapDreams, an image generation algorithm can be applied to a black-box neural network like other attribution methods discussed in Section 3. It performed a sensitivity analysis of the features by maximizing the activation of the target neuron by performing gradient ascent, i.e., finding the steepest slope of the function. A comparison between networks trained on real tumors and synthetic tumors revealed that the former was more sensitive to clinically relevant features and the latter was focusing on other features too. The network was found to be sensitive to intensity as well as sphericity in coherence with domain knowledge.

4.1.5. X-ray Imaging

In a recent study for detection of COVID-19 from chest X-ray images [47], a method called GSInquire was used to produce heatmaps for verifying the features learned by the proposed COVID-net model. GSInquire [48] was developed as an attribution method that outperformed prior methods like SHAP and Expected gradients in terms of the proposed new metrics—impact score and impact coverage. The impact score was defined as the percentage of features which impacted the model decision or confidence strongly. While impact coverage was defined in the context of the coverage of adversarially impacted factors in the input. Another study performed the analysis of uncertainty and interpretability for COVID-19 detection using chest X-rays. The heatmaps of the sample inputs for the trained model were generated using saliency maps, Guided GradCAM, GBP, and Class activation maps (CAM).

4.1.6. Skin Imaging

The features of a suite of 30 CNN models trained for melanoma detection [49] were compared using GradCAM and Kernel SHapley Additive exPlanations (SHAP). It was shown that even the models with high accuracy would occasionally focus on the features that were irrelevant for the diagnosis. There were differences in the explanations of the models that produced similar accuracy which was highlighted by the attribution maps of both the methods. This showed that distinct neural network architectures tend to learn different features. Another study [50] visualized the convolutional neural network (CNN) features for skin lesion classification. The features for the last two layers were visualized by rescaling the feature maps of the activations to the input size. The layers were observed to be looking at indicators like lesion borders and non-uniformity in color as well as risk factors like lighter skin color or pink texture. However, spurious features like artifacts and hair which have no significance were also learned indicating some extent of overfitting.

There are other studies using attribution based methods for diagnosis in addition to the more common imaging modalities discussed above. For example, a study performed uncertainty and interpretability analysis on CNNs for semantic segmentation of colorectal polyps, a precursor of rectal cancers [51]. Using GBP for heatmaps the convolutional neural network (CNN)s were found to be utilizing the edge and shape information to make predictions. Moreover, the uncertainty analysis revealed higher uncertainty in misclassified samples. There is plenty of opportunity for applying the explainability of deep learning methods to other modalities like laparoscopy and endoscopy e.g., [52]. An explainable model using SHapley Additive exPlanations (SHAP) attributions for hypoxemia, i.e., low blood oxygen tension prediction during surgery was presented in [53]. The study was performed for analyzing preoperative factors as well as in-surgery parameters. The resulting attributions were in line with known factors like BMI, physical status (ASA), tidal volume, inspired oxygen, etc.

The attribution based methods were one of the initial ways of visualizing neural networks and have since then evolved from simple class activation map and gradient-based methods to advanced techniques like Deep SHapley Additive exPlanations (SHAP). The better visualizations of these methods show that the models were learning relevant features in most of the cases. Any presence of spurious features was scrutinized, flagged to the readers, and brought adjustments to the model training methods. Smaller and task-specific models like [39] along with custom variants of the attribution methods can improve the identification of relevant features.

4.2. Non-Attribution Based

The studies discussed in this subsection approached the problem of explainability by developing a methodology and validating it on a given problem rather than performing a separate analysis using pre-existing attributions based methods like those previously discussed. These used approaches like attention maps, concept vectors, returning a similar image, text justifications, expert knowledge, generative modeling, combination with other machine learning methods, etc. It must be noted that the majority of these are still post-model but their implementation usually needs specific changes to the model structure such as in the attention maps or the addition of expert knowledge in case of rule-based methods. In this section, the studies are grouped by the explainability approach they took. Figure 5 shows a schematic of these methods according to the taxonomy discussed in Section 2. These are characterized in a hierarchical way using multiple taxonomies for a finer classification.
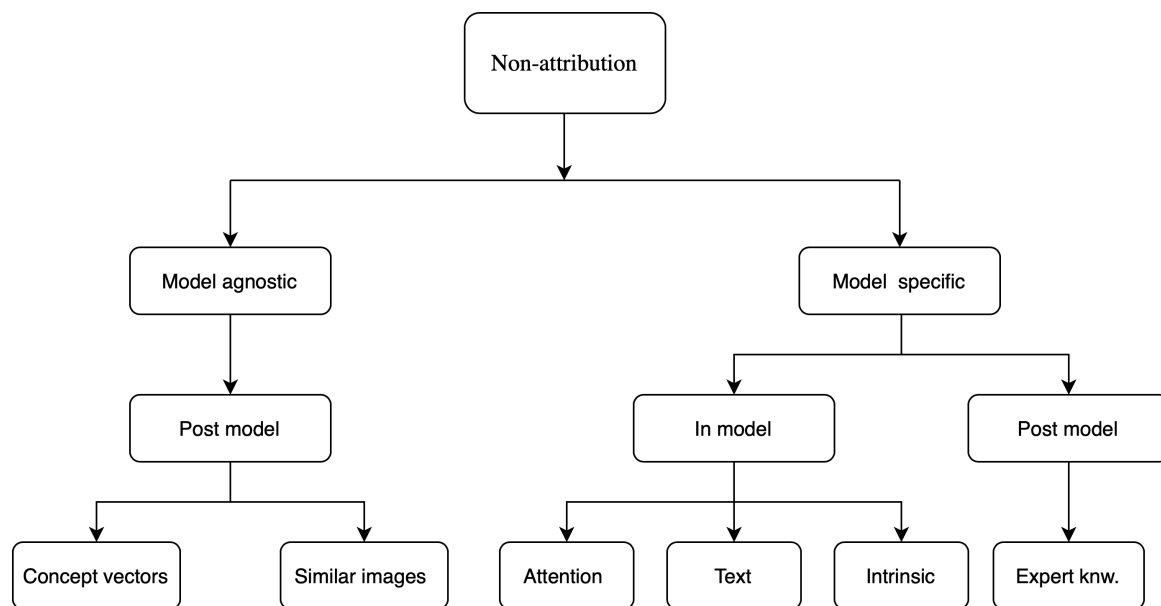

Figure 5. Classification of explainability methods that are not attribution based.

4.2.1. Attention Based

Attention is a popular and useful concept in deep learning. The basic idea of attention is inspired by the way humans pay attention to different parts of an image or other data sources to analyze them. More details about attention mechanisms in neural networks are discussed in [54]. An example of attention in medical diagnosis is given in [55]. Here, we discuss how attention-based methods can be used as an explainable deep learning tool for medical image analysis.

A network called MDNet was proposed [56] to perform a direct mapping between medical images and corresponding diagnostic reports. With an image model and a language model in it, the method used attention mechanisms to visualize the detection process. Using that attention mechanism, the language model found predominant and discriminatory features to learn the mapping between images and the diagnostic reports. This was the first work that exploited the attention mechanism to get insightful information from medical image dataset.

In [57] an interpretable version of U-Net [58] called SAUNet was proposed. It added a parallel secondary shape stream to capture important shape-based information along with the regular texture features of the images. The architecture used an attention module in the decoder part of the U-Net. The spatial and shape attention maps were generated using SmoothGrad to visualize the high activation region of the images.

4.2.2. Concept Vectors

A novel method called TCAV was proposed in [59] to explain the features learned by different layers to the domain experts without any deep learning expertise in terms of human-understandable concepts. It took the directional derivative of the network in the concept space much like that in the input feature space for saliency maps. It was tested to explain the predictions of diabetic retinopathy (DR) levels where it successfully detected the presence of microaneurysms and aneurysms in the retina. This provided justifications that were readily interpretable for the medical practitioners in terms of presence or absence of a given concept or physical structure in the image. However, many clinical concepts like the texture or the shape of a structure cannot be sufficiently described in terms of the presence or absence and need a continuous scale of measurement.

An extension of TCAV, which used the presence or absence of concepts, using Regression Concept Vectors (RCV) in the activation space of a layer was used to detect continuous concepts [60]. The task of the network was to detect tumors from breast lymph node samples. It was found that most of the relevant features like area and contrast were present in the early layers of the model. A further improvement over the TCAV used a new metric called Uniform unit Ball surface Sampling (UBS) [61] to provide layer-agnostic explanations for continuous and high dimensional features. It could explain high dimensional radiomics concepts across multiple layers which were validated using mammographic images. The model produced variations amongst the important concepts which were found to be lower across the layers of the SqueezeNet [62] compared to a baseline CNN with 3 dense layers explaining the better performance of the SqueezeNet.

4.2.3. Expert Knowledge

A vast majority of the research discussed in this review tried to correlate model features with expert knowledge using different approaches. Another approach was to use domain-specific knowledge to craft rules for prediction and explanation. An example of using task-specific knowledge to improve the results as well as the explanations were provided in [63] for brain midline shift (MLS) estimation using U-Net [58] based architecture and keypoints. It was reduced to the problem of detecting a midline using the model under domain constraints. The original midline was obtained using the endpoints and hence the shift from the predicted one was computed. The model also provided confidence intervals of the predictions making them more trustworthy for the end-user. Another study [64] used guidelines for rule-based segmentation of lung nodules followed by a perturbation analysis to compute the importance of features in each region. The explanations provided in terms of the regions already marked using rules were found to be more understandable for the users and showed the bias in data for improving the model. This method was then used to provide explanations at a global level for the entire dataset providing an overview of the relevant features.

4.2.4. Similar Images

Some studies provided similarly labeled images to the user as a reason for making a prediction for a given test image. A study [6] proposed analysis of layers of a 3D-convolutional neural network (CNN) using Gaussian mixture model (GMM) and binary encoding of training and test images based on their Gaussian mixture model (GMM) components for returning similar 3D images as explanations. The system returned activation wise similar training images using atlas as a clarification for its decision. It was demonstrated on 3D MNIST and an MRI dataset where it returned images with similar atrophy

conditions. However, it was found that the activation similarity depended on the spatial orientation of images in certain cases which could affect the choice of the returned images.

In a study on dermoscopic images, a triplet-loss and k nearest neighbors (kNN) search-based learning strategy was used to learn convolutional neural network (CNN) feature embeddings for interpretable classification [65]. The evidence was provided as nearest neighbors and local image regions responsible for the lowest distance between the test image and those neighbors. Another approach used monotonic constraints to explain the predictions in terms of style and depth two datasets—dermoscopy images and post-surgical breast aesthetics [66]. It concatenated input streams with constrained monotonic convolutional neural network (CNN) and unconstrained convolutional neural network (CNN) to produce the predictions along with their explanations in terms of similar images as well as complementary images. The system was designed for only binary classification.

4.2.5. Textual Justification

A model that can explain its decision in terms of sentences or phrases giving the reasoning can directly communicate with both expert and general users. A justification model that took inputs from the visual features of a classifier, as well as embeddings of the predictions, was used to generate a diagnostic sentence and visual heatmaps for breast mass classification [67]. A visual word constraint loss was applied in the training of the justification generator to produce justifications in the presence of only a limited number of medical reports. Such multimodal explanations can be used to obtain greater user confidence due to a similarity with the usual workflow and learning process.

4.2.6. Intrinsic Explainability

Intrinsic explainability refers to the ability of a model to explain its decisions in terms of human observable decision boundaries or features. These usually include relatively simpler models like regression, decision trees, and support vector machines (SVM) for a few dimensions where the decision boundaries can be observed. Recent studies to make deep learning model intrinsically explainable using different methods such as a hybrid with machine learning classifiers and visualizing the features in a segmentation space.

An example of the latter was presented in [68] using the latent space of the features of a variational autoencoder for classification and segmentation of the brain MRI of Alzheimer's patients. The classification was performed in a two-dimensional latent space using an multi layer perceptron (MLP). The segmentation was performed in a three-dimensional latent space in terms of the anatomical variability encoded in the discriminating features. This led to the visualization of the features of the classifier as global and local anatomical characteristics which were usually used for clinical decisions. A study for detection of autism pectrum disorder (ASD) from functional magnetic resonance imaging (fMRI) used a hybrid of deep learning and support vector machines (SVM) to perform explainable classification [69]. The support vector machines (SVM) was used as a classifier on the features of a deep learning model and the visualization of the decision boundary explained the model.

This subsection discussed a variety of non-attribution explainability methods but the list is not exhaustive as newer methods are published frequently due to high interest in the area. The design of these methods is more involved than the application of attribution based methods on the inputs of a trained model. Specific elements like concept vectors, expert-based rules, image retrieval methods need to be integrated often at a model training level. This added complexity can potentially provide more domain-specific explanations at the expense of higher design effort. Notably, a majority of these techniques are still a post-hoc step but for a specific architecture or domain. Moreover, we have limited our scope to medical imaging as that is the dominant approach for automated diagnosis because of the detailed information presented by the images. However, patient records also provide rich information for diagnosis and there were studies discussing their explainability. For example, in [70] a gated recurrent unit (GRU)-based recurrent neural network (RNN) for mortality prediction from diagnostic

codes from electronic healthcare record (EHR) was presented. It used hierarchical attention in the network for interpretability and visualization of the results.

Table 2. Applications of explainability in medical imaging.

Method	Algorithm	Model	Application	Modality
Attribution	Gradient*I/P, GBP, LRP, occlusion [36]	3D CNN	Alzheimer's detection	Brain MRI
	GradCAM, GBP [37]	Custom CNN	Grading brain tumor	Brain MRI
	IG [38]	Inception-v4	DR grading	Fundus images
	EG [39]	Custom CNN	Lesion segmentation for AMD	Retinal OCT
	IG, SmoothGrad [41]	AlexNet	Estrogen receptor status	Breast MRI
	Saliency maps [42]	AlexNet	Breast mass classification	Breast MRI
	GradCAM, SHAP [49]	Inception	Melanoma detection	Skin images
	Activation maps [50]	Custom CNN	Lesion classification	Skin images
	DeepDreams [46]	Custom CNN	Segmentation of tumor from liver	CT imaging
	GSInquire, GBP, activation maps [47]	COVIDNet CNN	COVID-19 detection	X-ray images
Attention	Mapping between image to reports [56]	CNN & LSTM	Bladder cancer	Tissue images
	U-Net with shape attention stream [57]	U-net based	Cardiac volume estimation	Cardiac MRI
Concept vectors	TCAV [59]	Inception	DR detection	Fundus images
	TCAV with RCV [60]	ResNet101	Breast tumor detection	Breast lymph node images
	UBS [61]	SqueezeNet	Breast mass classification	Mammography images
Expert knowledge	Domain constraints [63]	U-net	Brain MLS estimation	Brain MRI
	Rule-based segmentation, perturbation [64]	VGG16	Lung nodule segmentation	Lung CT
Similar images	GMM and atlas [6]	3D CNN	MRI classification	3D MNIST, Brain MRI
	Triplet loss, kNN [65]	AlexNet based with shared weights	Melanoma	Dermoscopy images
	Monotonic constraints [66]	DNN with two streams	Melanoma detection	Dermoscopy images
Textual justification	LSTM, visual word constraint [67]	Breast mass classification	CNN	Mammography images
Intrinsic explainability	Deep Hierarchical Generative Models [68]	Auto-encoders	Classification and segmentation for Alzheimer's	Brain MRI
	SVM margin [69]	Hybrid of CNN & SVM	ASD detection	Brain fMRI

5. Discussion

There has been significant progress in explaining the decisions of deep learning models, especially those used for medical diagnosis. Understanding the features responsible for a certain decision is useful for the model designers to iron out reliability concerns for the end-users to gain trust and make better judgments. Almost all of these methods target local explainability, i.e., explaining the decisions for a single example. This then is extrapolated to a global level by averaging the highlighted features, especially in cases where the images have the same spatial orientation. However, emerging methods like concept vectors (Section 4.2.2) provide a more global view of the decisions for each class in terms of domain concepts.

It is important to analyze the features of a black-box which can make the right decision due to the wrong reason. It is a major issue that can affect performance when the system is deployed in the real world. Most of the methods, especially the attribution based are available as open-source implementations. However, some methods like GSInquire [48] which show higher performance on some metrics are proprietary. There is an increasing commercial interest in explainability, and specifically the attribution methods which can be leveraged for a variety of business use cases.

The explainability methods have two different but overlapping objectives for the two different user groups. Deep learning practitioners can use them to design better systems by analyzing the model features and understanding the interactions between the model and the data. The clinical end-users can be provided with the explanations as a reasoning for the model decision and hence build confidence and trust in the model decision and also help identify potentially questionable decisions. A recent study compared the understanding of explanations amongst data scientists [71]. In this study common issues like missing data and redundant features were introduced and the data scientists were provided explanations of the trained models in order to identify the problems. The study reported over trust on the models as they tried to justify the issues as meaningful features. This is contrary to lower trust and acceptance from end-users who are wary of the black-box nature. It is notable that the experienced data scientists were able to use them effectively for understanding model and data issues.

Studies analyzing the effect of explanations on the decisions of the clinical end-users show in general positive outcomes [38]. There are studies comparing explainability methods quantitatively [36,39] which are discussed previously. The quantitative analysis focuses on theoretical correctness and robustness while missing out on actual clinical usefulness. There is a pertinent need to perform end-user based qualitative comparison of explanations for medical imaging applications. This can help to identify the most relevant techniques for explaining decisions to the clinicians. Such studies can be performed using expert agreement where a panel of experts can be asked to rate the explanations. A similar approach was used for deep learning based methods in [72] and for clinical diagnosis in [73]. We are currently working on a quantitative and qualitative analysis of various XAI methods in the diagnosis of retinal disease. The explanations will have quantitative comparisons along with qualitative evaluation by expert clinicians rating the explanations. This would also help to evaluate the overlap between the clinical knowledge acquired through clinical training and experience and the model features acquired for the pattern recognition task for a given dataset.

Studies have extended existing explainability methods to better suit the challenges of the medical imaging domain. For example, [39] proposed Expressive gradients (EG), an extension of commonly used IG to cover the retinal lesions better while [60] extended concept vectors from [59] for continuous concepts like texture and shape. Such studies lead to the advancement of the explainability domain and provided customization without designing new methods from scratch. Despite all these advances, there is still a need to make the explainability methods more holistic and interwoven with uncertainty methods. Expert feedback must be incorporated into the design of such explainability methods to tailor the feedback for their needs. Initially, any clinical application of such explainable deep learning methods is likely to be a human-in-the-loop (HITL) hybrid keeping the clinical expert in the control of the process. It can be considered analogous to driving aids like adaptive cruise control or lane keep

assistance in cars where the driver is still in control and responsible for the final decisions but with a reduced workload and an added safety net.

Another direction of work can be to use multiple modalities like medical images and patients' records together in the decision-making process and attribute the model decisions to each of them. This can simulate the diagnostic workflow of a clinician where both images and physical parameters of a patient are used to make a decision. This can potentially improve accuracy as well as explain the phenomena more comprehensively. To sum up, explainable diagnosis is making convincing strides but there is still some way to go to meet the expectations of end-users, regulators, and the general public.

Conflicts of Interest: The authors declare no conflict of interest.

Acronyms

AI	artificial intelligence
AMD	Age-related macular degeneration
ASD	autism pectrum disorder
CAD	Computer-aided diagnostics
CAM	Class activation maps
CNN	convolutional neural network
CNV	choroidal neovascularization
CT	computerized tomography
DME	diabetic macular edema
DNN	deep neural networks
DR	diabetic retinopathy
EG	Expressive gradients
EHR	electronic healthcare record
fMRI	functional magnetic resonance imaging
GBP	Guided backpropagation
GDPR	General Data Protection Regulation
GMM	Gaussian mixture model
GradCAM	Gradient weighted class activation mapping
GRU	gated recurrent unit
HITL	human-in-the-loop
IG	Integrated gradients
kNN	k nearest neighbors
LIFT	Deep Learning Important FeaTures
LRP	Layer wise relevance propagation
MLP	multi layer perceptron
MLS	midline shift
MRI	magnetic resonance imaging
OCT	optical coherence tomography
PCC	Pearson's correlation coefficient
RCV	Regression Concept Vectors
ReLU	rectified linear unit
RNN	recurrent neural network
SHAP	SHapley Additive exPlanations
SVM	support vector machines
TCAV	Testing Concept Activation Vectors
UBS	Uniform unit Ball surface Sampling

References

1. Jo, T.; Nho, K.; Saykin, A.J. Deep learning in Alzheimer's disease: Diagnostic classification and prognostic prediction using neuroimaging data. *Front. Aging Neurosci.* **2019**, *11*, 220. [CrossRef]
2. Hua, K.L.; Hsu, C.H.; Hidayati, S.C.; Cheng, W.H.; Chen, Y.J. Computer-aided classification of lung nodules on computed tomography images via deep learning technique. *OncoTargets Ther.* **2015**, *8*, 2015–2022.
3. Sengupta, S.; Singh, A.; Leopold, H.A.; Gulati, T.; Lakshminarayanan, V. Ophthalmic diagnosis using deep learning with fundus images–A critical review. *Artif. Intell. Med.* **2020**, *102*, 101758. [CrossRef] [PubMed]
4. Leopold, H.; Singh, A.; Sengupta, S.; Zelek, J.; Lakshminarayanan, V., Recent Advances in Deep Learning Applications for Retinal Diagnosis using OCT. In *State of the Art in Neural Networks*; El-Baz, A.S., Ed.; Elsevier: New York, NY, USA, 2020; in press.
5. Holzinger, A.; Biemann, C.; Pattichis, C.S.; Kell, D.B. What do we need to build explainable AI systems for the medical domain? *arXiv* **2017**, arXiv:1712.09923.
6. Stano, M.; Benesova, W.; Martak, L.S. Explainable 3D convolutional neural network using GMM encoding. In Proceedings of the Twelfth International Conference on Machine Vision, Amsterdam, The Netherlands, 16–18 November 2019; Volume 11433, p. 114331U.
7. Moccia, S.; Wirkert, S.J.; Kenngott, H.; Vemuri, A.S.; Apitz, M.; Mayer, B.; De Momi, E.; Mattos, L.S.; Maier-Hein, L. Uncertainty-aware organ classification for surgical data science applications in laparoscopy. *IEEE Trans. Biomed. Eng.* **2018**, *65*, 2649–2659. [CrossRef] [PubMed]
8. Adler, T.J.; Ardizzone, L.; Vemuri, A.; Ayala, L.; Gröhl, J.; Kirchner, T.; Wirkert, S.; Kruse, J.; Rother, C.; Köthe, U.; et al. Uncertainty-aware performance assessment of optical imaging modalities with invertible neural networks. *Int. J. Comput. Assist. Radiol. Surg.* **2019**, *14*, 997–1007. [CrossRef]
9. Meyes, R.; de Puiseau, C.W.; Posada-Moreno, A.; Meisen, T. Under the Hood of Neural Networks: Characterizing Learned Representations by Functional Neuron Populations and Network Ablations. *arXiv* **2020**, arXiv:2004.01254.
10. 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]
11. Stiglic, G.; Kocbek, P.; Fijacko, N.; Zitnik, M.; Verbert, K.; Cilar, L. Interpretability of machine learning based prediction models in healthcare. *arXiv* **2020**, arXiv:2002.08596.
12. Arya, V.; Bellamy, R.K.; Chen, P.Y.; Dhurandhar, A.; Hind, M.; Hoffman, S.C.; Houde, S.; Liao, Q.V.; Luss, R.; Mojsilović, A.; et al. One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques. *arXiv* **2019**, arXiv:1909.03012.
13. Ying, Z.; Bourgeois, D.; You, J.; Zitnik, M.; Leskovec, J. Gnnexplainer: Generating explanations for graph neural networks. In Proceedings of the Advances in Neural Information Processing Systems 32, Vancouver, BC, Canada, 8–14 December 2019; Volume 32, pp. 9240–9251.
14. Wold, S.; Esbensen, K.; Geladi, P. Principal component analysis. *Chemom. Intell. Lab. Syst.* **1987**, *2*, 37–52. [CrossRef]
15. Maaten, L.V.D.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
16. Safavian, S.R.; Landgrebe, D. A survey of decision tree classifier methodology. *IEEE Trans. Syst. Man Cybern.* **1991**, *21*, 660–674. [CrossRef]
17. 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]
18. Ancona, M.; Ceolini, E.; Öztireli, C.; Gross, M. Towards better understanding of gradient-based attribution methods for deep neural networks. *arXiv* **2017**, arXiv:1711.06104.
19. Alber, M.; Lapuschkin, S.; Seegerer, P.; Hägele, M.; Schütt, K.T.; Montavon, G.; Samek, W.; Müller, K.R.; Dähne, S.; Kindermans, P.J. iNNvestigate neural networks. *J. Mach. Learn. Res.* **2019**, *20*, 1–8.
20. 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; Springer: Cham, Switzerland; pp. 818–833.
21. Goodfellow, I.J.; Shlens, J.; Szegedy, C. Explaining and harnessing adversarial examples. *arXiv* **2014**, arXiv:1412.6572.

22. Lipovetsky, S.; Conklin, M. Analysis of regression in game theory approach. *Appl. Stoch. Model. Bus. Ind.* **2001**, *17*, 319–330. [CrossRef]

23. Simonyan, K.; Vedaldi, A.; Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv* **2013**, arXiv:1312.6034.

24. Springenberg, J.T.; Dosovitskiy, A.; Brox, T.; Riedmiller, M. Striving for simplicity: The all convolutional net. *arXiv* **2014**, arXiv:1412.6806.

25. 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*. [CrossRef] [PubMed]

26. Shrikumar, A.; Greenside, P.; Shcherbina, A.; Kundaje, A. Not just a black box: Learning important features through propagating activation differences. *arXiv* **2016**, arXiv:1605.01713.

27. 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. [CrossRef]

28. Sundararajan, M.; Taly, A.; Yan, Q. Axiomatic attribution for deep networks. In Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; Voume 70, pp. 3319–3328.

29. Kindermans, P.J.; Schütt, K.T.; Alber, M.; Müller, K.R.; Erhan, D.; Kim, B.; Dähne, S. Learning how to explain neural networks: Patternnet and patternattribution. *arXiv* **2017**, arXiv:1705.05598.

30. Shrikumar, A.; Greenside, P.; Kundaje, A. Learning important features through propagating activation differences. In Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; Voume 70, pp. 3145–3153.

31. Smilkov, D.; Thorat, N.; Kim, B.; Viégas, F.; Wattenberg, M. Smoothgrad: Removing noise by adding noise. *arXiv* **2017**, arXiv:1706.03825.

32. Chen, H.; Lundberg, S.; Lee, S.I. Explaining Models by Propagating Shapley Values of Local Components. *arXiv* **2019**, arXiv:1911.11888.

33. Yosinski, J.; Clune, J.; Bengio, Y.; Lipson, H. How transferable are features in deep neural networks? In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, USA, 8–13 December 2014; pp. 3320–3328.

34. Singh, A.; Sengupta, S.; Lakshminarayanan, V. Glaucoma diagnosis using transfer learning methods. In *Proceedings of the Applications of Machine Learning*; International Society for Optics and Photonics (SPIE): Bellingham, WA, USA, 2019; Volume 11139, p. 111390U.

35. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.

36. Eitel, F.; Ritter, K.; Alzheimer's Disease Neuroimaging Initiative (ADNI). 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, ML-CDS 2019, IMIMIC 2019*; Lecture Notes in Computer Science; Suzuki, K., et al., Eds.; Springer: Cham, Switzerland, 2019; Volume 11797. [CrossRef]

37. Pereira, S.; Meier, R.; Alves, V.; Reyes, M.; Silva, C.A. Automatic brain tumor grading from MRI data using convolutional neural networks and quality assessment. In *Understanding and Interpreting Machine Learning in Medical Image Computing Applications*; Springer: Cham, Switzerland, 2018; pp. 106–114.

38. Sayres, R.; Taly, A.; Rahimy, E.; Blumer, K.; Coz, D.; Hammel, N.; Krause, J.; Narayanaswamy, A.; Rastegar, Z.; Wu, D.; et al. Using a deep learning algorithm and integrated gradients explanation to assist grading for diabetic retinopathy. *Ophthalmology* **2019**, *126*, 552–564. [CrossRef]

39. Yang, H.L.; Kim, J.J.; Kim, J.H.; Kang, Y.K.; Park, D.H.; Park, H.S.; Kim, H.K.; Kim, M.S. Weakly supervised lesion localization for age-related macular degeneration detection using optical coherence tomography images. *PLoS ONE* **2019**, *14*, e0215076. [CrossRef]

40. Singh, A.; Sengupta, S.; Lakshminarayanan, V. Interpretation of deep learning using attributions: Application to ophthalmic diagnosis. In *Proceedings of the Applications of Machine Learning*; International Society for Optics and Photonics (SPIE): Bellingham, WA, USA, 2020; in press.

41. Papanastasopoulos, Z.; Samala, R.K.; Chan, H.P.; Hadjiiski, L.; Paramagul, C.; Helvie, M.A.; Neal, C.H. Explainable AI for medical imaging: Deep-learning CNN ensemble for classification of estrogen receptor status from breast MRI. In *Proceedings of the SPIE Medical Imaging 2020: Computer-Aided Diagnosis*; International Society for Optics and Photonics: Bellingham, WA, USA, 2020; Volume 11314, p. 113140Z.

42. Lévy, D.; Jain, A. Breast mass classification from mammograms using deep convolutional neural networks. *arXiv* **2016**, arXiv:1612.00542.

43. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 1–9.

44. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–6 December 2012; pp. 1097–1105.

45. Mordvintsev, A.; Olah, C.; Tyka, M. Inceptionism: Going Deeper into Neural Networks. Google AI Blog. 2015. Available online: https://ai.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html (accessed on 23 May 2020)

46. Couteaux, V.; Nempont, O.; Pizaine, G.; Bloch, I. Towards Interpretability of Segmentation Networks by Analyzing DeepDreams. In *Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support*; Springer: Cham, Switzerland, 2019; pp. 56–63.

47. Wang, L.; Wong, A. COVID-Net: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest radiography images. *arXiv* **2020**, arXiv:2003.09871.

48. Lin, Z.Q.; Shafiee, M.J.; Bochkarev, S.; Jules, M.S.; Wang, X.Y.; Wong, A. Explaining with Impact: A Machine-centric Strategy to Quantify the Performance of Explainability Algorithms. *arXiv* **2019**, arXiv:1910.07387.

49. Young, K.; Booth, G.; Simpson, B.; Dutton, R.; Shrapnel, S. Deep neural network or dermatologist? In *Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support*; Springer: Cham, Switzerland, 2019; pp. 48–55.

50. Van Molle, P.; De Strooper, M.; Verbelen, T.; Vankeirsbilck, B.; Simoens, P.; Dhoedt, B. Visualizing convolutional neural networks to improve decision support for skin lesion classification. In *Understanding and Interpreting Machine Learning in Medical Image Computing Applications*; Springer: Cham, Switzerland, 2018; pp. 115–123.

51. Wickstrøm, K.; Kampffmeyer, M.; Jenssen, R. Uncertainty and interpretability in convolutional neural networks for semantic segmentation of colorectal polyps. *Med Image Anal.* **2020**, *60*, 101619. [CrossRef] [PubMed]

52. Moccia, S.; De Momi, E.; Guarnaschelli, M.; Savazzi, M.; Laborai, A.; Guastini, L.; Peretti, G.; Mattos, L.S. Confident texture-based laryngeal tissue classification for early stage diagnosis support. *J. Med Imaging* **2017**, *4*, 034502. [CrossRef] [PubMed]

53. Lundberg, S.M.; Nair, B.; Vavilala, M.S.; Horibe, M.; Eisses, M.J.; Adams, T.; Liston, D.E.; Low, D.K.W.; Newman, S.F.; Kim, J.; et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat. Biomed. Eng.* **2018**, *2*, 749–760. [CrossRef]

54. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.

55. Bamba, U.; Pandey, D.; Lakshminarayanan, V. Classification of brain lesions from MRI images using a novel neural network. In *Multimodal Biomedical Imaging XV*; International Society for Optics and Photonics: Bellingham, WA, USA, 2020; Volume 11232, p. 112320K.

56. Zhang, Z.; Xie, Y.; Xing, F.; McGough, M.; Yang, L. Mdnet: A semantically and visually interpretable medical image diagnosis network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6428–6436.

57. Sun, J.; Darbeha, F.; Zaidi, M.; Wang, B. SAUNet: Shape Attentive U-Net for Interpretable Medical Image Segmentation. *arXiv* **2020**, arXiv:2001.07645.

58. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; Springer: Cham, Switzerland; pp. 234–241.

59. 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). *arXiv* **2017**, arXiv:1711.11279.

60. Graziani, M.; Andrearczyk, V.; Müller, H. Regression concept vectors for bidirectional explanations in histopathology. In *Understanding and Interpreting Machine Learning in Medical Image Computing Applications*; Springer: Cham, Switzerland, 2018; pp. 124–132.

61. Yeche, H.; Harrison, J.; Berthier, T. UBS: A Dimension-Agnostic Metric for Concept Vector Interpretability Applied to Radiomics. In *Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support*; Springer: Cham, Switzerland, 2019; pp. 12–20.

62. Iandola, F.N.; Han, S.; Moskewicz, M.W.; Ashraf, K.; Dally, W.J.; Keutzer, K. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5 MB model size. *arXiv* **2016**, arXiv:1602.07360.

63. Pisov, M.; Goncharov, M.; Kurochkina, N.; Morozov, S.; Gombolevsky, V.; Chernina, V.; Vladzymyrskyy, A.; Zamyatina, K.; Cheskova, A.; Pronin, I.; et al. Incorporating Task-Specific Structural Knowledge into CNNs for Brain Midline Shift Detection. In *Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support*; Springer: Cham, Switzerland, 2019; pp. 30–38.

64. Zhu, P.; Ogino, M. Guideline-Based Additive Explanation for Computer-Aided Diagnosis of Lung Nodules. In *Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support*; Springer: Cham, Switzerland, 2019; pp. 39–47.

65. Codella, N.C.; Lin, C.C.; Halpern, A.; Hind, M.; Feris, R.; Smith, J.R. Collaborative Human-AI (CHAI): Evidence-based interpretable melanoma classification in dermoscopic images. In *Understanding and Interpreting Machine Learning in Medical Image Computing Applications*; Springer: Cham, Switzerland, 2018; pp. 97–105.

66. Silva, W.; Fernandes, K.; Cardoso, M.J.; Cardoso, J.S. Towards complementary explanations using deep neural networks. In *Understanding and Interpreting Machine Learning in Medical Image Computing Applications*; Springer: Cham, Switzerland, 2018; pp. 133–140.

67. Lee, H.; Kim, S.T.; Ro, Y.M. Generation of Multimodal Justification Using Visual Word Constraint Model for Explainable Computer-Aided Diagnosis. In *Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support*; Springer: Cham, Switzerland, 2019; pp. 21–29.

68. Biffi, C.; Cerrolaza, J.J.; Tarroni, G.; Bai, W.; De Marvao, A.; Oktay, O.; Ledig, C.; Le Folgoc, L.; Kamnitsas, K.; Doumou, G.; et al. Explainable Anatomical Shape Analysis through Deep Hierarchical Generative Models. *IEEE Trans. Med. Imaging* **2020**. [CrossRef]

69. Eslami, T.; Raiker, J.S.; Saeed, F. Explainable and Scalable Machine-Learning Algorithms for Detection of Autism Spectrum Disorder using fMRI Data. *arXiv* **2020**, arXiv:2003.01541.

70. Sha, Y.; Wang, M.D. Interpretable predictions of clinical outcomes with an attention-based recurrent neural network. In Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, Boston, MA, USA, 20–23 August 2017; pp. 233–240.

71. Kaur, H.; Nori, H.; Jenkins, S.; Caruana, R.; Wallach, H.; Wortman Vaughan, J. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. In Proceedings of the CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, 25–30 April 2020; pp. 1–14. [CrossRef]

72. Arbabshirani, M.R.; Fornwalt, B.K.; Mongelluzzo, G.J.; Suever, J.D.; Geise, B.D.; Patel, A.A.; Moore, G.J. Advanced machine learning in action: Identification of intracranial hemorrhage on computed tomography scans of the head with clinical workflow integration. *NPJ Digit. Med.* **2018**, *1*, 1–7. [CrossRef]

73. Almazroa, A.; Alodhayb, S.; Osman, E.; Ramadan, E.; Hummadi, M.; Dlaim, M.; Alkatee, M.; Raahemifar, K.; Lakshminarayanan, V. Agreement among ophthalmologists in marking the optic disc and optic cup in fundus images. *Int. Ophthalmol.* **2017**, *37*, 701–717. [CrossRef]