applied sciences

MDPI

Article

Evaluating Explainable Artificial Intelligence for X-ray Image Analysis

Miquel Miró-Nicolau [1,2], Gabriel Moyà-Alcover [1,2] and Antoni Jaume-i-Capó [1,2,*]

[1] Computer Graphics and Vision and AI Group (UGiVIA), Research Institute of Health Sciences (IUNICS), Department of Mathematics and Computer Science, Universitat de les Illes Balears, 07122 Palma, Spain; miquel.miro@uib.cat (M.M.-N.); gabriel.moya@uib.es (G.M.-A.)

[2] Laboratory of Artificial Intelligence Applications (LAIA@UIB), Universitat de les Illes Balears, 07122 Palma, Spain

* Correspondence: antoni.jaume@uib.es

Abstract: The lack of justification of the results obtained by artificial intelligence (AI) algorithms has limited their usage in the medical context. To increase the explainability of the existing AI methods, explainable artificial intelligence (XAI) is proposed. We performed a systematic literature review, based on the guidelines proposed by Kitchenham and Charters, of studies that applied XAI methods in X-ray-image-related tasks. We identified 141 studies relevant to the objective of this research from five different databases. For each of these studies, we assessed the quality and then analyzed them according to a specific set of research questions. We determined two primary purposes for X-ray images: the detection of bone diseases and lung diseases. We found that most of the AI methods used were based on a CNN. We identified the different techniques to increase the explainability of the models and grouped them depending on the kind of explainability obtained. We found that most of the articles did not evaluate the quality of the explainability obtained, causing problems of confidence in the explanation. Finally, we identified the current challenges and future directions of this subject and provide guidelines to practitioners and researchers to improve the limitations and the weaknesses that we detected.

Keywords: explainable artificial intelligence; artificial intelligence; X-ray; decision support systems; neural networks; image analysis

Citation: Miró-Nicolau, M.; Moyà-Alcover, G.; Jaume-i-Capó, A. Evaluating Explainable Artificial Intelligence for X-ray Image Analysis. *Appl. Sci.* **2022**, *12*, 4459. https://doi.org/10.3390/app12094459

Academic Editor: Jan Egger

Received: 28 February 2022
Accepted: 26 April 2022
Published: 28 April 2022

1. Introduction

Artificial intelligence (AI) is used in many fields, and this usage has increased with the appearance of deep learning models. These models have achieved impressive results, but these results are frequently difficult to understand for the human user. To overcome the limitation of these models, also named black-box models, explainable artificial intelligence (XAI) emerged to be able to understand, trust, and effectively manage this emerging generation of artificially intelligent partners [1]. XAI approaches were proposed to make a shift towards more transparent AI, aiming to create a suite of techniques that produce more explainable models whilst maintaining high-performance levels [2].

Even with the increased importance of XAI, some basic questions remain without a consensus, including what explainable artificial intelligence is. In this paper, we used the definition proposed in [3] that defined it as follows: *Given an audience, an explainable artificial intelligence is one that produces details or reasons to make its functioning clear or easy to understand, and trust the behavior of intelligent systems*, as users can verify the reasons for the computer's presentations or assertions as true and can understand them, and, furthermore, as assessed in [4], anticipate circumstances in which the machine's recommendations will not be trustworthy, in which case the computer's recommendations should not be followed even though they appear trustworthy.

The appearance of XAI approaches is due to the dangers of the black-box models. Making decisions based on these models produced ethical and technical concerns caused by the fact that these models were not understood and, consequently, their decisions were not justified. Solving these dangers has an importance that depends on the field of study to be applied. In some of these fields, e.g., medicine, fixing these problems is vital. For verification of the results and ethical and legal issues, the non-explained prediction has not been sufficient to support a diagnosis and therefore has not been suitable for medical practice.

XAI algorithms overcome the caveats of black-box models, increasing their explainability. Most of these methods cannot be used without some model working as a backbone. The algorithms are typically used alongside black-box models and aim to improve them, but they are also fully independent from one other.

The main concern of any study about explainability should be XAI algorithms and not the AI models. We found multiple solutions aiming to do so. For example, one of the biggest fields of study is saliency methods, such as class activation maps (CAMs) [5] or attention, characterized by indicating the importance of each pixel for the task to be explained; however, we can also find completely different methods, such as those aiming to obtain textual explanations through auxiliary models. This diversity within explainability algorithms makes some kind of classification necessary to simplify further decision-making processes.

In the medical images field, the interest in radiography analysis has provoked a significant increase in requests for the analysis of X-ray images. The usage of AI to support radiological work is a promising field of study, as indicated in several recent reviews about the topic [6–8].

One of the problems of applying XAI to the analysis of X-ray images is a different kind of image that includes radiography analysis to support the diagnosis of multiple diseases, such as pneumonia [9,10] and bone diseases [11–13], among others. This trend was augmented with the appearance of COVID-19, as could be seen in the publication of several guidelines to perform this analysis, such as the one proposed by the Spanish Society of Emergency Radiology (SERAU) that had standardized the radiological report in the case of COVID-19 infection (http://serau.org/covid-19/ (accessed on 25 April 2022)). Furthermore, the publication of these guidelines also indicates the increased interest in the standardization of the prognosis of diseases and could be seen as an initial step for the development of XAI algorithms.

To sum up, and following the proposal by [2], the need for explanations in X-ray image analysis contexts is justified by the following:

1. Verification: A system must be interpreted and verified by medical experts.
2. Upgrade: Understanding a system allows for its improvement.
3. Discover: New knowledge can be generated based on XAI systems strategies.
4. Legal, ethical, and social aspects: Compliance with these issues is a priority for AI systems in medical image analysis contexts, as they affect humans.

As new XAI models have emerged and have been used for radiography analysis, the interpretation of results and the extraction of broader principles from existing work have become more challenging. One possible solution to this challenge is to adopt an evidence-based paradigm. To understand the role of evidence, we need to recognize that, across diverse study disciplines, there is a common need for methods that allow for the objective and consistent aggregation of outcomes in multiple empirical studies [14]. In this context, evidence is defined as the synthesis of the best scientific studies on a specific topic or research question. The primary method of synthesis is a systematic literature review (SLR), as proposed in [15]. In contrast to an expert review based on ad hoc literature selection, an SLR is a methodologically rigorous review of research findings. The aim of an SLR, as explained in [16], is not merely to aggregate all available evidence on a research question but also to enable the development of evidence-based guidelines for practitioners. Furthermore, this kind of review also allows for the identification of open questions about the subject of the research.

We herein present the findings from the SLR on explainable artificial intelligence (XAI) for X-ray image analysis. To our knowledge, this topic has not been systematically reviewed. We aim to determine and compare the method of XAI for radiography image analysis, to conduct further research in this area. In addition, we identified the current challenges and future directions of this subject, providing guidelines to practitioners and researchers for improving the limitations and the weaknesses that we detected.

This article is organized as follows. In Section 2, we describe the method used for our systematic literature review; this involves producing and following the rules of a protocol. Section 3 presents the results of our synthesis of the literature, including the geographical spread and publication details. Here, we report the results of our quality assessment and research questions (RQs). Subsequently, we discuss our key findings. In Section 4, we present some limitations of this study. In Section 5, we provide some recommendations for further research. The last part is Section 6, in which we present our conclusions.

2. Method

This study has been undertaken as a systematic literature review based on the guidelines proposed in [15], aiming to systematically identify the application of XAI to X-ray images. The review protocol includes these elements: a definition of the research questions, a search strategy for primary studies, study selection criteria, study quality assessment procedures, data extraction criteria, and a synthesis of the extracted data.

2.1. Research Questions (RQs)

This study addresses the following research questions:

- RQ1: What is the purpose of the article?
  With this question, we wanted to identify the aim of the study. In most cases, we identified which pathology the authors wanted to detect in the study.
- RQ2: What AI methods are used?
  We identified the type of AI algorithm used as a backbone for the XAI techniques.
- RQ3: What data are used?
  With this question, we wanted to know which dataset was used in each study.
- RQ4: Is code open access?
  With this question, we wanted to identify the studies with publicly available code.
- RQ5: Which type of XAI method is used?
  With this question, we classified the XAI methods used on the papers depending on their particular implementation.
- RQ6: Is a post hoc method or a model-based method used?
  Following the proposals of [17], we classified the studies depending on the nature of the explainability output.
- RQ7: What kind of explainability is obtained?
  Following the taxonomy proposed in [3], we classified the XAI methods depending on the output that was generated.
- RQ8: What metrics are used for the evaluation of the obtained explainability results?
  We wanted to identify which metrics were used to measure the quality of the explanation results.

2.2. Search Strategy

The search process was a manual search over multiple databases of conference proceedings and journals papers. To perform this task, we defined a query string (search terms) and we selected the papers of interest, as shown in Figure 1.

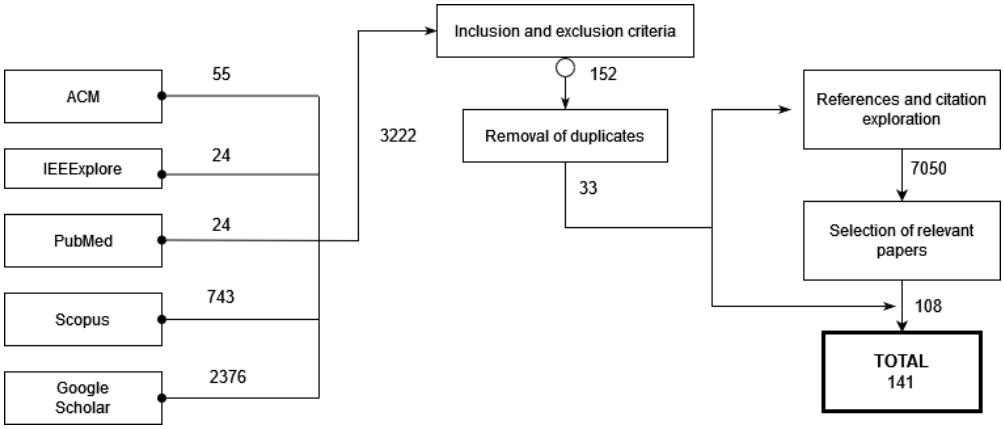


Figure 1. Research process.

2.2.1. Query Strings

Query strings were defined by deriving important terms from the RQs. These terms are "explainable artificial intelligence", "X-ray images", and "medical images". After a brief search, we found a few papers and extracted new significant keywords from them. These keywords are "XAI", "model explanation", "interpretable machine learning", and "explainability". By combining the keywords with the original query, we formed the final query string, as follows:

("XAI" OR "model explanation" OR "interpretable machine learning" OR "explainability" OR "explainable artificial inteligence") AND ("X-ray images" OR "medical images").

We used this search query in each selected database to find the primary studies.

2.2.2. Source Selection

This search, as can be seen in Figure 1, was performed on multiple digital databases: the ACM Digital Library, the IEEE Xplore Digital Library, Google Scholar, and SCOPUS. These databases were selected because they are the most popular academic indexes in the field of engineering and computer science. Furthermore, we also performed the search on PubMed because it is the citation indexing service for reference on databases regarding life sciences and biomedical information. To assure that we did not miss any important material, we performed secondary searches with the references and citations found in our primary results and first author publications for related works.

2.3. Selection Criteria

The most important inclusion criteria were whether the study used an XAI technique for the analysis of X-ray images. We only included studies that were peer-reviewed and written in English.

Otherwise, we excluded studies that did not use XAI. We also excluded those that did not apply these techniques to X-ray images. We did not find any systematic reviews; however, if we had, we would have excluded them.

2.4. Quality Assessment (QA)

Following the guidelines proposed in [15], we carried out the quality assessment for each included article. We did not find any standardized definition of how to measure the quality of this kind of study. We defined a set of QA questions aiming to assess the rigor, credibility, and relevance of the selected studies. These questions were based on previous methodologies. We selected three questions from those proposed in [18]. We selected three QA questions aiming to assert the relevance of the studies and their results. The CASP proposal explained how each of these questions can be measured:

QA1. Is there a clear statement about the aims and objectives of the research?

We wanted to know the goal and aim of the research. We focused on the description of the importance of the topic handled by the study and its relevance.

QA2. Is the research design appropriate to address the aims of the research?
We wanted to identify the justification of the research design. We checked whether the authors had discussed how they decided which methods should be used and why these were the best options. With this QA question and the previous one, we evaluated the RQ1 question.

QA3. Is the data analysis sufficiently rigorous?
This question aims to determine whether the analysis of the results is correct. We focused on whether the researchers analyzed the results numerically and whether they tried to overcome potential biases. This question allowed us to evaluate RQ8.

Another set of questions for quality assessment was obtained from [19]. The QA questions from this study aim to cover three aspects of the research: rigor, credibility, and relevance. We selected two questions from this study to, in addition to the previous ones, assert these three aspects. In this study, as in the CASP questionnaire, how each of the following questions can be evaluated is explained:

QA4. Is there an adequate description of the context in which the research was performed?
This question aims to identify whether the study correctly analyzes the context of the research. We assessed the identification and explanation of the existing methods and study.

QA5. Is there a clear statement of in the findings?
We wanted to determine whether the study provides credible results and clearly justified conclusions.

We defined a final question aiming to detect whether the results and algorithm are reproducible. This question is not found in any previous article. For this reason, we explained it in more detail:

QA6. Are the results easily reproducible?
We wanted to determine whether the results obtained were replicable. We assessed whether the algorithm and the data were openly available or not. We considered the results to be reproducible if the data and the original algorithm were available. If one of these two elements were not present, we considered the results to be partially reproducible. If neither of them were available, we considered the results to be non-reproducible. With this question, we assert the quality of RQ4 and RQ3.

Each question had three possible answers: "Yes", "Partly", and "No". These three answers had one associated value: Yes = 1, Partly = 0.5, and No = 0. The score for each study was obtained by summing the scores of the answers to each of the previous questions.

2.5. Data Extraction

From each study, we extracted the following information:

1. Title.
2. Source.
3. Year of publication.
4. Authors of the study.
5. Institutions and countries represented by the authors.
6. Keywords.
7. Aim of the study.
8. Pathology addressed by the study.
9. Datasets used in the study.
10. XAI technique used.
11. Kind of explainability addressed.
12. Metrics used to evaluate the XAI algorithm performance.

Data were organized in tables to easily present the basic information of each study. Tables were used to answer the research questions. These tables are available at https://github.com/explainingAI/SLR (accessed on 20 December 2021).

3. Results and Discussion

This section presents and discusses the findings of the review. First, an overview of the selected studies is presented. Next, the quality of the studies is defined through the already defined QA questions. Finally, the review findings of the RQs are reported and discussed.

3.1. Overview of the Selected Studies

In this review, we systematically evaluated 141 papers. Thirty-nine of them (27%) were published in conference proceedings; 102 (73%) in journals. All the papers we reviewed were published between 2017 and 2021, although we did not set any filter based on the publication year. These methods emerged in 2017, along with the increased usage of black-box models as convolutional neural networks in medical contexts. This interest further increased with the application of these algorithms to X-ray images caused by the emergence of COVID-19.

Figure 2 indicates the number of institutions per origin country of each study. Figure 3 shows the publishing year of each article. We found that most institutions only had one study considered in this review. Nevertheless, we found a set of them that stands out with multiple publications, including the NIH (National Institute of Health) with eight studies, Johns Hopkins University with seven studies, Stanford University with four articles, and the University of Toronto, Massachusetts General Hospital, and the University of California with three studies each.

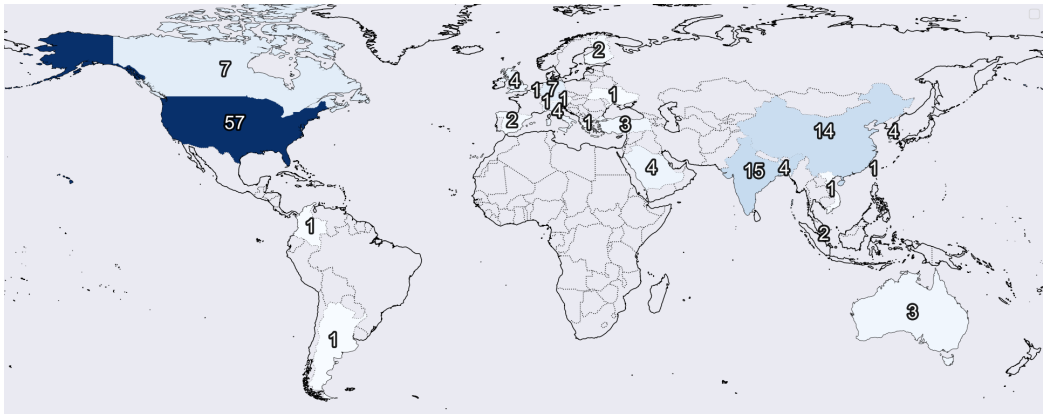


Figure 2. Map with the number of articles by country. A dark color indicates a higher number of published articles.

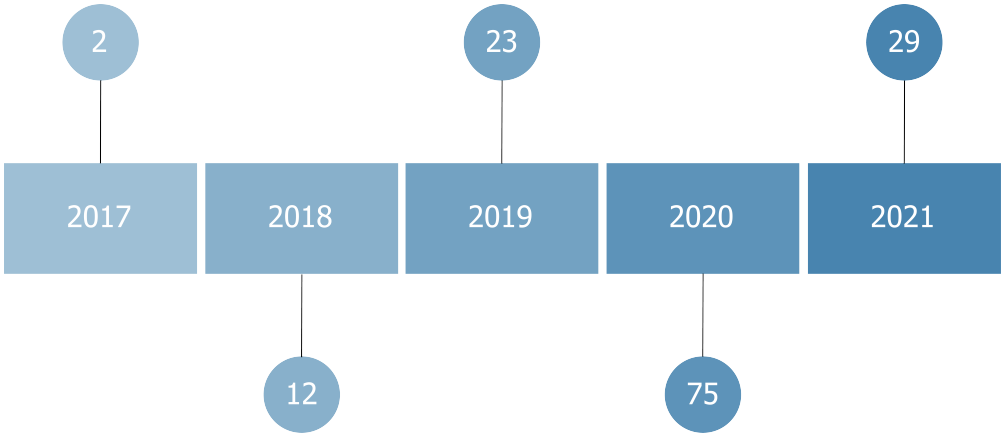


Figure 3. Number of published articles per year.

All the selected articles pertain to experimental research. No survey research was found. We filled in a form for each selected study https://github.com/explainingAI/SLR (accessed on 20 December 2021). To answer the RQs, we used plots and charts.

3.2. Quality Assessment

Using the previously defined QA questions, we obtained the results of each study's quality. The distribution of the punctuation can be seen in Figures 4 and 5: 2 articles were rated with 6 points (1.41%), 3 with 5.5 points (2.13%), 3 with 5 points (2.13%), 16 with 4.5 points (11.4%), 25 with 4 points (17.73%), 21 with 3.5 points (14.89%), 20 with 3 points (14.18%), 18 with 2.5 points (12.77%), 20 with 2 points (14.184%), 12 with 1.5 points (8.51%), and 1 with 1 point (8.51%).

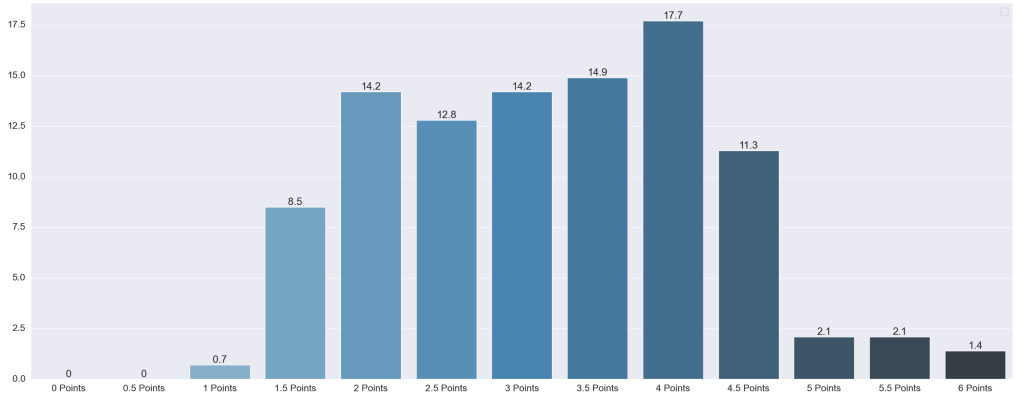


Figure 4. Total quality punctuation of each article. Vertical axis indicates the percentage of articles for a category.

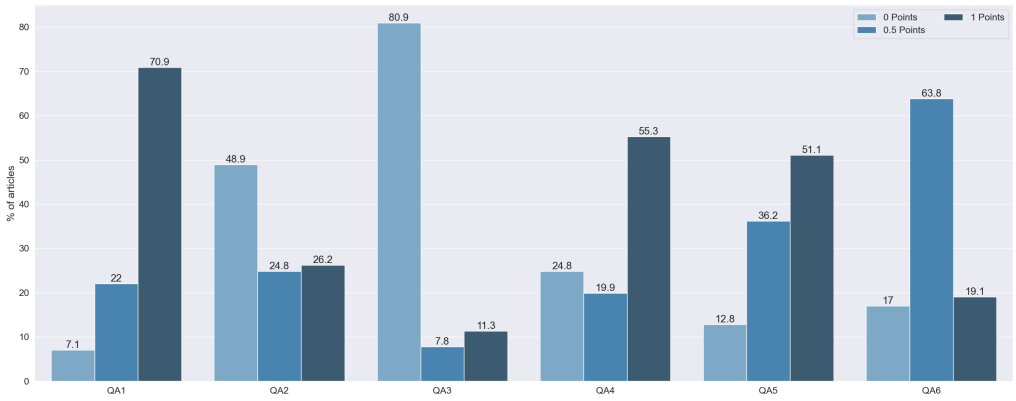


Figure 5. Total quality punctuation of each article by question. Vertical axis indicates the percentage of articles for a category.

QA1 and QA4 were the most fully answered questions, with 100 studies with 1 point on the first one and 78 with 1 point on the second one. These two questions tried to identify whether the studies provide suitable justification, taking into account the state of the art and the explanation of the topic. The importance of these questions is shown by the fact that we did not find any study that scored 0 points on both of these questions. Of the articles that did not have a full score for both questions, the majority did fail QA4, without explaining the state of the art, whereas only 10 did not indicate the aims and objectives of the research.

Sixty-nine studies had 0 points on QA2. These articles did not justify the selection of the explainable algorithm. Furthermore, they did not take into consideration any other technique apart from the one used.

One hundred and fourteen articles failed QA3. These papers were characterized as not using any kind of metric for the explainability results. From the rest of the methods,

we assessed whether they tried to overcome potential biases. Only 16 articles (9.76%) performed a rigorous data analysis without biases. On QA5, the majority of studies, 51.06% of them, had full punctuation. This score means that all of them made clear statements of the findings. Moreover, 36.17% of the articles only scored 0.5 points, penalizing the lack of clarity in the exposition of their findings. Finally, 12.76% of the articles failed this question due to the non-existence of any statement of the findings.

Finally, on QA6, 27 articles had full punctuation. Most of the studies (90 articles) had either open data or open code, with only 19 studies having neither. We observed that 81.56% of the articles made the data publicly available, while only 20.57% of the articles published their respective code.

We did not find any study with a complete failure to answer the QA questions (0 points in total). The articles with less punctuation scored 1 point. Furthermore, most of the reviewed articles (76.59%) had at least 2.5 points. This punctuation, half of the maximum score, can be seen as a quality threshold. Taking this into consideration, we can conclude that most articles had sufficient quality.

3.3. RQ1: What Is the Purpose of the Article?

With this question, we aimed to identify the main purpose of each study. Because we reviewed articles centered on X-ray images in a medical context, most of the articles were centered on the prediction or identification of a health problem.

We found 17 different purposes in the reviewed articles; see Figure 6. We classified these 19 goals into three classes: thoracic problems (86.52% of the reviewed articles), bone-related problems (9.93%), and others (3.55%). It is noteworthy that most of the articles were centered on the detection of thoracic diseases. This trend had increased in 2020 and 2021 with the appearance of COVID-19. We observed that the main thoracic disease handled by the reviewed articles was pneumonia (50.35%), whether it was COVID-19-related (42.55%) or not (7.80%). Other articles addressed this disease, but with a more general goal to detect multiple pulmonary diseases (28.37%). Further, 4.26% of the reviewed articles had the detection of tuberculosis as the main purpose. Finally, from the rest of the articles that handled thoracic diseases, we found multiple purposes: the detection of pneumothorax (1.42%), the detection of cystic fibrosis (0.71%), the detection of malposition of feeding tubes (0.71%), the segmentation of tracheal tubes (0.71%), and the detection of congestive heart failure (0.71%).

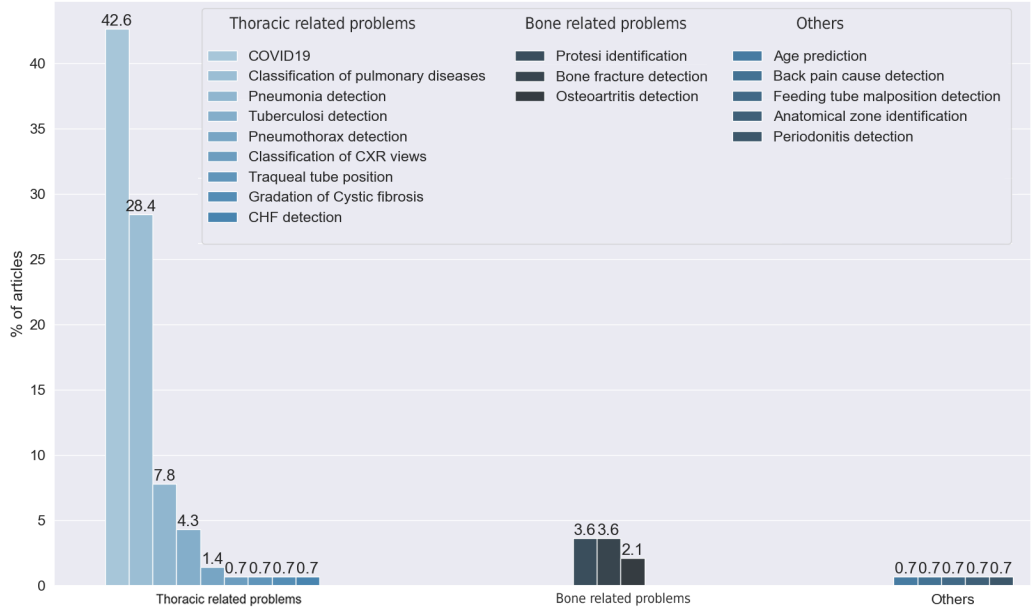


Figure 6. Answers to RQ1. Vertical axis indicates the percentage of articles for each category.

We found another set of articles centered on bone-related diseases. We classified these into three types, depending on the purposes addressed: the detection of bone fractures (3.55%), the identification of prosthesis type (3.55%), and the identification of osteoarthritis (2.13%).

Finally, one set of articles did not fall into any of the two previous categories: one study identified different anatomy zones from the image, and another one classified different CXR views. One study aimed to predict the age of the patients, another one diagnosed periodontitis, and one study recognized causes of shoulder pain.

Taking into consideration the results obtained from RQ1, we observed that the majority of the reviewed articles used X-ray images to detect either lung diseases (e.g., pneumonia) or bone-related problems. From these results, we can make two conclusions: that, in the medical community, apart from some particular cases, X-ray images are used for these two kinds of problems, and that these two kinds of problems are the only ones that are suitable for the usage of AI methods with X-ray images.

3.4. RQ2: What AI Methods Are Used?

With this question, we wanted to determine which backbone algorithm is used in each study. We found nine different AI methods used in the reviewed articles.

As we can see in Figure 7, 96.74% of the reviewed articles used multiple architectures of the convolutional neural network (CNN). We found that 77.30% of the articles used a standalone CNN. This kind of model is based on the proposal by [20]. These CNNs are characterized by the fact that they output a probability for each image to pertain to some predefined class. Moreover, 7.09% of the reviewed articles combined a CNN and a recurrent neural network (RNN). The CNN was designed to extract features from images, and the RNN was first introduced in [21] to handle data with time information. Both techniques were combined to generate text from the image. Another model used was the ensemble of multiple CNNs (2.84%). This approach consists of the combination of the output of multiple standalone CNNs. Further, 3.54% of the studies used a fully convolutional neural network (FCN), a special implementation of a CNN first introduced in [22] to develop a specialized algorithm for semantic segmentation. We also found that 2.13% of the articles used the method proposed in [23]. In this work, they presented a special kind of neural network called a siamese or twin neural network. These models are built upon two sub-neural networks, and the final result was obtained by a comparison of these two to find similarities between images. This allowed for a comparison between an unknown image with a known image. We found that 2.13% of the articles used multiple instance learning (MIL), a method formalized in [24], in combination with a CNN. We also found one study that used CNN multitasking, an AI method first proposed in [25] that improves generalization by using the domain information contained in the training signals of related tasks as an inductive bias.

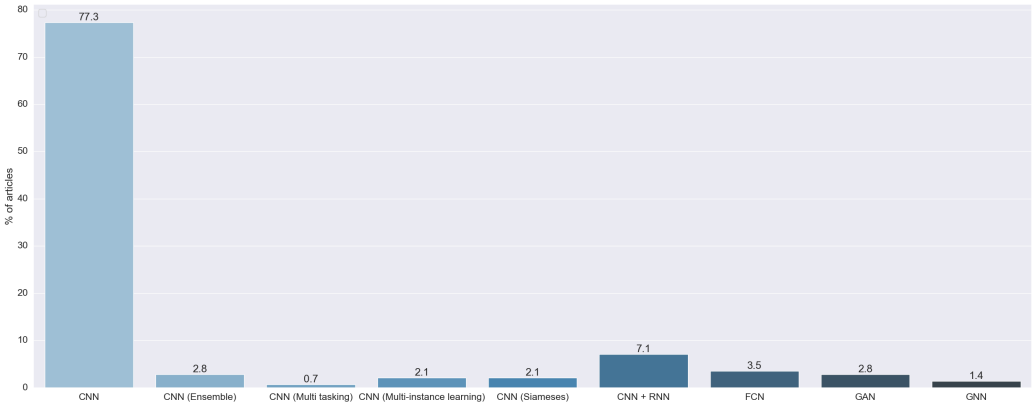


Figure 7. Answers to RQ2. Vertical axis indicates the percentage of articles for each category.

We found six studies that did not use a CNN as a backbone algorithm to perform the prediction. We found four articles that used a generative adversarial network (GAN). This technique was firstly introduced in [26] to generate realistic images. Two studies used a graph neural network (GNN), first introduced in [27], that combined a neural network with graph theory.

Taking into consideration the results obtained with RQ1 and comparing them with those obtained with this question, as shown in Figure 8a,b, we can assert that the purpose did not affect which AI method was used. We can see the predominance of the CNN in each purpose, with at least 75% of articles in each category using this technique. Furthermore, the value is even higher if we consider the different kinds of CNNs as the same AI methods. The predominance of this method is because it is specially designed to work with images without taking into consideration the type of image or the anatomical zone depicted in them.

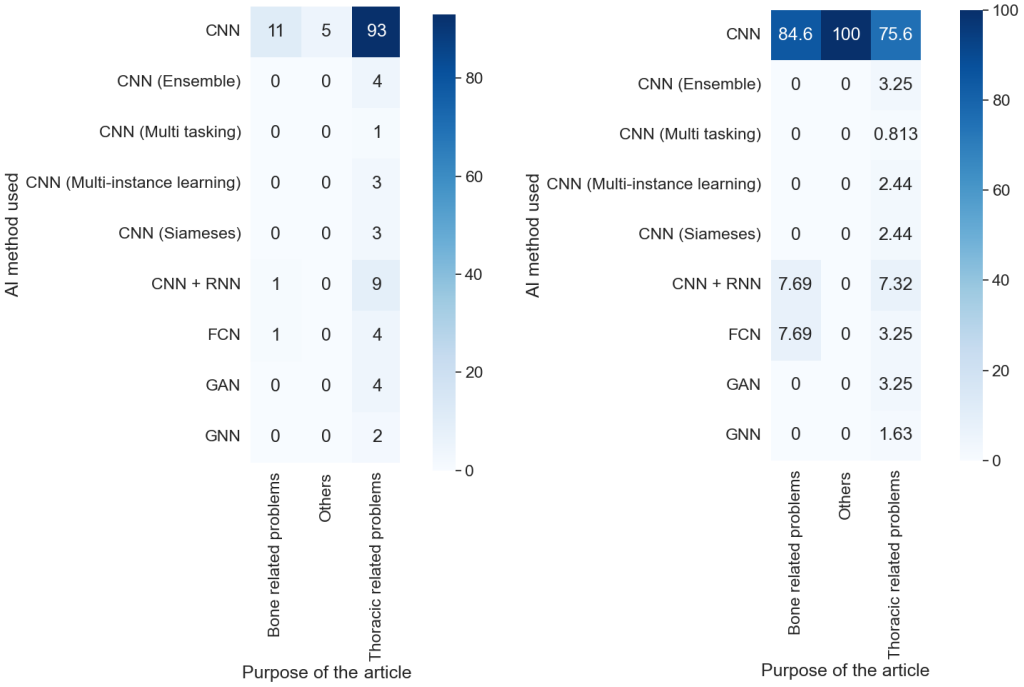


Figure 8. Heat maps comparing the results of RQ1 and RQ2. (**a**) In absolute values; (**b**) normalizing the results by column.

By answering this question, we found that all articles reviewed used deep learning methods. We did not find any method that combines traditional machine learning techniques with explainability. This was caused by the fact that the explainability concept was first defined to overcome the limitation of deep learning in black-box models.

3.5. RQ3: What Data Are Used?

The answers to this question allowed us to observe the large number of datasets used by the reviewed studies. First, we found that 18.43% of the articles did not indicate which data they used, or the data were not publicly available. The rest of the studies used 26 different datasets, as can be seen in Figure 9. Most of the articles used two or more of these datasets. We grouped the datasets into five categories depending on the goal of the dataset: pulmonary diseases, COVID-19, tuberculosis-related disease, bone-related diseases, and dental images. We distinguished COVID-19 and tuberculosis from the rest of the pulmonary diseases due to their relevance and the amount of interest in both diseases in the reviewed studies.

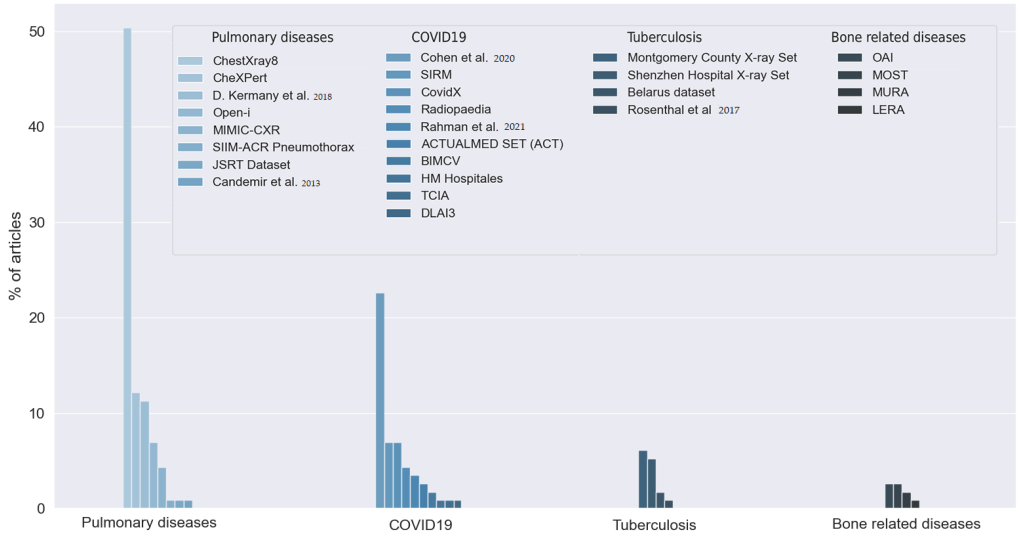


Figure 9. Datasets used by the reviewed articles. Answers to RQ3. Vertical axis indicates the percentage of articles for each category. D. Kermany et al. [28], Candemir et al. [29], Cohen et al. [30], Rahman et al. [31] and Rosenthal et al. [32].

The reviewed articles used 10 different datasets for COVID-19 detection. Twenty-six (22.61%) articles used the dataset proposed in [30], which provides a set of X-ray images of multiple diseases, including COVID-19. This dataset was built using data from multiple previously published datasets. The dataset from [33] was used in eight studies (6.12%). The authors in [34] published another COVID-19 dataset, the COVIDx. This dataset was used in eight articles (6.12%). Five studies (4.35%) used images available at [35], a public repository for X-ray images. The authors in [31] published a dataset containing more than 20,000 samples of X-ray images. Four articles used Rahman et al.'s dataset. Two studies used the BIMCV COVID-19 dataset, obtained in the Valencian region of Spain [36]. Three datasets were used by only one study: the COVID-19 dataset [37], from The Cancer Imaging Archive (TCIA), the dataset from [38], and the dataset from the 3rd Deep Learning and Artificial Intelligence Summer/Winter School (DLAI3), made available at Kaggle [39].

The studies that aim to detect tuberculosis used multiple datasets. The Montgomery County X-ray Set and the Shenzhen Hospital X-ray Set are both presented in [40] and contain X-ray images of patients with tuberculosis. Six studies used one or both of these datasets. Two studies used the Belarus dataset. One study also used the dataset published in [32].

We found eight different datasets for the detection of multiple pulmonary diseases. Fifty-eight articles (50.43% of the articles) used the ChestX-ray8 dataset. This dataset was firstly introduced in [41] and consists of 112,120 chest X-ray images. The authors in [42] proposed the CheXPert dataset, consisting of chest X-rays of healthy patients and patients with pulmonary diseases. The dataset was built with 224,316 chest radiographies. Fourteen studies used this dataset, representing 12.17% of the reviewed works. Thirteen studies (11.3%) used the dataset proposed in [28]. This dataset contains two kinds of chest X-ray images and optical computerized tomographies (OCTs). IU (Indiana University) X-ray images, also referred to as the Open-i dataset, are provided in [43], and eight studies used them (6.96%). Five studies used the data provided in [44], known as the MINIC-CXR collection. The main feature of this dataset is the combination of X-ray images with their respective radiology reports. One study used the pneumothorax dataset published in [45], and another one used the dataset published in [46], which contains two kinds of X-ray images with or without lung nodules. Only one study used the dataset proposed in [29]. This dataset is a combination of the JSRT, the Montgomery County set, and a set of images from India containing images of healthy and ill patients.

Nine studies used five different datasets centered on bone-related detection: The MURA dataset proposed in [47], the MOST dataset proposed in [48], and the OAI dataset presented in [49] were centered on osteoarthritis disease. One study used the LERA dataset proposed in [50].

Finally, one study used two datasets of dental images: the Suzhou Stomatological Hospital dataset and the Zhongshan dataset. Both datasets indicated whether the X-ray image was from a healthy patient or a patient with periodontitis.

We can see that the results obtained from this answer are fully compatible with those obtained with RQ1. The purpose of the article influenced the selection of the dataset. Both questions show the prevalence of pulmonary diseases as the main issue of the research.

3.6. RQ4: Is the Code Open Access?

The results of RQ4 show that most studies (79.29%) do not make the algorithm publicly available. The rest of them (20.71%) used Github to publish their code, except for one study that used Bitbucket. This small number of articles with open access codes is compatible with the fact that none of them made their data publicly available.

3.7. RQ5: Which Type of XAI Method Is Used?

The results of RQ5 allow us to observe that 18 studies did not indicate which explainable algorithm they used; for example, one study output a saliency map for explainability, but the authors did not indicate which method was used. We also observed that there were techniques that obtain explanations via the artificial intelligence algorithm itself (model-based methods). This special case was addressed with RQ2. We did not consider these studies for further analysis of this research question.

Figure 10 shows the prevalence of techniques that generate visual explanations. This was expected because we centered the search process on studies that analyze X-ray images. The most used techniques were class activation map (CAM) methods, which were used in 72 studies (64.86%). We considered CAM as a set of techniques based on the original work of [5]. In this work, they developed the original CAM method, aiming to identify which regions of the image were more important for a CNN to make its prediction. To achieve this, they used global average pooling, a well-known technique in the field of deep learning. Continuing the work of Zhou, the authors in [51] proposed GradCAM. Its main contribution was the generalization of the previous method through the use of gradients of the classes, so they did not need to modify the structure of the original network. GradCAM++ was a further improvement, proposed in [52], that aimed to increase the precision of the location of the objects and the explanation of multiple objects in the same image. The results of these three techniques were saliency maps.

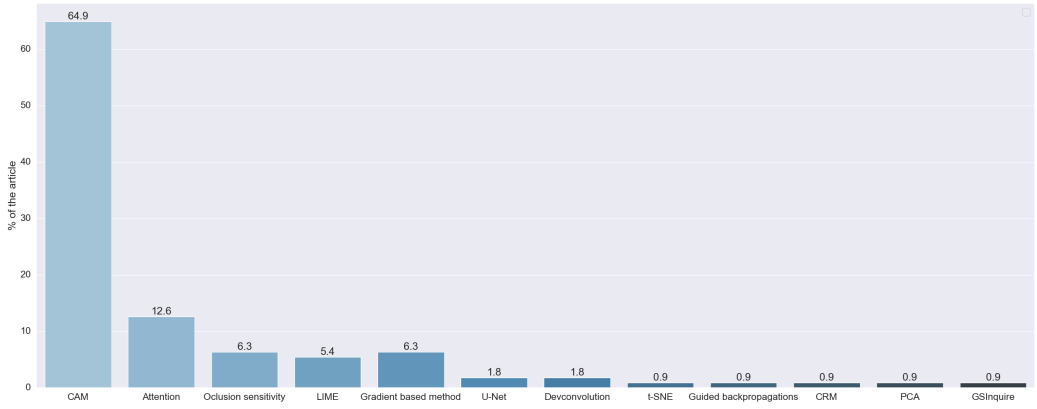


Figure 10. XAI techniques, obtained with RQ5. Vertical axis indicates the percentage of articles for each category.

The next most used type of method is based on attention, with 12.61% of the studies using them. These methods generated a saliency map through built-in recurrent methods. One of the first papers that proposed this technique is [53], which uses an RNN to select the most important part of the input image.

Agnostic methods were also used. These methods are fully independent of the backbone model, working even with unknown techniques. These methods were used by 11.71% of the reviewed studies. Occlusion sensitivity was the most used agnostic algorithm, with 6.31% of the studies using it. This method aimed to generate a saliency map by systematically occluding different portions of the input image and analyzing how this affects the performance of the algorithm. Another agnostic method is LIME, first introduced in [54], and was used in 5.41% of the studies. This method learns an interpretable model locally around the prediction of the black-box model. To be able to achieve this, the method defines the locality with occlusions methods. Because it uses a simpler model to explain a complex one, LIME is also considered a surrogate method.

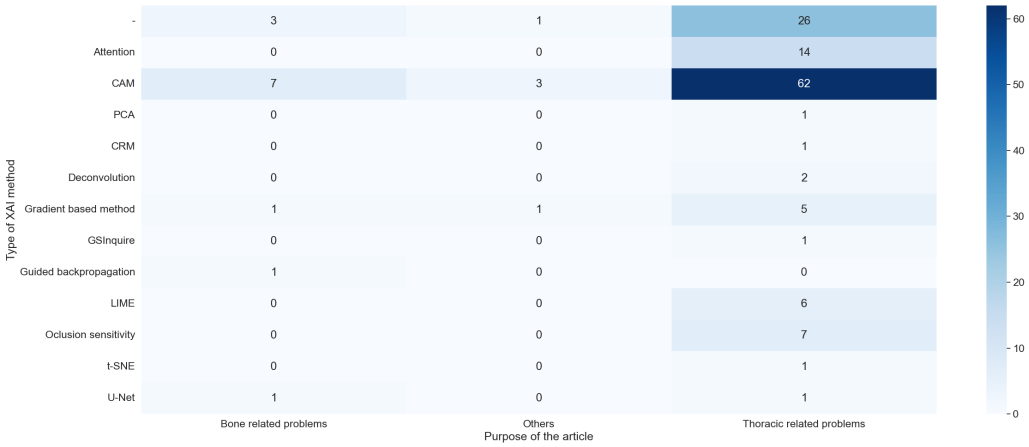
We also found that 6.31% of the studies used techniques based on the calculation of the gradient of the classification with respect to the input image, as proposed in [55,56] and [57]. Two studies used the U-Net, a neural network for segmentation introduced in [58], also to obtain saliency maps. Two articles used deconvolution, a well-known technique, to obtain these maps. The authors in [59] proposed a combination of deconvolution and backward pass. They called this approach guided backpropagation, which was used in one study. One study used the method proposed in [60], GSInquire, that combines methods to obtain a saliency map similar to GradCAM, with occlusion techniques to obtain and measure the importance of a part of the input data. One study used the technique proposed in [61], class-selective relevance mapping (CRM), which removes parts of the feature map of the last layer of the CNN to generate a saliency map.

Apart from the saliency map, one study accomplished visual explanations through dimensional reduction. To achieve this, it used the well-known t-SNE technique. Three studies combined two or more of these techniques, two that used CAM methods and LIME, and one that combined GradCAM with PCA.

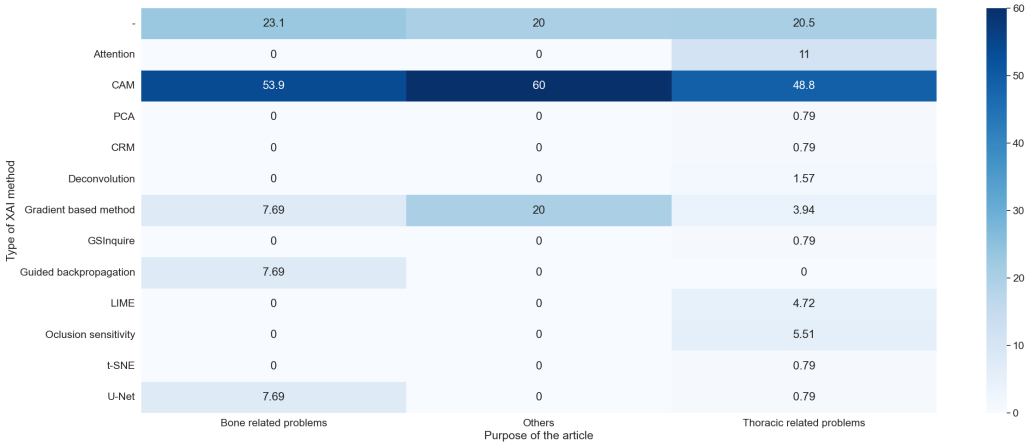
The explainable algorithm depends, to some extent, on the purpose and AI method used. These two features were analyzed in questions RQ1 and RQ2, and are compared with the results of this question in Figure 11a,b and Figure 12a,b, respectively.

We can easily see in Figure 11a,b that the predominance of CAM methods did not depend on the purpose of the article. It is also clear that the articles that handle thoracic problems used more diverse XAI methods. This diversity can be attributed mainly to two causes: It is the category with more articles, so it is expected that we find more, different, XAI methods, and there is a diversity of datasets available about thoracic problems, allowing for more complex methods of XAI.

Taking into consideration the results obtained with RQ2, we can assert that the studies that indicate the method used were those that generated visual explanations and used post hoc methods, as can be seen in Figure 12a,b. This is due to the lack of diversity in the rest of the explainability approaches; for example, all studies that aimed to generate text explanations combined CNN and RNN.

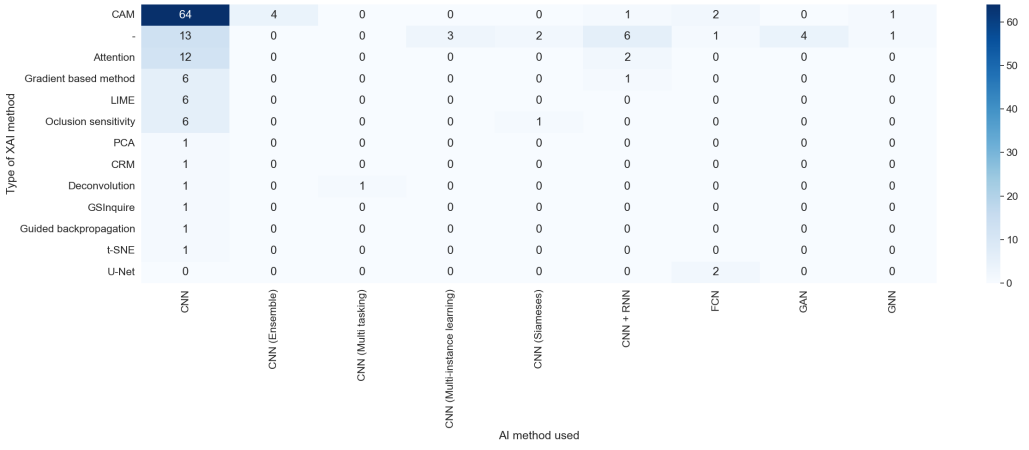


(**a**)



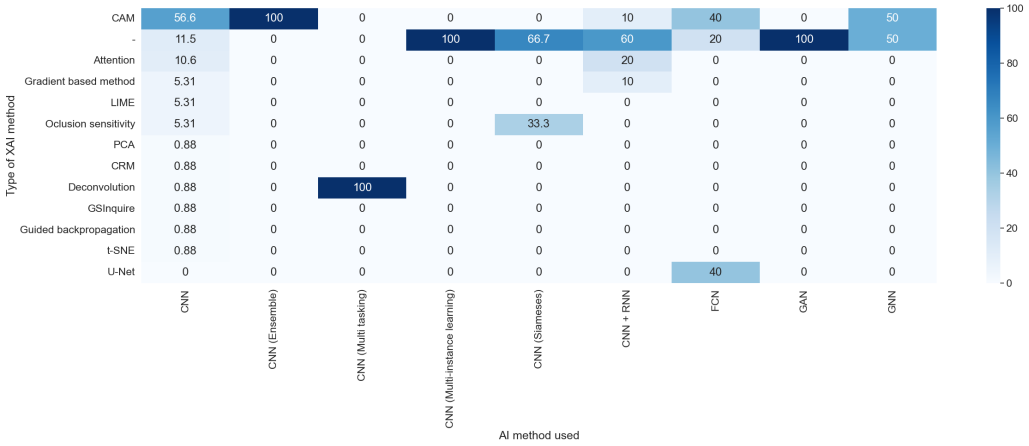
(**b**)

Figure 11. Heat maps comparing the results of RQ1 and RQ5. All articles that did not indicate the type of XAI method used are shown. (**a**) In absolute values; (**b**) normalized results by column.



(**a**)

Figure 12. *Cont.*



(**b**)

Figure 12. Heat maps comparing the results of RQ2 and RQ5. All articles that did not indicate the type of XAI method used are shown. (**a**) In absolute values; (**b**) normalized results by column.

3.8. RQ6: Post Hoc or Model-Based?

The work of [17] introduced a categorization of existing explainable techniques into model-based and post hoc categories. A model-based technique is defined as the construction of models that readily provide insight into the relationships they have learned. The main challenge of model-based explainability is to come up with models that are simple enough to be easily understood by the audience yet sophisticated enough to properly fit the underlying data. A post hoc technique is defined as the analysis of a trained model in order to provide insights into the learned relationships. When there is an interest in more general relationships learned by a model, e.g., relationships that are relevant for a particular class of responses or subpopulation, they use dataset-level explainability.

Figure 13 shows the different types of methods detected in the answers to RQ6. We can see that most of the studies, 70.92%, use post hoc techniques, while 25.53% use model-based methods. This difference is caused by the greater simplicity of the post hoc approach.

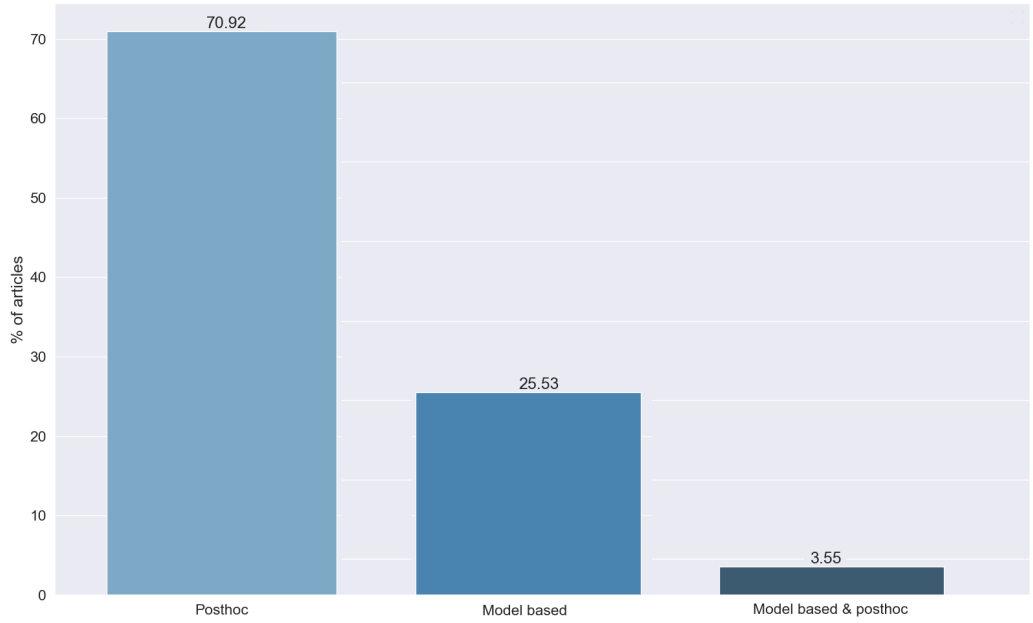
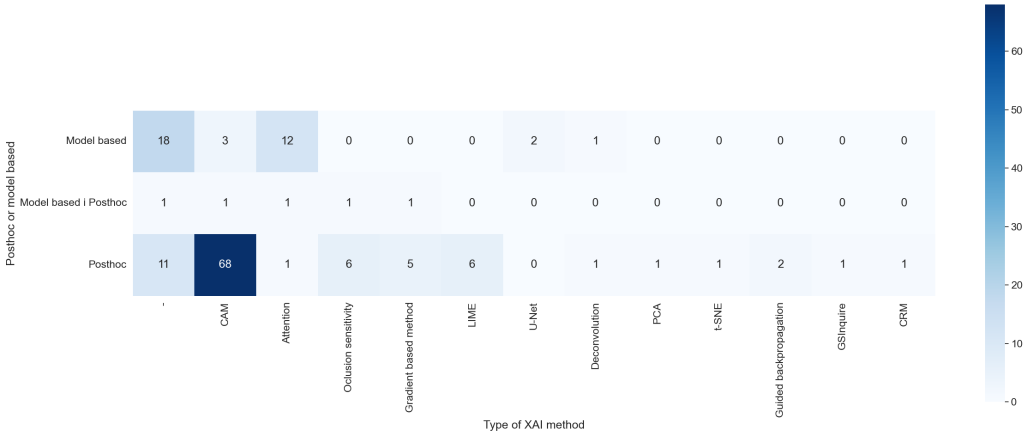


Figure 13. Type of method used, obtained from the answers to RQ6. Vertical axis indicates the percentage of articles for each category.

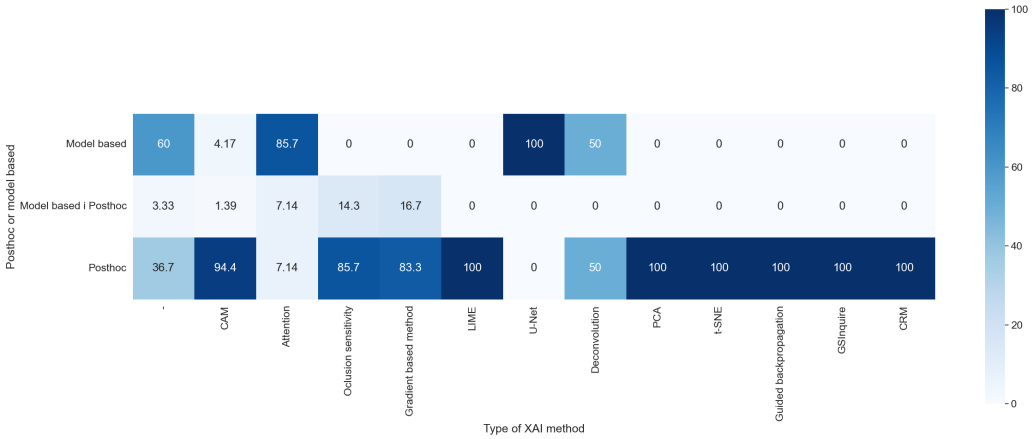
Finally, we found a set of methods (3.55%) that combined both approaches. To combine them, most articles used the results of the post hoc method as an input for the transparent model to gain explainability as a whole.

Figure 14a,b allowed us to check which XAI method is considered post hoc or model-based. We compared the results of RQ6 with those obtained with RQ5. Most of the algorithm is only in one category; for example, all techniques of features visualization were always considered as post hoc models, but some algorithms, even if they were a post hoc model, could be used as a foundation for model-based techniques. Accordingly, 4.17% of the studies that used CAM methods integrate them into models and, for this reason, are categorized as model-based. The results of this comparison were expected, as the XAI method is the element that determines if a method is post hoc or model-based.

Taking into consideration the results obtained with RQ1 and comparing them with those obtained with this question, as shown in Figure 15a,b, we can see that the model-based methods are predominantly used in the studies centered on thoracic problems. This can be explained by the availability of comparatively large and high-quality datasets with thoracic images, which allowed the use of more complex methods such as GANs or methods for text generation. These results are also compatible with, and similar to, those obtained to compare RQ1 and RQ5. The cause of this similarity is discussed in the previous paragraph.

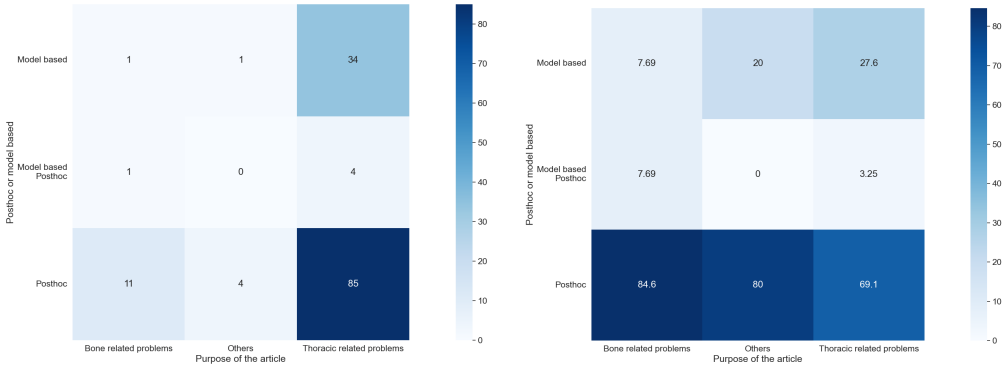


(**a**)



(**b**)

Figure 14. Heat maps comparing the results of RQ5 and RQ6. All articles that did not indicate which type of XAI method used are shown. (**a**) In absolute values; (**b**) normalized results by column, values in percentages.



(**a**) (**b**)

Figure 15. Heat maps comparing the results of RQ1 and RQ6. (**a**) In absolute values; (**b**) normalized results by column.

The larger use of post hoc models was caused by their inherent simplicity in comparison to model-based approaches. The usage of model-based models requires more extensive knowledge of the algorithm to accomplish the goal of both transparency and good performance. In contrast, post hoc methods are simpler to use and require less expertise, and can even be agnostic, meaning that they did not depend on the underlying method, such as LIME.

3.9. RQ7: What Kind of Explainability Is Obtained?

This question aims to classify the reviewed studies depending on their respective output. These outputs represented a different type of explainability, independently of the underlying technique used. Distinct techniques could generate similar outputs. Figure 16 indicates the results obtained from this research question.

The authors in [3] proposed a taxonomy for the different types of explainability. Based on this taxonomy, we identified three different types of explainability used for X-ray images:

- Visual explanations. Techniques that aim to visualize the model behavior. In this category, we include dimensional reduction techniques and saliency maps. The techniques of dimension reduction, such as t-SNE (introduced in [62]), allowed for human interpretations via the simplification of the handled data. Saliency maps were first introduced in [63], aiming to visualize and identify the significant visual features from an image for a model of artificial intelligence.
- Text explanations. A set of techniques based on text generation algorithms to obtain explanations in natural language.
- Explanations by example. Based on case-based reasoning (CBR), these are characterized by the search for a previously known image to explain the decision made by artificial intelligence.

The findings of this question are highly related to those obtained with RQ5. Each of the techniques used in the reviewed studies used one of the methods detected with RQ5.

We found that the kind of explainability most often used is a visual explanation. In particular, we found that 85.81% of the articles generated a saliency map. The predominance of this technique was expected, taking into consideration that we reviewed articles that analyzed X-ray images and that the technique was specially designed for image analysis.

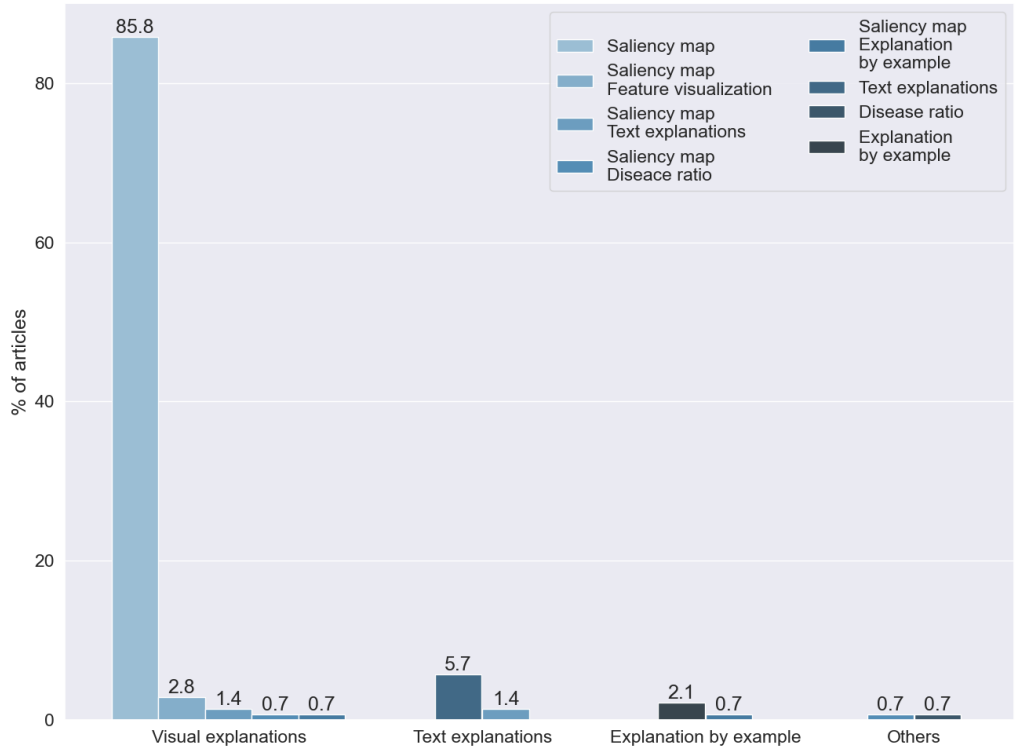


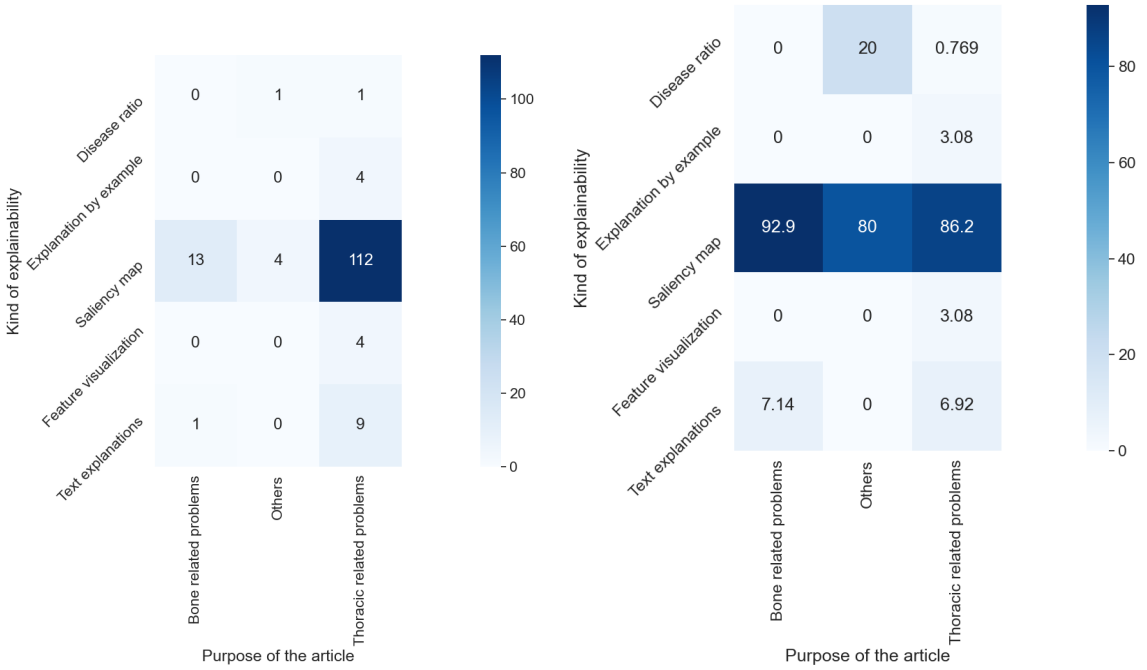
Figure 16. Kind of explainability found by answering RQ7. Vertical axis indicates the percentage of articles for each category.

Text explanations were the second most common approach for explainability, used for 5.67% of the studies, and 2.12% of the articles were based on techniques of explanation by example. One study proposed the creation of a numerical ratio to grade the probability of the presence of disease.

The rest of the articles combined two or more of the previous techniques: saliency maps and dimensional reduction techniques (2.836%), text explanations and saliency maps (1.418%), saliency maps and explanations by example (0.709%), and disease ratios and saliency maps (0.709%).

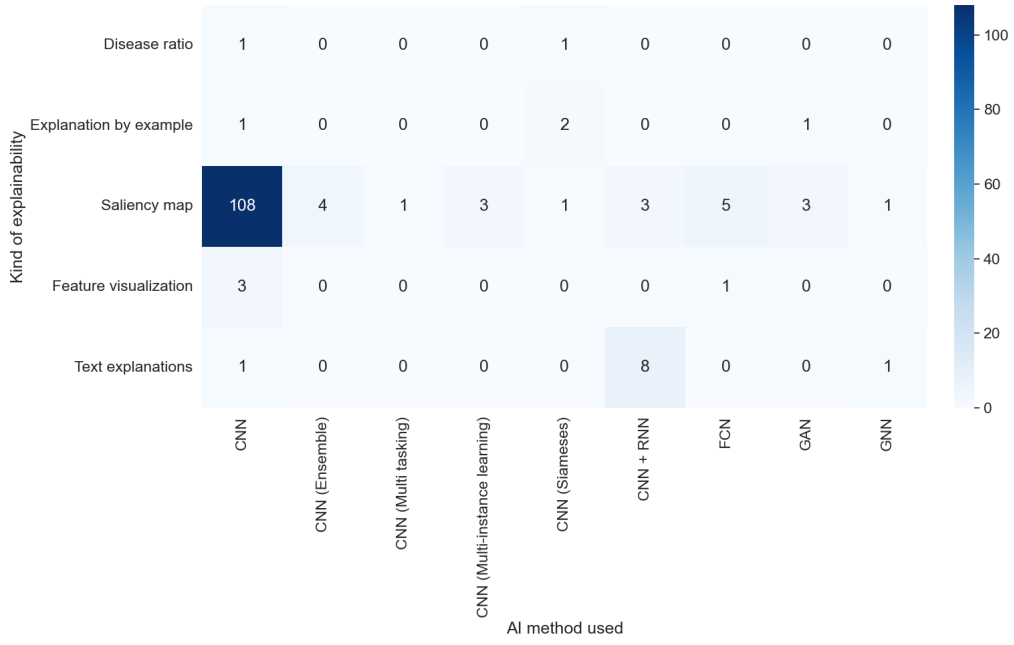
The results of this research question are highly related to those obtained in RQ1, RQ2, and RQ5. The relation between RQ5, RQ2, and RQ1 is discussed in previous sections; thus, it can be expected that similarities between them and RQ7 are found.

We found, in the comparison between RQ1 and RQ6, that the vast majority of model-based methods were used for thoracic problems. This relation is also observable in the relation between the results of RQ7 and RQ1, as can be seen in Figure 17a,b. Figure 18a,b allowed us to check if there was some relation between the AI method used and the kind of explainability obtained. We detected two different elements: a predominance of saliency maps, without a meaningful difference between different AI methods, and a relation between some kind of explainability and the AI method; e.g., text explanations are almost only obtained with a combination of a CNN and an RNN. In Figure 19b, we can see that the majority of articles that indicated which XAI method was used generated saliency maps. By contrast, in Figure 19a, it is clear that all articles that did not indicate the XAI algorithm or that used model-based approximations generated a vastly more diverse set of algorithms.



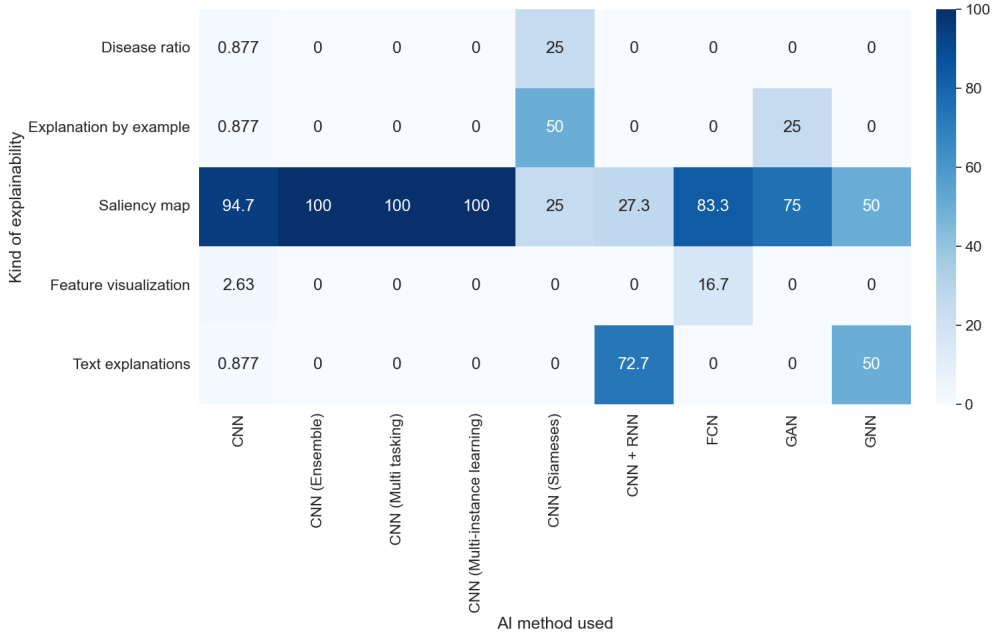
(**a**) (**b**)

Figure 17. Heat maps comparing the results of RQ1 and RQ7. (**a**) In absolute values; (**b**) normalized results by column, values in percentages.



(**a**)

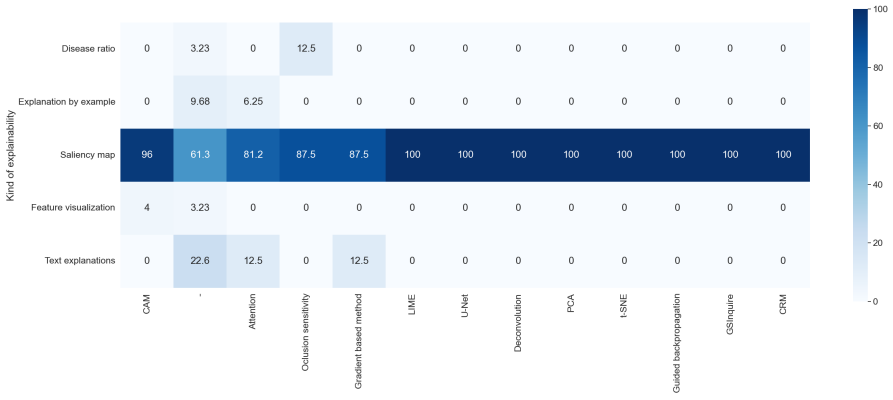
Figure 18. *Cont.*



(**b**)

Figure 18. Heat maps comparing the results of RQ2 and RQ7. (**a**) In absolute values; (**b**) normalized results by column, values in percentages.

(**a**)



(**b**)

Figure 19. Heat maps comparing the results of RQ5 and RQ7. (**a**) In absolute values; (**b**) normalized results by column, values in percentages.

3.10. RQ8: What Metrics Are Used for the Evaluation of the Obtained Explainability Results?

We can see that most (81.56%) of the reviewed articles did not evaluate the explainability of the method, as can be seen in Figure 20. For this reason, these articles did not use any metrics for explainability. This is caused by the nature of the labels of the dataset. Many datasets only have a binary label to indicate the presence of a disease, precluding the results of the explainability from being compared to any ground truth. For example, most of the methods based on visual explanation were unable to make comparisons with any verified information.

To analyze the outcome of this research question, we did not take into consideration the articles that did not make any measure of the results. The rest of the methods, as can be seen in Figure 21, used 11 different metrics. We divided these metrics into two main groups: metrics for segmentation and metrics for text generation.

Segmentation metrics were used by 28.947% of the articles analyzed. We classified four different metrics in this category: the intersection over union (IOU) developed in [64] and its multiple modifications, the mean intersection over union (mIOU), the intersection over bounding box (IoBB), and the Dice coefficient, proposed independently in both [65,66]. All of these metrics have been used for segmentation problems; in this context, they were used to measure the performance of visual explainability techniques.

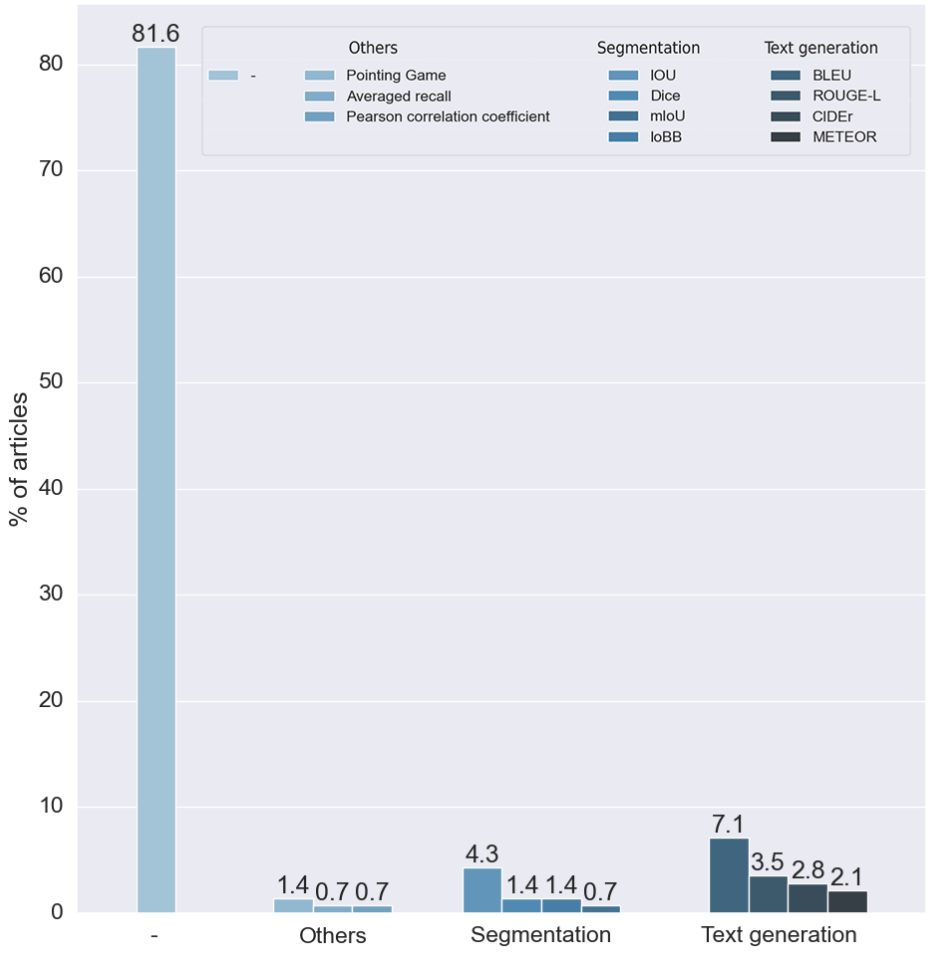


Figure 20. Metrics used for the evaluation results of RQ8. Vertical axis indicates the percentage of articles for each category.

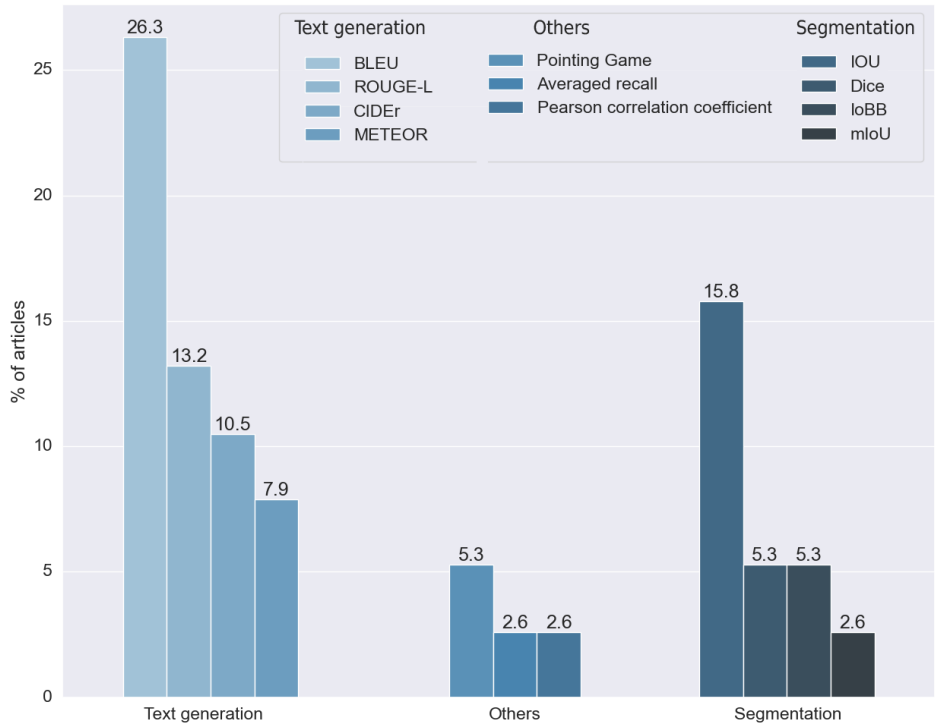


Figure 21. Metrics used for the evaluation results of RQ8, without considering the articles that did not indicate which metric was used. Vertical axis indicates the percentage of articles for each category.

Four different metrics were used in the text generation methods: the Bilingual Evaluation Understudy (BLEU), first introduced in [67], the Metric for the Evaluation of Translation with Explicit ORdering (METEOR), proposed in [68], ROUGE, a set of measures proposed in [69], and Consensus-Based Image Description Evaluation (CIDEr), presented in [70]. All these metrics were introduced to calculate the performance of standard text generation. For these reasons, they were used in articles that made explainability through text explanation, allowing the results obtained with expert annotations to be compared. Approximately 57% used one or more of these metrics.

We found four metrics in two different studies that were not suitable for any of the two proposed groups. Two studies used the pointing game benchmark, proposed in [71], which aims to *compare the spatial selectiveness of top-down saliency maps*. One study used averaged recall, a metric designed for the CBIR (content-based image retrieval) algorithm. Finally, one study used the Pearson coefficient to compare a ratio calculated by the model with one calculated manually by experts.

As we can observe, the metric used is highly related to the data used (RQ3) and the kind of explainability (RQ7). The metrics used depend on the output obtained, as in the case of text generation methods that use a specific kind of measure. Furthermore, they can only be applied to specific data types. For this reason, we can relate the datasets used to their respective metrics.

4. Limitations of the Review

This systematic review considered 141 studies to evaluate and assess the performance of various explainability algorithms in the X-ray image context. The limitation of this review is that only 16 of the 141 studies compared the results of multiple XAI methods. Therefore, the comparison between different algorithms is not conclusive. The other limitation is that the study selection bias threatens this review's validity. The selection of studies depends on the search strategy, literature sources, and selection criteria. We defined the search query using the RQs. Furthermore, we retrieved the relevant studies from five electronic databases. However, some relevant studies may not have used the terms related to the

RQs. Therefore, we may have excluded these studies. To reduce this threat, we manually scanned the references and citation list of each relevant study to search for other relevant studies that were not included in the initial search. We believe that the number of studies that we missed is small.

5. Guidelines for Practitioners and Scientists

The findings from this systematic literature review inspired us to define some guidelines for practitioners and scientists. The goal of these recommendations focuses on the fact that the field requires more robust research methods, more clarity in dissemination, more reproducibility, and more explainability metrics. The use of explainability should help to improve and upgrade AI models and to yield accomplishments in legal, ethical, and social aspects. From our results and discussion, we derived the following guidelines:

5.1. Reproducibility

Twenty-nine studies were fully reproducible. We recommend publishing the datasets and algorithms used. We also recommend publishing the raw data of the results to facilitate others researchers to compute other metrics. This would provide an accurate comparison between different approaches in solving the same problems. To facilitate this task, the authors could carry the proposed QA out and respond to the proposed RQ before submitting any work.

5.2. Explainability Metrics

In this SLR, 81.56% of the reviewed articles did not evaluate the explainability of the method and did not use any metrics for explainability. Among the studies that did use metrics, we observed that most of them were based on text explanation generation and that the existence of metrics helps to obtain more useful methods.

According to [4], what is missing is a standard procedure to measure, quantify, and compare the explainability of enhancing approaches that allow scientists to compare these different approaches. In the case of classification, the metrics are clear (raw data, accuracy, recall, and the F1 score, among others). According to [72], most of the work about explainability relies on the authors' intuition, and an essential point is to have metrics that describe the overall explainability with the aim to compare different models regarding their level of explainability.

The authors in [73] discussed specific methods for evaluating the performance of an explanation for an AI system. Regarding healthcare diagnosis, the authors in [74] explained that understanding the classification through appropriate explanations helps users to trust the system.

For diagnostic support using X-ray image analysis, experts need reasons or justifications for any result of the AI system, particularly when unexpected decisions are made [2]. Moreover, measuring explainability is not only for justifying decisions; it can also prevent errors. Understanding system vulnerabilities can help to avoid erroneous predictions of the system. Therefore, future studies must compare the performance of the explainability to analyze and discuss the results. All studies should use standard metrics to measure their performance. These metrics differ depending on the explainability type obtained. In the case of medical images, this comparison should be performed with expert information.

5.3. Upgrade and Discover

We did not find any study that applies the information obtained from the explainability to upgrade the original model. A model that can be explained and understood is one that can be more easily improved. Thus, XAI could be the foundation for ongoing iterations and improvements between humans and machines [2].

Moreover, XAI can be useful to discover new facts. If XAI models indicate something unknown in X-ray image analysis, experts must analyze it to determine whether it is a new discovery or an error of the system [2].

Therefore, future studies should concentrate on this approach to improve the results of the backbone model. It is also advisable to use the information obtained from the XAI algorithm to improve the knowledge of the disease and/or the diagnosis (if applicable). The explainability results can produce new insights into the original problem.

5.4. Legal, Ethical, and Social Aspects

None of the reviewed studies applied the information obtained from the explainability to detect biases or to check for accomplishments in legal, ethical, and social aspects. Compliance with these issues is a priority for AI systems in X-ray image analysis contexts because they affect humans and can help experts and society to adopt such systems.

In addition, X-ray image analysis using AI must offer to defend algorithmic decisions in an auditable and demonstrable way to be fair and ethical, which leads to trust-building [2]. Moreover, methods must comply with legislation, such as the General Data Protection Regulation (GDPR) in the EU, which includes the *right to explanation* [75]. Therefore, future studies should use explainability results to verify these aspects.

6. Conclusions

This systematic literature review investigated explainable artificial intelligence for X-ray image analysis. We defined and answered several research questions to analyze the state of the research in this area according to [15] guidelines. The primary findings are summarized as follows:

- RQ1. The different purposes found in the reviewed articles can be grouped into two main groups: methods that analyzed chest X-ray images and methods that analyzed bone X-ray images.
- RQ2. Many different AI models were used as a backbone for the XAI algorithms, but all of them, except one, used different implementations of CNNs. All studies used deep learning algorithms because of their good results in comparison to non-deep methods (e.g., decision trees and logistic regression) and the fact that most of these algorithms are explainable by themselves.
- RQ3. There were several datasets used in this context. All datasets used can be grouped depending on the anatomy region they describe. The results of this research question are coherent with those obtained with RQ1. We identified mostly two types of datasets: bone-related diseases and lung-related diseases.
- RQ4. Our findings show that the majority of the methods are not open. Of those that freely share their code, the most used platform was Github.
- RQ5. The reviewed studies used many different XAI methods. Most of them produced saliency maps. The usage of CAM techniques stands out, with more than half of the reviewed articles using them. The predominance of these techniques is highly related to the kind of data handled, the X-ray images, and the backbone methods (CNNs).
- RQ6. Using the taxonomy proposed in [3], we classified the reviewed studies by the method used: post hoc methods, model-based methods, or a combination of both. We compared them, and we detected the predominance of post hoc models. We found that the cause of this predominance was the simplicity of these methods compared to model-based algorithms.
- RQ7. We identified three different kinds of explainability obtained in the reviewed studies: visual explanation, text explanation, and explanation by example. All these three categories are highly related to the method used, which is based on RQ5.
- RQ8. Our findings show that most of the reviewed articles did not use any metrics to measure the quality of the explainability. Only 18.46% of the articles used them. The methods that used metrics can be divided between those based on text generation, which used metrics to compare the generated text with ground truths, and those that output visual explanations. In this case, the metrics used were those typically used for segmentation metrics. This research question allowed us to see the absence of

measurements in these methods and to see how metrics are heavily related to the data used (RQ3) and the kind of explanation (RQ7).

Apart from these research questions, we evaluated the quality of the reviewed articles utilizing a set of quality assessment (QA) questions. From the results of answering these questions, we found that the overall quality of the articles is satisfactory, but that there are some problems and weaknesses in the methods used.

From these findings, we have proposed guidelines to try to overcome these problems. Basically, our recommendations and guidelines focus on the fact that the field requires more robust research methods and more explainability metrics, which is evident from our investigation of RQ8 and our questions regarding different quality assessments. The fact that the results of XAI algorithms are measured subjectively showed that these results suffer from biases, as indicated in [72], making an objective comparison between different XAI methods and any measurement of the performance of the methods used impossible. At the same time, this lack of metrics made the algorithm incapable of carrying out an initial justification for the XAI algorithms proposed in [2]. Addadi and Berrada justified the need for an explainable algorithm in the medical field to verify, upgrade, discover, and handle legal and ethical aspects of the black-box models. Without any kind of measure of the performance, all previous objectives are not satisfied. An objective and simple metric that can objectively measure the confidence of the XAI methods are needed to overcome these problems.

Current Challenges and Future Directions

We obtained two main conclusions from the SLR. First, researchers that apply explainable methods do not evaluate them properly as they do not apply any metrics; this fact makes comparison with future advances in this field difficult. Second, results obtained from explainability are not used to improve the classifiers and to check the accomplishment of ethical and legal issues.

The current challenge is that with this lack of assessment and of improvement of the AI models from explainability results, there is no real applicability of the XAI models. As a future direction, researchers must assess the XAI methods to know the level of explanation and must improve the AI models from explainable results, in terms of improving the results and/or to ensure the accomplishment of ethical and legal issues. This paper intends to describe the state of the art of this field proposing a way to unify the future research.

Author Contributions: Conceptualization: M.M.-N., G.M.-A. and A.J.-i.-C.; methodology: M.M.-N., G.M.-A. and A.J.-i.-C.; validation: M.M.-N., G.M.-A. and A.J.-i.-C.; formal analysis: M.M.-N., G.M.-A. and A.J.-i.-C.; investigation: M.M.-N.; resources: G.M.-A. and A.J.-i.-C.; data curation: M.M.-N.; writing—original draft preparation: M.M.-N.; writing—review and editing: M.M.-N., G.M.-A. and A.J.-i.-C.; visualization: M.M.-N.; supervision: G.M.-A. and A.J.-i.-C.; project administration: A.J.-i.-C.; funding acquisition: A.J.-i.-C. All authors have read and agreed to the published version of the manuscript.

Funding: Project PID2019-104829RA-I00 "EXPLainable Artificial INtelligence systems for health and well-beING (EXPLAINING)" funded by MCIN/AEI/10.13039/501100011033. Miquel Miró-Nicolau benefited from the fellowship FPI_035_2020 from Govern de les Illes Balears.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data available at: https://github.com/explainingAI/SLR.

Conflicts of Interest: The authors declare that there are no conflict of interest.

References

1. Gunning, D.; Stefik, M.; Choi, J.; Miller, T.; Stumpf, S.; Yang, G.Z. XAI—Explainable artificial intelligence. *Sci. Robot.* **2019**, *4*, eaay7120. [CrossRef] [PubMed]
2. Adadi, A.; Berrada, M. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access* **2018**, *6*, 52138–52160. [CrossRef]
3. Barredo Arrieta, A.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; Garcia, 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]
4. Burkart, N.; Huber, M.F. A survey on the explainability of supervised machine learning. *J. Artif. Intell. Res.* **2021**, *70*, 245–317. [CrossRef]
5. Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; Torralba, A. Learning deep features for discriminative localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2921–2929.
6. Harris, M.; Qi, A.; Jeagal, L.; Torabi, N.; Menzies, D.; Korobitsyn, A.; Pai, M.; Nathavitharana, R.R.; Ahmad Khan, F. A systematic review of the diagnostic accuracy of artificial intelligence-based computer programs to analyze chest X-rays for pulmonary tuberculosis. *PLoS ONE* **2019**, *14*, e0221339. [CrossRef] [PubMed]
7. Abelaira, M.D.C.; Abelaira, F.C.; Ruano-Ravina, A.; Fernández-Villar, A. Use of conventional chest imaging and artificial intelligence in COVID-19 infection. A review of the literature. *Open Respir. Arch.* **2021**, *3*, 100078. [CrossRef]
8. Kwon, T.; Lee, S.P.; Kim, D.; Jang, J.; Lee, M.; Kang, S.U.; Kim, H.; Oh, K.; On, J.; Kim, Y.J.; et al. Diagnostic performance of artificial intelligence model for pneumonia from chest radiography. *PLoS ONE* **2021**, *16*, e0249399. [CrossRef]
9. Ordookhanians, A.; Li, X.; Nakandala, S.; Kumar, A. Demonstration of Krypton: Optimized CNN inference for occlusion-based deep CNN explanations. *Proc. VLDB Endow.* **2019**, *12*, 1894–1897. [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. Tiulpin, A.; Thevenot, J.; Rahtu, E.; Lehenkari, P.; Saarakkala, S. Automatic knee osteoarthritis diagnosis from plain radiographs: A deep learning-based approach. *Sci. Rep.* **2018**, *8*, 1–10. [CrossRef]
12. Rayan, J.C.; Reddy, N.; Kan, J.H.; Zhang, W.; Annapragada, A. Binomial classification of pediatric elbow fractures using a deep learning multiview approach emulating radiologist decision making. *Radiol. Artif. Intell.* **2019**, *1*, e180015. [CrossRef] [PubMed]
13. Karim, M.R.; Jiao, J.; Döhmen, T.; Cochez, M.; Beyan, O.; Rebholz-Schuhmann, D.; Decker, S. DeepKneeExplainer: Explainable Knee Osteoarthritis Diagnosis From Radiographs and Magnetic Resonance Imaging. *IEEE Access* **2021**, *9*, 39757–39780. [CrossRef]
14. Budgen, D.; Charters, S.; Turner, M.; Brereton, P.; Kitchenham, B.; Linkman, S. Investigating the applicability of the evidence-based paradigm to software engineering. In Proceedings of the 2006 International Workshop on Workshop on Interdisciplinary Software Engineering Research, Shanghai, China, 20 May 2006; pp. 7–14.
15. Kitchenham, B.; Charters, S. Guidelines for performing systematic literature reviews in software engineering version 2.3. *Engineering* **2007**, *45*, 1051.
16. Khosravi, P.; Ghapanchi, A.H. Investigating the effectiveness of technologies applied to assist seniors: A systematic literature review. *Int. J. Med. Inform.* **2016**, *85*, 17–26. [CrossRef]
17. Murdoch, W.J.; Singh, C.; Kumbier, K.; Abbasi-Asl, R.; Yu, B. Interpretable machine learning: Definitions, methods, and applications. *arXiv* **2019**, arXiv:1901.04592.
18. CASP. Critical Appraisal Skills Programme. CASP Qualitative Studies Checklist. 2018. Available online: http://casp-uk.net (accessed on 29 March 2021).
19. Dybå, T.; Dingsøyr, T. Empirical studies of agile software development: A systematic review. *Inf. Softw. Technol.* **2008**, *50*, 833–859. [CrossRef]
20. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
21. Rumelhart, D.E.; Hinton, G.E.; Williams, R.J. Learning representations by back-propagating errors. *Nature* **1986**, *323*, 533–536. [CrossRef]
22. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
23. Bromley, J.; Guyon, I.; LeCun, Y.; Säckinger, E.; Shah, R. Signature verification using a ïsiameseẗime delay neural network. *Adv. Neural Inf. Process. Syst.* **1993**, *6*, 737–744.
24. Dietterich, T.G.; Lathrop, R.H.; Lozano-Pérez, T. Solving the multiple instance problem with axis-parallel rectangles. *Artif. Intell.* **1997**, *89*, 31–71. [CrossRef]
25. Caruana, R. Multitask learning. *Mach. Learn.* **1997**, *28*, 41–75. [CrossRef]
26. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial networks. *arXiv* **2014**, arXiv:1406.2661.
27. Scarselli, F.; Gori, M.; Tsoi, A.C.; Hagenbuchner, M.; Monfardini, G. The graph neural network model. *IEEE Trans. Neural Networks* **2008**, *20*, 61–80. [CrossRef] [PubMed]
28. Kermany, D.; Zhang, K.; Goldbaum, M. Labeled optical coherence tomography (OCT) and Chest X-Ray images for classification. *Mendeley Data* **2018**, *2*. [CrossRef]

29. Candemir, S.; Jaeger, S.; Palaniappan, K.; Musco, J.P.; Singh, R.K.; Xue, Z.; Karargyris, A.; Antani, S.; Thoma, G.; McDonald, C.J. Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration. *IEEE Trans. Med. Imaging* **2013**, *33*, 577–590. [CrossRef] [PubMed]
30. Cohen, J.P.; Morrison, P.; Dao, L.; Roth, K.; Duong, T.Q.; Ghassemi, M. COVID-19 Image Data Collection: Prospective Predictions Are the Future. *arXiv* **2020**, arXiv:2006.11988.
31. Rahman, T.; Khandakar, A.; Qiblawey, Y.; Tahir, A.; Kiranyaz, S.; Kashem, S.B.A.; Islam, M.T.; Al Maadeed, S.; Zughaier, S.M.; Khan, M.S.; et al. Exploring the effect of image enhancement techniques on COVID-19 detection using chest X-ray images. *Comput. Biol. Med.* **2021**, *132*, 104319. [CrossRef]
32. Rosenthal, A.; Gabrielian, A.; Engle, E.; Hurt, D.E.; Alexandru, S.; Crudu, V.; Sergueev, E.; Kirichenko, V.; Lapitskii, V.; Snezhko, E.; et al. The TB Portals: An Open-Access, Web-Based Platform for Global Drug-Resistant-Tuberculosis Data Sharing and Analysis. *J. Clin. Microbiol.* **2017**, *55*, 3267–3282. [CrossRef]
33. SIRM. COVID-19 Database | SIRM. 2020. Available online: https://www.sirm.org/en/category/articles/covid-19-database/ (accessed on 20 April 2021).
34. Wang, L.; Lin, Z.Q.; Wong, A. Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. *Sci. Rep.* **2020**, *10*, 1–12. [CrossRef]
35. Gaillard, F. Radiopaedia. org, the Wiki-Based Collaborative Radiology Resource. 2014. Available online: https://radiopaedia.org/ (accessed on 20 April 2021).
36. De La Iglesia Vayá, M.; Saborit, J.M.; Montell, J.A.; Pertusa, A.; Bustos, A.; Cazorla, M.; Galant, J.; Barber, X.; Orozco-Beltrán, D.; García-García, F.; et al. Bimcv COVID-19+: A large annotated dataset of rx and ct images from covid-19 patients. *arXiv* **2020**, arXiv:2006.01174.
37. NIH. COVID-19—The Cancer Imaging Archive (TCIA) Public Access—Cancer Imaging Archive Wiki. 2020. Available online: https://wiki.cancerimagingarchive.net/display/public/covid-19 (accessed on 20 April 2021).
38. Hospitales, H. Covid Data Save Lives-HM Hospitales. 2021. Available online: https://www.hmhospitales.com/coronavirus/covid-data-save-lives/english-version (accessed on 30 April 2021).
39. Chan, J.H. DLAI3 Hackathon Phase3 COVID-19 CXR Challenge. 2020. Available online: https://www.kaggle.com/c/dlai3-phase3/datasets (accessed on 30 April 2021).
40. Jaeger, S.; Candemir, S.; Antani, S.; Wáng, Y.X.J.; Lu, P.X.; Thoma, G. Two public chest X-ray datasets for computer-aided screening of pulmonary diseases. *Quant. Imaging Med. Surg.* **2014**, *4*, 475. [PubMed]
41. Wang, X.; Peng, Y.; Lu, L.; Lu, Z.; Bagheri, M.; Summers, R.M. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2097–2106.
42. Irvin, J.; Rajpurkar, P.; Ko, M.; Yu, Y.; Ciurea-Ilcus, S.; Chute, C.; Marklund, H.; Haghgoo, B.; Ball, R.; Shpanskaya, K.; et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 29–31 January 2019; Volume 33, pp. 590–597.
43. Demner-Fushman, D.; Kohli, M.D.; Rosenman, M.B.; Shooshan, S.E.; Rodriguez, L.; Antani, S.; Thoma, G.R.; McDonald, C.J. Preparing a collection of radiology examinations for distribution and retrieval. *J. Am. Med. Inform. Assoc.* **2016**, *23*, 304–310. [CrossRef] [PubMed]
44. Johnson, A.E.; Pollard, T.J.; Greenbaum, N.R.; Lungren, M.P.; Deng, C.y.; Peng, Y.; Lu, Z.; Mark, R.G.; Berkowitz, S.J.; Horng, S. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. *arXiv* **2019**, arXiv:1901.07042.
45. SIIM. The Pneumothorax Challenge. 2019. Available online: https://siim.org/page/pneumothorax_challenge (accessed on 20 April 2021).
46. Shiraishi, J.; Katsuragawa, S.; Ikezoe, J.; Matsumoto, T.; Kobayashi, T.; Komatsu, K.I.; Matsui, M.; Fujita, H.; Kodera, Y.; Doi, K. Development of a digital image database for chest radiographs with and without a lung nodule: Receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. *Am. J. Roentgenol.* **2000**, *174*, 71–74. [CrossRef] [PubMed]
47. Rajpurkar, P.; Irvin, J.; Bagul, A.; Ding, D.; Duan, T.; Mehta, H.; Yang, B.; Zhu, K.; Laird, D.; Ball, R.L.; et al. Mura: Large dataset for abnormality detection in musculoskeletal radiographs. *arXiv* **2017**, arXiv:1712.06957.
48. Segal, N.A.; Nevitt, M.C.; Gross, K.D.; Hietpas, J.; Glass, N.A.; Lewis, C.E.; Torner, J.C. The Multicenter Osteoarthritis Study (MOST): Opportunities for rehabilitation research. *PM&R J. Inj. Funct. Rehabil.* **2013**, *5*, 647–654.
49. McGowan, J.A. Perspectives on the future of bone and joint diseases. *J. Rheumatol. Suppl.* **2003**, *67*, 62–64.
50. Varma, M.; Lu, M.; Gardner, R.; Dunnmon, J.; Khandwala, N.; Rajpurkar, P.; Long, J.; Beaulieu, C.; Shpanskaya, K.; Li, F-F.; et al. Automated abnormality detection in lower extremity radiographs using deep learning. *Nat. Mach. Intell.* **2019**, *1*, 578–583. [CrossRef]
51. 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 IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 618–626.
52. Chattopadhay, A.; Sarkar, A.; Howlader, P.; Balasubramanian, V.N. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, USA, 12–15 March 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 839–847.
53. Mnih, V.; Heess, N.; Graves, A.; Kavukcuoglu, K. Recurrent models of visual attention. *arXiv* **2014**, arXiv:1406.6247.

54. Ribeiro, M.T.; Singh, S.; Guestrin, C. " Why should I trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA USA, 13–17 August 2016; pp. 1135–1144.
55. Simonyan, K.; Vedaldi, A.; Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv* **2013**, arXiv:1312.6034.
56. Smilkov, D.; Thorat, N.; Kim, B.; Viégas, F.; Wattenberg, M. Smoothgrad: Removing noise by adding noise. *arXiv* **2017**, arXiv:1706.03825.
57. Rebuffi, S.A.; Fong, R.; Ji, X.; Vedaldi, A. There and back again: Revisiting backpropagation saliency methods. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 8839–8848.
58. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
59. Springenberg, J.T.; Dosovitskiy, A.; Brox, T.; Riedmiller, M. Striving for simplicity: The all convolutional net. *arXiv* **2014**, arXiv:1412.6806.
60. Lin, Z.Q.; Shafiee, M.J.; Bochkarev, S.; Jules, M.S.; Wang, X.Y.; Wong, A. Do explanations reflect decisions? A machine-centric strategy to quantify the performance of explainability algorithms. *arXiv* **2019**, arXiv:1910.07387.
61. Kim, I.; Rajaraman, S.; Antani, S. Visual interpretation of convolutional neural network predictions in classifying medical image modalities. *Diagnostics* **2019**, *9*, 38. [CrossRef] [PubMed]
62. Van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
63. Kadir, T.; Brady, M. Saliency, scale and image description. *Int. J. Comput. Vis.* **2001**, *45*, 83–105. [CrossRef]
64. Jaccard, P. The distribution of the flora in the alpine zone. 1. *New Phytol.* **1912**, *11*, 37–50. [CrossRef]
65. Dice, L.R. Measures of the amount of ecologic association between species. *Ecology* **1945**, *26*, 297–302. [CrossRef]
66. Sorensen, T.A. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons. *Biol. Skar.* **1948**, *5*, 1–34.
67. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. Bleu: A method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics, Philadelphia, PA, USA, 6–12 July 2002; pp. 311–318.
68. Banerjee, S.; Lavie, A. 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, Ann Arbor, MI, USA, 29 June 2005; pp. 65–72.
69. Lin, C.Y. Rouge: A package for automatic evaluation of summaries. In Proceedings of the Text Summarization Branches out, Barcelona, Spain, 25–26 July 2004; pp. 74–81.
70. Vedantam, R.; Lawrence Zitnick, C.; Parikh, D. Cider: Consensus-based image description evaluation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 4566–4575.
71. Zhang, J.; Bargal, S.A.; Lin, Z.; Brandt, J.; Shen, X.; Sclaroff, S. Top-down neural attention by excitation backprop. *Int. J. Comput. Vis.* **2018**, *126*, 1084–1102. [CrossRef]
72. Miller, T. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.* **2019**, *267*, 1–38. [CrossRef]
73. Hoffman, R.R.; Mueller, S.T.; Klein, G.; Litman, J. Metrics for explainable AI: Challenges and prospects. *arXiv* **2018**, arXiv:1812.04608.
74. Alam, L.; Mueller, S. Examining the effect of explanation on satisfaction and trust in AI diagnostic systems. *BMC Med. Inform. Decis. Mak.* **2021**, *21*, 1–15. [CrossRef] [PubMed]
75. Voigt, P.; Von dem Bussche, A. The eu general data protection regulation (gdpr). In *A Practical Guide*, 1st ed.; Springer: Cham, Switzerland, 2017; Volume 10, p. 3152676.