Computers in Biology and Medicine 156 (2023) 106668

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Survey of explainable artificial intelligence techniques for biomedical imaging with deep neural networks

Sajid Nazir [a,*], Diane M. Dickson [b], Muhammad Usman Akram [c]

[a] Department of Computing, Glasgow Caledonian University, Glasgow, UK
[b] Department of Podiatry and Radiography, Research Centre for Health, Glasgow Caledonian University, Glasgow, UK
[c] Computer and Software Engineering Department, National University of Sciences and Technology, Islamabad, Pakistan

ARTICLE INFO

Keywords:
Interpretable AI
Blackbox
Features
Supervised learning
Predictive models
Neural networks
Diagnostic imaging
Backpropagation

ABSTRACT

Artificial Intelligence (AI) techniques of deep learning have revolutionized the disease diagnosis with their outstanding image classification performance. In spite of the outstanding results, the widespread adoption of these techniques in clinical practice is still taking place at a moderate pace. One of the major hindrance is that a trained Deep Neural Networks (DNN) model provides a prediction, but questions about why and how that prediction was made remain unanswered. This linkage is of utmost importance for the regulated healthcare domain to increase the trust in the automated diagnosis system by the practitioners, patients and other stakeholders. The application of deep learning for medical imaging has to be interpreted with caution due to the health and safety concerns similar to blame attribution in the case of an accident involving autonomous cars. The consequences of both a false positive and false negative cases are far reaching for patients' welfare and cannot be ignored. This is exacerbated by the fact that the state-of-the-art deep learning algorithms comprise of complex interconnected structures, millions of parameters, and a 'black box' nature, offering little understanding of their inner working unlike the traditional machine learning algorithms. Explainable AI (XAI) techniques help to understand model predictions which help develop trust in the system, accelerate the disease diagnosis, and meet adherence to regulatory requirements.

This survey provides a comprehensive review of the promising field of XAI for biomedical imaging diagnostics. We also provide a categorization of the XAI techniques, discuss the open challenges, and provide future directions for XAI which would be of interest to clinicians, regulators and model developers.

1. Introduction

Deep Neural Networks (DNN) provide exceedingly better results in image classification tasks compared to other techniques in Artificial Intelligence (AI), and can often surpass the human domain experts. There has been a renewed interest in applications of AI techniques for imaging classification and segmentation for biomedical applications [1]. Convolutional Neural Networks (CNN) are state-of-the-art models for image segmentation and classification [2], although encoder-decoder transformer architectures have also been proposed [3–5]. In segmentation for medical imaging, the region of interest (RoI) such as a lesion needs to be delineated from the surrounding pixels by assigning a label to every pixel in the image so that the similar constituent objects in the image are grouped together. Classification is the process of assigning a label, such as benign or malignant to a medical image, for binary classification or one of the many classes, such as normal, pneumonia, COVID-19, for a multi classification problem. The images for biomedical applications are acquired through an array of imaging modalities such as conventional X-ray and ultrasound, and used for disease diagnosis and management, for example, use of fundus images for detection of eye diseases [6]. The democratization of AI has made possible the application of machine learning algorithms by domain experts, without the need of a deeper understanding of the underlying algorithms. AI can aid biomedical imaging as it can "see" additional and superior detailed information regarding tumor size, texture, shape etc. compared to, for example, human radiologists [7]. AI can potentially transform the image-based diagnosis in radiology [8]. Disease diagnosis through these techniques is inevitable to help overcome the shortage of, for example, 2000 radiologists in UK [9]. However, for the regulated healthcare domain, it is utmost important to comprehend, justify, and explain the

* Corresponding author.
E-mail address: sajid.nazir@gcu.ac.uk (S. Nazir).

https://doi.org/10.1016/j.compbiomed.2023.106668
Received 6 August 2022; Received in revised form 12 January 2023; Accepted 10 February 2023
Available online 18 February 2023

AI model predictions for a wider adoption of automated diagnosis.

Simpler deterministic models are often inherently explainable, for example, it is relatively easier to understand a decision made by a shallow decision tree in terms of the if-else rules [10,11]. However, the models currently gaining prominence for image analysis are based on neural network techniques. A DNN mimics the biological neuron connections in a human brain but this is without our full understanding of neuron interconnections and functioning of the human brain. In an object identification task, a child can easily learn to identify an animal in an image although it is not very well understood as to how a child recognizes an object. The DNN model or the algorithm architecture also depends on hyperparameters that define the performance of the trained model [12]. Neural networks require a lot of training data and time before they can produce acceptable results but are like a 'black-box' providing little or no insight into the decision-making process [13]. The model opaqueness makes it hard even for the model developers to understand how the model is functioning [14]. Explainability is required for problems that are not fully comprehensible by humans in terms of how to arrive at or mimic the process from the problem to a solution. Incorporating explainability into a model takes time and effort, and may affect its prediction accuracy. Thus, it should be reserved for areas where health and safety are important, such as healthcare, compared to, for example, an application classifying an email as spam.

Explainable AI (XAI) techniques aim to provide additional information about a model's decision thereby improving trust in model's decisions, as shown in Fig. 1 "An explainable model is one which provides explanations for its predictions at the human level for a specific task. An interpretable model is one for which some conclusions can be drawn about the internals/predictions of the model" [15]. However, there is often an explainability gap between the model's prediction and its explanations being understandable to a human. Various techniques and tools have been developed that can aid the model's understandability. An understanding of DNNs can be obtained by visualizing the layer wise activations and features of a trained network through the online tool [16]. The connection between machine vision and Natural Language Processing (NLP) for AI explainability was explored through image annotations and visual question-answering [17]. An XAI challenge aims to develop better XAI models for the financial services community as a collaboration between the organizers, Google and academic institutions [18]. The explainability or interpretability gap in current healthcare explainable systems, goes beyond imaging, and similar problems exist for contextual language models such as Bidirectional Encoder Representations from Transformers (BERT) [19]. DNN by itself cannot provide explanations that can be understood by domain experts or the end-user whereas symbolic AI methods such as knowledge graphs can be explainable [20]. XAI is starting to become synonymous with the use of deep learning techniques in the regulated domains. It is thus becoming significantly important to understand why and how a decision was made, why the model failed, how to avoid failures, and ultimately how to improve the model.

The increasing emphasis on XAI is evidenced by a number of recent survey papers in the area. A study provided a taxonomy and evaluation techniques for XAI [21]. A survey of visual analytics techniques covered the XAI adoption for interpreting DL models [22]. Some of the reviews have covered Machine Learning (ML) interpretability [23], trustworthiness for DNN [24], explanations types [25], objective quality evaluation of explanations [25], responsible AI [2], post-hoc explainability techniques [2], and evaluation of the model explanation [26]. Some of the XAI surveys address the applications in healthcare sector. A survey of XAI in machine learning based clinical decision support system (CDSS) [27], medical image analysis [28,29], interpretable methods for neural networks [30], AI interpretability in radiology [31], trustworthy AI for healthcare [32], AI techniques for COVID-19 [33,34] neuroimaging [35], and oncological radiomics, feature extraction basics and major interpretability methods [36].

In this paper, we use the term XAI to refer to and encompass all the similar terms such as interpretable, understandable and trustworthy AI, and define the term explainability for a CNN model to encompass any technique, or supplementary information that helps to understand the model's underlying decision process. The papers were selected by searching for articles published on explainable, interpretable, and trustworthy XAI techniques for biomedical applications on the popular databases of Web of Science, ScienceDirect, IEEE Explore, and Google Scholar.

In comparison to the other surveys above, we present a holistic and comprehensive survey of the XAI techniques and their application for biomedical imaging. This survey is unique to previously published surveys as we cover the XAI techniques from the perspective of its application to the different imaging modalities and medical specialties. The contributions of this paper are as follows:

- Provides a comprehensive review of the XAI techniques useful for the understanding of DNN models along with grouping of the techniques, highlighting their scope, applicability, and usage. We also provide a discussion of the various tools and frameworks that can help with explainability, and the evaluation of explainability.

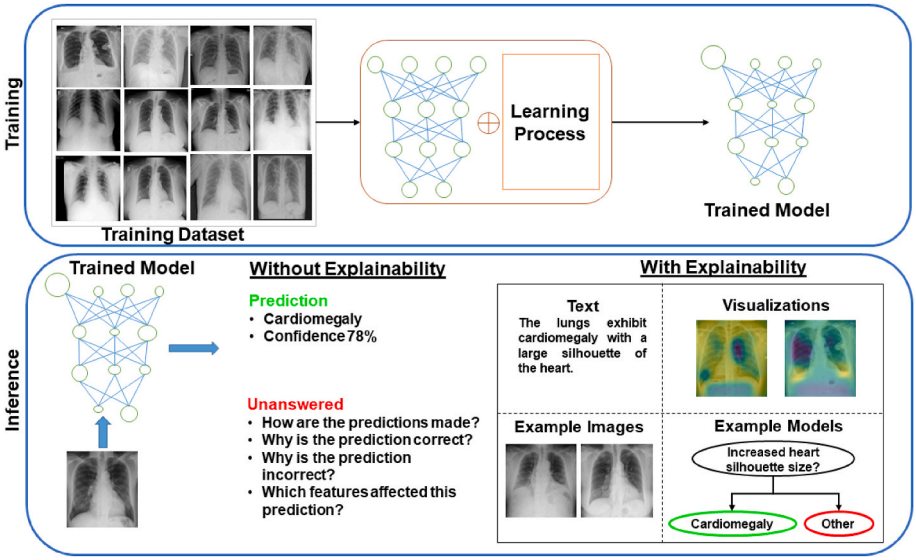

Fig. 1. XAI helps stakeholders to understand the model's decision.

- Provides the use of XAI techniques for image classification and segmentation tasks in different image modalities for the disease diagnosis in biomedical imaging domain. The techniques are covered according to the imaging modalities and summarized according to the medical specialties.
- Provides a discussion on the different imaging modalities and XAI techniques in medical imaging. The challenges and research directions for the adoption of XAI for biomedical imaging and how these challenges can be addressed are also covered.

The rest of the paper is organized as follows. The background information is provided in Section 2. Section 3 covers the XAI techniques for DNNs. Section 4 discusses the application of XAI techniques in biomedical imaging. The open challenges and recommendations are provided in Section 5. Finally, the conclusion is provided in Section 6.

2. Background

2.1. DNN in medical applications

Deep Learning is a sub-field of machine learning, and is based on neural networks, coming to the recent prominence due to the outstanding results on the image datasets. A lot of machine learning models are available, but these are distinct from a deep learning model in that the features are required to be extracted which is difficult and time consuming in clinical practice. It is also difficult to determine and translate the feature descriptors for a complex disease pattern [6]. In deep learning models the features can be learnt automatically [6]. This is convenient but herein lies a problem regarding the decision made by the model. Given a trained model, the prediction comprises of a class label (for example malignant tumor) and an accompanying confidence level [6]. The field of XAI is burgeoning in view of this challenge and is aimed at creating white box DNN models that are inherently explainable or to provide explanations for black box DNN models' predictions, where the model functioning is too complex or hidden and therefore not understood.

DNNs have been very successful in the image classification and segmentation tasks which are often required in a clinical practice. Despite the outstanding performance of DNN for classification and segmentation tasks, they have some limitations which need to be considered for healthcare applications. A DNN being a complex structure is more like a black box with little or no visibility into why and how a particular model's decision was made. Both a false positive and false negative pose a problem and can have serious consequences. A false positive in disease diagnosis would mean that presence of a disease is predicted when the patient does not have the disease, and the patient may have to endure unnecessary treatment. A false negative means that a patient has a disease but is not diagnosed correctly, and the disease continues to progress. Hence, false negative error may be costlier compared to a false positive in relation to medical image classification [37].

IBM Watson for health system was developed by IBM as a system with better performance compared to oncologists [38]. The performance of IBM Watson system for oncology had very successful performance and agreements with oncologists in some areas but poor in other areas, indicating adaptations to local practices and not extensible at a global scale [38]. The Watson system could not provide justifications for its recommendations in clinical practice [38].

The model bias can result in erroneous outputs due to faulty assumptions in the model development process. The biases in the algorithms, data, or humans can result in misdiagnosis of certain patient groups further compounding the inequalities [39]. Algorithm or model bias can be a risk in AI as the models can generate a wrongful output based on an intentional or unintentional bias [40]. The problem of bias can be addressed by an open sharing of the research process, methods, data and results, with such open sharing across institutions and funding agencies enabling open science [39]. The open models can be trained with local data from, for example, hospitals across the globe to mitigate bias and boost sample size [39].

During the model development, it is important to have data from unseen sources in the test set to ensure model generalizability [41]. A generalizable model should provide similar performance on the unseen data, to that obtained with the training data. The evaluation of deep learning techniques was performed across different hospitals, and for pneumonia screening it was found that the models had better internal (of hospitals) results compared to external [42]. The model during training ideally should learn the discriminating image features of the various classes but instead can also learn extraneous features [42]. It was demonstrated that deep learning models on chest X-ray image datasets were not relying on medical pathology but instead on confounding factors for COVID-19 detection [41,43].

The XAI techniques are needed to make the AI predictions understandable, trustworthy, and reliable. These techniques and the automated AI decisions will never replace a clinician but instead should be seen as an AI assistant aiding better human decisions through additional insights into the model's predictions. The use of XAI techniques can reduce the workload of the clinicians, help to make better diagnosis, and reduce the complications associated with a false positive and a false negative outcome.

2.2. Definitions and context

The terms relating to XAI are used interchangeably in different domains and literature [40]. By analyzing the corpus focusing on AI methods, the AI systems were categorized as: (1) Opaque, where the input to output mappings are invisible to the user, (2) Interpretable system, where a user can both see and understand how the inputs are mapped to output, and (3) Comprehensible systems, that provides some additional output, such as visualizations, text etc. [44]. The opaque systems are DNNs, which have a better performance and accuracy, but are not interpretable compared to the conventional AI systems.

2.2.1. Explainable AI

DARPA XAI program was launched in May 2017 envisioning three categories to improve explainability as deep explanation, interpretable models, and model induction [45]. The XAI program addresses two challenges of event classification and autonomous system policies, with explanation effectiveness being critical for both challenge areas [45]. DARPA is funding 13 different research groups, which are pursuing a range of approaches to making AI more explainable [46].

DARPA defines explainable AI as, "AI systems that can explain their rationale to a human user, characterize their strengths and weaknesses, and convey an understanding of how they will behave in the future. Naming this program explainable AI (rather than interpretable, comprehensible, or transparent AI, for example) reflects DARPA's objective to create more human-understandable AI systems through the use of effective explanations" [45].

"Explanation refers to the ability to accurately describe the mechanism, or implementation, that led to an algorithm's output, often so that the algorithm can be improved in some way" [47]. The four principles for XAI were described as [48]:

- *Explanation*: A system delivers or contains accompanying evidence or reason(s) for outputs and/or processes.
- *Meaningful*: A system provides explanations that are understandable to the intended consumer(s).
- *Explanation Accuracy*: An explanation correctly reflects the reason for generating the output and/or accurately reflects the system's process.
- *Knowledge Limits*: A system only operates under conditions for which it was designed and when it reaches sufficient confidence in its output.

European Union (EU) General Data Protection Regulation (GDPR) is legislating a "right to explanation" for EU citizens impacted by automated algorithmic decisions [49]. The EU legislation is likely to increase the focus of computer scientists towards developing algorithms and systems that will enable explanations and avoid discrimination [50].

2.2.2. Interpretable AI

There is little agreement on what is an interpretable ML system and how it should be evaluated [51]. There is a need to have a more formalized definition of interpretability in the context of DNNs [51]. An interpretable model can obtain transparency by providing the reasons behind their output [52].

A distinction between explanation and interpretation was provided as, "We find that interpretation refers to the ability to contextualize a model's output in a manner that relates it to the system's designed functional purpose, and the goals, values, and preferences of end users. In contrast, explanation refers to the ability to accurately describe the mechanism, or implementation, that led to an algorithm's output, often so that the algorithm can be improved in some way" [47].

2.2.3. Responsible AI

Responsible AI is concerned with the actions and decisions of AI systems that can have consequences of ethical nature [53]. Even a system deemed to be safe and built according to legislations can have negative consequences for societal and human well-being [53]. The AI systems should be considered as tools designed by the people where irrespective of the system's degree of autonomy and ability to learn, it cannot replace the human responsibility [53].

Responsible AI is concerned specifically with establishing cultural and ethical principles to implement fairness and accountability, to reduce bias, to promote fairness, to ensure privacy, and to facilitate explainability [2]. Responsible AI, in addition to explainability, should also consider the guidelines behind Responsible AI to establish fairness, accountability, and privacy when implementing AI models in real environments [2].

2.2.4. Trustworthy AI

The guidelines for trustworthy AI are published by EU High-Level Expert Group on AI [54] as, "the AI system's lifecycle should meet the seven key requirements for Trustworthy AI: (1) human agency and oversight, (2) technical robustness and safety, (3) privacy and data governance, (4) transparency, (5) diversity, non-discrimination and fairness, (6) environmental and societal well-being and (7) accountability".

While trustworthiness should be a property of every explainable system, this does not mean that every trustworthy system should also be explainable [2]. Self-explaining AI is proposed that produces a decision and an explanation. It provides a human understandable explanation and confidence level for each decision and which was deemed to motivate and lead towards trustworthy AI [55]. A self-explainable system could check its own inputs and outputs, and could indicate in case of errors coupled with a human interpretable explanation [55].

2.2.5. Accuracy and interpretability trade-off

DNN can provide very accurate predictive performance but owing to a large number of trainable parameters, layers, and the performance determined through the hyperparameters, these are hard to interpret or explain. For example, VGG-16 model requires 15.5 billion floating-point operations for a single image classification with approximately 138 million parameters [56]. However, it is the better predictive performance of DNN models compared to simpler models that has brought these to prominence. There is a trade-off between the accuracy and interpretability of the models, as shown in Fig. 2, where non-linear models with better performance, such as DNNs are less explainable, and linear models which are least accurate, such as decision trees, are more explainable [2,21,45]. It remains a challenge to build models that

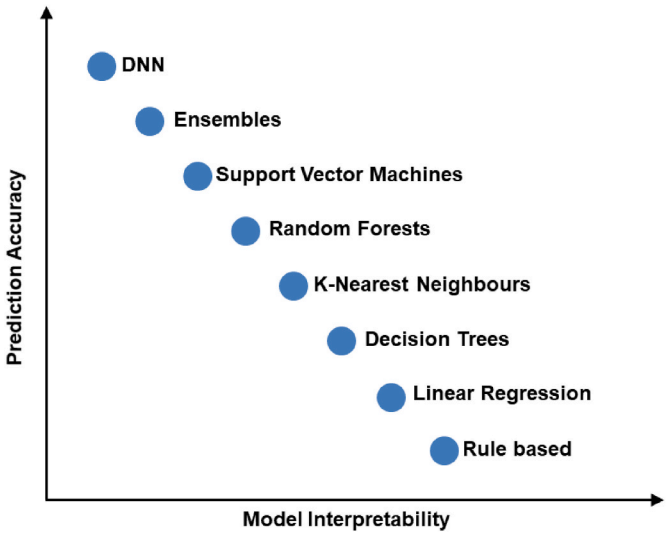

Fig. 2. Model's prediction accuracy and interpretability trade-off.

are both accurate and interpretable [52]. This trade-off can limit the usability of the more accurate models.

Therefore, it is important that models are not only interpretable but also accurate, a position held by neural networks. The nature of explanations is determined by the domain. For domains with a lack of knowledge about the underlying causal process, such as healthcare, it is prudent not to prefer explainability over diagnostic or predictive accuracy as this may affect the health and well-being of patients with some who get additional investigations and treatment and similarly those who are mjs- or undiagnosed [57]. This creates a dilemma in clinical practice as the model selected for use would logically have the greatest predictive performance.

The tradeoff in predictive accuracy and explanation is important to understanding the XAI models and to make appropriate choices for a trustworthy model [58]. The models with an intrinsic interpretability provide better explainability at lower predictive performance compared to post-hoc methods that in contrast have less explainability but a higher predictive performance [59]. Thus, there is an accuracy cost for intrinsic interpretable models [21]. The advances in post-hoc interpretability is helpful by approximating the black box models [58].

2.3. XAI stakeholders

Owing to a large variety of AI models and their intended purposes, a single strategy for explainability will probably not work well for the different stakeholders in the system [60]. A lack of consensus in the meaning of explainability originates due to different stakeholders, each having their own understanding of the terms [61]. The major stakeholders for XAI and their interests and concerns are described below:

2.3.1. AI experts and application developers

The details required by the AI experts would be much more compared to other stakeholders owing to the need to improve the model performance by changing its architecture and the choice of the various hyperparameters. The developers also have an interest and responsibility in ensuring that the system works correctly [61].

The application developers would like to use the XAI techniques to understand the inner working and model improvement [62]. Most XAI work is by the AI experts which might help engineers to build better AI systems but may be limited for other stakeholders especially the end-user [8]. Similarly, formal mathematical notations and techniques can unambiguously provide explanations for domain experts but are not targeted to the end users [63]. One of the ways forward could be based on the statistical correctness and fairness of the model [49].

2.3.2. Regulators

The regulators would most likely be interested in an overall performance and a combination of explanations of the model rather than specific cases of individual patients [25]. These have interest in an XAI system to be able to lay down standards and model validation and certification [2]. Similar to ISO 13485:2016 Medical devices standard for Quality management systems that prescribes the medical devices regulatory requirements, there is a need for certifications of DNN models for clinical adoption [64].

Certification frameworks for a formal validation of models could testify the model operation statistically and ensure that the model predictions meet the required regulatory standards such as ISO/IEC CD 23053 (2020) for establishing a framework for use of AI systems [49]. An Independent, Accredited Certification Program for Responsible AI RAII Certification uses six dimensions of responsible AI (System Operations, Explainability and Interpretability, Accountability, Consumer Protection, Bias and Fairness, and Robustness) to represent trust that an AI system has been designed, built, and deployed in line with the five Organization for Economic Co-operation and Development (OECD) principles [65] on AI [66].

2.3.3. Medical practitioners

Medical practitioners are interested in understanding the model by asking why a particular decision was made and how does it validate their own diagnosis. They are more likely to employ and trust a system if they can understand the reasoning of a prediction by a model. Failure to identify and address the concerns of the system users can lead to another "AI Winter", where a small market segment could change the research and development focus from XAI [61]. Clinicians are the primary users of XAI systems and they may have differing views but all have a need for the explainability for the patient [27].

2.3.4. Patients

The patients would be interested in knowing the soundness of the model's decision as perceived by their own chain of thoughts for their medical condition. This has gained an increased importance due to the recent emphasis on the patients' role for data ownership and the need for providing a personalized healthcare. The patients require simplified explanations for the clinical decision as the DNN models are very complex and cannot be explained to a lay user [49]. Patients require assurance that the XAI methods can be trusted, and are providing a valid and consistent prediction.

3. Explainability techniques for DNN

For DNN which have a black-box nature, it is not possible to directly understand the inner functioning. A heatmap can be used to represent the hidden information regarding the feature importance using colors. The heatmaps are commonly used to explain AI models and the quality of heatmaps can be appreciated using perturbation analysis [25]. Different techniques have been developed to understand the model classification at a pixel, superpixel, or feature level and can be categorized as:

- Visualization: These techniques are aimed at depicting the inner workings of the model as data visualizations. These can be further categorized as perturbation or gradient/backpropagation-based approaches.
- Visualization by Perturbation: These methods work by changing the input image and noticing the changes to the model prediction. This could be by occlusion or modifying the input by adding noise or by blurring.
- Visualization by Gradient/Backpropagation: The backpropagation-based methods use the backward pass to determine the image attributes that determine the model prediction.
- Explanation by Example: These explanations aid the model explainability by extracting examples similar to a prediction, generating additional features such as text, or using additional details.

The explainability techniques can be categorized depending on the scope, applicability and nature [21,29] as under:

- Scope:
  - Local: The methods with local scope are used to explain for a single outcome or prediction.
  - Global: The methods with a global scope try to explain the functionality of the whole model.
- Applicability:
  - `Model Specific: These methods work only for specific models only.`
  - Model Agnostic: These methods are independent of the model and can be used for any model architecture.
- Usage/Nature:
  - `Intrinsic: These are the models which have the explainability built-in in that it is easy to observe the model decision and features, treating it as a white-box, e.g., decision trees.`
  - Post hoc: These methods do not have access to the model parameters and weights, but are used to explain the model after it is built treating it as a black-box, e.g., DNN.

The XAI techniques for understanding of image-based models are described next and are summarized in Table 1. This section also covers the XAI libraries and frameworks, and describes the techniques for evaluating and improving the explainability techniques.

3.1. Visualization

The visualization techniques are aimed at highlighting the significance of pixels, image patches, neuron activations, salient gradients, or features that can help to understand the model decision-making process. The benefit is that the model's prediction can be compared against a clinician's decision, leading to an overall understanding, trust, and better decision making.

The Partial Dependence Plot (PDP) is a popular tool for visualizing black box models and can highlight the feature importance [67]. PDP was extended to propose a toolbox, Individual Conditional Expectation (ICE) plots [68]. ICE plots use the relation between the features of individual observations and the predicted response in order to improve the PDP [68]. The experiments with the proposed method showed improvements in visualizations [68]. The tool is available as an R package ICEBOX [68].

A visual explanation method termed CNN Fixations utilized feature dependencies across the layers during the forward pass to uncover the discriminative image locations that dictate the model prediction [69]. The method computed the important image locations (CNN Fixations-analogous to human eye fixations) and required no additional training, architectural change, or gradient computation [69]. The approach works by unraveling the underlying forward pass operation to find the important pixel locations [69]. The visual explanations for image captioning models with Long Short-Term Memory (LSTM) were also demonstrated [69].

3.2. Visualization by perturbation

Perturbation based visualization approaches work by changing the input to a DNN model to determine the significance of the features, pixels, or image areas to the prediction. Local Interpretable Model-agnostic Explanations (LIME) was proposed to explain any classifier's predictions by approximating it locally around the prediction [70]. It

Table 1
XAI TECHNIQUES.

Classification	Technique	Reference	Scope		Applicability		Usage	
			Local	Global	Model Specific	Model Agnostic	Post hoc	Intrinsic
Visualization	PDP	[67]		✓		✓	✓	
	Individual conditional expectation (ICE)	[68]	✓			✓	✓	
	CNN Fixations	[69]	✓			✓		✓
Visualization by Perturbation	Local Interpretable Model-Agnostic Explanations (LIME)	[70]	✓	✓		✓	✓	
	Autoencoder Based Approach for Local Interpretability (ALIME)	[71]	✓			✓	✓	
	RISE	[72]	✓		✓		✓	
	IASSA	[73]		✓		✓	✓	
	X-Caps	[15]	✓		✓			✓
	HihO (Hierarchical Occlusion)	[74]	✓			✓	✓	
	NICE (Neural Image Compression and Explanation)	[75]	✓			✓	✓	
	Similarity Difference and Uniqueness Method (SIDU)	[76]	✓			✓	✓	
Visualization by Gradient/ Backpropagation	Saliency Maps	[77]	✓			✓	✓	
	Class Activation Map (CAM)	[78]	✓		✓		✓	
	Grad-CAM	[79]	✓		✓		✓	
	Grad-CAM++	[80]	✓		✓		✓	
	Integrated Grad-CAM	[81]	✓		✓		✓	
	Eigen-CAM	[82]	✓			✓	✓	
	Layer-Wise Relevance Propagation (LRP)	[83]	✓	✓	✓		✓	
	Selective Layer-Wise Relevance Propagation (SLRP)	[84]	✓		✓		✓	
	Excitation Backpropagation	[85]		✓	✓		✓	
	Deep Learning Important FeaTures (DeepLIFT)	[86]	✓	✓		✓	✓	
	Deconvnet	[87]	✓		✓		✓	
	Testing Concept Activation Vectors (TCAV)	[88]	✓			✓	✓	
	DeepTaylor	[89]	✓			✓	✓	
	SHapley Additive exPlanation (SHAP)	[90]	✓	✓		✓	✓	
	Locality Guided Neural Network (LGNN)	[91]		✓		✓		✓
Explanation by Example	Text Explanations	[92,93]	✓			✓	✓	
	Image captioning	[94]		✓	✓		✓	
	Explainable Neural Network (XNN)	[95]		✓	✓		✓	
	MAGIX	[96]		✓		✓		
	BETA	[97]		✓		✓		
	Surrogate-ontologies-decision trees	[98]	✓			✓	✓	
	Graph neural network	[99,100]		✓	✓		✓	
	Counterfactuals	[101, 102]	✓			✓	✓	
	CoCoX	[103]		✓		✓	✓	
	Influence functions	[104]		✓		✓	✓	
	Bayesian Teaching	[8,105]		✓		✓	✓	
	PIECE	[106]	✓			✓	✓	

works by carrying out a random search by successively leaving out the super-pixels from the image. An example of text explanations with Support Vector Machine (SVM) and deep networks for images were also provided [70]. An autoencoder was used as a weighting function to improve the performance of LIME model that had a limitation of instability in the generated explanation due to the generated dataset [71]. The technique was termed Autoencoder-based Local Interpretability Model (ALIME) [71].

Randomized Input Sampling for Explanation of Black-box Models (RISE) generates an importance map based on the importance of each pixel to the prediction [72]. The model can also be used for captioning an image [72]. The authors also proposed two model evaluation metrics termed deletion and insertion implying removal and addition of pixels and measuring the corresponding effect [72]. The model was evaluated using the proposed automatic causal metrics, termed as deletion and insertion, showing better performance compared to Grad-CAM, Sliding Window, and LIME [72]. Iterative and Adaptive Sampling with Spatial Attention (IASSA) was proposed that used an iterative and adaptive sampling module to determine the saliency map of each pixel for the model prediction [73]. The proposed method was motivated by RISE and used deletion, insertion, Intersection over Union (IoU), F1-score and a pointing game score for the evaluation of saliency maps [73]. The method was shown to provide better results in highlighting the regions of interest in the saliency maps comparison with RISE and LIME [73].

Capsule networks can encode high-level image features in the capsule vector's dimensions. An explainable capsule network, X-Caps was proposed to predict based on human interpretable features by encoding high-level image attributes within the capsule vectors [15]. It was demonstrated that the proposed 2D capsule network system provided better visualizations and malignancy predictions compared to 3D CNN and achieved better diagnostic accuracy [15].

Occlusion means to mask part of an image and to observe the corresponding change to the classification. A novel hierarchical occlusion algorithm is proposed and is based on two observations; that the important features are localized in an image, and features collection in multiple scales can affect the prediction outcome [74]. The introduction of a hierarchy excluded the regions with less important features [74]. The algorithm requires an image and a classification model for its operation [74].

Natural Image Compression and Explanation (NICE) uses a mask over the entire image for each pixel in the image that captures the saliency based on its contribution to the prediction. The mask is used to subsample less important background pixels to a low resolution whereas the important pixels retain their resolution [75]. The sparse masks align

well with human intuitions as compared to backpropagation-based techniques [75].

A visual explainable system Similarity Difference and Uniqueness Method (SIDU) was proposed to estimate pixel saliency by extracting feature maps from the last CNN layer, and creating a difference mask to form the visual explanation of the prediction [76]. The method was motivated by Grad-CAM and RISE, however, unlike Grad-CAM it was gradient-free, and unlike RISE, it used a combined final mask instead of a random mask [76]. The proposed scheme was evaluated on clinical and general datasets and showed better visual explanations with a greater trust for the human expert [76].

3.3. Visualization by gradient/backpropagation

Backpropagation based visualization approaches use the back-propagated information in the backward pass to determine the significance of the image regions for the prediction. It was argued that the Gradient-Based methods are lacking as the neurons' gradients do not relate to the top-down saliency [85]. Thus, the gradient based methods lack in generalizability and robustness [85].

Saliency map highlights the relevant input features for model prediction [107]. This is achieved by testing a network repeatedly to determine which parts of the input effect the output [108]. The saliency map-based methods for image classification were evaluated using the proposed model parameter randomization and data randomization tests [107]. A study considered two visualization techniques, the first generated an image based on maximizing class score, whereas the second, computed a class saliency map [77]. It used saliency maps for both convolution and deconvolution, and a gradient-based visualization technique [77].

Another study extended the ability of a CNN from just localizing an object to identify the regions in the image that are aiding the discrimination [78]. The Class Activation Map (CAM) approach used global average pooling for generation class activation maps [78]. The weighted sum of the spatial averages of each unit's feature maps at the last convolutional layer were used for the final output generation [78]. Gradient-weighted Class Activation Mapping (Grad-CAM) was proposed for producing visual explanations by using the gradients in the final model layer to identify and highlight the important image regions [79]. It was formally proved by the authors that Grad-CAM generalized CAM for many CNN architectures [79]. The method was evaluated with human studies for class discrimination, trust, and faithfulness in explaining the learnt function [79]. The applicability of Grad-CAM for visual question answering, image captioning, and image classification were also demonstrated [79]. Trustworthy ML-based perception was obtained using Grad-CAM heatmaps [109].

Grad-CAM++ was proposed as a method to build on Grad-CAM and provided better visual explanations compared to it and was better in explaining multiple occurrences of the same class objects in a single image [80]. Grad-CAM++ was a more generalized method compared to CAM and Grad-CAM and a mathematical formulation and comparisons on multiple datasets was provided. The method was demonstrated to be better compared to Grad-CAM in both the objective evaluations and through human studies [80].

Integrated Grad-CAM was proposed to overcome the gradient issues in CAM based methods by taking advantage of the backpropagation techniques [81]. The proposed method was evaluated and compared with Grad-CAM and Grad-CAM++ [110] using Energy-based Point Game (EBPG) and Bounding box (Bbox) as object localization accuracy and feature visualization metrics [81]. The method Eigen-CAM was proposed to use principal components in order to provide visual explanation in the presence of an adversarial noise [82]. The proposed method was compared with Grad-CAM, Grad-CAM++, and CNN Fixations and the results showed an improvement for the weakly supervised object localization [82]. It was demonstrated that the proposed method could work in the presence of adversarial noise and irrespective of the model accuracy [82].

The pixel contributions visualized as heatmaps can help to verify the accuracy of the classification model [83]. A general solution using pixel-wise decomposition termed Layer-wise Relevance Propagation (LRP), was proposed by Ref. [83]. LRP has been used in different computer vision applications and originally assigned importance scores or relevancies to the different input dimension contributions to the model decision [56]. The LRP has some limitations such as the heatmaps were noisy and could be non-discriminative [84]. Selective LRP (SLRP) was proposed by combining gradient-based and relevance-based methods to overcome the limitation by producing clearer heatmap [84]. The model was validated on the ILSVRC2012 dataset [84]. The model was evaluated through a comparison with LRP, Contrastive LRP (CLRP), and Softmax Gradient LRP (SGLRP), and found to provide less noise, better class discrimination and held entire target objects [84].

Task-specific attention maps were generated by modelling CNN top-down attention and a new backpropagation scheme called Excitation Backprop [85]. The model was evaluated by measuring accuracy of localization task on MS COCO, PASCAL VOC07 and ImageNet datasets [85].

Deep Learning Important FeaTures (DeepLIFT) works using back-propagation of the neuron contributions in the network to every feature of the input. It compared each of the neuron activations with the reference activation and contribution scores were calculated based on the difference between the two values [86]. Using differences lets the information through even in case of zero gradient [86]. The method computed a 'difference-from-reference' to allow information to pass even with a zero gradient, and was evaluated on models trained with simulated genomic data and MNIST [86].

Deconvolutional Network (deconvnet) were proposed for unsupervised learning by changing the filters at various input patterns that could capture the mid and high-level features generalizing different classes [111]. Using a standard CNN, the proposed deconvnet mapped the feature activations in the intermediate layers back to the input pixel space to determine which image region caused that activation [87]. A deconvnet essentially performs the function of a CNN in reverse to map features to pixels [87]. It was shown through visualizations that the features are not random or uninterpretable patterns but rather have class discrimination and increasing invariance between layers [87].

Testing with Concept Activation Vectors (TCAV) used directional derivatives in order to capture the degree of a user-defined concept, such as presence of stripes, for the image classification, for zebra images [88]. The aim of the schemes was to provide explanations which are more human friendly [88].

The Deep Taylor decomposition utilized network structure to back-propagate the explanations from output to input layer [89]. Each neuron was regarded as a function to be decomposed on its input variables [89]. The method was evaluated on MNIST and a large convolutional network [89]. The generated heatmaps helped the role of input pixels for an unseen classification task [89].

SHapley Additive exPlanations (SHAP) was proposed as a unified framework for understanding complex model predictions by assigning each feature an importance value by combining six existing methods [90]. Classic Shapley Value Estimation is based on game theory for computing explanations [90]. The evaluation of proposed technique was found to be more consistent with human understanding and intuition with an improved computational efficiency compared to the other selected approaches [90].

An algorithm motivated by Self-Organizing Map (SOM), termed as Locality Guided Neural Network (LGNN) was proposed, that preserved the locality between neighboring neurons at each DNN layer [91]. The method was demonstrated using a CNN but can be applied to other DNNs [91]. The clusters of similar filters from one layer to another can highlight the filters responding to similar semantic concepts [91]. The backpropagation step allowed for the information to be shared between the neurons within the same layer [91].

3.4. Explanation by example

It is easy to understand a model's prediction by a user if similar example images can be provided from the dataset [28]. Use of exemplars, that is, the images representing each image class were used for image classification, in a system comprised of an association network and a classifier [112]. The input image was transformed using the association network to an associative image by the DNN. The similarity between the exemplars and the associative images explained the classifier decision [112]. Twinning of an opaque ANN model with a more interpretable model was proposed in an approach termed post-hoc explanation-by-example [113]. The aim was to use a more interpretable model twin for a DNN (black box), and map the feature map to the former from the latter. Evaluation of the proposed twin system was extensively carried out by creating four different twin models, such as CNN-case based reasoning (CNN-CBR) [113].

Image captioning can help in understanding the model's predictions and also help the practitioner with the diagnosis. Automatic generation of text corresponding to an image can aid the understanding of the image scene [92]. An attention-based model was proposed for caption generation that automatically learnt the content of an image [92]. The model was evaluated on Flickr8k, Flickr30k and MS COCO datasets [92]. Two forms of attention were used, a stochastic "hard" and a deterministic "soft" attention [92]. The proposed technique exploited learned attention for better interpretability, well-aligned to human intuition on the datasets with METEOR and BLEU metrics [92]. An algorithm for automatic location learning was proposed to use the location specific labels from the textual reports and a Bi-directional Long Short-Term Memory Recurrent Neural Network [114]. The algorithm comprised of two parts, a text to bounding box system, and a Guided Attention based Inference Network (GAIN) which used bounding box labels to improve the image level classification [114]. RSNA Pneumonia dataset was used for text to bounding box system, and GAIN network on MIMIC and NIH dataset with the in-house generated reports [114]. It was shown that the combined approach achieved better attention maps and a higher classification accuracy [114].

Image captioning was used to generate textual description for an image and involved computer vision and NLP, however the performance of image captioning improved with deep learning especially the encoder-decoder model [94]. The proposed model comprised of two parts; generation part for image caption generation, and an explanation part for the caption words for the region weight matrices [94]. MS COCO and Flicr30K datasets were used with BLEU and METOER as evaluation metrics [94].

Given a pair of image and textual questions, Visual Question Answering (VQA) techniques aim to generate the answers and are applicable to the highest CNN layers [115]. A Hierarchical Feature Network (HFnet) was proposed that showed that the low-level layer features are superior in comparison to low-level semantic questions [115]. Explaining Visual Classification using Attributes (EVCA) was proposed providing an explanation for classification prediction and a class label [116]. The system generated a textual justification with an RNN and image analysis with a CNN [116]. The system provided significant visual attributes for an expert, and extraction of images similar to input image for non-experts [116].

Explainable Neural Network (XNN) was proposed for explainability of DNN predictions by learning a DNN layer's high dimensional activation vector using an embedding to a low dimensional explanation space. A Sparse Reconstruction Autoencoder (SRAE) was proposed to learn the embeddings to the explanation space [95]. The DNN for prediction was associated with an XNN attached to the DNN intermediate layer [95]. The explanation space visualization with low-dimensional features was shown to improve the human machine communication [95]. The datasets CUB-200-2011 and Places365 dataset were used for a quantitative evaluation [95].

Model Agnostic Globally Interpretable Explanations (MAGIX) technique learns the if-then rules for explainability of models for classification problems [96]. The proposed method used genetic algorithm for determining conditions and evolving the rules [96]. The redundant rules were eliminated by ordering the rules, and considering the precision, and cover [96]. The method was evaluated on four datasets and a random forest classifier with 500 trees [96]. A hybrid Deep type-2 fuzzy system (D2FLS) for XAI was proposed to address the high dimensional input challenge of a DNN [108]. The method combined the autoencoders idea with interpretable type-2 fuzzy systems [108]. The fuzzy logic system used if-then rules and trained layer by layer to learn the important features similar to the stacked autoencoders [108].

Surrogate model is an explainable and simpler model that can be used to approximate the predictions of a black-box model which is otherwise difficult to interpret directly. An ontology describes a domain of interest using an appropriate logical language [98]. Use of ontologies was proposed to improve the explainability in the form of decision trees [98]. An algorithm TREPAN Reloaded was introduced to extract the surrogate model of decision tree for a DNN [98]. The evaluation of the proposed method showed that using the semantic information can improve the human understanding of explanations [98].

The method Black Box Explanations through Transparent Approximations (BETA) utilized an objective function to enable learning of compact decision sets for explaining the working of the model in the feature space [97]. The proposed method was evaluated and compared against LIME, Interpretable Decision Sets (IDS) and Bayesian Decision Lists (BDL) and the accuracy was found to be higher than IDS and BDL [97].

Graphs can capture complex information in a natural manner, especially for applications where there is an interplay of diverse information and relationships [100]. Counterfactual graphs were proposed to enable an exploration and interactive human-in-the-loop automated decision pipeline [100]. Knowledge graphs are a data structure that represent entities as nodes and their relationships using edges or links between them, organized as ontological schema [117]. A survey of knowledge graphs as a tool for XAI is provided covering rule-based machine learning, image recognition, recommender system, predictive tasks, and NLP [117]. A knowledge graph based eXplainable Neural-symbolic learning (X-NeSyL) methodology was proposed to learn deep and obtain symbolic representations [20]. The explainability measure was used for determining the alignment of the human expert and machine explanations [20]. A knowledge infused learning (K-iL) proposed for knowledge graphs can be used with DNN [118].

The knowledge was extracted from a trained DNN and represented as knowledge graph making it easier for explainability of the whole model [119]. The method was evaluated through image classification by determining the most significant neurons for predicting each class, and graph analysis to determine the classes group that were similar [119]. The graph represented the correlation between the activation values by the graph nodes and the graph edges (between input layer and hidden layer), and were plotted as heatmaps for each neuron [119].

A model decision can be supported by a counterfactual explanation of that prediction to describe the smallest change to the feature values that would change the prediction, and this could be used for both negative and positive automated decisions [101]. Conceptual and Counterfactual Explanations (CoCoX) was proposed to explain CNN model decisions based on sematic level features, termed fault-lines in Cognitive Psychology [103]. The nature of explanations being concept-based and counterfactual, were found to be relevant for both experts and non-experts [103]. An image classification task was chosen for the human evaluation of the proposed model for 'justified trust' and 'explanation satisfaction', and demonstrated better performance compared to other techniques [103].

Influence function is a technique from statistics and its use was proposed to identify training points responsible for a given prediction by an algorithm [104]. The influence functions measured the effect of local and small changes and were demonstrated to be useful for a variety of

applications, adversarial training examples, fixing mislabeled examples, and model understanding [104].

It was argued that simply presenting similar examples is not enough but rather criticism should also be presented, and Maximum Mean Discrepancy (MMD)-critic based on Bayesian model criticism framework for aiding interpretability was proposed [120]. Bayesian Model Criticism (BMC) can help identify how a model fails to explain data and thus aids model development [120]. An XAI approach using BEN (combinatorial software fault localization scheme) generated counterfactual explanations for the model predictions [121]. BEN was used to identify the significant image features that determined the prediction and if removed may change the predicted class [121]. The approach was validated using VGG16 model and could derive counterfactual explanations for 44 out of 50 images [121]. Comparison with SHAP showed that the proposed approach was contributing positively to the original decision [121].

What-if interactive questions can be used for building self-explaining systems. These questions are termed counterfactuals and are gaining in importance for XAI [100]. Plausible Exceptionality-based Contrastive Explanations (PIECE) was proposed for XAI comprising of a CNN for which explanations were required and a Generative Adversarial Network (GAN) generating semi-factual or counterfactual explanatory images [106]. It was shown that PIECE generated the best semi-factuals with a higher L_1 distance (test to explanation image) compared to other similar methods, such as Min-Edit [106]. The reason for image misclassification were investigated by generating counterfactual images using GAN [122]. It was shown that the attribution maps of the counterfactual images help in interpreting the misclassified images [122].

Bayesian teaching is a method for selecting examples and is based on the interactions between a teacher that provides examples to explain, and a learner explainee that draws inferences [105]. In this explanation-by-examples approach, the AI model and the training data is provided as input, and produces a small subset of training images affecting the inference [105]. Bayesian Teaching was proposed for XAI for modelling the human explainee, to validate explanations by determining the shift in explainee's inference [105]. Explainability can be quantified by building human learning models. In explanation by Bayesian Teaching, providing a probabilistic model, dataset, and an inference method, explained the model inference by returning a subset of examples [45]. A combination of explanation example and saliency maps were proposed for explaining pneumothorax diagnosis in CXR images using Bayesian Teaching [8]. The two explainability approaches are complementary, in that, saliency map highlights the pixels with significant contribution to the model prediction and explanation-by-examples selects the training images likely for prediction in Bayesian Teaching [8].

3.5. Libraries, tools, and frameworks

The tools and libraries for supporting XAI visualization can simplify interacting and interpreting of the DNN for the researchers and lay persons. The relevant tools are discussed in this section and a summary is provided in Table 2.

The *iNNvestigate* library addresses the challenge of a systematic comparison between the different DNNs through a common interface [123]. The implementations for LRP, PatternNet, and PatternAttribution besides many others such as SmoothGrad, GuidedBackprop are provided [123]. All of the implemented methods perform backpropagation and a quantitative evaluation using perturbation analysis [25] is provided [123].

A Visual Analytics Framework termed explAIner was developed as a TensorBoard plugin for XAI [124]. The Interactive and Explainable Machine Learning framework had three stages of model understanding, diagnosis of any model limitations using different XAI methods, and refinement of the models [124]. The framework was qualitatively evaluated on MNIST dataset with different types of human users and received very positive feedback [124].

Table 2
XAI Libraries, tools, and Frameworks.

Tool	Supported XAI Techniques	Description	Ref
iNNvestigate	Saliency Maps, SmoothGrad, IntegratedGradients, Deconvnet, GuidedBackprop, PatternNet, PatternAttribution, LRP, DeepTaylor	A common interface for a systematic comparison between many XAI backpropagation methods	[123]
explAIner	LIME, CAV, LRP, DeepTaylor, Saliency, Gradient, DeepLIFT, Grad-CAM, DecConvNet etc.	An interactive visual analytics framework integrated in TensorBoard for users to determine model limitations, and perform model optimization	[124]
Neuroscope	Gradient based methods, such as, CAM, Grad-CAM, Guided Grad-CAM, Grad-CAM++, Saliency Maps, GuidedBackProp	Visualizations for the CNN layers for image classification and semantic segmentation, supports networks created and trained with Keras and PyTorch	[125]
Interactive Graphical User Interface	Interactive visualizations for heatmaps but without support for a particular XAI technique	CNN and RNN model visualizations for different layers and inputs	[126]
Vertex XAI	Supports Sampled Shapley, Integrated Gradients, and XRAI (eXplanation with Ranked Area Integrals) methods	Unifies AutoML with AI as a unified API, and user interface. Provides example-based explanations to help users to refine data for improving model performance. Image classification and object detection are supported	[127]
SageMaker Clarify	Supports SHAP, PDP	Uses a model-agnostic feature attribution approach, heatmaps for feature importance are generated for understanding the model behavior, image classification and object detection are supported	[128]
IntrepretML toolkit	Glassbox models (Linear models, rule lists, GAM), black-box explainability (PDP, LIME etc.)	Open source Python package	[129]

An AI toolbox, Neuroscope, was proposed for visualizing classification and segmentation of CNN through an easy to use GUI providing visualizations of CNN layers [125]. The model's internal view can be visualized by providing insights into the predictions [125]. Different visualization methods can be selected for classification and segmentation, such as Grad-CAM, activation map and saliency map [125].

A web-based interactive approach to visualize the CNN allows to view the data flow and model architecture, the weights, layer processing and interpretable aspects of the whole model [126]. The interface provides visualization techniques for CNN, and RNN [126]. For CNN model, the weights for convolutional, pooling and fully-connected layers could be viewed [126]. The interface was evaluated using surveys for both experts of DNN and non-experts and showed the approach to be effective [126].

The major cloud platforms provide their own support for developing and understanding XAI models. The Google Cloud Platform (GCP) provides Vertex XAI that can help understand the classification model's outputs [127]. The Vertex AI supports both the Automated Machine Learning (AutoML) and custom trained image models [127]. The

Amazon Web Services (Amazon Cloud) has Amazon SageMaker Clarify model to help ML developers and other stakeholders to understand how the ML models make predictions [128]. Clarify uses an implementation of SHAP and produces PDP to depict the feature effect on the predicted outcome [128]. Microsoft provides IntrepretML toolkit for XAI for white-box and black-box models. An open-source toolkit Fairlearn provides integration with Azure Machine Learning for the SDK and the AutoML graphical user interface. The users can understand the models by the major influences and domain experts can also validate their models [129].

The AutoML is a powerful concept and a mechanism that abstracts the complex details required in developing an efficient DNN architecture by simplifying it as a search of the best DNN model through Network Architecture Search (NAS). This makes it simpler, cost-effective and quicker to develop an efficient DNN architecture.

3.6. Evaluation of explainability

While it is important to design and incorporate explainable techniques for imaging, it is equally important to check the performance and quality of the explanations to determine if it suffices for a given application and to evaluate it against a benchmark. Some work has been reported in the literature to evaluate the explanation quality to determine the XAI method suitability for different applications. Despite the large number of existing explainable models, there is no agreement on techniques for measuring or evaluating the explainability performance of an AI model. The evaluation methods can be classified as quantitative and qualitative. The quantitative approaches compute metrics for evaluation, whereas the qualitative approaches often use humans to rate the explainability. Compared to qualitative approaches through visual observations and comparisons, quantitative approaches can provide more objectivity and reliability in the process. A summary of the evaluation techniques is provided in Table 3.

Although heatmaps can be evaluated qualitatively, there is a requirement for an objective comparison [130]. The authors proposed two quantitative metrics termed, relevance mask accuracy, and relevance mass accuracy for evaluating visual explanations [130]. A framework (CLEVR) with visual question answering was proposed to compare ten explanation methods that can provide insights in understanding the different methods [130]. The proposed VQA-based evaluation was concluded to be realistic, and selective compared to the previous approaches. Among the evaluated XAI methods on the evaluation task, Deconvnet and Grad-CAM were found to be least accurate, whereas LRP, Integrated Gradients and Guided Backpropagation were found to be the most accurate methods [130]. Such evaluation techniques may grow in importance with the maturity and adoption of the XAI techniques.

Table 3
XAI EVALUATION TECHNIQUES.

Technique	Focus	Description	Ref
Insertion, deletion	Evaluation of interpretability	Insertion and removal of pixels	[72]
CLEVR	Quantitative evaluation of ten methods	VQA-based evaluation, database, framework and two metrics were proposed	[130]
GWAP	Improve human understanding of AI	Games generate useful data from the gameplay	[131]
Methodology	Saliency Methods	Determine suitability of methods to tasks	[107]
Taxonomy	Evaluation of interpretability	Proposed the three evaluation approaches	[51]
CMIE	Customizable model evaluation	In-model and post-model information for generating multi-dimensional interpretability evaluation	[132]
Modality-specific Feature Importance (MSFI) metric	Multi-modal medical imaging	Evaluation of 16 post-hoc methods	[133]
Z-inspection	Trustworthy AI	General inspection process	[134]

Two causal automatic evaluation metrics, deletion and insertion, were proposed for measuring explainability [72]. The deletion metric worked by removing the image information to change the model's decision, where removal of the important pixels will cause a drop and low Area Under the Curve (AUC) implies a good explanation. Similarly, the insertion metric worked by introducing pixels with a higher AUC corresponding to a better explanation [72]. The insertion and deletion as evaluation metric was used to compare the proposed SIDU method with RISE [76].

XAI was integrated with Games With A Purpose (GWAP) to assess the AI interpretation by humans [131]. GWAP are useful for validating large datasets for online gameplay [131]. A multiplayer GWAP was created for explaining DNN for image recognition [131]. A methodology for evaluation of explanation approaches was proposed in Ref. [107]. Although the approach could be applied to any explanation technique, the authors provided the results on saliency methods for neural network image classification [107]. The model parameter randomization and data randomization tests were proposed which have a wide applicability for explanation methods [107]. The results showed that a visual inspection of the explanations in itself may lack sensitivity to the model and data [107].

A taxonomy of the evaluation methods was provided as: Application-grounded evaluation, Human-grounded metrics, and Functionally-grounded evaluation [51]. Application-grounded evaluations involve humans with a real application, for example, clinicians on a diagnosing task can be provided with an evaluation of the model against the clinician's diagnosis. Human-grounded metrics allow lay humans to conduct simpler experiments, appropriate for explanation quality in general. Functionally-grounded evaluations does not involve humans and uses proxies for explanation quality with formal definition of interpretability [51].

The Customizable Model Interpretation Evaluation (CMIE) is a set of customizable evaluation methods of using the in-model and post-model information to generate multi-dimensional interpretability evaluation of the CNNs with different structures [132]. The model is customizable, can be used for any black box model and the model reports are interpretable by non-technical persons [132]. The model was evaluated both qualitatively and quantitatively using Interpretation trees as a surrogate model [132].

Although heatmaps are common in explanation, heatmaps' performance in explaining the multi-modal medical images are not well understood [133]. A Modality-specific Feature Importance Metric (MSFI) was proposed to aid in understanding if the 16 post-hoc selected methods fulfill the clinical requirements [133]. MSFI metric had better results compared to the other chosen post-hoc heatmap algorithms for multi-modal imaging tasks [133].

Z-inspection can be used for different AI domains such as healthcare and is a general inspection process [134]. It was proposed for evaluating trustworthy AI and can help inform the stakeholders about the technical, ethical and legal risks before the production of an AI system [134].

4. Explainability for biomedical imaging

Biomedical imaging covers a wide spectrum of imaging modalities, such as conventional x-rays and ultrasound. Medical image acquisition often requires very costly equipment and can involve substantial time. The availability of datasets for thoracic Computed Tomography (CT), chest radiographs and mammograms has driven the research for lung diseases and breast cancer [6]. The datasets, DNN models and XAI techniques cover a range of disease investigation and involve specialties such as radiology, and ophthalmology, however, radiology seems to have a large share and therefore a higher need for application of XAI

techniques [135].

In this section, we group use of XAI techniques according to different image modalities and provide a summary of techniques in Table 4.

4.1. Conventional imaging/X-rays

Conventional CXR is an inexpensive and readily available imaging modality [37,41]. CXR have been extensively used for COVID-19 diagnosis. COVID-19 detection on CXR images with a fine tuned EfficientNet-B7 achieved an accuracy of 95.01% and AUC scores of 0.950, which were better in comparison to many other DNN approaches for COVID-19 detection [161]. Grad-CAM was used for understanding the contribution of significant regions for the model's prediction [161]. A multiclass classification of NIH CXR dataset images into three classes, COVID-19, pneumonia, and normal used VGG-16 and VGG-19 CNN architecture for classification and Grad-CAM for interpretability [164]. It was found that any foreign substance in the lung cavity resulted in misclassification of images, and sometimes a correct classification was due to the wrong reasons [164]. A framework for multiclassification into three classes with an ensemble of CNN and Naive Bayes as a classifier used Grad-CAM for visualization [37]. XAI was used for the detection of pulmonary disease and COVID-19 with CXR images using transfer learning with VGG-16 model [251]. Two models were used, first to discern a pulmonary disease and then to establish if COVID-19 is manifested [251]. Grad-CAM was used to generate activation maps for COVID-19 highlighting the correct areas in the activation maps [251]. For the localization and severity grading of COVID-19, activation maps could be unsuitable for clinical use as these could provide inconsistent localization [159]. Infection maps were proposed for COVID-19 diagnosis from CXR images to overcome the issues [159]. The infection maps could also be used to determine the progression using the time series CXR data [159].

Counterfactual Generative Network (CGN) was proposed for medical image visualizations on a CXR dataset for COVID-19 [165]. The counterfactual lesion predictions were embedded as prior conditions to generate counterfactual lesion images [165]. In addition to the visual attribution maps, U-Net was used as a generator comprising an encoder and decoder for generating the counterfactual images [165]. Eight CNN models were used for COVID-19 CT image classification with LIME to help identify specific differentiating features [157].

An Explainable DNN ensemble-based system, DeepCOVIDExplainer was proposed for COVID-19 detection on CXR images from 'COVIDx v2.0' dataset [162]. The explainability of the models was investigated through LRP, Grad-CAM and Grad-CAM++ to identify the regions significant for the classification [162]. The best performing individual DNN architectures were VGG-19 and DenseNet-161 and the heatmaps generated by Grad-CAM++ were the most reliable in comparison with Grad-CAM and LRP [162].

An Explanatory clustering framework was proposed to group CXR images based on the extent of infection and severity level to aid the clinicians' decisions. DeepSHAP was used for visualizing the discriminating regions for the disease severity [163]. The model was used for COVID and pneumonia classification and achieved accuracies of 95% and 97% respectively [163]. VGG-19 with a transfer learning approach was proposed for COVID-19 detection by using the CXR and CT images from the three datasets, COVID-chest X-ray, SARS-COV-2 CT-scan and Chest X-ray images (Pneumonia) dataset [158]. The model performance was more accurate on CT compared to CXR images, and had an accuracy of 95.61% in COVID-19 detection. The model predictions were visualized using Grad-CAM [158]. A CNN model CovXNet was used for binary and multiclass classification with depth-wise convolution with varying dilation rates that aided extraction of diverse features from CXR [160]. The multiclass classification (COVID, normal, viral or bacterial pneumonia) achieved an accuracy of 90.2%. Grad-CAM was used to determine the CXR image regions contributing to the model's decision [160].

A Bayesian network, termed MulNet was proposed for pneumonia detection by exploiting multisource data combining DNN and an explainable model [166]. The trained CNN model provided binary classification for pneumonia, and was combined with 7-dimensional vector derived from the clinical reports, resulting in an 8-dimensional vector [166]. The proposed model was compared with SVM, RF, and Decision trees [166]. The authors concluded that more detailed and large-scale knowledge graphs could further improve the classification accuracy [166].

An attention guided network was proposed for lesion locations for multi-label thoracic disease classification in CXR images [168]. The method provided an average AUC score of 0.824 across all 14 pathologies. Grad-CAM was used for lesion localization [168]. A segmentation-based deep fusion network was proposed with two CNN classification models as feature extractors for extracting the discriminative regions which were used for thoracic disease classification on CXR images. The method achieved better classification performance compared to the other approaches based on ROC [169]. The lesion detection was assessed using CAM and was found to be reliable [169].

Explainability could be enhanced by gamification. Emergent languages involve multi-agent communication games with a sender/receiver coordinate to complete a task and have been shown to enhance the neural network capabilities [252]. Emergent Symbolic language are the cornerstones of Good Old-Fashioned AI (GOFAI) systems which are explainable [252]. The proposed system was used for deep medical image classification using CheXpert dataset and the results showed that the finer details of the input images were manifested in the generated symbols [252].

Paediatric pneumonia diagnosis was proposed using an ensemble of CNNs [167]. The proposed XAI technique was used to combine heatmaps from each of the ensemble models [167]. The method was used to determine the probability that CXR has consolidation or non-consolidation (other infiltrates), which could be helpful to the clinicians [167]. The ensembles were useful for improving the classification results for both bacterial and viral pneumonia [167]. However, the dataset was small with low-resolution images [167].

A pre-trained VGG16 model and a custom model were used for the classification of viral and bacterial pediatric pneumonia in CXR images [171]. The model activations were visualized using LIME and Grad-CAM and found to highlight the discriminative regions in the CXR [171]. Another study for pediatric pneumonia diagnosis used an ensemble of five different DNN models for classification of CXR on the XrPP dataset [167]. The study showed that the ensemble provided better results compared to transfer learning and achieved an AUC of 0.92 [167]. The heatmaps were obtained by combining the individual heatmaps for each of the ensemble models [167].

The explanations using Bayesian Teaching on CXR were obtained for pneumothorax diagnosis and showed that it can help medical experts to predict AI decisions [8]. The study combined the saliency maps and Bayesian Teaching and explored the relationship between AI classification, radiologist's diagnosis, and radiologist's predictions [8]. The evaluation used AlbuNet-34 and Kaggle dataset SIIM-ACR Pneumothorax Segmentation and had eight radiologists as participants [8].

The malignancy prediction in a lung cancer dataset were derived from capsule vector weights indicating the contribution of a visual attribute to nodule's malignancy [15]. The information from the child capsule to the parent capsule was routed using a new routing sigmoid function [12]. CXR images were used for lung cancer prediction using the proposed X-Caps XAI technique generating malignancy scores with explanations at human-level that were better interpretable by the radiologists [15].

Cardiomegaly was diagnosed using CNN and explainable feature map for explainability with CXR images [236]. CXR images were used for cardiac hyper-trophy readings [236]. Functional accuracy of CNN was used for model evaluation [236]. The feature map showed positive values for the portions with significant influence on the disease diagnosis [236].

Table 4
The Use of various Imaging modalities and XAI techniques for disease diagnosis.

Medical Specialty	Imaging Modality	Technique	Description	Ref
Neurology	CT angiography images	RNN, RFNN and autoencoders, Grad-CAM++	Radiological image biomarkers for stroke outcome prediction, MR CLEAN Registry dataset	[136]
	SPECT DaTSCAN	VGG-16, LIME	Early diagnosis of Parkinson's disease, Parkinson's Progression Markers Initiative (PPMI) database	[137]
	Single Photon Emission Computed Tomography (SPECT)	SHAP was found best performing out of six interpretation methods for four selected DNN models	Parkinson's Progression Markers Initiative(PPMI) database	[138]
	SPECT	Custom CNN, LRP	Classification of dopamine transporter (DAT) SPECT images	[139]
	Brain fluorodeoxyglucose (FDT) PET	3D CNN, t-SNE visualization	Alzheimer's and Parkinson's disease, ADNI database	[140]
	CT & Magnetic Resonance Imaging (MRI)	Classification and segmentation, SHAP	Study cohort of 155 people, CT scans with Siemens instruments, MRI with Siemens scanner	[141]
	MRI	3D CNN, LRP	Multiple Sclerosis on Alzheimer's Disease Neuroimaging Initiative (ADNI)	[142]
	MRI	CNN and SpinalNet, LIME	Alzheimer's Disease & gene expression multimodal data	[143]
	MRI	CNN, LRP	Alzheimer's disease, Alzheimer's Disease Neuroimaging Initiative (ADNI)	[144]
	MRI	DNN, Activation pattern maps	Voxel-based, region-based and patch-based on Alzheimer's Disease Neuroimaging Initiative(ADNI)	[145]
	MRI	3D CNN ensemble, SmoothGrad	Identification of brain regions contributing to brain aging	[146]
	MRI	CNN	Hierarchical Occlusion (HihO), Parkinson's Progression Markers Initiative (PPMI) cohort of diffusion weighted magnetic resonance imaging (DW-MRI)	[74]
	MRI	CNN, CAM	Parkinson's disease	[147]
	MRI	3D CNN, CAM	Parkinson's disease, PPMI dataset	[148]
	MRI	Subtractive Spatial Lightweight Convolutional Neural Network (SSLW-CNN), CAM	Brain MRI RadioPaedia database	[149]
	MRI	CNN, Grad-CAM and Guided Backpropagation	Brain Tumor, BraTS dataset	[150]
	MRI	CNN, Proposed MSFI visualization	Brain tumor, Glioma grading into lower-grade and higher-grade, BraTS 2020 dataset	[133]
	MRI	Three CNN models, Network dissection and Grad-CAM	Brain tumor segmentation, BraTS 2018 dataset	[151]
	MRI	CNN, Grad-CAM	Transfer learning with ResNet50, IXI dataset	[152]
	MRI	Susceptibility-Weighted Images using Relevance Analysis	Multiple Sclerosis (neuroimmunological disease) with DeepLift heatmaps	[153]
	MRI	2D CNN, Grad-CAM	Binary classification to identify wrap-around and Gibbs ringing in low-field brain MRI, IXI and T1-weighted brain images	[154]
	Ultrasound	CNN, Grad-CAM	Classification of fetal brain abnormalities, Private dataset	[155]
	Immunohistochemically-stained archival slides	CNN, Grad-CAM and Feature Occlusion	Classification of Alzheimer's disease	[156]
Respiratory	CXR, CT	Eight CNN models, LIME	NasNetMobile had the best results, open-source Kaggle datasets. 400 CT & CXR images	[157]
	CXR, CT	VGG-19, Grad-CAM	Binary classification on COVID-chest X-ray, SARS-COV-2 CT-scan and Chest X-ray images (Pneumonia) dataset	[158]
	CXR	Five DNNs, Infection map	Joint localization and severity grading, CXR dataset QaTa-COV19	[159]
	CXR	CNN CovXNet, Grad-CAM	Multiclass classification with depth-wise convolution with varying dilation rates, Three datasets	[160]
	CXR	EfficientNet B7, Grad-CAM	BIMCV COVID19+, RSNA, NIH, Montfort, and few other datasets	[161]
	CXR	Four CNN base learners with Naïve Bayes as meta-learner, Grad-CAM	Multiclass classification, Kaggle RSNA dataset	[37]
	CXR	DNN ensemble model, Grad-CAM++, LRP	Multiclass classification for COVID-19, COVIDx v2.0 dataset	[162]
	CXR	VGG-19 based model, DeepSHAP	Explanatory clustering framework, public datasets	[163]
	CXR	VGG-16 and VGG-19, Grad-CAM	Model fine tuning, NIH CXR dataset	[164]
	CXR	Counterfactual Generative Network	Explanations for thoracic images, Chest X-ray 14 and VinDr-CXR datasets	[165]
	CXR	Bayesian Teaching	Explaining pneumothorax diagnosis, SIIM-ACR Pneumothorax Segmentation Dataset-Kaggle	[8]
	CXR	DenseNet121, Bayesian network	Used a combination of CXR images and clinical data	[166]
	CXR	Ensemble of CNN, heatmaps	X-ray Pediatric-Pneumonia (XrPP), a public pediatric dataset of chest X-rays	[167]
	CXR	Capsule network X-Caps	Lung cancer, LIDC-IDRI dataset	[15]
	CXR	LLAGnet, Grad-CAM	Lesion locations for multi-label thoracic disease classification, ChestX-ray14 dataset	[168]
	CXR	Segmentation-based Deep Fusion Network, Two CNN classification models	Thoracic disease classification, CXR 14 dataset	[169]
	CXR	CNN, CAM	Binary classification of chest radiographs, 313719 institutional images	[170]
	Pediatric CXR	CNN, LIME and Grad-CAM	Classification as bacterial and viral pneumonia, public dataset	[171]
	Pediatric CXR	CNN ensemble, Combined heatmaps of the CNN ensemble models	Pediatric pneumonia diagnosis, XrPP dataset and another public dataset	[172]
	CT	t-SNE, Grad-CAM	SARSCoV-2 CT and COVID19-CT datasets	[173]

(continued on next page)

Table 4 (*continued*)

Medical Specialty	Imaging Modality	Technique	Description	Ref
	CT	DenseNet201, Convolutional LSTM and online tool at http://perceivelab.com/covid-ai.	Prior lung and lobe segmentation for COVID-19	[174]
	CT	LIME, SHAP, U-Net	CC-CCII data, and data from four hospitals, Explainable Diagnosis and Slice Integration module	[141]
	CT	Joint Classification and Segmentation (JCS), Activation mapping	COVID-CS dataset	[175]
	CT	CycleGAN activation maximization	Lung lesion malignancy classification on LIDC-IDRI dataset	[176]
	CT	Multi-scale Attention Network, Proposed two visualization techniques based on Grad-CAM	Classification of seven types of pulmonary textures on high resolution CT images	[177]
	CT	CNN and LSTM, Grad-CAM	Classification of Emphysema pattern, COPDGene dataset	[178]
	CT	CNN, Grad-CAM	Lung nodule malignancy prediction, Data from the National Cancer Institute Cancer Data Access System	[179]
	CT	CNN, Soft Activation Mapping for visualization	Lung nodule classification in low-dose CT images, LIDC-IDRI public dataset	[180]
	CT	3D CNN, Grad-CAM	Lung cancer prognostication, Seven datasets across five institutions	[181]
	CT	Deep Stacked Interpretable Sequencing Cell (SISC) architecture, Critical Response Maps for visualization	Classification of abnormal lung nodules for lung cancer prediction, LIDC-IDRI dataset	[182]
	CT	Four pre-trained DNN, Grad-CAM	Embedded low-quality chest CT images of COVID-19 pneumonia, public and private data	[183]
	Cytological images	Deep CNN, Grad-CAM	Classification of malignant cells, private dataset	[184]
Ophthalmic	Retinal fundus images	R–CNN, text report	Messidor-2, IDRiD, E-Ophtha dataset	[185]
	Retinal fundus images	CNN, Grad-CAM	APTOS, MESSIDOR, IDRiD dataset for Diabetic Retinopathy (DR) grading	[186]
	Retinal fundus images	CNN, Structural similarity of the visualization map	EyePACS and DIARETDB1	[187]
	Retinal fundus images	CNN, CAM	Quantitative assessment of image quality, SDRSP and IDRiD	[188]
	Retinal fundus images	Integrated Gradients	DR severity grading by ten ophthalmologists unassisted, grades only, and grades plus heatmaps	[189]
	Retinal fundus images	Custom DNN, Attention maps	Severity grading, Kaggle DR dataset, and five public test datasets	[190]
	Color Fundus Photography (CFP)	ExplAIn for segmentation and classification	Explanatory AI algorithm, OPHDIAT, EyePACS, IDRiD, DeepDR	[191]
	Retinal images	Patho-GAN	Kaggle Diabetic Retinopathy dataset	[192]
	Retinal fundus images	CNN, Grad-CAM	Segmentation and classification for Glaucoma diagnosis on mobile devices, Public datasets	[193]
	Retinal fundus images	CNN, Guided Grad-CAM	Laterality classification of fundus images, Private dataset	[194]
	Retina fundus images	Five CNN models, Grad-CAM	Glaucoma classification and localization, private dataset	[195]
	Retinal fundus images	ConvNet, Evidence Activation Mapping for visualization	Optic disc segmentation and disease localization, Feature aggregation at multiple scales for glaucoma diagnosis	[196]
	Retinal fundus images	DNN, TCAV	Interpretability of Diabetic retinopathy	[88]
	Retinal fundus Images	CNN, Grad-CAM and integrated heatmaps	Fundus Image Classification and Retinal Disease Localization, private dataset	[197]
	Optical Coherence Tomography (OCT)	Heatmaps, SqueezeNet, VGG-16	OCT-2017 dataset	[198]
	Optical Coherence Tomography (OCT)	Custom CNN model OCT-NET, CAM	Classification of diabetes-related retinal diseases, SERI-CUHK and A2A SD-OCT datasets	[199]
Breast cancer	MRI	CNN, Grad-CAM	DCE-MRI	[200]
	MRI	3D ResNet, CAM and COrrelation Attention Map (COAM)	Tumor classification and localization on DCE-MRI dataset	[201]
	MRI	SHAP	3-dimensional regression CNN for volumetric breast density, Dataset available on request	[202]
	Ultrasound	CycleGAN activation maximization	Breast lesion classification on BreastMNIST dataset	[176]
	Ultrasound	Grad-CAM	Public dataset	[203]
	Ultrasound	Bimodal model, Grad-CAM	Multimodal Multiview data augmented with heatmaps for malignancy, Data available on request	[204]
	Mammograms	Case Based Reasoning (CBR)	Breast Cancer Wisconsin (BCW), Mammographic Mass (MM) dataset, Breast Cancer (BC) datasets	[205]
	Mammograms	CNN and GAN, ICADx framework	DDSM dataset	[206]
	Histopathology images	Attention guided CNN, Proposed Guided activations	Classification on BACH microscopy dataset	[207]
	Histopathology images	Regression Concept Vectors (as an extension to TCAV)	Nuclei contrast and correlation were found relevant to the classification of breast tissue patches, Camelyon16 and Camelyon17 datasets	[208]
	Histopathology images	Four CNN trained on RoI, Occlusion for visualization	Multiclass classification, NIH projects' dataset	[209]
	Microarray datasets	Graph CNN, Graph LRP	Prediction of metastatic events in breast cancer, public dataset compiled from 10 microarray datasets	[210]
Dermatology	Dermoscopic RGB Images	CA-Net: Comprehensive Channel Attention CNN	Lesion segmentation, ISIC 2018	[211]
	Dermoscopic lesion images	DNN, LIME	Skin lesion classification, HAM10000 dataset	[212]
	Dermoscopic lesion images	Six CNN models, CAM	Skin disease diagnosis by fusing metadata and dermoscopy images, ISIC 2018 dataset	[213]

(*continued on next page*)

Table 4 (*continued*)

Medical Specialty	Imaging Modality	Technique	Description	Ref
	Dermoscopic RGB images	mutual bootstrapping deep convolutional neural networks (MB-DCNN) model, CAM	Lesion segmentation and classification, ISIC 2017 and PH2 datasets	[214]
	Dermoscopic lesion images	CNN, Feature maps for visualization	Binary classification of Skin Lesions, ISIC dataset	[215]
	Dermoscopic lesion images	Deep Image Priors, U-Net	Counterfactual explanations for ISIC dataset	[216]
	Dermoscopic lesion images	CNN and LSTM, Spatial attention visualization	Leverages the two sources of medical information with visualizations, ISIC 2017 and ISIC 2018 datasets	[217]
	Dermoscopic images	Concept Activation Vectors	Multi-modal melanoma classification framework ExAID, evaluated on PH2 and Derm7pt public datasets	[218]
Genitourinary	CT	Relational Functional Gradient Boosting (RFGB)	Non-CNN for Kidney cancer	[219]
	Ultrasound, MRI	Pre-trained DNNs, LIME	Binary classification for Prostate Cancer, Prostate MRI and US image data	[220]
	Pathological images	CNN, for Multi-scaled features	Cervical cancer detection, pathological images of a uterine cervix provided by Nagasaki University Hospital, Japan.	[221]
Gastrointestinal	Gastric X-Ray	Deep CNN, Grad-CAM	H. Pylori Infection	[222]
	CT	LSTM, Image captioning	MICCAI 2017 LiTS challenge dataset	[223]
	CT	CNN, Concept attribution	Liver tumor segmentation on LiTS database	[224]
	CT	Dual-Attention Dilated Residual Network (DADRN), Grad-CAM	Liver lesion classification, private dataset	[225]
	CT	CNN, Heatmaps	Liver lesion localization and classification, private dataset	[226]
	CT	3D DNN, Grad-CAM	AppendiXNet pre-trained on YouTube videos (Kinetics) and fine-tuned on 438 annotated CT scans	[227]
	MRI	CNN, Influence functions	Liver cancer lesion classification	[228]
	RGB Images	Four CNN models, Saliency, guided backpropagation, integrated gradients, input gradients, and DeepLIFT	Evaluation of cancer, Dataset-University Hospital Augsburg, Medizinische Klinik III, Germany, MICCAI	[229]
	Video Capsule Endoscopy (WCE)	LIME, SHAP, Contextual Importance and Utility (CIU)	Images obtained with Video Capsule Endoscopy for lesion detection, Red Lesion Endoscopy dataset	[230]
	Video Capsule Endoscopy	ResNet-34, CAM	Ulcer recognition on WCE dataset	[231]
	Histopathological images	Cumulative Fuzzy Class Membership Criterion (CFCMC)	Classification of colorectal cancer, Public dataset of Hematoxylin and Eosin (H&E) tissue slides	[232]
	Histopathological images	CNN, CAM	Differentiate between two subtypes of primary liver cancer, Cancer Genome Atlas' (TCGA) hepatocellular carcinoma (LIHC) and cholangiocarcinoma (CHOL) diagnostic FFPE WSI collections and Stanford University Medical Center dataset	[233]
	Histology	Four CNN models, CAM	Non-alcoholic fatty liver disease (NAFLD) and the progressive form of non-alcoholic steatohepatitis(NASH)	[234]
Musculoskeletal	Pediatric musculoskeletal radiographs	Five Deep CNN, CAM	Classification of pediatric musculoskeletal radiographs into five anatomical region, public images	[235]
Cardiology	CXR	ResNet, Explainable feature map	Cardiac hypertrophy classification, NIH CXR dataset	[236]
	CXR	Generative Visual Rationale (GVR) compared against other visualization techniques such as LIME	Congestive heart failure with private PACS dataset	[237]
	Coronary X-ray Angiography (CAG)	CNN and RNN, CAM and Grad-CAM	Stenosis classification and localization in X-ray Angiography, Multicenter dataset	[238]
	CT	3D CNN, Grad-CAM	Detection of coronary artery atherosclerosis, Coronary CT Angiography (CCTA) dataset	[239]
	CT	3D Attention Identical Dual Deep Network, 3D Grad-CAM	Coronary Artery Calcium (CAC) detection, ImageVU dataset	[240]
	MRI	Discovering and Testing with Concept Activation Vectors (D-TCAV)	Segmentation-MICCAI2017 dataset,	[241]
	MRI	3D CNN, CAM	Classification of brain dysmaturation from neonatal MRI, private data	[242]
	Ultrasound	CNN, Supervised Object Detection with Normal data Only (SONO)	Fetal ultrasound videos, the barcode-like timeline to capture clinical characteristics of each case.	[243]
	Electrocardiogram (ECG)	PSI, LIME, SHAP, with 1D CNN	MIT-BIH arrhythmia dataset	[244]
	ECG	Inception-Resnet-v1, Grad-CAM	Identification of local cardiac structures, and predict systemic phenotypes, Stanford Echocardiography Database	[245]
	ECG video data	Deep CNN, Grad-CAM	Morphological classification of mitral valve disease using eight CNN models, Tehran Heart Center Hospital dataset	[246]
	ECG	xECGNet, Attention maps	Detection and explainability of concurrent cardiac arrhythmias, CPSC2018	[247]
	ECG	Shapley, CNN, XGBoost	ECG-ViEW II	[248]
Fetal	Ultrasound	Various Saliency maps method	Regression CNN, HC18 dataset	[249]
	MRI	CA-Net: Comprehensive Channel Attention CNN	Segmentation of multiple organs from T2-weighted fetal MRI	[211]
Mortality prediction	Time series data	CNN, heatmaps	MIMIC-III database of Intensive Care Unit (ICU) records	[250]

Five deep CNN models were used for automatic classification of paediatric musculoskeletal radiographs of the five anatomical area (shoulder, elbow, pelvis, hand, and knee) and achieved a classification rate of 33 radiographs per second [235]. CAM was used to generate the heatmaps and the system provided good accuracy results despite a small dataset [235].

CXR is the most commonly used imaging modality for detecting COVID-19, or lung infections, however, in comparison to CT, it is less accurate. The most used explainable technique for X-rays seems to be Grad-CAM, as evidenced by Table 4. Although there are efforts to

incorporate explainability while analyzing the CXRs and X-rays, the algorithms are often not able to explain or mimic the exact experience of a human grader or radiologists which they utilize during X-ray examinations. The XAI can help in identifying all such factors which form the basis of a particular decision. This can also help in overcoming the challenges associated with the limited information present in an X-ray and actually highlighting how the density, shape, size and location of abnormalities affect the final diagnosis.

4.2. Computed Tomography (CT)

LIME was used to help identify the specific features that helped to differentiate CT images for COVID-19 and normal images using 8 CNN models [157]. CNN model for COVID-19 classification of CT images used t-SNE (dimensionality reduction) and Grad-CAM showing well separated clusters for COVID-19 and non-COVID-19 cases [173]. CT images of COVID-19 were classified with U-Net achieving better performance (accuracy, precision, specificity, and AUC) than other state-of-the-art methods, such as COVID-Net and COVNet [141]. LIME was used for explaining each prediction by dividing an image into super-pixels. SHAP method was also used for explainability to quantitatively estimate the contribution of each super-pixel [141].

A COVID-19 detection and lesion categorization system were proposed for CT-scans by combining segmentation and classification networks [174]. It was shown that the lung and lobe segmentation improved the classification results by over 6% [174]. A lung lesion malignancy classification was proposed using CycleGAN activation maximization for visualizations of the classifier's decision LIDC-IDRI dataset [176]. A Joint Classification and Segmentation (JCS) system was proposed for diagnosis and explainability of COVID-19 with CT diagnosis [175]. A large COVID-CS dataset was also constructed with patient and pixel level annotations [175]. The JSC identified and segmented the opacification areas, a basic CT feature of COVID-19 [175]. Activation maps were used for explainability of the predictions and Dice scores for segmentation [175]. The CT images from multiple sources were mixed to avoid overfitting and bias [175].

A DNN framework with four pre-trained DNNs was used for COVID-19 binary classification on CT images [183]. The framework was tested on low-quality CT images, and ResNet-50 performed best amongst the selected DNNs with an accuracy of 99.87% and a sensitivity of 99.58% [183]. Grad-CAM was used for visualizing the relevant regions in the model's prediction [183]. A multi-scale attention network was proposed comprising stacked attention modules and a multi-scale fusion module for classification of seven types of pulmonary textures on high resolution CT images [177]. An average classification accuracy of 94.78% was achieved. Grad-CAM was used to determine working of the proposed model [177].

A framework was proposed for segmenting CT volumes using a CNN and report generation using LSTM language model for liver tumor [223]. The public MICCAI 2017 LiTS challenge dataset was used for evaluating a supervised attention model for captioning [223]. A CNN and LSTM were used for classification of emphysema pattern on COPDGene CT dataset [178]. The model scores were calculated by combining DNN outputs and were compared against Grad-CAM visualisations and clinical parameters [178]. The model achieved agreement with visual emphysema scores in the test cohort [178].

CT images have been used for lung tumor localization and classification. A CNN ensemble was used for lung nodule malignancy prediction with three CNN architectures, each trained with seven different seeds for the initial weights [179]. The ensemble was found to improve the classification of the system and achieved an AUC of 90.29% in predicting nodules that would probably be diagnosed as cancerous in two years. Grad-CAM was used for visualizing the model activations for the input images [179]. Lung nodule localization and classification on low-dose CT images (LIDC-IDRI public dataset) was implemented using soft activation mapping for fine-grained features analysis with a CNN [180]. The visualisations of the lung features were obtained using CAM, Grad-CAM and the proposed visualization method. The results were incorporated to increase the radiologists' confidence level for clinical diagnosis [180]. A stacked interpretable architecture was proposed for the classification of abnormal lung nodules for lung cancer prediction on LIDC-IDRI dataset [182]. The proposed architecture provided better insights into the model's decisions and achieved better performance compared to the considered approaches. The proposed radiomic sequencer achieved correct predictions that can potentially improve clinical adoption [182].

A concept attribution method inspired by activation maximization (DeepDream) and similar to TCAV [88] was proposed for liver tumor segmentation on CT images LiTS database [224]. The method provided insights of the high-level feature importance in the trained CNN model [224]. CNN was used for automated localization and classification of liver lesions in spectral CT images. The RoIs from the localization tasks were used to train another CNN for the classification (healthy, hypodense metastasis, cyst) [226]. Heatmaps were used for visualisations and performed well for smaller lesion detection [226]. A Dual-attention Dilated Residual Network (DADRN) was proposed for liver lesion classification on a CT image dataset and achieved an accuracy comparable to RoI based methods [225], Grad-CAM was used for explanations of the classification results and the proposed system was found to identify the small size lesions [225].

A 3D DNN, AppendiXNet was proposed for detection of appendicitis with a small CT exam dataset [227]. The system was trained on YouTube videos (Kinetics) and fine-tuned on 438 annotated CT scans. Grad-CAM was used for the explainability of the model predictions [227]. The model performance was found comparable to similar 3D CNN models and had a high diagnostic performance for acute appendicitis detection [227].

Kidney cancer care was projected to cost around $5.1 billion in USA by 2020 [219]. Binary classification for renal cell carcinomas used large feature set with Relational Functional Gradient Boosting (RFGB) and provided better model performance in accuracy, F1-Score and AUC, compared to other techniques and a shallow neural network [219]. The data was from a study for multiphase Contrast-enhanced Computed Tomography (CECT) of the abdomen and pelvis with expert graded histology evaluation [219]. The radiomics based features comprised mean, contrast, image intensities, Fourier frequencies etc. RFGB found simpler rules for modelling the features and could be easily compared to a large single highly accurate model [219].

An adapted ResNet architecture was used on MR CLEAN Registry dataset of CT angiography images and determined that the automated DNN method was better than radiological image biomarkers for stroke outcome prediction and was helpful in the treatment selection [136]. Grad-CAM++ was used for visualization and the model was found to be useful for predicting stroke outcome [136]. Coronary CT angiography (CCTA) images were used for coronary artery atherosclerosis detection and localization on CCTA datasets with 3D-CNN [239]. The system was evaluated and showed an AUC of 0.91 [239]. The localization used Grad-CAM for deriving heatmap visualisations helpful to the physician in locating atherosclerosis location and likelihood [239]. 3D Attention Identical Dual Deep Network (AID-Net) was proposed for Coronary Artery Calcium (CAC) detection and 3D attention mechanisms were integrated for classification on ImageVU dataset [240]. 3D Grad-CAM was used for understanding the DNN prediction [240]. The proposed system achieved an AUC of 0.9627 and a detection accuracy of 0.9272 [240].

CT is used across many medical specialities such as respiratory and cardiology studies. Some genitourinary and gastrointestinal studies also use CT modality. There are different XAI techniques used for different medical imaging modalities but Grad-CAM seems to be the most widely used as evidenced by Table 4.

4.3. Magnetic resonance imaging (MRI)

Alzheimer is a neurodegenerative disease and affects around 10 million new cases worldwide [143]. Image and gene expression data were combined for a multimodal Alzheimer's disease detection [143]. CNN and SpinalNet were used for MRI images and KNN, Support Vector Classifier (SVC), and Xboost for the microarray gene expression data [143]. The significant genes were identified with LIME for explainability [143]. The ADNI MRI dataset was used for explaining Alzheimer's classification using LRP for visualizing CNN decisions [144]. It was shown that LRP was useful for explainability of Alzheimer's classification predictions and applicable to similar disease diagnosis using MRI data [144].

A method aimed at increasing the clinical accuracy for the early stages of Alzheimer's disease was proposed and showed that the voxel relationships in different regions can provide discriminative information for clinical diagnosis [145]. The method was validated on a weighted MRI dataset and provided better performance over Regional Mean Volume (RMV), and Hierarchical Feature Fusion (HFF) methods using the four metrics (accuracy, sensitivity, specificity, and AUC) for the binary and multiclass classification [145]. The method also provided representations of regional abnormalities for prediction interpretability [145].

Parkinson's is a neurodegenerative disease affecting the aging population. CNN was used to create the biomarkers for Parkinson's disease from Neuromelanin Sensitive Magnetic Resonance Imaging (NMS-MRI) and achieved better test accuracy compared to contrast ratio-based classification [147]. CAM was used to identify the relevant and most discriminative image areas [147]. 3D CNN was used for the Parkinson's disease detection and classification of MRI scans from the PPMI dataset [148]. The system achieved an accuracy of 95.29% and Receiver Operating Characteristics (ROC)-AUC of 0.98 for both classes. CAM method was used for the explainability of the model's predictions in selecting the relevant areas [148].

A framework used volumetric neonatal MRI for classification of brain dysmaturation with 3D CNN and CAM reporting a classification accuracy of 98.5% [242]. The brain aging was investigated by using 3D CNN ensemble for identification of predicting the contributory brain regions [146]. SmoothGrad implemented using iNNvestigate was used and the explanations were aggregated across samples [146]. From the aggregated explanations it was concluded that cisterns and ventricles governed the age predictions [146].

Multiple sclerosis is a neurological disease, usually assessed using MRI images to detect lesions in brain white matter [153]. CNN was used for classification of images and various attribution algorithms were used to generate heatmaps to visualize each voxel contribution to the class prediction [153]. Various methods were compared and DeepLIFT was determined to be preferable to LRP [153]. The 3D, T1-weighted multi-echo, gradient-echo images were obtained from an MRI scanner and it was concluded using relevance analysis that relevant voxels were around venous vessels [153]. 3D CNN were used for diagnosing the multiple sclerosis on MRI dataset and LRP was used for validating the model's decisions [142]. The study concluded that the framework and LRP provided transparency in the model's decision [142].

A Subtractive Spatial Lightweight CNN (SSLW-CNN) was proposed to develop explainability for medical analysis and to evaluate trust for brain tumor classification in MRI images [149]. The malignant tumor cells were divided into five classes [149]. The model used CAM for finding the relevant features and the final visualizations demonstrated how the system reached its decision on classifying tumors [149]. 3D CNN was used for tumor grading on BraTS dataset for expediting tumor assessment for the whole brain and a defined tumor region, and showed that the results with defined RoI achieved better accuracy [150]. The interpretability was evaluated with Grad-CAM and Guided Backpropagation for validating the trustworthiness [150]. A pre-trained ResNet50 model was used on IXI dataset of MRI slices for a multiclass classification into a glioblastoma, vestibular schwannoma, or no tumor [152]. Grad-CAM was used for visualisations and helped to identify the tumor location. The study concluded that by embedding model explainability in the early model training can save time and help clinicians in using a model's prediction for the clinical decisions [152].

A Modality Specific Feature Importance (MSFI) metric was proposed to encode explanation patterns and clinical images for modality specific feature localization [133]. Glioma grading was performed by classification into lower-grade and higher-grade gliomas on the BraTS 2020 multi-modal dataset [133]. MSFI was shown to perform better in comparison to the other chosen post-hoc heatmap algorithms for multi-modal imaging tasks [133].

The brain tumor segmentation was investigated for uncertainty and interpretability by training three different models on BraTS 2018 dataset [151]. The study was aimed at deriving the visualisations for the feature maps helpful for the medical diagnosis [151]. Network dissection and Grad-CAM were used for the activations and spatial attention to understand the model's localization and segmentation performance [151]. Comprehensive Attention CNN (CA-Net) was used for multi-class segmentation of fetal MRI [211]. The dataset comprised of 150 stacks of T2-weighted fetal MRI scans [211]. The evaluation of the proposed scheme showed significant performance improvement over U-Net, through the use of comprehensive attention guidance with spatial, channel, and scale attention modules [211]. CA-Net also provided better explainability by visualizing the attention weight maps, although it had around 15 times smaller model size [153].

Cardiac MRI (CMR) segmentation was conducted to infer the different cardiac conditions using a DNN, and Discovering and Testing with Concept Activation Vectors (D-TCAV) for interpretability of the significant features for the cardiac disease diagnosis [241]. The study used MICCAI 2017 dataset with the U-Net architecture [241].

A model fusion based approach was proposed by using MRI and ultrasound data. The choice of the MRI slice/image is important and the authors proposed selecting the best MRI slice for improved model performance [220]. MRI and CT images for hydrocephalus segmentation achieved significant improvement in thin-slice MRI and CT images [141]. The generalization and representation of the model were enhanced by reducing the model uncertainty [141]. Dice scores were used for evaluating the segmentation results [141]. Grad-CAM was used for visual explanations of the wrap-around model in low-field brain MRI images on two datasets showing an agreement between the radiologists and the models using Cohen's kappa values [154].

MRI is the most favored imaging modality for neurology and particularly useful for the neurodegenerative diseases. Some studies investigating breast cancer and cardiology have also reported the use of MRI. Again, Grad-CAM seems to be the most favored approach as evidenced by Table 4.

4.4. Ultrasound imaging

Ultrasound images have not been used as extensively in clinical AI analysis as the other imaging modalities [253]. It can be difficult to control the image quality due to manual operation, operator dependency and acoustic shadows [253].

Ultrasound and MRI images were used for prostate cancer classification and data fusion of several pre-trained deep learning approaches with various shallow machine learning approaches. LIME was used for explainability and the fusion approach was found to improve the model performance [220].

Ultrasound images were used for breast lesion classification on BreastMNIST dataset with CycleGAN activation maximization for visualizations of the model predictions [176]. The qualitative evaluation of the classifier was performed using accuracy, specificity, Matthews correlation coefficient, and AUC whereas the qualitative evaluation used online surveys. The method performed better compared to DeepTaylor, LRP, and DeepSHAP in terms of image quality and semantic meaning

[176]. A breast cancer ultrasound dataset was used to investigate the susceptibility to adversarial attacks [203]. Multi-Task Learning (MTL) was used for improving the classification accuracies [203]. The multidisciplinary DREAM challenge was launched to reduce the false positive rates in breast cancer diagnosis with winning solutions based on CNN [254]. Breast ultrasound and MRI can provide additional information for the mammographic diagnosis [254]. Multimodal Multiview ultrasound images were used for breast cancer risk assessment on 10,815 ultrasound images [204]. Grad-CAM was used for generating the heatmaps for guiding the human experts [204].

Fetal brain abnormalities can contribute to mental retardation and neurodevelopmental delay [155]. Fetal brain planes were classified as normal or abnormal using a CNN model on ultrasound images, with Grad-CAM for lesion localization and diagnosis [155]. Fetal head circumference estimation is a key biometric for determining fetal health and estimating gestational age. The estimation was performed using ultrasound images and Regression CNN [249]. The interpretability of the CNN classification was investigated by using various saliency map methods and the explanation methods were evaluated using perturbation analysis [249]. Supervised Object Detection with Normal data Only (SONO) architecture was proposed for the detection of abnormalities in fetal ultrasound videos [243]. The video data was converted to a barcode type timeline to aid the explainability which made it easier to identify the substructures [243].

Ultrasound has applications in breast cancer diagnosis, genitourinary and cardiology related medical problems. Although use of Grad-CAM and other XAI techniques have been reported in Table 4 but there is not a predominant single XAI technique used across the various imaging modalities.

4.5. Optical imaging

4.5.1. Dermatology

Automated skin lesion segmentation can help with an early diagnosis of skin disease [211]. Skin lesion segmentation with the proposed technique used ISIC 2018 which has 2594 images [211]. The results were compared with DenseASPP, RefineNet and DeepLabv3+, and showed comparable results [211]. Semantically significant pre-images were obtained by utilizing gradients from the loss estimators ISIC dataset for skin that had seven disease groups [216]. The proposed approach was shown to be effective for generating counterfactual explanations, which are an important tool to explain the model predictions [216].

A DNN was used as skin image classifier and LIME with the ABCD-rule for model explainability [212]. The ABCD-rule of dermoscopy is known to outperform other methods [212]. For explainability, the LIME model was fused with a previously introduced human medical algorithm by the authors [212].

An explanation framework termed ExAID was proposed for malignancy of skin lesions using Concept Activation Vectors (CAV) The framework used two classifiers, a concept-level classifier for dermatological concepts; and a disease level classifier for lesion diagnosis which were evaluated on PH2 and Derm7pt public datasets [218].

The use of metadata for improving the skin disease diagnosis was proposed with a multiplication-based data-fusion framework for skin image classification on smaller datasets [213]. Dermoscopic lesion images dataset ISIC' 2018 was used with six pre-trained CNN models for skin disease diagnosis by fusing metadata and dermoscopy images [213]. It was found that the age and location metadata improved the classification performance. The qualitative evaluation used the CAM method to identify the important image regions [213]. A similar concept was used by leveraging the taxonomic organization of skin lesions to develop a hierarchical neural network, and spatial and channel attention modules for visualizations for insights in dermatologists/decisions [217]. The dermoscopic lesion images from ISIC 2017 and ISIC 2018 datasets were used. It was shown that the hierarchical diagnosis performed better than the other models for classification, and identified the relevant image regions with attention maps [217].

A Deep CNN based model was proposed for simultaneous skin lesion segmentation and classification and found to perform better compared to the other models on Jaccard index and AUC on ISIC 2017 and PH2 datasets [214]. Another study used ISIC dataset for binary classification of skin lesions [215]. CNN feature maps activations were used to understand the CNN model activations in dermoscopic lesion images to increase the dermatologists' trust in the system [215].

4.5.2. Diabetic retinopathy

DR affects millions of people in the world and is one of the leading causes of blindness [187]. Glaucoma can be diagnosed with AI techniques applied to OCT images and fundus photographs [249]. A review of AI techniques for Glaucoma diagnosis and progression and the translation to clinical practice are covered in Ref. [249]. Four CNN models were used and visualizations by heatmaps determined the important features which were compared against the expert's judgement [187]. The datasets EyePACS and DIARETDB1 were used and although different CNN models had a consistent classification performance, the image feature disagreement between different architectures was 70% as the highest value [187].

A framework was proposed for multilabel classification of DR images which provided explainability by classifying each pixel [191]. The image classification was derived from a multivariate lesion segmentation and categorization of pathological lesions [191]. The results demonstrated a high classification accuracy and explainability [191]. The authors proposed and applied the ExplAIn algorithm to multiclass classification of DR fundus images [191]. The classification accuracy was evaluated at both the pixel and image level [191].

A Generative Adversarial Network (GAN) based visualization method, Patho-GAN was proposed for diabetic retinopathy detection and retinal image generation [192]. A methodology similar to Koch's Postulates for Evidence-Based Medicine (EBM) was used for interpretability [192]. These postulates provided the association between the disease and the pathogens [192]. The significant features for the prediction were identified by isolating the neuron activation patterns [192].

A diagnosis system was proposed using a CNN for segmentation and classification of glaucoma detection with fundus images for deployment to mobile devices [193]. The proposed system had three interpretability measures for the intermediate segmentation results, morphological features, and activation maps with Grad-CAM [193]. The system was considered useful for mass screening as it could be run without an Internet connection on a mobile device [193].

CNN was used to classify the laterality of the fundus images and Grad-CAM was used for identifying the significant image regions for the classification [194]. The uncertainty estimation and the visualizations were found helpful for the model explainability. The proposed CNN model achieved a better accuracy compared to VGG-16 and AlexNet, and was found to be comparable to clinicians' performance [194]. An uncertainty-aware system for severity grading of the retinal fundus images were proposed to aid the expert with the trust in the model's decision [190]. The explanations and uncertainty were provided for each prediction which was considered useful for a clinical setting [190].

The glaucoma classification and localization used CNN and Grad-CAM on fundus images [195,197]. Five pre-trained CNN models were trained and evaluated with the best model achieving an accuracy of 96% and a specificity of 100% for the optic disc [195]. The localization with Grad-CAM was found to vary between the different CNN architectures and the external datasets were found to provide lower accuracies compared to the authors' own dataset [195]. A human intervention step was included to improve the localization accuracy and a qualitative evaluation was performed to verify the classification and localization results [197]. Better integrated heatmaps were obtained compared to Grad-CAM with VGG-16 by the use of proposed dissimilarity

calculations and knowledge preservation [197].

Multi scale features were used to improve the performance of glaucoma diagnosis and showed better local region identification for clinical diagnosis on the challenging ORIGA fundus image datasets [196]. The proposed EAMNet achieved glaucoma diagnosis with 0.88 AUC and accurate optic disc segmentation (0.9 Adisc and 0.278 Edisc) [196].

Based on the progression of the disease, DR can be classified into one of four grades [179]. TCAV was used for predicting diabetic retinopathy on retinal fundus images on a scale of level 0 (no DR) to level 4 (proliferative) and consultation with medical experts showed that the technique can help experts to understand the model predictions and improve the model [88]. Integrated gradients explanations were used for improving the severity grading of DR on retinal fundus images which were also graded by ten ophthalmologists unassisted, grades only, and grades plus heatmaps, where grades and grades plus heatmaps were generated using a DNN [189]. The grading time was found to decrease across the automated grading, however, in general the grades plus heatmaps were found to be as effective as only grades [189]. A framework was proposed for assessing retinal fundus images quality with quantitative scores and qualitative visualization on SDRSP and IDRiD dataset [188]. The proposed system was found to be useful for the DR screening program [188].

A faster R–CNN ResNet was used for feature localization in fundus images [185]. The proposed system generated a report for each processed image containing the classification prediction, lesion detection, and text annotations for the expert [185]. The system used three public datasets, Messidor-2, IDRiD, and E-Ophtha totalling to 2314 images [185].

Optical Coherence Tomography (OCT) in retinal images used VGG-16 and SqueezeNet [198]. The DNN models were trained on the OCT-2017 dataset [198]. To identify the significant areas of the image for the prediction, an occlusion mask was moved across the image recording the classification probabilities with the most contributing areas that resulted in a high drop [198]. A DNN model was proposed for classification of diabetes-related retinal diseases on Optical Coherence Tomography (OCT) volumes and the model was found to perform better on accuracy, sensitivity, and specificity compared to other similar models, without requiring manual annotations [199].

An ensemble of CNN was used for an automated diagnosis of DR severity and Grad-CAM was used for explainability to indicate the image regions contributing to the DR grade prediction by the CNN [186]. Three datasets (APTOS, IDRiD, MESSIDOR) were used for the model evaluation [186]. Learning rate hyperparameter was tuned automatically by using Cyclical Learning Rates (CLR) between a lower and upper bound [186].

Ophthalmology is one of the most widely used application of AI and there are efforts towards explainability in this area but are still at very basic level. The explainable algorithms for ophthalmology especially for DR need to include more insights from the actual domain and these should be able to incorporate conventional grading systems like Early Treatment Diabetic Retinopathy Study (ETDRS) and thickness mapping etc. for more detailed insights. The ophthalmologists are interested in finding out the parameters responsible for a specific finding, rather than just knowing the final diagnosis. In fundus images, the actual variations in landmarks i.e. blood vessels, optic disc and fovea/macula region along with appearances of any new lesion are important. For OCT images, it is important to look into layer variants and lesions appearing in those layers. The explainability needs to focus on these features in order to have a detailed analysis and reasoning.

4.5.3. Cancer

A Cumulative Fuzzy Class Membership Criterion (CFCMC) technique was proposed that provided a human friendly explanation of the prediction decision. It provided threefold explanations, semantic explanation of misclassification possibility, training sample for the prediction, and training samples from conflicting classes [232]. The technique was used on a public dataset of Hematoxylin and Eosin (H&E) tissue slides of histopathological images for colorectal cancer [232].

Gastric cancer ranks as the fourth common type of cancer worldwide [231]. The LIME, SHAP, and CIU were used for explainability of the CNN predictions to Video Capsule Endoscopy (VCE) in-vivo gastric images [230]. The VCE procedure was used to determine red lesion segments for detecting bleeding and a single examination may have 10 h of video [230]. The explanations of the three XAI methods were evaluated by a human expert and the study was aimed at developing a human decision support system [230]. CNN based architecture with ResNet-34 as a base network was used for ulcer recognition and classification in capsule endoscopy dataset [231]. The system performed better on F1, F2 and ROC-AUC measures compared to the other chosen CNN networks. The proposed architecture performed better compared to the other CNN architectures for small ulcer lesions. The visualisations for the important image parts used CAM confirming the ulcer localization [231].

LIME was used for explainability of histological images for lymph node metastases classification of CNN on the Patch Camelyon (P-CAM) dataset [255]. The LIME explanations matched in general with the clinician annotations, and the CNN models seemed to focus on specific features of the images [255]. It seemed that CNN model had learnt to associate the tumors to the image structures in the centre as was noted in the generated heatmaps [255].

A multi-scale feature learning explainable method was proposed for CNN model decisions [221]. The method was used for cervical cancer diagnosis using the decision interpretation from the perspective of visualization and statistics, in pathological images of a uterine cervix provided by Nagasaki University Hospital, Japan [221]. Multi-scale feature dictionaries were introduced to capture the discriminative features over different scales to aid better diagnosis [221]. The results identified the important and contributing image learned features for the diagnosis [221].

Four CNN models were used for Barrett's esophagus along with five visualization techniques of Saliency, Guided Backpropagation, Integrated Gradients, input X gradients, and DeepLIFT for determining model predictions with experts' annotations [229]. The saliency was found to provide the best matches [229]. The focus was on the early detection and on clarifying the image regions important for the classification [229].

4.6. Nuclear imaging

Parkinson disease affects the nervous system and is a progressive degenerative neurological condition [137]. Single-photon Emission Computed Tomography (SPECT) DaTSCAN imaging is used for early intervention and treatment of the disease [137]. LIME was proposed for an early detection of the disease on DaTSCAN imaging [137]. The 41st slice of 3D Digital Imaging and Communications in Medicine (DICOM) volume were used as JPEG images and intensity normalization was performed to mitigate the effect of different acquisition parameters [137]. The proposed model was found to be useful for early diagnosis of Parkinson disease based on the accuracy and interpretability [137].

Four DNN architectures with six interpretation methods used SPECT images for the detection of Parkinson's disease. SHAP was found to generate better heatmaps out of the six interpretation methods for the four selected DNN models. For evaluating the interpretation, a commonly used technique, DICE coefficient was used for the Parkinson's Progression Markers Initiative (PPMI) database [138]. The Guided backpropagation performed best amongst the six selected interpretation methods [138].

LRP was used with CNN-based classification for DAT-SPECT images of Parkinson patients with unknown syndromes [139]. Some relevance maps were found to be inconsistent with the true class labels, however, the accuracy, sensitivity and specificity of CNN were 95.8%, 92.8% and 98.7% respectively [139]. It was recommended that as LRP provides explainability for CNN predictions on DAT-SPECT, it could be used for

clinical practice [139].

A deep learning based cognitive signature was developed for brain PET for Alzheimer's and Parkinson's disease [140]. The CNN model produced 128 features for each sample which were visualized using t-SNE for similarity. It was shown that using the imaging biomarkers provides an objective assessment of the cognitive impairment [140].

4.7. clinical/observational data

ECG is often used for cardiovascular imaging. An ECG interpretation solution for heartbeat classification was proposed using several model-agnostic methods, PSI, LIME and SHAP [244]. The work focused on sample-based explanations and was evaluated using 1-D Jaccard's index to measure the similarity of the classifier and the relevant subsequence from the XAI method [244]. The binary and multi-class models were based on five/four CNN blocks respectively [244]. CNN, RNN and XGBoost frameworks were used for diagnosing the risk of Acute Myocardial Infarction (AMI) using the ECG-ViEW II dataset [248]. The significant features were identified with Shapley, demonstrating a high relevance of age and sex [248]. CNN models were used with ECG data to identify cardiac features to predict systemic phenotypes affecting cardiovascular risk [245]. Grad-CAM was used for visualizations and identified RoI for predicting the systemic phenotypes which are otherwise considered difficult for the human experts [245]. Transthoracic Echocardiography images were used for an automatic Carpentier's functional classification of mitral valve diseases with eight deep CNN models [246]. In visualization with Grad-CAM, ResNeXt50 provided the best explainable results highlighting the RoI [246].

CXR images for Congestive Heart Failure (CHF) were used to train a neural network on a labelled dataset, and a generative model trained on unlabelled dataset to estimate B-type natriuretic peptide BNP as a marker for CHF [237]. Generative Visual Rationale (GVR) based on GAN were used for visualization and compared against other visualization techniques such as LIME, concluding that GVR can identify the confounding images [237]. An automated system for stenosis localization and severity classification into three classes in Coronary X-ray Angiography (CAG) utilized CNN and RNN models, with CAM and Grad-CAM for visualization [238]. The results showed a high sensitivity for stenosis positioning and a high accuracy for stenosis classification [238].

Whole slide images were used to automate the scoring of the four histological features in Non-alcoholic Fatty Liver Disease (NAFLD) at two different scales with four CNN models [234]. CAM was used for the visualization and the pathologist's Kleiner score were found to be consistent with the DNN model's scores [234].

A CNN model was used for automatic classification of malignant lung cells with cytological images and the results were found comparable to a cystopathologist [184]. Grad-CAM was used for the visualization of the important image areas for the classification and found to identify the malignant cells [184].

Multivariate time series data was analyzed with CNN for explainability with visualizations of relevant features over time [250]. The visualizations used heatmaps of the same shape as the time series data [250]. The in-hospital mortality was predicted based on the ICU records [250]. The salient and discriminant points on the heatmap were determined using *t*-test [250].

A CNN, xECGNet was proposed to fine tune the attention maps by adding regularization loss to the objective function [247]. The model evaluation showed better classification performance compared to the other state of the art methods and provided an improved visualization of Cardiac Arrhythmia (CA) types [247]. The model used a modified version of CAM for generating the attention maps and was trained and tested using the CPSC 2018 dataset [247].

XAI was used for understanding the classification of cardiotocography for fetal health status [256]. Light Gradient Boosting Machine algorithm was used for fetus health classification [256]. The dataset from Kaggle comprised of three classes and 21 features [256]. SHAP was used for the data explainability [256].

Early detection of cognitive decline was investigated in a smart home equipped with various sensors to monitor the occupants' behavior [257]. The system was termed as Health XAI and the collected information was shown on a clinician dashboard showing the anomalies, predictions and explanations [257]. For detecting anomalous behavior of a person, the anomaly feature vectors were created using the participant profile in the cloud-based system [257]. The XAI model also generated a natural language description of the explanations [257].

Immunohistochemically-stained archival slides were used for Alzheimer's disease classification with CNN [156]. Grad-CAM was used to identify the important visual features for the model's predictions and feature occlusion was performed to understand the image features contributing to plaque predictions, which were found to agree with the human interpretable neuropathology [156].

4.8. Cancer detection

A review of XAI approaches for cancer detection with MRI for DNNs concluded that explainability can be improved with counter examples, and intrinsically explainable models [258]. Prostate cancer is a common form of cancer in men and an early detection can significantly increase survival rates [220]. Multi-modal fusion to combine pre-trained DL models for cancer detection on ultrasound and MRI images was used to improve the classification accuracy [220]. The open source prostate MRI and ultrasound image data were used with LIME for explainability of the important features for benign and malignant classifications [220].

Liver cancer diagnosis used MRI images and CNN for lesion classification [228]. The feature scoring used Influence functions corresponding to the image features and confirmed that the most important imaging features were identified [228].

A visual Case Based Reasoning (CBR) used a database of previous known cases to determine the solution for a new query. CBR was proposed for visual reasoning on three public breast cancer datasets obtaining a classification accuracy similar to KNN but provided better explainability [205]. The explainability was provided using a visual interface for similarities between the query and similar cases, with scatter plots for quantitative, and rainbow boxes for qualitative similarities [205].

LRP was extended for application to Graph-CNN and tested for explainability for a breast cancer dataset [210]. The proposed method GLRP was applied to gene expression dataset [210]. The Graph CNN learnt features on weighted graphs with nodes and edges, unlike CNN that use a grid like structure [210]. GLRP provided explanations for Graph-CNN that were consistent with the clinical knowledge and identified drivers of tumor progression [210].

Histopathology images were used for breast cancer classification on BACH microscopy dataset using a proposed RoI guided soft attention network [207]. The results of the visualizations were aligned with that of expert pathologists without the post processing required by Grad-CAM [207]. The histopathology images were used for multi-scale localization of relevant RoI with four CNN trained on samples from pathologists' screenings, followed by a CNN for classification into four diagnostic classes [209]. The proposed system was found to be better compared to the methods using hand crafted features. The visualizations used occlusion and deconvolution [209]. Histopathology images were used for breast cancer grading using the proposed Regression Concept Vectors (as an extension to TCAV) showing the nuclei contrast as a relevant concept for tumor tissue detection [208]. The Camelyon16 and Camelyon17 datasets were used for the classification of breast tissue patches with a ResNet101 pre-trained model [208]. A CNN model was trained on histopathological images and used as an AI assistant to aid pathologists with the classification of the two liver cancer types of hepatocellular carcinoma and cholangiocarcinoma [233]. The study found an improvement to the overall prediction accuracy with the model assistance although the implemented model alone was not as accurate as

the pathologists [233].

A multi-input CNN architecture was proposed for predicting breast tumor response to chemotherapy on DCE-MRI images [200]. The visualizations with Grad-CAM were used to localize the tumor RoI and showed the important features to be in the peripheral regions [200]. 3D ResNet was used to identify tumor classification with Cosine Margin Sigmoid Loss (CMSL) and localization of cancer lesions with COrelation Attention Map (COAM) on DCE-MRI dataset [201]. Volumetric breast density estimation was assessed with 3-dimensional regression CNN on MRI images with explainability using SHAP [202]. The method's main advantage was concluded as not requiring voxel-level 3D segmentations [202].

An interpretable framework ICADx used a GAN to learn the relationship between the standard breast image classification system (BI-RADS) and malignancy [206]. GAN comprised of an interpretable diagnosis network and a synthetic lesion network [206]. The synthetic lesion generative network used adversarial learning to aid the interpretable diagnosis network designed using a CNN [206]. The proposed system was validated on a public dataset DDSM [206]. The model performance was evaluated showing that the proposed system interpreted prediction in terms of BI-RADS [206].

The timely detection and extent of the spread of cancerous cells is of utmost importance for deciding a treatment plan. A study investigating lung cancer used the DNN model to predict the future susceptibility of cancerous tissue development in lung nodules [179]. Such studies can be used to help preempt the disease onset and with a timely intervention can save precious lives.

5. Challenges and future directions

The XAI techniques can help in understanding the predictions of a deep learning model yet there are challenges to be overcome before these techniques become mainstream for healthcare applications. In this Section, we provide a discussion of the application and challenges of the XAI methods for biomedical applications and also provide the developments currently underway that will transform the XAI applications in biomedical imaging.

5.1. XAI applications in medical imaging

The medical imaging covers a diverse set of image acquisition in temporal and spatial domains to acquire the most relevant information aiding the diagnosis of the underlying medical condition. The image modalities and the perceptible image details these contain vary. However, the most common and preferred XAI method across the various medical imaging modalities seems to be Grad-CAM (Table 4), followed by LRP, LIME and SHAP and other methods. The increased use of XAI techniques is seen in neurology, respiratory and ophthalmic studies. In neurology, the most prevalent image modality is MRI. CXR images are most favored for respiratory studies due to the ease with which these can be obtained, with Grad-CAM as the most prevalent XAI methods. In ophthalmic studies, fundus images seem to be the dominant modality.

A lot of techniques have been proposed and work underway to make it possible for the DNN and XAI techniques to be widely adopted in the clinical practice. DNNs enable automated RoI selection without the time-consuming manual annotations [225,226]. It is important to associate the model's prediction uncertainty to aid the clinician for more attention to these [151,190,194].

The incorporation of the supplementary information in the model training can help to improve the model's prediction and explainability [156,217,226]. Similarly, the use of frameworks combining the automated segmentation and classification could for example, can use the segmentation results to guide the classification process [214]. XAI techniques using multimodal DNN models can help with the hidden patterns related to each modality in segmentation and classification tasks. The use of multiple scales can increase the localization and explainability [167,196,200,221]. Using the XAI techniques to assist the clinician can help to produce better diagnosis than the model or clinician alone [204,212,233,234]. Due to the paucity of the medical data, some research is focused on addressing the associated challenges [140,227]. A DNN model segmentation and classification was run as a mobile application for a wider accessibility for mass screening of glaucoma [193].

The explainability techniques are mostly based on visualizations that provide a qualitative indication by highlighting the significance of the image areas contributing to a model's decision. This can provide a degree of confidence in the model but without a standard way of evaluating the different models and their outcomes cannot be trusted for medical imaging. The creation of quantitative metrics for model comparison for explainability will give rise to the successful application of relevant explainability techniques to different image modalities. Such metrics can also meet the needs for the regulations and standards in assessing and approving the XAI techniques for the clinical practice.

There is a need to develop DNN models that are inherently interpretable and yet are accurate. Such XAI systems need to be incorporated into the clinical process workflows with intuitive interfaces. Model agnostic methods hold a promise of further development and utilization for explainability.

Similar to the progress in AutoML to create a reproducible and optimal model automatically based on the given constraints, there is a need to develop similar framework for automated XAI models where in response to the basic requirements of the user, an appropriate XAI technique can be automatically chosen and applied. Explanatory Interactive Learning (XIL) which adds the human in the model training loop [259] can also help to increase the adoption of the XAI techniques.

5.2. Understanding and improving explainability

The term 'explanation' as used in XAI is not very well defined [60,260,261]. A universally agreed definition of explanations is lacking and there has been some efforts in this direction by developing mathematical methods [25]. There is no agreement on what is meant by 'explanation' [262]. Efforts are underway by some research institutions and regulatory authorities to prescribe clear definitions, scope and applicability of explainable AI and other similar interacting terms. This differs for the different perspectives of a patient, clinician and engineer and the current definitions lack consistency with explainable AI representing a false hope to provide a patient level decision support [19]. The model's highlighted features might also differ in relevance to human identified features and decisions [19]. The adoption of AI in clinical practice would thus take more effort in making the XAI techniques to be more comprehensible.

The development of DNN and then having attempts to identify their internal workings and decision-making process has resulted in development of many useful techniques. However, it would be useful if the explainability is built into the model. Most of the studies relating to cancer detection use a post-hoc explainability technique whereas a development of intrinsic explanation techniques would be advantageous for adoption and explainability [258]. A multi-task capsule network, X-Caps was proposed to improve the explainability by using the high-level visual attributes of capsule networks to provide malignancy scores used by experts [15]. Automated heatmap analysis were integrated within the MLOps (Machine Learning Operations) pipeline for vision AI as a building block used to detect biased models by activations with invalid pixels in test images [109].

White-box modelling methods like XNN, scalable Bayesian rule lists, and monotonic GBM, provide little accuracy penalty for interpretability [264]. Developing an interpretable model from the very beginning will make explanation and fairness evaluation easier [264]. Future approaches may see a combination of intrinsic, rule-extraction and attribution methods to provide answers in a human interpretable language [10].

5.3. Data-centric XAI

The quality and quantity of the training data plays a major role in the trained model's performance and resultantly to the explainability that can be obtained from such a model. Large and high-quality data is required by DNN for better results without which the model may overfit, that is, becomes closely aligned to the limited set of datapoints [265]. This can be a challenge for imaging datasets where obtaining the diseased image samples is challenging due to limited number of samples compared to normal cases [6]. The data quality can be affected by noise and incorrect labelling [10,254]. The lack of data quality can also hinder the results of explanation by examples [120]. Denoising methods can improve the robustness of the DNN against small perturbations and quality degradations by using denoising methods to enhance the edges and outlines [266]. The image pre-processing using contrast-to-noise ratio, removing irrelevant regions etc. improved the DNN performance [267]. The quality of images improves the performance of DNN algorithms in correctly predicting a disease and therefore can reduce the explainability gap. The reduction of noise in the data can make it easier for the model to explain or interpret [268].

Deep learning models require sufficient training data to yield reproducible and useful results [37]. Without sufficient data, the AI models can suffer from overfitting and the data from different hospitals differ due to the variety of data collection policies and protocols [265]. Some studies reporting DNN medical imaging performance exceeding the radiologists were often using small datasets and can be considered a significant limitation [6,254]. Data collection for breast cancer covering all image variations is difficult considering other factors such as age, race, ethnicity etc. [6]. Biomedical imaging datasets sometime suffer from noise, whereas CT images are commonly affected by noise [265]. The predictions of DNN with the same architecture may be different due to data quality and the model initialization (random) for MRI images [269].

Some of the research has focussed on meeting the challenges of small datasets. Various techniques can be employed to address the lack of data, such as, data augmentation [34,37]. Additional realistic data could be generated through GANs using the latent data distributions [13,34,37]. LRP pruning was shown to provide good results for transfer learning with small datasets [56]. Decontextualized Hierarchical Representation Learning (DHRL) framework was proposed for small datasets by capturing of signals across the different spatial scales [270]. The framework was applied to study natural patterns with unsupervised deep learning methods [270]. Few-shot learning can work with very few training samples [13].

The labelled datasets used to train DNNs are often from multiple sources or maintained by different organisations owning the data. The funders and governments are emphasizing an open sharing of data for AI model training but most data are not representative, interoperable, and sufficiently diverse [39].

The imaging data captured in clinical settings with the exception of radiology may only be housed in the local department's systems [13]. New types of information and novel insights into the data are possible by combining the data from diverse sources [272]. Federated learning enhances the use of data privately stored in a local repository to develop a global DNN model by using local data to train a local model, and sharing model parameters with a server. This can facilitate a global repository of benchmark datasets for biomedical imaging. This would provide opportunities for measuring the accuracy and explainability of DNN against the same dataset and mitigate the different biases in the model and data.

5.4. Model centric XAI

The use of sufficient, representative, and diverse type of data for DNN models can help to reduce bias and enhance generalizability. Most ML studies on medical images use trained models that are often based on a single unit or hospital which brings in a selective bias and sometimes inclusion of low-quality scans [273]. Counterfactual algorithms have the potential to uncover biases in both the models and data [106]. The use of XAI techniques such as visualizations are helpful to identify the model bias and an experienced clinicians can identify any unsupported or bad DNN decisions.

Current AI success is attributable to models that provide better performance but are opaque or black box [45]. One of the stated advantages of DNN is that of an automated learning of the image features and overcoming the problems associated with the hand-crafted features. End to end automated XAI applications, such as, combining segmentation and classification, will help XAI adoption in clinical practice. The model architecture has many hyperparameters that need to be tried and re-tried for designing an architecture with an optimum performance. AutoML is a step towards the direction of automatically creating an optimum architecture and making it possible for the architectures to be reproducible and comparable. Such automated workflows are expected to make the process of incorporating XAI easier into clinical decisions by making it possible to select the best visualizations in response to a few simple questions. The automation will not only make these XAI techniques widely available but will enable medical practitioners to make better and informed decisions by leveraging the insights through XAI.

This can in future help towards models' certifications and standardizations such as by Food and Drug Administration (FDA) [274]. The AI Fairness 360 toolkit is an open-source library that can help detect and mitigate bias in machine learning models and is available in Python and R. It is aimed at translating the algorithmic model research to the actual practice in domains such as healthcare [275]. Similarly, Google's What-If Tool can be used for optimizing the model's AI fairness [276].

The adversarial inputs comprising a modified image can drive a model to misclassify and is termed as an adversarial attack. The DNN model can also be susceptible to an adversarial attack. It is important to devise techniques that can identify or prevent an adversarial attack [277]. Undetectable adversarial attacks can be devised to change the importance maps dramatically and interpretations of breast ultrasound images were shown to be susceptible to adversarial attacks similar to other image modalities [203]. The study of adversarial attacks can also help to develop better interpretation models that are more robust against such attacks [278]. The CNN fixation method was applied to detect an adversarial attack on sets of images with a clean and perturbed image, and even despite the labels being changed, the proposed method could still localize the regions correctly in both images [69]. A framework for trustworthy and XAI combined selected features of trusted oracles, blockchain and smart contracts to reduce adversarial attacks and bias [11]. A counterfactual approach, Global Explanations for Bias Identification (GEBI) was proposed for identification of bias in the data for image classification although it can be applied to other problems [279].

A case for scientific XAI for medical AI, proposed a top-down rather than the current bottom-up models [38]. It is argued that the proposed top-down approach may be technologically difficult but focuses on scientific explanations, whereas the current strategy leads to confounding classifications [38]. Complementary optimization procedures used to generate adversarial examples are similar to the proposed method that learns the imperceptible pattern of noise. Thus, the proposed method can be used for generating adversarial examples by creating abnormal masks. The learned masks can minimize adversarial label once applied back to the adversarial images, and can select the original predicted label from a clean image as top predicted class [280]. Thus, the proposed method can recover original predicted labels without any changes to the model or training [280].

5.5. Human-in-the-loop

The trained DNN model can provide predictions comparable to human expert in a fraction of the time, however, this is not sufficient to

fully delegate the medical diagnosis to an algorithm. Instead, DNN and XAI techniques can support the clinicians to make better and informed decisions. The human in the loop approach ideally allows for an interaction with the model to make the inferred changes to obtain better model predictions and with sufficient explainability.

Explainable systems that involve the human user in the decision process of the DNN are useful to determine answers to the 'why' and 'how' questions regarding the model's predictions [93]. A human-in-the-loop system was proposed for a video action recognition to make it easier to understand the learned features by non-expert human users [93]. The decisions affecting patients' lives cannot be entrusted fully to an algorithm, rather visualization approaches that can relate the original data and the visualizations are more important to support clinical decision making [261]. The clinical decision support system must be trustworthy, and should reinforce the human decision [27]. The use of XAI for a binary decision regarding presence or absence of a disease is of great help, allowing more complex cases which are difficult to classify to be determined in a clinical setting by a human expert [281]. The radiomics CT image signatures and DNN methods were used to gain insight into CT image features that are important for COVID-19 prediction [282]. The features from DL and radiologists were compared to gain an understanding of the ML algorithm interpretability for improving human diagnostic performance [282].

A systematic understanding of XAI and humans as a system is lacking and remains a challenge [8]. Approaches like Bayesian Teaching can help clinicians to use the explanations to understand the XAI diagnosis [8]. DNN systems have sensitivity similar to radiologists but often times these DNN can detect different lesion characteristics [6]. The radiologist's accuracy in characterising lesions can be much improved using a DNN [6]. Using Explainable Neural Network (XNN) along with a Sparse Reconstruction Autoencoder (SRAE) it was shown that for classification tasks which are difficult for humans, the understanding of a DNN can be improved by using an XNN for explainability [95].

The idea of an AI assistant was proposed to support paediatric radiologists by freeing them for other tasks and possibly bringing a balance between XAI and human intelligence that can meet legal and regulatory requirements [5]. However, human expertise cannot be dispensed and only simple, and time-consuming tasks should be directed to AI. The ultimate responsibility would still be with the human radiologists [5]. A human-machine collaboration approach was introduced for the image segmentation, subsequently reducing the work done by the clinicians, to speed-up the process, and have better segmentation masks [159]. The clinician provides masks for a subset of images, which are then used by the segmentation DNN for improvement by generating competing masks. Following clinician verification and further runs, the segmentation masks on larger datasets can generate enhanced quality [159]. Interactive ML was proposed for the medical domain with "human--in-the-algorithmic-loop" to solve problems that cannot be solved by automation or human alone [272].

The ultimate beneficiary of the model explainability is a human user. Explanatory Interactive Learning (XIL) added interaction by including the scientist in the model training loop [259]. The focus shifted to correcting the underlying reasons for a wrong model decision which can aid in understanding the confounding factors in image classification [259]. An eXplainable Neural-symbolic learning (X-NeSyL) methodology combined both symbolic and deep representations to assess the alignment between human and machine explanations using an explainability metric [20]. The method was evaluated by fine tuning Faster R–CNN on MonuMAI, and PASCAL-Part dataset [20]. The learned representation fusion helped the model to learn with a human expert [20].

5.6. Incorporating stakeholders in XAI

Many DNN models meet the explainability requirements of the system developers and specialists but not of the end users of those systems [59]. It is critical to focus on the needs of users while developing the model explainability [59]. DNNs have the potential to revolutionize biomedical imaging and pave the way for a personalized diagnostic [36]. The use of DNNs for radiomics has increased but there are prevailing challenges around the explainability of the classification and prediction decisions of these models [36]. The clinicians are more likely to use a DNN for diagnosis if the model provides insights into its decision and can be understood to be not being driven by any artifacts [283]. It was argued that XAI systems are being developed by the DNN scientists for themselves whereas these should be focused on the end-user [260, 284]. The models should be based on the explanations based on lessons from social sciences and psychology and developed jointly [284].

Similar to the role of an autonomous car driver, the final decision for image-based disease diagnosis would rest with the specialist whereas the XAI model insights should be used only to aid the decision. For increasing accountability, the use of machine learning models should be limited by regulators to only the ones that have been empirically validated to ensure that these are not used as mere tools for arbitrary interference by the stakeholders [57]. Thus, more efforts should be placed on the models and datasets that are aligned for the intended purpose [57]. Certification of DNN models by agencies external to the developer or user could help to eliminate risk of unauthorized changes to how the system works and can prove beneficial in case of litigation to prove the correctness of the system [49]. The challenge against an automated DNN decision could thus not require specialized technical knowledge of the AI system [49].

In order to use AI successfully requires a multi-disciplinary effort encompassing guarantees of ethical, legal and social aspects of AI and increasing awareness [5]. It is by breaking down these silos within which each domain expert operates to develop systems that address the needs for successful deployment in clinical practise. A multidisciplinary approach can develop understanding of the AI application in the society and markets [285]. It is important to collaborate with the clinicians during the design and evaluation of DNNs and to take the viewpoint of the clinicians and patients as to what is considered explainable by them [258].

It is important for the clinical practitioners and computer scientists to work together for better explainable systems. Some challenges in building multi-disciplinary collaborating teams comprising of engineers and radiologists are covered in Ref. [286]. The skillsets of each are relevant for the other in order to better understand the viewpoints and limitations [286]. The radiologists have the skills to identify and label abnormalities and disease, whereas software engineers have expertise in software packages, DNN architectures and data processing, and there has to be a mutual sharing and transfer of skills to develop a successful system [286]. It is important that the users of the XAI techniques are made aware of its applicability, limitations, and capabilities for an effective adoption.

5.7. Hybrid/fusion imaging for XAI

Multi-modality comprises of combining the different imaging modalities such as MRI, CT, and PET. In medical imaging context, it could not only include the different image modalities for diagnosis but increasingly includes patient attributes or time-series data [13]. Cardiac MRI and CT may be combined with the prior ECG [13]. The combination of the various data types, data mapping, and data fusion is the central problem in data science, and bringing multi-modal data together can provide new insights [272]. The information fusion and multimodal system can have a significant performance over the systems working with a single type of data [36,100].

A cloud-based healthcare framework for COVID-19 like pandemics combined imaging data such as CXR, CT, PET with blood pressure, cough sound, and body temperature, for training a DNN model [287]. Hybrid imaging provides additional information from the different image modalities that can significantly increase the investigation and identification of a disease [35]. Two significant issues faced for

multimodal imaging are the class imbalances and lack of sufficient labelled data [35]. The black box and opaque nature of deep neural networks precludes their usefulness despite excellent performance of multimodal methods [288].

The correlation between MRI and gene expression data for Alzheimer is known to exist and was used to implement a multimodal diagnostic model for diagnosis [143]. Multimodality technique was observed to provide better specificity compared to each of the modalities alone [143]. A GAN could also be trained to generate PET from CT and convert between MRI and CT [13]. The clinical acceptance of AI techniques would be aided by multi-modality and information fusion of both imaging and non-imaging data [34,170].

Multimodality has its own challenges of data representation, translation, alignment, and fusion [288]. It was shown that by aggregating the different visual explanations can help to understand the models thereby increasing the trust in the system [74].

5.8. Mathematical and graph based XAI approaches

The mathematical approaches can improve the interpretability of models, however, may not be suitable as concrete examples for the end-user [63]. A graph structure can be used to represent relationships by representing entities as nodes and the relations as edges or links and was proposed for data in imaging and NLP applications etc. using graphs and a Graph Neural Network (GNN) [99]. The model and the learning algorithm, output functions, comparison with Random Walks and Recursive Neural Networks, and complexity were derived mathematically [99]. It was shown through the model evaluation that it can be useful for many important applications [99].

Graph Neural Networks can learn from graph structures and can represent information capturing different types of relationships rather than just an ordered grid structure [100]. This could be useful to study complex diseases such as cancer where different modalities have an interplay towards the diagnosis and management of the disease [100]. An approach, XGNN, was proposed to understand the GNN at the model level [289]. The interpretation of DNN on graph data has not been much explored but the model level interpretations can explain the sub-graph patterns resulting in a prediction [289]. The proposed model was evaluated using the synthetic and real-world datasets [289]. For classification, the model first learnt features using multiple GCN layers, averaging features to obtain graph level embeddings, and using fully-connected layers for the classification [289].

Motivated by the need to develop hybrid intelligent systems, knowledge graphs can represent domain knowledge with XAI to develop trustworthy explanations [117]. The enhancement of knowledge representation at scale in knowledge graphs can aid the XAI requirements but requires addressing the misalignment between different knowledge graph resources [117].

A formal framework was proposed for learning explanations as meta-predictors in a model-agnostic method based on image perturbations to determine the part of image affecting the classifier decision. The model is somewhat similar to LIME as both have functions' output with reference to the neighborhood inputs around an image region generated by image perturbations [280]. The advantage of meta-predictor is that the ML algorithms can discover explanations automatically and that the prediction accuracy corresponds to meta-predictor's faithfulness [280].

5.9. Explainability to improve models

Explainability can also be used for model improvement. This can be seen as an ongoing cycle of the model explainability driving better understanding and development of improved models, which in turn results in developing better model explainability techniques. Model pruning using LRP was shown to be helpful to determine which filters and weights in a model were important [56]. It was shown with the proposed deconvnet that the model visualizations can be used to determine model problems and can help improve the models [87]. Deconvolution, Grad-CAM, and Guided Grad-CAM were used to display the network structure of CNN layers to develop a hyperparameter optimization strategy to improve the CNN model performance [62]. Explainability was used to improve the model's classification accuracy. The evidence of the model's prediction was associated with saliency map of the model's representation. Guided Zoom was proposed to improve the model classification accuracy by considering the evidence used in each of the k-predictions [290]. The results on four datasets demonstrated that the Guided Zoom improved the model's classification accuracy [290].

Influence functions have basis in statistics and were demonstrated to be useful for generation of visually similar training set attacks, determining dataset errors, understanding, and debugging the model behavior [104]. This understanding of the model helps to improve the model [104]. The influence functions require complex and expensive computations which was offset by approximating the influence functions by second-order optimization techniques [104].

CNN models can have a large number of learned weights or parameters. In order to utilize CNN for devices with limited storage and processing, such as edge computing, mobile agents, and autonomous agent applications, there is a need to reduce the size of the model [56]. The application of such optimization techniques can also help in the understanding and improvement of the models.

A framework termed Network Dissection was proposed to quantify the interpretability of CNN latent representations [291]. It was shown that interpretability of the representation learnt by the hidden units can be significantly affected by the different training techniques, thereby helping to understand the CNN model characteristics [291].

6. Conclusion

This paper provides a survey of the emerging field of Explainable Artificial Intelligence (XAI) techniques and their application in the understandability of the image classification and segmentation tasks for biomedical imaging. Explainable AI coupled with deep neural network models can help with both the detection and diagnosis of disease, by providing additional insights to the clinician. There is an urgent need to develop algorithms to add reasoning and explan ability by considering all the clinical parameters and the changes which led to a particular finding. By leveraging the decision-making process of a model, opportunities also exist to understand, improve, and develop better AI algorithms.

We see a lack of algorithm standardization for biomedical imaging with diverse and independent efforts for improvements. A centralized regulatory authority can harmonize the algorithmic development efforts to benchmark and ensure a consistent algorithm improvement. The proprietary algorithms should be made public or at least be subject to an approval process before being deployed. Similar to improving the deep learning model itself, a data centric approach can help to ensure better quality data in sufficient quantity for model training. The regulatory authorities can seek to develop benchmark biomedical data against which the different explainability techniques can be tested and validated. Federated learning is also promising for biomedical applications with different hospitals contributing their data to train and improve a deep learning model without violating data privacy.

Currently, the image-based diagnosis and prognosis in the different disciplines of biomedical imaging such as radiology and ophthalmology require expert opinion of practitioners who are already overburdened. The current situation can be improved by using deep learning techniques with automated localization and explainability for accelerating the decision-making process of the medical experts. We consider a human-in-the-loop as a pre-requisite for any XAI application based clinical decision. The XAI techniques should be aimed at supporting the radiologists and other medical practitioners for better use of scarce resources and to enhance service provision.

Increasingly, there are health and legal regulations that mandate

approval and compliance for the deep learning techniques which can be helped using XAI techniques. The overcoming of challenges to clinical adoption of XAI will foster growth of the field, and result in better health and safety measures for the patients through timely interventions by building trust in the model's predictions.

Funding

This work was partly supported by the Glasgow Caledonian University [number 10386 M7311-GCRF Cloud Based AI Grader] Global Challenges Research Fund (GCRF).

Authors' contributions

Sajid Nazir: Conceptualization, Writing of first draft, Data Resources, Methodology, Model Development and Validation. Diane M Dickson: Methodology, Review and editing of manuscript. Muhammad Usman Akram: Conceptualization, Methodology, Review and editing of manuscript.

Ethical approval

This article does not contain any data, or other information from studies or experimentation, with the involvement of human or animal subjects.

Conflicts of interest

None declared.

Declaration of competing interest

The authors declare that they have no conflict of interest.

References

[1] C.J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, K. D, Key challenges for delivering clinical impact with artificial intelligence, BMC Med. 17 (2019).
[2] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, F. Herrera, Explainable Artificial Intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI, Inf. Fusion 58 (2020) 82–115.
[3] M. Ahamed, A. Imran, Joint learning with local and global consistency for improved medical image segmentation, in: Annual Conference on Medical Image Understanding and Analysis, 2022.
[4] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale, 2010.
[5] R. Strudel, R. Garcia, I. Laptev, C. Schmid, Segmenter: transformer for semantic segmentation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021.
[6] H.P. Chan, R.K. Samala, L.M. Hadjiiski, C. Zhou, Deep learning in medical image analysis, Adv. Exp. Med. Biol. (2020) 3–21.
[7] E. Sorantin, M. Grasser, A. Hemmelmayr, S.H.F. Tschauner, V. Weiss, J. Lacekova, A. Holzinger, The augmented radiologist: artificial intelligence in the practice of radiology, Pediatr. Radiol. (2021) 1–13.
[8] T. Folke, S. Yang, S. Anderson, P. Shafto, Explainable AI for medical imaging: explaining pneumothorax diagnoses with Bayesian teaching, Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications III 11746 (2021) 644–664.
[9] New RCR census shows the NHS needs nearly 2,000 more radiologists [Online]. Available: https://www.rcr.ac.uk/posts/new-rcr-census-shows-nhs-needs-nearly-2000-more-radiologists. (Accessed 15 September 2022).
[10] G. Ras, M. van Gerven and P. Haselager, "Explanation Methods in Deep Learning: Users, Values, Concerns and Challenges," *Explainable And Interpretable Models in Computer Vision and Machine Learning,* pp. 19-36.
[11] M. Nassar, K. Salah, M. ur Rehman and D. Svetinovic, "Blockchain for explainable and trustworthy artificial intelligence," Wiley Interdisciplinary Reviews: Data Min. Knowl. Discov., vol. 10, no. 1.
[12] S. Nazir, S. Patel, D. Patel, Model optimisation techniques for convolutional neural networks, in: Handbook of Research on New Investigations in Artificial Life, *AI, and Machine Learning*, IGI Global, 2022, pp. 269–298.
[13] N. McCarthy, A. Dahlan, T. Cook, N. O'Hare, M. Ryan, B. St John, A. Lawlor, K. Curran, Enterprise imaging and big data: a review from a medical physics perspective, Phys. Med. 83 (2021) 206–220.
[14] P. Tripicchio, S. D'Avella, Is deep learning ready to satisfy industry needs? Procedia Manuf. 51 (2020) 1192–1199.
[15] R. LaLonde, D. Torigian and U. Bagci, "Encoding visual attributes in capsules for explainable medical diagnoses," in Medical Image Computing and Computer Assisted Intervention – MICCAI 2020.
[16] J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, H. Lipson, Understanding Neural Networks through Deep Visualization, 2015 [Online] Sep. 2022.
[17] M. Oussalah, AI explainability. A bridge between machine vision and Natural Language processing, in: International Conference on Pattern Recognition, 2021.
[18] [Online]. Available FICO Community, https://community.fico.com/s/explainable-machine-learning-challenge. (Accessed 15 September 2022).
[19] M. Ghassemi, L. Oakden-Rayner, A.L. Beam, The false hope of current approaches to explainable artificial intelligence in health care, The Lancet Digital Health 3 (11) (2021).
[20] N. Díaz-Rodríguez, A. Lamas, J. Sanchez, G. Franchi, I. Donadello, S. Tabik, D. Filliat, P. Cruz, R. Montes, F. Herrera, EXplainable Neural-Symbolic Learning (X-NeSyL) methodology to fuse deep learning representations with expert knowledge graphs: the MonuMAI cultural heritage use case, Inf. Fusion 79 (2022) 58–83.
[21] A. Adadi, M. Berrada, Peeking inside the black-box: a survey on explainable artificial intelligence (XAI), IEEE Access 6 (2018) 52138–52160.
[22] G. Alicioglu, B. Sun, A survey of visual analytics for Explainable Artificial Intelligence methods, Comput. Graph. 102 (2021) 502–520.
[23] P. Linardatos, V. Papastefanopoulos, S. Kotsiantis, A.I. Explainable, A review of machine learning interpretability methods, Entropy 23 (1) (2020).
[24] H. Xiaowei, D. Kroening, W. Ruan, J. Sharp, Y. Sun, E. Thamo, M. Wu, X. Yi, A survey of safety and trustworthiness of deep neural networks: verification, testing, adversarial attack and defence, and interpretability, Computer Science Review 37 (2020).
[25] W. Samek and K. R. Müller, "Towards explainable artificial intelligence," in Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Lecture Notes in Computer Science, vol. vol. 11700, Springer, Cham..
[26] A. Das and P. Rad, "Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey," [Online]. Available: https://doi.org/10.48550/arXiv.2006.11371. [Accessed 15 September 2022].
[27] A.M. Antoniadi, Y. Du, Y. Guendouz, L. Wei, C. Mazo, B.A. Becker, C. Mooney, Current challenges and future opportunities for XAI in machine learning-based clinical decision support systems: a systematic review, Appl. Sci. 11 (11) (2021).
[28] A. Singh, S. Sengupta, V. Lakshminarayanan, Explainable deep learning models in medical image analysis, Journal of Imaging 6 (6) (2020).
[29] B.H.v. d. Velden, H.J. Kuijf, K.G Gilhuijs, M.A. Viergever, Explainable artificial intelligence (XAI) in deep learning-based medical image analysis, Med. Image Anal. 79 (2022).
[30] E. Tjoa, C. Guan, A survey on explainable artificial intelligence (XAI): toward medical XAI, IEEE Transact. Neural Networks Learn. Syst. 32 (11) (2021) 4793–4813.
[31] M. Reyes, R. Meier, S. Pereira, C.A. Silva, F.-M. Dahlweid, H.v. Tengg-Kobligk, R. M. Summers, R. Wiest, On the interpretability of artificial intelligence in radiology: challenges and opportunities, Radiology: Artif. Intell. 2 (3) (2020).
[32] A.F. Markus, J.A. Kors, P.R. Rijnbeek, The role of explainability in creating trustworthy artificial intelligence for health health care: a comprehensive survey of the terminology, design choices, and evaluation strategies, J. Biomed. Inf. 113 (2021), 103655.
[33] T.T. Nguyen, Q.V.H. Nguyen, D.T. Nguyen, E.B. Hsu, S. Yang, P. Eklund, Artificial Intelligence in the Battle against Coronavirus (COVID-19): A Survey and Future Research Directions, 2021, https://doi.org/10.48550/arXiv.2008.07343 [Online]. Available:. (Accessed 15 September 2022).
[34] R. Karthik, R. Menaka, M. Hariharan, G.S. Kathiresan, AI for COVID-19 detection from radiographs: incisive analysis of state of the art techniques, IRBM 43 (5) (2021) 486–510.
[35] S. O'Sullivan, F. Jeanquartier, C. Jean-Quartier, A. Holzinger, D. Shiebler, P. Moon, C. Angione, Developments in AI and machine learning for neuroimaging, in: Artificial Intelligence and Machine Learning for Digital Pathology. Lecture Notes in Computer Science vol. 12090, Springer, Cham., 2020.
[36] P. Panagiotis, L. Brocki, N.C. Chung, W. Marchadour, F. Vermet, L. Gaubert, V. Eleftheriadis, D. Plachouris, D. Visvikis, G.C. Kagadis, M. Hatt, Artificial intelligence: deep learning in oncological radiomics and challenges of interpretability and data harmonization, Phys. Med. 83 (2021) 108–121.
[37] R.K. Singh, R. Pandey, R.N. Babu, COVIDScreen: Explainable Deep Learning Framework for Differential Diagnosis of COVID-19 Using Chest X-Rays," *Neural Computing And Applications*, 2021.
[38] J.M. Durán, Dissecting Scientific Explanation in AI (sXAI): A Case for Medicine and Healthcare, vol. 297, Artificial Intelligence, 2021.
[39] N. Norori, Q. Hu, F.M. Aellen, F.D. Faraci, A. Tzovara, Addressing bias in big data and AI for health care: a call for open science, Patterns 2 (10) (2021).
[40] C. Meske, E. Bunde, J. Schneider, M. Gersch, Explainable Artificial Intelligence: Objectives, Stakeholders, and Future Research Opportunities, Information Systems Management, 2020.
[41] K.B. Ahmed, G.M. Goldgof, R. Paul, D.B. Goldgof, L.O. Hall, Deep Learning Models May Spuriously Classify Covid-19 from X-Ray Images Based on Confounders, 2021, https://doi.org/10.48550/arXiv.2102.04300 [Online]. Available:. (Accessed 15 September 2022).

[42] J.R. Zech, M.A. Badgeley, M. Liu, A.B. Costa, J.J. Titano, E.K. Oermann, Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study, PLoS Med. 15 (11) (2018).
[43] A.J. DeGrave, J.D. Janizek, S.-I. Lee, AI for radiographic COVID-19 detection selects shortcuts over signal, Nat. Mach. Intell. 3 (7) (2021) 610–619.
[44] D. Doran, S. Schulz, T.R. Besold, What Does Explainable AI Really Mean? A New Conceptualization of Perspectives, 2017, https://doi.org/10.48550/arXiv.1710.00794 [Online]. Available:. (Accessed 15 September 2022).
[45] D. Gunning, D.W. Aha, DARPA's Explainable Artificial Intelligence Program, AI Magazine, 2019.
[46] W. Knight, The U.S. Military Wants its Autonomous Machines to Explain Themselves, MIT Technology Review, 2017.
[47] D.A. Broniatowski, Psychological Foundations of Explainability and Interpretability in Artificial Intelligence, NIST, 2021.
[48] P.J. Phillips, C.A. Hahn, P.C. Fontana, A.N. Yates, K. Greene, D.A. Broniatowski, M.A. Przybocki, NISTIR 8312: Four Principles of Explainable Artificial Intelligence, NIST, 2021.
[49] J. Gryz, M. Rojszczak, Black box algorithms and the rights of individuals: no easy solution to the 'explainability' problem, Internet Policy Review 10 (2021).
[50] B. Goodman and S. Flaxman, "European Union regulations on algorithmic decision-making and a "right to explanation"," AI Mag., vol. 38, no. 3, pp. 50-57.
[51] F. Doshi-Velez, B. Kim, Towards a Rigorous Science of Interpretable Machine Learning, 2017, https://doi.org/10.48550/arXiv.1702.08608 [Online]. Available:. (Accessed 15 September 2022).
[52] K.D. Abeyrathna, O.C. Granmo, M. Goodwin, Extending the tsetlin machine with integer-weighted clauses for increased interpretability, IEEE Access 9 (2021).
[53] Taking Responsibility, Responsible Artificial Intelligence. Artificial Intelligence: Foundations, Theory, and Algorithms, Springer, Cham., 2019.
[54] Ethics Guidelines for Trustworthy AI, Eurpean Commission, 2021.
[55] D.C. Elton, Self-explaining AI as an alternative to interpretable AI, in: International Conference on Artificial General Intelligence, 2020.
[56] S.-K. Yeom, P. Seegerer, S. Lapuschkin, A. Binder, S. Wiedemann, K.-R. Müller, W. Samek, Pruning by explaining: a novel criterion for deep neural network pruning, Pattern Recogn. 115 (2021).
[57] A.J. London, Artificial intelligence and black-box medical decisions: accuracy versus explainability, Hastings Cent. Rep. 49 (1) (2019) 15–21.
[58] A. Rai, Explainable AI: from black box to glass box, J. Acad. Market. Sci. 48 (1) (2019) 137–141.
[59] K. Bauer, O. Hinz, W.v. d. Aalst, C. Weinhardt, Expl(AI)n it to me – explainable AI and information systems research, Business & Information Systems Engineering 63 (2) (2021).
[60] A. Páez, The pragmatic turn in explainable artificial intelligence (XAI), Minds Mach. 29 (3) (2019) 441–459.
[61] A. Preece, D. Harborne, D. Braines, R. Tomsett and S. Chakraborty, "Stakeholders in Explainable AI," [Online]. Available: https://doi.org/10.48550/arXiv.1810.00184. [Accessed 15 September 2022].
[62] Y. Wang, Y. Wang, Z.C.H. Li, X. Tang, Y. Yang, CNN hyperparameter optimization based on CNN visualization and perception hash algorithm, in: 19th International Symposium on Distributed Computing and Applications for Business Engineering and Science, DCABES), 2020.
[63] A. Heuillet, F. Couthouis, N. Díaz-Rodríguez, Explainability in deep reinforcement learning, Knowl. Base Syst. 214 (2021).
[64] ISO 13485:2016, ISO - International Organization for Standardization, 2018.
[65] The OECD Artificial Intelligence (AI) Principles," oecd.Ai.
[66] Responsible.ai".
[67] Q. Zhao, T. Hastie, Causal interpretations of black-box models, J. Bus. Econ. Stat. 39 (1) (2021) 272–281.
[68] A. Goldstein, A. Kapelner, J. Bleich, E. Pitkin, Peeking inside the black box: visualizing statistical learning with plots of individual conditional expectation, J. Comput. Graph Stat. 24 (1) (2015) 44–65.
[69] K.R. Mopuri, U. Garg, R.V. Babu, C.N.N. Fixations, An unraveling approach to visualize the discriminative image regions, IEEE Trans. Image Process. 28 (5) (2019) 2116–2125.
[70] M.T. Ribeiro, S. Singh, C. Guestrin, Why should i trust you?': explaining the predictions of any classifier, in: Proc. 22nd ACM SIGKDD Int. Conf. Knowl, . Discovery Data Mining, 2016.
[71] S.M. Shankaranarayana, D. Runje, ALIME: autoencoder based approach for local interpretability, in: International Conference on Intelligent Data Engineering and Automated Learning, 2019.
[72] V. Petsiuk, A. Das and K. Saenko, "RISE: Randomized Input Sampling for Explanation of Black-box Models," [Online]. Available: https://doi.org/10.48550/arXiv.1806.07421. [Accessed 15 September 2022].
[73] B. Vasu, C. Long, Iterative and adaptive sampling with spatial attention for black-box model explanations, in: IEEE Winter Conference on Applications of Computer Vision, WACV), 2020.
[74] W.S. Monroe, F.M. Skidmore, D.G. Odaibo, M.M. Tanik, HihO: accelerating artificial intelligence interpretability for medical imaging in IoT applications using hierarchical occlusion, Neural Comput. Appl. 33 (2021) 6027–6038.
[75] X. Li, S. Ji, Neural image compression and explanation, IEEE Access 8 (2020) 214605–214615.
[76] S. M. Muddamsetty, N. S. J. Mohammad and T. B. Moeslund, "SIDU: similarity difference and uniqueness method for explainable AI," in IEEE International Conference on Image Processing (ICIP).
[77] K. Simonyan, A. Vedaldi and A. Zisserman, "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps," [Online]. Available: https://doi.org/10.48550/arXiv.1312.6034. [Accessed 15 September 2022].
[78] Y. Rao, J. Ni, H. Zhao, Deep learning local descriptor for image splicing detection and localization, IEEE Access 8 (2020) 25611–25625.
[79] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad Cam, Visual explanations from deep networks via gradient-based localization, in: IEEE International Conference on Computer Vision, ICCV), 2017.
[80] A. Chattopadhay, A. Sarkar, P. Howlader, V.N. Balasubramanian, Grad-CAM++: generalized gradient-based visual explanations for deep convolutional networks, in: IEEE Winter Conference on Applications of Computer Vision, WACV), 2018.
[81] S. Sattarzadeh, M. Sudhakar, K.N. Plataniotis, J. Jang, Y. Jeong, H. Kim, Integrated grad-cam: sensitivity-aware visual explanation of deep convolutional networks via integrated gradient-based scoring, in: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP), 2021.
[82] M.B. Muhammad, M. Yeasin, Eigen-CAM: class activation map using principal components, in: International Joint Conference on Neural Networks, IJCNN), 2020.
[83] 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) (2015).
[84] Y.J. Jung, S.H. Han, H.J. Choi, Explaining CNN and RNN using selective layer-wise relevance propagation, IEEE Access 9 (2021) 18670–18681.
[85] J. Zhang, S.A. Bargal, Z. Lin, J. Brandt, X. Shen, S. Sclaroff, Top-down neural attention by excitation Backprop, Int. J. Comput. Vis. 126 (2018) 1084–1102.
[86] A. Shrikumar, P. Greenside, A. Kundaje, Learning important features through propagating activation differences, in: International Conference on Machine Learning, 2017.
[87] M.D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: European Conference on Computer Vision, 2014.
[88] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, R. Sayres, Interpretability beyond feature attribution: quantitative testing with concept activation vectors (TCAV), in: International Conference on Machine Learning, 2018.
[89] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, K.-R. Müller, Explaining nonlinear classification decisions with deep Taylor decomposition, Pattern Recogn. 65 (2017) 211–222.
[90] S. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Adv. Neural Inf. Process. Syst. 30 (2017).
[91] R. Tan, N. Khan, L. Guan, Locality guided neural networks for explainable artificial intelligence, in: International Joint Conference on Neural Networks, IJCNN), 2020.
[92] K. Xu, J.L. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R.S. Zemel, Y. Bengio, Show, Attend and Tell: Neural Image Caption Generation with Visual Attention, International conference on machine learning, 2015.
[93] Y. Dong, H. Su, J. Zhu, B. Zhang, Improving interpretability of deep neural networks with semantic information, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017.
[94] S.-H. Han, M.-S. Kwon, H.-J. Choi, EXplainable AI (XAI) approach to image captioning, J. Eng. 2020 (13) (2020) 589–594.
[95] Z. Qi, S. Khorram, L. Fuxin, Embedding Deep Networks into Visual Explanations, Artificial Intelligence, 2020.
[96] N. Puri, P. Gupta, P. Agarwal, S. Verma, B. Krishnamurthy, MAGIX: Model Agnostic Globally Interpretable Explanations, 2018, https://doi.org/10.48550/arXiv.1706.07160 [Online]. Available:. (Accessed 15 September 2022).
[97] H. Lakkaraju, E. Kamar, R. Caruana, J. Leskovec, Interpretable & Explorable Approximations of Black Box Models, 2017, https://doi.org/10.48550/arXiv.1707.01154 [Online]. Available:. (Accessed 15 September 2022).
[98] R. Confalonieri, T. Weyde, T.R. Besold, F.M.d.P. Martín, Using ontologies to enhance human understandability of global post-hoc explanations of black-box models, Artif. Intell. 296 (2021).
[99] F. Scarselli, M. Gori, A. Tsoi, M. Hagenbuchner, G. Monfardini, The graph neural network model, IEEE Trans. Neural Network. 20 (1) (2009) 61–80.
[100] A. Holzinger, B. Malle, A. Saranti, B. Pfeifer, Towards multi-modal causability with Graph Neural Networks enabling information fusion for explainable AI, Inf. Fusion 71 (2021) 28–37.
[101] S. Wachter, B. Mittelstadt, C. Russell, Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR, SSRN Electronic Journal, 2017.
[102] S. Dandl, C. Molnar, M. Binder, B. Bischl, Multi-objective counterfactual explanations, in: International Conference on Parallel Problem Solving from Nature, 2020.
[103] A.R. Akula, Shuai Wang, S.-C. Zhu, CoCoX: generating conceptual and counterfactual explanations via fault-lines, in: The Thirty-Fourth AAAI Conference on Artificial Intelligence, 2020.
[104] P.W. Koh, P. Liang, Proceedings of the 34th International Conference on Machine Learning, PMLR, 2017.
[105] S.C. Yang, W.K. Vong, R.B. Sojitra, T. Folke, P. Shafto, Mitigating belief projection in explainable artificial intelligence via Bayesian teaching, Sci. Rep. 11 (1) (2021) 1–17.
[106] E.M. Kenny, M.T. Keane, On generating plausible counterfactual and semi-factual explanations for deep learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2021.
[107] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, B. Kim, Sanity checks for saliency maps, Adv. Neural Inf. Process. Syst. 31 (2018).
[108] R. Chimatapu, H. Hagras, M. Kern, G. Owusu, Hybrid deep learning type-2 fuzzy logic systems for explainable AI, in: IEEE International Conference on Fuzzy Systems, FUZZ-IEEE), 2020.

[109] M. Borg, R. Jabangwe, S. Åberg, A. Ekblom, L. Hedlund, A. Lidfeldt, Test automation with grad-CAM heatmaps - a future pipe segment in MLOps for vision AI?, in: IEEE International Conference on Software Testing, Verification and Validation Workshops ICSTW), 2021.

[110] A. Chattopadhay, A. Sarkar, P. Howlader, V. Balasubramanian, Grad-CAM++: generalized gradient-based visual explanations for deep, in: 2018 IEEE Winter Conference on Applications of Computer Vision, 2018.

[111] M.D. Zeiler, G.W. Taylor, R. Fergus, Adaptive deconvolutional networks for mid and high level feature learning, in: International Conference on Computer Vision, 2011.

[112] T. Kashima, R. Hataya, H. Nakayama, Visualizing association in exemplar-based classification, in: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP), 2021.

[113] E.M. Kenny, M.T. Keane, Explaining Deep Learning using examples: optimal feature weighting methods for twin systems using post-hoc, explanation-by-example in XAI, Knowl. Base Syst. 233 (2021).

[114] S. Kashyap, A. Karargyris, J. Wu, Y. Gur, A. Sharma, K. Wong, M. Moradi, T. Syeda-Mahmood, Looking in the right place for anomalies: explainable ai through automatic location learning, in: IEEE 17th International Symposium on Biomedical Imaging, ISBI), 2020.

[115] J. Hong, J. Fu, Y. Uh, T. Mei, H. Byun, Exploiting hierarchical visual features for visual question answering, Neurocomputing 351 (2019) 187–195.

[116] M.u. Hassan, P. Mulhem, D. Pellerin, G. Quénot, Explaining visual classification using attributes, in: International Conference on Content-Based Multimedia Indexing, CBMI), 2019.

[117] I. Tiddi, S. Schlobach, Knowledge Graphs as Tools for Explainable Machine Learning: a Survey, Artificial Intelligence, 2021.

[118] M. Gaur, K. Faldu, A. Sheth, Semantics of the black-box: can knowledge graphs help make deep learning systems more interpretable and explainable? IEEE Internet Computing 25 (1) (2021) 51–59.

[119] V. Horta, A. Mileo, International Conference on Database and Expert Systems Applications, 2019.

[120] B. Kim, R. Khanna, O. Koyejo, Examples are not enough, learn to criticize! criticism for interpretability, in: Proc. 29th Conf. Neural Inf. Process. Syst, , NIPS), 2016.

[121] J. Chandrasekaran, Y. Lei, R. Kacker and D. R. Kuhn, "A combinatorial approach to explaining image classifiers," in IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW), 20021.

[122] M. Suzuki, Y. Kameya, T. Kutsuna, N. Mitsumoto, Understanding the reason for misclassification by generating counterfactual images, in: 17th International Conference on Machine Vision Applications, MVA), 2021.

[123] M. Alber, S. Lapuschkin, P. Seegerer, M. Hägele, K. Schütt, G. Montavon, W. Samek, K. Müller, S. Dähne, P. Kindermans, iNNvestigate neural networks, J. Mach. Learn. Res. 20 (93) (2019) 1–8.

[124] T. Spinner, U. Schlegel, H. Schafer, M. El-Assady, IEEE Trans. Visual. Comput. Graph. 26 (1) (2020) 1064–1074.

[125] C. Schorr, P. Goodarzi, F. Chen, T. Dahmen, Neuroscope: an explainable AI toolbox for semantic segmentation and image classification of convolutional neural nets, Appl. Sci. 5 (2021).

[126] A. Vyas, P. Calyam, An interactive graphical visualization approach to CNNs and RNNs, in: IEEE Applied Imagery Pattern Recognition Workshop, AIPR), 2020.

[127] Introduction to Vertex explainable AI, Google Cloud, [Online]. Available: https://cloud.google.com/vertex-ai/docs/explainable-ai/overview. (Accessed 15 September 2022).

[128] Amazon SageMaker clarify model explainability - Amazon SageMaker," AWS [Online]. Available: https://docs.aws.amazon.com/sagemaker/latest/dg/clarify-model-explainability.html. (Accessed 15 September 2022).

[129] Responsible and trusted AI - cloud adoption framework, Microsoft, [Online]. Available: https://docs.microsoft.com/en-us/azure/cloud-adoption-framework/innovate/best-practices/trusted-ai. (Accessed 15 September 2022).

[130] L. Arras, A. Osman, W. Samek, CLEVR-XAI: a benchmark dataset for the ground truth evaluation of neural network explanations, Inf. Fusion 81 (2022) 14–40.

[131] L.B. Fulton, J.Y. Lee, Q. Wang, Z. Yuan, J. Hammer, A. Perer, Getting playful with explainable AI: games with a purpose to improve human understanding of AI, in: Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems, 2020.

[132] L. Fan, C. Liu, Y. Zhou, T. Zhang, Q. Yang, Interpreting and evaluating black box models in a customizable way, in: IEEE International Conference on Big Data, Big Data), 2020.

[133] W. Jin, X. Li, G. Hamarneh, Evaluating explainable AI on a multi-modal medical imaging task: can existing algorithms fulfill clinical requirements?, in: Association for the Advancement of Artificial Intelligence Conference AAAI), 2022.

[134] R. Zicari, J. Brodersen, J. Brusseau, B. Düdder, T. Eichhorn, T. Ivanov, G. Kararigas, P. Kringen, M. McCullough, F. Möslein, N. Mushtaq, Z-Inspection: a process to assess trustworthy AI, IEEE Transactions on Technology and Society 2 (2) (2021).

[135] L. Eldridge, What Is Radiology? Understanding Diagnostic, Interventional, and Therapeutic Radiology, 2021 [Online]. Available: https://www.verywellhealth.com/what-is-radiology-5085100. (Accessed 15 September 2022).

[136] A. Hilbert, L.A. Ramos, H.J. van Os, S. Olabarriaga, M. Tolhuisen, M.J. Wermer, R.S. Barros, I. van der Schaaf, D. Dippel, Y.W.E.M. Roos, W. van Zwam, Data-efficient deep learning of radiological image data for outcome prediction after endovascular treatment of patients with acute ischemic stroke, Comput. Biol. Med. 115 (2019).

[137] P.R. Magesh, R.D. Myloth, R.J. Tom, An explainable machine learning model for early detection of Parkinson's disease using LIME on DaTSCAN imagery, Comput. Biol. Med. 126 (2020).

[138] T. Pianpanit, S. Lolak, P. Sawangjai, T. Sudhawiyangkul, T. Wilaiprasitporn, Parkinson's disease recognition using SPECT image and interpretable AI: a tutorial, IEEE Sensor. J. 21 (20) (2021).

[139] M. Nazari, A. Kluge, I. Apostolova, S. Klutmann, S. Kimiaei, M. Schroeder, R. Buchert, Explainable AI to improve acceptance of convolutional neural networks for automatic classifcation of dopamine transporter SPECT in the diagnosis of clinically uncertain parkinsonian syndromes, Eur. J. Nucl. Med. Mol. Imag. 49 (4) (2021) 1176–1186.

[140] H. Choi, Y.K. Kim, E.J. Yoon, J.Y. Lee, D.S. Lee, Cognitive signature of brain FDG PET based on deep learning: domain transfer from Alzheimer's disease to Parkinson's disease, Eur. J. Nucl. Med. Mol. Imag. 47 (2) (2020) 403–412.

[141] G. Yang, Q. Ye, J. Xia, Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: a mini-review, two showcases and beyond, Inf. Fusion 77 (2022) 29–52.

[142] F. Eitel, E. Soehler, J. Bellmann-Strobl, A. Brandt, K. Ruprecht, R. Giess, J. Kuchling, S. Asseyer, M. Weygandt, J. Haynes, M. Scheel, Uncovering convolutional neural network decisions for diagnosing multiple sclerosis on conventional MRI using layer-wise relevance propagation, Neuroimage: Clinical 24 (2019).

[143] M.S. Kamal, A. Northcote, L. Chowdhury, N. Dey, R.G. Crespo, E. Herrera-Viedma, Alzheimer's patient analysis using image and gene expression data and explainable-AI to present associated genes, IEEE Trans. Instrum. Meas. 70 (2021) 1–7.

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

[145] E. Lee, J.-S. Choi, M. Kim, H.-I. Suk, Toward an interpretable Alzheimer's disease diagnostic model with regional abnormality representation via deep learning, Neuroimage 202 (2019).

[146] G. Levakov, G. Rosenthal, I. Shelef, T.R. Raviv, G. Avidan, From a deep learning model back to the brain—identifying regional predictors and their relation to aging, Hum. Brain Mapp. 41 (12) (2020) 3235–3252.

[147] S. Shinde, S. Prasad, Y. Saboo, R. Kaushick, J. Saini, P.K. Pal, M. Ingalhalikar, Predictive markers for Parkinson's disease using deep neural nets on neuromelanin sensitive MRI, Neuroimage: Clinical 22 (2019).

[148] S. Chakraborty, S. Aich, H.C. Kim, Detection of Parkinson's disease from 3T T1 weighted MRI scans using 3D convolutional neural network, Diagnostics 10 (6) (2020).

[149] A. Kumar, R. Manikandan, U. Kose, D. Gupta, S.C. Satapathy, Doctor's dilemma: evaluating an explainable subtractive spatial Lightweight convolutional neural network for brain tumor diagnosis, ACM Trans. Multimed Comput. Commun. Appl 17 (3s) (2021) 1–26.

[150] S. Pereira, R. Meier, V. Alves, M. Reyes, C. Silva, Automatic brain tumor grading from MRI data using convolutional neural networks and quality assessment, in: Understanding and Interpreting Machine Learning in Medical Image Computing Applications, 2018.

[151] P. Natekar, A. Kori, G. Krishnamurthi, Demystifying brain tumor segmentation networks: interpretability and uncertainty analysis, Front. Comput. Neurosci. 14 (2020).

[152] P. Windisch, P. Weber, C. Fürweger, F. Ehret, M. Kufeld, D. Zwahlen, A. Muacevic, Implementation of model explainability for a basic brain tumor detection using convolutional neural networks on MRI slices, Neuroradiology 62 (11) (2020).

[153] A. Lopatina, S. Ropele, R. Sibgatulin, J.R. Reichenbach, D. Güllmar, Investigation of deep-learning-driven identification of multiple sclerosis patients based on susceptibility-weighted images using relevance analysis, Front. Neurosci. 14 (2020).

[154] M. Jimeno, K. Ravi, Z. Jin, D. Oyekunle, G. Ogbole, S. Geethanath, ArtifactID: identifying artifacts in low-field MRI of the brain using deep learning, Magn. Reson. Imag. 89 (2022) 42–48.

[155] B. Xie, T. Lei, N. Wang, H. Cai, J. Xian, M. He, L. Zhang, H. Xie, Computer-aided diagnosis for fetal brain ultrasound images using deep convolutional neural networks, Int. J. Comput. Assist. Radiol. Surg. 15 (8) (2020) 1303–1312.

[156] Z. Tang, K.V. Chuang, C. DeCarli, L.W. Jin, L. Beckett, M.J. Keiser, B.N. Dugger, Interpretable classification of Alzheimer's disease pathologies with a convolutional neural network pipeline, Nat. Commun. 10 (1) (2019) 1–4.

[157] M.M. Ahsan, K.D. Gupta, M.M. Islam, S. Sen, M.L. Rahman, M.S. Hossain, COVID-19 symptoms detection based on NasNetMobile with explainable AI using various imaging modalities, Machine Learning and Knowledge Extraction 2 (4) (2020) 490–504.

[158] H. Panwar, P.K. Gupta, M.K. Siddiqui, R. Morales-Menendez, P. Bhardwaj, V. Singh, A Deep Learning and Grad-CAM Based Color Visualization Approach for Fast Detection of COVID-19 Cases Using Chest X-Ray and CT-Scan Images," *Chaos*, vol. 140, Solitons & Fractals, 2020.

[159] A. Degerli, M. Ahishali, M. Yamac, S. Kiranyaz, M.E.H. Chowdhury, K. Hameed, T. Hamid, R. Mazhar, M. Gabbouj, COVID-19 infection map generation and detection from chest X-ray images, Health Inf. Sci. Syst. 9 (1) (2021).

[160] T. Mahmud, M.A. Rahman, S.A. Fattah, CovXNet: a multi-dilation convolutional neural network for automatic COVID-19 and other pneumonia detection from chest X-ray images with transferable multi-receptive feature optimization, Comput. Biol. Med. 122 (2020).

[161] M. Chetoui, M.A. Akhloufi, Deep efficient neural networks for explainable COVID-19 detection on CXR images, in: International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, 2021.
[162] M. Karim, T. Döhmen, M. Cochez, O. Beyan, D. Rebholz-Schuhmann, S. Decker, Deepcovidexplainer: explainable COVID-19 diagnosis from chest X-ray images, in: IEEE International Conference on Bioinformatics and Biomedicine, BIBM), 2020.
[163] S. Ravi, S. Khoshrou, M. Pechenizkiy, ViDi: descriptive visual data clustering as radiologist assistant in COVID-19 streamline diagnostic [Online]. Available:, 2020 https://doi.org/10.48550/arXiv.2011.14871. (Accessed 15 September 2022).
[164] K.-S. Lee, J.Y. Kim, E.-t. Jeon, W.S. Choi, N.H. Kim, K.Y. Lee, Evaluation of scalability and degree of fine-tuning of deep convolutional neural networks for COVID-19 screening on chest X-ray images using explainable deep-learning algorithm, J. Personalized Med. 10 (4) (2020).
[165] J. Kim, M. Kim, Y. Ro, Interpretation of lesional detection via counterfactual generation, in: IEEE International Conference on Image Processing, ICIP), 2021.
[166] H. Ren, A.B. Wong, W. Lian, W. Cheng, Y. Zhang, J. He, Q. Liu, J. Yang, C. J. Zhang, K. Wu, H. Zhang, Interpretable pneumonia detection by combining deep learning and explainable models with multisource data, IEEE Access 9 (2021) 95872–95883.
[167] H. Liz, M. Sánchez-Montañés, A. Tagarro, S. Domínguez-Rodríguez, R. Dagan, D. Camacho, Ensembles of Convolutional Neural Network models for pediatric pneumonia diagnosis, Future Generat. Comput. Syst. 122 (2021) 220–233.
[168] B. Chen, J. Li, G. Lu, D. Zhang, Lesion location attention guided network for multi-label thoracic disease classification in chest X-rays, IEEE journal of biomedical and health informatics 24 (7) (2019) 2016–2027.
[169] H. Liu, L. Wang, Y. Nan, F. Jin, Q. Wang, J. Pu, SDFN: segmentation-based deep fusion network for thoracic disease classification in chest X-ray images, Comput. Med. Imag. Graph. 75 (2019) 66–73.
[170] J.A. Dunnmon, D. Yi, C.P. Langlotz, C. Ré, D.L. Rubin, M.P. Lungren, Assessment of convolutional neural networks for automated classification of chest radiographs, Radiology 290 (2) (2019) 537–544.
[171] S. Rajaraman, S. Candemir, G. Thoma, S. Antani, Visualizing and explaining deep learning predictions for pneumonia detection in pediatric chest radiographs, SPIE Medical Imaging 10950 (2019) 200–211.
[172] H. Liz, M. Sánchez-Montañés, A. Tagarro, S. Domínguez-Rodríguez, R. Dagan, D. Camacho, Ensembles of Convolutional Neural Network models for pediatric pneumonia diagnosis, Future Generat. Comput. Syst. 122 (2021) 220–233.
[173] H. Alshazly, C. Linse, E. Barth, T. Martinetz, Explainable COVID-19 detection using chest CT scans and deep learning, Sensors 21 (2) (2021).
[174] M. Pennisi, I. Kavasidis, C. Spampinato, V. Schinina, S. Palazzo, F.P. Salanitri, G. Bellitto, F. Rundo, M. Aldinucci, M. Cristofaro, a.P. Campioni, An explainable AI system for automated COVID-19 assessment and lesion categorization from CT-scans, Artif. Intell. Med. 118 (2021).
[175] Y.-H. Wu, S.-H. Gao, J. Mei, J. Xu, D.-P. Fan, R.-G. Zhang, M.-M. Cheng, JCS: an explainable COVID-19 diagnosis system by Joint classification and segmentation, IEEE Trans. Image Process. 30 (2021) 3113–3126.
[176] A. Katzmann, O. Taubmann, S. Ahmad, A. Mühlberg, M. Sühling, H. Groß, Explaining clinical decision support systems in medical imaging using cycle-consistent activation maximization, Neurocomputing (2021) 141–156.
[177] R. Xu, Z. Cong, X. Ye, Y. Hirano, S. Kido, T. Gyobu, Y. Kawata, O. Honda, N. Tomiyama, Pulmonary textures classification via a multi-scale attention network, IEEE journal of biomedical and health informatics 24 (7) (2019) 2041–2052.
[178] S.M. Humphries, A.M. Notary, J.P. Centeno, M.J. Strand, J.D. Crapo, E. K. Silverman, D.A. Lynch, Deep learning enables automatic classification of emphysema pattern at CT, Radiology 294 (2) (2020) 434–444.
[179] R. Paul, M. Schabath, R. Gillies, L. Hall, D. Goldgof, Convolutional Neural Network ensembles for accurate lung nodule malignancy prediction 2 years in the future, Comput. Biol. Med. 122 (2020).
[180] Y. Lei, Y. Tian, H. Shan, J. Zhang, G. Wang, M.K. Kalra, Shape and margin-aware lung nodule classification in low-dose CT images via soft activation mapping, Med. Image Anal. 60 (2020).
[181] A. Hosny, C. Parmar, T.P. Coroller, P. Grossmann, R. Zeleznik, A. Kumar, J. Bussink, R.J. Gillies, R.H. Mak, H.J. Aerts, Deep learning for lung cancer prognostication: a retrospective multi-cohort radiomics study, PLoS Med. 15 (11) (2018).
[182] D. Kumar, V. Sankar, D. Clausi, G.W. Taylor, A. Wong, SISC: end-to-end interpretable discovery radiomics-driven lung cancer prediction via stacked interpretable sequencing cells, IEEE Access 7 (2019) 145444–145454.
[183] H. Ko, H. Chung, W.S. Kang, K.W. Kim, Y. Shin, S.J. Kang, J.H. Lee, Y.J. Kim, N. Y. Kim, H. Jung, J. Lee, COVID-19 pneumonia diagnosis using a simple 2D deep learning framework with a single chest CT image: model development and validation, J. Med. Internet Res. 22 (6) (2020).
[184] A. Teramoto, A. Yamada, Y. Kiriyama, T. Tsukamoto, K. Yan, L. Zhang, K. Imaizumi, K. Saito, H. Fujita, Automated classification of benign and malignant cells from lung cytological images using deep convolutional neural network, Inform. Med. Unlocked 16 (2019).
[185] A. Kind, G. Azzopardi, An explainable AI-based computer aided detection system for diabetic retinopathy using retinal fundus images, in: International Conference on Computer Analysis of Images and Patterns, Springer, Cham., 2019, pp. 457–468.
[186] M. Shorfuzzaman, M.S. Hossain, A.E. Saddik, An explainable deep learning ensemble model for robust diagnosis of diabetic retinopathy grading, ACM Trans. Multimed Comput. Commun. Appl 17 (3s) (2021) 1–24.
[187] R. Reguant, S. Brunak, S. Saha, Understanding inherent image features in CNN-based assessment of diabetic retinopathy, Sci. Rep. 11 (1) (2021).
[188] Y. Shen, B. Sheng, R. Fang, H. Li, L. Dai, S. Stolte, J. Qin, W. Jia, D. Shen, Domain-invariant interpretable fundus image quality assessment, Med. Image Anal. 61 (2020).
[189] R. Sayres, A. Taly, E. Rahimy, K. Blumer, D. Coz, N. Hammel, J. Krause, A. Narayanaswamy, Z. Rastegar, D. Wu, S. Xu, Using a deep learning algorithm and integrated gradients explanation to assist grading for diabetic retinopathy, Ophthalmology 126 (4) (2019) 552–564.
[190] T. Araújo, G. Aresta, L. Mendonça, S. Penas, C. Maia, Â. Carneiro, A. M. Mendonça, A. Campilho, DR|GRADUATE: uncertainty-aware deep learning-based diabetic retinopathy grading in eye fundus images, Med. Image Anal. 63 (2020).
[191] G. Quellec, H.A. Hajj, M. Lamard, P.-H. Conze, P. Massin, B. Cochener, ExplAIn: explanatory artificial intelligence for diabetic retinopathy diagnosis, Med. Image Anal. 72 (2021).
[192] Y. Niu, L. Gu, Y. Zhao, F. Lu, Explainable diabetic retinopathy detection and retinal image generation, IEEE journal of biomedical and health informatics (2021) 1, 1.
[193] J. Martins, J.S. Cardoso, F. Soares, Offline computer-aided diagnosis for Glaucoma detection using fundus images targeted at mobile devices, Comput. Methods Progr. Biomed. 192 (2020).
[194] Y. Jang, J. Son, K.H. Park, S.J. Park, K.H. Jung, Laterality classification of fundus images using interpretable deep neural network, J. Digit. Imag. 31 (6) (2018) 923–928.
[195] M. Kim, J.C. Han, S.H. Hyun, O. Janssens, S. Van Hoecke, C. Kee, W. De Neve, Medinoid: computer-aided diagnosis and localization of glaucoma using deep learning, Appl. Sci. 9 (15) (2019).
[196] W. Liao, B. Zou, R. Zhao, Y. Chen, Z. He, M. Zhou, Clinical interpretable deep learning model for glaucoma diagnosis, IEEE journal of biomedical and health informatics 24 (5) (2019) 1405–1412.
[197] Q. Meng, Y. Hashimoto, S.I. Satoh, How to extract more information with less burden: fundus image classification and retinal disease localization with ophthalmologist intervention, IEEE J. Biomed. Health Informat. 24 (12) (2020) 3351–3361.
[198] A. Marginean, A. Groza, S. Nicoara, G. Muntean, R.R Slavescu, I. Letia, Towards balancing the complexity of convolutional neural network with the role of optical coherence tomography in retinal conditions, in: 15th International Conference on Intelligent Computer Communication and Processing, ICCP), 2019.
[199] O. Perdomo, H. Rios, F.J. Rodríguez, S. Otálora, F. Meriaudeau, H. Müller, F. A. González, Classification of Diabetes-Related Retinal Diseases Using a Deep Learning Approach in Optical Coherence Tomography," *Computer Methods and Programs in Biomedicine*, 2019, pp. 181–189.
[200] M. El Adoui, S. Drisis, M. Benjelloun, Multi-input deep learning architecture for predicting breast tumor response to chemotherapy using quantitative MR images, Int. J. Comput. Assist. Radiol. Surg. 15 (2020) 1491–1500.
[201] L. Luo, H. Chen, X. Wang, Q. Dou, H. Lin, J. Zhou, G. Li, P. Heng, Deep angular embedding and feature correlation attention for breast MRI cancer analysis, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, 2019.
[202] B. van der Velden, M. Janse, M. Ragusi, C. Loo, K. Gilhuijs, Volumetric breast density estimation on MRI using explainable deep learning regression, Sci. Rep. 10 (1) (2020) 1–9.
[203] H. Rasaee, H. Rivaz, Explainable AI and susceptibility to adversarial attacks: a case study in classification of breast ultrasound images, in: IEEE International Ultrasonics Symposium, IUS), 2021.
[204] X. Qian, J. Pei, H. Zheng, X. Xie, L. Yan, H. Zhang, C. Han, X. Gao, H. Zhang, W. Zheng, Q. Sun, L. Lu, K.K. Shung, Prospective assessment of breast cancer risk from multimodal multiview ultrasound images via clinically applicable deep learning, Nat. Biomed. Eng. 5 (6) (2021) 522–532.
[205] J.B. Lamy, B. Sekar, G. Guezennec, J. Bouaud, B. Séroussi, Explainable artificial intelligence for breast cancer: a visual case-based reasoning approach, Artif. Intell. Med. 94 (2019) 42–53.
[206] S.T. Kim, H. Lee, H.G. Kim, a.Y.M. Ro, ICADx: interpretable computer aided diagnosis of breast masses, Medical Imaging 2018: Computer-Aided Diagnosis 10575 (2018) 450–459.
[207] H. Yang, J. Kim, H. Kim, S. Adhikari, Guided soft attention network for classification of breast cancer histopathology images, IEEE Trans. Med. Imag. 39 (5) (2019) 1306–1315.
[208] M. Graziani, V. Andrearczyk, H. Müller, Regression concept vectors for bidirectional explanations in histopathology, in: Understanding and Interpreting Machine Learning in Medical Image Computing Applications, 2018.
[209] B. Gecer, S. Aksoy, E. Mercan, L.G. Shapiro, D.L. Weaver, J.G. Elmore, Detection and classification of cancer in whole slide breast histopathology images using deep convolutional networks, Pattern Recogn. 84 (2018) 345–356.
[210] H. Chereda, A. Bleckmann, K. Menck, J. Perera-Bel, P. Stegmaier, F. Auer, F. Kramer, A. Leha, T. Beißbarth, Explaining decisions of graph convolutional neural networks: patientspecific molecular subnetworks responsible for metastasis prediction in breast cance, Genome Med. 13 (1) (2021).
[211] R. Gu, G. Wang, T. Song, R. Huang, M. Aertsen, J. Deprest, S. Ourselin, T. Vercauteren, S. Zhang, CA-net: comprehensive attention convolutional neural networks for explainable medical image segmentation, IEEE Trans. Med. Imag. 40 (2) (2021) 699–711.
[212] F. Stieler, F. Rabe, B. Bauer, Towards domain-specific explainable AI:model interpretation of a skin image classifier using a human approach, in: IEEE/CVF

Conference on Computer Vision and Pattern Recognition Workshops, CVPRW), 2021.
[213] W. Li, J. Zhuang, R. Wang, J. Zhang, W.S. Zheng, Fusing metadata and dermoscopy images for skin disease diagnosis, in: IEEE 17th International Symposium on Biomedical Imaging, ISBI), 2020.
[214] Y. Xie, J. Zhang, Y. Xia, C. Shen, A mutual bootstrapping model for automated skin lesion segmentation and classification, IEEE Trans. Med. Imag. 39 (7) (2020) 2482–2493.
[215] P.V. Molle, M.D. Strooper, T. Verbelen, B. Vankeirsbilck, P. Simoens, B. Dhoedt, Visualizing convolutional neural networks to improve decision support for skin lesion classification, in: Understanding and Interpreting Machine Learning in Medical Image Computing Applications, 2018.
[216] V. Narayanaswamy, J.J. Thiagarajan, A. Spanias, Using deep image priors to generate counterfactual explanations, in: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP), 2021.
[217] C. Barata, M.E. Celebi, J.S. Marques, Explainable skin lesion diagnosis using taxonomies, Pattern Recogn. 110 (2021).
[218] A. Lucieri, M. Bajwa, S. Braun, M. Malik, A. Dengel, S. Ahmed, ExAID: A Multimodal Explanation Framework for Computer-Aided Diagnosis of Skin Lesions, Computer Methods and Programs in Biomedicine, 2022.
[219] G. Kunapuli, B.A. Varghese, P. Ganapathy, B. Desai, S. Cen, M. Aron, I. Gill, V. Duddalwar, A decision-support tool for renal mass classification," journal of digital imaging, J. Digit. Imag. 31 (6) (2018) 929–939.
[220] M.R. Hassan, M.F. Islam, M.Z. Uddin, G. Ghoshal, M.M. Hassan, S. Huda, G. Fortino, Prostate cancer classification from ultrasound and MRI images using deep learning based Explainable Artificial Intelligence, Future Generat. Comput. Syst. 127 (2022) 462–472.
[221] K. Uehara, M. Murakawa, H. Nosato, H. Sakanashi, Multi-scale exaplainable feature learning for pathological image analysis using convolutional neural networks, in: IEEE International Conference on Image Processing, ICIP), 2020.
[222] R. Ishii, X. Zhang, N. Homma, An interpretable DL-based method for diagnosis of H. Pylori infection using gastric X-ray images, in: IEEE 3rd Global Conference on Life Sciences and Technologies (LifeTech 2021), 2021.
[223] J. Tian, C. Li, Z. Shi and F. Xu, "A diagnostic report generator from CT volumes on liver tumor with semi-supervised attention mechanism," in International Conference on Medical Image Computing and Computer-Assisted Intervention.
[224] V. Couteaux, O. Nempont, G. Pizaine, I. Bloch, Towards interpretability of segmentation networks by analyzing deepdreams, in: Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support, 2019.
[225] X. Chen, L. Lin, D. Liang, H. Hu, Q. Zhang, Y. Iwamoto, X.H. Han, Y.W. Chen, R. Tong, J. Wu, A dual-attention dilated residual network for liver lesion classification and localization on CT images, in: IEEE International Conference on Image Processing, ICIP), 2019.
[226] N. Shapira, J. Fokuhl, M. Schultheiß, S. Beck, F.K. Kopp, D. Pfeiffer, J. Dangelmaier, G. Pahn, A.P. Sauter, B. Renger, A.A. Fingerle, Liver lesion localisation and classification with convolutional neural networks: a comparison between conventional and spectral computed tomography, Biomed. Phys. Eng. Express 6 (1) (2020).
[227] P. Rajpurkar, A. Park, J. Irvin, C. Chute, M. Bereket, D. Mastrodicasa, C. P. Langlotz, M.P. Lungren, A.Y. Ng, B.N. Patel, AppendiXNet: deep learning for diagnosis of appendicitis from a small dataset of CT exams using video pretraining, Sci. Rep. 10 (1) (2020) 1–7.
[228] C.J. Wang, C.A. Hamm, L.J. Savic, M. Ferrante, I. Schobert, T. Schlachter, M. Lin, J.C. Weinreb, J.S. Duncan, J. Chapiro, B. Letzen, Deep learning for liver tumor diagnosis part II: convolutional neural network interpretation using radiologic imaging features, Eur. Radiol. 29 (7) (2019) 3348–3357.
[229] L.A.d.S. Jr, R. Mendel, S. Strasser, A. Ebigbo, A. Probst, H. Messmann, J.P. Papa, C. Palm, Convolutional Neural Networks for the evaluation of cancer in Barrett's esophagus: explainable AI to lighten up the black-box, Comput. Biol. Med. 135 (2021).
[230] S. Knapič, A. Malhi, R. Saluja, K. Främling, Explainable artificial intelligence for human decision-support system in medical domain, Machine Learning and Knowledge Extraction 3 (3) (2021) 740–770.
[231] S. Wang, Y. Xing, L. Zhang, H. Gao, H. Zhang, Deep Convolutional Neural Network for Ulcer Recognition in Wireless Capsule Endoscopy: Experimental Feasibility and Optimization," *Computational And Mathematical Methods in Medicine*, 2019.
[232] P. Sabol, P. Sinčák, P. Hartono, P. Kočan, Z. Benetinová, A. Blichárová, Ľ. Verbóová, E. Štammovác, A. Sabolová-Fabianová, A. Jašková, Explainable classifier for improving the accountability in decision-making for colorectal cancer diagnosis from histopathological images, J. Biomed. Inf. 109 (2020).
[233] A. Kiani, B. Uyumazturk, P. Rajpurkar, A. Wang, R. Gao, E. Jones, Y. Yu, C. P. Langlotz, R.L. Ball, T.J. Montine, B.A. Martin, Impact of a deep learning assistant on the histopathologic classification of liver cancer, NPJ Digital Med. 3 (1) (2020) 1–8.
[234] F. Heinemann, G. Birk, B. Stierstorfer, Deep learning enables pathologist-like scoring of NASH models, Sci. Rep. 9 (1) (2019).
[235] P.H. Yi, T.K. Kim, J. Wei, J. Shin, F.K. Hui, H.I. Sair, G.D. Hager, J. Fritz, Automated semantic labeling of pediatric musculoskeletal radiographs using deep learning, Pediatr. Radiol. 49 (8) (2019) 1066–1070.
[236] H. Yoo, S. Han, K. Chung, Diagnosis support model of cardiomegaly based on CNN using ResNet and explainable feature map, IEEE Access 9 (2021) 55802–55813.
[237] J. Seah, J. Tang, A. Kitchen, F. Gaillard, A. Dixon, Chest radiographs in congestive heart failure: visualizing neural network learning, Radiology 290 (2) (2019) 514–522.
[238] C. Cong, Y. Kato, H.D. Vasconcellos, J. Lima, B. Venkatesh, Automated stenosis detection and classification in x-ray angiography using deep neural network, in: IEEE International Conference on Bioinformatics and Biomedicine, BIBM), 2019.
[239] S. Candemir, R.D. White, M. Demirer, V. Gupta, M.T. Bigelow, L.M. Prevedello, B. S. Erdal, Automated coronary artery atherosclerosis detection and weakly supervised localization on coronary CT angiography with a deep 3-dimensional convolutional neural network, Comput. Med. Imag. Graph. 83 (2020).
[240] Y. Huo, J.G. Terry, J. Wang, V. Nath, C. Bermudez, S. Bao, P. Parvathaneni, J. J. Carr, B.A. Landman, Coronary calcium detection using 3D attention identical dual deep network based on weakly supervised learning, Med. Imaging Image Process. 10949 (2019) 308–315.
[241] A. Janik, J. Dodd, G. Ifrim, K. Sankaran, K. Curran, Interpretability of a deep learning model in the application of cardiac MRI segmentation with an ACDC challenge dataset, Med. Imag. 2021: Image Process. 11596 (2021) 861–872.
[242] R. Ceschin, A. Zahner, W. Reynolds, J. Gaesser, G. Zuccoli, C.W. Lo, V. Gopalakrishnan, A. Panigrahy, A computational framework for the detection of subcortical brain dysmaturation in neonatal MRI using 3D Convolutional Neural Networks, Neuroimage 178 (2018) 183–197.
[243] M. Komatsu, A. Sakai, R. Komatsu, R. Matsuoka, S. Yasutomi, K. Shozu, A. Dozen, H. Machino, H. Hidaka, T. Arakaki, K. Asada, S. Kaneko, A. Sekizawa, R. Hamamoto, Detection of cardiac structural abnormalities in fetal ultrasound videos using deep learning, Appl. Sci. 11 (1) (2021).
[244] I. Neves, D. Folgado, S. Santos, M. Barandas, A. Campagner, L. Ronzio, F. Cabitza, H. Gamboa, Interpretable heartbeat classification using local model-agnostic explanations on ECGs, Comput. Biol. Med. 133 (2021).
[245] A. Ghorbani, D. Ouyang, A. Abid, B. He, J.H. Chen, R.A. Harrington, D.H. Liang, E.A. Ashley, J.Y. Zou, Deep learning interpretation of echocardiograms, NPJ Digital Med. 3 (1) (2020) 1–10.
[246] M. Vafaeezadeh, H. Behnam, A. Hosseinsabet, P. Gifani, Automatic morphological classification of mitral valve diseases in echocardiographic images based on explainable deep learning methods, Int. J. Comput. Assist. Radiol. Surg. 17 (2) (2022) 413–425.
[247] J. Yoo, T.J. Jun, Y.-H. Kim, xECGNet: fine-tuning attention map within convolutional neural network to improve detection and explainability of concurrent cardiac arrhythmias, Comput. Methods Progr. Biomed. 208 (2021), 106281.
[248] L. Ibrahim, M. Mesinovic, K.-W. Yang, M.A. Eid, Explainable prediction of acute myocardial infarction using machine learning and Shapley values, IEEE Access 8 (2020) 210410–210417.
[249] J. Zhang, C. Petitjean, S. Ainouz, Segmentation-based vs. Regression-based biomarker estimation: a case study of fetus head circumference assessment from ultrasound images, Journal of Imaging 8 (2) (2022).
[250] F. Viton, M. Elbattah, J.-L. Guérin, G. Dequen, Heatmaps for visual explainability of CNN-based predictions for multivariate time series with application to healthcare, in: International Conference on Healthcare Informatics, ICHI), 2020.
[251] L. Brunese, F. Mercaldo, A. Reginelli, A. Santone, Explainable deep learning for pulmonary disease and coronavirus COVID-19 detection from X-rays, Comput. Methods Progr. Biomed. 196 (2020), 105608.
[252] A. Chowdhury, A. Santamaria-Pang, J.R. Kubricht, P. Tu, Emergent Symbolic language based deep medical image classification, in: IEEE 18th International Symposium on Biomedical Imaging, ISBI), 2021.
[253] M. Komatsu, A. Sakai, A. Dozen, K. Shozu, S. Yasutomi, H. Machino, K. Asada, S. Kaneko, R. Hamamoto, Towards clinical application of artificial intelligence in ultrasound imaging, Biomedicines 9 (7) (2021) 720.
[254] E.P.V. Le, Y. Wang, Y. Huang, S. Hickman, F.J. Gilbert, Artificial intelligence in breast imaging, Clin. Radiol. 74 (5) (2019) 357–366.
[255] I.P.d. Sousa, M.M.B.R. Vellasco, E.C.d. Silva, Local interpretable model-agnostic explanations for classification of lymph node metastases, Sensors 19 (2019).
[256] P. Dwivedi, A.A. Khan, S. Mugde, G. Sharma, Diagnosing the major contributing factors in the classification of the fetal health status using cardiotocography measurements: an AutoML and XAI approach, in: 13th International Conference on Electronics, Computers and Artificial Intelligence (ECAI), 2021.
[257] E. Khodabandehloo, D. Riboni, A. Alimohammadi, HealthXAI: collaborative and explainable AI for supporting early diagnosis of cognitive decline, Future Generat. Comput. Syst. 116 (2021) 168–189.
[258] M.A. Gulum, C.M. Trombley, M. Kantardzic, A review of explainable deep learning cancer detection models in medical imaging, Appl. Sci. 11 (10) (2021) 4573.
[259] P. Schramowski, W. Stammer, S. Teso, A. Brugger, F. Herbert, X. Shao, H.-G. Luigs, A.-K. Mahlein, K. Kersting, Making deep neural networks right for the right scientific reasons by interacting with their explanations, Nat. Mach. Intell. 2 (8) (2020) 476–486.
[260] S.N. Payrovnaziri, Z. Chen, P. Rengifo-Moreno, T. Miller, J. Bian, J.H. Chen, X. Liu, a.Z. He, Explainable artificial intelligence models using real-world electronic health record data: a systematic scoping review, J. Am. Med. Inf. Assoc. 27 (7) (2020) 1173–1185.
[261] C. Gillmann, N.N. Smit, E. Gröller, B. Preim, A. Vilanova, T. Wischgoll, Ten open challenges in medical visualization, Comput. Graphics Appl. 41 (5) (2021) 7–15.
[262] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, ACM Comput. Surv. 51 (5) (2018) 1–42.
[264] P. Hall, N. Gill, An Introduction to Machine Learning Interpretability, O'Reilly Media, 2018.
[265] M.-H. Tayarani-N, Applications of Artificial Intelligence in Battling against Covid-19: A Literature Review," *Chaos*, Solitons & Fractals, 2020.

[266] J. Yim, K. Sohn, Enhancing the performance of convolutional neural networks on quality degraded datasets, in: International Conference on Digital Image Computing: Techniques and Applications, DICTA), 2017.
[267] M. Heidari, S. Mirniaharikandehei, A.Z. Khuzani, G. Danala, Y. Qiu, B. Zheng, Improving the performance of CNN to predict the likelihood of COVID-19 using chest X-ray images with preprocessing algorithms, Int. J. Med. Inf. 144 (2020), 104284.
[268] A. Paka, K. Gade, D. Farah, Model Performance Management with Explainable AI, O'Reilly Media, Inc., 2021.
[269] E. Thibeau-Sutre, O. Colliot, D. Dormont, N. Burgos, Visualization approach to assess the robustness of neural networks for medical image classification, in: Medical Imaging 2020: Image Processing, 2020.
[270] R.I. Etheredge, M. Schartl, A. Jordan, Decontextualized learning for interpretable hierarchical representations of visual patterns, Patterns 2 (2021).
[272] A. Holzinger, From machine learning to explainable AI, in: 2018 World Symposium on Digital Intelligence for Systems and Machines, DISA, 2018.
[273] N. Hampe, J.M. Wolterink, S.G. M.v. Velzen, T. Leiner, I. Išgum, Machine Learning for Assessment of Coronary Artery Disease in Cardiac CT: A Survey," *Frontiers In Cardiovascular Medicine*, vol. 6, 2019.
[274] Proposed Regulatory Framework for Modifications to Artificial Intelligence/ Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) -Discussion Paper and Request for Feedback".
[275] [Online]. Available: AI Fairness 360 (AIF360) https://github.com/Trusted-AI/AIF360. (Accessed 15 September 2022).
[276] Using the what-if tool, Google, [Online]. Available: https://cloud.google.com/ai-platform/prediction/docs/using-what-if-tool. (Accessed 15 September 2022).
[277] A.-K. Dombrowski, C.J. Anders, K.-R. Müller, P. Kessel, Towards robust explanations for deep neural networks, Pattern Recogn. 121 (2022).
[278] T.-T.-H. Le, H. Kang, H. Kim, Robust adversarial attack against explainable deep classification models based on adversarial images with different Patch sizes and perturbation ratios, IEEE Access 9 (2021) 133049–133061.
[279] A. Mikołajczyk, M. Grochowski, A. Kwasigroch, Towards Explainable Classifiers Using the Counterfactual Approach – Global Explanations for Discovering Bias in Data, 2020, https://doi.org/10.48550/arXiv.2005.02269 [Online]. Available:. (Accessed 15 September 2022).
[280] R.C. Fong, A. Vedaldi, Interpretable explanations of black boxes by meaningful perturbation, in: IEEE International Conference on Computer Vision, ICCV), 2017.
[281] A.S. Mursch-Edlmayr, W.S. Ng, A. Diniz-Filho, D.C. Sousa, L. Arnould, M. B. Schlenker, K. Duenas-Angeles, P.A. Keane, J.G. Crowston, H. Jayaram, Artificial intelligence algorithms to diagnose glaucoma and detect glaucoma progression: translation to clinical practice, Translat. Vision Sci. Technol. 9 (2) (2020) 55.
[282] H. Wang, L. Wang, E.H. Lee, J. Zheng, W. Zhang, S. Halabi, C. Liu, K. Deng, J. Song, K.W. Yeom, Decoding COVID-19 pneumonia: comparison of deep learning and radiomics CT image signatures, Eur. J. Nucl. Med. Mol. Imag. 48 (5) (2021) 1697, 1697.
[283] C. Gilvary, N. Madhukar, J. Elkhader, O. Elemento, The missing pieces of artificial intelligence in medicine, Trends Pharmacol. Sci. 40 (8) (2019) 555–564.
[284] T. Miller, P. Howe, L. Sonenberg, Explainable AI: Beware of Inmates Running the Asylum or: How I Learnt to Stop Worrying and Love the Social and Behavioural Sciences, 2017, https://doi.org/10.48550/arXiv.1712.00547 [Online]. Available: . (Accessed 15 September 2022).
[285] S. Larsson, F. Heintz, Transparency in artificial intelligence, Internet Policy Review 9 (2) (2020).
[286] T. Martín-Noguerol, F. Paulano-Godino, R. López-Ortega, J.M. Górriz, R. F. Riascos, A. Luna, Artificial intelligence in radiology: relevance of collaborative work between radiologists and engineers for building a multidisciplinary team, Clin. Radiol. 76 (5) (2020) 317–324.
[287] M.S. Hossain, G. Muhammad, N. Guizani, Explainable AI and mass surveillance system-based healthcare framework to combat COVID-I9 like pandemics, IEEE Network 34 (4) (2020) 126–132.
[288] G. Joshi, R. Walambe, K. Kotecha, A review on explainability in multimodal deep neural nets, IEEE Access 9 (2021) 59800–59821.
[289] H. Yuan, J. Tang, X. Hu, S. Ji, XGNN: towards model-level explanations of graph neural networks, in: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, CA, USA, 2020.
[290] S.A. Bargal, A. Zunino, V. Petsiuk, J. Zhang, K. Saenko, V. Murino, S. Sclaroff, Guided Zoom: zooming into network evidence to refine fine-grained model decisions, IEEE Trans. Pattern Anal. Mach. Intell. 43 (11) (2021) 4196–4202.
[291] D. Bau, B. Zhou, A. Khosla, A. Oliva, A. Torralba, Network dissection: quantifying interpretability of deep visual representations, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017.