European Journal of Radiology 162 (2023) 110786

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 – Beyond 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]

[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

Driven by recent advances in *Artificial Intelligence* (AI) and *Computer Vision* (CV), the implementation of AI systems in the medical domain increased correspondingly. This is especially true for the domain of medical imaging, in which the incorporation of AI aids several imaging-based tasks such as classification, segmentation, and registration. Moreover, AI reshapes medical research and contributes to the development of personalized clinical care. Consequently, alongside its extended implementation arises the need for an extensive understanding of AI systems and their inner workings, potentials, and limitations which the field of *eXplainable AI* (XAI) aims at. Because medical imaging is mainly associated with visual tasks, most explainability approaches incorporate *saliency-based* XAI methods. In contrast to that, in this article we would like to investigate the full potential of XAI methods in the field of medical imaging by specifically focusing on XAI techniques not relying on saliency, and providing diversified examples. We dedicate our investigation to a broad audience, but particularly healthcare professionals. Moreover, this work aims at establishing a common ground for cross-disciplinary understanding and exchange across disciplines between Deep Learning (DL) builders and healthcare professionals, which is why we aimed for a non-technical overview. Presented XAI methods are divided by a method's output representation into the following categories: Case-based explanations, textual explanations, and auxiliary explanations.

1. Introduction

In recent years, advances in *Deep Learning* (DL) coupled with increasing computational power and innovative algorithms have promoted the development of DL-based systems throughout several medical areas. Medical Imaging is one prominent example, referring to several different technologies used to view the human body to diagnose, monitor, prevent, or treat a medical condition. Each modality, such as *Computed Tomography* (CT), *Magnetic Resonance Imaging* (MRI), or *X-Ray,* provides different information about the studied body area regarding possible diseases, injuries, or the effectiveness of medical treatments. Assessment of this information often relies on performing medical imaging tasks, including classification, detection, segmentation, and registration. Ongoing research at the intersection of Artificial Intelligence (AI) and medical imaging focuses on facilitating possibly time-consuming and inconvenient medical tasks by upbringing feasible and appropriate DL-based solutions [1–3]. Although such systems have been shown to outperform humans in certain analytical tasks, their utilization continues to spark concerns and remains limited, mainly due to the lack of *explainability.* Consequently, DL systems are often categorized as *Black Box* techniques. It is helpful to understand the core idea behind DL *models* to justify this categorization. Put simply, a DL model (also called a neural network) can be described as a mathematical *parameterized* entity that can process a specific input (e.g., an X-Ray image) and

[*] 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).

https://doi.org/10.1016/j.ejrad.2023.110786
Received 10 January 2023; Received in revised form 3 March 2023; Accepted 14 March 2023
Available online 20 March 2023

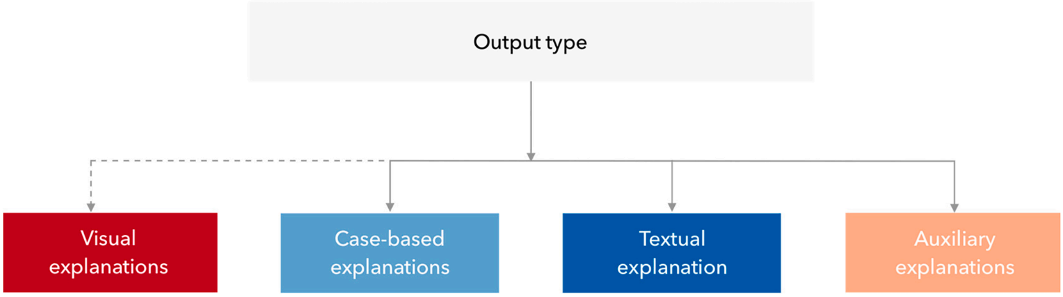

Fig. 1. Distinction of explanatory approaches by resulting output presentation form. Non-visual XAI approaches encompass auxiliary, case-based, and textual explanations. Visual explanations are not considered in this work but have been listed for completeness.

generate a corresponding output (e.g., a disease prediction). To capture the relation between a set of inputs (termed *training data*) and the desired outputs (termed *labels*), an elaborate and complex *model training* is required, during which a mapping function is approximated by iteratively estimating appropriate parameter values (parametrization). This estimation is calculated with a *loss function* which denotes how well a trained model fits a featured data set. With simplified input data like two-dimensional coordinates, the final mapping could be as easily depicted as a quadratic equation with few parameters. However, as medical imaging data is highly dimensional, mapping functions can become enormously complex, operating in the trillion parameter range [4]. Then such networks impede a direct interpretation of their predictions, mainly because of the inherent unpredictability of mapping functions coupled with high structural complexity. Consequently, the demand for interpretability and explainability of AI has experienced a tremendous resurgence over recent years, promoting the formation of new research fields, mainly known as eXplainable AI (XAI) [5,6]. Generally, XAI refers to all methods and approaches enabling human users to comprehend AI models. Since there is neither a mathematical nor a standardized definition of explainability and interpretability, in this paper, a non-mathematical definition from the social sciences perspective will be used as given by Miller [7]: “Interpretability is the degree to which a person can understand the cause *of a decision*”. Consequently, while interpretability focuses on the reasonings behind resulting outputs and helps uncover cause-and-effect relationships, *explainability* differs in the sense that it is associated with a system’s internal logic and procedures [5]. However, XAI is not a purely technological issue; instead, it relies on various medical, legal, ethical, and technological aspects that require thorough investigation. For instance, from the development point-of-view, explainability can be helpful to sanity-check DL models beyond mere performance and to identify severe errors before deploying tools into clinical validation or utilization. From the medical perspective, all systems – whether AI-powered or not – are incumbent to a rigorous validation process and medical certification [8]. However, random errors, systematic errors, and biases impede the development of DL tools with 100% diagnostic accuracy. And if bias is present, there will be prediction errors for inputs deviating from training data. Conclusively, random and systematic errors will occur in the clinical setting, even with a fully validated high-performing DL model. Even though this cannot be avoided, XAI can help to uncover such cases by providing a global (whole model) or a local (single prediction) model explanation and simultaneously safe-checking predictions that might be out-of-distribution. Regarding legal aspects, sensitive data-related issues such as privacy and security, patient consent, and anonymization also play an important role. For example, a prominent regulation in the European Union called *General Data Protection Regulation* enforces the right of patients to receive transparent information about a decision’s origin and requires the inclusion of XAI [9]. Consequently, the legal implications of establishing AI within healthcare are important, and the ongoing debate between innovation and regulation needs careful consideration [10].

To achieve a successful interplay between healthcare professionals and XAI, it is not only important to specifically tailor DL systems for the healthcare sector but also to introduce XAI as a powerful technique to healthcare professionals and provide a non-technical introduction to how XAI methods can help them handle novel DL systems. This review aims to fill that gap by providing a non-technical overview of common non-visual XAI methods that apply to medical imaging, along with their advantages, pitfalls, and limitations. While common saliency-based methods like GradCAM [11] project their explanations directly onto an input image, non-saliency-based methods usually provide more diversified explanations, e.g., by generating plots [12], textual descriptions [1] or confidence scores [13]. For ease of reading, we will label saliency-based methods as “visual” and non-saliency-based methods as “non-visual”. The contributions of this paper can be summarized as follows:

- Collection and categorization of non-visual XAI methods into distinctive categories defined by an XAI method’s expected output.
- For each category, a summarization of the methods’ functioning in a non-technical manner.
- Presentation of limitations, pitfalls, and potentials of the introduced XAI methods regarding implementation, evaluation, and interpretation (see Table 1 and Appendix A).
- Summarization of XAI’s current state of research in the field of medical imaging and outlining of future directions.

2. Non-Visual XAI methods in medical imaging

Since medical imaging is mainly associated with visual tasks, most explainability approaches incorporate *visual* XAI methods, including attribution and heat maps [9]. Even though visual XAI methods are considered easy to interpret and intuitive, some studies pointed out significant limitations. A major study by Adebayo et al. [14] investigated whether saliency methods are insensitive. This highly unwanted effect indicates that an explanation is unrelated to the model or data and does not explain anything. One example encompasses edge detectors because they just highlight areas with strong color changes within images and do not have a relation to a model prediction. Moreover, Ghorbani et al. [15] demonstrated saliency maps’ vulnerability to image perturbations and fragility to adversarial attacks leading to the question of how the *robustness* of visual XAI methods can be ensured. In contrast, Tomsett et al. [16] pointed out a lack of consistency concerning evaluation metrics in sanity-check studies concluding that despite the increasing effort, it remains challenging to evaluate visual explanations fully. Moreover, visual XAI methods represent only a minor subset of all possible methods. Therefore, even though non-visual XAI techniques might seem more specific and, in some cases, rely on method-specific knowledge for a correct interpretation, it is desirable to investigate the whole range of XAI methods against the background of exploring the full potential of XAI in the medical imaging domain. Consequently, apart from visual explanations, three subgroups of non-visual XAI methods

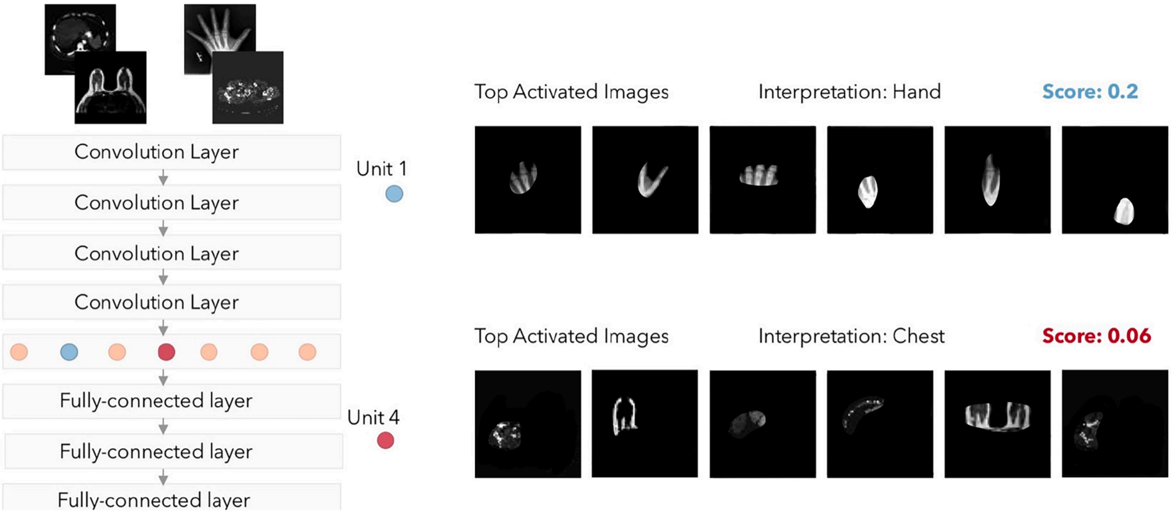

Fig. 2. Depiction on top-activated images for two exemplary units of a CNN based on [26] using the MedNIST dataset [27,28]. The top row represents activations for hands images, whereas the row below contains images of breasts and the abdomen. The visualizations serve primarily to clarify the procedure. The focus is on evaluating individual concepts concerning specific units, which is why this approach is listed among case-based methods.

were identified according to the type of result a user can expect when applying an XAI technique [17]: auxiliary explanations, case-based explanations, and textual explanations. A thorough introduction to these groups will be provided in subsequent sections. Importantly, they can be used as a classification scheme for XAI methods by their outcome, as depicted in Fig. 1. Deciding which type of informational representation is suitable for a given application is strongly specific to the end-users, technical circumstances, the desired explainability outcome, and - most importantly - what is perceived as beneficial concerning interpretability in the context of the medical domain. In this paper, we primarily refer to end-users as radiologists, clinicians, and doctors whose main need presumably might be the linking of knowledge and confirmation of diagnosis. For example, a radiologist applying a tumor detection model could possibly be more interested in receiving visual explanations on the original image to analyze the model's focus and how much trust can be put into its predictions. In contrast, a model that estimates a patient's overall survival chances might benefit from using *counterfactuals*, which are a case-based XAI method enabling "What-if"-assumptions and could, for example, yield how a diagnosis changed if specific clinical parameters were adjusted. However, during the development of such models, synthetic visualizations and auxiliary values may be more insightful [17]. Conclusively, each method has a unique potential to reveal helpful explanatory insight, and there are no recommendations or restrictions for selecting appropriate methods. Because of that, selecting an appropriate XAI method for the task at hand remains a challenge, especially considering the subjective perceptions of end-users alongside their individual needs [18,19] and relatively sparse assessment guidelines.

Recently, in [20], an extended investigation of real-world user needs for understanding AI was performed using an algorithm-informed XAI question bank. This work aims to clarify how end-user requirements should be understood, prioritized, and addressed to provide specific criteria. Miller [7] extensively examined the question of beneficial explanations, arguing that several XAI methods only consider the researcher's intuition, not including social aspects of how humans define, generate, select, evaluate, and present explanations. Miller argues that a good explanation is contrastive, selective, and truthful. On the one hand, it is advisable to know why a particular prediction was made instead of another (contrastive) because humans tend to devalue explanations that contradict their prior beliefs. On the other hand, an explanation of unexpected results by including several attributes can become overwhelming (selective). Nevertheless, such explanations must be as adequate as possible (truthful), which settles a trade-off between given requirements [7]. These challenges are commonly a question of finding the right balance for corresponding circumstances, the task at hand, the selected model type, and involved end-users and remain an essential aspect of ongoing research. Conclusively, several questions and challenges concerning XAI are most likely to be investigated in the context of a multi-disciplinary intersection of social and ethical sciences, end-users, and DL practitioners [7,21].

2.1. *Case-based explanations*

Case-based explanations may be highly diversified but share the same goal of providing insight based on specific examples, such as using similar input images, data samples, or counterfactuals to enable "What if"-assumptions [22]. Compared to visual and auxiliary explanations, this explanatory method is less explored. However, some research asserts that this approach is the most intuitive for human users to comprehend [7]. Furthermore, case-based explanations commonly aim at sample-based reasoning and uphold the potential of patient-individual explanation approaches within the medical domain. One prominent example of case-based methods is ***Testing with Concept Activation Vectors*** (**TCAV**) [23]. TCAV determines a model's sensitivity to an underlying high-level concept for a given class by training a linear classifier to separate the images containing the defined concept from those that do not. On the resulting hyperplane, TCAV utilizes directional derivatives to estimate the degree to which a defined concept is vital to a classification result. An exemplary intuitive question could be how sensitive a prediction of a zebra is to the presence of the concept "stripes". In this sense, a Concept Activation Vector (CAV) can be understood as a numerical representation generalizing a concept in the activation space of a neural network's specific layer, also called the *bottleneck*. For calculating a concept's CAV, two separate datasets must be generated: A concept dataset and a random dataset representing arbitrary data. For example, to define the concept "stripes", images of striped objects can be collected, whereas the arbitrary dataset can be a group of random images without stripes. A final quantitative explanation, also called the *TCAVQ* measure, represents the relative importance of each concept across all prediction classes, allowing for a global interpretation. One pitfall to be aware of when using the TCAV technique is the possibility of obtaining meaningless CAVs. A randomly selected set of images still produces a CAV; hence a significance test is recommended. Additionally, assembling and acquiring appropriate concepts often involve human supervision, which can be inconvenient for datasets without already labeled data.

Automated Concept-based Explanations **(ACE)** [24] extend TCAV in the very first step by providing an automated form of concept learning. Instead of requiring segmented images with pre-defined concepts, images of the same class are first cut into multiple segments by a semantic segmentation algorithm. This process is repeated at multiple resolutions to capture the complete hierarchy of possible concepts within an image (coarse-to-fine). The concepts are then represented as grouped pixels (segments). The resulting segments are passed through the model to generate activation space representations. Cohorts of similar segments are grouped using the euclidean distance as a similarity measure in the activation space of a pre-trained CNN. In contrast, outliers with a deficient similarity measure are removed. Finally, a TCAV concept-based importance score is calculated for each concept to retrieve the most significant ones regarding the classification task. As previously introduced, the idea behind the TCAV score is to estimate a concept's average positive effect on predicting a specific class. A significant advantage of ACE, compared to TCAV, is that it mitigates the need for human supervision, as manual labeling is not required. However, the method relies on the manual setting of parameters, which can affect the outcome, e.g., resulting in duplicated or mixed concepts [25].

A somewhat different approach called ***Network Dissection*** [26] is a general framework designed to quantify the interpretability of individual units (e.g., neurons or channels) within a neural network. The main idea is to measure the alignment between unit response and a set of concepts, assuming each network unit acts as a feature detector. The original work drew these concepts from a broad and dense segmentation data set called *Broden*. During the procedure of network dissection, the network is given images alongside pixel-wise segmentations as input, and for each unit, the top activated images are selected based on a corresponding activation map. This activation map is interpolated to match the resolution of the input image and subsequently thresholded, selecting all regions which exceed the threshold, simulating a binary segmentation as visualized in Fig. 2.

The resulting regions can be evaluated against the ground-truth segmentations from the Broden dataset by calculating the Intersection Over Union, which serves as a confidence score. The pixel-wise segmentations denote either low-level features such as patterns and structures or high-level features like objects-parts and scenes. Then, each unit is associated with the most confident label. By doing so, it can be analyzed which concepts are detected by which units. Fig. 2 shows that some of these concepts are mainly detected by one unit, while others are shared. With this technique, the authors confirmed that representations at different layers disentangle distinct categories of meaning and that different training techniques can significantly impact the representation learned by hidden units. As proposed by the authors of [26], results can be summarized in a Dissection report, e.g., containing the number of unique detectors within a network or histograms denoting which specific attributes the detectors represent. Furthermore, it is possible to detect concepts beyond the classes defined in the classification task if a pixel-wise labeled concept dataset is provided. However, channels are often not fully disentangled and can not be interpreted in isolation, so the interpretation ability of units activating to certain concepts should not be over-interpreted. Additionally, Network Dissection aligns human concepts with positive activations only. Negative activations of channels or activations below the arbitrarily defined threshold could also be significant [29].

Apart from concepts, another set of techniques aims at identifying *influential instances*. A prominent example is **Influence Functions (IF)** [30] proposed by Koh and Liang which approximate this during upweighting a specific sample by an infinitesimally small amount and measuring the change in loss and parameters. A mandatory condition is for the loss function to be 2nd-order differentiable. Wang et al. used IF to uncover relevant features of liver lesions on multiphase MRI by estimating the probability of the correct lesion classification deviating when altering specific learned features [31]. Conclusively, IF allow for understanding model behavior and detecting outliers in training data. A limitation is that IF only apply to models with a 2nd-order differentiable loss. Moreover, there is no agreed-upon threshold separating influential from non-influential instances.

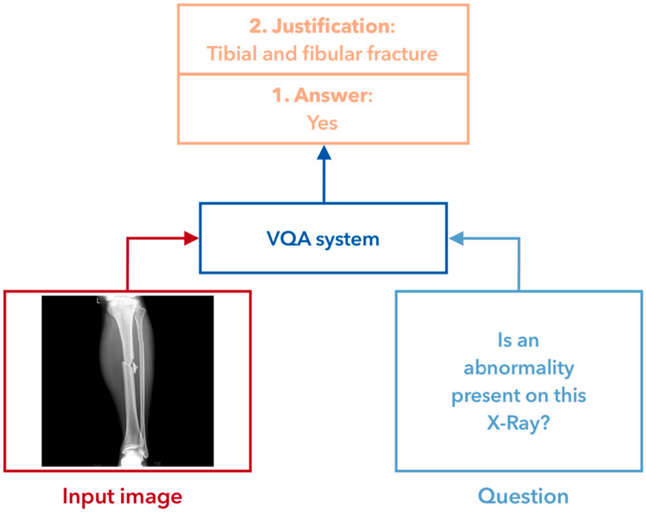
2. Justification:
Tibial and fibular fracture
1. Answer:
Yes
VQA system
Is an abnormality present on this X-Ray?
Input image
Question

Fig. 3. Exemplary application of a two-staged VQA system representing a possible application within the medical imaging domain.

The generative approaches inspire another subset of case-based XAI methods. For example, so-called *counterfactual* explanations aim to enable contrastive reasoning by altering the input image so that a given classifier would make a different prediction than its actual label. Given such an explanation, users are equipped with a contrary sample to compare the model's initial prediction against. Especially in medical contexts, with data often encompassing textual and structural information, counterfactual images can significantly contribute to the decision processes. In [32], a framework called ***GANterfactual*** was proposed to generate counterfactual image explanations based on adversarial image-to-image translation. More precisely, the authors developed a binary classifier to estimate whether a given X-ray image depicts lungs suffering from pneumonia. Subsequently, the authors trained a CycleGAN [33] modified with a custom counterfactual loss function, simultaneously incorporating the binary classifier [32]. In simple terms, the CycleGAN's primary focus is to apply minimally-required changes to the input image that cause the classifier to predict the contrary class. Given the use case of pneumonia, an input image showing pneumonia-infected lungs is transformed into another image depicting the same lung but without the characteristics causing the classifier to label the image as "pneumonia", possibly, resulting in a "healthy" counterpart. Moreover, a user study was conducted to evaluate the approach against the state-of-the-art methods of Layerwise Relevance Propagation [34] and Local Interpretable Model-agnostic Explanations [35]. The authors stated that counterfactual explanations led to significantly better results regarding mental models, explanation satisfaction, and trust. This method appears intuitively explainable due to its contrastive setting, especially for non-experts. However, a limitation of this method is its lack of flexibility and strong model-specificity since a slight change in the task would require the acquisition of a new dataset alongside another training of the CycleGAN and the binary classifier.

2.2. *Textual explanations*

The inability to explain a model prediction in semantically and visually meaningful ways is a well-known shortcoming of several existing computer-aided diagnosis approaches [1]. Conclusively, textual XAI approaches aim at depicting additional information through textual

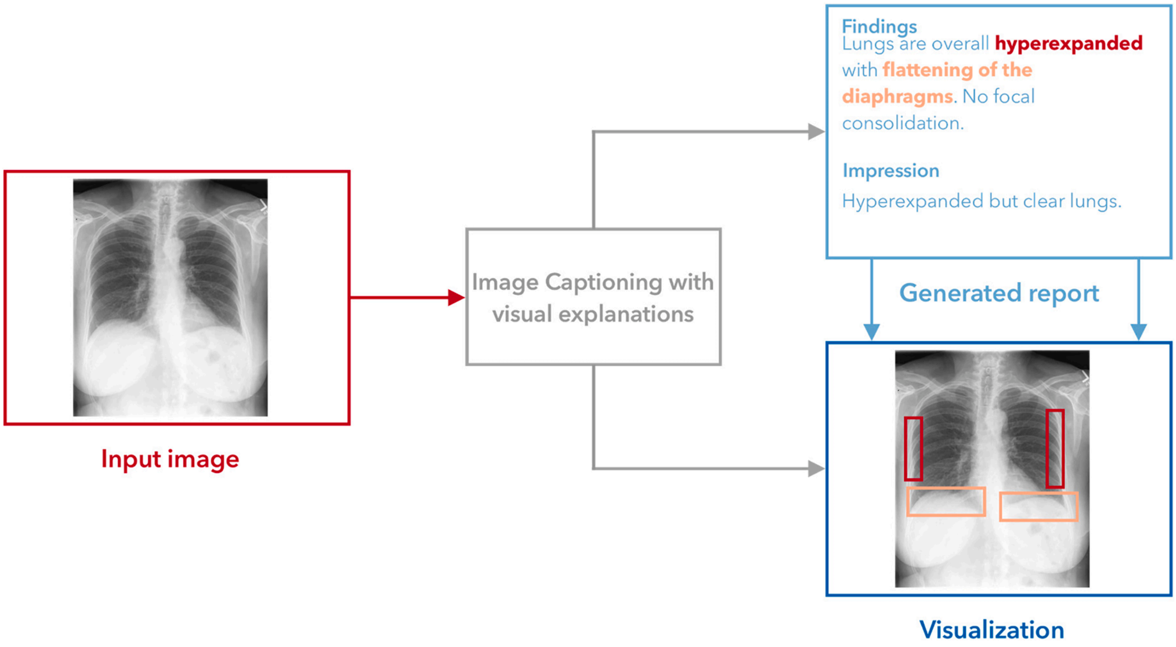
Input image
Image Captioning with visual explanations
Findings
Lungs are overall hyperexpanded with flattening of the diaphragms. No focal consolidation.
Impression
Hyperexpanded but clear lungs.
Generated report
Visualization

Fig. 4. Presentation of an exemplary AI system performing Image Captioning with visual explanations. The automated report generation is extended by visualization on the input image using bounding boxes to denote obtained findings (hyperexpansion and flattened diaphragms). The input image is taken from the IU X-Ray dataset [48].

explanations represented by natural language. Several methods have been proposed within studies to gain additional insight through a more specific and context-related quantification of images [1,36], with solutions ranging from the semantic annotation of relevant regions to the simulation of conversations based on AI agents. One fundamental mutuality is extending the available information basis (images) by additional sources of information concerning the given visual content. One desirable attribute of such systems is that questions can be *free-form* and *open-ended*, meaning that users can ask beyond binary yes-or-no questions. According to these requirements, the authors of [37] proposed a ***Visual Question Answering*** **(VQA)** task. Generally, VQA serves as a representative area of automatic image understanding. An exemplary depiction of such systems is shown in Fig. 3. VQA draws on research from several disciplines, such as Computer Vision, Natural Language Processing, and Knowledge Representation [37]. An important contribution of the VQA task is an accompanying dataset expanding another image dataset called Microsoft *Common Objects in Context* (COCO) [38]. The final VQA dataset contains ~ 0.25 million images, ~0.76 million questions, and ~ 10 million answers.

However, this leads to a challenging requirement of textual explanations, namely the dependence on sophisticated datasets, including images alongside descriptive semantic information. This challenge is specifically present in medical imaging, where several imaging modalities, anatomical areas, and disease entities must be considered. Ren and Zhou [36] developed a model called CGMVQA by using the ImageCLEF 2019 VQA-Med dataset [39], which can answer corresponding questions related to medical images. The model is not restricted to a specific disease and can be used for various image modalities and body regions.

Building on the aim of not only using single questions but allowing for a context-related dialog, Das et al. introduced ***Visual Dialog*** [40], an AI agent attempting a conversation with humans about an image's visual content. A human asks questions, e.g., what color an object is, and the AI agent tries to answer. The agent embeds the visual content and the dialog's history to develop a subsequent answer. More specifically, given a natural image, a dialog history, and a question, the agent tries to deduce a context and answer the question accurately. The Common Objects in Context (COCO) dataset [38], including natural images of multiple everyday objects, served as the basis for generating another dataset called *VisDial*. This dataset was explicitly crafted for the Visual Dialog task and contained dialogs with ten question–answer pairs on ~140 k COCO images, resulting in a total of ~1.4 M question–answer pairs. Given one image, a dialog history encompassing question–answer pairs, and a follow-up question formulated in natural language, the AI agent's goal is to provide a natural language answer to each question. For example, in [41], Visual Dialog was deployed in radiology specific to chest X-ray images bringing forth *RadVisDial*, the first publicly available dataset for visual dialog in radiology. The authors outlined AI agents' practical usefulness and clinical contribution to medical imaging regarding a radiologist's workflow. However, they also pointed out that X-rays are only one of the many data points available (e.g., medications, lab data, clinical data) for a patient's diagnosis and emphasize the importance of including these additional parameters in AI agents to overcome diagnostic limitations. Another radiologic dataset of interest that is not confined to X-ray images and allows for an interplay between visual components and semantic relations within radiologic images is the *Radiology Objects in COntext (ROCO*) [42]. Among several descriptive components, included images contain keywords and descriptions, which can, for example, be used for multimodal image representations in classification or natural sentence generation.

Another approach that does not require user questions but automatically generates an associated description of the image content based on an input image is the task of ***Image Captioning*** [43]. In this work, the authors utilized a modified deep recurrent architecture for sequence modeling [44] to employ a generative model for generating natural language sentences describing a given input image. Such approaches are especially meaningful in the context of automatic diagnosis report generation, which, in a wider sense, can be seen as a form of an image's visual content explanation. In general, diagnostic report generation is a time-consuming and knowledge-intensive task [45]. They summarize findings observed in medical images and are different from image captioning in that they are paragraphs (e.g., indication, findings, and impression) rather than sentences. Moreover, high precision is mandatory, and generated reports must focus on normal and abnormal findings related to medical characteristics rather than general descriptions [45,46]. Several studies investigated the fusion of image captioning with visual explanations, as shown in Fig. 4, called **Image Captioning with**

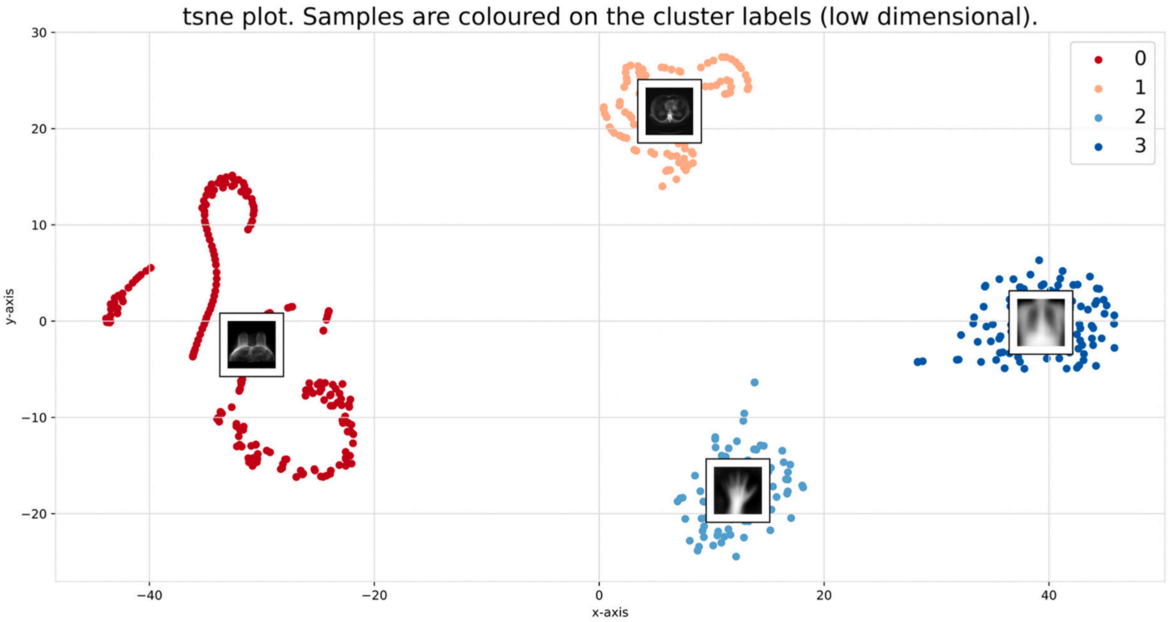

Fig. 5. t-SNE Clustering performed with clustimage [56] on radiologic images contained within the MedNIST dataset [27,28]. The contained data classes were *'BreastMRI' (0), 'AbdomenCT' (1), 'Hand' (2),* and *'CXR' (3).* For each cluster, an image was drawn from the embedding centroid. The detected clusters indicate a good separability between the classes.

Visual Explanations [47].

Zhang et al. [49] proposed an extensive framework named TandemNet, similar to Image Captioning [43], yielding visual attention maps corresponding to textual explanations on images. Lee et al. [50] used Image Captioning with visual explanations for breast mammograms by combining radiology reports with visual saliency maps. A similar approach called TieNet was shown by Wang et al. [51] on chest X-rays. Importantly the authors pointed out how different parts of textual explanations resulted in different saliency maps on the input image related to radiological findings.

Overall, it is noteworthy that systems like VQA or Image Captioning are not XAI methods as such but rather serve as an example of how an interplay between semantic and visual elements could serve as a rectification within the diagnostic process. Moreover, applying these methods does not require an in-depth understanding or method-specific interpretation.

2.3. Auxiliary explanations

Auxiliary measures mainly provide additional information, such as a statistical indicator for single predictions or whole models, and can, for example, be illustrated in tabular or graphical form. Even though the specific interpretation strongly depends on the context and its implementation, auxiliary methods are powerful techniques to provide condensed information for a single prediction or a whole model. Potential applications include i) prediction intervals denoting a prediction's variance [52], ii) plots illustrating uncertainty [53], or iii) importance scores [54]. A prominent scatter-plot-based method that can project high-dimensional data in a two- or three-dimensional space called ***T-distributed stochastic neighbor embedding*** **(t-SNE)** was introduced by van der Maaten and Hinton [12]. It builds upon conditional probabilities to express the distances between data points and find similarities. To facilitate this process, the similarities are usually measured within an embedding space, a relatively low-dimensional space into which input data represented by high-dimensional vectors (e.g., images or texts) can be projected. In simpler terms, t-SNE provides a depiction or intuition of how the data is arranged in a high-dimensional space. Algorithms such as t-SNE are often used to cluster input images based on their activation of neurons in a network. In [55], Rauber et al. use t-SNE to depict activations of hidden neurons and the learned data representations. It could be shown that these projections provide valuable feedback about the relationships between neurons and classes. As stated by the authors: "*This feedback may confirm the known, reveal the unknown, and prompt improvements along the classification pipeline, as we have shown through concrete examples.*" [55].

Moreover, not only neural activations but also the data itself can be directly translated into the embedding space. This is especially helpful when exploring new datasets or analyzing relationships regarding clusters and outliers within the data, as shown in Fig. 5. One pitfall to be aware of is how the interpretation of obtained results is conducted. Altering tunable hyperparameters, such as the *perplexity* (balancing the attention t-SNE gives to local and global aspects of the data), can drastically change the plot. Secondly, t-SNE plots cannot always visualize the relative sizes of clusters appropriately; hence, cluster size and distance should not be over-interpreted. An interactive examination of such effects can be tested in [57]. ***Uniform Manifold Approximation and Projection*** **(UMAP)** [58] is a successor of t-SNE, superior in run time performance, scalability regarding larger datasets, and perseverance of data structure. Graziani et al. [59] used UMAP to visualize layer activations of a classification model for Retinopathy of Prematurity. More precisely, the investigated misclassification errors using a 2D UMAP compression of the activations of specific layers. A major disadvantage of UMAP is its lack of maturity, as it is a relatively new technique, and best practices are not yet established [60].

Prototypes represent another set of auxiliary methods. Human classification of images or objects strongly relies on a subconscious comparison with *prototypical* parts or characteristics associated with certain objects. For example, when classifying cat breeds, the final decision is based on the existence or absence of common *prototypical parts* representing a specific breed, such as ear shape, fur length, or color. This

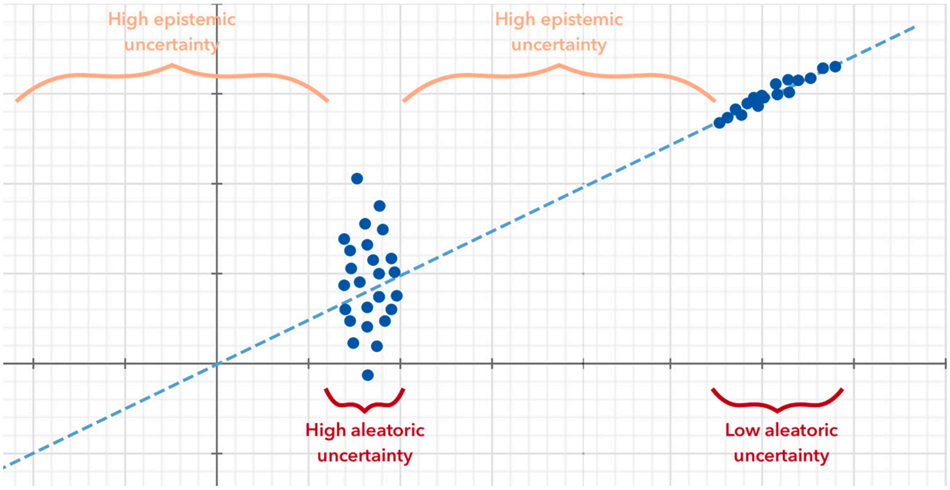

Fig. 6. Depiction of the interplay between epistemic and aleatoric uncertainty. The dashed line represents the approximated function by a model, whereas the blue dots represent samples from the training data. Figure partly inspired by [70].

reasoning is also common in complex identification tasks, e.g., when during the diagnosis of cancer, radiologists compare suspected tumors in radiologic images with prototypical tumor depictions [61]. Consequently, a prototype can be understood as a data instance representative of a specific class and can be used to describe data or develop an interpretable model [62,63]. In this sense, ***ProtoTree*** **(Neural Prototype Trees)** [62] provides a local and global explanation utilizing a decision tree that is explainable by design. Its hierarchical structure is intended to increase interpretability and lead to more insights regarding clusters in the data by exploiting the positive and the negative reasoning process. Instead of similarity scores multiplied by weights, the local explanation shows the trajectory of the entered image through the decision trees. In this work, the number of prototypes is comparatively reduced and not selected according to the class. A significant advantage stems from the fact that each node represents a *trainable* prototype represented by a tensor containing a specific image area. Hence, no manual formulation of prototypes or concepts is required. However, one shortcoming to be aware of is the interpretation of explanations containing no positive matches. For example, a model could predict the class “sparrow” if a sample lacked red feathers, a long beak, and wide wings. Therefore, checking for unambiguity and awareness of exceptional cases is highly recommended [64].

A comparable method is based on feature-space partitioning, referred to as ***TreeView*** [65]. A complex model is represented via hierarchical clustering of features according to the activation values of hidden neurons in such a way that each cluster comprises a set of neurons with a similar distribution of activations across the training set. Subsequently, the feature space is divided into subspaces by clustering similar neurons according to their activation distribution, with each cluster representing a specific factor. Given a prediction of which the actual label is known, a decision tree surrogate model is used to predict the label and traces the decision tree’s relevant nodes for its prediction. The results are depicted in a scatter plot, in which each column denotes if a sample belongs to a specific class. In [66], Sattigeri et al. employed disentangled generative models enabling unsupervised learning of high-level concepts to visualize the surrogate’s decision path within the input space. A fine-grained analysis of the high-level concepts within the input space can yield a different understanding of the connection between inputs and label spaces, enabling the analyst to validate their mental map between the spaces [65]. Overall, a visualization using TreeView enables a convenient transition between factors, class labels, and an input data space. However, an interpretation may be challenging without a thorough understanding of how this method works.

Another set of techniques is derived from conventional DL tools not capturing a model’s *uncertainty*. Uncertainty may appear in two forms, *aleatoric* uncertainty and *epistemic* uncertainty [67], with the former being caused by data or label noise and the latter being caused by sparsity or out-of-data distribution. While epistemic uncertainty is reducible by acquiring more data or optimizing the training process, aleatoric uncertainty is not because of a task’s intrinsic randomness and the fact that a model is only an approximation of it [13,17]. An overview is presented in Fig. 6. Commonly, predictive probabilities obtained as model output (e.g., the softmax output) are often falsely represented as model confidence [68]. However, a model can be uncertain in its predictions even with a high softmax output since different distributions can yield the same probability estimate. Ultimately, the distribution form characterizes the range of a system’s behavior [69]. In response, *uncertainty quantification* seeks to identify and combine sources of variability to define and characterize the range of a system’s possible behavior. This can be performed directly on a model’s weights or by locating outliers within the data and adjusting the model to capture these [17]. Explanation methods involving this uncertainty category enable the identification of model limitations and provide information about possible optimization approaches.

Bayesian Neural Networks **(BNNs)** [68,71,72] are stochastic ANNs trained using Bayesian inference. Bayesian Inference is a powerful framework that initially helps with overfitting in neural networks but also estimates how uncertain a model is by, for example, computing conditional probabilities of a model’s weights w.r.t. the training data. Bayesian networks offer a paradigm for interpretability based on probability theory [73] and are, therefore, intrinsically interpretable. BNNs can be designed in different ways regarding the selection of stochastic components, distribution approximation, and inference approaches [74]. Eaton-Rosene et al. [75] proposed a generalizable technique for quantifying uncertainty with Bayesian Neural Networks for semantic segmentation on the BraTS 2017 dataset [76], including 285 subjects with high- or low-grade gliomas. The networks predict a voxel’s probability of belonging to each segmentation class and generate calibrated confidence intervals of downstream biomarkers.

In contrast, there are also other methods for uncertainty estimation, justified by the fact that Bayesian approaches usually require complex modifications of the training process and can be computationally expensive. One prominent approach is to estimate uncertainty using ***Deep Ensembles*** [77], which quantify uncertainty by sampling from multiple models, where each model is trained separately on the same dataset. The models’ agreement represents the confidence in the overall model ensemble. Yang and Fevens [78] applied Monte Carlo dropout and ensembles to several tasks and modalities, including COVID-19 classification on X-rays, brain tumor classification on MRIs, and breast cancer detection from histopathological slides. An essential advantage of Deep Ensembles is the possibility to distinguish between aleatoric and epistemic uncertainty: The arrangements' mean can be interpreted as aleatoric uncertainty, while the agreements' standard deviation represents epistemic uncertainty, which is low if all mean estimates yield a similar value and grows if the mean estimates differ strongly. Such insights help evaluate the appropriateness of model architecture; for example, if low epistemic uncertainty but high aleatoric uncertainty is observed, the underlying architecture is likely too simple to approximate a given task well [79]. However, as pointed out by D'Angelo and Fortuin [80], a low diversity between models within an ensemble can negatively affect uncertainty estimates and the detection of out-of-distribution data. Even though uncertainty quantification is often treated as a distinct topic separated from explainability, its contribution to this area speaks against separation and, on the contrary, promotes the incorporation of uncertainty estimation as an intrinsic part of XAI research [17]. Moreover, as the authors of [75] pointed out, uncertainty modeling in DL systems makes such models safer for clinical use.

In summary, in Table 1 we present an overview of all the methods introduced in this context, including their strengths and weaknesses. In addition, we specify the explanation target, namely whether a method explains data (model input), internals (model weights, layers, units), or predictions (model output). Lastly, the explanation perspective is denoted as well, specifying if an XAI method provides global (regarding the whole model) or local (regarding specific data samples, internals, or predictions) explanations.

3. Summary

As frequently employed explainability approaches in medical imaging are mainly based on visual explanations relying on saliency maps [9], with this work, we opted for a diversified introduction to state-of-the-art non-visual methods to outline the full potential of XAI in medical imaging. Additionally, guided by the aim of establishing a common ground for cross-disciplinary understanding and exchange between DL practitioners and healthcare professionals, we aimed for a non-technical overview. To allow for an intuitive categorization of the presented XAI approaches, this overview was structured by method output into the following categories: Case-based explanations, textual explanations, and auxiliary explanations. As has been shown, these categories are often very different regarding their individual contribution to explainability and the underlying methodology.

Case-based explanations often allow for local and global analyses of models, including concept-based approaches. These can be specifically interesting for disease entities like melanoma, where the diagnosis process involves classification based on a set of characteristic lesion concepts as investigated for skin lesions classifiers by Lusscieri et al. [81]. Moreover, concerning adversarial approaches, this category also provides counterfactual explanations which are not only intuitively understandable for humans [7] but also pose an interesting research area in terms of cybersecurity, as such explanations arise from deceiving the model by applying minimally required changes to the input, simultaneously revealing models' susceptibility to attacks [82].

In relation to the generated output of the methods, textual explanations are less diversified than case-based explanations. Nevertheless,

Table 1
Overview of introduced explanations alongside their strengths, weaknesses, and explanation targets.

Method output	Method name	Strength	Weakness	Explanation target	Explanation perspective
Case-based	Influence functions [30]	Direct identification of data samples that fall out of distribution	Only applicable to models with a 2nd-order differentiable loss No agreed-upon threshold separating influential from non-influential instances	Data	Global | Local
	Network Dissection [26]	Inner workings Detection of concepts beyond classes in classification task Communication of inner workings in a non-technical way	Dependence on datasets with labels on pixel-level Only positive activations are considered	Internals	Global
	Testing with Concept Activation Vectors [23]	Explorative explanations beyond feature attribution	Manual formulation of concepts required	Predictions | Internals	Global | Local
	Automated concept-based explanations [24]	No manual labeling required	Results depending on selection of parameters	Predictions | Internals	Global | Local
	GANterfactuals [32]	Intuitive explainability by contrastive setting, even for non-experts	Dependence on binary classifier Lack of flexibility	Predictions	Local
Textual	Image Captioning [43]	Semantic relationship between images and textual descriptions is intuitive, even for non-experts		Data | Predictions	Local
	Visual Question Answering [37]	Free-form and open-ended		Data | Predictions	Local
	Visual Dialog [40]	Interactive and context-related dialog		Data	Local
	Image Captioning with Visual Explanations [47]	Interplay between an image's semantic content description and a visual explanation	Dependence on complex datasets including images and textual annotations	Data | Predictions	Local
Auxiliary	t-SNE [12]	Allows for explorative interpretation of activations within a neural network but also investigation	Sensitivity to tunable hyperparameters Cluster size and distance should not be over-interpreted	Data | Internals | Predictions	Global | Local
	UMAP [58]	Improved time performance, Better perseverance of data structure	Lack of maturity	Data | Internals | Predictions	Global | Local
	Deep Ensembles [77]	Allows for estimation of both, epistemic and aleatoric uncertainty	Low diversity of ensemble members affects uncertainty estimates	Data	Global
	ProtoTree [62]	No manual formulation of prototypes/ concepts is required	Possibility of ambiguous result	Predictions	Global | Local
	BNNs [71]	Intrinsically interpretable	Requires complex modifications to training process	Data	Global
	TreeView [65]	Transition between factors, class labels, and input data space	Interpretation requires a thorough understanding of the method	Predictions	Local

they enable a valuable linkage between medical images and semantic information. Also, their interpretation usually does not rely on a thorough understanding of the underlying internal functioning, as generated texts are easily interpretable. A significant advantage is that the interpretation of images, and thus, the generation of diagnostic reports or documentation, can also be automated, which can be particularly helpful for inexperienced or in-training healthcare professionals who want to confirm their diagnostic findings with the help of a support system. Coupled with visualizations, systems encompassing image captioning can even pose as an educative, as findings described within generated captions can be projected directly onto the initial image, justifying which image area is responsible for the generation of a specific caption part. However, diagnostic processes often rely on the assessment of multiple data points [41], such as different X-Ray views (lateral, oblique, or anteroposterior), comparisons between modalities, or evaluating changes over time (e.g., follow-up examinations). For these reasons, systems such as VQA or VisualDialog are clinically limited and rely on further optimization research that takes the presented requirements into account.

Lastly, auxiliary approaches yield explanations in the form of additional information, such as statistical indicators for single predictions or models with their final presentation form strongly varying as results can be, for example, provided as importance scores, visualizations of embedding or activation spaces, or as surrogate white-box models. Great potential is attributed to the quantification of uncertainty. The differentiation between epistemic and aleatoric uncertainty can be used by DL engineers, in particular, to determine whether, and if so, model or training adjustments are necessary. Regarding predictions, uncertainty estimations can uncover when a model does not have high confidence, indicating out-of-distribution samples. Among other things, these considerations are also important for differentiating whether a model acts according to an open-world or closed-world assumption. Differentiation of these assumptions can be crucial as open-world systems have the ability to denote low confidence when answering questions to which the answer is unknown, whereas closed-world systems would simply answer with the most likely answer (class) of a given subset as the opposite cannot be confirmed. Especially with medical questions, it would be incorrect to state that the patient does not suffer from a specific disease if there is no record reporting this unless more information is given to confirm this assumption. Out-of-distribution identification and closed-world assumptions are closely related, as models are usually not equipped with the ability to reject an input during training or inference if it is not represented well by the underlying data or falls out of distribution. Therefore, uncertainty quantification significantly contributes to model integrity within the medical domain.

Conclusively, even though non-visual XAI methods may appear less intuitive compared to prominent saliency-based approaches such as GradCAM [11], they significantly contribute to the diversification of XAI, allowing for the systematic uncovering of model flaws, data outliers, distributional discrepancies, and biases. An aspect that has so far been little researched is the clinical integration and empirical assessment of the impact of such XAI methods. Therefore, investigating this question could be of significant interest for future research.

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 Table A1).

Table A1
Listing of presented XAI methods ordered by their citations/year ratio (according to Google Scholar, see Appendix B Supplementary data) alongside a corresponding open-source repository link (if available). If no open-source repository was available, unofficial implementations were provided.

Method output	Method	Citations/ year ratio	Open source library/ dataset available
Case-based	Influence functions [30]	333	https://github.com/kohpangwei/influence-release
	Testing with Concept Activation Vectors [23]	215	https://github.com/tensorflow/tcav
	Network Dissection [26]	203	https://github.com/CSAILVision/NetDissect
	Automated concept-based explanations [24]	81	https://github.com/amiratag/ACE
	GANterfactuals [32]	8	https://github.com/hcmlab/GANterfactual
Textual	Image Captioning with Visual Explanations [47]	1257	https://github.com/zizhaozhang/tandemnet https://github.com/zizhaozhang/distill2
	Image Captioning [43]	770	https://github.com/sgrvinod/a-PyTorch-Tutorial-to-Image-Captioning
	Visual Question Answering [37]	524	https://github.com/GT-Vision-Lab/VQA_LSTM_CNN https://github.com/jiasenlu/HieCoAttenVQA
	Visual Dialog [40]	146	https://github.com/batra-mlp-lab/visdial-challenge-starter-pytorch
Auxiliary	t-SNE [12]	1816	https://lvdmaaten.github.io/tsne/
	UMAP [58]	1384	https://github.com/lmcinnes/umap
	BNNs [68]	962	https://github.com/yaringal/DropoutUncertaintyExps
	Deep Ensembles [77]	581	https://github.com/SamsungLabs/pytorch-ensembles
	ProtoTree [62]	33	https://github.com/M-Nauta/ProtoTree
	TreeView [65]	7	–

Appendix B. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ejrad.2023.110786.

References

[1] Z. Zhang, Y. Xie, F. Xing, M. McGough, and L. Yang, “MDNet: A Semantically and Visually Interpretable Medical Image Diagnosis Network,” presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6428–6436. Accessed: Apr. 07, 2022. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2017/html/Zhang_MDNet_A_Semantically_CVPR_2017_paper.html.

[2] 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.

[3] S. Koitka, M.S. Kim, M. Qu, A. Fischer, C.M. Friedrich, F. Nensa, Mimicking the radiologists’ workflow: Estimating pediatric hand bone age with stacked deep neural networks, Med. Image Anal. 64 (Aug. 2020), 101743, https://doi.org/10.1016/j.media.2020.101743.

[4] 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.

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

[6] F. Doshi-Velez and 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.

[7] 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.

[8] D. Higgins, V.I. Madai, From bit to bedside: a practical framework for artificial intelligence product development in healthcare, Adv. Intellig. Syst. 2 (10) (2020) 2000052, https://doi.org/10.1002/aisy.202000052.

[9] 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.

[10] J. Amann, A. Blasimme, E. Vayena, D. Frey, V. I. Madai, and the Precise4Q consortium, Explainability for artificial intelligence in healthcare: a multidisciplinary perspective, BMC Medical Informatics and Decision Making, vol. 20, no. 1, p. 310, Nov. 2020, doi: https://doi.org/10.1186/s12911-020-01332-6.

[11] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and 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.

[12] Van der Maaten, Laurens, Hinton, Geoffrey, Visualizing data using t-SNE., ., vol. 9, no. 86, pp. 2579–2605, 2008.

[13] Y. Kwon, J.-H. Won, B.J. Kim, M.C. Paik, Uncertainty quantification using Bayesian neural networks in classification: Application to biomedical image segmentation, Comput. Stat. Data Anal. 142 (Feb. 2020), 106816, https://doi.org/10.1016/j.csda.2019.106816.

[14] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, B. Kim, Sanity checks for saliency maps, in: Proceedings of the 32nd International Conference on Neural Information Processing Systems.

[15] A. Ghorbani, A. Abid, J. Zou, Interpretation of Neural Networks Is Fragile, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, no. 01, Art. no. 01, Jul. 2019, doi: https://doi.org/10.1609/aaai.v33i01.33013681.

[16] R. Tomsett, D. Harborne, S. Chakraborty, P. Gurram, and A. Preece, "Sanity checks for saliency metrics," presented at the AAAI Conference on Artificial Intelligence, Feb. 2020. Accessed: Nov. 14, 2022. [Online]. Available: https://research.ibm.com/publications/sanity-checks-for-saliency-metrics.

[17] 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.

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

[19] F. Hohman, M. Kahng, R. Pienta, D.H. Chau, Visual Analytics in Deep Learning: An Interrogative Survey for the Next Frontiers, IEEE Trans. Vis. Comput. Graph. 25 (8) (Aug. 2019) 2674–2693, https://doi.org/10.1109/TVCG.2018.2843369.

[20] Q.V. Liao, D. Gruen, S. Miller, Questioning the AI: Informing Design Practices for Explainable AI User Experiences, in: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, New York, NY, USA: Association for Computing Machinery, 2020, pp. 1–15. Accessed: Jun. 06, 2022. [Online]. Available: https://doi.org/10.1145/3313831.3376590.

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

[22] M.T. Keane, B. Smyth, Good counterfactuals and where to find them: a case-based technique for generating counterfactuals for explainable AI (XAI), in: in Case-Based Reasoning Research and Development, Cham, 2020, pp. 163–178, https://doi.org/10.1007/978-3-030-58342-2_11.

[23] B. Kim et al., Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV), in: Proceedings of the 35th International Conference on Machine Learning, Jul. 2018, pp. 2668–2677. Accessed: Feb. 14, 2022. [Online]. Available: https://proceedings.mlr.press/v80/kim18d.html.

[24] 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.

[25] D. Sauter, G. Lodde, F. Nensa, D. Schadendorf, E. Livingstone, M. Kukuk, Validating Automatic Concept-Based Explanations for AI-Based Digital Histopathology, Sensors, vol. 22, no. 14, Art. no. 14, Jan. 2022, doi: https://doi.org/10.3390/s22145346.

[26] D. Bau, B. Zhou, A. Khosla, A. Oliva, A. Torralba, Network Dissection: Quantifying Interpretability of Deep Visual Representations, presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6541–6549. Accessed: Apr. 07, 2022. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2017/html/Bau_Network_Dissection_Quantifying_CVPR_2017_paper.html.

[27] J. Yang, R. Shi, 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: https://doi.org/10.1109/ISBI48211.2021.9434062.

[28] 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.

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

[30] 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.

[31] C.J. Wang, et al., Deep learning for liver tumor diagnosis part II: convolutional neural network interpretation using radiologic imaging features, Eur Radiol 29 (7) (Jul. 2019) 3348–3357, https://doi.org/10.1007/s00330-019-06214-8.

[32] S. Mertes, T. Huber, K. Weitz, A. Heimerl, E. André, GANterfactual—Counterfactual Explanations for Medical Non-experts Using Generative Adversarial Learning, Front. Artificial Intelligence, vol. 5, 2022, Accessed: Jul. 12, 2022. [Online]. Available: https://www.frontiersin.org/articles/10.3389/frai.2022.825565.

[33] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired Image-To-Image Translation Using Cycle-Consistent Adversarial Networks," presented at the Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2223–2232. Accessed: Apr. 07, 2022. [Online]. Available: https://openaccess.thecvf.com/content_iccv_2017/html/Zhu_Unpaired_Image-To-Image_Translation_ICCV_2017_paper.html.

[34] 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.

[35] 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: https://doi.org/10.1145/2939672.2939778.

[36] F. Ren, Y. Zhou, CGMVQA: A New Classification and Generative Model for Medical Visual Question Answering, IEEE Access 8 (2020) 50626–50636, https://doi.org/10.1109/ACCESS.2020.2980024.

[37] 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.

[38] T.-Y. Lin, et al., Microsoft COCO: Common Objects in Context, in: in Computer Vision – ECCV Cham, 2014, pp. 740–755, https://doi.org/10.1007/978-3-319-10602-1_48.

[39] A. Ben Abacha, S. Hasan, V. Datla, J. Liu, D. Demner-Fushman, H. Müller, VQA-Med: Overview of the Medical Visual Question Answering Task at ImageCLEF 2019, Lect. Notes Comput. Sci (Sep. 2019).

[40] A. Das et al., "Visual Dialog," presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 326–335. Accessed: Apr. 07, 2022. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2017/html/Das_Visual_Dialog_CVPR_2017_paper.html.

[41] O. Kovaleva *et al.*, Towards Visual Dialog for Radiology, in: Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing, Online, Jul. 2020, pp. 60–69. doi: https://doi.org/10.18653/v1/2020.bionlp-1.6.

[42] O. Pelka, S. Koitka, J. Rückert, F. Nensa, C.M. Friedrich, Radiology Objects in COntext (ROCO): A Multimodal Image Dataset, in: in Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis, Cham, 2018, pp. 180–189, https://doi.org/10.1007/978-3-030-01364-6_20.

[43] 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.

[44] S. Hochreiter, J. Schmidhuber, Long Short-Term Memory, Neural Comput. 9 (8) (Nov. 1997) 1735–1780, https://doi.org/10.1162/neco.1997.9.8.1735.

[45] S. Yang, J. Niu, J. Wu, X. Liu, Automatic Medical Image Report Generation with Multi-view and Multi-modal Attention Mechanism, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 12454 LNCS, pp. 687–699, 2020, doi: https://doi.org/10.1007/978-3-030-60248-2_48.

[46] J. Yuan, H. Liao, R. Luo, J. Luo, Automatic Radiology Report Generation Based on Multi-view Image Fusion and Medical Concept Enrichment, in: in Medical Image Computing and Computer Assisted Intervention – MICCAI, Cham, 2019, pp. 721–729, https://doi.org/10.1007/978-3-030-32226-7_80.

[47] K. Xu et al., Show, attend and tell: neural image caption generation with visual attention, in: Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37, Lille, France, Jul. 2015, pp. 2048–2057.

[48] D. Demner-Fushman, et al., Preparing a collection of radiology examinations for distribution and retrieval, J. Am. Med. Inform. Assoc. 23 (2) (Mar. 2016) 304–310, https://doi.org/10.1093/jamia/ocv080.

[49] Z. Zhang, P. Chen, M. Sapkota, L. Yang, TandemNet: Distilling Knowledge from Medical Images Using Diagnostic Reports as Optional Semantic References, in: in Medical Image Computing and Computer Assisted Intervention – MICCAI, Cham, 2017, pp. 320–328, https://doi.org/10.1007/978-3-319-66179-7_37.

[50] H. Lee, S.T. Kim, Y.M. Ro, Generation of Multimodal Justification Using Visual Word Constraint Model for Explainable Computer-Aided Diagnosis, in: in Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support, Cham, 2019, pp. 21–29, https://doi.org/10.1007/978-3-030-33850-3_3.

[51] X. Wang, Y. Peng, L. Lu, Z. Lu, R. M. Summers, TieNet: Text-Image Embedding Network for Common Thorax Disease Classification and Reporting in Chest X-Rays, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Jun. 2018, pp. 9049–9058. doi: https://doi.org/10.1109/CVPR.2018.00943.

[52] T. Pearce, A. Brintrup, M. Zaki, A. Neely, High-Quality Prediction Intervals for Deep Learning: A Distribution-Free, Ensembled Approach, in: Proceedings of the 35th International Conference on Machine Learning, Jul. 2018, pp. 4075–4084. Accessed: Dec. 12, 2022. [Online]. Available: https://proceedings.mlr.press/v80/pearce18a.html.

[53] M.S. Ayhan, P. Berens, Test-time Data Augmentation for Estimation of Heteroscedastic Aleatoric Uncertainty in Deep Neural Networks, presented at the Medical Imaging with Deep Learning, Apr. 2018. Accessed: Apr. 07, 2022. [Online]. Available: https://openreview.net/forum?id=rJZz-knjz.

[54] W. Jin, X. Li, G. Hamarneh, Evaluating Explainable AI on a Multi-Modal Medical Imaging Task: Can Existing Algorithms Fulfill Clinical Requirements? arXiv, Mar. 12, 2022. doi: https://doi.org/10.48550/arXiv.2203.06487.

[55] P.E. Rauber, S.G. Fadel, A.X. Falcão, A.C. Telea, Visualizing the Hidden Activity of Artificial Neural Networks, IEEE Trans. Vis. Comput. Graph. 23 (1) (Jan. 2017) 101–110, https://doi.org/10.1109/TVCG.2016.2598838.

[56] E. Taskesen, Python package clustimage is for unsupervised clustering of images. Nov. 2021. Accessed: Dec. 10, 2022. [Online]. Available: https://erdogant.github.io/clustimage.

[57] M. Wattenberg, F. Viégas, I. Johnson, How to Use t-SNE Effectively, accessed Apr. 13, 2022, Distill (Oct. 13, 2016.), http://distill.pub/2016/misread-tsne.

[58] L. McInnes, J. Healy, N. Saul, L. Großberger, UMAP: Uniform Manifold Approximation and Projection, J. Open Source Software 3 (29) (Sep. 2018) 861, https://doi.org/10.21105/joss.00861.

[59] M. Graziani et al., Improved interpretability for computer-aided severity assessment of retinopathy of prematurity, in: Medical Imaging 2019: Computer-Aided Diagnosis, Mar. 2019, vol. 10950, pp. 450–460. doi: https://doi.org/10.1117/12.2512584.

[60] S. Nanga, et al., Review of dimension reduction methods, J. Data Anal. Informat. Process. 9 (3) (2021), https://doi.org/10.4236/jdaip.2021.93013.

[61] A. Holt, I. Bichindaritz, R. Schmidt, P. Perner, Medical applications in case-based reasoning, Knowl. Eng. Rev. 20 (3) (Sep. 2005) 289–292, https://doi.org/10.1017/S0269888906000622.

[62] M. Nauta, R. van Bree, and C. Seifert, "Neural Prototype Trees for Interpretable Fine-Grained Image Recognition," presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 14933–14943. Accessed: Jul. 12, 2022. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2021/html/Nauta_Neural_Prototype_Trees_for_Interpretable_Fine-Grained_Image_Recognition_CVPR_2021_paper.html?ref=https://githubhelp.com.

[63] C. Chen, O. Li, C. Tao, A.J. Barnett, J. Su, C. Rudin, in: This looks like that: deep learning for interpretable image recognition, Curran Associates Inc., Red Hook, NY, USA, 2019, pp. 8930–8941.

[64] D. Rymarczyk, Ł. Struski, M. Górszczak, K. Lewandowska, J. Tabor, and B. Zieliński, "Interpretable Image Classification with Differentiable Prototypes Assignment." arXiv, Dec. 06, 2021. Accessed: Jun. 03, 2022. [Online]. Available: http://arxiv.org/abs/2112.02902.

[65] J. J. Thiagarajan, B. Kailkhura, P. Sattigeri, and K. N. Ramamurthy, "TreeView: Peeking into Deep Neural Networks Via Feature-Space Partitioning," arXiv: 1611.07429 [cs, stat], Nov. 2016, Accessed: Apr. 07, 2022. [Online]. Available: http://arxiv.org/abs/1611.07429.

[66] P. Sattigeri, K. N. Ramamurthy, J.J. Thiagarajan, B. Kailkhura, Treeview and Disentangled Representations for Explaining Deep Neural Networks Decisions, in: 2020 54th Asilomar Conference on Signals, Systems, and Computers, Nov. 2020, pp. 284–288. doi: https://doi.org/10.1109/IEEECONF51394.2020.9443487.

[67] A.D. Kiureghian, O. Ditlevsen, Aleatory or epistemic? Does it matter? Struct. Saf. 31 (2) (Mar. 2009) 105–112, https://doi.org/10.1016/j.strusafe.2008.06.020.

[68] Y. Gal, Z. Ghahramani, Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning, in: Proceedings of The 33rd International Conference on Machine Learning, Jun. 2016, pp. 1050–1059. Accessed: Jul. 08, 2022. [Online]. Available: https://proceedings.mlr.press/v48/gal16.html.

[69] M.C. Darling, D.J. Stracuzzi, Toward Uncertainty Quantification for Supervised Classification, SAND–2018-0032, 1527311, Jan. 2018. doi: https://doi.org/10.2172/1527311.

[70] M. Abdar, et al., A review of uncertainty quantification in deep learning: Techniques, applications and challenges, Information Fusion 76 (Dec. 2021) 243–297, https://doi.org/10.1016/j.inffus.2021.05.008.

[71] J. Lampinen, A. Vehtari, Bayesian approach for neural networks—review and case studies, Neural Netw. 14 (3) (Apr. 2001) 257–274, https://doi.org/10.1016/S0893-6080(00)00098-8.

[72] D.M. Titterington, Bayesian Methods for Neural Networks and Related Models, Stat. Sci. 19 (1) (Feb. 2004) 128–139, https://doi.org/10.1214/088342304000000099.

[73] B. Mihaljević, C. Bielza, P. Larrañaga, Bayesian networks for interpretable machine learning and optimization, Neurocomputing 456 (Oct. 2021) 648–665, https://doi.org/10.1016/j.neucom.2021.01.138.

[74] L.V. Jospin, H. Laga, F. Boussaid, W. Buntine, M. Bennamoun, Hands-On Bayesian Neural Networks—A Tutorial for Deep Learning Users, IEEE Comput. Intell. Mag. 17 (2) (May 2022) 29–48, https://doi.org/10.1109/MCI.2022.3155327.

[75] Z. Eaton-Rosen, F. Bragman, S. Bisdas, S. Ourselin, M.J. Cardoso, Towards Safe Deep Learning: Accurately Quantifying Biomarker Uncertainty in Neural Network Predictions, in: in Medical Image Computing and Computer Assisted Intervention – MICCAI, Cham, 2018, pp. 691–699, https://doi.org/10.1007/978-3-030-00928-1_78.

[76] B.H. Menze, et al., The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS), IEEE Trans Med Imaging 34 (10) (Oct. 2015) 1993–2024, https://doi.org/10.1109/TMI.2014.2377694.

[77] B. Lakshminarayanan, A. Pritzel, C. Blundell, Simple and scalable predictive uncertainty estimation using deep ensembles, in:.

[78] S. Yang, T. Fevens, Uncertainty Quantification and Estimation in Medical Image Classification, in: in Artificial Neural Networks and Machine Learning – ICANN, Cham, 2021, pp. 671–683, https://doi.org/10.1007/978-3-030-86365-4_54.

[79] M.E.E. Khan, A. Immer, E. Abedi, M. Korzepa, Approximate Inference Turns Deep Networks into Gaussian Processes, in: Advances in Neural Information Processing Systems, 2019, vol. 32. Accessed: Jun. 08, 2022. [Online]. Available: https://proceedings.neurips.cc/paper/2019/hash/b3bbccd6c008e727785cb81b1aa08ac5-Abstract.html.

[80] F. D'Angelo, V. Fortuin, "Repulsive Deep Ensembles are Bayesian," presented at the Neural Information Processing Systems, Jun. 2021. Accessed: Dec. 18, 2022. [Online]. Available: https://www.semanticscholar.org/paper/Repulsive-Deep-Ensembles-are-Bayesian-D'Angelo-Fortuin/be5491660a61d60606aaec8dc0e7e046fb930110.

[81] A. Lucieri, M.N. Bajwa, S.A. Braun, M.I. Malik, A. Dengel, S. Ahmed, On Interpretability of Deep Learning based Skin Lesion Classifiers using Concept Activation Vectors, in: 2020 International Joint Conference on Neural Networks (IJCNN), Jul. 2020, pp. 1–10. doi: https://doi.org/10.1109/IJCNN48605.2020.9206946.

[82] X. Ma, et al., Understanding adversarial attacks on deep learning based medical image analysis systems, Pattern Recogn. 110 (Feb. 2021), 107332, https://doi.org/10.1016/j.patcog.2020.107332.