applied sciences

MDPI

Article

Explainable-AI in Automated Medical Report Generation Using Chest X-ray Images

Saad Bin Ahmed *, Roberto Solis-Oba and Lucian Ilie

Department of Computer Science, Middlesex College, Western University, London, ON N6A 5B7, Canada
* Correspondence: sahme532@uwo.ca

Abstract: The use of machine learning in healthcare has the potential to revolutionize virtually every aspect of the industry. However, the lack of transparency in AI applications may lead to the problem of trustworthiness and reliability of the information provided by these applications. Medical practitioners rely on such systems for clinical decision making, but without adequate explanations, diagnosis made by these systems cannot be completely trusted. Explainability in Artificial Intelligence (XAI) aims to improve our understanding of why a given output has been produced by an AI system. Automated medical report generation is one area that would benefit greatly from XAI. This survey provides an extensive literature review on XAI techniques used in medical image analysis and automated medical report generation. We present a systematic classification of XAI techniques used in this field, highlighting the most important features of each one that could be used by future research to select the most appropriate XAI technique to create understandable and reliable explanations for decisions made by AI systems. In addition to providing an overview of the state of the art in this area, we identify some of the most important issues that need to be addressed and on which research should be focused.

Keywords: explainable AI; class activation map; linear interpretable model-agnostic; layer-wise relevance propagation

check for updates

Citation: Ahmed, S.B.; Solis-Oba, R.; Ilie, L. Explainable-AI in Automated Medical Report Generation Using Chest X-ray Images. *Appl. Sci.* **2022**, *12*, 11750. https://doi.org/10.3390/app122211750

Academic Editor: Valentin Sebastian Schäfer

Received: 24 September 2022
Accepted: 2 November 2022
Published: 18 November 2022

1. Introduction

Machine learning models are powerful tools for the solution of complex problems. However many of these models are so complex that we do not really understand how they work [1,2]. Hence, for machine learning models that are used in critical applications, such as medical diagnosis, it is essential that the models provide a transparent and clear explanation of how they reached a particular decision. Without an explanation of how a model works, it cannot be trusted [1].

Explainability in artificial intelligence (XAI) is a relatively new field that aims at providing explanations on how an AI model works and how it makes its decisions. Incorporating explanations into an AI model does not directly try to improve the performance of the model, but rather to give insights into how and why a model produces a particular output. Understanding an AI model includes knowledge of the role of each parameter of the model, which factors affect the model's output and how the model's parameters and input influence the output [3].

XAI is fundamental for the development and adoption of AI prediction systems for healthcare and other critical applications, as it provides the necessary elements for transforming a mysterious, incomprehensibly complex black box system into a trustworthy and efficient tool. This paper summarizes some of the most relevant work in the field of XAI, specifically in relation to deep-neural-network-based models for chest X-ray image analysis and automated medical report generation. Medical imaging techniques are extensively used for diagnosing illnesses [4–7]. However, each image needs to be carefully examined by an experienced healthcare professional, who then needs to spend time writing a report explaining their findings. This specialized and time-consuming activity can delay or even prevent timely treatment for some patients [8].

Automating the analysis of medical images and writing the corresponding medical reports would alleviate these problems and free some time for healthcare professionals, who can then focus on treatment and patient care. There has been alot of interest recently in the development of AI models for image analysis and automatic medical report generation [9,10].

Image analysis is an essential part of the automated medical report generation process since it helps to extract meaning from an image that can then be reported through text describing the image contents.

Convolutional neural networks (ConvNets) have been demonstrated to be an effective machine learning technique for image analysis especially in hyperspectral image classification and channel reduction [11,12] and textual image analysis [13]. ConvNets consist of neurons grouped into interconnected layers [14], as further discussed in Section 2.

ConvNets analyze images and extract useful information from them that can be fed to a natural language generator to produce a description of their contents understandable to humans [10,15]. This process falls under the umbrella of image captioning.

The use of XAI in healthcare is still a relatively new notion, with plenty of room for improvement of existing techniques or introduction of new ones. In the health sciences, the explanations given by deep learning models are very important as they aid clinicians in understanding diagnoses and making decisions [16,17].

Medical practitioners are concerned about the health of their patients, so they wish to utilize AI-assisted solutions with confidence. AI systems that perform effectively are desirable and must be trustworthy.

This survey's main contribution is to provide a comparative examination of XAI approaches used in medical image analysis and automated medical report generation. There is extensive research on XAI techniques for medical image analysis, but less work has been conducted on XAI for automated medical report generation. Automated medical report generation helps diagnose patients and reduces the amount of work for doctors. Such systems require the development of sophisticated and complex applications that make sensitive decisions, and therefore, we require explainable AI models that justify the output produced by those systems. Consequently, this study explores XAI methods that could be extended to systems for automating the generation of medical reports.

We think that XAI could be used to increase the accuracy of automated medical report generation from medical images, as reported accuracies of existing methods are quite low [9,18,19]. If AI models are created that integrate explainable AI, researchers will be able to focus their efforts on certain characteristics or factors of a model that may help to enhance their accuracy.

The contributions of this paper are the following.

1. We categorize and organize current research on the design and use of XAI techniques to generate explanations for AI models used in the analysis of X-ray images and in the automatic generation of medical reports.
2. This high-level overview of research on XAI in this field helps identify some of the most important issues with existing XAI models and highlight the importance of collaborative efforts between clinicians, practitioners and system designers. We hope this paper will help focus researchers on these issues and eventually lead to the creation of effective, accurate, efficient and highly understandable and reliable AI systems for healthcare.

The rest of this paper is organized into four broad categories. The first part provides some background information. The second part focuses on explanation approaches for medical chest X-ray image analysis. The third part describes explanation approaches proposed for automated medical report generation. The last part discusses the pros and cons of available XAI methods and in which circumstances a particular XAI method would be the most suitable.

2. Background

Medical images provide extensive information that can be used to diagnose diseases and track the progress of patients. Chest X-ray images necessitate a thorough examination by radiologists, who then document the results of their analyses in full-text reports. To generate accurate reports, radiologists must have expertise in diagnosing medical images. Nonetheless, many reports do not provide a conclusive diagnostic due to the large number of potential diagnoses. Furthermore, the amount of time it takes radiologists to prepare full-text reports is an issue of concern. In modern-day hospitals, automated medical imaging techniques are commonly employed to help alleviate these problems.

Convolutional-Neural-Network-based systems are commonly used for medical image analysis. ConvNets are powerful feature extractors able to discover relevant information in images without the need for human intervention.

2.1. Convolutional Neural Networks

Artificial neural networks (ANNs), or just neural networks, imitate the behavior of the human brain by allowing computers to learn how to recognise relevant components of a problem and how these components interact; this information aids in the solution of difficult problems. An ANN consists of nodes or neurons grouped into layers and connections between them; neuron connections have associated weights.

Neurons receive information from other neurons or from external sources, process the information and pass it to other neurons or external sources through their connections.

A ConvNet is an ANN that can extract features from an input image, classify them and identify patterns (see Figure 1).

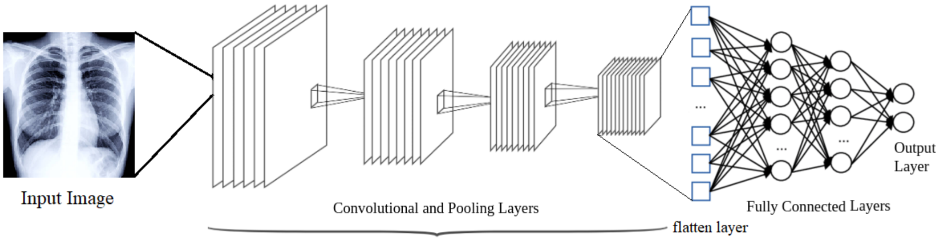

Figure 1. Architecture of Convolutional Neural Networks.

A ConvNet is a deep learning network consisting of the following kinds of neuron layers:

- **Input Layer:** The input layer reads the image that will be processed.
- **Convolutional Layers:** Convolutional layers process the input image to extract features from it. A convolutional layer applies a set of linear *kernels* or *filters* to its input to produce a set of so-called *convolved features*. A non-linear activation function is applied to the convolved features to produce the output of the convolutional layer, called a *feature* or *activation map*. Filters are designed to detect changes in an image's intensity values to recognise spatial patterns such as edges. The more convolutional layers a ConvNet has, the more complex spatial features a ConvNet can detect in an image. Some XAI techniques are aimed at visualizing feature maps to gain a better understanding of the image features that a ConvNet discovers and uses to reach conclusions about what an image represents.
- **Pooling Layers:** Pooling is a technique for reducing the size of feature maps to speed up computation. In the pooling layer, the feature maps produced from the previous layer are down-sampled so that new feature maps with a condensed resolution can be generated. The input spatial dimension is greatly reduced by this layer. The most common types of pooling are max pooling and average pooling in which a group of values from a feature map is replaced with the maximum or the average value in the group, respectively. In a ConvNet, a convolutional layer is usually followed by a pooling layer.

- **Flatten Layer:** Feature maps are flattened to create a long continuous linear vector from all the 2-Dimensional arrays. A fully connected layer uses the flattened matrix as input for image classification.
- **Fully-connected Layers:** These layers appear at the end of ConvNets. They process the feature maps computed by the other layers to determine relationships between high-level image features.
- **Output Layer:** In a neural network that performs multi-class classification, the output layer consists of a set of scores giving the likelihood of the image belonging to each one of the classes that the ConvNet can identify. The softmax function is commonly used in ConvNets to compute these scores.

Given a large number of medical images depicting different classes of abnormalities, a ConvNet can learn the key characteristics of each class, but it lacks the ability to provide explanations about how particular classes are detected.

When dealing with data of a sensitive nature, such as medical data, we can see the desire to better understand how complex algorithms process it. Without such understanding, it is very difficult to trust the information produced by these algorithms.

2.2. Attention Mechanisms in Deep Learning

Visual selective attention enables us to focus on the most important parts of a scene and allows us to efficiently extract useful information from it. According to cognitive science, the abundance of information restricts the human ability to comprehend it, so we must focus on a small part of it [20]. To study the human visual perception process, researchers have developed models of visual selective attention that simulate the human visual system.

The study of attention mechanisms has made huge advances in the past few years in areas such as natural language processing [21] and image processing [22]; attention mechanisms mimic the perceptual mechanisms in the human brain. Most research combining deep learning algorithms with visual attention mechanisms focus on the use of masks; masking identifies the key features in an image. Deep neural networks can learn about the regions in an image that are most relevant for a specific task and, hence, on which they must focus their attention.

Attention mechanisms in image processing compute an *attention map,* which is the matrix representing the importance that each part of the image has for a particular task. In a ConvNet, the input can be re-weighted with that map before feeding it to the convolutional layers so the ConvNet can focus on the most relevant parts of an image. The input is encoded so that it can be fed to the convolutional and pooling layers. If all the network states used to encode the input are used to produce the attention map, the corresponding attention mechanism is called global; if only some of the states are used, the mechanism is called local attention.

The weights in an attention map are between zero and one with higher weights given to the parts of an image that are more relevant to a ConvNet. If the weights are continuous values between zero and one, the attention mechanism is called *soft attention*; if the weights are only either zero or one, the attention mechanism is called *hard attention*. Soft attention mechanisms are deterministic, while hard attention mechanism are stochastic [23]. Soft attention mechanisms can be further divided into spatial attention [24], channel attention [25] and self attention [26].

Self attention was introduced in transformer-based architectures [27]. Self-attention focuses on a single context, and it is commonly used in NLP tasks [28]. In multi-head attention, multiple attention modules run in parallel, and this allows focus to be simultaneously centered on diverse parts of the input and multiple relationships between the input components to be discovered [26].

The layers of a convolutional neural network process the input image and generate new channels from the input's initial three channels, Each channel contains different information, and a channel attention module assigns weights to these channels reflecting their relevance, so the ConvNet can focus on the channels with higher weights [25]. Convolutional neural networks also use spatial attention modules to identify the most relevant locations of an image [24].

2.3. Explainability in AI

Deep learning models have shown to be reliable, highly effective and accurate in a wide range of research fields, but we do not exactly know how these models make predictions and why specific conclusions are reached; these are concerns that limit our trust in them. We can think of deep learning models as black boxes that receive input and produce output, but we do not understand the complex processes that happen inside. We would like to have models that are reliable, accurate and transparent, so they can be trusted. There are several reasons why explainability in AI is essential:

1. Enhances understanding of models output. Users of AI models can make informed decisions if they understand how the models work.
2. Reduces the number of errors. Explainability can help spot model anomalies that allow us to design more accurate models. Explainability also helps to learn from mistakes and train the models to prevent them.
3. Provides clarity about models output and strengthens our confidence in them. This is essential to build trust and have AI models adopted and accepted.

XAI explanations can be classified into the following three categories [3,29]: visual explanations, textual explanation and example-based explanations.

1. **Visual Explanations**
   A visual explanation of a medical image is vital to a reliable analysis of the image. Visual explanations that show important parts of an image and can be used to justify decisions are called *saliency* or *heat maps* [5,23,30].
2. **Textual Explanations**
   Text-based explanations provide descriptions of model output. A description may consist of a simple labeling of an image's contents or an entire medical report [19,31].
3. **Example-based Explanations**
   In example-based explanations, examples are provided to help understand why predictions are made by deep learning models. For instance, a previously diagnosed patient who had the same symptoms as a new patient can be used to understand a diagnosis made by the model [32].

XAI methods can be classified into two groups: ante hoc and post hoc. The term post hoc refers to methods that are used to generate explanations for a model's output using the trained model, whereas ante hoc techniques create explanations during the training stage of the model.

XAI methods can also be classified as global or local. A local approach provides explanations for the output produced by a deep learning model for a single, specific instance. Most local approaches are model agnostic, which means that they do not require knowledge of how the deep learning model works. Global approaches aim to explain the logic behind the functioning of a deep learning model, and therefore, these approaches are model-dependent.

3. XAI Approaches for Chest X-ray Image Analysis and Report Generation

A number of techniques have been proposed in XAI to explain how a deep learning model works and how it infers conclusions from the given input. Table 1 summarized details of XAI techniques presented in recent years by considering chest X-rays image analysis. In this section, we describe some of the explanation techniques that have been used in AI models that analyze chest X-rays.

3.1. Class Activation Mapping

In ConvNet-based image classification models, *class activation maps* (CAM) are used to highlight the most relevant or discriminative portions of an image that are used by a model to identify disorders in chest X-rays. A class activation map serves as a visual explanation of a ConvNet model that can assist radiologists in determining whether the decisions made by the model are based on the processing of the correct features of chest X-ray images. Class activation maps can also assist in the detection of data bias.

Some ConvNet layers preserve spatial information and are capable of detecting different objects in an image, such as bones, organs or tumors. The feature or activation maps of a ConvNet measure the importance of each part of an image in detecting a particular object. To construct a CAM, the average of each feature map of the ConvNet's last layer is computed, and these average values are fed to a fully connected layer to assign a weight to each image feature reflecting its importance in the computation of the output (see Figure 2).

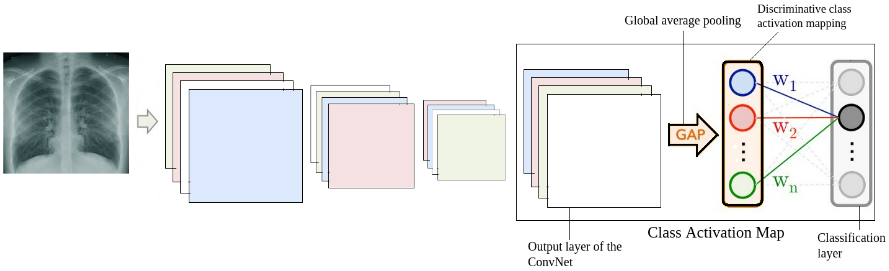

Figure 2. Conceptual view of class activation mapping; w_1, w_2 ... w_n represent the weights of the feature maps. Class activation mapping identifies the most relevant features in the input image used for prediction.

Each feature map is mapped back to the input X-ray by assigning a colour to each region of the X-ray based on the weight assigned to that region by the feature map. Colours are used to highlight the parts of the image that are most significant for detecting the object selected by a feature map. Similarly, the collection of all ConvNet feature maps can be mapped back to the image by using a linear combination of the feature maps using the weights determined by the fully connected layer. The resulting heat map highlights the areas of the input X-rays that are more relevant to the ConvNet's output.

Let $f_k(x, y)$ be the value of the kth feature map for position (x, y) of the input image and w_k be the weight of the kth feature map. The class activation map $M(x, y)$ is defined as,

$$M(x, y) = \sum_k w_k f_k(x, y) \tag{1}$$

Several CAM-based methods have been proposed in the literature to create visual explanations of ConvNet-based models using linear combinations of feature maps. These methods differ in the manner in which the weights for the feature maps are computed. A theoretical study of the best algorithms for computing the values for these weights is presented in [2]. Below, we review some of the recent research on chest X-ray image analysis that uses CAMs and its variations to explain ConvNet models.

A learning system for identification of pneumothorax in chest X-rays using the deep convolutional neural network ResNet-152 [33] is described in [34]. CAM heat map analysis was used to highlight the parts of an X-ray that are most important to the predictions of the model. It is helpful for radiologists to see what parts of an image are the focus of the neural network as this assists them in figuring whether the neural network bases its predictions on the areas of an image that are most relevant to a particular diagnosis.

CheXNet [30] is a deep learning model to detect and locate 14 different diseases on chest X-ray images. The ChestX-ray 14 dataset was used to train a 121-layer densely connected convolutional neural network that has performance comparable to that of experienced radiologists. CheXNet was also used to predict lung cancer from chest X-ray images [5] and for thoracic disease classification [35]. Transfer learning was used twice in [5] to create a more accurate model for lung cancer detection. This application of transfer learning led to the computation of class activation maps that accurately show the most salient locations on the X-rays that the model uses for making predictions.

A method is proposed in [36] to improve understanding of the features that most heavily influence the decisions of neural network classifiers, through the use of adversarial robust optimization. The invariance of a model to perturbations on its inputs is referred to as adversarial robustness. Feature understanding and interpretability is significant in X-ray analysis because it helps explain why a classifier made a diagnosis. When models are adversarially trained, CAMs reveal a substantially broader set of interpretable features.

Variations of Class Activation Mapping

Class activation maps can be used only with ConvNets with a specific architecture in which feature maps directly transfer to the output softmax layers, and hence CAMs can only be used to explain the decisions of a limited number of ConvNet types.

Grad-CAM [37] is a generalization of CAM that works with a wider range of ConvNets. Grad-CAM assigns weights to the feature maps based on the gradient information from the last convolutional layer. These weights are calculated by averaging the gradients across the spatial dimensions of the input image. Let y^c be the score that a ConvNet computes for the probability that an input X-ray image displays disease or anomaly c. Grad-CAM computes the weights w_k in (2) for anomaly c as,

$$w_k = \frac{1}{z} \sum_x \sum_y \frac{\partial y^c}{\partial f_k(x,y)} \tag{2}$$

where z is the number of pixels in the feature map.

The lesion-location guided network LLAGnet [38] integrates two different attention mechanisms: Region level attention (RLA) and channel level attention (CLA) into a unified framework in order to focus on the discriminative features of lesion locations as suggested by professional radiologists. Grad-CAMs are used in LLAGnet to construct class discriminative heat maps which can identify the approximate spatial location of each candidate disease in a chest X-ray image.

Many researchers have been working on image analysis of chest X-rays, particularly after the COVID-19 pandemic. An individual with COVID-19 can suffer from many types of respiratory illness, from a simple cold to pneumonia as a result of this disease. Chest X-rays have become even more important since COVID-19 as they are used as a diagnostic tool for assessing the state of the lungs.

The deep learning architecture CovXNet [39] was designed to predict pneumonia caused by COVID-19 in chest X-ray images. X-ray images with different resolutions are used to train several ConvXNets. A meta learner uses the predictions made by the ConvXNets to generate a final output. Grad-CAM was integrated with the ConvXNets to generate heat maps used to interpret the learning of the network from a clinical perspective. The heat maps provide important information about the underlying reasons for the presence of pneumonia.

Covid-SDNet [40] is a ConvNet-based model for categorizing COVID-19 cases as severe, moderate, mild and absent from X-ray images. Grad-CAM was used to highlight the regions of an input X-ray image that triggered a prediction and also the regions that show a counterfactual explanation suggesting a different classification.

A system that integrates image processing, Guided Grad-CAM, ConvNets and risk management is presented in [41] to detect COVID-19 in chest radiography images. Guided Grad-CAM combines Grad-CAM and guided backpropagation to create high resolution heat maps that visualize at the pixel level the most important areas of an X-ray image for a ConvNet (see Figure 3).

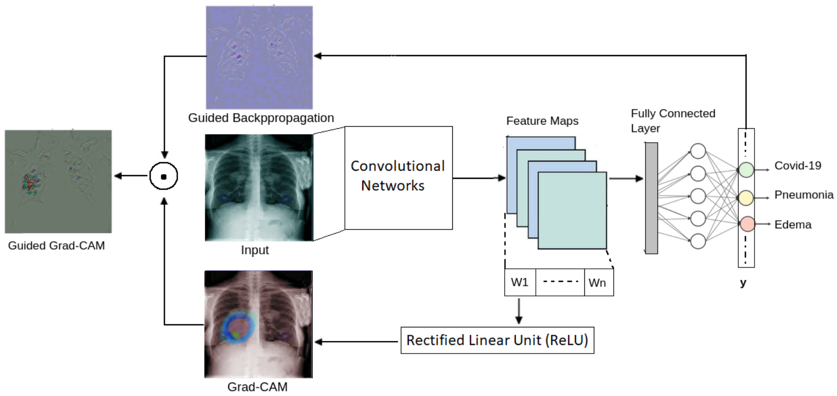

Figure 3. Conceptual view of guided gradient class activation mapping. Guided Grad CAM combines gradient information from the last convolutional layer and guided backpropagation for creating high resolution heat maps.

One of the shortcomings of Grad-CAM is that the heat maps that it produces might be distorted due to the gradients being backpropagated to the input. For the task of object detection and classification, Grad-CAM has poor performance when localizing multiple objects in the same image. Furthermore, for images containing a single object of interest, the heat maps produced by Grad-CAM often do not capture the entire object.

Grad-CAM++ [42] is a generalization of Grad-CAM that incorporates each pixel's contribution to the final output. Grad-CAM++ improves on Grad-CAM by providing better object localization and accurate detection of multiple objects in a single image. Grad-CAM++ computes the weights w_k in (1) for class or anomaly c as,

$$w_k = \sum_x \sum_y \alpha_{xy}^{kc} \, relu(\frac{\partial y^c}{\partial f_k(x,y)}) \tag{3}$$

where α_{xy}^{kc} are the weights for the pixel-wise gradients for class c and feature map k, and *relu* is rectified linear unit activation function. Work on pixel-space visualization, such as deconvolution [43] and guided backpropogation [44], have shown that positive gradients are very important in producing accurate saliency maps. An activation map f_k with a positive gradient implies that an increase in intensity at location (x, y) would have a positive influence on the classification score y_c. Based on this, in GradCAM++ a linear combination of the partial derivatives of each pixel in an activation map f_k represents the importance of that map.

Three XAI methods, Grad-CAM, Grad-CAM++ and Integrated Gradients [45] were used in multiple neural network architectures trained to detect pathologies in X-rays in [46]. The accuracy of the heat maps that they produced was compared to segmentations made by human experts. An explainable deep neural network called DeepCovidExplainer for automatic detection of COVID-19 symptoms from chest X-rays is presented in [47]. Grad-CAM++was used to highlight class discriminating regions in X-rays. Other variations of class activation maps include Score-CAM [48], LIFT-CAM [2] and Ablation-CAM [49].

3.2. *Attention-Based Explanation*

In the field of deep learning, the concept of attention has attracted a lot of interest due to its powerful influence on the learning ability of deep neural networks. Studies have been conducted on developing attention-based models that can explain decisions made by neural network models, allowing humans to trust these decisions.

Attention is undoubtedly one of the most powerful ideas in the field of cognitive science. Attention focuses on relevant features of input data while fading out the non-relevant ones. Attention allows a neural network to spend more computational power on the relevant features, which represent the critical portions of the data as shown in Figure 4.

Using attention, a neural network can focus on valuable portions of the input and learn the relationships between them.

The concept of attention is implemented in natural language processing (NLP) systems through transformers [26], which have revolutionised the field of NLP. Medical report generation is an NLP problem which will be discussed in the next section. In image analysis, the notion of attention is incorporated in ConvNets using attention modules.

Table 1. Overview of explainable AI methods for chest X-ray image analysis. The last column explains the disease or anomaly predicted by a model, and the first column indicates the XAI method used to explain the decisions of the model.

Explainable AI Techniques	Studies	Year	Chest X-ray Analysis
Class Activation Mapping (CAM) and its variations. CAM creates a heat map reflecting the importance of the feature maps.	Saporta et al. [46]	2021	COVID-19
	Paul et al. [34]	2020	Pneumothorax identification
	Mahmud et al. [39]	2020	COVID-19
	Tabik et al. [40]	2020	COVID-19
	Lin et al. [41]	2020	COVID-19
	Karim et al. [47]	2020	COVID-19
	Khakzar et al. [36]	2019	Classification of Chest pathologists
	Dunnmon et al. [50]	2019	Labelling of Chest X-ray pathologies
	Sedai et al. [6]	2018	Detection of Pathology diseases
	Rajpurkar et al. [30]	2018	Thoracic disease classification
	Ausawalaithong et al. [5]	2018	Detection of Lung Cancer
Attention-based Explanation. It focuses on the most relevant features of the input and uses them to explain predictions.	Park et al. [51]	2021	COVID-19 Detection
	Ouyang et al. [7]	2020	Multiple pathologies
	Liu et al. [8]	2019	Thoracic disease classification
	Wang et al. [52]	2019	Classification of Thoracic Diseases
	Pesce et al. [53]	2019	Pulmonary lesions
	Huang et al. [54]	2019	Diagnose Chest Pathology
	Guan et al. [55]	2018	Emphysema Detection
	Ypsilantis et al. [56]	2017	Enlarged Heart
Local Interpretable Model-Agnostic Explanations (LIME). LIME simplifies prediction models to create explanations that are simpler to understand.	Ahsan et al. [57]	2021	COVID-19
	Kamal et al. [58]	2021	COVID-19
	Dixit et al. [59]	2021	COVID-19
	Punn et al. [60]	2021	COVID-19
	Teixeira et al. [61]	2021	COVID-19
	Kundu et al. [62]	2021	Pneumonia detection
	Ahsan et al. [63]	2020	COVID-19 including multiple pathologies
Layer-wise Relevance Propagation (LRP). LRP propagates the predictions backwards through the layers of the model to compute the importance of each part of the input.	Bassi et al. [64]	2022	COVID-19
	Samsom et al. [65]	2021	Pneumonia detection
	Bassi et al. [66]	2021	COVID-19
	Bassi et al. [67]	2021	COVID-19, lung segmentation
	Karim et al. [47]	2020	COVID-19

In the multi-label chest X-ray image classification problem, the discriminative features of different pathologies must be learned. In general, chest X-rays could contain information of various anomalies, so critical clues are required to classify and localize the different abnormalities in lung regions.

Attention has been used with Fully ConvNets (FCNs) in [7,54] to design a multi-attention convolutional neural network for automatic disease detection in chest X-ray images. Fully ConvNets are an adaptation of the DenseNet-121 model that can process spatial information [54]. FCNs create multiple attention maps for each pathology category being considered via a collection of correlated convolutions followed by a mean-pooling process. Because each channel in an image shows a specific visual symptom for a disease class, the channels have the ability to represent huge intra-class variability. This intra-class variability helps to generate explanations using heat maps based on spatial attention maps.

A recurrent attention model is proposed in [53] that uses reinforcement learning to focus on the parts of an X-ray image that are likely to display pulmonary lesions.

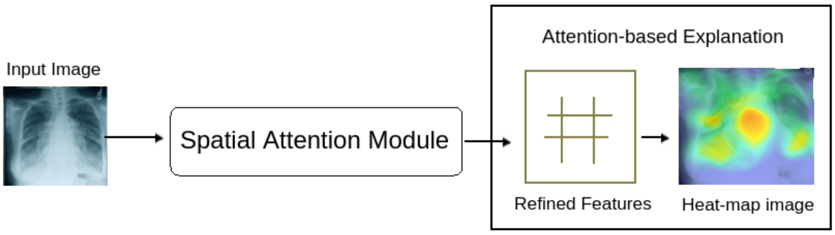

Figure 4. Attention-based explanation. The refined features are highlighted through attention feature maps and represented via a heat map image.

Spatial attention maps were used to predict the regions where a pathology might exist. This improves the accuracy of classification.

A stochastic attention-based approach is presented in [56] to predict which areas in chest X-rays should be visually explored to look for a specific radiological anomaly: enlarged heart.

Thorax diseases typically occur in localised disease-specific regions. Irrelevant noisy regions in chest X-ray images are those regions which either do not portray any information about the disorder or do not present a clear depiction of an image. These irrelevant noisy regions may have a negative effect on ConvNets that are trained with whole X-ray images. An attention guided ConvNet is described in [55] which learns where noisy regions are and uses that information to accurately identify regions showing the presence of a disorder. A heat map generated with an attention guided ConvNet serves as a guide to crop out the noisy regions of an X-ray image.

KGZNet, a knowledge guided deep neural network for automatic diagnosis of thoracic diseases is proposed in [52]. KGZNet is a zoom neural network that is trained on hierarchically organized partitions of X-ray images and guided by human medical expertise. According to [52], thoracic diseases are typically limited within certain lung regions. A lung lesion is learnt through the analysis of lung images guided by an attention heat map. Disease-specific CAM attention heat maps focused on locating specific disorders were used to visualize suspicious lesions in chest X-ray images.

A novel deep learning method to diagnose COVID-19 in chest X-ray images using a self-supervised learning approach with a convolutional attention module is presented in [51]. By using Score-CAM [48], they identified the causes of misclassified cases. Score-CAM heat maps were generated based on a convolutional attention mechanism.

3.3. Local Interpretable Model-Agnostic Explanations (LIME)

LIME is an interpretability method that is model agnostic; this means that LIME explains why an AI model makes a particular prediction without making any assumptions about the architecture of the model. For the case of ConvNets, these models might base their decisions on a large number of features. Model explanations need to be understandable to humans, so explanations must be based on a subset of features that humans can comprehend and relate to the predictions made by a model. Therefore, LIME approximates a ConvNet with a simpler interpretable model that behaves like the ConvNet for a specific prediction (this is called *local interpretability*).

LIME works by first partitioning an image into blocks of homogeneous regions consisting of pixels with similar attributes (such as color and brightness). These blocks are called *superpixels*. Then, a set S of new images is created by graying out a random selection of superpixels. A ConvNet is then used on each image of S to make a prediction, and a weight is assigned to each superpixel denoting its importance for making the prediction. In an image, the superpixels with the largest weight are highlighted, yielding a heat map showing the most important parts of the image over which the ConvNet based its prediction.

A number of works have been proposed for COVID-19 detection since the pandemic began in 2020. To reduce dependency on limited COVID-19 test kits, an alternative is the use of screening systems for chest X-rays. Two studies on the effectiveness of COVID-19 detection using different ConvNet-based models were presented in [57,63]; LIME was used to identify the features that the ConvNets used to distinguish patients with COVID-19 from patients without COVID-19.

A ConvNet-based system to identify lesions in X-rays is presented in [62]. In this work, a combination of predictions from different classifiers was used to detect abnormalities using frontal and lateral X-rays. Radiologists highlighted the regions of X-rays they would focus on to make a diagnosis, and this human-made highlighting was compared to that produced by Grad-CAM and LIME.

A spiking neural network technique is presented in [58] to detect COVID-19 positive cases using a spike neural network (SNN) with supervised synaptic learning. Three additional works on the use of ConvNets to detect COVID-19 using X-ray images are presented in [59–61]. All these works use LIME to identify the regions of an X-ray showing a COVID-19 infection.

Explanations obtained through LIME have demonstrated their importance in COVID-19 related research. In analyzing chest X-rays, localization and segmentation are considered crucial parts of the deep learning process as prediction of various chest pathologies can be effectively explained by LIME as reported by recent research discussed in this section.

3.4. Layer-Wise Relevance Propagation (LRP)

Layer-wise relevance propagation unravels the prediction of a deep neural network by propagating the prediction backwards through the layers of the network to compute relevance scores for the pixels of the input image. This backward propagation is performed as follows. For each neuron i in the last layer, the neural network computes an output X_i through its activation function. This output X_i is the relevance score for neuron i. Consider now a neuron j in network layer L_l with relevance score R^l_j. This relevance score is backpropagated to the neurons k in the previous layer L_{l-1} that provide input to j so that

$$R^l_j = \sum_{k \in L_{l-1}} r_{jk} \tag{4}$$

where r_{jk} is the fraction of the relevance score of neuron j transferred to neuron k. Equation (4) defines a conservation property, so that the total relevance score of the neurons in each layer is the same, and it is equal to the value of the prediction $p(X)$ computed by the neural network for image X.

The relevance score for a neuron k in layer L_{l-1} is computed as follows:

$$R^{l-1}_k = \sum_{j \in L_l} r_{jk} \tag{5}$$

The relevance scores of pixels are obtained from the relevance scores of the neurons in the first layer. The relevance scores are visualized as a heat map. The functions r_{jk} in Equation (4) are called *propagation rules*. Different propagation rules have been proposed, including LRP-0 [68], the epsilon rule [68], LPR-$\alpha\beta$ [68], the Z^+ rule [69] and the gamma rule [70].

Deep Covid-Explainer, a neural network ensemble for automatic detection of COVID-19 symptoms from chest X-rays is described in [47]. Class discriminating regions are highlighted using Grad-CAM++ and LRP to provide explanations and to identify critical regions on patients' chests.

In [65], two ConvNets, VGG16 and ResNet60, are used to detect pneumonia caused by the COVID-19 virus. LRP, LIME and Grad-CAM are used to generate explanations for the predictions made by the two models.

In [66], a model for detecting pneumonia and COVID-19 is presented that uses deep neural networks trained with transfer learning. LRP was used to discover that the words and letters printed in X-rays can influence the predictions of the model. In [67], a COVID-19 detection model was designed that consists of a segmentation module and a 201-layer ConvNet. LRP was used to generate heat maps which were correlated with the Brixia scoring system used by radiologists to measure the severity of COVID-19 in different lung regions.

ISNet [64] is a ConvNet-based system that is able to perform segmentation and classification as a single process. ISNet introduced the concept of relevance segmentation in LRP maps to minimize background relevance.

4. XAI Approaches in Medical Report Generation

Automated medical report generation from chest X-rays has the potential to improve patient clinical diagnosis. Automated report generation is a special type of image captioning problem. The sentences generated by image captioning are usually short and describe the most prominent visual elements of an image. This cannot fully represent the rich information of an image, but it can help train deep learning models to associate parts of an image with words. Deep learning models for image captioning that use attention mechanisms are effective and accurate [71–73].

An auto report generator can potentially relieve doctors of a considerable amount of work by assisting them in drafting medical reports. We have discussed the role of explanations in chest X-ray image analysis. In this section, we explain why XAI is a significant part of automated medical report generation.

4.1. Image Captioning with Visual Explanations

The image captioning problem combines elements of natural language processing (NLP) and image processing. There are several works on the use of deep neural networks for image captioning that use visual explanations for the models predictions. A mutli-model neural network called TandemNet is presented in [74] which can detect bladder cancer and produce a diagnostic report. TandemNet uses ResNet to analyze images and long-short term memory (LSTM) networks to model report sentences. A dual-attention module is used to train the system using images and text.

Through the interplay of semantic information with visual information, TandemNet is taught to distill the most relevant features of an image. Attention maps are used to visualize how TandemNet uses image and text information to support its predictions.

TieNet [75] is a system for predicting thoracic diseases and automatically generating diagnostic reports. TieNet uses a ResNet for image analysis and LSTM networks for text processing, and it integrates multi-level attention models for the most significant words in a report and regions in an image.

A system is proposed in [76] that generates explanations for the predictions made by a deep-neural-network-based diagnosis system. The justification generator provides explanations consisting of heat maps highlighting the most relevant regions in an image and textual reports indicating the significance of the heat maps.

4.2. Textual Explanations with Concept Activation Vectors

Concept activation vectors (CAVs) [77] are designed to provide an interpretation of the inner working of a deep neural network using human understandable concepts. The state of a neural network can be represented as a vector space V_n, where vectors correspond to input features and neuron outputs. This vector space is difficult to understand for humans who are more adept at working with concepts. CAVs provide a translation between V_n and V_k, where V_k is a vector space in which vectors correspond to human understandable concepts.

In [78], a ConvNet model using variational auto-encoders (VAE) [79] is presented that detects cardiac diseases in temporal sequences of cardiac magnetic resonance (MR) segmentations. CAVs allow us to identify clinically known biomarkers that are associated with cardiac disorders. Hence, when the model classifies images, it also provides interpretable concepts relevant to the classification and relates them to the corresponding parts of the images.

The CAV model was extended in [80] through the addition of regression concept vectors (RCVs); while CAV models indicate whether a concept is present or not in an explanation of a deep learning model's prediction, regression concept vectors express continuous measures of that concept. RCVs are especially useful when investigating continuous features such as tumor size. The use of RCVs to generate explanations for the decisions of the breast cancer detection ConvNet [80] gives a better understanding of why the ConvNet classifies some areas of an image as cancerous and others as healthy.

In [81], a framework is proposed for generating explanations for ConvNet decisions using RCVs. The framework allows explanation generation for multi-class classification tasks, and it improves the learning stage through the removal of spatial dependencies of the convolutional feature maps.

4.3. Other Textual Explanation Techniques

A hierarchical model for text processing with multi-attention is presented in [31]. This work identified two main aspects in automated medical report generation. The first one is related to identifying regions in an X-ray image that show a pathology and describing this information in textual form. To address both issues, a novel multi-attention hierarchical model is proposed that focuses on the image's channels and spatial information and a word embedding method that incorporates the patient's medical history.

As part of the efforts to develop a radiologist-interpretable algorithm for lung cancer prediction, ref. [82] presents a hierarchical semantic convolutional neural network model (HSCNN) for detecting malignant nodules in CT scans. When analyzing and detecting a malignant nodule, HSCNN considers five nodule properties: calcification, margin, subtlety, texture and sphericity. In addition to the diagnosis prediction, these five nodule properties help explain the final malignancy prediction.

In [8], the authors presented a domain-based system for generating chest X-ray radiology reports. Based on predictions about topics that will be discussed in the report, their model then generates conditional sentences corresponding to these topics. With reinforcement learning, the resulting system is fine-tuned for both clinical accuracy and readability. The attention mechanism was embedded in their presented model, and attention maps were generated as an output with a highlighted portion of an image that corresponds to its description.

5. Discussion of Explainable AI Approaches

In the development of high stakes decision-making systems, such as computer-aided diagnostic systems, it is of fundamental importance that we understand how those systems work and how they make decisions; without this knowledge, these systems cannot be trusted. XAI is rapidly becoming one of the mainstream subjects in AI as it provides the foundations for understanding complex AI models, a necessary requisite for deploying these systems in critical applications. XAI is still in its early stages. New and better explanation techniques are being developed, and we expect that they will revolutionize the healthcare field.

Automated medical report generation can be subdivided into two problems, which are related but come from two separate domains of study. Analysis of medical images is one aspect of the problem that is related to computer vision. The other half of the problem belongs to natural language processing (NLP). There is no standard single XAI technique available that can be applied simultaneously to computer vision and NLP models. When analysis of medical images needs to be conducted on parts of an image, class activation maps (CAM) and its variants have proven to be important techniques for explaining the decisions of ConvNet classifiers. In situations where we do not have understanding of the AI model which needs to be trained, model-agnostic techniques, such as LIME, are helpful.

Attention mechanisms are part of neural architectures and are capable of dynamically highlighting relevant features of the input data, which in the case of image analysis focuses on specific parts of an image and in the case of NLP on a particular sequence of textual elements. If we are interested in finding the role of each pixel to the training of an AI model, then LRP is the best-suited technique as it propagates the output back through the network until reaching the input layer using the network weights and neural activation created by the forward-pass.

A summary of XAI techniques used in medical image analysis and report generation is depicted in Table 2. Being post hoc techniques, CAM, Grad-CAM, LIME, LRP and CAV use trained networks to generate explanations, whereas attention-based explanations and explanations generated for image captioning are ante hoc techniques. LIME and CAV are techniques that generate global explanations. CAV can also generate local explanations.

Table 2. List of explainable AI techniques used for medical image analysis and report generation. Checkmarks in the last 4 columns indicate whether each explainable AI technique is post hoc or ante hoc and if it generates global and local explanation.

Techniques	Study	Post-hoc	Ante-hoc	Global	Local
Class Activation Mapping and its variants	[2,5,33–41,46,47,50]	✓	-	-	✓
Attention-based Explanations	[7,8,26,51–56]	-	✓	-	✓
Local Interpretable Model-Agnostic Explanations (LIME)	[57–63]	✓	-	✓	-
Layer-wise Relevance Propagation (LRP)	[47,64–70]	✓	-	-	✓
Image Captioning with Visual Explanations	[74–76]	-	✓	-	✓
Concept Activation Vectors	[77–79,81]	✓	-	✓	✓

Two important open research problems that we encountered through our study are the following.

- An integrated XAI framework is required for automated medical report generation. The framework should integrate the explainability aspect for both image processing and text generation. Currently, existing XAI methods deal with only one aspect of the automated report generation process.
- A reasoning mechanism is required to provide quantitative, and not just qualitative explanations of the decisions of a model. This will be helpful to understand and improve the accuracy of AI models.

5.1. Improving AI Models through XAI

XAI methods help explain the decisions made by AI models and can help enhance them. The reflective neural network Reflective-Net introduced in [83] uses a reflection process to improve its accuracy. In this reflection process, first a classifier makes a prediction based on an input I, and it generates an explanation E. Then, the input I and explanation E are given to a reflective network that refines the prediction. Training the reflective network with correct and incorrect explanations helps increase its accuracy.

XAI techniques have been found to be highly effective in improving deep learning model performance as described in [84]. MobilNet, a ConvNet to detect metal surface defects, was used in [84] on a dataset of images containing images with super imposed text and company logos. Through the use of LRP, it was demonstrated that the performance of the model of mode was consistently negatively affected by that unwanted information. The LRP analysis showed that the model learned to identify patterns in the text and logos rather than the actual surface defects. The performance of the model was greatly improved by removing the superimposed text and logos from the images before training.

5.2. Challenges in Explainable AI

In order to ensure the production of accurate automated medical reports, we should use a multidisciplinary approach and take into account input from the report generation system designers, the users of the system and anyone who will be affected by the system. In spite of the fact that XAI can assist in identifying problems with medical data, the existence of unstructured medical data remains a challenge for the development of useful AI-based systems.

There are several problems with existing XAI techniques. Two of the most important problems are:

1. Difficulty for humans to understand saliency maps,
2. Lack of quantitative methods to evaluate the correctness and completeness of explanations.

To address the first challenge, we note that there is no best universal XAI technique. Some techniques provide understandable and accurate explanations for some prediction models but not for others. So, a careful selection of the XAI methods is essential for improving the quality of explanations. In addition, it is very important, as mentioned above, to involve system users in the design of XAI techniques, as user input is geared towards improving explanations understandability.

The second challenge highlights the need to design accurate metrics to evaluate XAI techniques [1]. Currently, studies on this area are mainly based on subjective measurements, such as user satisfaction, clarity of descriptions and trust in the system [1]. An overview of metrics for evaluating explainability properties (i.e., clarity, breadth, parsimony, completeness and soundness) is discussed in [85]. There is an overall lack of universally accepted quantitative evaluation metrics for XAI techniques, so additional research in this direction is needed.

6. Conclusions

A report generated by an automated medical report generator must be trustworthy, easy to understand and accurate in order to be used effectively in practice. The quality of the explanations on how the report was generated and how its diagnoses were reached is a key factor to meet these goals. Having a system that is explainable allows developers to identify any shortcomings or inefficiencies and clinicians to be confident in the decisions they make with the help of these systems.

Although many studies have been conducted on the use of XAI in the medical field, there was not any work summarizing research on the use of XAI for automated medical report generation. XAI techniques have been experimented and discussed with reference to medical image analysis, but the role of XAI in NLP models with reference to medical report generation have not been extensively explored. This paper summarizes some of the most relevant research in the use of XAI for image analysis of chest X-ray images and automatic medical report generation. We also list some of the current challenges in XAI research and mention some ways in which the performance of AI models can be improved through the use of XAI.

Author Contributions: This research topic is conceptualized by S.B.A.; Information is collected and written in paper form by S.B.A. and R.S.-O. Initial-draft of paper is written by S.B.A. Formal analysis of presented work is done by S.B.A., R.S.-O. and L.I. Writing—reviewing and editing is performed by S.B.A., R.S.-O. and L.I. Funding is acquired by R.S.-O. and L.I. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partially supported by a Discovery Grant (RGPIN-2020-06423) from the Natural Sciences and Engineering Research Council of Canada to R.S.-O. and a Discovery Grant (R3143A01) from the Natural Sciences and Engineering Research Council of Canada to L.I.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Arrieta, A.B.; Rodríguez, N.D.; Ser, J.D.; Bennetot, A.; Tabik, S.; Barbado, A.; García, S.; Gil-Lopez, S.; Molina, D.; Benjamins, R.; et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Inf. Fusion* **2020**, *58*, 82–115. [CrossRef]
2. Jung, H.; Oh, Y. Towards better explanations of class activation mapping. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 1316–1324.
3. Adadi, A.; Berrada, M. Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access* **2018**, *6*, 52138–52160. [CrossRef]
4. Feng, Y.; Teh, H.S.; Cai, Y. Deep learning for chest radiology: A review. *Curr. Radiol. Rep.* **2019**, *7*, 24. [CrossRef]
5. Ausawalaithong, W.; Thirach, A.; Marukatat, S.; Wilaiprasitporn, T. Automatic lung cancer prediction from chest X-ray images using the deep learning approach. In Proceedings of the 11th Biomedical Engineering International Conference (BMEICON), Chiang Mai, Thailand, 21–24 November 2018; pp. 1–5.
6. Sedai, S.; Mahapatra, D.; Ge, Z.; Chakravorty, R.; Garnavi, R. 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, Switzerland, 2018; Volume 11046, pp. 267–275 .
7. Ouyang, X.; Karanam, S.; Wu, Z.; Chen, T.; Huo, J.; Zhou, X.S.; Wang, Q.; Cheng, J.-Z. Learning hierarchical attention for weakly-supervised chest X-ray abnormality localization and diagnosis. *IEEE Trans. Medicalimaging* **2021**, *40*, 2698–2710. [CrossRef] [PubMed]
8. Guanxiong, L.; Tzu-Ming Harry, H.; Matthew, M.; Willie, B.; Wei-Hung, W.; Peter, S.; Marzyeh, G. Clinically accurate chest X-ray report generation. *Mach. Learn. Healthc.* **2019**, *106*, 249–269.
9. Alfarghaly, O.; Khaled, R.; Elkorany, A.; Helal, M.; Fahmy, A. Automated radiology report generation using conditioned transformers. *Inform. Med. Unlocked* **2021**, *24*, 100557. [CrossRef]
10. Brunese, L.; Mercaldo, F.; Reginelli, A.; Santone, A. Explainable deep learning for pulmonary disease and coronavirus COVID-19 detection from X-rays. *Comput. Methods Programs Biomed.* **2020**, *196*, 105608. [CrossRef]
11. Cao, J.; Li, X. A 3D 2D convolutional neural network model for hyperspectral image classification. *arXiv* **2021**, arXiv:2111.10293.
12. Shamsolmoali, P.; Zareapoor, M.; Yang, J. Convolutional neural network in network (cnnin): Hyperspectral image classification and dimensionality reduction. *IET Image Process.* **2019**, *13*, 246–253. [CrossRef]
13. Ahmed, S.B.; Naz, S.; Razzak, M.I.; Yousaf, R. Deep learning based isolated arabic scene character recognition. In Proceedings of the 2017 1st International Workshop on Arabic Script Analysis and Recognition (ASAR), Nancy, France, 3–5 April 2017; pp. 46–51.
14. Simonyan, K.; Vedaldi, A.; Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. In Proceedings of the 2nd International Conference on Learning Representations, ICLR, Banff, AB, Canada, 14–16 April 2014.
15. Xue, Y.; Xu, T.; Long, L.R.; Xue, Z.; Antani, S.; Thoma, G.R.; Huang, X. Multimodal recurrent model with attention for automated radiology report generation. In Proceedings of the Medical Image Computing and Computer Assisted Intervention—MICCAI 2018, Granada, Spain, 16–20 September 2018; pp. 457–466.
16. Al-muzaini, H.A.; Al-yahya, T.N.; Benhidour, H. Automatic arabic image captioning using rnn-lstm-based language model and cnn. *Int. Adv. Comput. Sci. Appl.* **2018**, *9*, 2018. [CrossRef]
17. Wang, H.; Wang, H.; Xu, K. Evolutionary recurrent neural network for image captioning. *Neurocomputing* **2020**, *401*, 249–256. [CrossRef]
18. Irvin, J.; Rajpurkar, P.; Ko, M.; Yu, Y.; Ciurea-Ilcus, S.; Chute, C.; Marklund, H.; Haghgoo, B.; Ball, R.L.; Shpanskaya, K.S.; et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; pp. 590–597.
19. Chen, Z.; Song, Y.; Chang, T.-H.; Wan, X. Generating radiology reports via memory-driven transformer. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, 19–20 November 2020.
20. Cowan, N. Working memory underpins cognitive development, learning, and education. *Educ. Psychol. Rev.* **2014**, *26*, 197–223. [CrossRef] [PubMed]
21. Galassi, A.; Lippi, L.; Torroni, P. Attention, please! A Critical Review of Neural Attention Models in Natural Language Processing. *arXiv* **2019**, arXiv:1902.02181. Available online: http://arxiv.org/abs/1902.02181 (accessed on 10 September 2022).
22. Xue, Y. Attention based image compression post-processing convlutional neural network. In *CVPR Workshops*; Computer Vision Foundation/IEEE: Piscataway, NJ, USA, 2019.
23. Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.C.; Salakhutdinov, R.; Zemel, R.S.; Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention. *ICML* **2015**, *1392*, 2048–2057.
24. Jaderberg, M.; Simonyan, K.; Zisserman, A.; Kavukcuoglu, K. Spatial transformer networks. *Adv. Neural Inf. Process.* **2015**, *28*, 2017–2025.
25. Chen, L.; Zhang, H.; Xiao, J.; Nie, L.; Shao, J.; Liu, W.; Chua, T.-S. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6298–6306.

26. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is All you Need. In Proceedings of the Advances in Neural Information Processing Systems 30 (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; Volume 30, pp. 5998–6008.
27. Shang, Y.; Xu, N.; Jin, Z.; Yao, X. Capsule network based on self-attention mechanism. In Proceedings of the 2021 13th International Conference on Wireless Communications and Signal Processing (WCSP), Changsha, China, 20–22 October 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 1–4.
28. Berg, A.; O'Connor, M.; Cruz, M.T. Keyword transformer: A self-attention model for keyword spotting. In Proceedings of the Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association, Brno, Czech Republic, 30 August–3 September 2021; pp. 4249-4253.
29. Murdoch, W.J.; Singh, C.; Kumbier, K.; Abbasi-Asl, R.; Yu, B. Definitions, methods, and applications in interpretable machine learning. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 71–80. [CrossRef]
30. Rajpurkar, P.; Irvin, J.; Ball, R.L.; K, Y.B.Z.; Mehta, H.; Duan, T.; Ding, D.; Bagul, A.; Langlotz, C.P. Deep learning for chest radiograph diagnosis: A retrospective comparison of the chexnext algorithm to practicing radiologists. *PLoS Med.* **2018**, *15*, e1002686. [CrossRef]
31. Huang, X.; Yan, F.; Xu, W.; Li, M. Multi-attention and incorporating background information model for chest X-ray image report generation. *IEEE Access* **2019**, *7*, 154808–154817. [CrossRef]
32. Hoffer, E.; Ailon, N. *Deep Metric Learning Using Triplet Network*; SIMBAD: Copenhagen, Denmark, 2015; pp. 84–92.
33. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
34. Yi, P.H.; Kim, T.K.; Yu, A.C.; Bennett, B.; Eng, J.; Lin, C.T. Can i outperform a junior resident? comparison of deep neural network to first-year radiology residents for identification of pneumothorax. *Emerg. Radiol.* **2020**, *27*, 367–375. [CrossRef]
35. Liu, H.; Wang, L.; Nan, Y.; Jin, F.; Wang, Q.; Pu, J. Sdfn: Segmentation-based deep fusion network for thoracic disease classification in chest X-ray images. *Comput. Med. Imaging Graph* **2019**, *75*, 66–73. [CrossRef]
36. Khakzar, A.; Albarqouni, S.; Navab, N. Learning interpretable features via adversarially robust optimization. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Shenzhen, China, 13–17 October 2019.
37. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 618–626.
38. Chen, B.; Li, J.; Lu, G.; Zhang, D. Lesion location attention guided network for multi-label thoracic disease classification in chest X-rays. *IEEE J. Biomed. Health Inform.* **2020**, *24*, 2016–2027. [CrossRef] [PubMed]
39. Mahmud, F.S.; Rahman, T. 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.* **2020**, *122*, 103869. [CrossRef]
40. Tabik, S.; Gómez-Ríos, A.; Martín-Rodríguez, J.L.; Sevillano-García, I.; Rey-Area, M.; Charte, D.; Guirado, E.; Suárez, J.L.; Luengo, J.; Valero-González, M.A.; et al. Covidgr dataset and covid-sdnet methodology for predicting COVID-19 based on chest X-ray images. *IEEE J. Biomed. Health* **2020**, *24*, 3595–3605. [CrossRef] [PubMed]
41. Lin, T.-C.; Lee, H.-C. COVID-19 chest radiography images analysis based on integration of image preprocess, guided grad-cam, machine learning and risk management. In Proceedings of the 4th International Conference on Medical and Health Informatics, Kamakura City, Japan, 14–16 August 2020; pp. 281–288.
42. Chattopadhyay, A.; Sarkar, A.; Howlader, P.; Balasubramanian, V.N. Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision, WACV 2018, Lake Tahoe, NV, USA, 12–15 March 2018; pp. 839–847.
43. Zeiler, M.D.; Fergus, R. Visualizing and understanding convolutional networks. In Proceedings of the Computer Vision—ECCV 2014, Zurich, Switzerland, 6–12 September 2014; Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T., Eds.; Springer International Publishing: Cham, Switzerland, 2014; pp. 818–833.
44. Springenberg, J.T.; Dosovitskiy, A.; Brox, T.; Riedmiller, M.A. Striving for simplicity: The all convolutional net. *arXiv* **2015**, arXiv:1412.6806.
45. Sundararajan, M.; Taly, A.; Yan, Q. Axiomatic attribution for deep networks. In Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6–11 August 2017; Precup, D., Teh, Y.W., Eds.; pp. 3319–3328.
46. Saporta, A.; Gui, X.; Agrawal, A.; Pareek, A.; Truong, S.Q.; Nguyen, C.D.; Ngo, V.-D.; Seekins, J.; Blankenberg, F.G.; Ng, A.Y.; et al. Benchmarking saliency methods for chest X-ray interpretation. *Nat. Mach. Intell.* **2022**, *4*, 867–878. [CrossRef]
47. Karim, M.R.; Döhmen, T.; Cochez, M.; Beyan, O.; Rebholz-Schuhmann, D.; Decker, S. Deepcovidexplainer: Explainable COVID-19 diagnosis from chest X-ray images. In Proceedings of the 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Seoul, Republic of Korea, 16–19 December 2020; pp. 1034–1037.
48. Wang, H.; Wang, Z.; Du, M.; Yang, F.; Zhang, Z.; Ding, S.; Mardziel, P.; Hu, X. Score-cam: Score-weighted visual explanations for convolutional neural networks. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, 14–19 June 2020; pp. 111–119.

49. Desai, S.; Ramaswamy, H.G. Ablation-cam: Visual explanations for deep convolutional network via gradient-free localization. In Proceedings of the 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), Snowmass, CO, USA, 1–5 March 2020; pp. 972–980.
50. Dunnmon, J.; Yi, D.; Langlotz, C.; Ré, C.; Rubin, D.; Lungren, M. Assessment of convolutional neural networks for automated classification of chest radiographs. *Radiology* **2019**, *290*, 537–544. [CrossRef]
51. Park, J.; Kwak, I.-Y.; Lim, C. A deep learning model with self-supervised learning and attention mechanism for COVID-19 diagnosis using chest X-ray images. *Electronics* **2021**, *10*, 1996. [CrossRef]
52. Wang, K.; Zhang, X.; Huang, S. Kgznet: Knowledge-guided deep zoom neural networks for thoracic disease classification. In Proceedings of the 2019 IEEE International Conference on Bioinformatics and Biomedicine, San Diego, CA, USA, 18–21 November 2019; pp. 1396–1401.
53. Pesce, E.; Withey, S.J.; Ypsilantis, P.-P.; Bakewell, R.; Goh, V.; Montana, G. Learning to detect chest radiographs containing pulmonary lesions using visual attention networks. *Med. Image Anal.* **2019**, *53*, 26–38. [CrossRef]
54. Huang, Z.; Fu, D. 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), Chongqing, China, 24–26 May 2019; pp. 294–299.
55. Guan, Q.; Huang, Y.; Zhong, Z.; Zheng, Z.; Zheng, L.; Yang, Y. Diagnose like a radiologist: Attention guided convolutional neural network for thorax disease classification. *arXiv* **2018**, arXiv:1801.09927.
56. Ypsilantis, P.-P.; Montana, G. Learning what to look in chest X-rays with a recurrent visual attention model. *arXiv* **2017**, arXiv:1701.06452.
57. Ahsan, M.M.; Nazim, R.; Siddique, Z.; Huebner, P. Detection of COVID-19 patients from ct scan and chest X-ray data using modified mobilenetv2 and lime. *Healthcare* **2021**, *9*, 1099. [CrossRef] [PubMed]
58. Kamal, M.S.; Chowdhury, L.; Dey, N.; Fong, S.J.; Santosh, K. Explainable ai to analyze outcomes of spike neural network in COVID-19 chest X-rays, in In Proceedings of the 2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Melbourne, Australia, 17–20 October 2021; pp. 3408–3415.
59. Dixit, A.; Mani, A.; Bansal, R. Covidetect-desvm: Explainable framework using differential evolution algorithm with svm classifier for the diagnosis of COVID-19. In Proceedings of the 2021 4th International Conference on Recent Developments in Control, Automation Power Engineering (RDCAPE), Noida, India, 7–8 October 2021; pp. 339-334.
60. Punn, N.S.; Agarwal, S. Automated diagnosis of COVID-19 with limited posteroanterior chest X-ray images using fine-tuned deep neural networks. *Appl. Intell.* **2020**, *51*, 2689–2702.
61. Teixeira, L.; Pereira, R.M.; Bertolini, D.; Oliveira, L.S.; Nanni, L.; Cavalcanti, G.; Costa, Y. Impact of lung segmentation on the diagnosis and explanation of COVID-19 in chest X-ray images. *Sensors* **2021**, *21*, 7116. [CrossRef]
62. Luis, V.; Flávio, A.; Santos, N.P.; João, A.; Manuel, R.T.J.; Rodrigo, V. A ensemble methodology for automatic classification of chest X-rays using deep learning. *Comput. Biol. Med.* **2022**, *145*, 105442.
63. Ahsan, M.M.; Gupta, K.D.; Islam, M.M.; Sen, S.; Rahman, M.L.; Hossain, M.S. COVID-19 symptoms detection based on nasnetmobile with explainable ai using various imaging modalities. *Mach. Knowl. Extr.* **2020**, *2*, 490–504. [CrossRef]
64. Bassi, P.R.; Cavalli, A. ISNet: Costless and Implicit Image Segmentation for Deep Classifiers, with Application in COVID-19 Detection. *arXiv* **2022**, arXiv:2202.00232.
65. Samsom, Q. Generating explanations for chest medical scan pneumonia predictions. *Covid Inf. Commons Stud. Pap. Chall.* **2021**. Available online: https://academiccommons.columbia.edu/doi/10.7916/d8-t9np-xk59 (accessed on 1 July 2022).
66. Bassi, P.R.; Attux, R. A deep convolutional neural network for COVID-19 detection using chest X-rays. *Res. Biomed. Eng.* **2022**, *38*, 139–148. [CrossRef]
67. Pedro, B.; de Faissol, A.R. COVID-19 detection using chest X-rays: Is lung segmentation important for generalization? *arXiv* **2021**, arXiv:2104.06176.
68. Bach, S.; Binder, A.; Montavon, G.; Klauschen, F.; Müller, K.-R.; Samek, W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE* **2015**, *10*, e0130140. [CrossRef]
69. Montavon, G.; Lapuschkin, S.; Binder, A.; Samek, W.; Müller, K.-R. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recognit.* **2017**, *65*, 211–222. [CrossRef]
70. Montavon, G.; Binder, A.; Lapuschkin, S.; Samek, W.; Müller, K.-R. Layer-wise relevance propagation: An overview. *Explain. AI* **2019**, *11700*, 193–209.
71. Lu, J.; Xiong, C.; Parikh, D.; Socher, R. Knowing when to look: Adaptive attention via A visual sentinel for image captioning. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 375–383.
72. Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; Zhang, L. Bottom-up and top-down attention for image captioning and VQA. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 6077–6086.
73. Fang, H.; Gupta, S.; Iandola, F.N.; Srivastava, R.K.; Deng, L.; Dollár, P.; Gao, J.; He, X.; Mitchell, M.; Platt, J.C.; et al. From captions to visual concepts and back. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015.

74. Zhang, Z.; Chen, P.; Sapkota, M.; Yang, L. Tandemnet: Distilling knowledge from medical images using diagnostic reports as optional semantic references. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2017*; Springer International Publishing: Berlin/Heidelberg, Germany, 2017; pp. 320–328.
75. Wang, X.; Peng, Y.; Lu, L.; Lu, Z.; Summers, R.M. Tienet: Text-image embedding network for common thorax disease classification and reporting in chest X-rays. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 9049–9058.
76. Lee, H.; Kim, S.T.; Ro, Y.M. Generation of multimodal justification using visual word constraint model for explainable computer-aided diagnosis. In Proceedings of the Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support, Shenzhen, China, 17 October 2019; pp. 21–29.
77. Kim, B.; Wattenberg, M.; Gilmer, J.; Cai, C.J.; Wexler, J.; Viégas, F.B.; Sayres, R. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmaessan, Stockholm, Sweden, 10–15 July 2018; Volume 80, pp. 2673–2682.
78. Clough, J.R.; Öksüz, I.; Puyol-Antón, E.; Ruijsink, B.; King, A.P.; Schnabel, J.A. Global and local interpretability for cardiac mri classification. In Proceedings of the 22nd International Conference MICCAI, Shenzhen, China, 13–17 October 2019; pp. 656–664.
79. Kingma, D.P.; Welling, M. Auto-encoding variational bayes. *arXiv* **2014**, arXiv:1312.6114.
80. Pereira, S.; Meier, R.; Alves, V.; Reyes, M.; Silva, C.A. Understanding and interpreting machine learning in medical image computing applications. *Lect. Notes Comput. Sci.* **2018**, *11038*, 1–148.
81. Graziani, M.; Andrearczyk, V.; Marchand-Maillet, S.; Müller, H. Concept attribution: Explaining cnn decisions to physicians. *Comput. Biol. Med.* **2020**, *123*, 103865.
82. Shen, S.; Han, S.X.; Aberle, D.R.; Bui, A.A.T.; Hsu, W. An interpretable deep hierarchical semantic convolutional neural network for lung nodule malignancy classification. *Expert Syst. Appl.* **2019**, *128*, 84–95. [CrossRef] [PubMed]
83. Schneider, J.; Vlachos, M. Reflective-net: Learning from explanations. *arXiv* **2020**, arXiv:2011.13986.
84. Bento, V.; Kohler, M.; Diaz, P.; Mendoza, L.A.F.; Pacheco, M.A.C. Improving deep learning performance by using explainable artificial intelligence (xai) approaches. *Discov. Artif. Intell.* **2021**, *1*, 9. [CrossRef]
85. Zhou, J.; Gandomi, A.H.; Chen, F.; Holzinger, A. Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics* **2021**, *10, 593*. [CrossRef]