Medical Image Analysis 79 (2022) 102470

Contents lists available at ScienceDirect

Medical Image Analysis

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

Explainable artificial intelligence (XAI) in deep learning-based medical image analysis

Bas H.M. van der Velden*, Hugo J. Kuijf, Kenneth G.A. Gilhuijs, Max A. Viergever

Image Sciences Institute, University Medical Center Utrecht, Q.02.4.45, P.O. Box 85500, Utrecht, GA 3508, the Netherlands

ARTICLE INFO

Article history:
Received 22 July 2021
Revised 15 March 2022
Accepted 2 May 2022
Available online 4 May 2022

Keywords:
Explainable artificial intelligence
Interpretable deep learning
Medical image analysis
Deep learning
Survey

ABSTRACT

With an increase in deep learning-based methods, the call for explainability of such methods grows, especially in high-stakes decision making areas such as medical image analysis. This survey presents an overview of explainable artificial intelligence (XAI) used in deep learning-based medical image analysis. A framework of XAI criteria is introduced to classify deep learning-based medical image analysis methods. Papers on XAI techniques in medical image analysis are then surveyed and categorized according to the framework and according to anatomical location. The paper concludes with an outlook of future opportunities for XAI in medical image analysis.

1. Introduction

Deep learning has invoked tremendous progress in automated image analysis. Before that, image analysis was commonly performed using systems fully designed by human domain experts. For example, such image analysis system could consist of a statistical classifier that used handcrafted properties of an image (i.e., features) to perform a certain task. Features included low-level image properties such as edges or corners, but also higher-level image properties such as the spiculated border of a cancer. In deep learning, these features are learned by a neural network (in contrast to being handcrafted) to optimally give a result (or output) given an input. An example of a deep learning system could be the output 'cancer' given the input of an image showing a cancer.

Neural networks typically consist of many layers connected via many nonlinear intertwined relations. Even if one is to inspect all these layers and describe their relations, it is unfeasibly to fully comprehend how the neural network came to its decision. Therefore, deep learning is often considered a 'black box'. Concern is mounting in various fields of application that these black boxes may be biased in some way, and that such bias goes unnoticed. Especially in medical applications, this can have far-reaching consequences.

There has been a call for approaches to better understand the black box. Such approaches are commonly referred to as interpretable deep learning or explainable artificial intelligence (XAI) (Adadi and Berrada, 2018; Murdoch et al., 2019). These terms are commonly interchanged; we will use the term XAI. Some notable XAI initiatives include those from the United States Defense Advanced Research Projects Agency (DARPA), and the conferences on Fairness, Accountability, and Transparency by the Association for Computing Machinery (ACM FAccT).

The stakes of medical decision making are often high. Not surprisingly, medical experts have voiced their concern about the black box nature of deep learning (Jia et al., 2020), which is the current state of the art in medical image analysis (Litjens et al., 2017; Meijering, 2020; Shen et al., 2017). Furthermore, regulations such as the European Union's General Data Protection Regulation (GDPR, Article 15) require the right of patients to receive meaningful information about how a decision was rendered.

Researchers in medical imaging are increasingly using XAI to explain the results of their algorithms. Something can be considered a good explanation if it gives insight into how a neural network came to its decision and/or can make the decision understandable. In this survey, we aim to give a comprehensive overview of papers using XAI in medical image analysis. We chose to focus solely on papers that used deep learning-based XAI in medical image analysis.

The search strategy for inclusion of papers was as follows: We used the search query "(explainable deep learning OR interpretable deep learning OR XAI OR interpretable machine learning OR explainable machine learning) AND (medical imaging OR medical image analysis)" in SCOPUS. We included papers from peer reviewed journals and conferences. We analyzed the query results

* Corresponding author.
E-mail address: bvelden2@umcutrecht.nl (B.H.M. van der Velden).

https://doi.org/10.1016/j.media.2022.102470

using the Active learning for Systematic Reviews toolbox (van de Schoot et al., 2021). This toolbox uses active learning to sort papers from most relevant to least relevant, while being updated by user input. Furthermore, we had discussions with colleagues, and used a snowballing approach – investigating papers referenced by the included papers and papers that refer to the included papers. We read the title and the abstract of each of these papers, and browsed paper content if we were not sure whether to include the paper. In case of multiple publications by the same authors on the same subject, we chose the journal publication or the most recent publication in case of multiple conference publications. Papers up to October 2020 are included in the survey.

The survey is structured as follows: We will first introduce the taxonomy of XAI and describe a framework to classify XAI techniques in Section 2. In Section 3, the discussed papers are characterized according to this XAI framework. We will discuss applications of XAI techniques in medical image analysis. In case of multiple papers using the same technique, we will discuss some early adopters and summarize the rest of the papers in the tables. Since XAI techniques often originate from computer vision, we will elaborate on papers that adapted XAI techniques from computer vision by adding domain knowledge from the medical imaging field. The papers are grouped in the tables according to explanation method and according to anatomical location. This survey adds to the review of Reyes et al. (2020); since they mainly discussed techniques in computer vision, without extensively evaluating the adaptation of such techniques throughout medical image analysis. Furthermore, we describe if and how techniques from computer vision have been adapted specifically for medical image analysis. This survey adds to the review of Huff et al. (2021), since they mostly focused on examples of visual explanation, while our survey aims for a more holistic approach including non-visual explanation, critiques on XAI, and methods for evaluating XAI. Additionally, we systematically survey papers, reflecting the current status of the field of XAI in medical imaging. In Section 4, we discuss the pros and cons of the discussed XAI techniques. The survey is concluded in Section 5 by discussing the state of the art of XAI in medical image analysis and an outlook of the opportunities of XAI.

2. Explainable artificial intelligence (XAI) framework

In this section, we will give a brief overview of Explainable Artificial Intelligence (XAI) techniques found in deep learning for medical image analysis. For exhaustive surveys focused solely on XAI, please refer to Adadi and Berrada (2018) and Murdoch et al. (2019).

We will distinguish XAI techniques based on three criteria: model-based versus post hoc, model-specific versus model-agnostic, and global versus local (i.e., the scope of the explanation). The framework of these three criteria is adapted from the surveys of Adadi and Berrada (2018) and Murdoch et al. (2019) and is depicted in Fig. 1. The following paragraphs will describe these criteria.

2.1. Model-based versus post hoc explanation

The first distinction we make is model-based explanation versus post hoc explanation (Fig. 1).

2.1.1. Model-based explanation

Model-based explanation refers to models, e.g. a linear regression model or a support vector machine, that are simple enough to be understood, but sophisticated enough to fit a relationship between input and output well (Murdoch et al., 2019). These are often the traditional machine learning models. Examples of model-based explanation enforce the use of a limited amount of features (i.e., sparsity), or enforce a human to be able to internally reason about the model's entire decision-making process (i.e., simulatability) (Murdoch et al., 2019). For example, models that enforce sparsity such as the least absolute shrinkage and selection operator (LASSO, Tibshirani (1996), force many coefficients to zero. Hence, a select subset of features leads to an output, making the inner construct of this model explainable.

Since the focus of our survey is on XAI methods for deep learning, model-based explanation by enforcing sparsity or simulatability is infeasible. Deep learning uses a deep neural network, typically with thousands to millions of weights, which is neither sparse, nor suited for a human to internally simulate and reason about the models entire decision making. However, one of the methods mentioned by Murdoch et al. (2019) was model-based feature engineering, i.e., automated approaches for constructing explainable features.

2.1.2. Post hoc explanation

Analyzing a trained model (i.e., a neural network in deep learning) to achieve insight into learned relationships is referred to as post hoc explanation. An important distinction between post hoc explanation and model-based explanation is that the former trains a neural network and subsequently attempts to explain the behavior of the ensuing black box network, whereas the latter forces the neural network to be explainable.

Methods that provide post hoc explanation include inspection of learned features, feature importance, and interaction of features (Abbasi-Asl and Yu, 2017; Olden et al., 2004; Tsang et al. 2018; as well as visual explanation by saliency maps (Selvaraju et al., 2017; Simonyan et al., 2013; Springenberg et al., 2014; Zeiler and Fergus, 2014; Zhou et al., 2016).

2.2. Model-specific versus model-agnostic explanation

The distinction between model-specific and model-agnostic explanation is related to that between model-based and post hoc explanation (Adadi and Berrada, 2018), but there are some nuanced differences.

2.2.1. Model-specific explanation

Model-specific explanation methods are limited to particular classes of models. For example, such a method may use attributes that are specific to a type of neural network. A drawback is that by aiming at model-specific explanation, we limit our choice of neural networks, thereby potentially excluding a neural network that could better fit the output to the input data.

Model-based explanation is by definition model-specific (Adadi and Berrada, 2018), but model-specific explanation is not necessary model-based. Some post hoc saliency mapping techniques are examples of techniques that are specific to a certain class of convolutional neural networks (CNNs), but are not model-based explanation methods (Murdoch et al., 2019).

2.2.2. Model-agnostic explanation

Model-agnostic explanation is independent of the choice of the type of neural network, operating solely on the input and the output of the neural network. By perturbing the input, the user can inspect what the change is in the output of the neural network. This can therefore explain which regions are driving the output. Model-agnostic explanation is naturally post hoc.

2.3. Scope of explanation

The scope of an explanation distinguishes between explanation for an entire model (global) versus explanation for a single output (local).

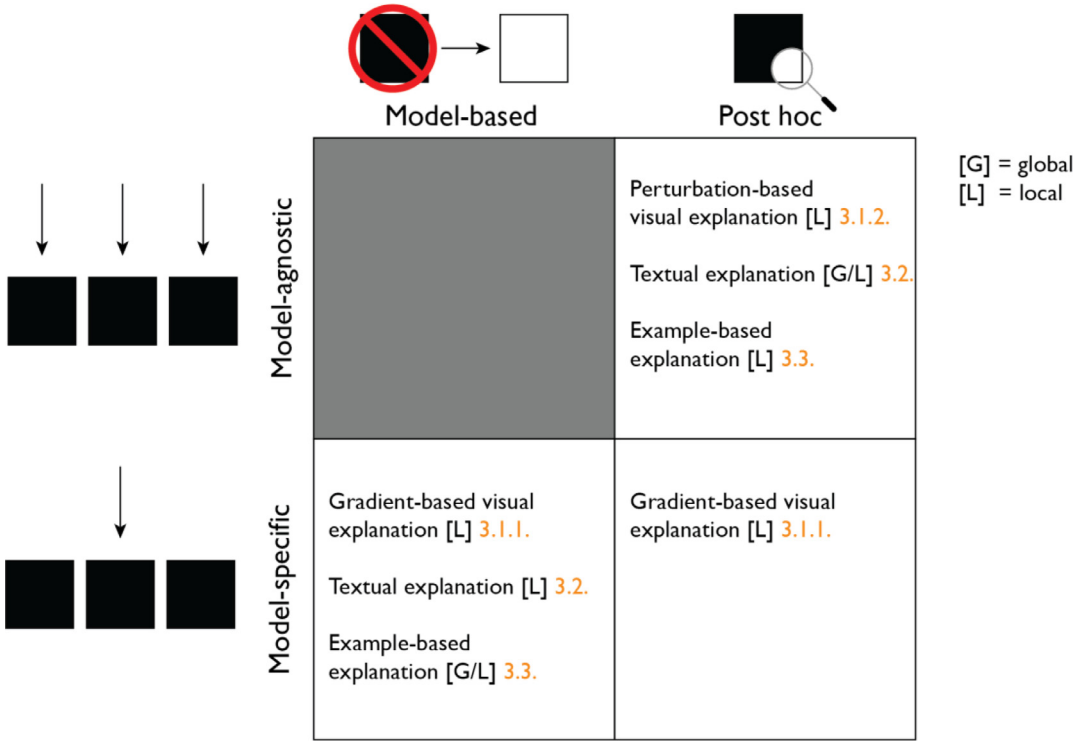

Fig. 1. The eXplainable Artificial Intelligence (XAI) framework proposed in this paper. A rough overview of XAI techniques (discussed in Section 3) is classified according to this framework. The orange number refers to the section number in the manuscript where the XAI technique is described.

2.3.1. Global explanation

Global explanation, also called dataset-level explanation, provides general relationships learned by the neural network. For example, global explanation could provide feature importance scores at the dataset level, i.e., how much do features contribute to the output across the entire dataset (Olden et al., 2004). As an illustration, one might observe from a neural network that – or even how much – high blood pressure increases the risk of a cardiac event. Another example of global explanation could be visualization of learned filters, i.e., which features are extracted by the neural network and to what extent are they meaningful to the task at hand (Olah et al., 2017; Zeiler and Fergus, 2014).

2.3.2. Local explanation

Local explanation provides explanation of a single input. In the example of cardiac risk, an input would be a single person. Local explanation would therefore explain why blood pressure is important to the risk of cardiac event for that single person, whereas global explanation would describe the relation of blood pressure with risk of cardiac events across the entire dataset. Another example of a local explanation could be a saliency map pinpointing to a brain tumor on magnetic resonance imaging (MRI) to explain which part of the MRI mainly contributed to the classifier output 'tumor'. Since this explains which part of the image drives the classifier to its output 'tumor' for that single person, this is a local explanation.

3. XAI in medical image analysis

In this section, we will present which XAI techniques are used in medical image analysis, and we will discuss adaptations of the methods typically seen in computer vision. We categorize the explanation methods into three types: visual, textual, and example-based; and we will classify each method according to the framework of model-based versus post hoc, model-specific versus model-agnostic, and global versus local explanation (Fig. 1). Table 1 provides an overview of the most frequently used techniques and shows their connections according to the taxonomy defined in Section 2.

3.1. Visual explanation

Visual explanation, also called saliency mapping, is the most common form of XAI in medical image analysis (Fig. 2). Saliency maps show the important parts of an image for a decision. Most saliency mapping techniques use backpropagation-based approaches, but some use perturbation-based or multiple instance learning-based approaches. These approaches will be discussed below. An overview of papers using saliency maps in medical imaging is shown in Table 2.

3.1.1. Backpropagation-based approaches

(Guided) backpropagation and deconvolution: Some of the earliest techniques to create saliency maps highlighted pixels that had the highest impact on the analysis output. Examples included visualization of partial derivatives of the output on pixel level (Simonyan et al., 2013), deconvolution (Zeiler and Fergus, 2014), and guided backpropagation (Springenberg et al., 2014). These techniques provided local, model-specific (only for CNNs), post hoc explanation. These techniques have been used in medical image analysis. For example, de Vos et al. (2019) estimated the amount of coronary artery calcium per cardiac or chest computed tomography (CT) image slice, and used deconvolution to visualize from where in the slice the decision was based on.

Class activation mapping (CAM): Zhou et al. (2016) introduced Class Activation Mapping (CAM). They replaced the fully connected layers at the end of a CNN by global average pooling on the last convolutional feature maps. The class activation map was a weighted linear sum of presence of visual patterns (captured by the filters) at different spatial locations. This technique

Table 1
Overview of eXplainable AI (XAI) techniques used in medical image analysis, classified by the framework from Section 2.

Technique	Section	Authors	Model- based	Post hoc	Model-specific	Model-agnostic	Global	Local
Visual explanation	3.1.							
Backpropagation-based approaches	3.1.1							
Backpropagation	3.1.1.1.	Simonyan et al. (2013)		√	√			√
Deconvolution	3.1.1.1.	Zeiler and Fergus (2014)		√	√			√
Guided backpropagation	3.1.1.1.	Springenberg et al. (2014)		√	√			√
Class activation mapping (CAM)	3.1.1.2.	Zhou et al. (2016)		√	√			√
Gradient-weighted class activation mapping (Grad-CAM)	3.1.1.3.	Selvaraju et al. (2017)		√	√			√
Layer-wise relevance propagation (LRP)	3.1.1.4.	Bach et al. (2015)		√	√			√
Deep SHapley Additive exPlanations (Deep SHAP)	3.1.1.5.	Lundberg and Lee (2017)		√	√	√*	√*	√
Trainable attention	3.1.1.6.	Jetley et al. (2018)	√		√			√
Perturbation-based approaches	3.1.2							
Occlusion sensitivity	3.1.2.1.	Zeiler and Fergus (2014)		√		√		√
Local Interpretable Model-agnostic Explanations (LIME)	3.1.2.2.	Ribeiro et al. (2016)		√		√		√
Meaningful Perturbation	3.1.2.3.	Fong and Vedaldi (2017)		√		√		√
Prediction difference analysis	3.1.2.4.	Zintgraf et al. (2017)		√		√		√
Textual explanation	3.2.							
Image captioning	3.2.1.	Vinyals et al. (2015)	√		√			√
Image captioning with visual explanation	3.2.2.	Zhang et al. (2017a)	√		√			√
Testing with Concept Activation Vectors (TCAV)	3.2.3.	Kim et al. (2018)		√		√	√	√
Example-based explanation	3.3.							
Triplet networks	3.3.1.	Hoffer and Ailon (2015)	√		√		√	√
Influence functions	3.3.2.	Wei Koh and Liang (2017)		√		√	√	
Prototypes	3.3.3	Chen et al. 2019	√		√			√

* Deep Shapley Additive exPlanations are post hoc and model-specific because of the optimization method, but Shapley Additive exPlanations can also be global and model-agnostic.

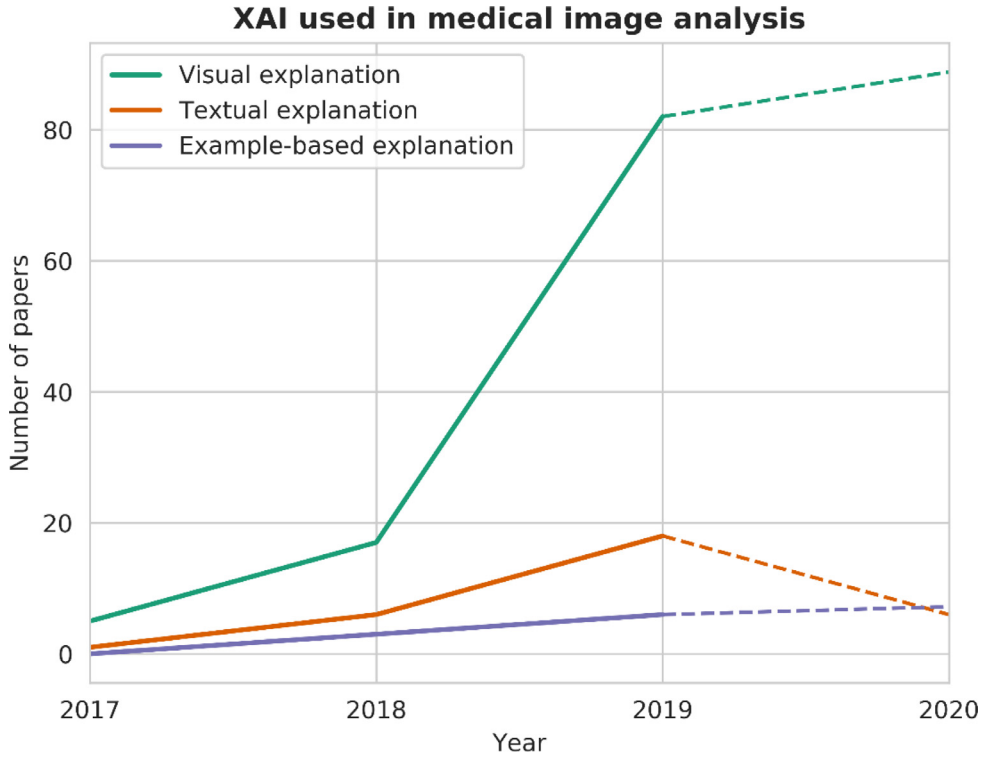

Fig. 2. Number of papers published per year in medical image analysis, for the three types of XAI techniques. Most papers use a visual explanation. The y-axis shows the number of papers included in this survey, the x-axis shows the year these papers were published in. The dashed line for 2020 is an extrapolation given the situation on October 31, 2020.

provided local, model-specific, post hoc explanation. Several researchers used this technique in medical imaging (Table 2).

CAMs have also been used in medical image analysis in ensembles of CNNs. For example, Jiang et al. (2019) constructed an ensemble of Inception-V3, ResNet-152, and Inception-ResNet-V2 to distinguish fundus images of healthy subjects or patients with mild diabetic retinopathy from those with moderate or severe diabetic retinopathy; and provided a weighted combination of the resulting CAMs for localization of diabetic retinopathy. Lee et al. (2019b) constructed CAMs of the output of an ensemble of four CNNs: VGG-16, ResNet-50, Inception-V3, and Inception-ResNet-V2, for the detection of acute intracranial hemorrhage.

Since medical images often contain information at multiple scales, multi-scale CAMs have also been proposed.

Table 2
Papers that provide visual explanation. For readability, the papers are sorted on anatomical location and only the first paper dealing with that anatomical location shows the location name. The column 'Main XAI technique used/based on' describes which visual explanation technique from Section 3.1 was used, or which technique the method in the corresponding paper is based on. When multiple visual explanation techniques have been applied, the most recent technique based on Table 1 has been noted. CAM = class activation mapping, CT = computed tomography, LIME = local interpretable model-agnostic explanations, LRP = Layer-wise relevance propagation, MRI = magnetic resonance imaging, OCT = optical coherence tomography, PET = positron emission tomography, SHAP = Shapley additive explanations.

Anatomical location	Authors (year)	Modality	Main XAI technique used/based on
Bladder	Woerl et al. (2020)	Histology	CAM
Brain	Ahmad et al. (2019)	MRI	CAM
	Baumgartner et al. (2018)	MRI	CAM
	Böhle et al. (2019)	MRI	LRP
	Ceschin et al. (2018)	MRI	CAM
	Chakraborty et al. (2020)	MRI	CAM
	Choi et al. (2020)	PET/CT	CAM
	Dang and Chaudhury (2019)	MRI	LRP
	Dubost et al. (2019b)	MRI	Guided backpropagation
	Dubost et al. (2019a)	MRI	Occlusion sensitivity
	Dubost et al. (2020)	MRI	Trainable attention
	Eitel et al. (2019)	MRI	LRP
	Fuchigami et al. (2020)	CT	Backpropagation
	Gao et al. 2019	MRI	Deconvolution
	Gao et al. (2019)	MRI	CAM
	Grigorescu et al. (2019)	MRI	LRP
	Hilbert et al. (2019)	MRI	Grad-CAM
	Kim and Ye (2020)	MRI	Grad-CAM
	Kubach et al. (2020)	Histology	Guided Grad-CAM
	Lee et al. (2019b)	CT	CAM
	Li et al. 2019b	MRI	CAM
	Lian et al. (2019)	MRI	Trainable attention
	Liao et al. (2020)	MRI	Grad-CAM
	Lin et al. (2019)	Ultrasound	CAM
	Natekar et al. (2020)	MRI	Grad-CAM
	Ng et al. (2018)	MRI	CAM
	Pereira et al. (2018)	MRI	Grad-CAM
	Pominova et al. (2018)	MRI	Grad-CAM
	Rezaei et al. (2020)	MRI	Backpropagation
	Saab et al. (2019)	CT	Multiple instance learning
	Seo et al. (2020)	MRI	Prediction difference analysis
	Shahamat and Saniee Abadeh (2020)	MRI	Occlusion sensitivity
	Shinde et al. (2019a)	MRI	CAM
	Shinde et al. (2019b)	MRI	CAM
	Tang et al. 2019	Histology	Grad-CAM
	Wang et al. 2020c	MRI	Guided backpropagation
	Wei et al. (2019)	MRI	Backpropagation
	Windisch et al. (2020)	MRI	Grad-CAM
	Xie et al. (2020)	Ultrasound	Grad-CAM
	Xu et al. (2019)	MRI	Trainable attention
	Xu et al. (2019)	MRI	LRP
	Ye et al. (2019)	CT	Grad-CAM
	Zintgraf et al. (2017)	MRI	Prediction difference analysis
Breast	Akselrod-Ballin et al. (2019)	X-ray	Meaningful perturbation
	El Adoui et al. (2020)	MRI	Grad-CAM
	Gecer et al. (2018)	Histology	Occlusion sensitivity
	Huang et al. (2020)	X-ray	CAM
	Kim et al. (2020)	Ultrasound	CAM
	Lee and Nishikawa (2019)	X-ray	CAM
	Luo et al. (2019)	MRI	CAM
	Maicas et al. (2019)	MRI	Multiple instance learning
	Obikane and Aoki (2020)	Histology	Grad-CAM
	Papanastasopoulos et al. (2020)	MRI	Integrated gradient
	Qi et al. (2019)	Ultrasound	CAM
	van der Velden et al. (2020)	MRI	SHAP
	Wang et al. (2018)	X-ray	Trainable attention
	Xi et al. (2019)	X-ray	CAM
	Yang et al. (2019)	Histology	Trainable attention
	Yi et al. (2019)	X-ray	CAM
	Zhou et al. 2020	Ultrasound	CAM
Cardiovascular	Candemir et al. (2020)	CT	Grad-CAM
	Cong et al. (2019)	X-ray	Grad-CAM
	Gessert et al. (2019)	OCT	Guided backpropagation
	Huo et al. (2019)	CT	Grad-CAM
	Patra and Noble (2020)	Ultrasound	Grad-CAM
	de Vos et al. (2019)	CT	Deconvolution

(*continued on next page*)

Table 2 (*continued*)

Anatomical location	Authors (year)	Modality	Main XAI technique used/based on
Chest	Ausawalaithong et al. (2018)	X-ray	CAM
	Brunese et al. (2020)	X-ray	Grad-CAM
	Chen et al. (2019)	X-ray	Grad-CAM
	Dunnmon et al. (2019)	X-ray	CAM
	Guo et al. (2020)	CT	CAM
	He et al. (2017)	Histology	Grad-CAM
	Hosny et al. (2018)	CT	Grad-CAM
	Huang and Fu (2019)	X-ray	CAM
	Humphries et al. (2020)	CT	Grad-CAM
	Khakzar et al. (2019)	X-ray	CAM
	Ko et al. (2020)	CT	Grad-CAM
	Kumar et al. (2019a)	CT	CAM
	Lei et al. (2020)	CT	CAM
	Li et al. 2019d	X-ray	Multiple instance learning
	Liu et al. 2019f	X-ray	CAM
	Mahmud et al. (2020)	X-ray	Grad-CAM
	Paul et al. 2020	CT	Grad-CAM
	Pesce et al. (2019)	X-ray	Trainable attention
	Philbrick et al. (2018)	CT	Grad-CAM
	Qin et al. (2020)	PET/CT	Grad-CAM
	Rajaraman et al. (2019)	X-ray	LIME
	Rajpurkar et al. (2018)	X-ray	CAM
	Schwab et al. (2020)	X-ray	Multiple instance learning
	Sedai et al. (2018)	X-ray	CAM
	Singla et al. (2018)	CT	Trainable attention
	Tang et al. (2019)	CT	CAM
	Tang et al. (2020)	X-ray	CAM
	Teramoto et al. (2019)	Histology	Grad-CAM
	van Sloun and Demi (2019)	Ultrasound	Grad-CAM
	Wang et al. 2019	X-ray	CAM
	Xu et al. 2019	CT	Grad-CAM
	Paul et al. (2020)	X-ray	CAM
	Zhu and Ogino (2019)	CT	SHAP
Dental	Vila-Blanco et al. (2020)	X-ray	Grad-CAM
Eye	Ahmad et al. 2019	Fundus photography	CAM
	Araújo et al. (2020)	Fundus photography	Multiple instance learning
	Costa et al. (2019)	Fundus photography	Multiple instance learning
	Jang et al. (2018)	Fundus photography	Guided Grad-CAM
	Jiang et al. (2019)	Fundus photography	CAM
	Kim et al. (2019)	Fundus photography	Grad-CAM
	Kumar et al (2019b)	Fundus photography	CAM
	Li et al. 2019a	Fundus photography	Trainable attention
	Liao et al. (2019)	Fundus photography	CAM
	Liu et al. (2019)	Fundus photography	CAM
	Martins et al. (2020)	Fundus photography	Grad-CAM
	Meng et al. (2020)	Fundus photography	Grad-CAM
	Narayanan et al. (2020)	Fundus photography	CAM
	Perdomo et al. (2019)	OCT	CAM
	Quellec et al. (2020)	Fundus photography	Backpropagation
	Shen et al. (2020)	Fundus photography	CAM
	Thakoor et al. (2019)	OCT	Grad-CAM
	Tu et al. (2020)	Fundus photography	CAM
	Wang et al. 2020a	OCT	Grad-CAM
	Wang et al. 2020b	CT	CAM
	Wang et al. 2019b	Fundus photography	CAM
	Zhang et al. (2019)	Fundus photography	Grad-CAM
	Zhou et al. (2020)	OCT	CAM
Female reproductive system	Gupta et al. (2020)	Histology	Grad-CAM
	GV and Reddy (2019)	Histology	Grad-CAM
	Sun et al. (2020)	Histology	CAM
Gastrointestinal	Chen et al. 2019	CT	Grad-CAM
	Everson et al. (2019)	Endoscopy	CAM
	García-Peraza-Herrera et al. (2020)	Endoscopy	CAM
	Heinemann et al. (2019)	Histology	CAM
	Itoh et al. (2020)	Endoscopy	Grad-CAM
	Kiani et al. (2020)	Histology	CAM
	Korbar et al. (2017)	Histology	Grad-CAM
	Kowsari et al. (2020)	Histology	Grad-CAM
	Lee et al. 2020	Ultrasound	Backpropagation
	Malhi et al. (2019)	Endoscopy	LIME
	Rajpurkar et al. (2020b)	CT	Grad-CAM
	Shapira et al. (2020)	CT	Multiple instance learning
	Wang et al. (2020)	MRI	Grad-CAM
	Wang et al. 2019a	Endoscopy	CAM
	Wickstrøm et al. (2020)	Endoscopy	Guided backpropagation
	Yan et al. (2020)	Histology	CAM
	Zhu et al. (2020)	Histology	Trainable attention

(*continued on next page*)

Table 2 (*continued*)

Anatomical location	Authors (year)	Modality	Main XAI technique used/based on
Lymph nodes	Ji (2019)	Histology	Grad-CAM
Musculoskeletal	Bien et al. (2018)	MRI	CAM
	Chang et al. (2020)	MRI	CAM
	Cheng et al. (2019)	X-ray	Grad-CAM
	Gupta et al. 2020	X-ray	Grad-CAM
	Jamaludin et al. (2017)	MRI	Guided backpropagation
	Kim et al. 2020	X-ray	Backpropagation
	Paul et al. (2019)	X-ray	CAM
	Zhang et al. (2020)	X-ray	Grad-CAM
	Zhao et al. (2018)	X-ray	CAM
	von Schacky et al. (2020)	X-ray	Grad-CAM
Prostate	Silva-Rodríguez et al. (2020)	Histology	CAM
	Yang et al. (2017)	MRI	CAM
Skin	Barata et al. (2020)	Dermatoscopy	Trainable attention
	Bian et al. (2019)	Photography	Backpropagation
	Li et al. (2020)	Dermatoscopy	CAM
	Li et al. 2019c	Photography	Prediction difference analysis
	Xie et al. 2020	Photography	CAM
	Yan et al. 2019	Dermatoscopy	Trainable attention
	Young et al. (2019)	Dermatoscopy	SHAP
	Zunair and Hamza (2020)	Photography	Grad-CAM
Skull	Kim et al. 2019b	X-ray	CAM
Thyroid	Lee et al. (2020)	CT	Grad-CAM
	Wang et al. 2019	Ultrasound	Attention
	Wang et al. 2020	Ultrasound	CAM
Multiple	Chan et al. (2019)	Histology	Grad-CAM
	Huang and Chung (2019)	Histology	CAM
	Hägele et al. (2020)	Histology	LRP
	Kermany et al. (2018)	Multiple	Occlusion sensitivity
	Kim et al. 2019	Multiple	CAM
	Langner et al. (2019)	MRI	Grad-CAM
	Meng et al. (2019)	Ultrasound	Trainable attention
	Schlemper et al. (2019)	CT	Trainable attention
	Tang (2020)	Multiple	CAM
	Upadhyay and Banerjee (2020)	Multiple	Grad-CAM

Liao et al. (2019) concatenated feature maps at three scales which were subsequently provided as input for the global average pooling. The provided activation maps showed higher resolution than single-scale maps, and were better at identifying small structures on fundus images of the retina. Shinde et al. (2019a) concatenated the feature maps of each layer before max-pooling and also gave those as input to a global average pooling layer. Their 'High Resolution' CAMs provided accurate localizations of brain tumors on MRI. García-Peraza-Herrera et al. (2020) proposed extracting CAMs at multiple resolutions. They showed that the CAMs at high resolution were accurate in highlighting interpapillary capillary loop patterns in endoscopy images, which were relatively small compared to the entire image.

Gradient-weighted class activation mapping (Grad-CAM): Selvaraju et al. (2017) introduced Gradient-weighted Class Activation Mapping (Grad-CAM), which is a generalization of CAM. Grad-CAM can work with any type of CNN to produce post hoc local explanation, whereas CAM specifically needs global average pooling. The authors also introduced guided Grad-CAM, an element-wise multiplication between guided backpropagation and Grad-CAM. Grad-CAM and Guided Grad-CAM have been used in medical image analysis. For example, Ji (2019) used Grad-CAM to show on which areas of histology lymph node sections a classifier based its decision of metastatic tissue; Kowsari et al. (2020) used it to pinpoint small bowel enteropathies on histology; and Windisch et al. (2020) used Grad-Cam to show which areas of brain MRI made the classifier decide on the presence of a tumor.

Layer-wise relevance propagation (LRP): Bach et al. (2015) introduced layer-wise relevance propagation (LRP). LRP uses the output of the neural network, e.g. a classification score between 0 and 1, and iteratively backpropagates this throughout the network. In each iteration (i.e., each layer), LRP assigns a relevance score to each of the input neurons from the previous layers. These distributed relevance scores must equal the total relevance score of its source neuron, according to the conservation law.

LRP has been used in medical image analysis. For example, Böhle et al. (2019) used LRP for identifying regions responsible for Alzheimer's disease from brain MR images. They compared the saliency maps provided by LRP with those provided by guided backpropagation, and found that LRP was more specific in identifying regions known for Alzheimer's disease.

Deep SHapley Additive exPlanations (Deep SHAP): Lundberg and Lee (2017) proposed a unified approach for explaining predictions by using SHapley Additive exPlanations (SHAP). This model-agnostic approach used Shapley values (Shapley, 2016), a concept from game theory. Shapley values determine the marginal contribution of every feature to the model's output individually. A downside of Shapley values is that they are resource-intensive to compute, since they require assessment of many permutations.

By combining DeepLIFT with Shapley values, Lundberg and Lee (2017) proposed a fast method to approximate Shapley values for CNNs called Deep SHAP. Deep SHAP has been used in medical image analysis. For example, van der Velden et al. (2020) used a regression CNN to estimate the volumetric breast density from breast MRI. Deep SHAP was used to explain which parts of the image had a positive contribution and a which parts a negative contribution to the density estimation.

Trainable attention: While many of the previously mentioned techniques highlighted what regions of the image the network focuses on, i.e. to where the attention was directed, Jetley et al. (2018) proposed a trainable attention mechanism. This trainable attention method highlighted where and in what proportion the network payed attention to input images for classification,

and used this attention to further amplify relevant areas and suppress irrelevant areas.

In medical imaging, Schlemper et al. (2019) used trainable attention and introduced grid attention. The rationale behind this was that most objects of interest in medical images are highly localized. By using grid attention, the trainable attention captured the anatomical information in medical images. They demonstrated high performance for both segmentation and localization, by adding the attention gates to a UNET (Ronneberger et al., 2015) and a variant of VGG (Simonyan and Zisserman, 2014). The attention coefficients were used to explain on which areas of the image the network focused.

3.1.2. Perturbation-based approaches

3.1.2.1. Occlusion sensitivity

Perturbation-based techniques perturb the input image to assess the importance of certain areas of that image for the task under consideration. Zeiler and Fergus (2014) used an occlusion sensitivity analysis to visualize which parts of the image were most important for classification. For example, they showed that an image of a dog holding a tennis ball was correctly classified by the dog's breed, except if the face of the dog was occluded, which yielded the incorrect classification 'tennis ball'.

3.1.2.2. Local interpretable model-agnostic explanations (LIME)

Ribeiro et al. (2016) introduced Local Interpretable Model-agnostic Explanations (LIME). LIME provides local explanation by replacing a complex model locally with simpler models, for example by approximating a CNN by a linear model. By perturbing the input data, the output of the complex model changes. LIME uses the simpler model to learn the mapping between the perturbed input data and the change in output. The similarity of the perturbed input to the original input is used as a weight, to ensure that explanations provided by the simple models with highly perturbed inputs have less effect on the final explanation. In images, Ribeiro et al. (2016) implemented the perturbations using superpixelsxxxxxxxxxxx is inclluded in the hyperlink"?>Achanta et al., 2012), rather than individual pixels, to show which regions were important for explaining a classification.

LIME has been used by several researchers in medical image analysis. For example, Malhi et al. (2019) used LIME to explain which areas in gastral endoscopy images contained bloody regions.

3.1.2.3. Meaningful perturbation

Fong and Vedaldi (2017) introduced meaningful perturbation, where they perturbed the input image to detect changes in the predictions of a trained neural network. Rather than using perturbations such as occlusion sensitivity that block out parts of the image, they suggested simulating naturalistic or plausible effects, leading to more meaningful perturbations, and consequently to more meaningful explanations. They opted for three types of local perturbations, namely a constant value, noise, or blurring.

Uzunova et al. (2019) stated that the perturbations proposed by Fong and Vedaldi (2017) were not suited for medical images. Replacing areas of a medical image with a constant value is implausible, and medical images naturally tend to be noisy and blurry. They proposed to replace pathological regions with a healthy tissue equivalent using a variational autoencoder (VAE). They showed that the perturbations by the VAE pinpoint pathological regions in diverse imaging studies as optical coherence tomography images of the eye (pathology consisted of intraretinal fluid, subretinal fluid, and pigment epithelium detachments), and MRI of the brain (pathology consisted of stroke lesions). Furthermore, they showed that using a VAE yielded better localization of pathology compared with using simple blurring or constant-value perturbations.

Lenis et al. (2020) used similar reasoning as Uzunova et al. (2019), and used inpainting to replace pathological regions with healthy tissue equivalents. They showed that the perturbations created by inpainting outperformed backpropagation and Grad-CAM in pinpointing masses in breast mammography and tuberculosis on chest X-rays, based on the Hausdorff distance between thresholded heatmaps derived from the saliency maps and the ground truth labels at pixel level.

3.1.2.4. Prediction difference analysis

Zintgraf et al. (2017) adapted prediction difference analysis (Robnik-Šikonja and Kononenko, 2008) for generating saliency maps. If each pixel in an image is considered a feature, prediction difference analysis assigns a relevance value to each pixel, by measuring how the prediction changes if the pixel is considered unknown. Zintgraf et al. (2017) expanded this by adding conditional sampling, which means that they only analyzed pixels that are hard to predict by simply investigating neighboring pixels, and by adding multivariable analysis, which means that they analyzed patches of connected pixels instead of single pixels. They included an analysis of brain MRI of patients with HIV versus healthy controls, yielding explanation of the classifier's decision.

Seo et al. (2020) used prediction difference analysis in combination with superpixels (or supervoxels for 3D) on multiple scales. These multiscale supervoxel-based saliency maps provided explanations that the authors described as visually pleasing since they follow image edges. The saliency maps explained which regions were informative for a classifier to distinguish between Alzheimer's disease patients and normal controls.

3.1.3. Multiple instance learning-based approaches

Multiple instance learning can be used for visualizing explanations. In multiple instance learning, training sets consist of bags of instances (Dietterich et al., 1997). These bags are labeled, but the instances are not. In medical image analysis, multiple instance learning can for example be done using a patch-based approach: An image represents the bag, and patches from that image represent the instances (Cheplygina et al., 2019).

Several researchers have used this approach to pinpoint which instances in the bag are responsible for the classification. For example, Schwab et al. (2020) localized critical findings in chest X-ray using such a patch-based approach. Each image patch received a prediction, and the predictions were overlaid on the image to visualize on which areas the classifier based its decision. Araújo et al. (2020) used multiple instance learning to explain which areas of a fundus photograph were important for diabetic retinopathy. They assessed the severity of the disease using an ordinal scale with grades from 0 to 5. Using a patch-based approach, they provided visual explanation maps for each diabetic retinopathy grade.

3.2. Textual explanation

Textual explanation is a form of XAI that provides textual descriptions. Such descriptions include relatively simple characteristics (e.g. 'spiculated mass'), up to entire medical reports. We will describe three types of textual explanation: image captioning, image captioning with visual explanation, and testing with concept attribution.

An overview of papers using textual explanation in medical imaging is shown in Table 3.

3.2.1. Image captioning

Vinyals et al. (2015) provided textual explanation for images using an end-to-end image captioning framework. They coupled a

Table 3
Papers that provide textual explanation. For readability, the papers are sorted on anatomical location and only the first paper dealing with that anatomical location shows the location name. The column 'Main XAI technique used/based on' describes which textual explanation technique from Section 3.2 was used, or which technique the method in the corresponding paper is based on. CT = computed tomography, TCAV = testing with concept activation vectors

Anatomical location	Authors (year)	Modality	Main XAI technique used/based on
Bladder	Zhang et al. (2017b)	Histology	Image captioning with visual explanation
Breast	Kim et al. 2019a	X-ray	Image captioning with visual explanation
	Lee et al. (2019a)	X-ray	Image captioning with visual explanation
	Sun et al. (2019)	X-ray	Image captioning
Cardiovascular	Clough et al. (2019)	MRI	TCAV
Chest	Gasimova (2019)	X-ray	Image captioning
	Kashyap et al. (2020)	X-ray	Image captioning with visual explanation
	Li et al. (2019)	X-ray	Image captioning with visual explanation
	Nunes et al. (2019)	X-ray	Image captioning with visual explanation
	Rodin et al. (2019)	X-ray	Image captioning with visual explanation
	Shen et al. (2019)	CT	Other textual explanation
	Singh et al. (2019)	X-ray	Image captioning
	Spinks and Moens (2019)	X-ray	Image captioning
	Tian et al. (2019)	X-ray	Image captioning
	Wang et al. 2019c	X-ray	Image captioning with visual explanation
	Wu et al. (2018)	CT	TCAV
	Yan et al. (2019)	CT	Other textual explanation
	Yang et al. 2020	X-ray	Image captioning
	Yin et al. (2019)	X-ray	Image captioning
	Yuan et al. (2019)	X-ray	Image captioning with visual explanation
Eye	Kim et al. (2018)	Fundus photography	TCAV
Female reproductive system	Ma et al. (2018)	Histology	Image captioning with visual explanation
Gastrointestinal	Tian et al. (2018)	CT	Image captioning with visual explanation
Kidney	Maksoud et al. (2019)	Histology	Image captioning
Musculoskeletal	Koitka et al. (2020)	X-ray	Image captioning
Multiple	Allaouzi et al. (2018)	Multiple	Image captioning
	Graziani et al. (2020)	Multiple	TCAV
	Jing et al. 2018	Multiple	Image captioning with visual explanation
	Pelka et al. (2019)	X-ray	Image captioning
	Zeng et al. (2020)	Multiple	Image captioning

convolutional neural network for encoding of the image, with a recurrent neural network – specifically a long-short term memory net (LSTM) (Hochreiter and Schmidhuber, 1997) – for textual encoding. They used human-generated sentences as ground truth for training, and used the bilingual evaluation understudy (BLEU) metric for evaluation. The BLEU-metric describes the precision of word N-grams, i.e. a sequence of N words, between generated and reference sentences (Papineni et al., 2002).

Singh et al. (2019) used an image captioning framework to provide textual explanation for chest X-rays. They used word-embedding databases Global Vectors (GloVe) (Pennington et al., 2014) and the radiology variant RadGloVe (Zhang et al., 2018) to train the LSTM, and used the aforementioned BLEU metric as well as variants METEOR, CIDER, and ROUGE (Banerjee and Lavie, 2005; Lin, 2004; Vedantam et al., 2015). As expected, higher performance was reached in the generated radiology report when both RadGloVe and GloVe were used instead of just GloVe.

3.2.2. Image captioning with visual explanation

Several researchers combined image captioning with visual explanation. Zhang et al. (2017a) introduced a framework that used dual attention, both for text and for imaging. They used a similar approach as with image captioning, i.e. an encoder for the image and an LSTM for the text, but added dual attention. This facilitated high-level interactions between image and text predictions, and yielded visual attention maps corresponding with textual explanation in Histology images.

Wang et al. 2018 used a similar approach, and showed in their chest X-ray example that different parts of the textual explanation led to different areas of saliency mapping in the image. They showed a saliency map of the chest with multiple regions corresponding to different radiological findings.

Lee et al. (2019a) showed image captioning with visual explanation for breast mammograms. They added a visual word constraint loss to the text-generating LSTM, to ensure that the provided explanations follow the correct jargon of breast mammography reports. They showed that adding this loss aids in generating better textual explanation. Furthermore, they linked the radiology reports to visual saliency maps.

3.2.3. Testing with concept activation vectors (TCAV)

Concept attributions provide explanation corresponding to high-level concepts that humans find easy to understand (Kim et al., 2018). Using Testing with Concept Activation Vectors (TCAV), Kim et al. (2018) presented human-friendly linear explanations of the internal state of neural networks, yielding global explanation of the networks in terms of human-understandable concepts. These concepts can be provided after training of the neural network as a post hoc analysis. The TCAV algorithm uses user-defined sets of examples of a concept and of random non-concept examples. Such a concept might be 'stripes' to assess whether an image contained a zebra, or 'spiculated mass' to assess whether an image contained a cancer. TCAV quantified the sensitivity of a trained model to such concepts using concept activation vectors (CAVs). The response of test cases to these CAVs was then used to measure the sensitivity to that concept. The authors showed feasibility of TCAV on a medical image processing example, by relating physician annotations such as 'microaneurysm' to diabetic retinopathy in fundus imaging.

Clough et al. (2019) identified cardiac disease in cine-MRI by classifying the latent space of a VAE. They used TCAV to show which clinically known biomarkers were related to cardiac disease. Furthermore, they reconstructed images with low peak ejection rate – a characteristic that might be related to cardiac disease – by adding the CAV to the latent space.

Table 4
Papers that provide example-based explanation. For readability, the papers are sorted on anatomical location and only the first paper dealing with that anatomical location shows the location name. The column 'Main XAI technique used/based on' describes which example-based explanation technique from Section 3.3 was used, or which technique the method in the corresponding paper is based on. CT = computed tomography, MRI = magnetic resonance imaging.

Anatomical location	Authors (year)	Modality	XAI technique used/based upon
Brain	Li et al. 2019d	MRI	Examples from the latent space
Breast	Uehara et al. (2019)	Histology	Prototypes
Chest	LaLonde et al. (2020)	CT	Examples from the latent space
	Silva et al. (2020)	X-ray	Examples from the latent space
Gastrointestinal	Peng et al. (2019)	Histology	Triplet network
	Wang et al. (2019)	MRI	Influence functions
Skin	Codella et al. (2018)	Dermatoscopy	Triplet network
	Sarhan et al. (2019)	Dermatoscopy	Examples from the latent space
Thyroid	Chen et al. (2020)	Histology	Examples from the latent space
	Li et al. 2020	Ultrasound	Prototypes
Multiple	Biffi et al. (2020)	MRI	Examples from the latent space
	Choudhary et al. (2019)	Histology	Triplet network
	Silva et al. (2018)	Multiple	Examples from the latent space
	Yan et al. (2018)	CT	Triplet network
	Yang et al. (2020)	Histology	Examples from the latent space with visual explanation

Graziani et al. (2020) expanded on TCAV by introducing regression concept vectors. The main addition was that, while TCAV indicated the presence or absence of binary concepts, regression concept vectors indicated continuous-valued measures of a concept. This can be useful when investigating a continuous concept such as tumor size. Graziani et al. (2020) showed that by using regression concept vectors, they could for example explain why the network classified one area of a breast histopathology image as cancer and another as healthy: Both areas of the image scored high on the concept 'contrast', but the concept 'nuclei area', referring to a clinically used system for evaluating cell size, was different between healthy and cancerous regions.

3.2.4. Other tel explanation techniques

Shen et al. (2019) used what they called a hierarchical semantic CNN to predict malignancy of lung nodules on CT. They classified five textual descriptions of image characteristics representative of lung nodule malignancy that are typically assessed by a radiologist. The task of finding textual descriptions was combined with the main task of classifying lung nodule malignancy. Although their hierarchical semantic CNN did not significantly outperform a normal CNN in predicting nodule malignancy, the method did provide human-interpretable characteristics of the nodules.

3.3. Example-based explanation

Example-based explanation is an XAI technique that provides examples relating to the data point that is currently being analyzed. This can be useful when trying to explain why a neural network came to a decision, and is related to how humans reason. For example, when a pathologist examines a biopsy of a patient that shows similarity with an earlier patient examined by the pathologist, the clinical decision may be enhanced by knowing the assessment of that earlier biopsy.

Example-based explanation often optimizes the hidden layers deep in the neural network (i.e., the latent space) in such a way that similar points are close to each other in this latent space, while dissimilar points are further away in the latent space.

An overview of papers using example-based explanation in medical imaging is shown in Table 4.

3.3.1. Triplet network

Several papers provided example-based explanation using a triplet network (Hoffer and Ailon, 2015). A triplet network consists of three identical networks with shared parameters. By feeding these networks three input samples, the network calculates two values consisting of the L_2 distances between the representations in the latent space (i.e., embedded representations) of these input samples. This allows learning of useful representations by unsupervised comparison of samples. When analyzing a data point, inspection of neighbors in this embedded representation will provide examples of data points that are similar to the data point that is being analyzed, which can provide explanation why the network came to its output.

Peng et al. (2019) used example-based explanation in colorectal cancer histology. They first trained a CNN using a triplet loss, hashing, and k hard-negatives to learn an embedding that preserves similarity. In testing, a coarse-to-fine search yielded the 10 nearest examples from a testing database related to the input image. This provided explanation on which images similar to the image that was being analyzed the network based a decision.

Yan et al. (2018) utilized a radiological picture archiving and communication systems (PACS) to extract 32000 clinically relevant lesions from the entire body. To learn relevant lesion embeddings, they trained a triplet network with three supervision cues: lesion size, lesion anatomical location (e.g. lung, liver, or kidney), and relative coordinate of the lesion in the body. These embeddings showed good separation based on anatomical location (e.g., liver lesions were separated from lung lesions), and could accurately retrieve example-based explanation from a test set.

Codella et al. (2018) also used a triplet loss but combined it with global average pooling, the technique used in CAM. Consequently, they could not only extract example-based explanation, but they also provided query activation maps and search result activation maps. In other words, a visual explanation showed which region of the input image the network used to generate the example-based explanation. They demonstrated this technique in dermatology images of melanoma.

3.3.2. Influence functions

Wei Koh and Liang (2017) proposed to use influence functions to explain on which inputs from a training set a decision was based. They did so by investigating what would happen in case an input from the training set would not be available or would be changed. Since it is expensive to assess this by perturbation, they provided an efficient approximation using influence functions (Cook and Weisberg, 1980). This implementation of influence functions is related to SHAP in the sense that they both allow efficient computation of feature importance.

Wang et al. (2019) used influence functions to explain which classifications of liver lesions on multiphase MRI were associ-

ated with which radiological characteristics. This global explanation provided insight into the neural network's behavior. For example, the class 'benign cyst' was most often associated with the radiological finding 'thin-walled mass'. Since the network did not only output the class label but also the corresponding radiological characteristics, this explanation could enhance user trust in the output of the network.

3.3.3. Prototypes

Chen et al. 2019 proposed to use typical examples as explanation (i.e., prototypes), which they described as 'this-looks-like-that'. The method reflected case-based reasoning that humans perform. For example, when a person explains why a picture contains a car, they can internally reason that this is a car because it looks like a car they have seen before. A prototype layer was added to the neural network, which grouped training inputs according to their classes in the latent space. A prototype was picked for each class, consisting of a typical example of that class. During testing, the method utilized parts of the test image that resembled these trained prototypes. The output was a weighted combination of the similarities to these prototypes. Hence, the explanation was an actual computation of the neural network, not a post hoc approximation.

Uehara et al. (2019) used prototypes to explain why a neural network classified patches of histology images as cancer or as not-cancer. The network was able to identify on which parts of the image it based its decision, and to what extent these parts of the image were similar to prototypical examples learned from the training set.

3.3.4. Examples from the latent space

Sarhan et al. (2019) proposed learning disentangled representations of the latent space using a residual adversarial VAE with a total correlation constraint. This adversarial VAE enhanced the fidelity of the reconstruction and provided more detailed descriptions of underlying generative characteristics of the data. When analyzing reconstructions by traversing through the latent space, they showed that their method yielded reconstructions that were more true to human-interpretable concepts such as lesion size, lesion eccentricity, and skin color compared with a regular VAE.

Biffi et al. (2020) provided a framework for explainable anatomical shape analysis using a ladder VAE (Sønderby et al., 2016). They coupled this ladder VAE with a multi-layered perceptron, enabling the network to train end-to-end for classification tasks. By doing this, the highest level of the latent space was enforced to be low-dimensional (2D or 3D), which meant that these learned latent spaces could be directly visualized without the need of further dimensionality reduction after training. They provided dataset-level explanation using these low-dimensional latent spaces to visualize differences in shape for hypertrophic cardiomyopathy versus healthy controls on cardiac MRI, and for Alzheimer's disease versus healthy controls on brain MRI by visualizing the shape of the hippocampus.

Silva et al. (2018) proposed example-based explanation that showed similar and dissimilar cases foraesthetic results of breast surgery on photos, and for skin images on dermoscopy. They identified these examples using a nearest neighbor search in latent space: The nearest neighbor of the same class was considered the most similar case, and the nearest neighbor of the other class was considered the most dissimilar case. Their explanation also included rule extraction from meta-features (e.g. the color of a skin lesion or the visibility of scars). They proposed three criteria to measure the validity of the rule-extracted explanation, namely: (1) completeness, i.e. the explanation should be general enough to be applied to more than one observation; (2) correctness, i.e. if the explanation itself was considered a model, it should correctly identify which class it belongs to; and (3) compactness, i.e. the explanation should be succinct.

In later work, Silva et al. (2020) combined example-based explanation with saliency mapping. First, they trained a baseline CNN to classify chest X-rays into pleural effusion versus non-pleural effusion. After that, the CNN was fine-tuned on saliency maps. In testing, a nearest neighbor search between the latent space of the test image and a curated 'catalogue' set of images was performed. Adding the saliency map yielded more consistent examples than extracting examples without the saliency map (i.e., the baseline CNN).

Sabour et al. (2017) showed that by replacing the scalar feature maps from convolution neural networks by vectorized representations (i.e., capsules), they were able to encode high-level features of images. Capsules were basically subcollections of neurons in a layer. These were linked to subcollections of neurons in subsequent layers, forming a capsule network. This capsule network was optimized using dynamic routing. In short, higher level capsules were activated if their corresponding lower-level capsules are active. This correspondence was described by routing coefficients, which summed to one for each capsule. The coefficients were iteratively (i.e., dynamically) updated when the capsule network received new input data. For the MNIST digits dataset, Sabour et al. (2017) found that these capsules learn human-interpretable features such as scale, thickness, and skew.

LaLonde et al. (2020) used capsules for lung cancer diagnosis, while also predicting visual attributes such as sphericity, lobulation, and texture. Since these visual attributes were not necessarily mutually exclusive, as was the case in MNIST (a digit cannot be a two and a nine at the same time), they adapted the dynamic routing algorithm accordingly. Specifically, the routing coefficients did not have to sum to one in their implementation. LaLonde et al. (2020) showed that their implementation was indeed able to predict these visual attributes as well as lung nodule malignancy.

4. Pros and cons of XAI techniques

All XAI techniques described in Section 3 have pros and cons, influencing how one would choose from the various options. We will structure these pros and cons in the categories ease of use, validity, robustness, computational cost, necessity to fine-tune, and open-source availability. An overview of these pros and cons per method from Table 1 is given in Table 5.

4.1. Ease of use

We define the ease of use by the potential of XAI techniques to be 'plug-and-play'. Post hoc model agnostic techniques have the highest ease of use. These methods generally consist of perturbation-based visual explanation techniques such as occlusion sensitivity. These techniques can be used on any trained neural network to provide a visual explanation. Model-based techniques typically have lowest ease of use, since the explanation is embedded in the design of the neural network.

4.2. Validity

We define validity by whether the explanation is correct and corresponds to what the end-user expects. In case of visual explanation, this can be assessed for example by asking a radiologist whether the explanation points towards the pathology that the neural network was designed to classify.

Research on quantifying validity of XAI is sparse, and currently focuses on visual explanation. Arun et al. (2021) aimed to quan-

Table 5
Pros and cons of XAI techniques. Pros are depicted by +, cons by -. The letters in the column Open source (original paper) refer to the URL below the table.

Technique	Ease of use	Validity	Robustness	Computational needs	No fine-tuning required	Open-source (original paper)	Open-source (captum.ai)
Visual explanation							
Backpropagation-based approaches							
Backpropagation	+	-	+	-	+	-	+
Deconvolution	+	n.t.	n.t.	-	+	-	+
Guided backpropagation	+	-	inc.	-	+	-	+
Class activation mapping (CAM)	+	n.t.	-	-	+	a	-
Gradient-weighted class activation mapping (Grad-CAM)	+	+/-	-	-	+/-	b	+
Layer-wise relevance propagation (LRP)	+	n.t.	+	-	+/-	-	+
Deep SHapley Additive exPlanations (Deep SHAP)	+	n.t.	n.t.	-	+/-	c	+
Trainable attention	+/-	n.t.	n.t.	+	-	d	-
Perturbation-based approaches							
Occlusion sensitivity	+	n.t.	-	+	-	-	+
Local Interpretable Model-agnostic Explanations (LIME)	+	n.t.	n.t.	+	-	e	+
Meaningful Perturbation	+	n.t.	n.t.	+	-	f	-
Prediction difference analysis	+	n.t.	n.t.	+	-	g	-
Textual explanation							
Image captioning	+/-	n.t.	n.t.	+	-	-	-
Image captioning with visual explanation	+/-	n.t.	n.t.	+	-	h	-
Testing with Concept Activation Vectors (TCAV)	+	n.t.	n.t.	n.t.	+/-	i	-
Example-based explanation							
Triplet networks	+/-	n.t.	n.t.	+	-	j	-
Influence functions	+	n.t.	n.t.	n.t.	+/-	k	-
Prototypes	+/-	n.t.	n.t.	+	-	l	-

n.t. = not tested by studies on that criterion.
inc. = inconclusive results between studies on that criterion.
a https://github.com/zhoubolei/CAM
b https://github.com/Cloud-CV/Grad-CAM
c https://github.com/slundberg/shap
d https://github.com/saumya-jetley/cd_ICLR18_LearnToPayAttention
e https://github.com/marcotcr/lime
f https://github.com/ruthcfong/perturb_explanations
g https://github.com/lmzintgraf/DeepVis-PredDiff
h https://github.com/zizhaozhang/tandemnet
i https://github.com/tensorflow/tcav
j https://github.com/eladhoffer/TripletNet
k https://github.com/kohpangwei/influence-release
l https://github.com/cfchen-duke/ProtoPNet

tify the validity of visual explanation techniques using the SIIM-ACR Pneumothorax Segmentation and RSNA Pneumonia Detection databases (Society for Imaging Informatics in Medicine and American College of Radiology, 2019; Radiological Society of North America, 2018). They compared four of the methods discussed in this paper: backpropagation, guided backpropagation, Grad-CAM, and guided Grad-CAM. Of these methods, Grad-CAM showed the highest validity. Note that this study solely focuses on chest X-rays. Therefore, more research is needed to investigate the validity of visual explanation techniques in other modalities and anatomical locations.

In case of textual explanation, validity can be assessed by comparing the generated textual explanation to the ground truth text. In case of example-based explanation, validity can be assessed by comparing relevant characteristics of found examples, such as patient or clinicopathological characteristics. To the best of our knowledge, there have not been such rigorous studies on validity performed for textual explanation and for example-based variation as there are for visual explanation (Arun et al., 2021). Hence, more research in this area is desired.

4.3. Robustness

The robustness of XAI techniques can be assessed by intentionally changing certain aspects of the deep learning framework and measuring the effect of these changes to the given explanation.

The robustness is mainly quantified for visual explanation techniques, using parameter randomization tests and data randomization tests.

The parameter randomization test compares visual explanation from a trained CNN with visual explanation from a randomly initialized untrained CNN of the same architecture. If the explanation depends on the learned parameters of the CNN (the desired situation), the two explanations should differ substantially. If the two explanations are similar, the visual explanation technique is insensitive to the properties of the CNN.

The data randomization test compares visual explanation from a trained CNN with visual explanation from a CNN trained on the same dataset but with randomly imputed labels. If the explanation depends on the data labels (the desired situation), the two explanations should differ substantially. If the two explanations are similar, the visual explanation does not depend on the relationship between images and labels.

Adebayo et al. (2018) performed these two tests for many visual explanation methods including backpropagation, guided backpropagation, Grad-CAM, and guided Grad-CAM. They showed that guided backpropagation and guided Grad-CAM provided a similar visual explanation in both tests, and might be emphasizing edges. Hence, caution is advised when using such methods for visualization.

Eitel and Ritter (2019) evaluated the robustness of visual explanation techniques guided backpropagation, layer-wise relevance

propagation, and occlusion sensitivity in medical images over multiple training runs, specifically for the classification of Alzheimer's disease using brain MRI. They found that layer-wise relevance propagation and guided backpropagation produced the most coherent visual explanation. This was not fully in line with the results of Adebayo et al. (2018).

Arun et al. (2021) performed similar analyses. Their results showed that guided backpropagation and Grad-CAM passed the parameter randomization test.

These conflicting results demonstrate that more research is desired for visual explanation techniques in medical image analysis. For textual and example-based XAI, such rigorous comparison studies have not yet been performed.

4.4. Computational cost

Computational cost of XAI is seldom reported in papers, but can be assessed by comparing how these explanation techniques work.

Since model-based techniques embed the explanation in the design of the neural network, it is obvious that these explanations are relatively costly to produce.

For visual explanation techniques, there is a clear distinction between backpropagation-based and perturbation-based techniques with respect to their computational needs. Backpropagation-based techniques typically make a single pass back through the neural network, which is relatively fast. Perturbation-based techniques require, however, extensive perturbation of input images to measure the influence of these perturbations on the output. Therefore, these techniques are generally more computationally-expensive. This can especially be the case in 3-dimensional, 4-dimensional, and/or multi-modality images, which often occur in medical image analysis.

The computational costs of the post hoc textual explanation TCAV and the post hoc example-based explanation of influence functions in medical image analysis has not rigorously been reported.

4.5. Necessity of fine-tuning

Some explanation techniques require no fine-tuning of parameters while others require fine-tuning of parameters associated with the XAI technique.

Since model-based techniques embed the explanation in the design of the neural network, it is obvious that fine-tuning of the network will influence the explanation.

For visual explanation, most backpropagation techniques have a limited number of parameters to tune. For example, in Grad-CAM, the user needs to choose at which layer to inspect the activation and in Deep SHAP, one needs to choose samples from the training set to calculate a background signal.

Perturbation-based visual explanation techniques often require a choice of the perturbation. For example, both occlusion sensitivity and LIME require the user to define the size and shape of the occluded areas. In meaningful perturbation, the user has to define what kind of perturbation technique is deemed best.

The post hoc textual explanation TCAV requires some fine-tuning with respect to the concepts that will be tested. The post hoc example-based explanation technique of influence functions requires definition of the functions of which the influence is to be measured.

4.6. Open-source availability

Most XAI techniques are available from open source. Often, code is available from the authors of the original paper. Many techniques are also implemented in XAI packages such as captum.ai. An overview of open-source availability of XAI techniques is given in Table 5.

5. Discussion

5.1. Overview

We have discussed 223 papers on eXplainable Artificial Intelligence (XAI) for deep learning in medical image analysis. We categorized the papers based on the XAI-frameworks proposed by Adadi and Berrada (2018) and Murdoch et al. (2019). Some trends were noticeable in the surveyed papers. The majority of the papers used post hoc explanation as contrasted with model-based explanation, i.e., the explanation was provided on a neural network that had already been trained, instead of being incorporated in neural network training. Both model-specific (e.g., specifically designed for CNNs) and model-agnostic explanation methods were used. Furthermore, most of the papers investigated provided local explanation rather than global explanation, i.e., the explanation was provided per case (e.g. per patient), rather than on a dataset-level (e.g. for all patients). Since we focus on deep learning in medical image analysis, these trends were to be expected. Most readily available XAI methods suitable for CNNs are saliency mapping techniques, which often provide post hoc, model-specific, and local explanation. Furthermore, post hoc XAI methods can be used after a neural network has been trained, making them more accessible than model-based XAI.

We categorized the papers based on anatomical location and modality of medical imaging. We found that most papers focus on chest or brain and on MRI (Fig. 3). This is comparable to what Litjens et al. (2017) found for deep learning methods in medical imaging in general. This trend is likely due to publicly available datasets in these organs and modalities, and not a reflection of how well explainable these organs and modalities are.

5.2. Evaluation of XAI

We have described several XAI techniques and their applications in medical image analysis, but how does one evaluate whether an XAI technique provides good explanation? Unlike measures of performance commonly used in medical image analysis, such as accuracy, Dice coefficient, or an ROC analysis; success criteria of explanation are more difficult to define. Doshi-Velez and Kim (2017) proposed a framework for the evaluation of explainability, consisting of three evaluation methods: application-grounded evaluation, human-grounded evaluation, and functionally-grounded evaluation.

5.2.1. Application-grounded evaluation

Application-grounded evaluation uses human experiments within a real application. In other words, let domain experts test the explanation. In medical image analysis this might involve a radiologist inspecting whether example-based explanations are actually good examples based on the many images the radiologist has seen in their many years of experience. The advantage of application-grounded evaluation is that it directly tests the objective that the system was built for. The disadvantage is that it is a costly evaluation.

5.2.2. Human-grounded evaluation

Human-grounded evaluation uses simpler human experiments that maintain the essence of the target application. In other words, let laypersons test the explanation or a proxy of the explanation. For example, when explaining the location and size of a cancer, this might involve a crowdsourcing project where laypersons judge the quality of saliency maps. Since it uses laypersons instead of

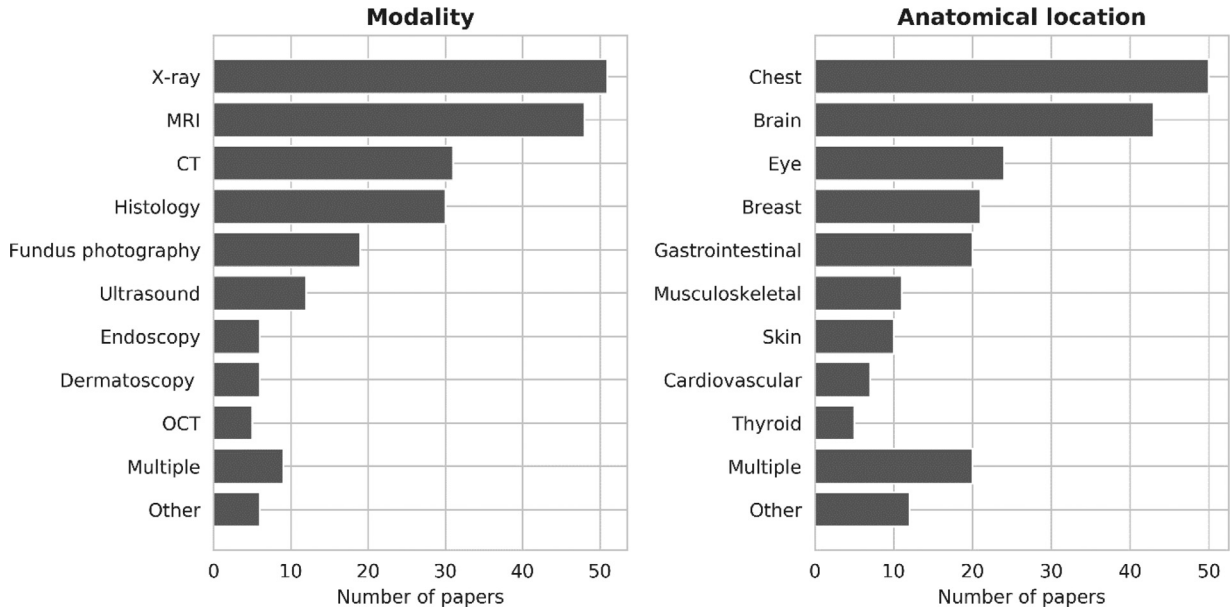

Fig. 3. Papers included in this survey, categorized by modality (left) and anatomical location (right). Papers discussing multiple modalities or anatomical locations were grouped as 'multiple'. Modalities or anatomical locations that were used in fewer than five papers were grouped as 'other'.

highly trained domain experts, the advantage of human-grounded evaluation is that it is less costly, while still receiving general notions of the quality of an explanation. The disadvantage is that the assessment of the quality of an explanation is a proxy of the actual quality.

5.2.3. Functionally-grounded evaluation

Functionally-grounded evaluation does not use human experiments, but uses other proxies to assess the quality of the explanation. These proxies may include measurements that have already been validated using human users. In our example of explaining the location and size of a cancer, this might involve comparing the explanation with manually drawn tumor delineations of a radiologist. The advantages of functionally-grounded evaluation stated by Doshi-Velez and Kim (2017) include that they are relatively cheap to acquire. This is, however, not necessarily the case in medical image analysis, since acquiring for example manual annotations is a very resource intensive process. When these manual annotations do already exist, e.g. when using curated data from a challenge, evaluation of explanations are easily extracted, and can be automatically extracted multiple times. This can be useful, for example in the development phase of explanation methods.

5.2.4. Evaluation of XAI in medical image analysis

Evaluation of XAI as proposed above is currently not yet standard practice in papers in medical image analysis. Furthermore, in medicine a good explanation can differ between areas of expertise of the person for whom the explanation is given. For example, a visual explanation pinpointing where disease is located could be a sufficient explanation for a radiologist or a medical image analysis researcher. However, clinicians such as an oncologist, neurologist, or hematologist would probably like to have XAI added to their clinical decision-making framework. Such framework would also incorporate the patient's history, previous and current treatments, treatment options, and expected effects or outcomes.

5.3. Critique on XAI

Rudin (2019) advised caution when using a black box with explanation for high-stakes decision making. Rudin raised several issues with explaining black boxes. For example, XAI may provide an explanation that is not completely faithful to what the original model computes: If the explanation explains 90% true to the model, that means that 10% is untrue (Rudin, 2019). Furthermore, an explanation may not make sense or provide enough detail to understand what the black box is doing. For example, a saliency map of the class with the highest probability may look similar to a saliency map of a class with a lower probability. Rudin therefore advices to use interpretable model-based XAI instead, such as the prototype network discussed in Section 3.3.3.

Critiques also often focus on the robustness of XAI techniques, as discussed in Section 4.

5.4. Outlook

Since high stakes decision-making is intertwined with medicine, we are convinced that XAI will be increasingly important. We have investigated the trends, and noticed that an increasing amount of papers contain a holistic approach, combining multiple forms of explanation. Examples of such more holistic approaches include combinations of textual explanation and visual explanation (e.g. Graziani et al., 2020), or combinations of example based explanation and visual explanation (e.g. Wang et al., 2019).

Future directions of XAI in medical image analysis may include biological explanation. Several researchers have predicted biological processes from imaging features using deep learning. For example, Matsui et al. (2020) predicted the molecular subtype of lower-grade gliomas on multimodal brain imaging, and Zhu et al. (2019) predicted the molecular subtype luminal A of breast cancer on MRI. These analyses used a biological target to train the neural network. However, performing such analysis the other way around, for example by performing a pathway analysis on imaging phenotypes (e.g. Bismeijer et al. (2020), not deep learning), could provide interesting biological explanation.

XAI may also be useful to aid physicians in the diagnostic process or in identifying unknown information from medical images. For example, a study on the diagnosis of tuberculosis on chest X-rays showed that 10 out of the 13 participating physicians (77%) had better diagnostic accuracy when assessing chest X-rays with an XAI providing a visual explanation compared to assessing the chest X-ray without XAI (Rajpurkar et al., 2020a).

It is likely that XAI in medical imaging will increasingly include domain information. To reach this goal, physicians should be included when designing task-specific interpretation methods (Fan et al., 2021). Active collaboration among physicians, theoretical researchers, medical imaging experts, and medical image analysis experts will be an important avenue for future development of deep learning methods (Fan et al., 2021).

Other directions of XAI in medical image analysis may include the link between causality and XAI. Typical medical image analysis consists of correlation rather than causation. Causality describes the relation between cause and effect, and can be mathematically described (Pearl, 2009). Current XAI techniques that aim to be free of bias such as prototypes are potentially still sensitive to differences in training population, which might hamper generalizability. Castro et al. (2020) describe how causal reasoning may be useful to assess biases in the data. DeGrave et al. (2021) gave an example how dataset bias can be detected using XAI: In studies that distinguish between X-rays of patients who were Coronavirus disease 2019 (COVID-19)-positive and of patients who were COVID-19-negative, they used visual explanation to demonstrate that high performance of the deep learning models was actually attributed to how the datasets were composed, rather than to actual COVID-19 detection in the X-rays. van Amsterdam et al. (2019) show an example of eliminating bias using causality, yielding unbiased prediction of prognosis for patients with lung cancer. It would be of interest to incorporate such analyses in explanation of medical images, as Chattopadhyay et al. (2019) have done for visual explanation of MNIST data.

There is no consensus on a priori estimations for required sample size for XAI and deep learning in medical imaging in general (Balki et al., 2019). Given the costly nature of acquiring medical imaging datasets in terms of money, time, and patient burden, it is desired to have guidelines describing what minimum sample sizes would be required for which XAI techniques.

5.5. Limitations

We derived our XAI framework from the frameworks of Adadi and Berrada (2018) and Murdoch et al. (2019). Other frameworks also exist, such as the framework by Kim et al. that divides XAI in pre-, during-, and post-model explanation. During- and post-model explanation are captured by our XAI framework with model-based and post hoc explanation. Pre-model explanation mainly focuses on the structure of a dataset, such as inspecting outliers. One could state that an example-based explanation that utilizes the latent distributions of a dataset could be perceived as a pre-model explanation. We have, however, not made this distinction, since in deep learning, these latent distributions are discovered by training a neural network.

We tried to be as comprehensive as possible with the inclusion of papers in our survey. However, XAI often is a technique used to support methods, and keywords are often not mentioned in the title or body of papers (Rudin, 2019). Therefore, we cannot guarantee that we covered all the work in the field. Nevertheless, we provided the search strategy to be as transparent as possible about the selection of papers.

6. Conclusion

This paper surveyed 223 papers using explainable artificial intelligence (XAI) in deep-learning based medical image analysis, classified according to an XAI framework, and categorized according to anatomical location and imaging technique. The paper discussed how to evaluate XAI, current critiques on XAI, and future perspectives for XAI in medical image analysis.

Additional information

This work was partially funded by the Dutch Cancer Society (KWF) Grant No.: 10755. We have no conflicts of interest.

Declaration of Competing Interest

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

References

Abbasi-Asl, Reza, Yu, Bin, 2017. Structural compression of convolutional neural networks arXiv preprint arXiv:1705.07356.

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Süsstrunk, S., 2012. SLIC superpixels compared to state-of-the-art superpixel methods. IEEE Trans. Pattern Anal. Mach. Intell. 34, 2274–2281. doi:10.1109/TPAMI.2012.120.

Adadi, A., Berrada, M., 2018. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). IEEE Access 6, 52138–52160. doi:10.1109/ACCESS.2018.2870052.

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., Kim, B., 2018. Sanity checks for saliency maps. Advances in neural information processing systems, p. 31.

Ahmad, A., Sarkar, S., Shah, A., Gore, S., Santosh, V., Saini, J., Ingalhalikar, M., 2019. Predictive and discriminative localization of IDH genotype in high grade gliomas using deep convolutional neural nets. In: Proceedings of the IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), pp. 372–375.

Ahmad, M., Kasukurthi, N., Pande, H., 2019. Deep learning for weak supervision of diabetic retinopathy abnormalities. In: Proceedings of the IEEE 16th International Symposium on Biomedical Imaging (ISBI), pp. 573–577.

Akselrod-Ballin, A., Chorev, M., Shoshan, Y., Spiro, A., Hazan, A., Melamed, R., Barkan, E., Herzel, E., Naor, S., Karavani, E., Koren, G., Goldschmidt, Y., Shalev, V., Rosen-Zvi, M., Guindy, M., 2019. Predicting breast cancer by applying deep learning to linked health records and mammograms. Radiology 292, 331–342. doi:10.1148/radiol.2019182622.

Allaouzi, I., Ben Ahmed, M., Benamrou, B., Ouardouz, M., 2018. Automatic caption generation for medical images. In: Proceedings of the 3rd International Conference on Smart City Applications, pp. 1–6.

Araújo, T., Aresta, G., Mendonça, L., Penas, S., Maia, C., Carneiro, Â., Mendonça, A.M., Campilho, A., 2020. DR|GRADUATE: uncertainty-aware deep learning-based diabetic retinopathy grading in eye fundus images. Med. Image Anal. 63. doi:10.1016/j.media.2020.101715.

Arun, N., Gaw, N., Singh, P., Chang, K., Aggarwal, M., Chen, B., Hoebel, K., Gupta, S., Patel, J., Gidwani, M., 2021. Assessing the (un) trustworthiness of saliency maps for localizing abnormalities in medical imaging. Radiol. Artif. Intell., e200267.

Ausawalaithong, W., Thirach, A., Marukatat, S., Wilaiprasitporn, T., 2018. Automatic lung cancer prediction from chest X-ray images using the deep learning approach. In: Proceedings of the 11th Biomedical Engineering International Conference (BMEICON), pp. 1–5.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., Samek, W., 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS One 10, 1–46. doi:10.1371/journal.pone.0130140.

Balki, I., Amirabadi, A., Levman, J., Martel, A.L., Emersic, Z., Meden, B., Garcia-Pedrero, A., Ramirez, S.C., Kong, D., Moody, A.R., Tyrrell, P.N., 2019. Sample-size determination methodologies for machine learning in medical imaging research: a systematic review. Can. Assoc. Radiol. J. doi:10.1016/j.carj.2019.06.002.

Banerjee, S., Lavie, A., 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pp. 65–72.

Barata, C., Celebi, M.E., Marques, J.S., 2020. Explainable skin lesion diagnosis using taxonomies. Pattern Recognit. doi:10.1016/j.patcog.2020.107413.

Baumgartner, C.F., Koch, L.M., Tezcan, K.C., Ang, J.X., Konukoglu, E., 2018. Visual feature attribution using Wasserstein GANs. In: Proceedings of the 31st Meeting of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2018, Switzerland, pp. 8309–8319. doi:10.1109/CVPR.2018.00867 IEEE Computer Society, Computer Vision Lab, ETH Zurich.

Bian, Z., Xia, S., Xia, C., Shao, M., 2019. Weakly supervised vitiligo segmentation in skin image through saliency propagation. In: Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 931–934.

Bien, N., Rajpurkar, P., Ball, R.L., Irvin, J., Park, A., Jones, E., Bereket, M., Patel, B.N., Yeom, K.W., Shpanskaya, K., et al., 2018. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: development and retrospective validation of MRNet. PLoS Med. 15, e1002699.

Biffi, C., Doumou, G., Duan, J., Prasad, S.K., Cook, S.A., O Regan, D.P., Rueckert, D., Cerrolaza, J.J., Tarroni, G., Bai, W., De Marvao, A., Oktay, O., Ledig, C., Le Folgoc, L., Kamnitsas, K., 2020. Explainable anatomical shape analysis through deep hierarchical generative models. IEEE Trans. Med. Imaging doi:10.1109/tmi.2020.2964499, 1–1.

Bismeijer, T., van der Velden, B.H.M., Canisius, S., Lips, E.H., Loo, C.E., Viergever, M.A., Wesseling, J., Gilhuijs, K.G.A., Wessels, L.F.A., 2020. Radiogenomic analysis of

breast cancer by linking mri phenotypes with tumor gene expression. Radiology 296, 277–287. doi:10.1148/radiol.2020191453.
Böhle, M., Eitel, F., Weygandt, M., Ritter, K.Initiative, on behalf of the A.D.N., 2019. Layer-wise relevance propagation for explaining deep neural network decisions in MRI-based Alzheimer's disease classification. Front. Aging Neurosci. 10. doi:10.3389/fnagi.2019.00194.
Brunese, L., Mercaldo, F., Reginelli, A., Santone, A., 2020. Explainable deep learning for pulmonary disease and coronavirus COVID-19 detection from X-rays. Comput. Methods Progr. Biomed. 196. doi:10.1016/j.cmpb.2020.105608.
Candemir, S., White, R.D., Demirer, M., Gupta, V., Bigelow, M.T., Prevedello, L.M., Erdal, B.S., 2020. Automated coronary artery atherosclerosis detection and weakly supervised localization on coronary CT angiography with a deep 3-dimensional convolutional neural network. Comput. Med. Imaging Graph. 83. doi:10.1016/j.compmedimag.2020.101721.
Castro, D.C., Walker, I., Glocker, B., 2020. Causality matters in medical imaging. Nat. Commun. 11, 1–10. doi:10.1038/s41467-020-17478-w.
Ceschin, R., Zahner, A., Reynolds, W., Gaesser, J., Zuccoli, G., Lo, C.W., Gopalakrishnan, V., Panigrahy, A., 2018. A computational framework for the detection of subcortical brain dysmaturation in neonatal MRI using 3D convolutional neural networks. Neuroimage 178, 183–197.
Chakraborty, S., Aich, S., Kim, H.C., 2020. Detection of Parkinson's disease from 3T t1 weighted MRI scans using 3D convolutional neural network. Diagnostics 10, 402.
Chan, L., Hosseini, M.S., Rowsell, C., Plataniotis, K.N., Damaskinos, S., 2019. Histosegnet: Semantic segmentation of histological tissue type in whole slide images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10662–10671.
Chang, G.H., Felson, D.T., Qiu, S., Guermazi, A., Capellini, T.D., Kolachalama, V.B., 2020. Assessment of knee pain from MR imaging using a convolutional Siamese network. Eur. Radiol. 1–11.
Chattopadhyay, A., Manupriya, P., Sarkar, A., Balasubramanian, V.N., 2019. Neural network attributions: A causal perspective. In: International Conference on Machine Learning. PMLR, pp. 981–990.
Chen, B., Li, J., Lu, G., Zhang, D., 2019. Lesion location attention guided network for multi-label thoracic disease classification in chest X-rays. IEEE J. Biomed. Health Inform. 24, 2016–2027.
Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., Su, J.K., 2019. This looks like that: deep learning for interpretable image recognition, in: Wallach, H., Larochelle, H., Beygelzimer, A., d\textquotesingle Alché-Buc, F., Fox, E., Garnett, R. (Eds.), Advances in Neural Information Processing Systems 32. Curran Associates, Inc., pp. 8930–8941.
Chen, P., Shi, X., Liang, Y., Li, Y., Yang, L., Gader, P.D., 2020. Interactive thyroid whole slide image diagnostic system using deep representation. Comput. Methods Programs Biomed. 195. doi:10.1016/j.cmpb.2020.105630.
Chen, X., Lin, L., Liang, D., Hu, H., Zhang, Q., Iwamoto, Y., Han, X.H., Chen, Y.W., Tong, R., Wu, J., 2019. A dual-attention dilated residual network for liver lesion classification and localization on CT images. In: Proceedings of the IEEE International Conference on Image Processing (ICIP), pp. 235–239.
Cheng, C.T., Ho, T.Y., Lee, T.Y., Chang, C.C., Chou, C.C., Chen, C.C., Chung, I.F., Liao, C.H., 2019. Application of a deep learning algorithm for detection and visualization of hip fractures on plain pelvic radiographs. Eur. Radiol. 29, 5469–5477.
Cheplygina, V., de Bruijne, M., Pluim, J.P.W., 2019. Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. Med. Image Anal. 54, 280–296. doi:10.1016/j.media.2019.03.009.
Choi, H., Kim, Y.K., Yoon, E.J., Lee, J.Y., Lee, D.S., 2020. 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. Imaging 47, 403–412.
Choudhary, A., Wu, H., Tong, L., Wang, M.D., 2019. Learning to evaluate color similarity for histopathology images using triplet networks. In: Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, pp. 466–474.
Clough, J.R., Oksuz, I., Puyol-Antón, E., Ruijsink, B., King, A.P., Schnabel, J.A., 2019. Global and local interpretability for cardiac MRI classification. In: Proceedings of the 22nd International Conference on Medical Image Computing and Computer-Assisted Intervention MICCAI 2019 doi:10.1007/978-3-030-32251-9_72.
Codella, N.C.F., Lin, C.C., Halpern, A., Hind, M., Feris, R., Smith, J.R., 2018. Collaborative human-AI (CHAI): Evidence-based interpretable melanoma classification in dermoscopic images. In: Proceedings of the 1st International Workshop on Machine Learning in Clinical Neuroimaging, MLCN 2018 doi:10.1007/978-3-030-02628-8_11, 1st Int. Work. Deep Learn. Fail. DLF 2018, 1st Int. Work. Interpret. Mach. Intell. Med. Image Comput. iMIMIC.
Cong, C., Kato, Y., Vasconcellos, H.D., Lima, J., Venkatesh, B., 2019. Automated Stenosis Detection and Classification in X-ray Angiography Using Deep Neural Network. In: Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 1301–1308.
Cook, R.D., Weisberg, S., 1980. Characterizations of an empirical influence function for detecting influential cases in regression. Technometrics 22, 495. doi:10.2307/1268187.
Costa, P., Araujo, T., Aresta, G., Galdran, A., Mendonca, A.M., Smailagic, A., Campilho, A., 2019. EyeWeS: weakly supervised pre-trained convolutional neural networks for diabetic retinopathy detection. In: Proceedings of the 16th International Conference on Machine Vision Applications, MVA 2019. Portugal doi:10.23919/MVA.2019.8757991, Institute of Electrical and Electronics Engineers Inc., INESC TEC.
Dang, S., Chaudhury, S., 2019. Novel relative relevance score for estimating brain connectivity from fMRI data using an explainable neural network approach. J. Neurosci. Methods 326. doi:10.1016/j.jneumeth.2019.108371.
de Vos, B.D., Wolterink, J.M., Leiner, T., de Jong, P.A., Lessmann, N., Isgum, I., 2019. Direct automatic coronary calcium scoring in cardiac and chest CT. IEEE Trans. Med. Imaging 38, 2127–2138. doi:10.1109/TMI.2019.2899534.
DeGrave, A.J., Janizek, J.D., Lee, S.I., 2021. AI for radiographic COVID-19 detection selects shortcuts over signal. Nat. Mach. Intell. 3, 610–619.
Dietterich, T.G., Lathrop, R.H., Lozano-Pérez, T., 1997. Solving the multiple instance problem with axis-parallel rectangles. Artif. Intell. 89, 31–71. doi:10.1016/s0004-3702(96)00034-3.
Doshi-Velez, F., Kim, B., 2017. Towards a rigorous science of interpretable machine learning arXiv preprint arXiv:1702.08608.
Dubost, F., Adams, H., Bortsova, G., Ikram, M.A., Niessen, W., Vernooij, M., de Bruijne, M., 2019a. 3D regression neural network for the quantification of enlarged perivascular spaces in brain MRI. Med. Image Anal. 51, 89–100.
Dubost, F., Adams, H., Yilmaz, P., Bortsova, G., van Tulder, G., Ikram, M.A., Niessen, W., Vernooij, M.W., de Bruijne, M., 2020. Weakly supervised object detection with 2D and 3D regression neural networks. Med. Image Anal. 65, 101767.
Dubost, F., Yilmaz, P., Adams, H., Bortsova, G., Ikram, M.A., Niessen, W., Vernooij, M., De Bruijne, M., 2019b. Enlarged perivascular spaces in brain MRI: Automated quantification in four regions. Neuroimage 185, 534–544.
Dunnmon, J.A., Yi, D., Langlotz, C.P., Ré, C., Rubin, D.L., Lungren, M.P., 2019. Assessment of convolutional neural networks for automated classification of chest radiographs. Radiology 290, 537–544.
Eitel, F., Ritter, K., 2019. Testing the robustness of attribution methods for convolutional neural networks in MRI-based Alzheimer's disease classification. In: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Springer, pp. 3–11. doi:10.1007/978-3-030-33850-3_1.
Eitel, F., Soehler, E., Bellmann-Strobl, J., Brandt, A.U., Ruprecht, K., Giess, R.M., Kuchling, J., Asseyer, S., Weygandt, M., Haynes, J.D., Scheel, M., Paul, F., Ritter, K., 2019. Uncovering convolutional neural network decisions for diagnosing multiple sclerosis on conventional MRI using layer-wise relevance propagation. NeuroImage Clin. 24. doi:10.1016/j.nicl.2019.102003.
El Adoui, M., Drisis, S., Benjelloun, M., 2020. Multi-input deep learning architecture for predicting breast tumor response to chemotherapy using quantitative MR images. Int. J. Comput. Assist. Radiol. Surg. 15, 1491–1500.
Everson, M., Herrera, L.C.G.P., Li, W., Luengo, I.M., Ahmad, O., Banks, M., Magee, C., Alzoubaidi, D., Hsu, H.M., Graham, D., Vercauteren, T., Lovat, L., Ourselin, S., Kashin, S., Wang, H.P., Wang, W.L., Haidry, R.J., 2019. Artificial intelligence for the real-time classification of intrapapillary capillary loop patterns in the endoscopic diagnosis of early oesophageal squamous cell carcinoma: A proof-of-concept study. United Eur. Gastroenterol. J. 7, 297–306. doi:10.1177/2050640618821800.
Fan, F.L., Xiong, J., Li, M., Wang, G., 2021. On interpretability of artificial neural networks: a survey. IEEE Trans. Radiat. Plasma Med. Sci. 5, 741–760.
Fong, R.C., Vedaldi, A., 2017. Interpretable Explanations of Black Boxes by Meaningful Perturbation. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV).
Fuchigami, T., Akahori, S., Okatani, T., Li, Y., 2020. A hyperacute stroke segmentation method using 3D U-Net integrated with physicians' knowledge for NCCT. In: Medical Imaging 2020: Computer-Aided Diagnosis, 11314. International Society for Optics and Photonics.
Gao, K., Shen, H., Liu, Y., Zeng, L., Hu, D., 2019. Dense-CAM: visualize the gender of brains with MRI images. In: Proceedings of the International Joint Conference on Neural Networks (IJCNN), pp. 1–7.
Gao, Y., Zhang, Y., Wang, H., Guo, X., Zhang, J., 2019. Decoding behavior tasks from brain activity using deep transfer learning. IEEE Access 7, 43222–43232. doi:10.1109/ACCESS.2019.2907040.
García-Peraza-Herrera, L.C., Everson, M., Lovat, L., Wang, H.P., Wang, W.L., Haidry, R., Stoyanov, D., Ourselin, S., Vercauteren, T., 2020. Intrapapillary capillary loop classification in magnification endoscopy: open dataset and baseline methodology. Int. J. Comput. Assist. Radiol. Surg. 15, 651–659. doi:10.1007/s11548-020-02127-w.
Gasimova, A., 2019. Automated enriched medical concept generation for chest X-ray images. In: Proceedings of the Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support: Second International Workshop, iMIMIC 2019, and 9th International Workshop, ML-CDS , Held in Conjunction with MICCAI doi:10.1007/978-3-030-33850-3_10.
Gecer, B., Aksoy, S., Mercan, E., Shapiro, L.G., Weaver, D.L., Elmore, J.G., 2018. Detection and classification of cancer in whole slide breast histopathology images using deep convolutional networks. Pattern Recognit. 84, 345–356.
Gessert, N., Latus, S., Abdelwahed, Y.S., Leistner, D.M., Lutz, M., Schlaefer, A., 2019. Bioresorbable scaffold visualization in IVOCT images using CNNs and weakly supervised localization. In: Medical Imaging 2019: Image Processing, 10949. SPIE, pp. 606–612.
Graziani, M., Andrearczyk, V.S.M.M., Müller, H., 2020. Concept attribution: explaining CNN decisions to physicians. Comput. Biol. Med. 123. doi:10.1016/j.compbiomed.2020.103865.
Grigorescu, I., Cordero-Grande, L., David Edwards, A., Hajnal, J.V, Modat, M., Deprez, M., 2019. Investigating image registration impact on preterm birth classification: An interpretable deep learning approach. In: Proceedings of the

1st International Working Smart Ultrasound Imaging, SUSI 2019 doi:10.1007/978-3-030-32875-7_12, 4th Int. Work. Preterm, Perinat. Paediatr. Image Anal. PIPPI 2019, held conjunction with 22nd Int. Conf. Med. Imaging Comput.

Guo, H., Kruger, M., Wang, G., Kalra, M.K., Yan, P., 2020. Multi-task learning for mortality prediction in LDCT images. Med. Imag. Comput. Aided Diagn., 113142C.

Gupta, M., Das, C., Roy, A., Gupta, P., Pillai, G.R., Patole, K., 2020. Region of interest identification for cervical cancer images. In: Proceedings of the IEEE 17th International Symposium on Biomedical Imaging (ISBI), pp. 1293–1296.

Gupta, V., Demirer, M., Bigelow, M., Sarah, M.Y., Joseph, S.Y., Prevedello, L.M., White, R.D., Erdal, B.S., 2020. Using transfer learning and class activation maps supporting detection and localization of femoral fractures on anteroposterior radiographs. In: Proceedings of the IEEE 17th International Symposium on Biomedical Imaging (ISBI), pp. 1526–1529.

GV, K.K., Reddy, G.M., 2019. Automatic classification of whole slide pap smear images using CNN with PCA based feature interpretation. In: Proceedings of the CVPR Workshops, pp. 1074–1079.

Hägele, M., Seegerer, P., Lapuschkin, S., Bockmayr, M., Samek, W., Klauschen, F., Müller, K.R., Binder, A., 2020. Resolving challenges in deep learning-based analyses of histopathological images using explanation methods. Sci. Rep. 10. doi:10.1038/s41598-020-62724-2.

He, J., Shang, L., Ji, H., Zhang, X., 2017. Deep learning features for lung adenocarcinoma classification with tissue pathology images. In: Proceedings of the International Conference on Neural Information Processing, pp. 742–751.

Heinemann, F., Birk, G., Stierstorfer, B., 2019. Deep learning enables pathologist-like scoring of NASH models. Sci. Rep. 9, 1–10.

Hilbert, A., Ramos, L.A., van Os, H.J.A., Olabarriaga, S.D., Tolhuisen, M.L., Wermer, M.J.H., Barros, R.S., van der Schaaf, I., Dippel, D., Roos, Y., et al., 2019. Data–efficient deep learning of radiological image data for outcome prediction after endovascular treatment of patients with acute ischemic stroke. Comput. Biol. Med. 115, 103516.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9, 1735–1780.

Hoffer, E., Ailon, N., 2015. Deep metric learning using triplet network. In: Proceedings of the International Workshop on Similarity-Based Pattern Recognition, pp. 84–92.

Hosny, A., Parmar, C., Coroller, T.P., Grossmann, P., Zeleznik, R., Kumar, A., Bussink, J., Gillies, R.J., Mak, R.H., Aerts, H.J.W.L., 2018. Deep learning for lung cancer prognostication: a retrospective multi-cohort radiomics study. PLoS Med. 15, e1002711.

Huang, Y., Chung, A.C.S., 2019. Evidence localization for pathology images using weakly supervised learning. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 613–621.

Huang, Z., Fu, D., 2019. Diagnose chest pathology in X-ray images by learning multi-attention convolutional neural network. In: Proceedings of the IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC), pp. 294–299.

Huang, Z., Zhu, X., Ding, M., Zhang, X., 2020. Medical image classification using a light-weighted hybrid neural network based on PCANet and DenseNet. IEEE Access. 8, 24697–24712.

Huff, D.T., Weisman, A.J., Jeraj, R., 2021. Interpretation and visualization techniques for deep learning models in medical imaging. Phys. Med. Biol. 66, 04TR01.

Humphries, S.M., Notary, A.M., Centeno, J.P., Strand, M.J., Crapo, J.D., Silverman, E.K., Lynch, D.A.of COPD (COPDGene) Investigators, G.E., 2020. Deep learning enables automatic classification of emphysema pattern at CT. Radiology 294, 434–444.

Huo, Y., Terry, J.G., Wang, J., Nath, V., Bermudez, C., Bao, S., Parvathaneni, P., Carr, J.J., Landman, B.A., 2019. Coronary calcium detection using 3D attention identical dual deep network based on weakly supervised learning. In: Proceedings of the Medical Imaging Image Processing.

Itoh, H., Lu, Z., Mori, Y., Misawa, M., Oda, M., Kudo, S.E., Mori, K., 2020. Visualising decision-reasoning regions in computer-aided pathological pattern diagnosis of endoscytoscopic images based on CNN weights analysis. In: HK, H., MA, M. (Eds.), Medical Imaging 2020: Computer-Aided Diagnosis. SPIE, Graduate School of Informatics, Nagoya University, Furo-cho, Chikusa-ku, Nagoya, Aichi, 464-8601, Japan doi:10.1117/12.2549532.

Jamaludin, A., Kadir, T., Zisserman, A., 2017. SpineNet: automated classification and evidence visualization in spinal MRIs. Med. Image Anal. 41, 63–73.

Jang, Y., Son, J., Park, K.H., Park, S.J., Jung, K.H., 2018. Laterality classification of fundus images using interpretable deep neural network. J. Digit. Imaging 31, 923–928. doi:10.1007/s10278-018-0099-2.

Jetley, S., Lord, N.A., Lee, N., Torr, P., 2018. Learn to Pay Attention. Proceeding of the International Conference on Learning Representations.

Ji, J., 2019. Gradient-based Interpretation on Convolutional Neural Network for Classification of Pathological Images. In: Proceeding of the International Conference on Information Technology and Computer Application, ITCA, pp. 83–86. doi:10.1109/ITCA49981.2019.00026 Institute of Electrical and Electronics Engineers Inc., No.2 High School of East China Normal University, Shanghai, China.

Jia, X., Ren, L., Cai, J., 2020. Clinical implementation of AI technologies will require interpretable AI models. Med. Phys. 47, 1–4. doi:10.1002/mp.13891.

Jiang, H., Yang, K., Gao, M., Zhang, D., Ma, H., Qian, W., 2019. An Interpretable Ensemble Deep Learning Model for Diabetic Retinopathy Disease Classification. In: Proceeding of the 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC, pp. 2045–2048. doi:10.1109/EMBC.2019.8857160 2019. Institute of Electrical and Electronics Engineers Inc., Beijing Zhizhen Internet Technology Co., Ltd, China.

Jing, B., Xie, P., Xing, E., 2018. On the Automatic Generation of Medical Imaging Reports. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 2577–2586.

Kashyap, S., Karargyris, A., Wu, J., Gur, Y., Sharma, A., Wong, K.C.L., Moradi, M., Syeda-Mahmood, T., 2020. Looking in the right place for anomalies: explainable Ai through automatic location learning. In: Proceeding of the 17th IEEE International Symposium on Biomedical Imaging, ISBI. I, pp. 1125–1129. doi:10.1109/ISBI45749.2020.9098370 EEE Computer Society, IBM Research Almaden.

Kermany, D.S., Goldbaum, M., Cai, W., Valentim, C.C.S., Liang, H., Baxter, S.L., McKeown, A., Yang, G., Wu, X., Yan, F., Dong, J., Prasadha, M.K., Pei, J., Ting, M., Zhu, J., Li, C., Hewett, S., Dong, J., Ziyar, I., Shi, A., Zhang, R., Zheng, L., Hou, R., Shi, W., Fu, X., Duan, Y., Huu, V.A.N., Wen, C., Zhang, E.D., Zhang, C.L., Li, O., Wang, X., Singer, M.A., Sun, X., Xu, J., Tafreshi, A., Lewis, M.A., Xia, H., Zhang, K., 2018. Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell 172, 1122–1131. doi:10.1016/j.cell.2018.02.010, e9.

Khakzar, A., Albarqouni, S., Navab, N., 2019. Learning interpretable features via adversarially robust optimization. Proceeding of the 22nd International Conference on Medical Image Computing and Computer-Assisted Intervention MICCAI 2019 doi:10.1007/978-3-030-32226-7_88.

Kiani, A., Uyumazturk, B., Rajpurkar, P., Wang, A., Gao, R., Jones, E., Yu, Y., Langlotz, C.P., Ball, R.L., Montine, T.J., et al., 2020. Impact of a deep learning assistant on the histopathologic classification of liver cancer. NPJ Digit. Med. 3, 1–8.

Kim, B.H., Ye, J.C., 2020. Understanding graph isomorphism network for rs-fMRI functional connectivity analysis. Front. Neurosci. 14, 630.

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., Sayres, R., 2018. Interpretability beyond feature attribution: quantitative testing with concept activation vectors (TCAV). In: Proceeding of the 35th International Conference on Machine Learning, ICML 2018. International Machine Learning Society (IMLS), pp. 4186–4195.

Kim, C., Kim, W.H., Kim, H.J., Kim, J., 2020. Weakly-supervised US breast tumor characterization and localization with a box convolution network. Proceeding of the Medical Imaging: Computer-Aided Diagnosis.

Kim, I., Rajaraman, S., Antani, S., 2019. Visual interpretation of convolutional neural network predictions in classifying medical image modalities. Diagnostics 9, 38.

Kim, M., Han, J.C., Hyun, S.H., Janssens, O., Van Hoecke, S., Kee, C., De Neve, W., 2019. Medinoid: computer-aided diagnosis and localization of glaucoma using deep learning. Appl. Sci. 9, 3064.

Kim, S.T., Lee, J.H., Ro, Y.M., K, M., HK, H., 2019a. Visual evidence for interpreting diagnostic decision of deep neural network in computer-aided diagnosis. Proceeding of the Medical Imaging: Computer-Aided Diagnosis SPIE, School of Electrical Engineering, KAIST, Daejeon, 34141, South Korea doi:10.1117/12.2512621.

Kim, Y., Choi, D., Lee, K.J., Kang, Y., Ahn, J.M., Lee, E., Lee, J.W., Kang, H.S., 2020. Ruling out rotator cuff tear in shoulder radiograph series using deep learning: redefining the role of conventional radiograph. Eur. Radiol. 30, 2843–2852.

Kim, Y., Lee, K.J., Sunwoo, L., Choi, D., Nam, C.M., Cho, J., Kim, J., Bae, Y.J., Yoo, R.E., Choi, B.S., et al., 2019b. Deep learning in diagnosis of maxillary sinusitis using conventional radiography. Invest. Radiol. 54, 7–15.

Ko, H., Chung, H., Kang, W.S., Kim, K.W., Shin, Y., Kang, S.J., Lee, J.H., Kim, Y.J., Kim, N.Y., Jung, H., et al., 2020. 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, e19569.

Koitka, S., Kim, M.S., Qu, M., Fischer, A., Friedrich, C.M., Nensa, F., 2020. Mimicking the radiologists' workflow: estimating pediatric hand bone age with stacked deep neural networks. Med. Image Anal. 64. doi:10.1016/j.media.2020.101743.

Korbar, B., Olofson, A.M., Miraflor, A.P., Nicka, C.M., Suriawinata, M.A., Torresani, L., Suriawinata, A.A., Hassanpour, S., 2017. Looking under the hood: deep neural network visualization to interpret whole-slide image analysis outcomes for colorectal polyps. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 69–75.

Kowsari, K., Sali, R., Ehsan, L., Adorno, W., Ali, A., Moore, S., Amadi, B., Kelly, P., Syed, S., Brown, D., 2020. HMIC: hierarchical medical image classification, a deep learning approach. Information 11. doi:10.3390/INFO11060318.

Kubach, J., Muhlebner-Fahrngruber, A., Soylemezoglu, F., Miyata, H., Niehusmann, P., Honavar, M., Rogerio, F., Kim, S.H., Aronica, E., Garbelli, R., Vilz, S., Popp, A., Walcher, S., Neuner, C., Scholz, M., Kuerten, S., Schropp, V., Roeder, S., Eichhorn, P., Eckstein, M., Brehmer, A., Kobow, K., Coras, R., Blumcke, I., Jabari, S., 2020. Same same but different: A Web-based deep learning application revealed classifying features for the histopathologic distinction of cortical malformations. Epilepsia 61, 421–432. doi:10.1111/epi.16447.

Kumar, D., Sankar, V., Clausi, D., Taylor, G.W., Wong, A., 2019a. SISC: end-to-end interpretable discovery radiomics-driven lung cancer prediction via stacked interpretable sequencing cells. IEEE Access 7, 145444–145454. doi:10.1109/ACCESS.2019.2945524.

Kumar, D., Taylor, G.W., Wong, A., 2019b. Discovery radiomics with CLEAR-DR: interpretable computer aided diagnosis of diabetic retinopathy. IEEE Access 7, 25891–25896. doi:10.1109/ACCESS.2019.2893635.

LaLonde, R., Torigian, D., Bagci, U., 2020. Encoding visual attributes in capsules for explainable medical diagnoses. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 294–304.

Langner, T., Wikström, J., Bjerner, T., Ahlström, H., Kullberg, J., 2019. Identifying morphological indicators of aging with neural networks on large-scale whole-body MRI. IEEE Trans. Med. Imaging 39, 1430–1437.

Lee, H., Kim, S.T., Ro, Y.M., 2019a. Generation of multimodal justification using visual word constraint model for explainable computer-aided diagnosis. In: Proceedings of the 2nd International Workshop on Interpretability of Machine Intelligence in Medical Image Computing, iMIMIC 2019, and the 9th International Workshop on Multimodal Learning for Clinical Decision Support, ML-CDS 2019, held in conjunction with the 22nd International Conference on Medical Imaging and Computer-Assisted Intervention, MICCAI 2019 doi:10.1007/978-3-030-33850-3_3, held conjunction with 22nd Interna.

Lee, H., Yune, S., Mansouri, M., Kim, M., Tajmir, S.H., Guerrier, C.E., Ebert, S.A., Pomerantz, S.R., Romero, J.M., Kamalian, S., Gonzalez, R.G., Lev, M.H., Do, S., 2019b. An explainable deep-learning algorithm for the detection of acute intracranial haemorrhage from small datasets. Nat. Biomed. Eng. 3, 173–182. doi:10.1038/s41551-018-0324-9.

Lee, J., Nishikawa, R.M., 2019. Detecting mammographically occult cancer in women with dense breasts using deep convolutional neural network and radon cumulative distribution transform. J. Med. Imaging 6, 44502.

Lee, Jeong Hoon, Ha, E.J., Kim, D., Jung, Y.J., Heo, S., Jang, Y.H., An, S.H., Lee, K., 2020. Application of deep learning to the diagnosis of cervical lymph node metastasis from thyroid cancer with CT: external validation and clinical utility for resident training. Eur. Radiol. 3066–3072.

Lee, Jeong Hyun, Joo, I., Kang, T.W., Paik, Y.H., Sinn, D.H., Ha, S.Y., Kim, K., Choi, C., Lee, G., Yi, J., et al., 2020. Deep learning with ultrasonography: automated classification of liver fibrosis using a deep convolutional neural network. Eur. Radiol. 30, 1264–1273.

Lei, Y., Tian, Y., Shan, H., Zhang, J., Wang, G., Kalra, M.K., 2020. Shape and margin-aware lung nodule classification in low-dose CT images via soft activation mapping. Med. Image Anal. 60, 101628.

Lenis, D., Major, D., Wimmer, M., Berg, A., Sluiter, G., Bühler, K., 2020. Domain aware medical image classifier interpretation by counterfactual impact analysis. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 315–325.

Li, C.Y., Liang, X., Hu, Z., Xing, E.P., 2019. Knowledge-driven encode, retrieve, paraphrase for medical image report generation. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 6666–6673.

Li, L., Xu, M., Liu, H., Li, Y., Wang, X., Jiang, L., Wang, Z., Fan, X., Wang, N., 2019a. A large-scale database and a CNN model for attention-based glaucoma detection. IEEE Trans. Med. Imaging 39, 413–424.

Li, M., Kuang, K., Zhu, Q., Chen, X., Guo, Q., Wu, F., 2020. IB-M: A Flexible Framework to Align an Interpretable Model and a Black-box Model. In: Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 643–649.

Li, Q., Xing, X., Sun, Y., Xiao, B., Wei, H., Huo, Q., Zhang, M., Zhou, X.S., Zhan, Y., Xue, Z., et al., 2019b. Novel iterative attention focusing strategy for joint pathology localization and prediction of MCI progression. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 307–315.

Li, W., Zhuang, J., Wang, R., Zhang, J., Zheng, W.S., 2020. Fusing metadata and dermoscopy images for skin disease diagnosis. In: Proceedings of the IEEE 17th International Symposium on Biomedical Imaging (ISBI), pp. 1996–2000.

Li, X., Wu, J., Chen, E.Z., Jiang, H., 2019c. From deep learning towards finding skin lesion biomarkers. In: Proceedings of the 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 2797–2800.

Li, Y., Shafipour, R., Mateos, G., Zhang, Z., 2019d. Mapping brain structural connectivities to functional networks via graph encoder-decoder with interpretable latent embeddings. In: Proceedings of the 7th IEEE Global Conference on Signal and Information Processing, GlobalSIP. Rochester, United States doi:10.1109/GlobalSIP45357.2019.8969239, 2019. Institute of Electrical and Electronics Engineers Inc., University of Rochester, Dept. of Electrical and Computer Engineering.

Li, Z., Wang, C., Han, M., Xue, Y., Wei, W., Li, L.J., Fei-Fei, L., 2019d. Thoracic disease identification and localization with limited supervision. Adv. Comput. Vis. Pattern Recognit. doi:10.1007/978-3-030-13969-8_7.

Lian, C., Liu, M., Wang, L., Shen, D., 2019. End-to-end dementia status prediction from brain mri using multi-task weakly-supervised attention network. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 158–167.

Liao, L., Zhang, X., Zhao, F., Lou, J., Wang, L., Xu, X., Zhang, H., Li, G., 2020. Multi--branch deformable convolutional neural network with label distribution learning for fetal brain age prediction. In: Proceedings of the IEEE 17th International Symposium on Biomedical Imaging (ISBI), pp. 424–427.

Liao, W., Zou, B., Zhao, R., Chen, Y., He, Z., Zhou, M., 2019. Clinical interpretable deep learning model for glaucoma diagnosis. IEEE J. Biomed. Health Informat. doi:10.1109/jbhi.2019.2949075.

Lin, C.Y., 2004. Rouge: A package for automatic evaluation of summaries. Text summarization branches out, pp. 74–81.

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

Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., van der Laak, J.A.W.M., van Ginneken, B., Sánchez, C.I., 2017. A survey on deep learning in medical image analysis. Med. Image Anal. doi:10.1016/j.media.2017.07.005.

Liu, C., Han, X., Li, Z., Ha, J., Peng, G., Meng, W., He, M., 2019. A self-adaptive deep learning method for automated eye laterality detection based on color fundus photography. PLoS One 14. doi:10.1371/journal.pone.0222025.

Liu, H., Wang, L., Nan, Y., Jin, F., Wang, Q., Pu, J., 2019f. SDFN: Segmentation-based deep fusion network for thoracic disease classification in chest X-ray images. Comput. Med. Imaging Graph. 75, 66–73.

Lundberg, S.M., Lee, S.I., 2017. A unified approach to interpreting model predictions. In: Proceedings of the Advances in Neural Information Processing Systems.

Luo, L., Chen, H., Wang, X., Dou, Q., Lin, H., Zhou, J., Li, G., Heng, P.A., 2019. Deep angular embedding and feature correlation attention for breast MRI cancer analysis. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 504–512.

Ma, K., Wu, K., Cheng, H., Gu, C., Xu, R., Guan, X., 2018. A pathology image diagnosis network with visual interpretability and structured diagnostic report. Proceedings of the 25th International Conference on Neural Information Processing ICONIP 2018 doi:10.1007/978-3-030-04224-0_24.

Mahmud, T., Rahman, M.A., Fattah, S.A., 2020. 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, 103869.

Maicas, G., Bradley, A.P., Nascimento, J.C., Reid, I., Carneiro, G., 2019. Pre and post-hoc diagnosis and interpretation of malignancy from breast DCE-MRI. Med. Image Anal. 58, 101562.

Maksoud, S., Wiliem, A., Zhao, K., Zhang, T., Wu, L., Lovell, B., 2019. CORAL8: concurrent object regression for area localization in medical image panels. In: Proceedings of the 22nd International Conference Medical Image Computing and Computer Assisted Intervention MICCAI 2019 doi:10.1007/978-3-030-32239-7_48.

Malhi, A., Kampik, T., Pannu, H., Madhikermi, M., Framling, K., 2019. Explaining machine learning-based classifications of *in-vivo* gastral images. In: Proceedings of the International Conference on Digital Image Computing: Techniques and Applications, DICTA doi:10.1109/DICTA47822.2019.8945986, 2019. Institute of Electrical and Electronics Engineers Inc., Department of Computer Science, Aalto University Finland, Finland.

Martins, J., Cardoso, J.S., Soares, F., 2020. Offline computer-aided diagnosis for Glaucoma detection using fundus images targeted at mobile devices. Comput. Methods Programs Biomed. 192. doi:10.1016/j.cmpb.2020.105341.

Matsui, Y., Maruyama, T., Nitta, M., Saito, T., Tsuzuki, S., Tamura, M., Kusuda, K., Fukuya, Y., Asano, H., Kawamata, T., Masamune, K., Muragaki, Y., 2020. Prediction of lower-grade glioma molecular subtypes using deep learning. J. Neurooncol. 146, 321–327. doi:10.1007/s11060-019-03376-9.

Meijering, E., 2020. A bird's-eye view of deep learning in bioimage analysis. Comput. Struct. Biotechnol. J. doi:10.1016/j.csbj.2020.08.003.

Meng, Q., Hashimoto, Y., Satoh, S., 2020. How to extract more information with less burden: Fundus image classification and retinal disease localization with ophthalmologist intervention. IEEE J. Biomed. Health Inform. 24, 3351–3361.

Meng, Q., Sinclair, M., Zimmer, V., Hou, B., Rajchl, M., Toussaint, N., Oktay, O., Schlemper, J., Gomez, A., Housden, J., et al., 2019. Weakly supervised estimation of shadow confidence maps in fetal ultrasound imaging. IEEE Trans. Med. Imaging 38, 2755–2767.

Murdoch, W.J., Singh, C., Kumbier, K., Abbasi-Asl, R., Yu, B., 2019. Definitions, methods, and applications in interpretable machine learning. Proc. Natl. Acad. Sci. USA 116, 22071–22080. doi:10.1073/pnas.1900654116.

Narayanan, B.N., Hardie, R.C., De Silva, M.S., Kueterman, N.K., 2020. Hybrid machine learning architecture for automated detection and grading of retinal images for diabetic retinopathy. J. Med. Imaging 7, 34501.

Natekar, P., Kori, A., Krishnamurthi, G., 2020. Demystifying brain tumor segmentation networks: interpretability and uncertainty analysis. Front. Comput. Neurosci. 14. doi:10.3389/fncom.2020.00006.

Ng, H.G., Kerzel, M., Mehnert, J., May, A., Wermter, S., 2018. Classification of MRI migraine medical data using 3D convolutional neural network. In: Proceedings of the International Conference on Artificial Neural Networks, pp. 300–309.

Nunes, N., Martins, B., André da Silva, N., Leite, F., J Silva, M., 2019. A multimodal deep learning method for classifying chest radiology exams. In: Proceedings of the Conference on Artificial Intelligence EPIA 2019 doi:10.1007/978-3-030-30241-2_28.

Obikane, S., Aoki, Y., 2020. Weakly Supervised domain adaptation with point supervision in histopathological image segmentation. In: Proceedings of the 5th Asian Conference on Pattern Recognition, ACPR 2019, pp. 127–140.

Olah, C., Mordvintsev, A., Schubert, L., 2017. Feature visualization. Distill 2, e7.

Olden, J.D., Joy, M.K., Death, R.G., 2004. An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. Ecol. Modell. 178, 389–397. doi:10.1016/j.ecolmodel.2004.03.013.

Papanastasopoulos, Z., Samala, R.K., Chan, H.P., Hadjiiski, L., Paramagul, C., Helvie, M.A., Neal, C.H., 2020. Explainable AI for medical imaging: Deep-learning CNN ensemble for classification of estrogen receptor status from breast MRI. In: HK, H., MA, M. (Eds.), Medical Imaging 2020: Computer-Aided Diagnosis. SPIE. University of Michigan, 1500 E. Medical Center Drive, Ann Arbor, MI 48109-5842, United States doi:10.1117/12.2549298.

Papineni, K., Roukos, S., Ward, T., Zhu, W.J., 2002. BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 311–318.

Patra, A., Noble, J.A., 2020. Incremental Learning of Fetal Heart Anatomies Using Interpretable Saliency Maps. In: Proceedings of the 23rd Conference Medical Image Underst. Anal. MIUA 2019 doi:10.1007/978-3-030-39343-4_11.

Paul, H.Y., Kim, T.K., Alice, C.Y., Bennett, B., Eng, J., Lin, C.T., 2020. Can AI outperform a junior resident? Comparison of deep neural network to first-year radiology residents for identification of pneumothorax. Emerg. Radiol. 27, 367–375.

Paul, H.Y., Kim, T.K., Wei, J., Shin, J., Hui, F.K., Sair, H.I., Hager, G.D., Fritz, J., 2019. Automated semantic labeling of pediatric musculoskeletal radiographs using deep learning. Pediatr. Radiol. 49, 1066–1070.

Paul, R., Schabath, M., Gillies, R., Hall, L., Goldgof, D., 2020. Convolutional neural network ensembles for accurate lung nodule malignancy prediction 2 years in the future. Comput. Biol. Med. 122, 103882.
Pearl, J., 2009. Causality. Cambridge University Press.
Pelka, O., Nensa, F., Friedrich, C.M., 2019. Variations on branding with text occurrence for optimized body parts classification. In: Proceedings of the 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 890–894.
Peng, T., Boxberg, M., Weichert, W., Navab, N., Marr, C., 2019. Multi-task learning of a deep K-nearest neighbour network for histopathological image classification and retrieval. In: Proceedings of the 22nd International Conference on Medical Image Computing and Computer-Assisted Intervention MICCAI 2019 doi:10.1007/978-3-030-32239-7_75.
Pennington, J., Socher, R., Manning, C.D., 2014. Glove: global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543.
Perdomo, O., Rios, H., Rodríguez, F.J., Otálora, S., Meriaudeau, F., Müller, H., González, F.A., 2019. Classification of diabetes-related retinal diseases using a deep learning approach in optical coherence tomography. Comput. Methods Programs Biomed. 178, 181–189. doi:10.1016/j.cmpb.2019.06.016.
Pereira, S., Meier, R., Alves, V., Reyes, M., Silva, C.A., 2018. Automatic brain tumor grading from MRI data using convolutional neural networks and quality assessment. In: Understanding and Interpreting Machine Learning in Medical Image Computing Applications. Springer, pp. 106–114.
Pesce, E., Joseph Withey, S., Ypsilantis, P.P., Bakewell, R., Goh, V., Montana, G., 2019. Learning to detect chest radiographs containing pulmonary lesions using visual attention networks. Med. Image Anal. 53, 26–38. doi:10.1016/j.media.2018.12.007.
Philbrick, K.A., Yoshida, K., Inoue, D., Akkus, Z., Kline, T.L., Weston, A.D., Korfiatis, P., Takahashi, N., Erickson, B.J., 2018. What does deep learning see? Insights from a classifier trained to predict contrast enhancement phase from CT images. Am. J. Roentgenol. 211, 1184–1193. doi:10.2214/AJR.18.20331.
Pominova, M., Artemov, A., Sharaev, M., Kondrateva, E., Bernstein, A., Burnaev, E., 2018. Voxelwise 3d convolutional and recurrent neural networks for epilepsy and depression diagnostics from structural and functional MRI data. In: Proceedings of the 2018 IEEE International Conference on Data Mining Workshops (ICDMW), pp. 299–307.
Qi, X., Zhang, L., Chen, Yao, Pi, Y., Chen, Yi, Lv, Q., Yi, Z., 2019. Automated diagnosis of breast ultrasonography images using deep neural networks. Med. Image Anal. 52, 185–198.
Qin, R., Wang, Z., Jiang, L., Qiao, K., Hai, J., Chen, J., Xu, J., Shi, D., Yan, B., 2020. Fine-grained lung cancer classification from PET and CT images based on multidimensional attention mechanism. Complexity 2020.
Quellec, G., Lamard, M., Conze, P.H., Massin, P., Cochener, B., 2020. Automatic detection of rare pathologies in fundus photographs using few-shot learning. Med. Image Anal. 61, 101660.
Radiological Society of North America, 2018. Pneumonia detection challenge. https://www.rsna.org/education/ai-resources-and-training/ai-image-challenge/rsna-pneumonia-detection-challenge-2018
Rajaraman, S., Candemir, S., Thoma, G., Antani, S., 2019. Visualizing and explaining deep learning predictions for pneumonia detection in pediatric chest radiographs. Med. Imaging 2019. Comput. Aided Diagn., 109500S.
Rajpurkar, P., Irvin, J., Ball, R.L., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C.P., et al., 2018. Deep learning for chest radiograph diagnosis: a retrospective comparison of the CheXNeXt algorithm to practicing radiologists. PLoS Med. 15, e1002686.
Rajpurkar, P., O'Connell, C., Schechter, A., Asnani, N., Li, J., Kiani, A., Ball, R.L., Mendelson, M., Maartens, G., van Hoving, D.J.others, 2020a. CheXaid: deep learning assistance for physician diagnosis of tuberculosis using chest x-rays in patients with HIV. NPJ Digit. Med. 3, 1–8.
Rajpurkar, P., Park, A., Irvin, J., Chute, C., Bereket, M., Mastrodicasa, D., Langlotz, C.P., Lungren, M.P., Ng, A.Y., Patel, B.N., 2020b. AppendiXNet: deep learning for diagnosis of appendicitis from a small dataset of CT exams using video pretraining. Sci. Rep. 10, 1–7.
Reyes, M., Meier, R., Pereira, S., Silva, C.A., Dahlweid, F.M., Tengg-Kobligk, H.V., Summers, R.M., Wiest, R., 2020. On the interpretability of artificial intelligence in radiology: challenges and opportunities. Radiol. Artif. Intell. 2, e190043.
Rezaei, M., Uemura, T., Näppi, J., Yoshida, H., Lippert, C., Meinel, C., 2020. Generative synthetic adversarial network for internal bias correction and handling class imbalance problem in medical image diagnosis. Medical Imaging 2020. Comput. Aided Diagn., 113140E.
Ribeiro, M.T., Singh, S., Guestrin, C., 2016. Why should i trust you?" Explaining the predictions of any classifier. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Association for Computing Machinery, New York, New York, USA, pp. 1135–1144. doi:10.1145/2939672.2939778.
Robnik-Šikonja, M., Kononenko, I., 2008. Explaining classifications for individual instances. IEEE Trans. Knowl. Data Eng. 20, 589–600. doi:10.1109/TKDE.2007.190734.
Rodin, I., Fedulova, I., Shelmanov, A., Dylov, D.V, 2019. Multitask and multimodal neural network model for interpretable analysis of X-ray images. In: Proceedings of the 019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 1601–1604.
Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Springer Verlag, pp. 234–241. doi:10.1007/978-3-319-24574-4_28.
Rudin, C., 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat. Mach. Intell. 1, 206–215. doi:10.1038/s42256-019-0048-x.
Saab, K., Dunnmon, J., Goldman, R., Ratner, A., Sagreiya, H., Ré, C., Rubin, D., 2019. Doubly weak supervision of deep learning models for head CT. In: Proceedings of the 22nd Int. Medical Image Computing and Computer Assisted Intervention MICCAI 2019 doi:10.1007/978-3-030-32248-9_90.
Sabour, S., Frosst, N., Hinton, G.E., 2017. Dynamic routing between capsules. Advances in neural information processing systems, p. 30.
Sarhan, M.H., Eslami, A., Navab, N., Albarqouni, S., 2019. Learning interpretable disentangled representations using adversarial VAEs. In: Proceedings of the 1st MICCAI Domain Adaptation and Representation Transfer and Medical Image Learning with Less Labels and Imperfect Data doi:10.1007/978-3-030-33391-1_5, 1st Int. Work. Med. Image Learn. with Less Labels Imperfect Data, MIL3ID 2019, held conjunction with 22nd Int. Conf. Med..
Schlemper, J., Oktay, O., Schaap, M., Heinrich, M., Kainz, B., Glocker, B., Rueckert, D., 2019. Attention gated networks: learning to leverage salient regions in medical images. Med. Image Anal. 53, 197–207. doi:10.1016/j.media.2019.01.012.
Schwab, E., Goossen, A., Deshpande, H., Saalbach, A., 2020. Localization of critical findings in chest X-ray without local annotations using multi-instance learning. In: Proceedings of the 17th IEEE International Symposium on Biomedical Imaging, ISBI 2020. IEEE Computer Society, Clinical Informatics, Solutions Services, Philips Research North America, Cambridge, MA, United States, pp. 1879–1882. doi:10.1109/ISBI45749.2020.9098551.
Sedai, S., Mahapatra, D., Ge, Z., Chakravorty, R., Garnavi, R., 2018. Deep multiscale convolutional feature learning for weakly supervised localization of chest pathologies in x-ray images. In: International Workshop on Machine Learning in Medical Imaging. Springer, Cham, pp. 267–275.
Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision, pp. 618–626.
Seo, D., Oh, K., Oh, I.S., 2020. In: Regional Multi-Scale Approach for Visually Pleasing Explanations of Deep Neural Networks, 8. IEEE Access, pp. 8572–8582. doi:10.1109/ACCESS.2019.2963055.
Shahamat, H., Saniee Abadeh, M., 2020. Brain MRI analysis using a deep learning based evolutionary approach. Neural Netw. 126, 218–234. doi:10.1016/j.neunet.2020.03.017.
Shapira, N., Fokuhl, J., Schultheiß, M., Beck, S., Kopp, F.K., Pfeiffer, D., Dangelmaier, J., Pahn, G., Sauter, A.P., Renger, B., et al., 2020. Liver lesion localisation and classification with convolutional neural networks: a comparison between conventional and spectral computed tomography. Biomed. Phys. Eng. Express 6, 15038.
Shapley, L.S., 2016. A value for n-person games. In: Contributions to the Theory of Games (AM-28), Volume II, 17. Princeton University Press, pp. 307–318. doi:10.1515/9781400881970-018.
Shen, D., Wu, G., Suk, H.I., 2017. Deep learning in medical image analysis. Annu. Rev. Biomed. Eng. 19, 221–248. doi:10.1146/annurev-bioeng-071516-044442.
Shen, S., Han, S.X., Aberle, D.R., Bui, A.A., Hsu, W., 2019. An interpretable deep hierarchical semantic convolutional neural network for lung nodule malignancy classification. Expert Syst. Appl. 128, 84–95. doi:10.1016/j.eswa.2019.01.048.
Shen, Y., Sheng, B., Fang, R., Li, H., Dai, L., Stolte, S., Qin, J., Jia, W., Shen, D., 2020. Domain-invariant interpretable fundus image quality assessment. Med. Image Anal. 61. doi:10.1016/j.media.2020.101654.
Shinde, S., Chougule, T., Saini, J., Ingalhalikar, M., 2019a. HR-CAM: Precise localization of pathology using multi-level learning in CNNS. In: Proceedings of the 22nd International Conference on Medical Image Computing and Computer-Assisted Intervention ; MICCAI 2019 doi:10.1007/978-3-030-32251-9_33.
Shinde, S., Prasad, S., Saboo, Y., Kaushick, R., Saini, J., Pal, P.K., Ingalhalikar, M., 2019b. Predictive markers for Parkinson's disease using deep neural nets on neuromelanin sensitive MRI. NeuroImage Clin. 22, 101748.
Silva-Rodríguez, J., Colomer, A., Sales, M.A., Molina, R., Naranjo, V., 2020. Going deeper through the Gleason scoring scale: an automatic end-to-end system for histology prostate grading and cribriform pattern detection. Comput. Methods Programs Biomed. 195, 105637.
Silva, W., Fernandes, K., Cardoso, M.J., Cardoso, J.S., 2018. Towards complementary explanations using deep neural networks. In: Understanding and Interpreting Machine Learning in Medical Image Computing Applications. Springer, pp. 133–140.
Silva, W., Poellinger, A., Cardoso, J.S., Reyes, M., 2020. Interpretability-guided content-based medical image retrieval. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 305–314.
Simonyan, K., Vedaldi, A., Zisserman, A., 2013. Deep inside convolutional networks: visualising image classification models and saliency maps. In: Proceedings of the 2nd International Conference on Learning Representations ICLR 2014 - Workshop Track Proceedings.
Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. In: Proceedings of the International Conference on Learning Representations, pp. 1–14.
Singh, S., Karimi, S., Ho-Shon, K., Hamey, L., 2019. From chest X-rays to radiology reports: a multimodal machine learning approach. 2019 International Conference on Digital Image Computing: Techniques and Applications, DICTA 2019. Institute of Electrical and Electronics Engineers Inc., Department of Computing, Macquarie University, Sydney, Australia doi:10.1109/DICTA47822.2019.8945819.

Singla, S., Gong, M., Ravanbakhsh, S., Sciurba, F., Poczos, B., Batmanghelich, K.N., 2018. Subject2Vec: generative-discriminative approach from a set of image patches to a vector. 21st International Conference on Medical Image Computing and Computer-Assisted Intervention MICCAI 2018 doi:10.1007/978-3-030-00928-1_57.

Society for Imaging Informatics in Medicine, American College of Radiology, 2019. Pneumothorax segmentation [WWW Document] URL.

Sønderby, C.K., Raiko, T., Maaløe, L., Sønderby, S.K., Winther, O., 2016. Ladder variational autoencoders, in: Lee, D.D., Sugiyama, M., Luxburg, U. V, Guyon, I., Garnett, R. (Eds.), Advances in Neural Information Processing Systems 29. Curran Associates, Inc., pp. 3738–3746.

Spinks, G., Moens, M.F., 2019. Justifying diagnosis decisions by deep neural networks. J. Biomed. Inform. 96. doi:10.1016/j.jbi.2019.103248.

Springenberg, J.T., Dosovitskiy, A., Brox, T., Riedmiller, M., 2014. Striving for simplicity: the all convolutional net. In: Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings.

Sun, H., Zeng, X., Xu, T., Peng, G., Ma, Y., 2020. Computer-Aided Diagnosis in Histopathological Images of the Endometrium Using a Convolutional Neural Network and Attention Mechanisms. IEEE J. Biomed. Health Inform. 24, 1664–1676. doi:10.1109/JBHI.2019.2944977.

Sun, L., Wang, W., Li, J., Lin, J., 2019. Study on medical image report generation based on improved encoding-decoding method. In: Proceedings of the International Conference on Intelligent Computing, pp. 686–696.

Tang, C., 2020. Discovering Unknown Diseases with Explainable Automated Medical Imaging. In: Proceedings of the Annual Conference on Medical Image Understanding and Analysis, pp. 346–358.

Tang, R., Tushar, F.I., Han, S., Hou, R., Rubin, G.D., Lo, J.Y., 2019. Classification of chest CT using case-level weak supervision. Medical Imaging 2019: Computer-Aided Diagnosis, 1095017.

Tang, Y.X., Tang, Y.B., Peng, Y., Yan, K., Bagheri, M., Redd, B.A., Brandon, C.J., Lu, Z., Han, M., Xiao, J., et al., 2020. Automated abnormality classification of chest radiographs using deep convolutional neural networks. NPJ Digit. Med. 3, 1–8.

Tang, Z., Chuang, K.V, DeCarli, C., Jin, L.W., Beckett, L., Keiser, M.J., Dugger, B.N., 2019. Interpretable classification of Alzheimer's disease pathologies with a convolutional neural network pipeline. Nat. Commun. 10. doi:10.1038/s41467-019-10212-1.

Teramoto, A., Yamada, A., Kiriyama, Y., Tsukamoto, T., Yan, K., Zhang, L., Imaizumi, K., Saito, K., Fujita, H., 2019. Automated classification of benign and malignant cells from lung cytological images using deep convolutional neural network. Inform. Med. Unlocked 16, 100205.

Thakoor, K.A., Li, X., Tsamis, E., Sajda, P., Hood, D.C., 2019. Enhancing the accuracy of glaucoma detection from OCT probability maps using convolutional neural networks. In: Proceedings of the 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 2036–2040.

Tian, J., Li, C., Shi, Z., Xu, F., 2018. A diagnostic report generator from CT volumes on liver tumor with semi-supervised attention mechanism. Proceedings of the 21st International Conference on Medical Image Computing and Computer-Assisted Intervention MICCAI 2018 doi:10.1007/978-3-030-00934-2_78.

Tian, J., Zhong, C., Shi, Z., Xu, F., 2019. Towards automatic diagnosis from multi–modal medical data. In: Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support. Springer, pp. 67–74.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. 58, 267–288.

Tsang, M., Cheng, D., Liu, Y., 2018. Detecting statistical interactions from neural network weights. In: Proceedings of the International Conference on Learning Representations.

Tu, Z., Gao, S., Zhou, K., Chen, X., Fu, H., Gu, Z., Cheng, J., Yu, Z., Liu, J., 2020. SUNet: a lesion regularized model for simultaneous diabetic retinopathy and diabetic macular edema grading. In: Proceedings of the 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), pp. 1378–1382.

Uehara, K., Murakawa, M., Nosato, H., Sakanashi, H., 2019. Prototype-based interpretation of pathological image analysis by convolutional neural networks. In: Proceedings of the Asian Conference on Pattern Recognition, pp. 640–652.

Upadhyay, U., Banerjee, B., 2020. Compact representation learning using class specific convolution coders-application to medical image classification. In: Proceedings of the 020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), pp. 1266–1270.

Uzunova, H., Ehrhardt, J., Kepp, T., Handels, H., 2019. Interpretable explanations of black box classifiers applied on medical images by meaningful perturbations using variational autoencoders. In: Medical Imaging 2019: Image Processing, 10949. SPIE, pp. 264–271.

van Amsterdam, W.A.C., Verhoeff, J.J.C., de Jong, P.A., Leiner, T., Eijkemans, M.J.C., 2019. Eliminating biasing signals in lung cancer images for prognosis predictions with deep learning. npj Digit. Med. 2, 1–6. doi:10.1038/s41746-019-0194-x.

van de Schoot, R., de Bruin, J., Schram, R., Zahedi, P., de Boer, J., Weijdema, F., Kramer, B., Huijts, M., Hoogerwerf, M., Ferdinands, G., et al., 2021. An open source machine learning framework for efficient and transparent systematic reviews. Nat. Mach. Intell. 3, 125–133.

van der Velden, B.H.M., Janse, M.H.A., Ragusi, M.A.A., Loo, C.E., Gilhuijs, K.G.A., 2020. Volumetric breast density estimation on MRI using explainable deep learning regression. Sci. Rep. 10. doi:10.1038/s41598-020-75167-6.

van Sloun, R.J.G., Demi, L., 2019. Localizing B-lines in lung ultrasonography by weakly supervised deep learning, *in-vivo* results. IEEE J. Biomed. Heal. Inform. 24, 957–964.

Vedantam, R., Lawrence Zitnick, C., Parikh, D., 2015. Cider: consensus-based image description evaluation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4566–4575.

Vila-Blanco, N., Carreira, M.J., Varas-Quintana, P., Balsa-Castro, C., Tomas, I., 2020. Deep neural networks for chronological age estimation from OPG images. IEEE Trans. Med. Imaging 39, 2374–2384.

Vinyals, O., Toshev, A., Bengio, S., Erhan, D., 2015. Show and tell: a neural image caption generator. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3156–3164.

von Schacky, C.E., Sohn, J.H., Liu, F., Ozhinsky, E., Jungmann, P.M., Nardo, L., Posadzy, M., Foreman, S.C., Nevitt, M.C., Link, T.M., et al., 2020. Development and validation of a multitask deep learning model for severity grading of hip osteoarthritis features on radiographs. Radiology 295, 136–145.

Wang, C.J., Hamm, C.A., Savic, L.J., Ferrante, M., Schobert, I., Schlachter, T., Lin, M.D., Weinreb, J.C., Duncan, J.S., Chapiro, J., Letzen, B., 2019. Deep learning for liver tumor diagnosis part II: convolutional neural network interpretation using radiologic imaging features. Eur. Radiol. 29, 3348–3357. doi:10.1007/s00330-019-06214-8.

Wang, H., Feng, J., Zhang, Z., Su, H., Cui, L., He, H., Liu, L., 2018. Breast mass classification via deeply integrating the contextual information from multi-view data. Pattern Recognit 80, 42–52. doi:10.1016/j.patcog.2018.02.026.

Wang, J., Cui, Y., Shi, G., Zhao, J., Yang, X., Qiang, Y., Du, Q., Ma, Y., Kazihise, N.G.F., 2020. Multi-branch cross attention model for prediction of KRAS mutation in rectal cancer with t2-weighted MRI. Appl. Intell. 50, 2352–2369.

Wang, J., Zhang, R., Wei, X., Li, X., Yu, M., Zhu, J., Gao, J., Liu, Z., Yu, R., 2019. An attention-based semi-supervised neural network for thyroid nodules segmentation. In: Proceedings of the 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 871–876.

Wang, K., Zhang, X., Huang, S., 2019. KGZNet: Knowledge-guided deep zoom neural networks for thoracic disease classification. In: Proceedings of the 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 1396–1401.

Wang, L., Zhang, L., Zhu, M., Qi, X., Yi, Z., 2020. Automatic diagnosis for thyroid nodules in ultrasound images by deep neural networks. Med. Image Anal. 61, 101665.

Wang, R., Fan, D., Lv, B., Wang, M., Zhou, Q., Lv, C., Xie, G., Wang, L., 2020a. OCT image quality evaluation based on deep and shallow features fusion network. In: Proceedings of the 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), pp. 1561–1564.

Wang, S., Xing, Y., Zhang, L., Gao, H., Zhang, H., 2019a. Deep convolutional neural network for ulcer recognition in wireless capsule endoscopy: experimental feasibility and optimization. Comput. Math. Methods Med. 2019.

Wang, Xi, Chen, H., Ran, A.R., Luo, L., Chan, P.P., Tham, C.C., Chang, R.T., Mannil, S.S., Cheung, C.Y., Heng, P.A., 2020b. Towards multi-center glaucoma OCT image screening with semi-supervised joint structure and function multi-task learning. Med. Image Anal. 63, 101695.

Wang, X, Liang, X., Jiang, Z., Nguchu, B.A., Zhou, Y., Wang, Y., Wang, H., Li, Y., Zhu, Y., Wu, F., Gao, J.H., Qiu, B., 2020c. Decoding and mapping task states of the human brain via deep learning. Hum. Brain Mapp 41, 1505–1519. doi:10.1002/hbm.24891.

Wang, X., Peng, Y., Lu, L., Lu, Z., Summers, R.M., 2018. TieNet: text-image embedding network for common thorax disease classification and reporting in chest X-rays. In: Proceedings of the 31st Meeting of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2018. IEEE Computer Society, United States, pp. 9049–9058. doi:10.1109/CVPR.2018.00943 Department of Radiology and Imaging Sciences, Clinical Center.

Wang, X., Xu, M., Li, L., Wang, Z., Guan, Z., 2019b. Pathology-aware deep network visualization and its application in glaucoma image synthesis. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 423–431.

Wang, X, Zhang, Y., Guo, Z., Li, J., 2019c. A computational framework towards medical image explanation. In: Proceedings of the 7th Joint International Workshop on Knowledge Representation for Health Care Process doi:10.1007/978-3-030-37446-4_10, Inf. Syst. Heal. Care, KR4HC/ProHealth 2019 1st Work. Transparent, Explain. Affect. AI Med. Syst. TEAAM 2019 held conjuncti.

Koh, P.W., Liang, P., 2017. Understanding black-box predictions via influence functions. In: International conference on machine learning. PMLR, pp. 1885–1894.

Wei, W., Poirion, E., Bodini, B., Durrleman, S., Ayache, N., Stankoff, B., Colliot, O., 2019. Predicting PET-derived demyelination from multimodal MRI using sketcher-refiner adversarial training for multiple sclerosis. Med. Image Anal. 58, 101546.

Wickstrøm, K., Kampffmeyer, M., Jenssen, R., 2020. Uncertainty and interpretability in convolutional neural networks for semantic segmentation of colorectal polyps. Med. Image Anal. 60, 101619.

Windisch, P., Weber, P., Fürweger, C., Ehret, F., Kufeld, M., Zwahlen, D., Muacevic, A., 2020. Implementation of model explainability for a basic brain tumor detection using convolutional neural networks on MRI slices. Neuroradiology doi:10.1007/s00234-020-02465-1.

Woerl, A.C., Eckstein, M., Geiger, J., Wagner, D.C., Daher, T., Stenzel, P., Fernandez, A., Hartmann, A., Wand, M., Roth, W., et al., 2020. Deep learning predicts molecular subtype of muscle-invasive bladder cancer from conventional histopathological slides. Eur. Urol. 78, 256–264.

Wu, B., Zhou, Z., Wang, J., Wang, Y., 2018. Joint learning for pulmonary nodule segmentation, attributes and malignancy prediction. In: Proceedings of the 15th IEEE International Symposium on Biomedical Imaging, ISBI 2018. IEEE Computer Society, China, pp. 1109–1113. doi:10.1109/ISBI.2018.8363765 Nat'l Engineering Laboratory for Video Technology Cooperative Medianet Innovation Center, Key Laboratory of Machine Perception (MoE) Sch'l of EECS, Peking University, Beijing, 100871.

Xi, P., Guan, H., Shu, C., Borgeat, L., Goubran, R., 2019. An integrated approach for medical abnormality detection using deep patch convolutional neural networks. Vis. Comput. 1–14.

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

Xie, Y., Zhang, J., Xia, Y., Shen, C., 2020. A mutual bootstrapping model for automated skin lesion segmentation and classification. IEEE Trans. Med. Imaging 39, 2482–2493.

Xu, H., Dong, M., Lee, M.H., O'Hara, N., Asano, E., Jeong, J.W., 2019. Objective detection of eloquent axonal pathways to minimize postoperative deficits in pediatric epilepsy surgery using diffusion tractography and convolutional neural networks. IEEE Trans. Med. Imaging 38, 1910–1922. doi:10.1109/TMI.2019.2902073.

Xu, R., Cong, Z., Ye, X., Hirano, Y., Kido, S., Gyobu, T., Kawata, Y., Honda, O., Tomiyama, N., 2019. Pulmonary textures classification via a multi-scale attention network. IEEE J. Biomed. Heal. Inform. 24, 2041–2052.

Yan, C., Xu, J., Xie, J., Cai, C., Lu, H., 2020. Prior-Aware CNN with multi-task learning for colon images analysis. In: Proceedings of the 17th IEEE International Symposium on Biomedical Imaging, ISBI 2020. IEEE Computer Society, Nanjing University of Information Science Technology, Nanjing, China, pp. 254–257. doi:10.1109/ISBI45749.2020.9098703.

Yan, K., Peng, Y., Sandfort, V., Bagheri, M., Lu, Z., Summers, R.M., 2019. Holistic and comprehensive annotation of clinically significant findings on diverse CT images: Learning from radiology reports and label ontology. In: Proceedings of the 32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2019. IEEE Computer Society, United States, pp. 8515–8524. doi:10.1109/CVPR.2019.00872 Imaging Biomarkers and Computer-Aided Diagnosis Laboratory, Clinical Center, National Institutes of Health, Bethesda, MD 20892.

Yan, K., Wang, X., Lu, L., Zhang, L., Harrison, A.P., Bagheri, M., Summers, R.M., 2018. Deep lesion graphs in the wild: relationship learning and organization of significant radiology image findings in a diverse large-scale lesion database. In: Proceedings of the 31st Meeting of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2018. IEEE Computer Society, United States, pp. 9261–9270. doi:10.1109/CVPR.2018.00965 Imaging Biomarkers and Computer-Aided Diagnosis Laboratory, National Institutes of Health Clinical Center, 10 Center Drive, Bethesda, MD 20892.

Yan, Y., Kawahara, J., Hamarneh, G., 2019. Melanoma Recognition via Visual Attention. In: Proceedings of the 26th International Conference on Information Processing in Medical Imaging, IPMI 2019 doi:10.1007/978-3-030-20351-1_62.

Yang, H., Kim, J.Y., Kim, H., Adhikari, S.P., 2019. Guided soft attention network for classification of breast cancer histopathology images. IEEE Trans. Med. Imaging 39, 1306–1315.

Yang, P., Zhai, Y., Li, L., Lv, H., Wang, J., Zhu, C., Jiang, R., 2020. A deep metric learning approach for histopathological image retrieval. Methods 179, 14–25.

Yang, S., Niu, J., Wu, J., Liu, X., 2020. Automatic medical image report generation with multi-view and multi-modal attention mechanism. In: Proceedings of the International Conference on Algorithms and Architectures for Parallel Processing, pp. 687–699.

Yang, X., Wang, Z., Liu, C., Le, H.M., Chen, J., Cheng, K.T.T., Wang, L., 2017. Joint detection and diagnosis of prostate cancer in multi-parametric MRI based on multimodal convolutional neural networks. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 426–434.

Ye, H., Gao, F., Yin, Y., Guo, D., Zhao, P., Lu, Y., Wang, X., Bai, J., Cao, K., Song, Q., et al., 2019. Precise diagnosis of intracranial hemorrhage and subtypes using a three-dimensional joint convolutional and recurrent neural network. Eur. Radiol. 29, 6191–6201.

Yi, P.H., Lin, A., Wei, J., Yu, A.C., Sair, H.I., Hui, F.K., Hager, G.D., Harvey, S.C., 2019. Deep-learning-based semantic labeling for 2D mammography and comparison of complexity for machine learning tasks. J. Digit. Imaging 32, 565–570. doi:10.1007/s10278-019-00244-w.

Yin, C., Qian, B., Wei, J., Li, X., Zhang, X., Li, Y., Zheng, Q., 2019. Automatic generation of medical imaging diagnostic report with hierarchical recurrent neural network. In: Proceedings of the IEEE International Conference on Data Mining (ICDM), pp. 728–737.

Young, K., Booth, G., Simpson, B., Dutton, R., Shrapnel, S., 2019. Deep neural network or dermatologist? In: Proceedings of the 2nd International Workshop on Interpretability of Machine Intelligence in Medical Image Computing, IMIMIC 2019 doi:10.1007/978-3-030-33850-3_6, 9th Int. Work. Multimodal Learn. Clin. Decis. Support. ML-CDS 2019, held conjunction with 22nd Interna.

Yuan, J., Liao, H., Luo, R., Luo, J., 2019. Automatic radiology report generation based on multi-view image fusion and medical concept enrichment. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 721–729.

Zeiler, M.D., Fergus, R., 2014. Visualizing and understanding convolutional networks. In: Lecture Notes in Computer Science. Springer Verlag, pp. 818–833. doi:10.1007/978-3-319-10590-1_53 (Including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics).

Zeng, X., Wen, L., Xu, Y., Ji, C., 2020. Generating diagnostic report for medical image by high-middle-level visual information incorporation on double deep learning models. Comput. Methods Progr. Biomed. 197, 105700.

Zhang, B., Tan, J., Cho, K., Chang, G., Deniz, C.M., 2020. Attention-based cnn for kl grade classification: data from the osteoarthritis initiative. In: Proceedings of the IEEE 17th International Symposium on Biomedical Imaging (ISBI), pp. 731–735.

Zhang, R., Tan, S., Wang, R., Manivannan, S., Chen, J., Lin, H., Zheng, W.S., 2019. Biomarker localization by combining CNN classifier and generative adversarial network. In: Proceedings of the 22nd International Conference on Medical Image Computing and Computer-Assisted Intervention. MICCAI 2019 doi:10.1007/978-3-030-32239-7_24.

Zhang, Y., Ding, D.Y., Qian, T., Manning, C.D., Langlotz, C.P., 2018. Learning to summarize radiology findings. In: Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis. Association for Computational Linguistics, Brussels, Belgium, pp. 204–213. doi:10.18653/v1/W18-5623.

Zhang, Z., Chen, P., Sapkota, M., Yang, L., 2017a. TandemNet: distilling knowledge from medical images using diagnostic reports as optional semantic references. Proceedings of the 20th International Conference on Medical Image Computing and Computer-Assisted Intervention. MICCAI 2017 doi:10.1007/978-3-319-66179-7_37.

Zhang, Z., Xie, Y., Xing, F., McGough, M., Yang, L., 2017b. MDNet: a semantically and visually interpretable medical image diagnosis network. In: Proceedings of the30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017. Institute of Electrical and Electronics Engineers Inc., University of Florida, United States, pp. 3549–3557. doi:10.1109/CVPR.2017.378.

Zhao, C., Han, J., Jia, Y., Fan, L., Gou, F., 2018. Versatile framework for medical image processing and analysis with application to automatic bone age assessment. Journal of Electrical and Computer Engineering 2018.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A., 2016. Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2921–2929.

Zhou, K., Gao, S., Cheng, J., Gu, Z., Fu, H., Tu, Z., Yang, J., Zhao, Y., Liu, J., 2020. Sparse-gan: Sparsity-constrained generative adversarial network for anomaly detection in retinal oct image. In: Proceedings of the IEEE 17th International Symposium on Biomedical Imaging (ISBI), pp. 1227–1231.

Zhou, L.Q., Wu, X.L., Huang, S.Y., Wu, G.G., Ye, H.R., Wei, Q., Bao, L.Y., Deng, Y.B., Li, X.R., Cui, X.W., et al., 2020. Lymph node metastasis prediction from primary breast cancer US images using deep learning. Radiology 294, 19–28.

Zhu, P., Ogino, M., 2019. Guideline-based additive explanation for computer-aided diagnosis of lung nodules. In: Proceedings of the 2nd International Workshop on Interpretability of Machine Intelligence in Medical Image Computing, IMIMIC 2019 doi:10.1007/978-3-030-33850-3_5, 9th Int. Work. Multimodal Learn. Clin. Decis. Support. ML-CDS 2019, held conjunction with 22nd Interna.

Zhu, Z., Albadawy, E., Saha, A., Zhang, J., Harowicz, M.R., Mazurowski, M.A., 2019. Deep learning for identifying radiogenomic associations in breast cancer. Comput. Biol. Med. 109, 85–90. doi:10.1016/j.compbiomed.2019.04.018.

Zhu, Z., Ding, X., Zhang, D., Wang, L., 2020. Weakly-supervised balanced attention network for gastric pathology image localization and classification. In: Proceedings of the IEEE 17th International Symposium on Biomedical Imaging (ISBI), pp. 1–4.

Zintgraf, L.M., Cohen, T.S., Adel, T., Welling, M., 2017. Visualizing deep neural network decisions: prediction difference analysis. In: Proceedings of the 5th International Conference on Learning Representations, ICLR 2017. University of Amsterdam, Netherlands International Conference on Learning Representations, ICLR.

Zunair, H., Hamza, A.B., 2020. Melanoma detection using adversarial training and deep transfer learning. Phys. Med. Biol. 65, 135005.