European Journal of Radiology 162 (2023) 110787

Contents lists available at ScienceDirect

European Journal of Radiology

journal homepage: www.elsevier.com/locate/ejrad

Review

Explainable AI in medical imaging: An overview for clinical practitioners – Saliency-based XAI approaches

Katarzyna Borys [a,e,*], Yasmin Alyssa Schmitt [a], Meike Nauta [a,f], Christin Seifert [a], Nicole Krämer [b,c], Christoph M. Friedrich [d,g], Felix Nensa [a,e,1]

[a] Institute for Artificial Intelligence in Medicine, University Hospital Essen, Girardetstraße 2, 45131 Essen, Germany
[b] Department of Social Psychology, Media and Communication, University of Duisburg-Essen, Forsthausweg 2, 47057 Duisburg, Germany
[c] Research Center "Trustworthy Data Science and Security", Otto-Hahn-Straße 14, 44227 Dortmund, Germany
[d] Department of Computer Science, University of Applied Sciences and Arts Dortmund, Emil-Figge-Straße 42, 44227 Dortmund, Germany
[e] Institute of Diagnostic and Interventional Radiology and Neuroradiology, University Hospital Essen, Hufelandstraße 55, 45147 Essen, Germany
[f] Data Management & Biometrics Group, University of Twente, Drienerlolaan 5, 7522 NB Enschede, The Netherlands
[g] Institute for Medical Informatics, Biometry, and Epidemiology (IMIBE), Zweigertstraße 37, 45130 Essen, Germany

ARTICLE INFO

Keywords:
Explainable AI
Medical imaging
Radiology
Black-Box
Explainability
Interpretability

ABSTRACT

Since recent achievements of Artificial Intelligence (AI) have proven significant success and promising results throughout many fields of application during the last decade, AI has also become an essential part of medical research. The improving data availability, coupled with advances in high-performance computing and innovative algorithms, has increased AI's potential in various aspects. Because AI rapidly reshapes research and promotes the development of personalized clinical care, alongside its implementation arises an urgent need for a deep understanding of its inner workings, especially in high-stake domains. However, such systems can be highly complex and opaque, limiting the possibility of an immediate understanding of the system's decisions. Regarding the medical field, a high impact is attributed to these decisions as physicians and patients can only fully trust AI systems when reasonably communicating the origin of their results, simultaneously enabling the identification of errors and biases. Explainable AI (XAI), becoming an increasingly important field of research in recent years, promotes the formulation of explainability methods and provides a rationale allowing users to comprehend the results generated by AI systems. In this paper, we investigate the application of XAI in medical imaging, addressing a broad audience, especially healthcare professionals. The content focuses on definitions and taxonomies, standard methods and approaches, advantages, limitations, and examples representing the current state of research regarding XAI in medical imaging. This paper focuses on saliency-based XAI methods, where the explanation can be provided directly on the input data (image) and which naturally are of special importance in medical imaging.

1. Introduction

In recent years, the number of *Artificial Intelligence* (AI) based applications for research and clinical care in medicine has increased dramatically, with medical imaging clearly being the focus of such developments [1]. Among the AI methods, *Deep Learning* (DL) has proven significant success in the field of computer vision, enabling promising solutions to medical-related image tasks, such as image preprocessing, registration, detection, and segmentation. To a certain extent, resulting systems have the potential to improve diagnostic accuracy or even exceed human performance [2]. The main advantage of a DL system is the versatile range of applications it can be deployed in, mainly because of the vast amount of successive layers within an *Artificial Neural Network* (ANN), a fundamental element of DL that enables variably derived mapping functions between input data and the desired outputs (e.g., prediction). The mapping for a neural network is approximated

* Corresponding author at: Institute for Artificial Intelligence in Medicine, Institute of Diagnostic and Interventional Radiology and Neuroradiology, University Hospital Essen, Hufelandstraße 55, 45147 Essen, Germany.
E-mail address: Katarzyna.Borys@uk-essen.de (K. Borys).

[1] Given their role as Associated Editor/Guest Editor, Prof. Felix Nensa had no involvement in the peer-review of this article and has no access to information regarding its peer-review.

https://doi.org/10.1016/j.ejrad.2023.110787
Received 12 January 2023; Received in revised form 3 March 2023; Accepted 14 March 2023
Available online 21 March 2023

during an elaborate training phase. However, the very nature of such networks often impedes a direct interpretation of the results, mainly because of the inherent unpredictability of mapping functions coupled with high structural complexity. The most relevant aspect concerning recent optimizations and advances of DL was initially limited to functional improvement [3,4]. Past years' research focused on implementing innovative and powerful system architectures, pursuing the goal of providing the best possible solutions to several tasks. This led to increasingly opaque and complex systems. At the same time, explainability and interpretability suffered under this trend, resulting in increased difficulty in understanding the prediction process and inner workings of emerging solutions. Consequently, the complexity of newly published AI solutions continues to increase, in some cases already operating in the trillion parameter range [5]. While such large parameter spaces lead to ever-increasing performance and expand application areas, the human ability to understand, comprehend and interpret how these systems work no longer keeps up, as depicted in Fig. 1. Consequently, the corresponding perspective on DL typically remains limited to a Black-Box principle impeding its implementation within the medical domain.

Especially where decisions made with the help of AI can have severe consequences, efforts are usually made to control and supervise these systems by human experts. This applies extraordinarily to the healthcare sector, where the consequences of such decisions can mean life and death. Moreover, even though the lack of explainability and interpretability does not hinder an outstanding performance of DL approaches, nor (depending on the domain) is it mandatory, the introduction of methods to better understand AI can be beneficial in many aspects. First and foremost, the need for explainability is strongly related to the scope and potential impact of a system's decisions in the application domain context. Consequently, in high-stake domains, such as law, autonomous driving, or healthcare, the question of "*Why*?" becomes existential. Especially against the background that the use of such systems aims at an interaction between healthcare professionals and AI-based support systems, which directly impacts patients and their health, it becomes obvious why the origin of decisions must be explainable. In addition, interpretability and explainability enable the linking of clinicians' and doctors' domain expertise with specific results. Moreover, explainability not only aims at functional benefits but also adds to clinical confidence and consistent compliance with legal and ethical requirements. These refer to regulations such as the European Union's General Data Protection Regulation, which enforces the right of patients to receive transparent information about a decision's origin [6], or the European AI Act, which introduces a regulatory and legal framework for AI systems. Lastly, identifying errors, limitations, and potential biases is essential in developing and applying AI systems and must be addressed accordingly [7]. For that reason, after the zeal to achieve exceedingly high performance, the demand for interpretability and explainability of AI has experienced a tremendous resurgence over recent years, promoting the formation of a new research field known as *eXplainable AI* (XAI) [8,9]. Generally, XAI refers to techniques that enable XAI stakeholders [10] to understand an AI algorithm and its decisions better. As proposed by Gilpin et al. [11] and later incorporated by Nauta et al. [12], this can be achieved by disclosing its reasoning, functioning, or behavior in human-understandable terms. However, depending on domain knowledge and AI expertise, the understandability, and need for explainability may vary across XAI stakeholder groups. In this paper, we primarily refer to *end-users,* including radiologists, clinicians, and doctors whose main need presumably might be the linking of knowledge and confirmation of diagnosis. For completeness, we disclosed an overview of all XAI stakeholder groups within the appendices (Appendix A). As depicted in Fig. 2, several approaches lead to an explanation, with the starting point depending on whether a given AI model is explainable by nature (called *White-Box models*) or not (named *Black-Box model*).

In this review, we identify key concepts and present an overview of established saliency-based (also called *visual*) XAI methods mined from diverse literature to provide end-users in the medical imaging domain (radiologists, clinicians, and doctors) an understanding of XAI's potential in a non-technical manner. The paper's contributions can be summarized as follows:

- Presenting a taxonomy for categorizing saliency-based XAI methods into well-defined groups to provide an intuitive overview.
- Categorization and summarization of the core components and algorithms of established XAI methods using the proposed taxonomy.
- Outlining of the limitations, pitfalls, and potentials of included visual XAI methods concerning practical implementation, evaluation, and interpretation and providing accompanying implementations in medical imaging.
- Summarization of the current state of XAI research in medical imaging and pointing out future directions.

1.1. Defining AI, ML, and DL

Recently, the promotion and usage of AI-based solutions have become common, leading to an interchangeable use of the terms "artificial intelligence", "machine learning", and "deep learning". Even though they are closely related, noteworthy differences must be considered. AI is a broad field comprising *Machine Learning* (ML) and DL but also covers various other disciplines, not focusing on any learning mechanisms at all. Generally, problems solved by general-operating AI systems can be phrased mathematically by a list of formal rules. The real potential of AI proved to be solving tasks that can be performed by humans easily but are difficult to define formally. In other words, AI focuses on transferring tasks, usually performed by humans, to the computational domain, traditionally realized by formulating explicit rule sets. A subset of AI is ML, which does not focus on imitating human capabilities by being programmed explicitly but instead adapts human learning mechanisms. In order to learn how to solve a given problem, a so-called ML model must be *trained* with a sufficiently large amount of data, representing the knowledge base from which rules can be automatically derived. In this context, a model refers to a parameterizable entity obtained during the training process to gain the ability to perform a pre-defined task such as pattern recognition, regression, or classification. Despite successful implementations, traditional ML limitations are attributed to feature extraction from data requiring human involvement. Additionally, complex tasks can exceed the capacities of ML models and

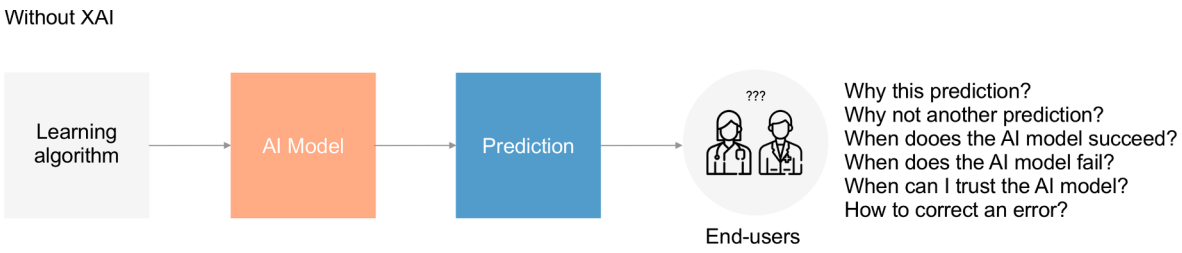

Fig. 1. Abstracted depiction of the interaction between AI models and end-users, which is not extended by the use of XAI.

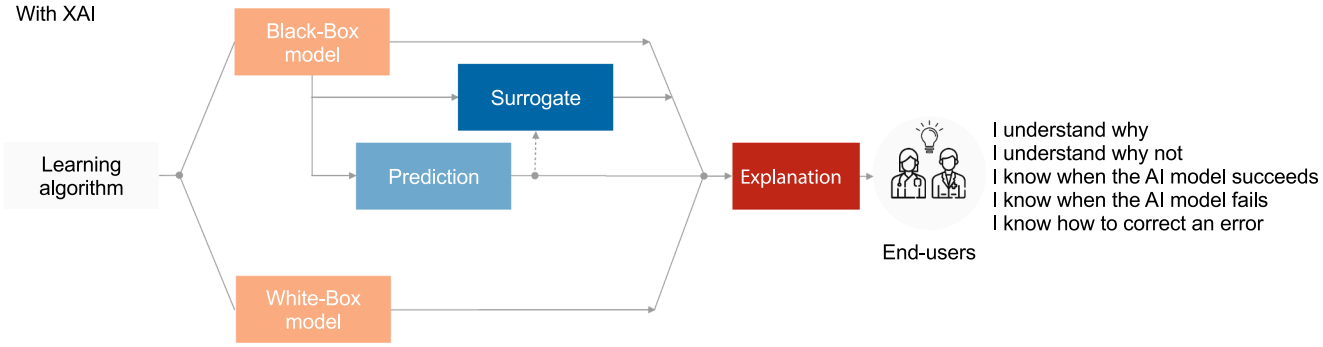

Fig. 2. Distinction of a deployment setup concerning an AI system with XAI. White-Box models are explainable by nature and can directly yield explanations. In contrast, Black-Box models must be explained globally (the model as a whole) or locally (a model's single prediction). Both views can optionally include the training of a subsequent surrogate model. Surrogate models are explainable models trained to approximate the predictions of an opaque Black-Box model.

not be adequately solved. Those limitations are what DL, a domain within ML comprising representational learning methods, excels at. As current DL algorithms are inspired by computational models of biological learning, one of the names that DL models go by is ANNs [13]. However, even though ANNs have partially been applied to understand brain function, it is essential to state that they can not be mistaken for realistic modeling of biological brain functionality. The modern definition of DL transcends the neuroscientific perspective by proposing a general learning approach comprising multiple composition levels representing automatically extracted features [3]. Consequently, DL can be defined as a mathematical framework based on deep ANNs for automatically deriving representations from given data [10]. DL proved to be a suitable approach for solving challenging and highly complex problems by automating the feature extraction process. Automatically extracted feature compositions allow a condensed representation, reflecting a hierarchy of structures in the data containing specific features – such as shapes, edges, and textures in the context of image data. Understanding how AI is implemented into visual imagery interpretation requires understanding another currently relevant DL network architecture: *Convolutional Neural Networks* (CNNs) [14].

These are ANNs that are designed for image data. In general, this architecture benefits from its ability to drastically reduce the number of parameters within the network inspired by biological perception mechanisms. Within a CNN, convolutional layers successively use learnable filters on input images to produce so-called *feature maps* denoting the detection of certain features in input images, such as shapes, colors, or structures [15]. Subsequent *pooling layers* transform each obtained feature map into a more condensed depiction. During inference, these filters represent the weights and connections within the CNN that can parametrize complex mapping functions from input images to specific outputs, also called *predictions* (e.g., "*1*" indicating the existence of a disease in the input image and "*0*" denoting its absence) as exemplarily depicted in Fig. 3 for a binary classification task. The forward pass (or forward propagation) refers to the calculation of intermediate variables (including predictions) from input to output. In contrast, during *backpropagation,* the order is traversed from the output to the input, yielding intermediate variables (partial derivatives) required while calculating the gradient with respect to specific parameters. Backpropagation is used for weight adjustment during training but can also be used during inference. For example, if the network from Fig. 3 predicted class "*1*" (fracture), backpropagation could be applied to uncover which image features (stored in the learned feature maps) contributed to the model's decision.

1.2. XAI-based taxonomies

Similar to AI, ML, and DL, in the research field of XAI, the terms "explainability" and "interpretability" are often used synonymously, without distinctive taxonomy or definition [16]. Concerning a clear taxonomic demarcation, there is neither a mathematical nor a fixed definition of interpretability in any other sense [17]. As stated by Lipton, "the term explainability holds no agreed-upon meaning, and yet machine learning conferences frequently publish papers which wield the term in a quasi-mathematical way", which underlines the ambiguity of the term [16]. However, for the context of this paper, a non-mathematical definition by Miller will be used, who studied the topic from the perspective of social sciences [18]: "*Interpretability is the degree to which a person can understand the cause of a decision*". Generally, interpretability is concerned chiefly with the intuition behind produced outputs of a model to identify cause-and-effect relationships. In contrast, explainability is associated with a system's internal logic and procedures [19]. As outlined in Fig. 4, the grouping of XAI methods throughout this

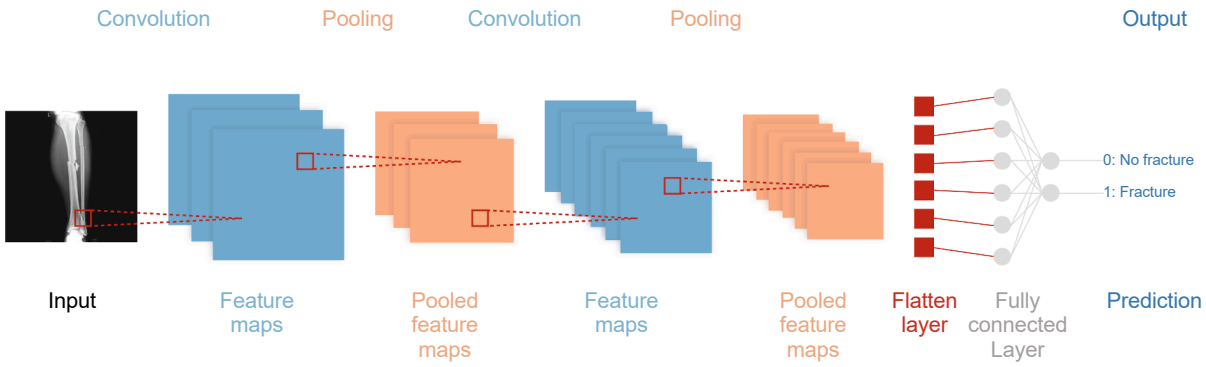

Fig. 3. Architectural setup of a CNN solving a binary classification task for detecting fractures containing Convolution layers to generate feature maps by applying filters (red rectangles with dashed lines) and Pooling layers to reduce the feature maps' dimensionality. The filters are moved iteratively across the whole image, capturing relevant image features. The specific filter values are learned during model training and are often called the weights of the CNN.

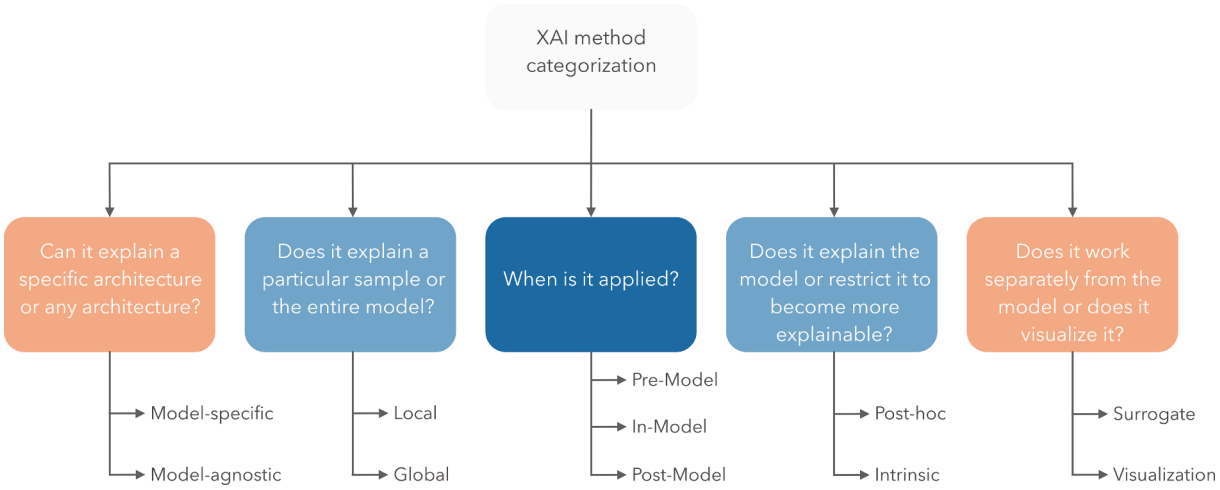

Fig. 4. Distinction between common XAI-related attributes used for categorizing explanations methods.

paper is conducted from the taxonomy proposed by the authors of [3]. This taxonomy scheme is primarily used to characterize the presented XAI methods within this paper. Even though a characterization of common properties serves as a distinctive classification, those properties are not necessarily mutually exclusive.

2. Visual XAI methods in medical imaging

In medical imaging, several frequently employed explainability approaches are based on visual (saliency-based) explanations related to a model's intermediate or final output [6]. For saliency-based approaches, the main idea is to exploit spatial information preserved through convolutional layers of a model, analyzing which parts of an image lead to a resulting decision. Conclusively an image's *salient* parts, with the highest attribution to the prediction, are then highlighted. When the attributions of all input features are arranged together to have the same shape as the input sample, they form so-called *attribution maps,* which usually are represented as heatmaps in which one color indicates a feature's positive contribution to the activation of the target output, while another color indicates a suppressing effect [20]. In technical terms, the generation of attribution maps can be broken down into *perturbation-based* or *back-propagation-based* methods [3]. Commonly, both categories are implemented as post-hoc visualization approaches and provide local explanations, as will be shown using concrete examples in the upcoming sections.

In order to determine which trends exist regarding the implementation of XAI methods in medical imaging, we performed a PubMed analysis based on a manual categorization using Rayyan [22] of all methods into *visual* (saliency-based) and *non-visual* (textual, auxiliary, and case-based). Textual approaches contain methods to semantically explain a prediction [23–25], whereas auxiliary measures refer to methods providing insights in tabular or graphical form, such as feature importances or statistical indicators. Lastly, case-based explanations can help to identify task-related concepts [26] or influential samples [27]. As indicated by the results in Fig. 5, among existing XAI methods, visual explanations seem to be the most frequent choice in medical image analysis which might be attributed to the ease of understanding and interpretation. Generally, visual explanations enable ascertaining whether a model's decision-making is similar to that of a radiologist and if relevant anatomical reference points are learned. If noticeable differences occur, such methods can be used to identify possible sources of error and biases quickly. However, the contribution of visual XAI approaches is highly diversified and not confined to bias identification only but also allows for transparent systems and diagnosis confirmation.

A vast amount of papers explaining DL models in medical image diagnosis use saliency-based (visual) methods, which is presumably due to their model-agnostic plug-and-play nature alongside conveniently available open-source implementations (see Appendix C) [3]. As justified by the trend in Fig. 5, this work's content is confined to visual XAI approaches. Moreover, we compared the Top 5 (measured by the

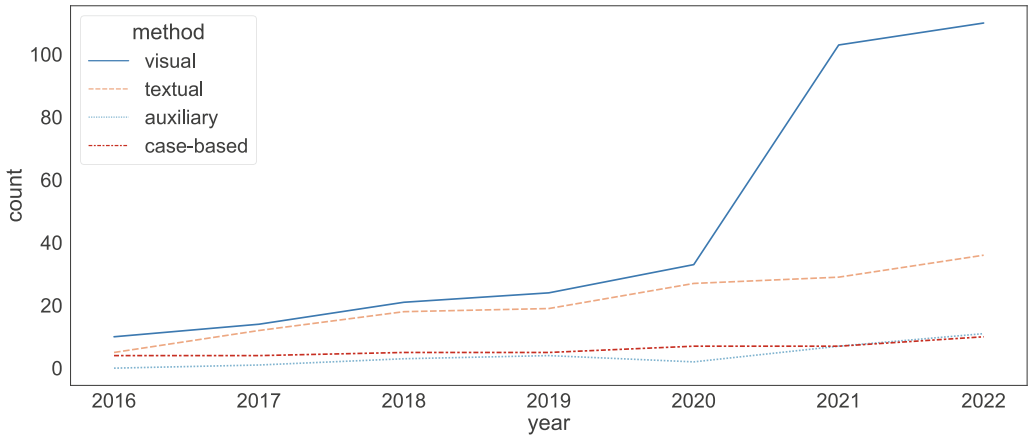

Fig. 5. Annual development of saliency-based (visual) and other non-visual (frequently divided into textual, example-based- case-based [21]) XAI methods applied in medical images, based on the cumulative number of citations (see Supplementary File 1).

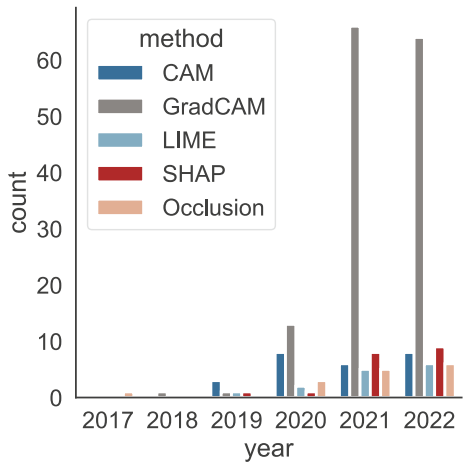

Fig. 6. Annual development of Top 5 saliency-based XAI methods applied in medical image analysis based on the total number of citations.

citation/year-ratio) visual XAI methods in medical imaging, with GradCAMs being the most frequent, as shown in Fig. 6. All displayed methods will be explained in detail in the subsequent sections.

2.1. Perturbation-based explanations

Perturbation is a technique to analyze the effect of altering the input features on the output of an AI model. That is achieved by removing, masking, or modifying certain input features, then performing the forward pass (computing the model's prediction) and measuring the deviation from the initial prediction. Generally, if image dimensions are independent, then perturbation is perfectly suitable, as the marginal effect of each dimension can be measured precisely. Unfortunately, specifically for image data, dimensions and pixels are strongly interdependent. Having an X-ray image of the lungs, removing a single pixel would probably not cause a significant change in the image's prediction class. The reason is that a pixel's information can be reconstructed based on its neighbors. Therefore, removing whole patches instead of individual pixels is more effective in achieving an effect. Common patch sizes are 5×5, or 10×10, depending on the size of target features and computational capacities [3].

For example, **Occlusion** [28], which Zeiler and Fergus first used, is a technique to generate attribution maps showing which parts of the image positively or negatively contribute to the score for a specific class. When applying occlusion, the input image is divided into rectangular patches of a particular size which are iteratively replaced with a predefined baseline. The attribution map's resolution is confined to the chosen patch size. A systematic perturbation of the input image is to monitor the effect on the output. This concept exploits that the most significant parts will strongly impact the output when altered. The implementation is model-agnostic as the method does not require access to the model's internal components. Tang et al. [29] generated occlusion maps to interpret models developed to classify pathologies from histopathological slides related to Alzheimer's disease, such as cored plaques, diffuse plaques, and cerebral amyloid angiopathy. However, a potential drawback is a high computational expense in terms of required computation time compared to other methods [30]. A similar approach, namely **Local Interpretable Model-agnostic Explanations (LIME)** [31], is a prevalent method using a comparative approach to occlusion, which involves training a secondary, interpretable surrogate model – e.g., linear regression or decision tree – to explain individual predictions made by the primary, opaque model. LIME's approach to image data is based on producing variations of the images by contiguously grouping pixels of similar intensity into so-called *superpixels*. These superpixels, which can be of any arbitrary shape, are then "switched off" iteratively by replacing the entries in each superpixel with a user-defined value (e.g., a specific intensity like 0 or the median). Identified relevant regions can then be shown delimited or highlighted on the input image. Magesh et al. [32] developed a model for the binary classification of DaT scan images as having Parkinson's disease or not and provided LIME explanations. Like occlusion, LIME is a model-agnostic and local method that can be applied to any model. The modularity and extensibility of this method compensate for the fact that it requires an appropriate specification of several parameters (i.e., surrogate model, image segmentation algorithm to derive the contiguous patches/ super-pixels, a similarity function for comparing original and perturbed sample, and the number of generated samples) [30]. Due to the number of tunable parameters, minor alterations can lead to significant differences in results. Therefore, concerning reproducibility, awareness of these parameters plays an important role. Moreover, in some cases, obtained attribution maps can convey a rather low image quality, as superpixels are coarse. However, other methods can yield improved quality. For example, **Meaningful Perturbations** [33] proposes optimized perturbation masks that blur the image as little as necessary while maximally decreasing a class score, posing an optimization problem. This differs from other visualization techniques using heuristics like high positive gradient values as an indicator of relevance. Instead of blocking out image parts (which can lead to out-of-distribution data), they suggested using naturalistic effects, such as blurring, noising, or constant value, to obtain more realistic perturbation and insightful explanations. However, Uzunova et al. [34] reasoned that this approach is unsuitable for medical imaging since replacement with constant values is implausible. Medical images naturally contain noise and blur, leading to inconsistencies. Instead, they proposed replacing pathologies on Optical Coherence Tomography (OCT) images and brain lesion MRIs with their healthy-looking equivalent using a variational autoencoder (VAE) and analyzed the effects on classification. Results indicate that the VAE's perturbations realistically pinpoint pathological regions in both imaging studies, OCT and MRI. Moreover, the VAE's localization of pathologies is superior to blurring or constant-value perturbations. Another extension of Meaningful Perturbation was pursued in **Real-time Perturbations** [35] by proposing a single forward-pass method of estimating an optimal perturbation mask. This specific task was delegated to a second neural network for which the generation of perturbation masks was defined as a learnable task of saliency detection. Moreover, **Extremal Perturbations** [36] aim to identify optimal perturbation masks by proposing a new operator called *smooth* max. In this context, optimal (extremal) perturbations refer to those areas with a maximal effect on the network's output among all perturbations. In contrast, so-called **SHapley Additive exPlanations (SHAP)** [37] utilize additive feature attribution methods based on Shapley values. These refer to a method from a coalitional game theory that estimates how to fairly distribute the "payout" (prediction) among the model's features [38]. In this case, the"players" are the individual features of a given instance, and the"pay-off" is the corresponding prediction minus the current average prediction for all instances. In general, it combines the idea of LIME with Shapley values. In practice, the input data is divided into coalitions and permuted by their presence or absence, on which a linear model is trained. However, the way the new instances are weighted differs from the LIME method using the Shapley value estimation of the corresponding coalitions. A downside of the original approach is that it scales exponentially in the number of features. Consequently, the authors of [37] proposed **KernelSHAP**, an approximate, computationally feasible method inspired by local surrogate models. For SHAP, global interpretations are consistent with the local explanations as Shapley values are the unit of global interpretations. However, resulting values can be misinterpreted, and – like many perturbation-based methods – it ignores feature dependence

(e.g., correlation). For example, the algorithm would assign a large weight to a feature when given a group of highly correlated features [A, B, C] and one arbitrary representative of the group, e.g., A. Consequently, B and C would hardly provide any additive information than A. Features B and C are redundant, and the model's decision-making process will rely heavily on feature A. Consequently, features B and C will score poorly when SHAP is used to explain the model. If SHAP assigned high-importance scores to features B and C, it would not be faithful to the model, even though they might represent similar statistical relations as A concerning the label [38]. Generally, correlation bias occurs because of how the ML algorithm trains the model, not how SHAP estimates feature importance. Young et al. [39] investigated the usage and explainability of KernelSHAP on dermoscopic images. Another perturbation-based approach employed in **Randomized Input Sampling for Explanation of Black-box Models (RISE)** [40] occludes input images using random occlusion patterns that are produced by sampling small binary masks (e.g., 7×7) and then interpolating them to larger resolutions. The model's response to each masked image can be captured by subsampling the images, revealing significant image areas. The final attribution map is a linear combination of the binary masks using weighted sums of the predicted scores multiplied with the masks. One limitation of RISE and Occlusion is that they do not consider objects' morphology and produce only approximate results. Moreover, as pointed out by Cooper et al. [41], both RISE and Meaningful Perturbation rely on pre-defined parameters, such as the number of epochs or generated masks and kernel size. Therefore, it can be difficult to determine optimal parameter values upfront to trade off accuracy and efficiency. Consequently, Cooper et al. [41] proposed a new technique called **Hierarchical Perturbation (HiPe)**, which extends RISE. During their analyses, the authors' key findings were that a large amount of computation occurs when regions with little effect on the model output are iteratively perturbed or when random perturbation region selection causes spatially similar or overlapping regions. To counteract these findings, salience thresholding was applied to perform model-agnostic saliency mapping of comparable quality and significantly faster than existing methods at a fraction of the computational cost. HiPe does this by focusing on perturbing the most salient regions with increasing resolution while ignoring regions that do not change the model's output. Respectively, regions with little impact are discarded [41].

2.2. Backpropagation-based explanations

In contrast, in *backpropagation-based* methods, an XAI algorithm performs one or more forward passes through the network and yields attribution maps calculating partial derivatives during the backpropagation stage to estimate the impact of gradients, weights, and activations. Therefore, this category subdivides further saliency maps, relevance maps, and class activation maps [42]. Simonyan et al. [43] were the first to propose a method utilizing backpropagation called **Saliency Map Visualization**. It calculates the partial derivative's absolute value regarding the target output class w.r.t. the network's input image. The idea is that a gradient's value indicates input features (pixels) that have the highest impact on the output. Respectively, pixels of the input image can be highlighted based on the amount of the positive gradient they receive, indicating their contribution. However, in some cases, the gradient is cut at zero when the input to the ReLU activation function during the forward pass is negative, resulting in a saturation problem [44]. This method is also known as **Vanilla Gradient** and can be seen as a generalization [38] of the later introduced **Deconvolution Networks (DeconvNets)** by Zeiler and Fergus [28]. DeconvNet's fundamental idea is to visualize neural activations of individual layers by reversing the network flow and setting the gradients to zero during the backward pass. Therefore, DeconvNets produce a CNN's opposite result with filters and unpooling operations, creating an activation map of neural (feature) activity. However, this method is confined to max-pooling layers, while the unpooling uses only an approximate inverse of the convolution features. De Vos et al. [45] predicted the slice-wise coronary artery calcium amount on CTs and subsequently applied deconvolution to visualize which slice region contributed to the prediction. A modification of this approach led to a new method called **Guided BackPropagation (GBP)** [46] which is based on the idea that neurons act as feature detectors. It is called "guided" because, during backpropagation, the activated neurons are chosen (guided) by setting negative gradients to zero, revealing image areas on which significant features exist. GBP is generic enough to work with several network architectures, like convolutional neural networks, generative adversarial networks, and fully-connected networks [47]. Adebayo et al. [48] showed one limitation: in some cases, GBP could be invariant to data and models, as in their evaluation, it performed partial input recovery. This is problematic, as recovering image parts does not explain predictions but emphasizes image features. Dubost et al. [49] introduced an automated method to quantify perivascular spaces on MRIs, which can indicate cerebral small vessel disease. To investigate if the neural networks learned the structures of interest or focused on other perivascular spaces' correlated features, the authors used GBP. Another visual explanation technique is **Layer-Wise Relevance Propagation (LRP)** [50]. Intuitively, LRP uses a network's weights and activations resulting from the forward pass and propagates the output back through the network until the input layer while assigning a relevance score to each input neuron from a preceding layer in each iteration. Put shortly, LRP is built upon the decomposition of the model's decision and generates distributed relevance scores that denote the connection between the activations of a given neuron and its input. Böhle et al. [51] used this method to identify regions from brain MRIs indicating Alzheimer's disease. Moreover, they compared saliency maps generated with LRP against those generated with Guided Backpropagation and concluded that LRP was more accurate in identifying relevant regions. One potential drawback of LRP is that resulting heatmaps can be sensitive to specific parameters; e.g., in the case of LRP, this can be the β value used to produce them [51].

Following the work on LRP, Shrikumar et al. [44] proposed **Deep Learning Important FeaTures (DeepLiFT),** which uses a reference image alongside the input image. This enables it to compare each neuron's activation to a reference activation and estimate the attribution according to the deviation. Moreover, separate consideration of positive and negative attributions can reveal dependencies that other approaches might not capture. De Souza et al. [52] applied DeepLiFT, among other methods, to highlight important regions for classifying early-cancerous tissues on endoscopic images in Barrett's esophagus-diagnosed patients. A limitation of DeepLIFT is that it is not *implementation invariant,* meaning that two identical models with different implementations but identical predictions can lead to deviating explanations [53]. Also, in response to a lack of sharpness in many attribution maps, Shrikumar et al. [54] proposed **Input * Gradient,** which calculates attribution by considering the output's partial derivatives w.r.t. the input and multiplying these with the input (pixel values). However, similar to Vanilla Gradients, this method can also encounter saturating gradients [55]. Stating that most gradient-based techniques lack certain "axioms" which are desirable characteristics for gradient-based approaches, Sundararajan et al. [53] point out that LRP, DeconvNets, GBP, and DeepLift have specific back-propagation logic that violates some axioms. For example, the axiom *completeness* requires an attribution method to completely justify the output, in the sense that attributions add up to the difference between the output of an input and a corresponding baseline (e.g., an all-zero vector) [56]. Therefore, Sundararajan et al. [56] proposed **Integrated Gradients (IG**), which in contrast to Input * Gradient, does not compute a single derivative evaluated at the input, but instead computes the average gradient by varying the input along a provided baseline (a

starting point that does not contain any information for the model, e.g., a black image). In this method, completeness [56], is satisfied: The attributions sum up to the target output minus the target output that was evaluated at the baseline. Nauta et al. [12] defined a more extensive list of axioms with 12 desired properties, such as completeness, compactness, and correctness. Wargnier-Dauchelle et al. [57] used IG on MRIs for Multiple Sclerosis classification and showed that using IG-based tissue probability maps instead of raw MRIs as model input led to more accurate classifiers. Lastly, a special technique to make gradient-based explanations less noisy, called **SmoothGrad**, was proposed by Smilkov et al. [58]. During this procedure, multiple versions of an input image are produced by adding noise to it. Subsequently, resulting pixel attribution maps are averaged. The fundamental idea is to "smooth out" fluctuations occurring in derivatives within neural networks (e.g., Vanilla Gradients heatmaps contain much noise caused by fluctuations in the output activation derivatives [59]). However, SmoothGrad is not a standalone XAI method but rather an extension to stabilize any gradient-based approach. The work of [60] deployed SmoothGrad to explain the DL-based classification of pancreatic tissue on histopathologic slides. Besides developments contributing to better performance in image recognition and localization, there are also approaches concerning architectural modifications for networks to become more interpretable.

So far, we have outlined methods for visualizing attribution maps of models. However, beyond these maps, CNNs, for example, have a stack of fully connected layers (on top of the last layer) converting feature maps to a pre-softmax score, which is not easily interpretable. As a countermeasure, Zhou et al. [61] proposed a procedure for generating so-called **Class Activation Mappings (CAMs)** by replacing the fully connected layers (except the last softmax-layer) with a Global Average Pooling (GAP) on each feature map of the (last) CNN layer. Consequently, a resulting class activation map denotes the direct impact of activation at a spatial point *(x, y)* regarding its corresponding class c. A disadvantage of CAM is its dependence on a particular architecture that includes GAP and one fully connected layer generating the prediction (CNNs). An extended version of CAM is **Gradient-weighted CAM (Grad-CAM)** [62] and uses gradients w.r.t. a target class c. More specifically, the GAP can be replaced with any differentiable neural network layers (needed for gradient calculation). Similar to CAM, the authors of Grad-CAM argue, "we can expect the last convolutional layers to have the best compromise between high-level semantics and detailed spatial information." [62]. The main difference between CAM and Grad-CAM is how feature maps are weighted to generate the final heatmap, whereas the pursued goal remains unchanged. In the work of Hosch et al. [63], GradCAMs identified that a model with outstanding accuracy incorrectly used image markings to classify a patient's position from chest X-Rays which rigorously compromised the model's integrity within the medical field. Such scenarios are called a *Clever Hans phenomenon* [64] or *shortcut learning* [65] and can often go unnoticed without analyzing the prediction process. After removing all image markings and re-training, the model focused on medically relevant reference areas such as the shoulders. Generally, Grad-CAM maps can be computed faster than the occlusion map. However, Grad-CAM maps usually have a lower spatial resolution than occlusion maps. Therefore, they can miss fine details since the underlying resolution of Grad-CAM matches the spatial resolution of the last convolutional feature map. Also, the occurrence of multiple objects from the same class can result in inconsistent coverage. Aiming at improvement, numerous extensions, such as **GradCAM++** [66], **Score-CAM** [67], or **Guided GradCAM** [62], have been developed. Opposed to attribution maps, as proposed by Das and Rad [42], *activation maximization* (a technique to maximally stimulate the activation of certain neurons) can be expanded to a global explanation method by using a so-called **Class Model Visualization** [43]. The main idea is to generate image visualizations that, in terms of activation maximization, are representative of a specific class or layer used by a network. This can be achieved in two manners: One way is to revert the network flow and guide it to alternate an image in such a way as to provide an interpretation for a target class. For example, when investigating which input would result in the network predicting the class "Banana", the model can be provided with a random noise image, which is then modified towards what the network predicts as a banana. Due to its generative nature, this method is also called **Deep Dream**. Another way is to feed the network an arbitrary image and let it enhance whatever was detected. Generally, this is done layer-wise because each layer comprises features at different abstraction levels and is known as Inceptionism [68]. Commonly, lower layers are sensitive to fundamental features such as edges or patterns, while higher layers are sensitive to more sophisticated features or even detected objects. However, even though such visualizations are insightful for natural images, they can be challenging to interpret in medical images, generally limiting the application of this method. Based on the concept of DeepDream, Couteaux et al. [69] investigated how segmentation networks can be interpreted in terms of feature sensitivity (robustness) or indifference during tumor segmentation in liver CT images. To achieve this, an image is forwarded in the network iteratively to retrieve a gradient belonging to an arbitrary neuron activation from the resulting output map. The image and segmentation features are determined to update the image following the gradient for the next iteration, allowing the evolution of "deep dreamed" features. This provides insights into the sensitivity and robustness of specific high-level features [69].

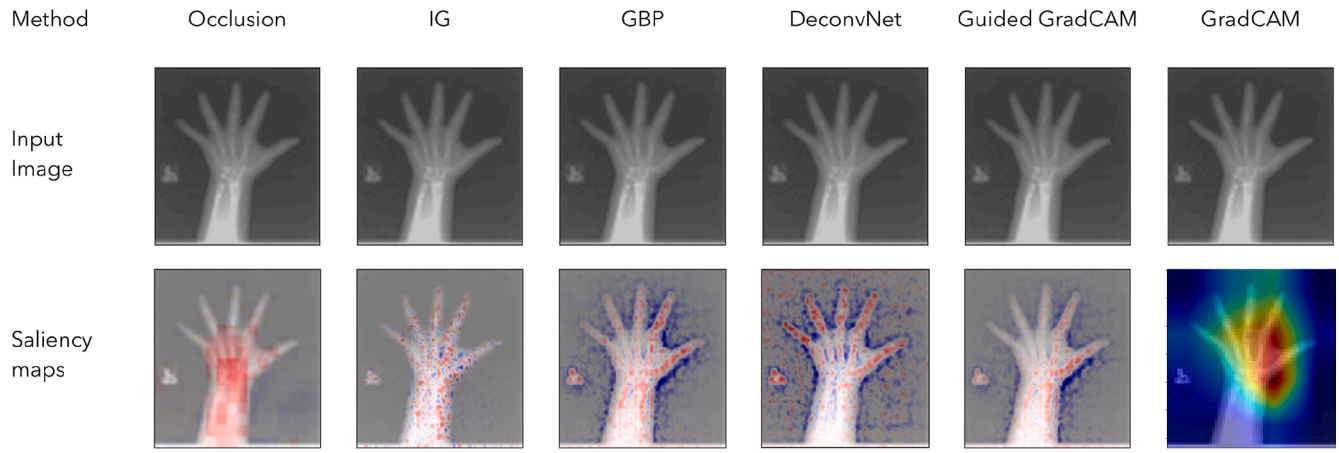

Fig. 7. Overview of a subset of backpropagation-based methods from captum [70] applied to DenseNet121 [71] trained on MedNIST [70,72] to classify images into the classes *'AbdomenCT'*, *'BreastMRI'*, *'CXR'*, *'ChestCT'*, *'Hand'*, and *'HeadCT'*. The displayed image was classified as *'Hand'* with a prediction score of 0.98. Red areas denote positive contribution, whereas blue areas indicate negative contribution.

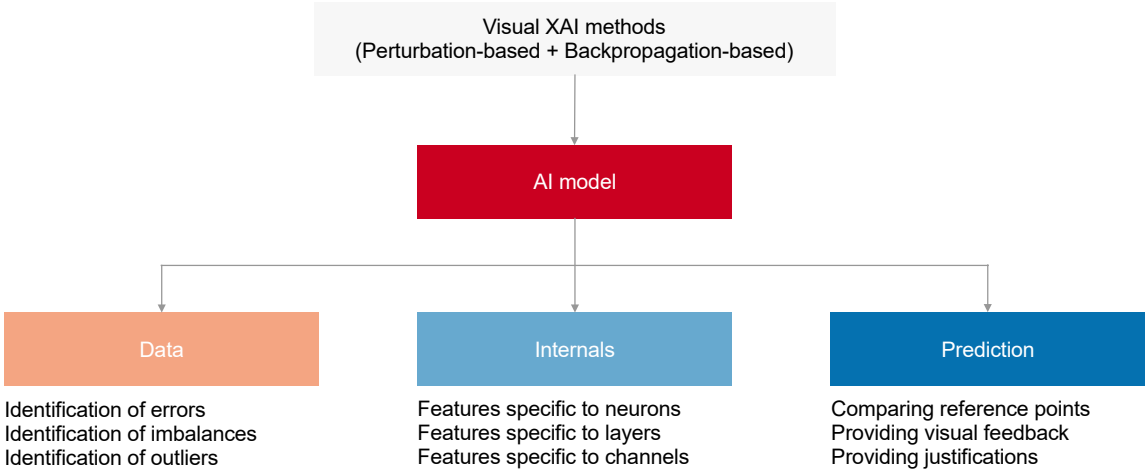

Fig. 8. A summarizing depiction of explanation targets pursued during the application of visual XAI methods. Applying visual XAI methods to an AI model enables the explanation of data, model internals, or model prediction.

3. Conclusion

Visual XAI methods enable the identification of potentials, limitations, and biases in the application of AI in medical imaging. For example, by comparing medically known reference points and the reference points used by a model during prediction. This helps to understand if the model has learned relevant features; otherwise, its integrity in the medical domain could be considered limited. Moreover, since a model is trained on data, visual XAI methods also allow assessing the data. For example, imbalances or flaws can be identified by applying activation maximization techniques either by constructing images from the noise that correspond to a specific class or by investigating what features a model's layer focuses on. Lastly, visual XAI methods enable predictions to be justified and contribute to the transparency of this process. Consequently, three explanatory goals pursued by visual XAI methods were identified, as depicted in Fig. 8. Generally, visual XAI methods seem to play a predominant role in medical imaging, with GradCAM currently being the most popular method (as presented in Fig. 6), presumably due to the ease of implementation and intuitiveness of interpretation. Even though the implementation of XAI in medical research is increasing, its usage in the clinical workflow does not reflect that trend. We reason that this is because no scientific recommendation guidelines are known yet, and satisfactory factors constituting an explanation and interpretation have yet to be agreed on by the XAI research community in cooperation with healthcare professionals. For example, in Fig. 7, several explanation methods for a model decision are presented, and it is evident that each method highlights slightly different image areas. Moreover, in each image, an X-Ray specific marking is captured (left to the hand), which could indicate a Clever Hans phenomenon, as such markings are not present in the other classes containing CTs and MRIs. Consequently, the question arises of which explanation best suits the given classification task. However, providing an answer to that question remains challenging. Intuitively, an interesting measure would be the accordance between a method's highlighted image areas and a radiologist's reference areas. However, inter- and intraobserver variability among healthcare professionals [73] impedes the direct usage of such measures. One important work in this area was presented by Xie et al. [74] in CheXplain, who conducted a physician-centered design to outline recommendations for human-centered medical AI development confining their results to four predominant categories: motivation vs. constraint and explanation vs. justification. Importantly, one mentioned limitation is that survey questions were mainly speculative due to the limited incorporation of AI in current medical practices. Moreover, even though XAI methods can contribute to explainability and interpretability, no solid scientific validation framework is known, reflected by a limited number of studies investigating the impact on clinical accuracy, relevance, and acceptance. Especially in medical imaging, validation often focuses on the accordance between a model prediction and the ground truth, concluding that a match indicates similar reasoning between the model and clinicians (reference points). However, little emphasis is laid on measuring the correspondence between the model's output and the results of an accompanying XAI method. Therefore, Venugopal et al. [75] define an *Explainability Failure* as a case where the classification generated by an AI algorithm matches with the study-level ground truth, but the explanation output is inadequate to justify the algorithm's output.

This is just one example of the complexity XAI evaluation can incur and the difficulty of comparing methods. A thorough investigation concerning the quantitative evaluation of XAI methods was proposed by the work of Nauta et al. [12], identifying 12 conceptual aspects, introduced as *Co-12 properties,* that serve as a categorization scheme for reviewing the evaluation practice. The *coherence* of an explanation w.r.t. human intuition is only one of them, as it has been shown that only relying on the plausibility of explanations can be misleading [12,76]. The authors consider interpretability a multi-faceted characteristic and state that quantitatively measuring interpretability should lead to a multi-dimensional assessment indicating to which extent specific properties are satisfied. Conclusively, in order to enable a successful interplay between AI and healthcare professionals in the future, domain-specific knowledge needs to be integrated into AI, and, in turn, healthcare professionals need to gain further knowledge of what XAI is and how methods from this research area can contribute to their handling of AI-based systems. Moreover, to enable the implementation of XAI into the medical imaging domain, the XAI and medical community has yet to agree on standardized approaches concerning implementation guidelines and validation methods. Consequently, this review attempted to fill that gap by identifying and presenting established and commonly used XAI methods for medical imaging alongside their practical impact, shortcomings, and limitations (Appendix B, Appendix C) in a non-technical manner to be comprehensive for a broader audience but specifically for healthcare professionals.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A

(See Fig. A1)

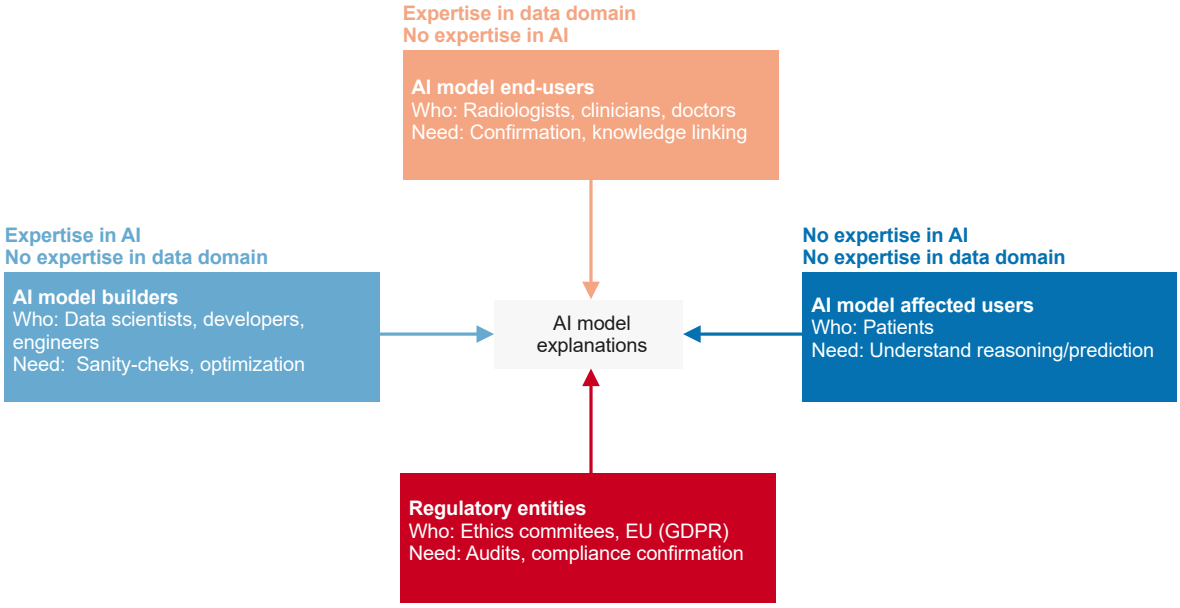

Figure showing different XAI stakeholder groups for AI models deployed in the medical domain, which depending on their domain knowledge and expertise, seek different explanation needs. The overview is partially inspired by [8].

Appendix B

(See Table B1)

Table B1
Listing of all included XAI methods ordered by a descending year-citation-ratio (according to google Scholar, see Supplementary File 2 in Appendix D) and divided by their explanation outcome. In addition, strengths, weaknesses, and open-source implementation are provided. Methods with a comparatively low number of citations are denoted with an asterisk.

Signal flow	Method	Strength	Weakness	Citations/year ratio
Perturbation-based	Occlusion [28]	No access to mode internals required, so model-agnostic	High computational cost Approximate results of object morphology	1,900
	SHAP [37]	Denotes positive and negative contributions	High computational expense	1,770
	LIME [31]	Denotes positive and negative contributions	Relies on pre-defined parameters High computational cost	1,597
	Meaningful Perturbations [33]	Perturbation mask generated based on optimization approach	Blur and noise naturally contained in medical images	209
	RISE [40]	Realistic perturbations	Approximate results of object morphology Relies on pre-defined parameters	123
	Real-time Perturbations [35]	Individual neural network generated perturbation mask	High implementational effort	77
	Extremal Perturbations [36]	Investigates the perturbation's effect w.r.t. its size	Blur and noise naturally contained in medical images	67
Backpropagation-based	Grad-CAM [62]	Intuitive, no explicit labeling of data needed	Multiple occurences of same class objects result in partial coverage	2,007
	DeconvNet [28]	Fast computation	Uses only an approximate inverse of the convolution features	1,900
	CAM [61]	Intuitive, no explicit labeling of data needed	Architectural dependence Inaccurate localization of heatmap	1,074
	Deep Dream [43]	Versatile range of implementations (neurons, layers, images, atlases)	Results can be challenging to interpret, especially for medical images	597
	Vanilla Gradient [43]	Straightforward and direct computation of gradients	Saturating gradients	597
	Integrated Gradients [56]	satisfies mathematical axioms regarding sensitivity and invariance	Results may differ, depending on selected baseline and number of iterations	572
	GBP [46]	Applicable to several network architectures	In some cases, invariant to data and model	559
	Deep LiFT [44]	Counters the saturation during Input*Gradient	Does not satisfy implementation invariance	446
	LRP [50]	Fast and scalable computation	Results are sensitive to specific parameters	396
	SmoothGrad [58]	Reduces visual noise	Results depend on hyper-parameters: σ, (the noise level) and n (number of samples to average over)	234
	Input * Gradient [54]	Simple process	Saturating gradients	78

Appendix C

(See Table C1)

Table C1
Listing of presented XAI methods in alphabetical order alongside a corresponding open-source repository link (if available).

Signal flow	Method	Open source library available	Surrogate	Model-agnostic
Perturbation-based	Extremal Perturbations [36]	https://github.com/facebookresearch/TorchRay	×	×
	LIME [31]	https://github.com/marcotcr/lime	✓	✓
	Meaningful Perturbations [33]	https://github.com/ruthcfong/perturb_explanations	×	✓
	Occlusion [28]	https://github.com/keiserlab/plaquebox-papers	×	✓
	Real-time Perturbations [35]	https://github.com/PiotrDabkowski/pytorch-saliency	×	✓
	RISE [40]	https://github.com/eclique/RISE	×	✓
	SHAP [37]	https://github.com/slundberg/shap	✓	✓
Backpropagatio-based	CAM [61]	https://github.com/zhoubolei/CAM	×	×
	DeconvNet [28]	https://github.com/pytorch/captum (PyTorch)	×	×
	Deep Dream [43]	https://github.com/pytorch/captum	×	×
	DeepLift [44]	https://github.com/kundajelab/deeplift	×	×
	Grad-CAM [62]	https://github.com/ramprs/grad-cam	×	×
	Input * Gradient [54]	https://github.com/pytorch/captum (PyTorch)	×	×
	Integrated Gradients [56]	https://github.com/ankurtaly/Integrated-Gradients	×	×
	LRP [50]	https://github.com/sebastian-lapuschkin/lrp_toolbox	×	×
	Vanilla Gradient [43]	https://github.com/pytorch/captum (PyTorch)	×	×
	SmoothGrad [58]	https://github.com/pair-code/saliency	×	×
	GBP [46]	https://github.com/pytorch/captum	×	×

Appendix D. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ejrad.2023.110787.

References

[1] S. Benjamens, P. Dhunnoo, B. Meskó, The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database, Art. no. 1, NPJ Digit. Med. 3 (1) (Sep. 2020), https://doi.org/10.1038/s41746-020-00324-0.

[2] S.M. McKinney, et al., International evaluation of an AI system for breast cancer screening, Art. no. 7788, Nature 577 (7788) (Jan. 2020), https://doi.org/10.1038/s41586-019-1799-6.

[3] A. Singh, S. Sengupta, V. Lakshminarayanan, Explainable Deep Learning Models in Medical Image Analysis, Art. no. 6, J. Imaging 6 (6) (Jun. 2020), https://doi.org/10.3390/jimaging6060052.

[4] W. Samek, G. Montavon, A. Binder, S. Lapuschkin, K.-R. Müller, Interpreting the Predictions of Complex ML Models by Layer-wise Relevance Propagation, *arXiv: 1611.08191 [cs, stat]*, Nov. 2016, Accessed: Apr. 07, 2022. [Online]. Available: http://arxiv.org/abs/1611.08191.

[5] W. Fedus, B. Zoph, N. Shazeer, Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity, J. Mach. Learn. Res. 23 (120) (2022) 1–39.

[6] B.H.M. van der Velden, H.J. Kuijf, K.G.A. Gilhuijs, M.A. Viergever, Explainable artificial intelligence (XAI) in deep learning-based medical image analysis, Med. Image Anal. 79 (Jul. 2022), 102470, https://doi.org/10.1016/j.media.2022.102470.

[7] A.M. Antoniadi, et al., Current Challenges and Future Opportunities for XAI in Machine Learning-Based Clinical Decision Support Systems: A Systematic Review, Art. no. 11, Appl. Sci. 11 (11) (Jan. 2021), https://doi.org/10.3390/app11115088.

[8] A. Barredo Arrieta, et al., Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, Inform. Fusion 58 (Jun. 2020) 82–115, https://doi.org/10.1016/j.inffus.2019.12.012.

[9] F. Doshi-Velez, B. Kim, Towards A Rigorous Science of Interpretable Machine Learning, *arXiv:1702.08608 [cs, stat]*, Mar. 2017, Accessed: Apr. 07, 2022. [Online]. Available: http://arxiv.org/abs/1702.08608.

[10] R. Tomsett, D. Braines, D. Harborne, A. Preece, S. Chakraborty, Interpretable to Whom? A Role-based Model for Analyzing Interpretable Machine Learning Systems. arXiv, Jun. 20, 2018. doi: 10.48550/arXiv.1806.07552.

[11] L. Gilpin, D. Bau, B. Yuan, A. Bajwa, M. Specter, L. Kagal, "Explaining Explanations: An Overview of Interpretability of Machine Learning," Oct. 2018, pp. 80–89. doi: 10.1109/DSAA.2018.00018.

[12] M. Nauta *et al.*, From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI. arXiv, May 31, 2022. Accessed: Jun. 20, 2022. [Online]. Available: http://arxiv.org/abs/2201.08164.

[13] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT Press, 2016.

[14] Y. Lecun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proc. IEEE 86 (11) (Nov. 1998) 2278–2324, https://doi.org/10.1109/5.726791.

[15] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet Classification with Deep Convolutional Neural Networks, in: *Advances in Neural Information Processing Systems*, 2012, vol. 25. Accessed: Jul. 05, 2022. [Online]. Available: https://papers.nips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html.

[16] Z.C. Lipton, In machine learning, the concept of interpretability is both important and slippery, *Machine Learning*, p. 28.

[17] M. Bellucci, N. Delestre, N. Malandain, C. Zanni-Merk, Towards a terminology for a fully contextualized XAI, Proc. Comput. Sci. 192 (Jan. 2021) 241–250, https://doi.org/10.1016/j.procs.2021.08.025.

[18] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, Artif. Intell. 267 (Feb. 2019) 1–38, https://doi.org/10.1016/j.artint.2018.07.007.

[19] F. Chollet, Deep Learning with Python, second ed., Simon and Schuster, 2021.

[20] M. Ancona, E. Ceolini, C. Öztireli, M. Gross, Towards better understanding of gradient-based attribution methods for Deep Neural Networks, in *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, Toulon, France, 2018. [Online]. Available: https://openreview.net/forum?id=Sy21R9JAW.

[21] M. Pocevičiūtė, G. Eilertsen, C. Lundström, Survey of XAI in Digital Pathology, in: A. Holzinger, R. Goebel, M. Mengel, H. Müller (Eds.), Artificial Intelligence and Machine Learning for Digital Pathology: State-of-the-Art and Future Challenges, Springer International Publishing, Cham, 2020, pp. 56–88, https://doi.org/10.1007/978-3-030-50402-1_4.

[22] M. Ouzzani, H. Hammady, Z. Fedorowicz, A. Elmagarmid, Rayyan—a web and mobile app for systematic reviews, Syst. Rev. 5 (1) (2016) 210, https://doi.org/10.1186/s13643-016-0384-4.

[23] A. Agrawal, et al., VQA: Visual Question Answering, Int. J. Comput. Vision 123 (1) (May 2017) 4–31, https://doi.org/10.1007/s11263-016-0966-6.

[24] Y. Dong, H. Su, J. Zhu, B. Zhang, Improving Interpretability of Deep Neural Networks With Semantic Information, presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4306–4314. Accessed: Apr. 07, 2022. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2017/html/Dong_Improving_Interpretability_of_CVPR_2017_paper.html.

[25] O. Vinyals, A. Toshev, S. Bengio, D. Erhan, Show and Tell: A Neural Image Caption Generator, presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3156–3164. Accessed: Dec. 12, 2022. [Online]. Available: https://www.cv-foundation.org/openaccess/content_cvpr_2015/html/Vinyals_Show_and_Tell_2015_CVPR_paper.html.

[26] A. Ghorbani, J. Wexler, J. Y. Zou, B. Kim, Towards Automatic Concept-based Explanations, in *Advances in Neural Information Processing Systems*, 2019, vol. 32. Accessed: Jun. 03, 2022. [Online]. Available: https://proceedings.neurips.cc/paper/2019/hash/77d2afcb31f6493e350fca61764efb9a-Abstract.html.

[27] P. W. Koh, P. Liang, "Understanding Black-box Predictions via Influence Functions," in: *Proceedings of the 34th International Conference on Machine Learning*, Jul. 2017, pp. 1885–1894. Accessed: Sep. 20, 2022. [Online]. Available: https://proceedings.mlr.press/v70/koh17a.html.

[28] M.D. Zeiler, R. Fergus, "Visualizing and Understanding Convolutional Networks", Computer Vision – ECCV Cham 2014 (2014) 818–833, https://doi.org/10.1007/978-3-319-10590-1_53.

[29] Z. Tang, et al., Interpretable classification of Alzheimer's disease pathologies with a convolutional neural network pipeline, Nat Commun 10 (1) (Dec. 2019) 2173, https://doi.org/10.1038/s41467-019-10212-1.

[30] I. Kakogeorgiou, K. Karantzalos, Evaluating explainable artificial intelligence methods for multi-label deep learning classification tasks in remote sensing, Int. J. Appl. Earth Obs. Geoinf. 103 (Dec. 2021), 102520, https://doi.org/10.1016/j.jag.2021.102520.

[31] M. T. Ribeiro, S. Singh, C. Guestrin, "'Why Should I Trust You?': Explaining the Predictions of Any Classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, Aug. 2016, pp. 1135–1144. doi: 10.1145/2939672.2939778.

[32] P.R. Magesh, R.D. Myloth, R.J. Tom, An Explainable Machine Learning Model for Early Detection of Parkinson's Disease using LIME on DaTSCAN Imagery, Comput. Biol. Med. 126 (Nov. 2020), 104041, https://doi.org/10.1016/j.compbiomed.2020.104041.

[33] R. C. Fong, A. Vedaldi, "Interpretable Explanations of Black Boxes by Meaningful Perturbation," presented at the Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 3429–3437. Accessed: Jul. 28, 2022. [Online]. Available: https://openaccess.thecvf.com/content_iccv_2017/html/Fong_Interpretable_Explanations_of_ICCV_2017_paper.html.

[34] H. Uzunova, J. Ehrhardt, T. Kepp, H. Handels, "Interpretable explanations of black box classifiers applied on medical images by meaningful perturbations using variational autoencoders," in *Medical Imaging 2019: Image Processing*, Mar. 2019, vol. 10949, pp. 264–271. doi: 10.1117/12.2511964.

[35] P. Dabkowski, Y. Gal, "Real Time Image Saliency for Black Box Classifiers," in *Advances in Neural Information Processing Systems*, 2017, vol. 30. Accessed: Jul. 28, 2022. [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/0060ef47b12160b9198302ebdb144dcf-Abstract.html.

[36] R. Fong, M. Patrick, A. Vedaldi, "Understanding Deep Networks via Extremal Perturbations and Smooth Masks," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2019, pp. 2950–2958. doi: 10.1109/ICCV.2019.00304.

[37] S. M. Lundberg, S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems*, 2017, vol. 30. Accessed: Apr. 14, 2022. [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.

[38] C. Molnar, *Interpretable Machine Learning*. Accessed: Apr. 12, 2022. [Online]. Available: https://christophm.github.io/interpretable-ml-book/.

[39] K. Young, G. Booth, B. Simpson, R. Dutton, S. Shrapnel, "Deep Neural Network or Dermatologist?," in *Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support*, Cham, 2019, pp. 48–55. doi: 10.1007/978-3-030-33850-3_6.

[40] V. Petsiuk, A. Das, and K. Saenko, "RISE: Randomized Input Sampling for Explanation of Black-box Models.," in *British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018*, 2018, p. 151. [Online]. Available: http://bmvc2018.org/contents/papers/1064.pdf.

[41] J. Cooper, O. Arandjelović, D.J. Harrison, Believe the HiPe: Hierarchical perturbation for fast, robust, and model-agnostic saliency mapping, Pattern Recogn. 129 (Sep. 2022), 108743, https://doi.org/10.1016/j.patcog.2022.108743.

[42] A. Das and P. Rad, "Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey.," *CoRR*, vol. abs/2006.11371, 2020, [Online]. Available: https://arxiv.org/abs/2006.11371.

[43] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps," presented at the 2nd International Conference on Learning Representations, {ICLR} 2014, Banff, AB, Canada, April 14-16, 2014, Workshop Track Proceedings, 2014.

[44] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, Sydney, NSW, Australia, Aug. 2017, pp. 3145–3153.

[45] B.D. de Vos, J.M. Wolterink, T. Leiner, P.A. de Jong, N. Lessmann, I. Išgum, Direct Automatic Coronary Calcium Scoring in Cardiac and Chest CT, IEEE Trans. Med. Imaging 38 (9) (Sep. 2019) 2127–2138, https://doi.org/10.1109/TMI.2019.2899534.

[46] J. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for Simplicity: The All Convolutional Net," presented at the ICLR (workshop track), 2015. Accessed: Jan. 06, 2023. [Online]. Available: https://lmb.informatik.uni-freiburg.de/Publications/2015/DB15a/.

[47] S. Mostafa, D. Mondal, M. A. Beck, C. P. Bidinosti, C. J. Henry, and I. Stavness, "Leveraging Guided Backpropagation to Select Convolutional Neural Networks for Plant Classification," *Frontiers in Artificial Intelligence*, vol. 5, 2022, Accessed: Jan. 05, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/frai.2022.871162.

[48] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, Dezember 2018, pp. 9525–9536.

[49] F. Dubost, et al., Enlarged perivascular spaces in brain MRI: Automated quantification in four regions, Neuroimage 185 (Jan. 2019) 534–544, https://doi.org/10.1016/j.neuroimage.2018.10.026.

[50] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, W. Samek, On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation, PLoS One 10 (7) (Oct. 2015) e0130140.

[51] M. Böhle, F. Eitel, M. Weygandt, and K. Ritter, "Layer-Wise Relevance Propagation for Explaining Deep Neural Network Decisions in MRI-Based Alzheimer's Disease Classification," *Frontiers in Aging Neuroscience*, vol. 11, 2019, Accessed: Apr. 07, 2022. [Online]. Available: https://www.frontiersin.org/article/10.3389/fnagi.2019.00194.

[52] L.A. de Souza, et al., Convolutional Neural Networks for the evaluation of cancer in Barrett's esophagus: Explainable AI to lighten up the black-box, Comput. Biol. Med. 135 (Aug. 2021), 104578, https://doi.org/10.1016/j.compbiomed.2021.104578.

[53] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, Sydney, NSW, Australia, Aug. 2017, pp. 3319–3328.

[54] A. Shrikumar, P. Greenside, A. Shcherbina, and A. Kundaje, "Not Just a Black Box: Learning Important Features Through Propagating Activation Differences." arXiv, Apr. 11, 2017. doi: 10.48550/arXiv.1605.01713.

[55] E. Prakash, A. Shrikumar, and A. Kundaje, "Towards More Realistic Simulated Datasets for Benchmarking Deep Learning Models in Regulatory Genomics.," in *Machine Learning in Computational Biology Meeting, MLCB 2021, online, November 22-23, 2021.*, 2021, pp. 58–77. [Online]. Available: https://proceedings.mlr.press/v165/prakash22a.html.

[56] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic Attribution for Deep Networks." arXiv, Jun. 12, 2017. doi: 10.48550/arXiv.1703.01365.

[57] V. Wargnier-Dauchelle, T. Grenier, F. Durand-Dubief, F. Cotton, and M. Sdika, "A More Interpretable Classifier For Multiple Sclerosis," in *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, Apr. 2021, pp. 1062–1066. doi: 10.1109/ISBI48211.2021.9434074.

[58] D. Smilkov, N. Thorat, B. Kim, F. B. Viégas, and M. Wattenberg, "SmoothGrad: removing noise by adding noise.," *CoRR*, vol. abs/1706.03825, 2017, [Online]. Available: http://arxiv.org/abs/1706.03825.

[59] I. Palatnik de Sousa, M. M. B. R. Vellasco, and E. Costa da Silva, "Explainable Artificial Intelligence for Bias Detection in COVID CT-Scan Classifiers," *Sensors*, vol. 21, no. 16, Art. no. 16, Jan. 2021, doi: 10.3390/s21165657.

[60] M. Kriegsmann, K. Kriegsmann, G. Steinbuss, C. Zgorzelski, A. Kraft, M.M. Gaida, Deep Learning in Pancreatic Tissue: Identification of Anatomical Structures, Pancreatic Intraepithelial Neoplasia, and Ductal Adenocarcinoma, Art. no. 10, Int. J. Mol. Sci. 22 (10) (Jan. 2021), https://doi.org/10.3390/ijms22105385.

[61] B. Zhou, A. Khosla, À. Lapedriza, A. Oliva, A. Torralba, "Learning Deep Features for Discriminative Localization," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2921–2929, 2016, doi: 10.1109/CVPR.2016.319.

[62] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, "Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization," presented at the Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 618–626. Accessed: Apr. 07, 2022. [Online]. Available: https://openaccess.thecvf.com/content_iccv_2017/html/Selvaraju_Grad-CAM_Visual_Explanations_ICCV_2017_paper.html.

[63] R. Hosch, L. Kroll, F. Nensa, S. Koitka, Differentiation Between Anteroposterior and Posteroanterior Chest X-Ray View Position With Convolutional Neural Networks, Rofo 193 (2) (Feb. 2021) 168–176, https://doi.org/10.1055/a-1183-5227.

[64] S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, K.-R. Müller, Unmasking Clever Hans predictors and assessing what machines really learn, Art. no. 1, Nat Commun 10 (1) (Mar. 2019), https://doi.org/10.1038/s41467-019-08987-4.

[65] R. Geirhos, et al., Shortcut learning in deep neural networks, Art. no. 11, Nat Mach Intell 2 (11) (Nov. 2020), https://doi.org/10.1038/s42256-020-00257-z.

[66] A. Chattopadhyay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM ++: Improved Visual Explanations for Deep Convolutional Networks," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Mar. 2018, pp. 839–847. doi: 10.1109/WACV.2018.00097.

[67] H. Wang *et al.*, "Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Seattle, WA, USA, Jun. 2020, pp. 111–119. doi: 10.1109/CVPRW50498.2020.00020.

[68] "Inceptionism: Going Deeper into Neural Networks," *Google AI Blog*, 2015. https://ai.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html (accessed Jun. 16, 2022).

[69] V. Couteaux, O. Nempont, G. Pizaine, and I. Bloch, "Towards Interpretability of Segmentation Networks by Analyzing DeepDreams," in *Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support*, Cham, 2019, pp. 56–63. doi: 10.1007/978-3-030-33850-3_7.

[70] N. Kokhlikyan *et al.*, "Captum: A unified and generic model interpretability library for PyTorch," *arXiv [cs.LG]*, 2020, [Online]. Available: http://arxiv.org/abs/2009.07896.

[71] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, pp. 2261–2269. doi: 10.1109/CVPR.2017.243.

[72] J. Yang, R. Shi, and B. Ni, "MedMNIST Classification Decathlon: A Lightweight AutoML Benchmark for Medical Image Analysis.," in *18th IEEE International Symposium on Biomedical Imaging, ISBI 2021, Nice, France, April 13-16, 2021*, 2021, pp. 191–195. doi: 10.1109/ISBI48211.2021.9434062.

[73] H. Al-Khawari, R.P. Athyal, O. Al-Saeed, P.N. Sada, S. Al-Muthairi, A. Al-Awadhi, Inter- and intraobserver variation between radiologists in the detection of abnormal parenchymal lung changes on high-resolution computed tomography, Ann Saudi Med 30 (2) (2010) 129–133, https://doi.org/10.4103/0256-4947.60518.

[74] Y. Xie, M. Chen, D. Kao, G. Gao, and X. "Anthony" Chen, "CheXplain: Enabling Physicians to Explore and Understand Data-Driven, AI-Enabled Medical Imaging Analysis," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing*

Systems, New York, NY, USA, Apr. 2020, pp. 1–13. doi: 10.1145/3313831.3376807.

[75] V.K. Venugopal, R. Takhar, S. Gupta, V. Mahajan, Clinical Explainability Failure (CEF) & Explainability Failure Ratio (EFR) – Changing the Way We Validate Classification Algorithms, J Med Syst 46 (4) (Mar. 2022) 20, https://doi.org/10.1007/s10916-022-01806-2.

[76] A. Jacovi and Y. Goldberg, "Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness?," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, Jul. 2020, pp. 4198–4205. doi: 10.18653/v1/2020.acl-main.386.