Explainable artificial intelligence to increase transparency for revolutionizing healthcare ecosystem and the road ahead

Article *in* Network Modeling Analysis in Health Informatics and Bioinformatics · December 2023

DOI: 10.1007/s13721-023-00437-y

CITATIONS
3

READS
104

3 authors:

Sudipta Roy

143 PUBLICATIONS **2,498** CITATIONS

SEE PROFILE

Debojyoti Pal

Government College of Engineering and Leather Technology

10 PUBLICATIONS **88** CITATIONS

SEE PROFILE

Tanushree Meena

Indian Institute of Technology (Banaras Hindu University) Varanasi

16 PUBLICATIONS **183** CITATIONS

SEE PROFILE

Network Modeling Analysis in Health Informatics and Bioinformatics (2024) 13:4
https://doi.org/10.1007/s13721-023-00437-y

REVIEW ARTICLE

Explainable artificial intelligence to increase transparency for revolutionizing healthcare ecosystem and the road ahead

Sudipta Roy[1] · **Debojyoti Pal**[1] · **Tanushree Meena**[1]

Received: 25 July 2023 / Revised: 8 November 2023 / Accepted: 9 November 2023

Abstract
The integration of deep learning (DL) into co-clinical applications has generated substantial interest among researchers aiming to enhance clinical decision support systems for various aspects of disease management, including detection, prediction, diagnosis, treatment, and therapy. However, the inherent opacity of DL methods has raised concerns within the healthcare community, particularly in high-risk or complex medical domains. There exists a significant gap in research and understanding when it comes to elucidating and rendering transparent the inner workings of DL models applied to the analysis of medical images. While explainable artificial intelligence (XAI) has gained ground in diverse fields, including healthcare, numerous unexplored facets remain within the realm of medical imaging. To better understand the complexities of DL techniques, there is an urgent need for rapid advancement in the field of eXplainable DL (XDL) or eXplainable Artificial Intelligence (XAI). This would empower healthcare professionals to comprehend, assess, and contribute to decision-making processes before taking any actions. This viewpoint article conducts an extensive review of XAI and XDL, shedding light on methods for unveiling the "black-box" nature of DL. Additionally, it explores the adaptability of techniques originally designed for solving problems across diverse domains for addressing healthcare challenges. The article also delves into how physicians can interpret and comprehend data-driven technologies effectively. This comprehensive literature review serves as a valuable resource for scientists and medical practitioners, offering insights into both technical and clinical aspects. It assists in identifying methods to make XAI and XDL models more comprehensible, enabling wise model choices based on particular requirements and goals.

Keywords Artificial intelligence · Deep learning · Explainable AI · Medical image

1 Introduction

A substantial improvement has been made across multiple areas of medicine and healthcare using AI and DL in the past few decades in many multimodal images (Roy et al. 2022). However, many interdisciplinary medical imaging (MI) researchers are losing interest in using black-box AI and DL methods although DL is giving very high accuracy. In response to the difficulties of interpretation of black boxes to the co-clinical community, researchers are working toward XDL to make comfortable and high-risk use cases in MI (Roy et al. 2022; Meena and Roy 2022). The quantity of papers published in the XAI in MI related papers over the year from PubMed search is shown in Fig. 1.

The concept of explainability in its current state remains abstract, lacks a clear consensus, and is characterized by imprecision (Patrício et al. 2022). Considering this, the novel approaches to enhance the explainability of DL techniques before being viable to utilize them in co-clinical use cases. In most of the previous research, the indirect analysis of the decision procedure of the existing model was primarily focused. These methods were applied on random network model without having an extra modification to help their effectiveness in the initial days of XDL. However, the post hoc explainable approaches endure numerous shortcomings considering the way of representation of explanations. Despite the recent popularity of diverse types of explainable

✉ Sudipta Roy
Sudipta1.roy@jioinstitute.edu.in

✉ Tanushree Meena
tanushree.meena@jioinstitute.edu.in

Debojyoti Pal
debojyoti.pal@jioinstitute.edu.in

[1] Artificial Intelligence and Data Science, Jio Institute, Navi Mumbai 410206, India

Published online: 22 December 2023

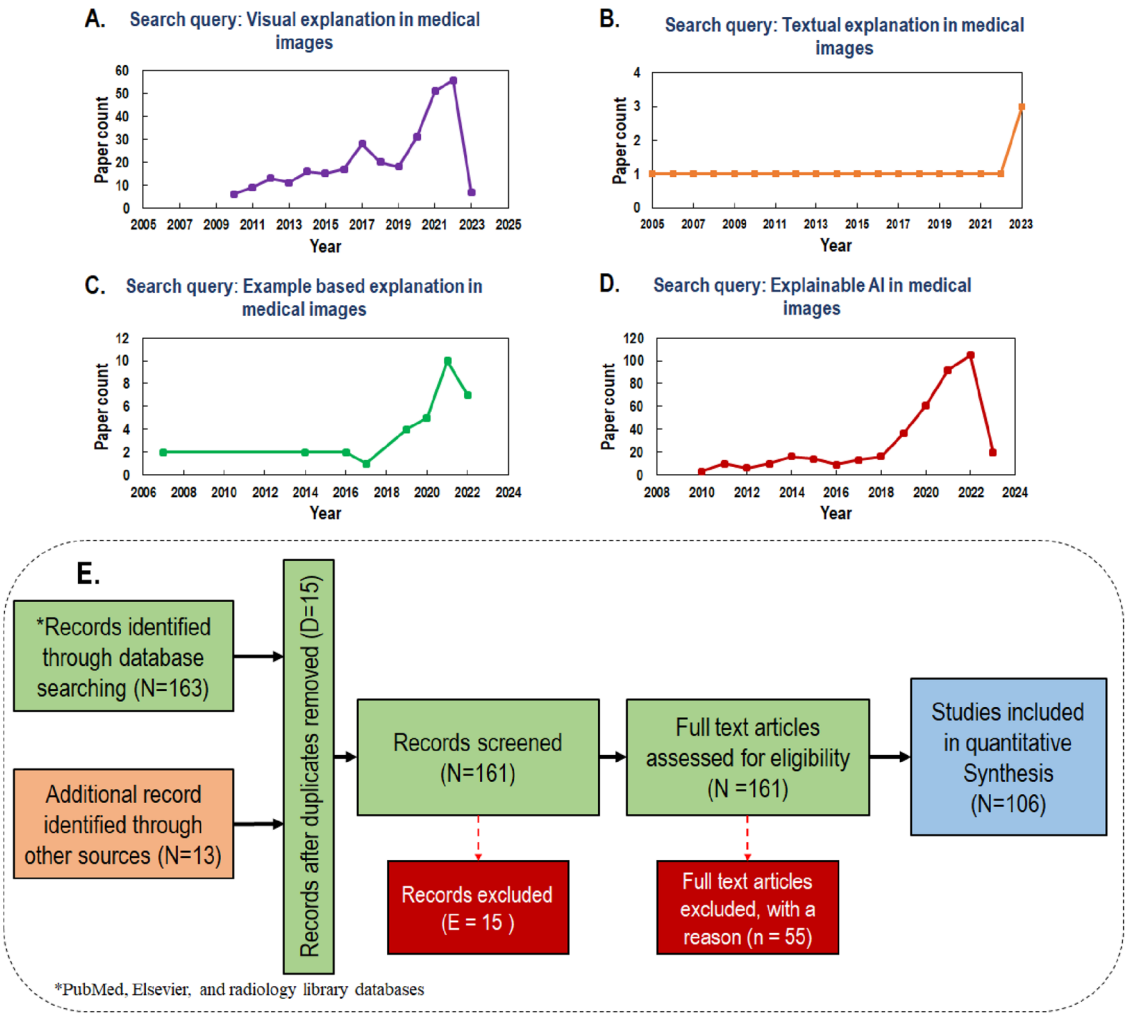

Fig. 1 Research articles published per year from 2005 on PubMed in XAI and MI related areas. **A** After using the search key "*Visual explanation in medical images*", **B** After using the search key "*Textual explanation in medical images*", **C** After using the search key "*Example-based explanation in medical images*", **D** After using the search key "*Explanation AI in medical images*", and **E** A schematic review (PRISMA) process for the study

models, the existing surveys on the interpretability of the DL method for medical image analysis (MIA) have not carefully assessed the advancement that happened in this novel area (Patrício et al. 2022).

An extensive systematic review search was conducted regarding the explainability of the deep learning model in the field of medical imaging. This is a retrospective study that combines and interprets the acquired data. All articles were retrieved from PubMed, Elsevier, and radiology library databases. The keywords used to search the articles are the combination of the words like "Visual explanation in medical images", B) After using the search key "Textual explanation in medical images", C) After using the search key "Example-based explanation in medical images", and D) After using the search key "Explanation AI in medical images" or "Textual explanation in medical images" or "Example-based explanation in medical images" or "Explanation AI in medical images". The search was conducted in March 2023.

The following key questions are answered in this review:

- What are existing techniques for explainability in medical imaging?
- How a black box can be interpreted?
- How does the interpretable model in radiology immensely help medical practitioners?
- What are the current challenges in the medical domain because of the black-box nature?
- What are the future research prospects in this field?

To point out these issues, we carefully studied recent work on XDL that were used in healthcare and mostly

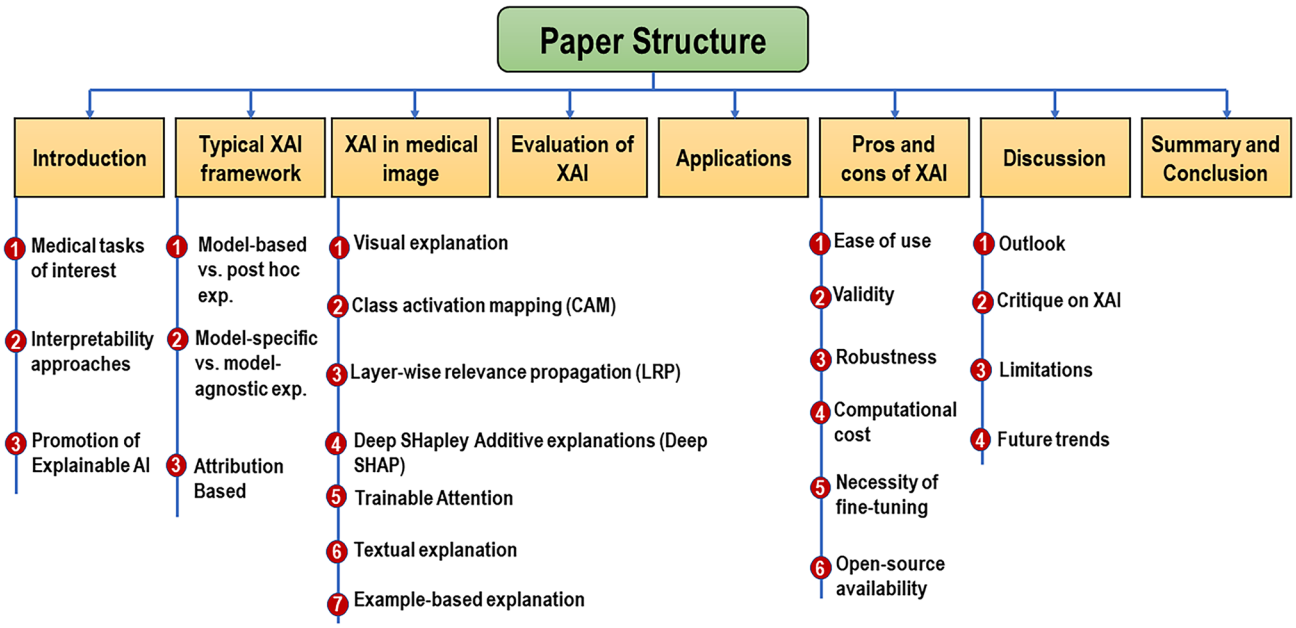

Fig. 2 The overall structure of paper in the tree diagram

on MIA. In a specific way, this review has the following contributions:

- A comprehensive analysis, conclusion derived, uses, and major contribution of XDL in MI.
- The performance metrics frequently used in assessment XDL for pictorial and textual descriptions.
- The XDL methods for the superiority of the textual interpretation in MIA.
- The forthcoming research possibility on XDL with a special focus on medical images and related areas.
- The useability for XDL in medical images.

The overall organization of this study is shown in Fig. 2 below.

1.1 Medical tasks of interest

The XDL methods develop the clarity of DL methods, which leads to further assurance in healthcare decision-making, and more tangible to adopt for clinical use. The benefit of XDL to healthcare professionals by gaining detailed insight into DL methods reaches the decision-making solutions using MI, as shown in Fig. 3.

The main objective of DL-based methods in medical and healthcare settings is to reach very high accuracy. The end user must trust, interpret, and indicate the user's response. In particular, the prototype should have accomplished adequate performance at their mission in a real-world clinical data setting which may not be used during their training procedure (Giuste et al. 2022). The faith in the XDL method is essential, specifically when the visual response is provided to the operator on significant model prediction. The solution is not enough if the visualization does not properly reflect the decision-making. The decision can be helpful in treatment planning, monitoring, diagnostics, precision medicine, and critical care as well by recognizing the avenue of implementation if a detailed explanation is given.

1.2 Interpretability approaches

There are many options to develop an XAI framework. One possible technical skeleton of XAI is shown in Fig. 4. This comprises of two parts: in part 1, the methodological implementation of DL to achieve the targets, and in part 2, the explainable blocks of each module and/or layer (Giuste et al. 2022).

The images are forward transmitted through a convolutional neural network (CNN) generated convolutional feature maps (FM) and other traditional computer model from a given input image and passing over a job implicit computation to find the desired decision-making system (Xing et al. 2022a, b; Meena and Sarawadekar 2023; Roy et al. 2017b; Meswal et al. 2023). Then, the output would be shown to healthcare specialists to review the outcomes and request a clarification if necessary (Zeineldin et al. 2022a). Finally, a visual explanation block needs to be provided to understand the effects of DL networks in each layer of the final output. This could be used as a state-of-the-art (SOTA) XAI pipeline.

In recent years, the traditional way of precision diagnostics is rapidly using various imaging like X-ray, CT scan, MRI, PET, and pathological imaging to make the clinician easier to evaluate tons of images (Roy and Shoghi 2019; Roy and Bandyopadhyay 2016). The automation of clinical decision support (CDS) systems outperforms the traditional models, but widespread adaption is still a barrier due to the lack of explainability. In the framework of estimation of results, the XDL-supported diagnosis via medical images is extremely needed to remove the risk (Zeineldin et al. 2022a). The most common XDL approach in outcome projection using traditional supervised machine learning is the feature scoring (FS) interpretation techniques such as saliency, and feature attribution (FA) could be used as evidence-supporting tools

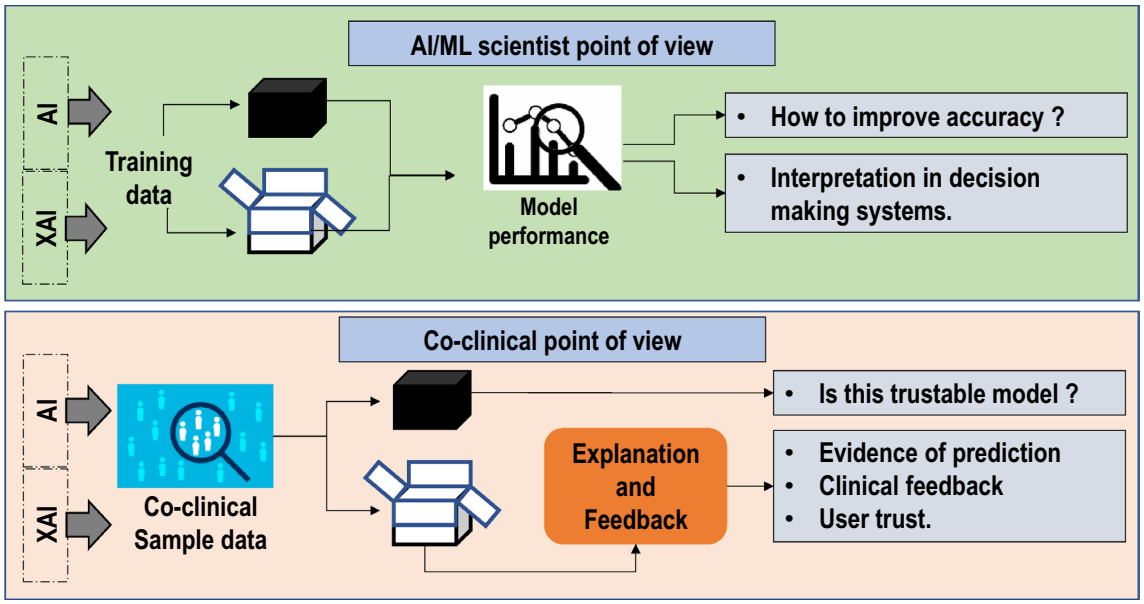

Fig. 3 The clinician's advantage using XAI/XDL after having understanding into the models to grasp the results. The XAI helps to upsurge the transparency of automated representations to make more assured decisions (Giuste et al. 2022)

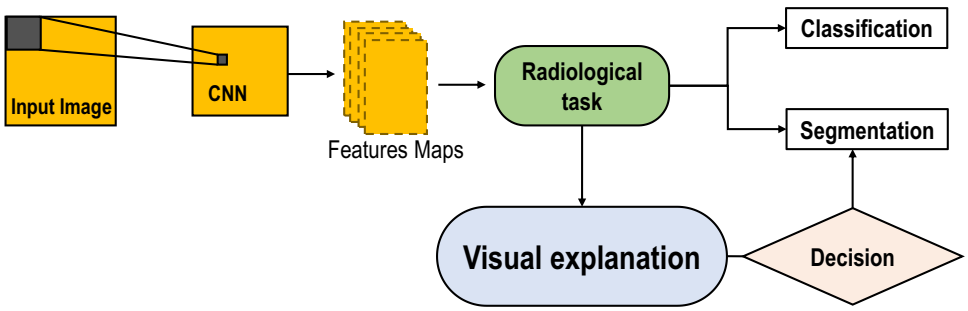

Fig. 4 The pipeline of NeuroXAI framework (Zeineldin et al. 2022a)

Table 1 Overview of different XAI approaches

Model	Typical feature scoring
Perturbation	• Identifies the difference between the input and reference image of the model result • The area is considered the correct label if the ablation of the area results in a change
Activation	The activation approach identifies the region with the highest neuron activation on the forward pass for the specific model
Gradient	Features having the highest effect on gradients throughout backward loss
Mixed	Gradients measured on the backward pass to upsurge activation-based resolution
Attention	• Learn important regions during the training to custom attention map in the ultimate model • Class explicit features optimization in training to emphasize attention

specifically, given input, to know the importance of scores. The typical FS is classified as perturbation, activation, gradient, combination of both activation and gradient, and attention approaches, and their corresponding schematic (Giuste et al. 2022) is shown in Table 1.

1.3 Promotion of the explainable AI

The first question that comes to our mind is whether XDL based method capable of fulfilling the healthcare community requirements in MI? The question comes to our mind due to the invisible, difficult-to-understand, and data-driven approach by DL methods (Pal et al. 2022). Another question comes to our mind about the capability of DL tools as reliable outcomes, as they are often impenetrable, and not fully understandable and that makes it difficult to get reliable decision-making outcomes. In that scenario, XDL along with trustworthiness, responsibility, privacy preservation, and validity could help the wide-scale application of XDL in healthcare decision-making. This would enable researchers and policymakers to make informed decisions regarding precision for widespread clinical use (Yanga et al. 2021).

The foundation of the progression of the decision system with both clinical level XDL and methodological XDL could be well addressed by the trustable, reproducibility, privacy, responsibility, and explainable (see in Fig. 5) to participate in the ethical level into design level operation. One high-level pathway toward XDL by the European Union is shown in Fig. 5.

Similarly, the United States Défense Research Project Agency decided to work on the new research effort by targeting more explainability in the DL models (see next Fig. 6). The XDL is a robust area with many studies going on in promising and numerous original plans developing that create a massive influence on DL expansion in multiple ways (see in Fig. 6).

The XDL aims to examine the decision-making process and evaluate its strengths and weaknesses, as well as propose an approach for future use. The XAI should enable researchers to comprehend the underlying insights of AI-based methods in healthcare by providing explanations for how these methods achieve their outcomes. To achieve greater visibility and trustworthiness, a suggested module of the XAI model has been added to the existing model, as depicted in Fig. 6. It is crucial to evaluate both generalizability and human experience to achieve a validated prediction. Without interpretable surrogate unit addition, deep learning may cause anxiety among operators and reduce the usability of the sophisticated model. The model should explain the question about the "reason of output(s)", "reliability and trust-ability of the model", "when it could success and fail", and "rectification of error". A typical feedback loop is needed for further XAI advancement which contains each stage from train, quality assurance, deploy, forecast, cross-validation, debug and monitoring, and these seven steps is shown in Fig. 7.

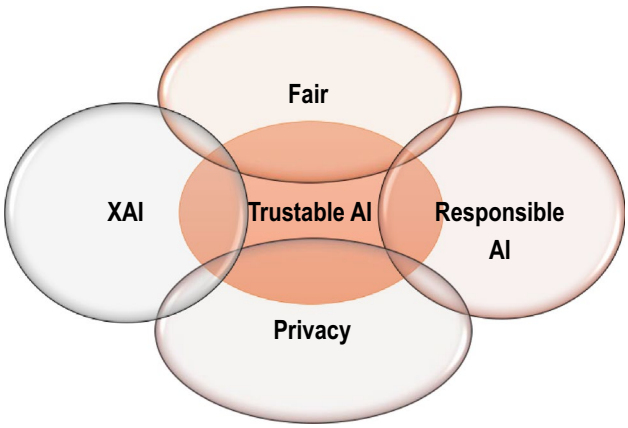

Fig. 5 The Venn diagram of trustable XAI (Yanga et al. 2021)

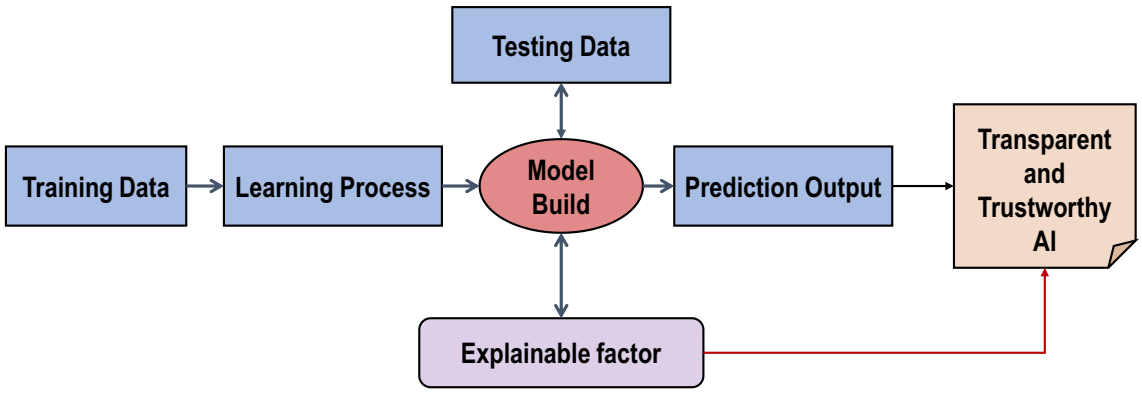

Fig. 6 The block diagram of schema of the explainable alternate module to accomplish a transparent and trustworthy model (Yanga et al. 2021)

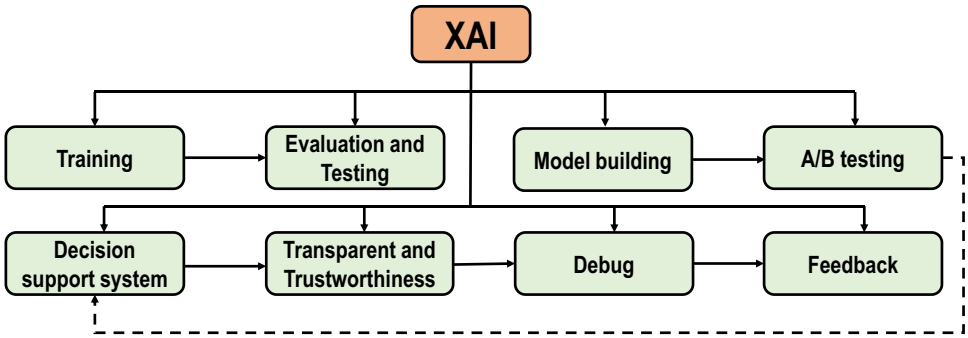

Fig. 7 A typical XAI feedback loop development for clinical practice (Yanga et al. 2021)

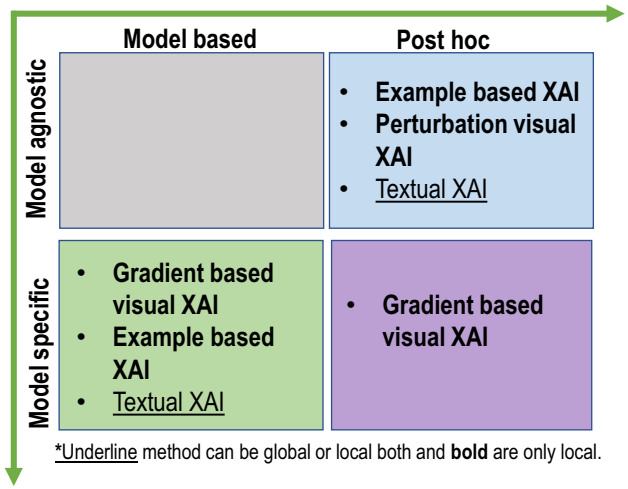

Fig. 8 The explainable AI structure under global and local model

There are different types of XAIs used in the in research, technology, innovation, public, and policy debates. The first option is interpretable in AI technology point of view, the second option is explainable that offers explanation for wide range of users for decision-making system, the third option is transparency which assess the accessibility of the model, the fourth option is Justifiability to increase the understandability of case-to-case outcome, and contestability to allow arguments against the decision. In traditional AI-based method (Roy et al. 2020), the expandability decreases with the increase of model performance. The typical figure of the XAI exploitability Vs. performance.

2 Typical XAI framework

The main XAI model can be divided based on model-based versus (vs.) post hoc, model specific vs. model agnostic, and global vs. local explanation tasks. The complete structure of these three principles was designed from a surveys paper (Velden et al. 2022a) and is shown in Fig. 8.

2.1 *Model based vs.* post hoc

Model-based explanation models encompass regression and traditional classifiers. These models are designed to represent the correlation between input and output in a straightforward manner, making them easy to comprehend (Velden et al. 2022a). With a limited number of features, healthcare professionals can readily interpret the model's insights. However, deep learning utilizes complex neural networks with numerous hidden layers and dense weights, making it challenging for developers and users to track and predict the reasoning behind decision-making. Post hoc trains a DL and consequently make an attempt to clarify the actions to explore black-box nature network to be explainable. Learning the relationship of the insight model in a lucid and simple manner is very important. Many recent methods include post hoc explanation like interaction, importance, learned and relationship of the features, and step by step saliency maps in their XAI models (Velden et al. 2022a). The first distinction of model-based vs. post hoc explanation (Fig. 8).

2.2 Model specific vs. agnostic

Model specific explanation generally use features that are very specific to a particular type of DL network and mostly limited to particular classes. Model based is a model-specific explanation method (Adadi and Berrada 2018), but the reverse is not necessarily true. A saliency mapping technique is not model-based explanation method for a few classes of CNNs, but they are (Velden et al. 2022a) post hoc.

Model agnostic is completely self-determining in the selection of network types, and it works only on the I/O of the net. By changing the input, the operator can inspect the change in the result of net in terms of explanation. By default, this agnostic is a post hoc model in nature. The scope of clarification separates between an explanation for a global vs. a local for this model-specific and agnostic's model.

A global explanation is a dataset-wide interpretation that provides overarching insights learned by a deep learning model. This type of explanation calculates the importance

of each contributing feature across the entire dataset. For instance, a global explanation may reveal how high blood pressure (BP) is potentially related to a cardiac incident or identify the significant features learned by a DL network for a particular event, along with visualizations of the corresponding features. (Olah et al. 2017).

A local explanation pertains to the interpretation of a single input, such as a particular individual's risk for a cardiac incident. For instance, a local explanation can provide an understanding of why BP is critical to the risk of cardiac arrest for a specific patient. In contrast, a global explanation depicts the relationship between BP and cardiac risk across the entire dataset. It highlights the significance of BP as a contributing feature to the likelihood of cardiac incidents, but it does not provide insights into the unique characteristics of individual patients.

2.3 Attribution based

Recent papers use attribution approaches for explaining DL network in precision diagnostics (Singh et al. 2020). The DL practitioners can use readily available tools and focus on the incremental developments to generate better understanding explanations model. The difficulty of transmission value of attribution to each input feature of a net led to the implementation of various techniques. Few examples of the visualization of attribution maps for multiple images is shown in Fig. 9. The positively and negatively influenced features to the target neurons are usually marked with different colors to identify the contributing features. The frequently used attribution methods like DeepTaylor (Montavon et al. 2017), DeepExplain (Ancona et al. 2017), LIFT, and Deep SHapley Additive eXplanations (SHAP) offer only positive facts that suitable for a specific set of responsibilities. The attribution techniques could be very helpful for unboxing the CNN-DL without major alterations.

3 XAI in MIA

In this section, the XAI procedures in practice used in MIA by computer vision (CV) and DL method is discussed. The XAI methods have been categorized into 3

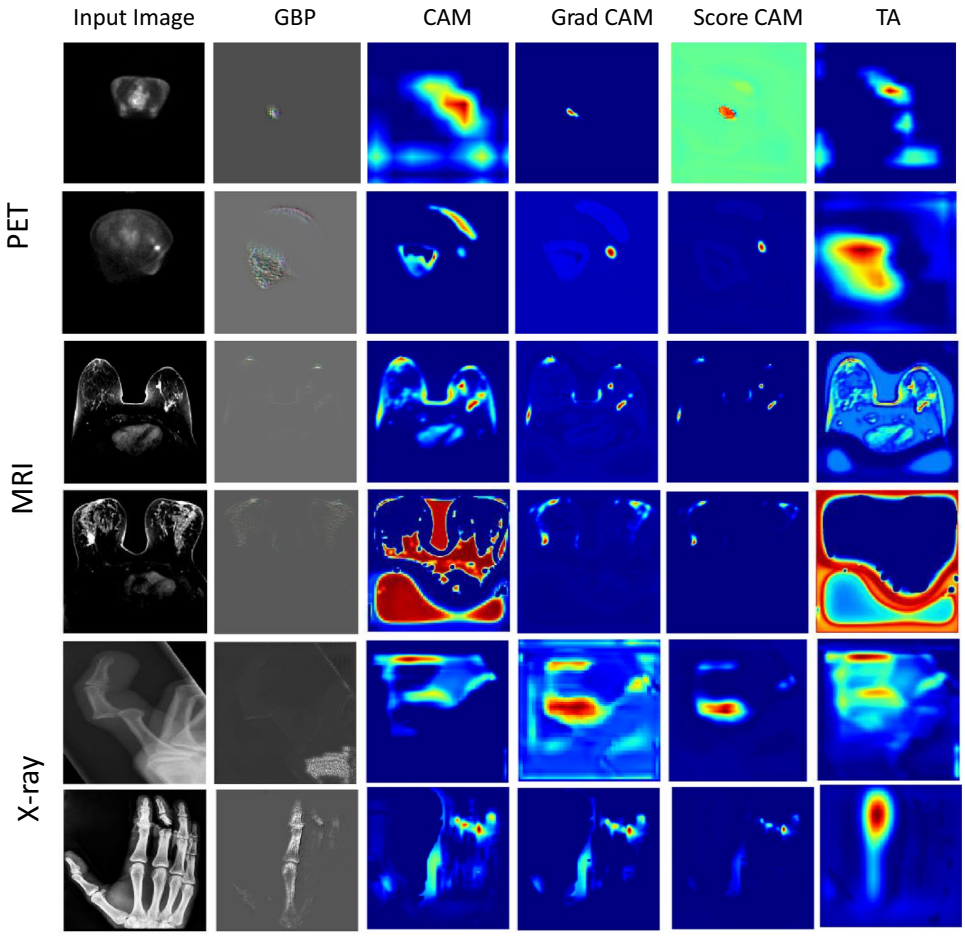

Fig. 9 Different types of XAI (GBP, CAM, Grad-CAM, Score CAM, and TA) used in MIA. The focused region are visualized in major cases

Table 2 The XAI procedures used in MIA

Techniques	Model based	Post hoc	Model specific	Model agnostic	Global	Local
Visual						
Backpropagation		■	■			■
Deconvolution		■	■			■
Guided backpropagation		■	■			■
Class activation map (CAM)		■	■			■
Gradient CAM (Grad-CAM)		■	■			■
Layer wise relevance propagation (LRP)		■	■			■
Deep Shapley Additive explanation (DeepSHAP)		■	■	■	■	■
Trainable attention	■		■			■
Perturbation based						
Occlusion sensitivity		■		■		■
Local interpretable model agnostics explanation (LIME)		■		■		■
Meaningful perturbation		■		■		■
Prediction difference analysis		■		■		■
Textual						
Image captioning (IC)	■		■			■
Testing with concept activation vectors (TCAV)		■		■	■	■
Example based						
Triplet networks	■		■		■	■
Influence function		■		■	■	
Prototypes	■		■			■

types: visual based, textual based, and example based; where each approach is graded as per the perspective of model-based, model-specific, model-agnostic, post hoc, global, and local explanation and their possible combinations (Fig. 8) (Velden et al. 2022a). The outline of regularly used methods and their families corresponding to the categorization defined is shown in Table 2.

3.1 Visual explanation

Saliency mapping (SM) is very popular form of visual explanation of XAI in MIA and shown in (Fig. 8). The SM exhibits the most significant components of image to understand the decision-making procedure. Major chunks of SM methods use backpropagation approaches, but few of SM also uses perturbation-based approaches. A distribution of recent work in medical images (MI) using SM is displayed in Table 2.

3.1.1 Guided backpropagation and deconvolution

Some of the most basic procedures to generate SM emphasized pixels and/or voxels that gave maximum effect on the output generation or testing. Examples consist of visualization of partial derivatives of the output on pixel level, deconvolution, and guided backpropagation (Springenberg et al. 1412). These methods providing specific to model, local, post hoc clarification.

3.1.2 Perturbation-based approaches

3.1.2.1 LIME The LIME (Ribeiro et al. 2016) offers local interpretability by dividing a complicated model in multiple smaller simpler versions like estimating a CNN. In technical term, the results of complicated model altered by perturbing the input image. The LIME utilizes the easier DL model to understand the planning between the perturbed I/P and altered in O/P. In images, the perturbations using super resolution was integrated without individual pixels to show the critical region for interpreting a DL technique. The LIME is used by many research scholars to show the interpretability of MIA. For example, used a LIME based method was used (Malhi et al. 2019) to explain areas that have bloody regions in gastral endoscopy images.

3.1.2.2 Meaningful perturbation Fong and Vedaldi (2017) detects variations in the projections of DL network from perturbed the input image. Initially, it was not useful for medical images but later on, simulating naturalistic was used to zoom more important perturbations, and subse-

quently to more important explanations. The local perturbations, like a steady value, degradation, noise, resolution, quality, and blurring was also used to make it robust and useful for the MI purpose as medical images are tends to have effect of noise and low resolution. For example, the perturbations by variational autoencoder (VAE) identify the pathological sections in various imaging studies such as pathology consisted of intraretinal fluid from optical coherence tomography (OCT) of eye, and pathology comprised of stroke lesions from MRI of the brain.

3.1.2.3 Prediction difference analysis (PDA) Zintgraf et al. 2017) produces saliency maps using adaptive PDA techniques. They determine how the likelihood change over on the different analysis that contribute to the estimate even for unknown pixels. In recent days, multiple scales super voxel PDA is using (Seo et al. 2019) SM generation to deliver explanation. The saliency maps were used to clarify the informative regions for classifier to differentiate between normal patients and having Alzheimer's patients.

3.1.3 Multi-instance learning

Is used for expandability visualization techniques. The training sets contain of bags of occurrences, where patches from that image correspond to the occurrences. These bags are labeled, but not the occurrences. In explainable MIA, instant based learning approach are used to determine the particular occurrences in the bag that are accountable for the classification. For example, in recent time instant learning-based explainable method was used to locate serious findings from X-ray (Schwab et al. 2020) of chest. The forecasts were covered on the source image to visualize important areas of the classifier based on its decision. Later on, multiple instances were used to visually explain the important area diabetic retinopathy grade from a fundus image.

3.2 CAM

The CAM is a weighted summation of visual shapes presence at different spatial position. In practice, CAM is replaced at the end of CNNs using global average pooling on each layer of the DL network. This CAM offers all local, post hoc and model-specific interpretation. The computer scientist used CAMs in MI and MIA in ensemble of CNN. For example, an ensemble of Inception, and ResNet (Jiang et al. 2019) was made to observe the healthy fundus images with moderate diabetic retinopathy (DR) to serious diabetic retinopathy (DR). Multi scale CAMs was proposed as medical images (MI) often contain multiple scales information. Some researchers point out that the CAMs are extremely accurate at high resolution by highlighting proper location in endoscopy images.

3.2.1 Grad-CAM

Gradient-weighted CAM (called Grad-CAM) (Selvaraju et al. 2017) is a simplified version of CAM expandability. The CAM explicitly needs global average pooling (GAP) and the Grad-CAM needs CNN to generate post hoc local explanation. An extension version of grad CAM was developed between guided backpropagation and Grad-CAM using element wise multiplication, and this is known is guided grad CAM. All these CAMs are now being used by MIA researchers to show the exploitability and interpretability of the model. For example, small bowel enteropathies on histology (Kowsari et al. 2020) and presence of a brain tumor from MRI were shown in the present study (Windisch et al. 2020) using the grad CAMs.

3.3 LRP

The LRP iteratively backpropagates its DL output (typically class between 0 and 1) in each iteration and LRP sends an importance score to every input neuron from its immediate layers. As per the law of conservation, these importance scores must be equal with the total importance score of its foundation neuron. For example, LRP was used for detecting locations that are accountable for Alzheimer's disorder from brain magnetic resonance images (Böhle et al. 2019; Roy et al. 2017a) They found LRP-guided backpropagation was better in identifying regions of Alzheimer's compared to saliency maps.

3.4 Deep SHapley Additive eXplanations (deep SHAP)

The Deep SHAP taken a notion of Shapley values from game theory, where Shapley values define the trivial influence of every feature to reach decision (Ho et al. 2021). Deep Shapley is a versatile method for interpreting the predictions of complex machine learning models, as it is not limited to any particular type of model architecture. This model-agnostic approach is made possible by a unified framework that can handle a wide variety of models, including convolutional neural networks, recurrent neural networks, and transformers. Moreover, Deep Shapley provides a measure of uncertainty for each Shapley value, which is crucial for assessing the reliability of the interpretation. However, Shapley values require consideration of many different transformations, which can make the computation of Shapley values resource-intensive. To address this, a faster Deep Shapley technique was proposed specifically for estimating the Shapley values for convolutional neural networks. Deep Shapley is widely used to identify the specific parts of an image that have a positive or negative contribution to a decision-making system. This makes it an invaluable tool for identifying

and understanding the features that are most important for a given task (Velden et al. 2022b).

3.5 Trainable attention (TA)

This TA technique emphasized where and what fraction the DL pays interest to the input images for prediction and classification. The Trainable attention also further expand important areas and suppress the inappropriate areas of the images. A grid TA was introduced in MI (Schlemper et al. 2019) to capture the functional information of image. The attention coefficients of DL network were used to describe the focused areas of the image. This method can be used for various DL net like UNET and different variation of VGG with very high performance.

3.6 Textual explanation

Textual explanations include comparatively straightforward characteristics. There are mainly 2 categories of textual XAI methods: image captioning (IC) and testing with concept attribution vectors (TCAV).

3.6.1 The textual explanation was suggested by an end-to-end IC structure in Singh et al. (2019)

They (Singh et al. 2019) have utilized human generated sentences as a reference for training purpose, and the precision of word N-grams metric for evaluation. An image captioning method for explainable purpose was used (Singh et al. 2019) on chest X-rays data set. As expected, higher accuracy was achieved in the generated radiology report when both radiology variant global vectors and global vectors were used to train the LSTM DL network instead of just global vectors.

3.6.2 TCAV

Offer high-level explainable vision to human (Kim et al. 2019). The testing with TCAV procedure works on user-specific sets of examples. The TCAV evaluate the sensitivity and specificity of a pretrained model to such ideas using CAV. The practicability of TCAV on a MIA is to relating doctor annotations with the automation annotation such as chest X-ray, and cardiac disease (Clough et al. 2019) by classifying the latent space.

3.7 Example based

The example-based explanation (Velden et al. 2022a) gives examples involving the previous data points that are currently being studied. For example, tumor biopsy of a patient may have some resemblance with the previous patient analyzed by the medical practitioner. The previous knowledge will help to enhance the clinical decision and the similar things happened when we explain DL network to came to a decision. The example-based interpretability optimizes number of hidden layers in DL network that have similar features to each other's layers in the latent space. The following are some examples based explainable network:

3.7.1 Triplet network

A triplet network is an example-based interpretable technique that comprises of three similar networks with mutual parameters. The network computes two values consist of the *L2* spaces between the representations in latent space of these input samples by supplying 3 input samples. This allows unsupervised assessment of samples after learning of useful representations from the latent space. The study on colorectal cancer histology (Peng et al. 2019) used a novel approach for explaining the decision-making process of a neural network. By analyzing similar images, the network was able to make a decision, and the researchers provided a visual explanation similar to CAM. This method was later applied to other medical imaging fields, including radiology, pathology, and cardiovascular imaging. Additionally, the researchers introduced query and search activation maps, which allow for the extraction of features that are relevant to explainability, based on lesion embeddings. This approach provides a more comprehensive and interpretable explanation of the neural network's decision-making process and can help identify important features for diagnosis and treatment.

3.7.2 Influence functions

An influence functions can be implemented to give explanation on the important and non-important features based on the decision made using a DL network (Roy and Bandyopadhyay 2013). So that the authors could examine what would happen if a training set were not available or altered. This implementation of IF is like SHAP, both allow inexpensive computation for significance feature selection.

4 Evaluation of XAI methods

In any computer science applications, quantitative and qualitative validation plays a vital role to measure the performance of XAI model. A professional healthcare professional should evaluate the attention map created to distinguishes highly diagnostic importance region of an image. In general process to evaluate XAI from model-based and healthcare perspectives is given in in Fig. 10. For example, healthcare practitioners can be inquired to distinguish among the cases that involved clinical images and those images with visual

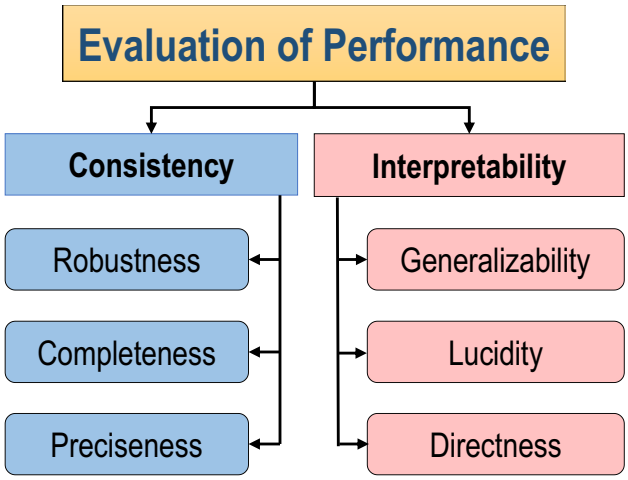

Fig. 10 Evaluation of model explainability. The presentation transparency, easiness and generalizability are the important criteria for the evaluation (Giuste et al. 2022)

explanation. However, managing such findings are costly and time consuming, especially for medical images (Giuste et al. 2022). The concept of the occlusion (Zeiler and Fergus 2014) was first conducted on input images, where gray square using DL was systematically included in the images. Another procedure of pixel flipping can also be applied for the evaluation for the features importance interpretation method. On the other hand, Randomized Input Sampling for Explanation (RISE) (Petsiuk et al. 1806), could be very much useful for the perturb method. Apart from these studies, R^2 for global surrogate model, could also be helpful to assess the model explainability. An adversarial attack can also be introduced (Lin et al. 2019) to estimate the strength of explainability of XAI to detect if their backdoor triggers present or not. Researchers also used saliency map output using intersection over union, recovery rate, and difference quantifiable method (Table 3).

5 Applications

The major work in summarized manner of XAI applied in MI are shown in Table 4.

6 Arguments for and against of XAI

The XAI methods are not error free like other techniques. In this section, we will discuss about pros and cons, and their several opportunities. We have chosen the following key characteristics to discuss for and against several XAI.

6.1 Ease of use

The post hoc agnostic XAI gets the highest rank in ease-of-use characteristics, and usually comprise of perturbation-centered visual explainability. These procedures could be used as a 'plug-and-play' on any DL network to deliver a visual clarification. This post hoc-based procedures usually have lowest level of ease of use as justifications are inserted within the design of DL network.

6.2 Validity

The validity is defined by the end users based on the explanation accuracy. In recent days, a visual explanation (Arun et al. 2008) was used to detect the pneumothorax and corresponding localization. The Grad-CAM demonstrated the highest validity compared to other three: backpropagation (BP), guided BP, and guided Grad-CAM that shown on the radiograph work. When it comes to textual explanation in XAI, the validity of experiments is determined by how well the explanation aligns with the referenced text. On the other hand, in example-based XAI, the validity of experiments is assessed by evaluating relevant characteristics against clinicopathological patient characteristics. However, there have been relatively fewer rigorous experiments conducted on textual and example-based validation, compared to visual explanation. Despite this, more research has been conducted on visual explanation in XAI, as it is often more intuitive and easier to understand for users. This has led to the development of various visual explanation methods that can help improve the interpretability and transparency of machine learning models.

6.3 Robustness

The robustness of XAI is evaluated by modifying some attributes of DL framework purposely and the determining the impact of these alterations to the given explanation. In general, parameter and data randomization tests are used with visual explanation comparisons from a trained deep CNN with a random initialized untrained deep CNN on the same framework. If the two interpretations differ significantly, then the visual explanation is considered vulnerable to the attributes and parameters of the Deep CNN. For instance, in a study on Alzheimer's disease classification from MRI brain scans, researchers assessed the robustness of visual interpretation by using guided backpropagation, layer-wise significance, and occlusion sensitivity over multiple training runs. Among these methods, layer-wise and guided backpropagation produced the most consistent visual interpretations (Arun et al. 2008; Kraus et al. 2021). This approach can help ensure that the XAI method is reliable and

Table 3 Research contribution in XAI on different body part

Organ	References	Technique	Modality	Contribution
Bone	Pierson et al. (2021)	CAM	X-ray	The goal of the model is to use X-ray images to forecast the level of knee injury and the amount of pain experienced
	Fu et al. (2021)	Attention	CT	The study introduces a multimodal spatial attention module (MSAM). It uses an attention mechanism to focus on the area of interest
Chest	Fan et al. (2022)	Grad-CAM	Ultrasound	The paper proposes a semi-supervised model based on attention mechanism and disentangled. It then uses Grad-CAM to improve model's explainable
	Lu et al. (2022)	Grad-CAM	CT	It proposes a neighboring aware graph neural network (NAGNN) for COVID-19 detection based on chest CT images
	Moncada-Torres et al. (2021)	SHAP	CT	In this paper, it compares the performance of different ML methods (RSFs, SSVMs, and XGB and CPH regression) and uses SHAP value to interpret the models
	Abeyagunasekera et al. (2022)	LIME, SHAP	X-ray	The study proposes a unified pipeline to improve explainability for CNN using multiple XAI methods
Lung	Born et al. (2021)	CAM	Ultrasound, X-ray	It uses three kinds of lung ultrasound images as datasets, and two networks, VGG-16 and VGG-CAM, to classify three kinds of pneumonia
	Jia et al. (2021)	CAM	X-ray, CT	The study improves two models, one of them based on MobileNet to classify COVID-19 CXR images, the other one is ResNet for CT image classification
	Song et al. (2021)	CAM	CT	It selects healthy and COVID-19 patient's data for training DRE-Net model
	Wang et al. (2021b)	Grad-CAM	CT	It proposes a method of deep feature fusion. It achieves better performance than the single use of CNN
	Wang et al. (2021c)	Grad-CAM	X-ray	It provides a computer-aided detection, which is composed of the Discrimination-DL and the Localization-DL, and uses Grad-CAM to locate abnormal areas in the image
	Wu et al. (2021)	Grad-CAM	CT	It shows a classifier based on the Res2Net network. The study uses Activation Mapping to increase the interpretability of the overall Joint Classification and Segmentation system
	Haghanifar et al. (2022)	Grad-CAM, LIME	X-ray	This work provides a COVID-19 X-ray dataset, and proposes a COVID-CXNet based on CheXNet using transfer learning
	Punn and Agarwal (2021)	Grad-CAM, LIME	X-ray, CT	It compares five DL models and uses the visualization method to explain NASNetLarge
	Alsinglawi et al. (2022)	SHAP	Electronic Health Record	The study introduces a predictive length of stay framework to deal with imbalanced EHR datasets
	Duell et al. (2021)	SHAP, LIME, Scoped Rules	Electronic Health Record	The study provides a comparison among three feature-based XAI techniques on EHR dataset. The results show that the use of these techniques cannot replace human experts

Table 3 (continued)

Organ	References	Technique	Modality	Contribution
Breast	Shen et al. (2021)	CAM	X-ray	It proposes a globally aware multiple instance classifier (GMIC) that uses CAM to identify the most informative regions with local and global information
	Wang et al. (2021d)	Attention heat map	X-ray	It provides the triple-attention learning A3 Net model to diagnose 14 chest diseases
Brain	Xie et al. (2020)	CAM	ultrasound	The purpose of this research is to develop computer-aided diagnosis algorithms for five common fetal brain abnormalities, which may provide assistance to doctors for brain abnormalities detection in antenatal neuro sonographic assessment
	Windisch et al. (2020)	Grad-CAM	MRI	Making the decisions of a network more explainable helped to identify potential bias and choose appropriate training data
	Gao et al. (2019)	CAM	MRI	Brain regions with significant differences between men and women are found with the proposed method, which can be used for future brain imaging studies
	Liao et al. (2020)	Grad-CAM	MRI	an effective framework based on a deformable convolutional neural network for fetal brain age prediction

can provide accurate and consistent explanations even in the face of changes or variations in the DL framework or dataset.

6.4 Computational cost

The computing cost is considered as per time and space requirements to execute the program and or algorithms. The model prognostics explainable methods are costlier in general among all other method. The BP-based methods usually get a single pass back through the DL framework is comparatively faster for visual explanation, whereas rigorous perturbation of inputs medical images to quantify the impact of perturbations on the outcome. The computation cost of post hoc textual TCAV and example-based have not been explored much yet.

6.5 Necessity of fine-tuning

Major explainable work must be fine-tuned before use as they are sensitive toward application. For example, the exact layer is needed to examine the activation of Grad-CAM based explainable method and this part could be modified by the programmer themselves. In another example, one requires to decide samples from the training data to compute the background signal in SHAP. In significant perturbation, the user must define the kind of perturbation procedure is considered for the best. The description of influence functions is required is to be quantify the textual part for post hoc method.

6.6 Open-source availability

The source code of major explainable AI methods is accessible in the open-source forum. There are many techniques also executed in captum.ai like XAI packages.

7 Discussion

The AI method has remarkable influence on healthcare and medicine patients will have maximized benefit from it. However, in spite of its instinctive demand, XAI in patient level monitoring and diagnostic decision-making is not progressed up to the mark. While XAI provides better visualization and interpretation, there is currently no normative validation to support its black-box prognostications without examination. XAI methods are currently seen as more of a methodological explanation model than an explainable healthcare model. (Ghassemi and Oakden-Rayner 2021).

In general, explainability gives better visualization and interpretation, but current XAI does not support any normative validation to allow their black-box prognostications without examination. The useful approach to rationalize the AI-based decision-making systems (Roy et al. 2018) must be detailed, cautious, meticulous safety and through multiple validation efforts, instead of local clarifications from a complex AI method. Although, few of XAI is adaptive at assessment and authenticating in several kinds of black-box systems, as many drugs and device's function, in effect, as

Table 4 The diverse applications areas of XAI in MI (Singh et al. 2020)

Type	Procedure	DL method	Area applied	Imaging
Attribution	Gradient, GBP, LRP, Occlusion (Eitel and Ritter 2019)	3D CNN	Alzheimer disease recognition	MRI (T2) of Brain
	Integrated Gradient (Jogani et al. 2022)	2D CNN models	Classify lung cancer	histopathological images
	Grad-CAM, GBP (Pereira et al. 2018)	Modified CNN	Identifying Different grad of tumor	Brain MRI
	Salient Region (Miwa et al. 2023)	Deep learning model—CNNs	To know and explain predictors	Multimodal images—in synthetic and real datasets
	IG (Shrikumar et al. 2017)	Inception-v4	DR grading	Fundus images
	Multiple gradient CAM (Zeineldin et al. 2022b)	3D CNN	for automatic brain glioma grading	MRI-T1w, T2w, T1Gd, and FLAIR, from multiple source
	EG (Yang et al. 2019)	Modified CNN	Target lesion segmentation	Retinal OCT
	Grad-CAM (Yamashita et al. 2018)	Ensemble CNN model	oxygen requirement forecast for COVID patients	chest radiograph
	Smooth Grad and IG (Papanastasopoulos et al. 2020)	ALexNet	Estrogen receptor level	MRI of breast
	Modified SHAPLEY (Wang et al. 2021a)	Deep unfolding high-resolution Net	to achieve feature detection	ultrasound images
	SM (Lévy and Jain 2016)	ALexNet	categorization of breast mass	MR T2 and T1
	Grad-CAM, LIME, and SHAP(Bhandari et al. 2022)	Light-weight CNN	Explanatory categorization into 3 class	Chest X-ray images
	Grad-CAM, SHAP (Young et al. 2019)	Inception Net	Malignancy identification	Skin images
	Guided Grad-CAM (Xu et al. 2021)	Residual Learning network	Diagnosing Fungal Keratitis	Microscopy images
	Activation maps (Molle et al. 2018)	Modified CNN	Lesion identification and categorization	Dermatology images
	Grad-CAM (Gozzi et al. 2022)	Transfer Learning	Detection multi-thorax anomaly	Chest X-ray
	DeepDreams (Couteaux et al. 2019)	Custom CNN	Detection and segmentation of tumor	Liver CT images
	LIME (Ribeiro et al. 2016)	Semi-Supervised DL	Detection of COVID-19 from chest	Radiograph
	Guided BackProp, DeepLift (Jin et al. 2023)	Fully supervised DL	Clinical decision-making system	Multi-modal images
	Class activation mapping (CAM) (Kim et al. 2022)	Based on quantitative similarity	Auto-labeling	chest X-ray images
	GSInquire, GBP, Activation (Wang and Wong 2020)	COVIDNet CNN	COVID-19 detection	X-ray images
	Fine-grained textual explanations (Ahmed et al. 2022)	Arbitrary Deep Learning (DL)	For computer-aided detection of skin lesions	Skin images
Attention	Mapping images to report (Zhang et al. 2017)	CNN and LSTM	Bladder cancer	Tissue image
	Occlusion based, and attention based (Hu et al. 2022)	Deep learning	Interventional tools in response to COVID-19	X-ray and skin
	Shape attention stream (Sun et al. 2020)	U-net	Cardiac volume measurement	MRI

Table 4 (continued)

Type	Procedure	DL method	Area applied	Imaging
Concept vectors	TCAV (Kim et al. 2017)	Inception	DR detection	Fundus images
	TCAV with RCV (Graziani et al. 2018)	ResNet101	Breast tumor identification	Lymph node images
	ExplAIn (Quellec et al. 2021)	trained from end to end with image supervision only	Diabetic retinopathy diagnosis	Color Fundus Photography
	UBS (Yeche et al. 2019)	SqueezeNet	Breast mass classification	Mammography images
	SHAP (SHapley Additive exPlanations) (Shibu et al. 2023)	1-D CNN and LSTM	Predict brain states	Functional near-infrared spectroscopy
Expert knowledge	Domain constraints (Pisov et al. 2019)	U-net	Brain MLS estimation	Brain MRI
	Rule-based classification (Ge et al. 2023)	DenseNet-121	Severity of gastroesophageal reflux disease	Endoscopic images
	Rule-based segmentation (Zhu and Ogino 2019)	VGG-16	Lung nodule detection	Lung CT
Similar images	GMM and atlas (Stano et al. 2019)	3D CNN	MRI classification	3D MNIST, Brain MRI
	LIME (Abir et al. 2022)	Transfer Learning Method	Diagnosing and Anticipating Leukemia	Microscopic images
	Triplet loss, kNN (Codella et al. 2018)	AlexNet with shared weights	Melanoma detection	Dermoscopy image
	Monotonic constraints (Silva et al. 2018)	DNN with two streams	Melanoma detection	Dermoscopy image

black boxes but the number is not too high (Patrício et al. 2205). An explainable method must be capable of serve as a helpful tool for detailed analysis in addition to algorithmic inventory, on which the suitable audience and critic should not be the subject's expert of AI, but rather the developers, auditors, and controllers from these associated systems. We have summarized the incorporation of XAI in academic, and clinical research settings, along with its benefits in Fig. 11.

Some other points to leverage XAI model to enable the understanding of precision diagnostics solutions (Giuste et al. 2022) such as outlook, critique, limitations, and future trends are described below.

7.1 Outlook

The future XAI tendencies in MIA will be more focused toward biological interpretation. Various investigators are predicting different biological significance from MI using DL network, but interesting biological explanation still uncovered yet (Velden et al. 2022a). Then, XAI will be very useful aid to the medical practitioners in distinguishing unknown information from MIA. In recent study shows that the diagnosis of tuberculosis from plain radiograph had better precision when evaluating with a visual XAI model compared to without XAI (Rajpurkar et al. 2020; Samek and Müller 2019).

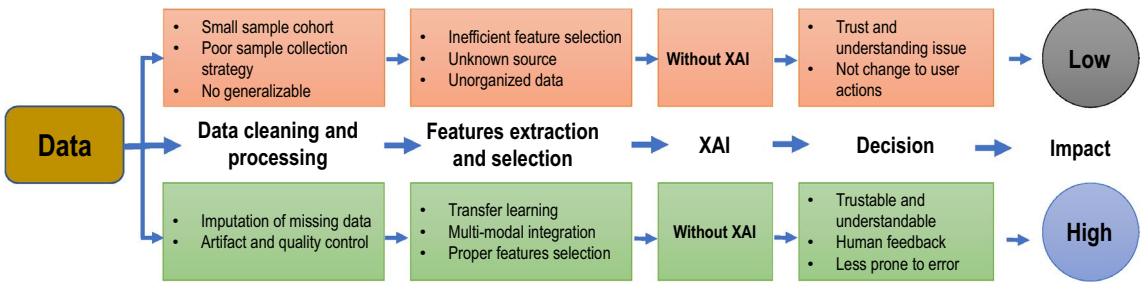

Fig. 11 The insights and common issues of XAI in industry-academia research (Giuste et al. 2022)

7.2 Critique on XAI

Sometimes, only a technical explanation does not deliver sufficient information to unbox the black box in medical images. For example, it is difficult to distinguish between a saliency map with the highest and lowest probability class, as they look very close (Velden et al. 2022a). Therefore, the use of interpretable method is more suitable than the use of XAI (Rudin 2019). Several other critical aspects like the robustness, and interpretability with respect to the healthcare need to be more focused during the implementation. The XAI must explain the 100% truth about the model, otherwise it will not be very effective model.

7.3 Challenges and limitations of revolutionizing healthcare industry

The XAI model predominantly focused on the explainability of either each layer of the Deep CNN, and/or visualization of object of interest and/or set of features that contributes to the models. But does not explore enough on the reason of the outcome or to solve the needs of healthcare professionals. For example, identification and reason of triple negative breast cancer from radiological imaging is very much needed using any AI-based method. The complete path of diagnostics from radiological images is still missing in XAI (Velden et al. 2022a). Revolutionizing the healthcare industry is a complex endeavor with various challenges and limitations. While advancements in technology and medical research have developed significantly, still there are several limitations that must be overcome. Here are some of the key challenges and considerations (Kraus et al. 2021):

1. *Regulatory intervention*: Healthcare is heavily regulated in most countries to ensure patient safety and data security. Navigating the regulatory landscape, obtaining approvals, and complying with changing regulations can be time consuming and costly.
2. *Data privacy and security*: The healthcare industry deals with sensitive patient data, making privacy and security paramount. Ensuring the protection of electronic health records (EHRs) and other medical data is an ongoing challenge, especially with the rise of cyber threats.
3. *Interoperability*: Many healthcare systems use different software and standards, making it difficult for different systems to communicate and share data seamlessly. Achieving interoperability is essential for providing comprehensive patient care.
4. *Resistance to change*: Healthcare professionals and institutions may resist adopting new technologies or changing established practices. Overcoming this resistance and ensuring that healthcare workers are trained to use new tools and systems is crucial.
5. *Ethical concerns*: Advancements in healthcare, such as gene editing and AI diagnostics, raise ethical questions about who should have access to certain technologies and how they should be used.
6. *Patient engagement and education*: Educating patients about new healthcare options and engaging them in their own care can be challenging. Many patients may not have the knowledge or resources to make informed decisions about their health.

The road ahead in revolutionizing the healthcare industry involves addressing these challenges while leveraging emerging technologies and innovative approaches. Some of the key strategies include:

1. *Collaboration*: Encouraging collaboration between healthcare providers, researchers, technology companies, and policymakers can help overcome many of the challenges mentioned above.
2. *Healthcare telemedicine and remote monitoring*: Giving patients access to their health information enables them to take charge of their health. Expanding telemedicine and remote monitoring capabilities can improve access to care, especially in remote areas.
3. *Education and training*: Providing ongoing education and training for healthcare professionals is essential for ensuring they are prepared to use new technologies effectively.
4. *Healthcare policy reform*: Policymakers must work to create a regulatory environment that supports innovation while ensuring patient safety and data security.

Revolutionizing the healthcare industry is a long-term endeavor that requires a multidisciplinary approach and a commitment to addressing the challenges that arise along the way. It's an ongoing process that should prioritize improving patient outcomes, reducing costs, and increasing access to high-quality care.

7.4 Future trends in localization

The image segmentation can be extremely explainable. In the present scenario, isolating regions and identification of multi-thorax disease from chest X-ray (Kabiraj et al. 2022; Chakraborty et al. 2023; Pal et al. 2023) images. One example of XAI technique used in segmentation to improve COVID-19 identification. The XAI framework was trained to recognize ground glass region on the smallest pixel level (Giuste et al. 2022). However, the output produced in segmentation does not produce any insight and clarity on why the model made the outcomes it did.

8 Summary and conclusion

This study reviewed 113 research articles where explainability in MIA through the use of AI and categorized to anatomical behavior and technique. In this paper, we studied XAI methods that enhanced its domain adoption based on lessons understood from medical images (MI) and also future trends are given with definite insights. This paper reviewed how to assess XAI, current criticisms on XAI, and future viewpoints for XAI in MIA using DL. In XAI, the clinicians, data scientists, and healthcare consultants can take advantages by the model clarity, and from explanation empowered by XAI to unbox the black-box decision-making systems. This also improves the credibility and responsibility of AI-based method and encourage their acceptance in the clinical healthcare workflow.

The use of explanations beyond visualization is one of the current challenging tasks. Current XAI permit us to increase understandings into operation. But these current studies are still restricted in various circumstances. The visualization using heatmap is a first order information which shows the important features for estimate, but relationship among those features is still not clear or not focused yet. In many medical applications, such relationship can solve many tasks specific problem or cause of the disease. The low level of abstraction of XAI is another limitation. The AI scientists must need to interpret the XAI to understand actions of models to make it sense for medical practitioners. Thus, the meta interpretation/explainable is needed that accumulate evidence from low level model's behavior and increase the human understandable level. The creation of more innovative meta-XAI is a worthwhile topic of research for future. However, the optimization of XAI for medical decision-making and diagnostics use is still a challenge that needs more advance level research. Finally, the use of XAI beyond visualization with reduced model performance is a challenge.

Acknowledgements This research work was supported by the RFIER—Jio Institute research "Computer Vision in Medical Imaging (CVMI)" project under the "*AI for All*" research center."

Author contributions SR wrote the main manuscript. TM and DP revised and added many sections of the manuscript. All authors prepared Figures and Tables. All authors reviewed the manuscript.

Funding The work is supported by the RFIER—Jio Institute research grant # 2022/33185004.

Data availability Not applicable.

Declarations

Conflict of interest Authors have no conflict to declare.

Ethical approval Not applicable.

References

Abeyagunasekera SHP, Perera Y, Chamara K, Kaushalya U, Sumathipala P, Senaweera O (2022) LISA: Enhance the explainability of medical images unifying current XAI techniques. In Proceedings of the 2022 IEEE 7th International Conference for Convergence in Technology (I2CT), Mumbai, India, 7–9 April 2022; pp. 1–9

Abir WH, Uddin MF, Khanam FR, Tazin T, Khan MM, Masud M, Aljahdali S (2022) Explainable AI in diagnosing and anticipating leukemia using transfer learning method. Comput Intell Neurosci. https://doi.org/10.1155/2022/5140148

Adadi A, Berrada M (2018) Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access 6:52138–52160

Alsinglawi B, Alshari O, Alorjani M, Mubin O, Alnajjar F, Novoa M, Darwish O (2022) An explainable machine learning framework for lung cancer hospital length of stay prediction. Sci Rep 12:607

Ancona M, Ceolini E, Öztireli C, Gross M (2017) "Towards better understanding of gradient-based attribution methods for deep neural networks." arXiv preprint arXiv:1711.06104

Arun N, Gaw N, Singh P, Chang K, Aggarwal M, Chen B (2020) "Assessing the (Un) trustworthiness of saliency maps for localizing abnormalities in medical imaging. arXiv." arXiv preprint arXiv:2008.02766

Bhandari M, Shahi TB, Siku B, Neupane A (2022) Explanatory classification of CXR images into COVID-19, Pneumonia and Tuberculosis using deep learning and XAI. Comput Biol Med 150:106156

Böhle M, Eitel F, Weygandt M, Ritter K (2019) Layer-wise relevance propagation for explaining deep neural network decisions in MRI-based Alzheimer's disease classification. Front Aging Neurosci 11:194

Born J, Wiedemann N, Cossio M, Buhre C, Brändle G, Leidermann K, Goulet J, Aujayeb A, Moor M, Rieck B et al (2021) Accelerating detection of lung pathologies with explainable ultrasound image analysis. Appl Sci 11:672

Chakraborty S, Kumar K, Reddy BP, Meena T, Roy S (2023) An Explainable AI based Clinical Assistance Model for Identifying Patients with the Onset of Sepsis," 2023 IEEE 24th International Conference on Information Reuse and Integration for Data Science (IRI), Bellevue, WA, USA pp. 297–302. https://doi.org/10.1109/IRI58017.2023.00059

Clough JR, Oksuz I, Puyol-Antón E, Ruijsink B, King AP, Schnabel JA (2019) "Global and local interpretability for cardiac MRI classification." In Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part IV 22, pp. 656–664. Springer International Publishing

Codella NC, Lin CC, Halpern A, Hind M, Feris R, Smith JR (2018) Collaborative human-AI (CHAI): evidence-based interpretable melanoma classification in dermoscopic images. Understanding and interpreting machine learning in medical image computing applications. Springer, Cham, pp 97–105

Couteaux V, Nempont O, Pizaine G, Bloch I (2019) Towards interpretability of segmentation networks by analyzing DeepDreams. Interpretability of machine intelligence in medical image computing and multimodal learning for clinical decision support. Springer, Cham, pp 56–63

Duell J, Fan X, Burnett B, Aarts G, Zhou SMA (2021) Comparison of Explanations Given by Explainable Artificial Intelligence Methods on Analysing Electronic Health Records. In Proceedings of the 2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI), Athens, Greece, 27–30 July 2021; pp. 1–4

Eitel F, Ritter K (2019) Alzheimer's Disease Neuroimaging Initiative (ADNI). Testing the Robustness of Attribution Methods for Convolutional Neural Networks in MRI-Based Alzheimer's Disease Classification. In Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support, ML-CDS 2019, IMIMIC 2019; Lecture Notes in Computer Science; Suzuki, K., et al., Eds.; Springer: Cham, Switzerland, 2019; Volume 11797

Fan Z, Gong P, Tang S, Lee CU, Zhang X, Song P, Chen S, Li H (2022) Joint localization and classification of breast tumors on ultrasound images using a novel auxiliary attention-based framework. arXiv 2022. arXiv:2210.05762

Fong RC, Vedaldi A (2017) "Interpretable explanations of black boxes by meaningful perturbation." In Proceedings of the IEEE international conference on computer vision, pp. 3429–3437

Fu X, Bi L, Kumar A, Fulham M, Kim J (2021) Multimodal spatial attention module for targeting multimodal PET-CT lung tumor segmentation. IEEE J Biomed Health Inf 25:3507–3516

Gao K, Shen H, Liu Y, Zeng L, Hu D (2019) "Dense-CAM: Visualize the Gender of Brains with MRI Images," 2019 International Joint Conference on Neural Networks (IJCNN), Budapest, Hungary, pp. 1–7. https://doi.org/10.1109/IJCNN.2019.8852260

Ge Z, Wang B, Chang J, Yu Z, Zhou Z, Zhang J, Duan Z (2023) Using deep learning and explainable artificial intelligence to assess the severity of gastroesophageal reflux disease according to the Los Angeles Classification System. Scand J Gastroenterol. https://doi.org/10.1080/00365521.2022.2163185. (**Epub ahead of print. PMID: 36625026**)

Ghassemi M, Oakden-Rayner L, Beam AL (2021) The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit Health 3:e745–e750

Giuste F, Shi W, Zhu Y, Naren T, Isgut M, Sha Y, Tong L, Gupte M, Wang MD (2022) Explainable artificial intelligence methods in combating pandemics: a systematic review. IEEE Reviews in Biomedical Engineering, vol. XX, no. X

Gozzi N, Giacomello E, Sollini M, Kirienko M, Ammirabile A, Lanzi P, Loiacono D, Chiti A (2022) Image embeddings extracted from CNNs outperform other transfer learning approaches in classification of chest radiographs. Diagnostics (basel) 12(9):2084. https://doi.org/10.3390/diagnostics12092084. (**PMID:36140486;PMCID:PMC9497580**)

Graziani M, Andrearczyk V, Müller H (2018) Regression concept vectors for bidirectional explanations in histopathology. Understanding and interpreting machine learning in medical image computing applications. Springer, Cham, pp 124–132

Haghanifar A, Majdabadi MM, Choi Y, Deivalakshmi S, Ko S (2022) COVID-cxnet: detecting COVID-19 in frontal chest X-ray images using deep learning. Multimed Tools Appl 81:30615–30645

Ho T-H, Park S-E, Xuanming Su (2021) A bayesian level-k model in n-person games. Manag Sci 67(3):1622–1638

Hu B, Vasu B, Hoogs A (2022) "X-MIR: EXplainable Medical Image Retrieval," 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 2022, pp. 1544–1554, doi: https://doi.org/10.1109/WACV51458.2022.00161

Jia G, Lam HK, Xu Y (2021) Classification of COVID-19 chest X-ray and CT images using a type of dynamic CNN modification method. Comput Biol Med 134:104425

Jiang H, Yang K, Gao M, Zhang D, Ma H, Qian W (2019) "An interpretable ensemble deep learning model for diabetic retinopathy disease classification." In 2019 41st annual international conference of the IEEE engineering in medicine and biology society (EMBC), pp. 2045–2048. IEEE

Jin W, Li X, Fatehi M, Hamarneh G (2023) Generating post-hoc explanation from deep neural networks for multi-modal medical image analysis tasks. MethodsX. https://doi.org/10.1016/j.mex.2023.102009

Jogani V, Purohit J, Shivhare I, Shrawne SC (2022) "Analysis of Explainable Artificial Intelligence Methods on Medical Image Classification." arXiv preprint arXiv:2212.10565

Kabiraj A, Meena T, Reddy PB, Roy S (2022) "Detection and Classification of Lung Disease Using Deep Learning Architecture from X-ray Images." In Advances in Visual Computing: 17th International Symposium, ISVC 2022, San Diego, CA, USA, October 3–5, 2022, Proceedings, Part I, pp. 444–455. Cham: Springer International Publishing

Kim B, Wattenberg M, Gilmer J, Cai C, Wexler J, Viegas F, Sayres R (2017) Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). arXiv 2017. arXiv:1711.11279

Kim ST, Lee JH, Ro YM (2019) "Visual evidence for interpreting diagnostic decision of deep neural network in computer-aided diagnosis." In Medical Imaging 2019: Computer-Aided Diagnosis, vol. 10950, pp. 139–147. SPIE

Kim D, Chung J, Choi J, Succi MD, Conklin J, Longo MGF, Ackman JB, Little BP, Petranovic M, Kalra MK, Lev MH, Do S (2022) Accurate auto-labeling of chest X-ray images based on quantitative similarity to an explainable AI model. Nat Commun 13(1):1867. https://doi.org/10.1038/s41467-022-29437-8. (**PMID:35388010;PMCID:PMC8986787**)

Kowsari K, Sali R, Ehsan L, Adorno W, Ali A, Moore S, Amadi B, Kelly P, Syed S, Brown D (2020) Hmic: Hierarchical medical image classification, a deep learning approach. Information 11(6):318

Kraus S, Schiavone F, Pluzhnikova A, Invernizzi AC (2021) Digital transformation in healthcare: analyzing the current state-of-research. J Bus Res 123:557–567

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

Liao L et al. (2020) "Multi-branch deformable convolutional neural network with label distribution learning for fetal brain age prediction," 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), Iowa City, IA, USA, 2020, pp. 424-427. https://doi.org/10.1109/ISBI45749.2020.9098553.

Lin Z, Li S, Ni D, Liao Y, Wen H, Jie Du, Chen S, Wang T, Lei B (2019) Multi-task learning for quality assessment of fetal head ultrasound images. Med Image Anal 58:101548

Lu S, Zhu Z, Gorriz JM, Wang SH, Zhang YD (2022) NAGNN: Classification of COVID-19 based on neighboring aware representation from deep graph neural network. Int J Intell Syst 37:1572–1598

Lucieri A, Bajwa MN, Braun SA, Malik MI, Dengel A, Ahmed S (2022) ExAID: a multimodal explanation framework for computer-aided diagnosis of skin lesions. Comput Methods Programs Biomed 215:106620

Malhi A, Kampik T, Pannu H, Madhikermi M, Främling K (2019) "Explaining machine learning-based classifications of in-vivo gastral images." In 2019 Digital Image Computing: Techniques and Applications (DICTA), pp. 1–7. IEEE

Meena T, Kabiraj A, Reddy PB, Roy S (2023) "Weakly Supervised Confidence Aware Probabilistic CAM multi-Thorax Anomaly Localization Network," 2023 IEEE 24th International Conference on Information Reuse and Integration for Data Science (IRI), Bellevue, WA, USA, pp. 309–314. https://doi.org/10.1109/IRI58017.2023.00061.

Meena T, Roy S (2022) Bone fracture detection using deep supervised learning from radiological images: a paradigm shift. Diagnostics 12(10):2420

Meena T, Sarawadekar K (2023) Seq2Dense U-Net: analyzing sequential inertial sensor data for human activity recognition using

dense segmentation model. IEEE Sens J 23(18):21544–21552. https://doi.org/10.1109/JSEN.2023.3301187

Meswal H, Kumar D, Gupta A et al (2023) A weighted ensemble transfer learning approach for melanoma classification from skin lesion images. Multimed Tools Appl. https://doi.org/10.1007/s11042-023-16783-y

Miwa D, Duy VN, Takeuchi I (2023) "Valid P-value for deep learning-driven salient region." arXiv preprint arXiv:2301.02437

Moncada-Torres A, van Maaren MC, Hendriks MP, Siesling S, Geleijnse G (2021) Explainable machine learning can outperform Cox regression predictions and provide insights in breast cancer survival. Sci Rep 11:6968

Montavon G, Lapuschkin S, Binder A, Samek W, Müller K-R (2017) Explaining nonlinear classification decisions with deep taylor decomposition. Pattern Recogn 65:211–222

Olah C, Mordvintsev A, Schubert L (2017) Feature visualization. Distill 2(11):e7

Pal D, Reddy PB, Roy S (2022) Attention UW-Net: a fully connected model for automatic segmentation and annotation of chest X-ray. Comput Biol Med 150:106083

Pal D, Meena T, Roy S (2023) "A fully connected reproducible SE-UResNet for multiorgan chest radiographs segmentation," 2023 IEEE 24th International Conference on Information Reuse and Integration for Data Science (IRI), Bellevue, WA, USA, 2023, pp. 261–266, doi: https://doi.org/10.1109/IRI58017.2023.00052

Papanastasopoulos Z, Samala RK, Chan HP, Hadjiiski L, Paramagul C, Helvie MA, Neal CH (2020) Explainable AI for medical imaging: Deep-learning CNN ensemble for classification of estrogen receptor status from breast MRI. In Proceedings of the SPIE Medical Imaging 2020: Computer-Aided Diagnosis; International Society for Optics and Photonics: Bellingham, WA, USA, 2020; Volume 11314, p. 113140Z

Patrício C, Neves JC, Teixeira LF. Explainable deep learning methods in medical imaging diagnosis: a survey. arXiv:2205.04766v2. [eess.IV] 13 Jun 2022

Peng T, Boxberg M, Weichert W, Navab N, Marr C (2019) "Multi-task learning of a deep k-nearest neighbour network for histopathological image classification and retrieval." In Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part I 22, pp. 676–684. Springer International Publishing

Pereira S, Meier R, Alves V, Reyes M, Silva CA (2018) Automatic brain tumor grading from MRI data using convolutional neural networks and quality assessment. Understanding and interpreting machine learning in medical image computing applications. Springer, Cham, pp 106–114

Petsiuk V, Das A, Saenko K (2018) "Rise: randomized input sampling for explanation of black-box models." arXiv preprint arXiv:1806.07421

Pierson E, Cutler DM, Leskovec J, Mullainathan S, Obermeyer Z (2021) An algorithmic approach to reducing unexplained pain disparities in underserved populations. Nat Med 27:136–140

Pisov M, Goncharov M, Kurochkina N, Morozov S, Gombolevsky V, Chernina V, Vladzymyrskyy A, Zamyatina K, Cheskova A, Pronin I et al (2019) Incorporating task-specific structural knowledge into CNNs for brain midline shift detection. Interpretability of machine intelligence in medical image computing and multimodal learning for clinical decision support. Springer, Cham, pp 30–38

Punn NS, Agarwal S (2021) Automated diagnosis of COVID-19 with limited posteroanterior chest X-ray images using fine-tuned deep neural networks. Appl Intell 51:2689–2702

Quellec G, Al Hajj H, Lamard M, Conze PH, Massin P, Cochener B (2021) ExplAIn: explanatory artificial intelligence for diabetic retinopathy diagnosis. Med Image Anal 72:102118

Rajpurkar P, Oconnell C, Schechter A, Asnani N, Li J, Kiani A, Ball RL et al (2020) CheXaid: deep learning assistance for physician diagnosis of tuberculosis using chest x-rays in patients with HIV. NPJ Dig Med. https://doi.org/10.1038/s41746-020-00322-2

Ribeiro MT, Singh S, Guestrin C (2016) "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, pp. 1135–1144

Roy S, Bandyopadhyay SK (2013) "Abnormal regions detection and quantification with accuracy estimation from MRI of brain." In 2013 2nd International Symposium on Instrumentation and Measurement, Sensor Network and Automation (IMSNA), pp. 611–615. IEEE

Roy S, Bandyopadhyay SK (2016) A new method of brain tissues segmentation from MRI with accuracy estimation. Procedia Comput Sci 85:362–369

Roy S, Shoghi KI (2019) "Computer-aided tumor segmentation from T2-weighted MR images of patient-derived tumor xenografts." In Image Analysis and Recognition: 16th International Conference, ICIAR 2019, Waterloo, ON, Canada, August 27–29, 2019, Proceedings, Part II 16, pp. 159–171. Springer International Publishing

Roy S, Bhattacharyya D, Bandyopadhyay SK, Kim TH (2017a) An iterative implementation of level set for precise segmentation of brain tissues and abnormality detection from MR images. IETE J Res 63(6):769–783

Roy S, Bhattacharyya D, Bandyopadhyay SK, Kim TH (2017b) An effective method for computerized prediction and segmentation of multiple sclerosis lesions in brain MRI. Comput Methods Programs Biomed 140:307–320. https://doi.org/10.1016/j.cmpb.2017.01.003

Roy S, Bhattacharyya D, Bandyopadhyay SK, Kim TH (2018) Heterogeneity of human brain tumor with lesion identification, localization, and analysis from MRI. Inform Med Unlocked 13:139–150

Roy S, Whitehead TD, Quirk JD, Salter A, Ademuyiwa FO, Li S, An H, Shoghi KI (2020) Optimal co-clinical radiomics: Sensitivity of radiomic features to tumour volume, image noise and resolution in co-clinical T1-weighted and T2-weighted magnetic resonance imaging. EBioMedicine 59:102963

Roy S, Meena T, Lim SJ (2022) Demystifying supervised learning in healthcare 4.0: a new reality of transforming diagnostic medicine. Diagnostics 12(10):2549

Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 1(5):206–215

Samek W, Müller KR (2019) "Towards explainable artificial intelligence". Explainable AI: interpreting, explaining and visualizing deep learning. Springer, pp 5–22

Schlemper Jo, Oktay O, Schaap M, Heinrich M, Kainz B, Glocker B, Rueckert D (2019) Attention gated networks: learning to leverage salient regions in medical images. Med Image Anal 53:197–207

Schwab E, Gooßen A, Deshpande H, Saalbach A (2020) "Localization of critical findings in chest X-ray without local annotations using multi-instance learning." In 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), pp. 1879–1882. IEEE

Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2017) "Grad-cam: Visual explanations from deep networks via gradient-based localization." In Proceedings of the IEEE international conference on computer vision, pp. 618–626

Seo D, Kanghan Oh, Il-Seok Oh (2019) Regional multi-scale approach for visually pleasing explanations of deep neural networks. IEEE Access 8:8572–8582

Shen Y, Wu N, Phang J, Park J, Liu K, Tyagi S, Heacock L, Kim SG, Moy L, Cho K et al (2021) An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization. Med Image Anal 68:101908

Shibu CJ, Sreedharan S, Arun KM, Kesavadas C, Sitaram R (2023) Explainable artificial intelligence model to predict brain states from fNIRS signals. Front Hum Neurosci 19(16):1029784. https://doi.org/10.3389/fnhum.2022.1029784. (**PMID:36741783;PMCID:PMC9892761**)

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

Silva W, Fernandes K, Cardoso MJ, Cardoso JS (2018) Towards complementary explanations using deep neural networks. Understanding and interpreting machine learning in medical image computing applications. Springer, Cham, pp 133–140

Singh S, Karimi S, Ho-Shon K, Hamey L (2019.) "From chest x-rays to radiology reports: a multimodal machine learning approach." In 2019 Digital Image Computing: Techniques and Applications (DICTA), pp. 1–8. IEEE

Singh A, Sengupta S, Lakshminarayanan V (2020) Explainable deep learning models in medical image analysis. J Imaging 6(6):52. https://doi.org/10.3390/jimaging6060052. (**PMID:34460598;PMCID:PMC8321083**)

Song Y, Zheng S, Li L, Zhang X, Zhang X, Huang Z, Chen J, Wang R, Zhao H, Chong Y et al (2021) Deep learning enables accurate diagnosis of novel coronavirus (COVID-19) with CT images. IEEE/ACM Trans Comput Biol Bioinform 18:2775–2780

Springenberg JT, Dosovitskiy A, Brox T, Riedmiller M (2014) "Striving for simplicity: The all convolutional net." arXiv preprint arXiv:1412.6806

Stano M, Benesova W, Martak LS (2019) Explainable 3D convolutional neural network using GMM encoding. In Proceedings of the Twelfth International Conference on Machine Vision, Amsterdam, The Netherlands, 16–18 November 2019; Volume 11433, p. 114331U.

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

Van der Velden BH, Kuijf HJ, Gilhuijs KG, Viergever MA (2022) Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. Med Image Anal 79:102470

Van Molle P, Der Strooper M, Verbelen T, Vankeirsbilck B, Simoens P, Dhoedt B (2018) Visualizing convolutional neural networks to improve decision support for skin lesion classification. Understanding and interpreting machine learning in medical image computing applications. Springer, Cham, pp 115–123

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

Wang Z, Zhu H, Ma Y, Basu A (2021) "XAI Feature Detector for Ultrasound Feature Matching," 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Mexico, 2021, pp. 2928–2931, https://doi.org/10.1109/EMBC46164.2021.9629944

Wang SH, Govindaraj VV, Górriz JM, Zhang X, Zhang YD (2021b) COVID-19 classification by FGCNet with deep feature fusion from graph convolutional network and convolutional neural network. Inf Fusion 67:208–229

Wang Z, Xiao Y, Li Y, Zhang J, Lu F, Hou M, Liu X (2021c) Automatically discriminating and localizing COVID-19 from community-acquired pneumonia on chest X-rays. Pattern Recognit 110:107613

Wang H, Wang S, Qin Z, Zhang Y, Li R, Xia Y (2021d) Triple attention learning for classification of 14 thoracic diseases using chest radiography. Med Image Anal 67:101846

Windisch P, Weber P, Fürweger C, Ehret F, Kufeld M, Zwahlen D, Muacevic A (2020) Implementation of model explainability for a basic brain tumor detection using convolutional neural networks on MRI slices. Neuroradiology 62:1515–1518

Windisch P et al (2020) Implementation of model explainability for a basic brain tumor detection using convolutional neural networks on MRI slices. Neuroradiology 62(11):1515–1518. https://doi.org/10.1007/s00234-020-02465-1

Wu YH, Gao SH, Mei J, Xu J, Fan DP, Zhang RG, Cheng MM (2021) JCS: an explainable covid-19 diagnosis system by joint classification and segmentation. IEEE Trans Image Process 30:3113–3126

Xie B, Lei T, Wang N et al (2020) Computer-aided diagnosis for fetal brain ultrasound images using deep convolutional neural networks. Int J CARS 15:1303–1312. https://doi.org/10.1007/s11548-020-02182-3

Xing H, Xiao Z, Zhan D, Luo S, Dai P, Li K (2022a) SelfMatch: robust semisupervised time-series classification with self-distillation. Int J Intell Syst 37(11):8583–8610

Xing H, Xiao Z, Rong Qu, Zhu Z, Zhao B (2022b) An efficient federated distillation learning system for multitask time series classification. IEEE Trans Instrum Meas. https://doi.org/10.1109/TIM.2022.3201203,71,(1-12)

Xu F, Jiang L, He W, Huang G, Hong Y, Tang F, Lv J, Lin Y, Qin Y, Lan R, Pan X, Zeng S, Li M, Chen Q, Tang N (2021) The clinical value of explainable deep learning for diagnosing fungal keratitis using in vivo confocal microscopy images. Front Med (lausanne) 14(8):797616. https://doi.org/10.3389/fmed.2021.797616. (**PMID:34970572;PMCID:PMC8712475**)

Yamashita R, Nishio M, Do RKG et al (2018) Convolutional neural networks: an overview and application in radiology. Insights Imaging 9:611–629. https://doi.org/10.1007/s13244-018-0639-9

Yang HL, Kim JJ, Kim JH, Kang YK, Park DH, Park HS, Kim HK, Kim MS (2019) Weakly supervised lesion localization for age-related macular degeneration detection using optical coherence tomography images. PLoS ONE 14:e0215076

Yang G, Ye Q, Xia J (2021) "Unbox the Black-box for the medical explainable AI via multi-modal and multi-centre data fusion: a mini-review, two showcases and beyond". arXiv:2102.01998v1 [cs.AI] 3 Feb 2021

Yeche H, Harrison J, Berthier T (2019) UBS: a dimension-agnostic metric for concept vector interpretability applied to radiomics. Interpretability of machine intelligence in medical image computing and multimodal learning for clinical decision support. Springer, Cham, pp 12–20

Young K, Booth G, Simpson B, Dutton R, Shrapnel S (2019) Deep neural network or dermatologist? Interpretability of machine intelligence in medical image computing and multimodal learning for clinical decision support. Springer, Cham, pp 48–55

Zeiler MD, Fergus R (2014) "Visualizing and understanding convolutional networks." In Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part I 13, pp. 818–833. Springer International Publishing

Zeineldin RA, Karar ME, Elshaer Z et al (2022a) Explainability of deep neural networks for MRI analysis of brain tumors. Int J CARS 17:1673–1683. https://doi.org/10.1007/s11548-022-02619-x

Zeineldin RA, Karar ME, Elshaer Z, Coburger J, Wirtz CR, Burgert O, Mathis-Ullrich F (2022b) Explainability of deep neural networks for MRI analysis of brain tumors. Int J Comput Assisted Radiol Surg 17(9):1673–1683

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

Zhu P, Ogino M (2019) Guideline-based additive explanation for computer-aided diagnosis of lung nodules. Interpretability of machine

intelligence in medical image computing and multimodal learning for clinical decision support. Springer, Cham, pp 39–47
Zintgraf LM, Cohen TS, Adel T, Welling M (2017) "Visualizing deep neural network decisions: Prediction difference analysis." arXiv preprint arXiv:1702.04595