Systematic Review

Tertiary Review on Explainable Artificial Intelligence: Where Do We Stand?

Frank van Mourik [1], Annemarie Jutte [1,2], Stijn E. Berendse [1], Faiza A. Bukhsh [1] and Faizan Ahmed [1,2,*]

1   Department of Computer Science, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands; a.m.p.jutte@utwente.nl (A.J.); f.a.bukhsh@utwente.nl (F.A.B.)
2   Ambient Intelligence Research Group, Saxion University of Applied Sciences, 7513 AB Enschede, The Netherlands
*   Correspondence: faizan.ahmed@utwente.nl

Abstract: Research into explainable artificial intelligence (XAI) methods has exploded over the past five years. It is essential to synthesize and categorize this research and, for this purpose, multiple systematic reviews on XAI mapped out the landscape of the existing methods. To understand how these methods have developed and been applied and what evidence has been accumulated through model training and analysis, we carried out a tertiary literature review that takes as input systematic literature reviews published between 1992 and 2023. We evaluated 40 systematic literature review papers and presented binary tabular overviews of researched XAI methods and their respective characteristics, such as the scope, scale, input data, explanation data, and machine learning models researched. We identified seven distinct characteristics and organized them into twelve specific categories, culminating in the creation of comprehensive research grids. Within these research grids, we systematically documented the presence or absence of research mentions for each pairing of characteristic and category. We identified 14 combinations that are open to research. Our findings reveal a significant gap, particularly in categories like the cross-section of feature graphs and numerical data, which appear to be notably absent or insufficiently addressed in the existing body of research and thus represent a future research road map.

Keywords: explainable AI; computer vision; time series; neural network; deep learning

1. Introduction

Explainable artificial intelligence (XAI) is a set of methods used to tackle the interpretability problem [1] by providing users explanations on how a model came to its conclusion. By providing these additional insights into its reasoning or internal workings, the model's transparency is increased, resulting in higher trust from the user.

In the context of deep learning, the model learns from the data, and the internal learning of the model is, generally, a black box. There is a constant call to make these black-box models more interpretable. It is especially crucial to understand how the model made certain decisions for critical systems. It is also essential for systems that are not critical, but where their black-box nature induces biases or other ethical dilemmas.

One example of the usage of XAI in daily life is reasons why a mortgage is approved or denied by a bank. This benefits two parties: the applicant for the mortgage and the bank. The applicant benefits because accepting a possible denial may be easier if a reason is provided. At the same time, the bank benefits from increased insight into how their model behaves and can therefore avoid biases or other ethical issues induced by the model [2].

Another example concerns healthcare [3], where a heavy debate is ongoing on AI implementation as "explainability is not a purely technological issue, instead it invokes a host of medical, legal, ethical, and societal questions that require thorough exploration" [3]. Another example is the bias in diagnosis based on X-ray data. The diagnosis is biased toward under-served-populations (see Seyyed-Kalantari et al. [4]).

Embracing the need for XAI, in April 2021 the European Commission proposed a regulation to ensure that complex "high-risk AI systems shall be designed and developed in such a way to ensure that their operation is sufficiently transparent to enable users to interpret the system's output and use it appropriately" [5] [Article 13]. Half a year later, in September 2021, Brazil followed by approving a similar proposal.

Although the terms "explainable AI" and "interpretable AI" are often confused as synonyms, they have subtle differences. Interpretable AI refers to the characteristic of an AI system that allows humans to understand the processes it uses to make decisions or predictions. In other words, "we will consider interpretability as the possibility of understanding the mechanics of a Machine Learning model, but not necessarily knowing why [6]". On the other hand, XAI focuses on the overall understanding of the data, including the unobserved cases. This also includes certain feature values not present in the data or some data points that have not occurred. In this paper, we do not use both terms in their strict meaning; instead, we focus on the larger class of XAI while keeping in mind that the cited paper may have used these terms interchangeably.

Numerous XAI methods have been developed over the past 5 to 10 years [7], ranging from model-agnostic methods (those that can be applied to all existing and future models) to model-specific intrinsic methods (those that can only be applied to a single or subset of methods). New methods and techniques are actively researched and the number of scientific papers describing these techniques is rapidly increasing.

The current rapid publication rate calls for higher-tier research that summarizes and synthesizes the current state of the art and highlights existing gaps. The most commonly used method is conducting systematic literature reviews that summarize and synthesize the current state of research on a topic [8].

Systematic literature reviews (SLRs) aim to summarize multiple individual research articles, whereas meta-reviews are used to summarize both individual research articles as well as existing SLRs. The top-tier evidence synthesis method is called a *tertiary review*, which is a systematic review of systematic literature reviews [9].

In this article, we will conduct a tertiary review of existing XAI methods and their characteristics, such as input data type or model type. Furthermore, we aim to categorize these methods into well-defined boxes to create a clear overview of the existing literature. Example categories could be heat maps, graphs, or decision trees.

Therefore, the goal of this paper is to provide a mapping of XAI categories with their characteristics and fill this mapping with existing XAI methods. This mapping will result in 2D matrices, with each cell indicating whether an XAI method in that category has or has not been researched with that characteristic.

We want to emphasize that this paper is aimed at meta-level reviews where the focus is on summarizing existing literature reviews and finding open research directions. There are multiple literature reviews (and most of them are cited in the coming section) written discussing an exciting research topic within the realm of explainable artificial intelligence (as it is detailed in this paper). However, to the best of our knowledge, this paper is the first attempt to combine knowledge from existing literature reviews into a single article. During the final editing of this paper, we came across the recent paper work of Saeed and Omlin [10] where they provided a scoping review of the field of XAI. Their review presents challenges and research directions in XAI as discussed in the research papers. Their approach is different from ours. Instead of focusing on the discussion points mentioned in primary review papers, we catalog which research directions have and have not been explored and extract new research gaps from this. Furthermore, this paper specifically focuses on XAI techniques, whereas Saeed and Omlin [10] discuss challenges in XAI in a broader sense.

To clarify the terminology used in this research, the terms used throughout this paper are defined in Section 2. In Section 3, the research method used for this tertiary review is explained. In Section 4, we synthesize the findings of the review. We then describe the

limitations of our research guidelines for future work in Section 5. Finally, we conclude our findings in Section 6.

2. Terminology

As the XAI research area is still young, a consensus has not been reached yet for every term that is used. Therefore, this section serves as a glossary that we will use for the remainder of this article.

1. ML: abbreviation of machine learning.
2. XAI: abbreviation of explainable artificial intelligence, also used as "Explainable AI", "Explainable Machine Learning", and "Interpretable AI".
3. Articles: any peer-reviewed scientific paper or article.
4. ML techniques: ML techniques are the areas in which machine learning models are applied, such as natural language processing, speech processing, computer vision, and predictive analysis.
5. ML domains: ML domains are application domains wherein ML models are being applied, like healthcare, retail, financial/banking, and industry.
6. Explainable AI methods: methods developed to increase the explainability and/or interpretability of either the outcomes of a model or the model itself.
7. Systematic Literature Review (abbreviation: SLR): an SLR identifies, selects, and critically appraises research in order to answer a clearly formulated question (https://libguides.csu.edu.au/review/Systematic accessed on 27 August 2024).
8. Tertiary Review: a systematic approach to summarizing information from existing SLRs and meta-reviews.

3. Review Methodology

This section describes the search strategy, search string, and inclusion and exclusion criteria. A complete overview of their application, from research question formation to collected studies, is provided to ensure a rigorous and reproduceable approach for this SLR.

3.1. Research Questions and High-Level Synthesis

Over the last five years, numerous systematic literature reviews on XAI have been conducted (as we shall see in the coming section).

We aim to present a high-level synthesis based on these literature reviews focused on answering the following questions:

- RQ 1: What are distinguishing characteristics of XAI methods?
- RQ 2: What are the distinct categories that can be established to classify various XAI methods based on their shared characteristics and methodologies?
- RQ 3: Which combinations of XAI categories and characteristics have been researched?

The tertiary review (see [9]) method is considered above a systematic literature review in the hierarchy of evidence synthesis methods, as seen in Figure 1. Since there are no well-defined guidelines for conducting tertiary reviews, we adapted the guidelines of Kitchenham et al. [8], where only existing systematic literature reviews and meta-analyses are taken into account, instead of individual articles.
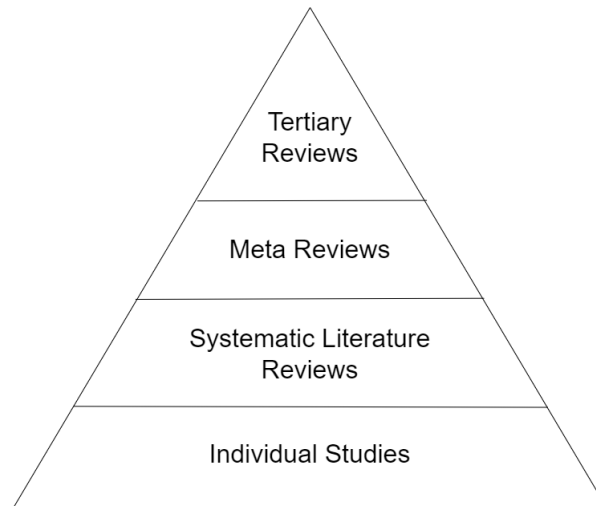
Figure 1. Hierarchy of evidence synthesis methods, based on Fusar-Poli and Radua [9].

3.2. Literature Retrieval and Selection

To retrieve literature, we need to craft a query. To craft the search query, we adapted a fine-tuning-based approach, where we started with a query and fine-tuned it by looking at the results returned by the database Scopus [11]. Scopus was chosen since it contains articles from a broader area of science and engineering and due to the ease of retrieving the returned results.

Based on the research question, we started with the combination of "Explainable AI" and "Systematic Literature Review", together with their respective synonyms, as the initial search query. After a couple of iterations, it resulted in a more comprehensive search query, as given in Listing 1.

Listing 1. Search query.

```
    (( systematic   AND   review ) OR (meta AND analysis AND review ))
AND
    (( explainable AND ai )OR( explainable AND artificial AND intelligence )
    OR ( xai )OR( interpretable AND ai )
    OR( explainable AND machine AND learning ))
```

As can be seen in Listing 1, the synonyms used for "explainable AI" are "explainable artificial intelligence", "xai", "interpretable AI", and "explainable machine learning". The synonyms used for "systematic literature review" are "systematic review" (which makes the search term "systematic literature review" redundant), and "meta-analysis review". We search the broader area of science and engineering instead of confining ourselves to computer science. This choice is made because relevant articles could be focused on a specific domain and would then generally be published in sources associated with the application domain.

To reduce the bias of database choice, this query was run in multiple databases resulting in varying results. Executing this query string on Scopus resulted in 92 articles. After applying the same query to IEEE Xplorer (12 results), ACM (four results), ScienceDirect (80 results), and Springer (41 results), 229 articles were retrieved. The search query is adapted to match the specific needs of each database.

Furthermore, one article was added manually, as this article is not yet published but does fit the search query, resulting in 230 articles, as seen in Figure 2.
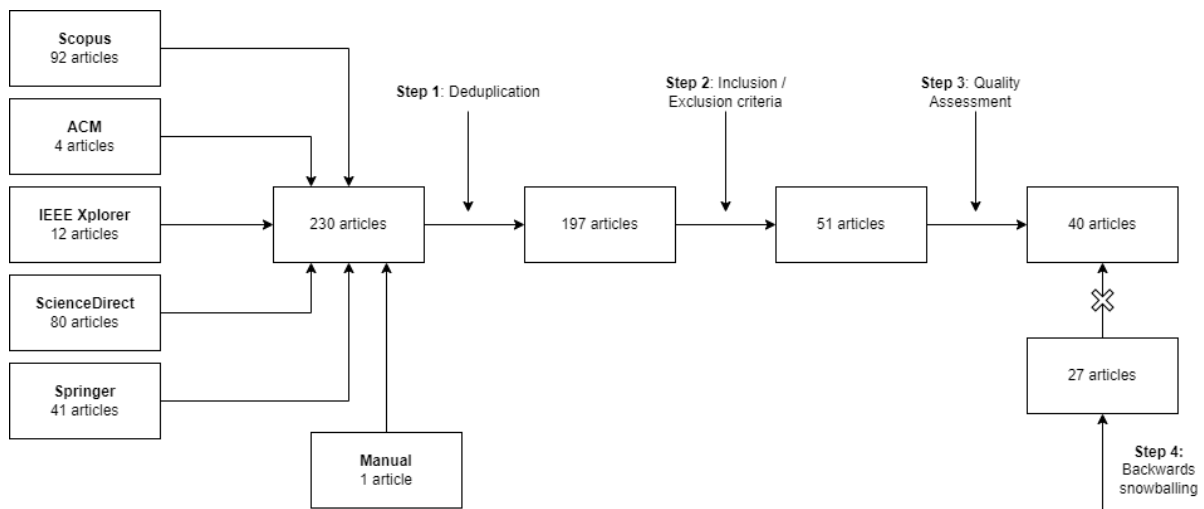
Figure 2. PRISMA flow diagram of the tertiary review.

3.3. Literature Selection

The proceeding subsection describes the steps followed based on the guidelines outlined in [8] to select only the literature relevant to our research questions.

3.3.1. Step 1: Duplicate Removal

Some of the resulting articles appear in multiple databases. Hence, the duplicates were removed. In total, 33 duplicates were removed, resulting in 197 unique articles.

3.3.2. Step 2: Inclusion Criteria

To find the most relevant articles, the following inclusion and exclusion criteria are defined:

- **IC1:** include only articles performing a systematic literature review or meta-analysis on XAI methods.
  Motivation: this is the aim of our tertiary review.
- **IC2:** include only articles fully available through the queried databases or manually added, using the University of Twente academic access.
  Motivation: fully available papers are needed to summarize the papers properly.
- **IC3:** include only articles written in English.
  Motivation: as the research is performed in English, only English-written articles are taken into account.
- **IC4:** include only scientific peer-reviewed papers.
  Motivation: Scientific papers that went through a peer-review process ensure a level of credibility and quality.

The initial selection step included abstract and title screening. The first author read the remaining 197 articles after duplicate removal, and the fourth and fifth authors critically evaluated this process. After the title and abstract screening, 51 articles were selected. IC1 provided the foundation for the exclusion of 146 articles. One major discovery of the title and abstract screening is the significant number of articles that contain the relevant keywords. However, upon closer examination, these articles did not substantively address the concepts in question or contribute novel methodologies to the field.

3.3.3. Step 3: Quality Assessment

To further strengthen the screening process, we defined quality criteria. It is possible that an article may inadvertently be included after the initial title and abstract screening, even if it is not completely relevant to our research focus. This could occur when the article employs certain terminology from our search query in a manner that is primarily attention-grabbing, thereby giving an impression of relevance that may not be entirely accurate.

Additionally, there are cases wherein an article discusses the topic of explainable artificial intelligence (XAI) but does not engage in a detailed exploration of specific XAI methodologies. We adopted these two considerations as key aspects of our quality assessment criteria. From the 51 articles, 11 were excluded via the quality assessment. An overview of the excluded papers with their exclusion reasoning can be found in Appendix A.

### 3.3.4. Step 4: Backwards Snowballing

To mitigate the risk of missing important articles, we used the backwards snowballing technique (see [12]). This technique allowed us to add articles found in the reference list of included articles when they matched our inclusion criteria. In total, we identified 27 additional articles by performing this technique. Due to time constraints, these articles were not explicitly read in full but instead were added in Appendix B. However, these papers are already included at a meta level since they are referenced by the papers that were already included in this review. Therefore, we have not used these papers explicitly in our results.

4. Results

In this section, the results gathered from the 40 included articles are presented. We start with general insights into these articles in Section 4.1. In the proceeding section, the selected articles are used to answer the research questions defined in Section 3. For RQ 1, the identified characteristics can be found in Section 4.2. For RQ 2, the identified categories are presented in Section 4.3. Finally, we give an overview of which combinations of characteristics and categories have been researched, according to the included articles, in the "XAI research matrix". This matrix serves to answer RQ 3 and can be found in Section 4.4.

4.1. Article Insights

As mentioned in Section 3.1, we expected the number of systematic literature reviews on XAI to increase in recent years. As shown in Figure 3, which illustrates the number of included articles per publication year, this hypothesis is correct. A rapid increase in SLRs on XAI can be observed. The value for 2023 is not representative, since the papers included are either published before 20 December 2022, or are scheduled to be published later in 2022 or early 2023; therefore, it is grayed out.
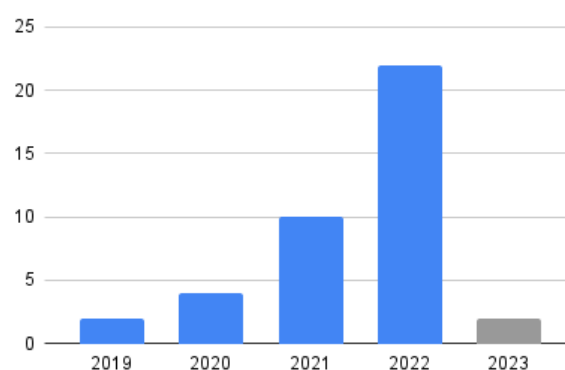
Figure 3. Overview of included articles per publication year.

4.2. XAI Methods Characteristics

In this section, we present the characteristics of XAI methods found in the selected articles. Out of the 41 included papers, we identified five papers [5,13–16] which are found to be most relevant to categorize XAI methods based on their characteristics.

Vilone and Longo [13] introduced five dimensions to characterize XAI methods that we adopt: explainability method (they refer to this as "stage"), scope, problem type, input data type, and explanation data type (they refer to this as "output format"). Fur-

thermore, whereas they split the post hoc explainability method into model-specific and model-agnostic methods, we incorporate this division into the separate model dependency characteristic.

European Commission [5], Saleem et al. [15] and Groen et al. [14] have presented comparable characteristics. However, they used more compact, less granular, sets of characteristics. For our purposes, including more characteristics will increase the granularity of the categorization, which allows us to specifically identify research gaps. Therefore, we choose to follow Vilone and Longo [13].

Finally, Guidotti et al. [16] characterized explainability methods by the model type. We added this characteristic to the list by Vilone and Longo [13].

In the following subsection, each category is briefly described.

4.2.1. Explainability Method

There are two primary means by which a method can increase the explainability of a machine learning model, namely intrinsic and post hoc explainability. Figure 4 (adapted from [17]) illustrates the difference between intrinsic and post hoc explainability and they are further described below.
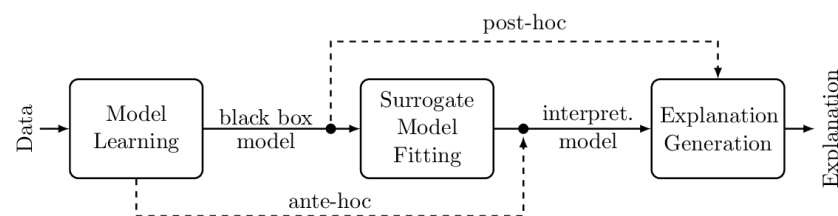
Figure 4. Visual overview of intrinsic (ante hoc) and post hoc explainability [17].

Intrinsic explainable models are designed to be naturally interpretable, meaning that their decision-making process is transparent and understandable from the outset. This intrinsic transparency is a key feature, as it allows users and developers to understand how and why specific decisions are made by the model.

In contrast, post hoc explainability methods are applied after the model has been trained. These methods attempt to shed light on the decision-making process of complex, often black-box models like deep neural networks. While post hoc methods can be quite effective in providing insights, they have limitations. One of the primary challenges is that they do not alter the underlying model to make it intrinsically more interpretable; instead, they try to interpret the model's outputs or internal state after the fact.

4.2.2. Model Dependency

XAI methods that provide intrinsic explainability are always model-specific, as explainability is integrated into the model itself rather than being applied separately. In contrast, post hoc explainability methods can potentially be applied to all existing machine learning models, making them model-agnostic. Since only post hoc methods can be model-agnostic, we consider the *model dependency* characteristic as an elaboration on post hoc explainability. *Model-specific* methods are tied to a specific set of machine learning models, whereas *model-agnostic* methods can be applied to virtually any machine learning model.

4.2.3. Scope

The scope of an explanation refers to what part of the inferential process of a model is explained by an XAI method. There are two possible scopes for an explanation, namely local and global. In *local explanation*, the method explains how the model arrives at the prediction for a single data input, stating nothing about the overall (inner) working of the model. On the other hand, global explanation methods provide insights into the inner workings of the model, for instance by identifying patterns in the data and potential bias in the model. Global methods provide an overview of the overall working of the machine learning model that may not provide useful information for a single instance.

4.2.4. Problem Type

Each machine learning model has a task it should perform or a problem it should solve. A wide variety of tasks exist, but in this review, we will only consider the two tasks that are by far the most popular [18], namely classification and regression.

In classification, one is focused on categorizing data points into predefined categories such as "yes/no", "cat/dog", or "high, medium, low". Note that segmentation, a common computer vision problem, can be considered as an extension of classification problem, since in segmentation we assign a category at the pixel level.

Classification has the task of predicting discrete variables and in contrast regression predicts continuous variables such as room temperature, mortality probability, or stock price.

4.2.5. Input Data Type

For each XAI method, we determine what type(s) of input data it can handle. We identify the following input data types:

- *Visual data:* this mostly includes images and videos.
- *Textual data:* this includes natural language and is unstructured.
- *Tabular/structured data:* this is the data that can be stored in spreadsheets or databases. Usually, the data type of each column is the same, mostly categorical or numerical data.
- *Time-series data:* represent an event over time. This data type records time as one dimension while the second dimension is the numerical or categorical representation of an event. The representative examples of this data type are sales data or weather forecasts.
- *Data-agnostic:* this includes models that can handle any type of data as their input.

It is important to note that we categorize videos as visual data rather than time-series data. This classification is based on the similarity between machine learning models designed for video data and those used for processing visual data. Therefore, our approach aligns videos with visual data models to leverage their inherent processing capabilities.

4.2.6. Explanation Data Type

Input data type determines the input that a certain XAI technique can handle. On the other hand, the produced explanation can be represented using different data types. We categorize explanation data types into the most commonly (see e.g., [19]) utilized types for clarity and specificity.

- *Visual data:* In this instance, explanation is returned as a visual object, for instance, a heatmap or saliency map.
- *Textual data:* Explanation is provided in the form of text generation via natural language processing.
- *Numerical data:* This is one of the common time outputs. The returned number can represent different things depending on the context and application. For example, computed importance score for a (sub)set of features is returned or it could be probabilities of outcomes, or influence percentages per input feature.
- *Rule-based data:* The explanation is returned in the form of decision rule(s) that explain *why* a certain prediction was made, possibly combined with a set of counterfactual rules which would have led to another prediction.

4.2.7. Model Type

ML models can be categorised based on their internal working mechanism. The categorization varies depending on the use case. We base the ML model types on Guidotti et al. [16]. The model types we identified are the following:

- Deep neural networks are machine learning models consisting of a network of nodes (the *neurons*) that compute a prediction, using their weighted average. There are many different types of neural networks such as convolutional neural network and recurrent neural network.

- Tree ensemble models work by combining (ensembling) multiple tree-based models by bagging and/or boosting, such as XGBoost or Random Forest models.
- Support vector machines work by finding an optimal hyperplane in high-dimensional space that separates the data.

Note that the above list is not exhaustive; however, it covers the most used machine learning models.

4.3. Explanator-Based XAI Categories

XAI methods provide output in a certain form, referred to as an explanator (see [16]). One can classify XAI methods on the basis of the explanator used. Such a classification first appeared in the works of Guidotti et al. [16] and was later extended and updated by Nauta et al. [19]. The classification presented below is an adapted version of Nauta et al. [19]. We do not consider localization and white-box as classes, since white-box is used as a synonym for intrinsic methods, which are already covered in the earlier section. We included conterfactuals as categories since they have gained more attention recently and deserve to be a separate class. The classification is given below.

- *Graphs:* It is an explanation in the form of a graph, consisting of nodes and edges, like Knowledge Graphs [20].
- *Feature importance:* It is a list of numbers showing the relevance and importance of each feature of the input data. Two of the most popular XAI methods (LIME (For documentation visit https://lime-ml.readthedocs.io/en/latest/ accessed on 27 August 2024) [21] and SHAP (Detailed documentation can be found here: https://shap.readthedocs.io/en/latest/contributing.html accessed on 27 August 2024) [22]) are part of this category.
- *Heatmaps:* It is an image indicating the relevant parts of the input data, for example, indicating the part of an image that contributed the most to the classification of the image. There are several variants [23] available which are now part of standard machine learning libraries (there are several implementations available; the implementations documented here cover a number of variants: https://github.com/jacobgil/pytorch-grad-cam accessed on 27 July 2024).
- *Rule-based explanators:* It is a rule, or set of rules, that determine why a prediction was made. The paper [24] describes a rule-based XAI package.
- *Decision Trees:* It is a rooted set of rules with a decision at every node in the tree. The earliest work in this regard is the paper [25].
- *Disentanglement:* It is a breakdown of the high-dimensional input features into lower-dimensional features, which may be directly interpretable (see, e.g., [26]).
- *Feature plot:* It is a plot that illustrates the correlation of multiple input features. There are two notable algorithms, namely partial dependence plot and conditional expectation plots (see, e.g., [27]). Both of them are part of scikit-learn (https://scikit-learn.org/stable/modules/partial_dependence.html accessed on 27 July 2024).
- *Counterfactuals:* It is an artificially generated example data point, similar to the input data point, where the predictive model gives a different prediction (see, e.g., [28]).
- *Prototypes:* It is a set of data examples that shows what the predictive model considers to be similar to the input, for example, its nearest neighbors. One well-known technique is MMD-critics [29]. The implementation can be found on the authors' github.
- *Text generation:* It is an explanation in the form of natural language, for example, using generated text (see, for example, [30]).
- *Representation Synthesis:* It is an artificial visualization produced as an explanation of the predictive model, like synthetic data samples. Feature visualization [31] is one such example.
- *Representation Visualizations:* It is a visualization produced to explain the predictive model, for example, principle component analysis.

4.4. XAI Knowledge Grid

In this section, we integrate the XAI characteristics and categories outlined in Sections 4.2 and 4.3. Our objective is to map and visualize the researched combinations in the existing literature and identify areas where research is lacking. Note that lacking research does not necessarily imply that these combinations have never been researched. It may be the case that the combination is researched but it has not appeared in the selected SLRs. Therefore, these recommendation should be taken with caution. Moreover, XAI is an active area of research, and since this paper is a meta study based on published SLRs, it may be the case that there is already published research since the inception of this paper or the selected SLRs (see also Section 3). Additionally, we will address certain combinations deemed currently unfeasible to research, as defined and discussed in Section 4.4.6).

The combinations of characteristics and categories are visualized as a matrix. Each cell in each matrix is filled with a referenced paper (if that paper mentions the combination as having been researched), an "*O*" (open for research), or an "*X*" (currently considered unfeasible), which provides an overview of the current research opportunities in this field.

In the remainder of this section, we will first explain which existing papers contributed significantly to our review. Afterwards, in Sections 4.4.2–4.4.5, we will display and explain the result matrices of this tertiary review. Finally, in Section 4.4.6, we will explain the "unfeasible" cells in more detail.

4.4.1. Key Papers

In 2018, Guidotti et al. [16] were the first to systematically review existing methods "for explaining black box models". They created a classification of the main methods concerning the explanation and type of black-box model used. The goal of the paper was to offer researchers tools to find methods most useful for their work. As the research of Guidotti et al. [16] can be seen as a "predecessor" of our paper, many XAI methods from their reviewed primary papers were used to fill the matrices of this paper.

In 2021, Vilone and Longo [13] extended the classification created by Guidotti et al. [16] by adding the "output formats" dimension, resulting in a practical tool for scholars to select the most suitable type of explanation format for their problem. This extension on Guidotti's classification provided great added value to our tertiary review while filling the matrix.

In 2022, European Commission [5] conducted a systematic literature review on recently developed XAI methods, together with their domains, tasks, and models. They found that, in recent years, XAI methods were mostly developed for deep learning and ensemble models. Additionally, they reviewed the problem type of XAI methods, like classification and/or regression, in combination with the explanation data type. Especially on these last two characteristics, the research of European Commission [5] significantly contributed to our matrices.

Lastly, in 2023, Nauta et al. [19] conducted a systematic review on the evaluation of XAI methods. Even though evaluation is not the aim of our review, Nauta et al. [19] also provided an application (visit https://utwente-dmb.github.io/xai-papers/ accessed on 27 July 2024) where papers that introduce XAI methods are indexed and labeled according to characteristics and categories. As Nauta et al. [19] based their classification on Guidotti et al. [16] and expanded this initial overview, their application contributed to filling the last open cells of the matrices.

4.4.2. Explainability Method, Model Dependency, and Scope

The first research grid or matrix we present is concerned with the explanability method, model dependency, and scope. The matrix is presented in Table 1. It is essential to point out that post hoc scope is further classified as "model specific" and "model agnostic". It is evident from the Table 1 that there is a substantial interest for most combinations of the explainability method, model dependency, and scope. Ante hoc counterfactuals have not been researched; however, this combination is currently deemed unfeasible, as

counterfactuals are by definition generated after the model has computed its prediction. For the scope characteristic, all combinations have been explored to a certain degree.

We would like to emphasize that Table 1 should not be misinterpreted as indicating a scarcity of future research prospects in this area. Its binary representation, signifying whether specific combinations have been researched, merely serves as a foundation for future exploration. References are included as starting points for further inquiry. Moreover, there is a significant need for future research to not only innovate but also refine existing XAI techniques, ensuring continuous advancement in the field.

Table 1. Researched matrix on explainability method, model dependency (model-specific (MS) and MA (model-agnostic), and scope. "*O*" (open for research), or an "X" (currently considered unfeasible).

XAI Types Categories	Expl. Method and Model Dep.			Scope	
	Ante Hoc	Post Hoc MS	MA	Local	Global
Graphs	[32]	[32]	[33]	[34]	[5]
Feature importance	[35]	[36]	[37]	[35]	[35]
Heatmaps	[36]	[35]	[19]	[35]	[38]
Rule-based explanators	[36]	[36]	[16]	[36]	[36]
Decision Trees	[36]	[36]	[16]	[36]	[5]
Disentanglement	[5]	[5]	[19]	[5]	[5]
Feature plots	[5]	[5]	[16]	[5]	[38]
Counterfactuals	X	[39]	[5]	[28]	[28]
Prototypes	[33]	[19]	[19]	[5]	[5]
Text generation	[19]	[13]	[19]	[13]	[13]
Representation Synthesis	[19]	[33]	[19]	[5]	[5]
Representation Visualizations	[19]	[19]	[19]	[5]	[5]

4.4.3. Input Data Type

In this subsection, we discuss the categorisation of XAI on the basis of the input data type. As can be seen in Table 2, for the input data type characteristic, most combinations have been researched. We have identified three possible research directions and two unfeasible research directions.

Table 2. Researched matrix on input data type, describing visual, textual, numeric (Num), categorical (Cat), time-series (TS), and data-agnostic (DA) methods. "*O*" (open for research), or an "X" (currently considered unfeasible).

XAI Types Categories	Input Data Type				
	Visual	Textual	Num/Cat	TS	DA
Graphs	[32]	[32]	[32]	*O*	[19]
Feature importance	[35]	[35]	[16]	[5]	[16]
Heatmaps	[35]	[19]	X	[19]	X
Rule-based explanators	[13]	[13]	[16]	[13]	[16]
Decision Trees	[19]	[19]	[16]	[5]	[16]
Disentanglement	[19]	[19]	[19]	[19]	[19]
Feature plots	[19]	[19]	[16]	[19]	*O*
Counterfactuals	[28]	[28]	[40]	[5]	[28]
Prototypes	[16]	[19]	[19]	[19]	[16]
Text generation	[13]	[13]	[13]	[13]	*O*
Representation Synthesis	[19]	*O*	[19]	[19]	[19]
Representation Visualizations	[19]	[19]	[19]	[19]	[19]

Disregarding the few combinations currently deemed to be unfeasible, four combinations are open for future research. Combinations that are deemed unfeasible to research are detailed further in Section 4.4.6.

It is evident from Table 2 that visual and numerical data have gained much attention, reflecting their wide spread use. Sequential data can be seen as a variant of these two data types in the form of video and time series data. Although there exists some research for sequential data, it is limited and requires more focus and attention from researchers.

Especially, time series data, which require methods unique to their nature, present an open area for research.

Data-agnostic methods are another interesting open research direction. Although data-agnostic methods exists for most input types, these techniques require further refinement. Also, there are clear research gaps to develop data-agnostic type methods for feature plots and text generation. The final open direction in this area is representation synthesis-based explanation methods for textual data.

4.4.4. Explanation Data Type

The explanation data type can often be deduced from the XAI categories. Consequently, it is understandable that some research areas are currently seen as unfeasible, as can be seen from Table 3. Certain combinations require innovative breakthroughs to become viable, whereas others face inherent conflicts; for example, numerical/categorical explanations are unsuitable for heatmaps, as heatmaps, by definition, represent visual data types.

While many combinations are currently infeasible, some avenues remain open for research, particularly textual explanations for XAI types like prototypes, representation synthesis, and visualizations (see Table 3). We emphasize again that just because a combination has been researched does not mean the exploration is complete. Often, the progress in these areas is just beginning, leaving many questions open and necessitating further investigation. Similarly, the existence of research must be taken with caution, since there could be ongoing research that has not appeared in the used SLR and thus in this paper.

Table 3. Researched matrix on explanation data type, describing visual, textual, numeric (Num), categorical (Cat), and rule-based explanations. "*O*" (open for research), or an "X" (currently considered unfeasible).

XAI Types Categories	Explanation Data Type			
	Visual	**Textual**	**Num/Cat**	**Rules**
Graphs	[32]	[32]	[32]	[5]
Feature importance	[35]	[35]	[5]	[5]
Heatmaps	[35]	X	X	X
Rule-based explanators	[5]	[5]	[5]	[5]
Decision Trees	[5]	X	X	[5]
Disentanglement	[5]	[5]	[5]	[5]
Feature plots	[5]	X	X	X
Counterfactuals	[28]	[28]	[40]	[5]
Prototypes	[5]	*O*	[5]	[5]
Text generation	X	[13]	X	X
Representation Synthesis	[5]	*O*	[5]	[5]
Representation Visualizations	[5]	*O*	[5]	[5]

4.4.5. Problem Type and Model Type

The final interesting set of combinations deals with the problem type and model types. The results are provided in Table 4. Since most ML problems are broadly categorized as either classification or regression problem, it is not surprising that three is existing research for both categories. It would be interesting to focus on sub-tasks such as segmentation or detection to enhance the problem type categories. With a finer granularity of problem type, we can discover more open questions.

Similarly, the notion of black-box models gained more attention after the dawn of deep learning models and lead to surge in XAI research (see, e.g., [41]). Naturally, there is existing research for each of XAI type categories for (D)NN. However, it is an active area of research and there are many directions that are open to research.

With the dawn of deep learning, the use of traditional ML techniques is declining. However, for use cases wherein a limited data or computational resources are available, these techniques are relevant. For tree-ensembles and SVM, there are multiple combinations that are open to further research. Especially, the research can be interesting in the context of SVM, which is a popular method of choice for ML with limited data (see, e.g., [42]). It

is interesting that the combination of tree-ensembles and heatmaps is not explored yet, although both are used widely in ML applications. During the review process of this paper, we came across recent research that explored this combination (see [43]). Once more, the existence of research does not imply the research area is closed. Instead, it indicates the active nature of research in XAI. Finally, the model-agnostic representation synthesis is an interesting open direction.

Table 4. Researched matrix on problem type and model type. Problem types include classification (Class.) and regression (Regr.). Model types include (deep) neural networks ((D)NN), tree-ensembles (TE), support vector machine (SVM), and model-agnostic (MA) models. "*O*" (open for research), or an "*X*" (currently considered unfeasible).

	Problem Type		Model Type			
XAI Types Categories	**Class.**	**Regr.**	**(D)NN**	**TE**	**SVM**	**MA**
Graphs	[5]	[32]	[32]	[44]	[19]	[33]
Feature importance	[35]	[5]	[37]	[37]	[37]	[37]
Heatmaps	[38]	[35]	[35]	*O*	[19]	[19]
Rule-based explanators	[36]	[5]	[16]	[16]	[16]	[16]
Decision Trees	[5]	[5]	[16]	[16]	[38]	[16]
Disentanglement	[5]	[19]	[5]	[19]	[19]	[19]
Feature plots	[38]	[5]	[19]	*O*	*O*	[16]
Counterfactuals	[28]	[28]	[39]	[39]	[5]	[5]
Prototypes	[5]	[33]	[45]	[19]	[19]	[19]
Text generation	[13]	[13]	[13]	[13]	*O*	[19]
Representation Synthesis	[5]	[5]	[33]	[19]	*O*	*O*
Representation Visualizations	[5]	[19]	[19]	*O*	*O*	[19]

4.4.6. Unfeasible Cells

In the above tables, some cells are labeled as "currently unfeasible to research". We will briefly explain why they are deemed impossible. Firstly, counterfactuals are by definition generated after the model has made its prediction; therefore, the explanation method "intrinsic explainability" is not an option.

Next, heatmaps and feature plots are by definition visual explanations. Therefore, all cells with explanation data type "not visual" are deemed unfeasible. Furthermore, by our definition, heatmaps cannot be used on tabular or data-agnostic input data.

There are some more combinations with the explanation data type which are deemed unfeasible. Firstly, decision trees are by definition either visual or state the node rules of the decision tree; therefore, textual and numerical or categorical explanations are deemed impossible. Secondly, for methods in the "text generation" category, the explanation data type needs to be "textual" by definition; hence, other combinations are deemed impossible.

5. Discussion

This section will discuss possible threats to the validity of this tertiary review. Furthermore, we provide some recommendations for future tertiary reviews based on limitations that we encountered.

5.1. Future Research Directions

This review aims to provide a comprehensive overview of the current state of research in explainable artificial intelligence (XAI), identifying key areas ripe for further investigation. Note that we have followed the matrix-based methodological approach where one can assume that "If an XAI category on the vertical axis has not been matched with a set of variables on the horizontal axis, it indicates a potential research gap to explore—unless it is technically infeasible", which may lead to the limited conclusion. Here, we would like to highlight that the proposed combinations are not exhaustive, but they are a starting point for an in-depth exploration direction.

The gaps outlined in the tables of Section 4.4 guide our proposed research directions. From the research grid presented in Section 4.4, one can easily identify 14 open research directions and an equal number of combinations that are unfeasible for research.

We have used 12 XAI-type categorizations that span the current XAI research landscape. Future research should explore whether the given XAI type categorization is complete. Since XAI is rapidly evolving, one can expect that the list of XAI type categories will enlarge in the future. One noticeable opportunity for progress is developing visualization methods that are accessible to non-experts. This need is echoed in the works of [46,47], emphasizing the potential for making complex AI models more understandable to a broader audience.

We observe a notable research gap in the utilization of various data types, particularly time-series data, point cloud data, and other dynamic data forms. Time-series data, critical in fields like finance, healthcare, and environmental studies, presents unique challenges and opportunities for XAI. This is complemented by the emerging relevance of point cloud data in sectors like autonomous vehicles and railway digital twins [48,49]. Time-series data are currently actively researched (see, e.g., [43,50]), while point cloud data still need more attention, though there are some attempt in this direction [51,52]. Expanding XAI research to include these diverse data types, alongside unsupervised and semi-supervised learning techniques, will broaden the scope and applicability of XAI methods.

An important aspect, as highlighted by [53,54], is the prevalent use of model-agnostic methods that create local surrogate models. These methods need to be refined to more accurately reflect the intricacies of the original "black box" models they are interpreting. Improving the accuracy and reliability of these model-agnostic methods, especially in their treatment of locality, is essential for the development of more transparent AI systems.

A pivotal direction for future research is enhancing the capabilities of existing XAI methods. For instance, addressing the computational limitations of widely used techniques like SHAP [55] is crucial. Research should focus on making these methods more computationally efficient and applicable for real-world industry scenarios. This enhancement will ensure that existing techniques remain relevant and useful in rapidly evolving AI landscapes.

Despite theoretical advancements, a gap exists in the practical application of XAI methods in industry. Future research must bridge this gap by refining XAI techniques to suit diverse industrial needs, considering aspects such as computational efficiency, usability, and scalability.

While our review successfully navigates around the issues of double counting in tertiary reviews, it primarily offers a binary overview of covered and uncovered topics in XAI. An extension of this work through a quantitative analysis of the volume of research in each identified gap could offer a more comprehensive understanding of the distribution and depth of current research efforts in XAI.

5.2. Threats to Validity and Limitations

To perform a tertiary review on a field that has existed for roughly six years, various assumptions are required, and certain limitations apply.

To conduct a comprehensive tertiary review, the choice of search query terms played a crucial role. Initially, "meta-analysis review" was employed as the primary term for our queries. However, upon reflection, the term "meta-review" may have been more inclusive, potentially capturing a broader spectrum of existing literature. While this alternative could have also increased the number of false positives, such instances could have been effectively filtered out during the title and abstract screening stage. Adopting this approach may have resulted in the inclusion of more relevant articles, thereby enriching our review.

A significant limitation we encountered was restricted access to some of the existing literature. To counter this, we utilized the institutional access provided by the University of Twente, which facilitated the collection of literature beyond what is openly accessible. Despite these efforts, certain papers remained inaccessible in their entirety, leading to their exclusion from our review. This restriction may have consequently narrowed the scope of our analysis, impacting the comprehensiveness of our findings.

Conducting a tertiary review inherently involves reliance on systematic literature reviews of primary research, which adds a layer of abstraction to our conclusions. This method means that our insights are indirectly shaped by the depth and interpretations presented in these secondary sources. Such a reliance could introduce variations or potential misinterpretations in our analysis, stemming from the methodologies, interpretations, and selection bias used in both the primary studies and the systematic literature reviews.

5.3. Limitations

The presented methods and techniques are extracted from the published systematic literature review; therefore, current state-of-the-art techniques could be missed from our analysis. For example, the recent proposal that combines LIME and evolutionary algorithms [56] is not included in our analysis because it has not yet been reported in any systematic literature review.

Another limitation of our work is an absence of certain combinations in the matrices. Due to the rapidly evolving nature of XAI, certain combinations may have been missed for reasons such as their absence in the selected SLRs or because they were recently reported in the literature and are therefore not part of the selected SLRs. Some examples are already reported in Section 4.4, albeit not exhaustively. Therefore, possible research gaps should be taken with caution.

We would like to emphasize an inherent limitation of tertiary reports, namely selection bias (see, e.g., [57]), as they tend to include only studies that meet specific criteria or report positive outcomes. Since a tertiary review is a meta study that extracts knowledge from systematic reviews, the selection bias is potentially enhanced. This often results in missing details from primary studies and an over-representation of successful or widely accepted methods while excluding those that were less effective or industry-oriented.

In this paper, we consciously refrained from describing XAI methods in detail since that would add more volume to the paper and there are other sources that describe them much better. For a brief description of the most well-known techniques, we refer to the chapter [58] and a detailed description can be found in the wonderful book by Molnar [27].

6. Conclusions

This comprehensive tertiary review systematically synthesized XAI methodologies, distilling key characteristics and categorizing them into a grid framework. Our analysis of 40 systematic literature reviews from 1992 to 2023 has revealed significant gaps in the field, particularly in under-researched areas of XAI characteristics and categories. We identified 14 open research directions and a similar number for research directions that are unfeasible to research. These findings underscore the necessity for targeted research to bridge these gaps, enriching the body of knowledge in XAI. Also, it emphasizes the need to further refine the existing methods and develop new techniques for the other underdeveloped areas in the XAI landscape. Furthermore, this study highlights the diverse nature of XAI methods, ranging from intrinsic to post hoc explainability. The implications of our findings are far-reaching, offering a road map for future research and development in XAI, which is crucial for the advancement of transparent, accountable, and ethical AI systems. While our study provides a foundational understanding of the current state of XAI research, it also acknowledges its limitations, including potential selection biases and the scope of the literature reviewed. This work serves as a call to action for the research community to delve deeper into the unexplored territories of XAI, fostering innovation and progress in this vital field.

In conclusion, the future of XAI research lies in its expansion to unexplored domains, diversification of data types and methodologies, and the bridging of the gap between theoretical research and practical, industry-oriented applications. This directional shift will not only enrich the field of XAI but also ensure its relevance and applicability in solving real-world problems.

Appendix A. Quality Assessment Papers

Table A1. Full-text read articles with exclusion reason.

Article	Included?	Exclusion Reason
Haque et al. [59]	✓	
Groen et al. [14]	✓	
Chen et al. [60]	✓	
Ahmed et al. [35]	✓	
Loh et al. [61]	✓	
Doh et al. [62]	✓	
Okolo et al. [63]	✓	
Yuan and Bertini [64]	✓	
Adak et al. [37]	✗	Irrelevant
Hauser et al. [45]	✓	
Chou et al. [28]	✓	
Hakkoum et al. [65]	✓	
European Commission [5]	✓	
Nimmy et al. [66]	✓	
Askr et al. [34]	✓	
Rajabi and Etminani [44]	✓	
Engelmann et al. [67]	✓	
Giuste et al. [68]	✓	
Vo et al. [69]	✗	Irrelevant
Li et al. [33]	✓	
Salahuddin et al. [70]	✓	
Tiddi and Schlobach [32]	✓	
Nor et al. [71]	✗	Irrelevant
Vilone and Longo [72]	✓	
Vilone and Longo [13]	✓	
Antoniadi et al. [36]	✓	
Wells and Bednarz [73]	✓	
Darias et al. [40]	✓	
Nazar et al. [74]	✓	
Chakrobartty and El-Gayar [38]	✓	
Chazette et al. [75]	✓	
Alamri and Alharbi [76]	✓	
Stepin et al. [77]	✓	
Payrovnaziri et al. [78]	✓	
Chatzimparmpas et al. [79]	✓	
Grossberg [80]	✓	
Kadam and Ahirrao [81]	✗	Does not introduce XAI methods
Cirqueira et al. [82]	✗	Does not introduce XAI methods
Naiseh et al. [83]	✓	
Chromik and Schuessler [84]	✗	Does not introduce XAI methods
Anjomshoae et al. [85]	✓	
Seeliger et al. [86]	✓	
Neto et al. [87]	✗	Irrelevant
Saleem et al. [15]	✓	
Emaminejad and Akhavian [88]	✗	Does not introduce XAI methods
Weber et al. [89]	✓	
Siering et al. [90]	✗	Irrelevant
Heckler et al. [91]	✗	Irrelevant
Guidotti [39]	✓	
Nunes and Jannach [92]	✗	Does not introduce XAI methods
Nauta et al. [19]	✓	

Appendix B. Backwards Snowballing

Table A2. Resulting articles from backwards snowballing, including references paper, based on title and abstract screening.

Newly Found Article	Referenced by
Guidotti et al. [16]	Nauta et al. [19]
Adadi and Berrada [41]	Nauta et al. [19]
Mary [93]	Chazette et al. [75]
Linardatos et al. [94]	Antoniadi et al. [36], Haque et al. [59], Nazar et al. [74]
Nunes and Jannach [92]	Chazette et al. [75], Naiseh et al. [83], Nauta et al. [19]
Gulum et al. [95]	Chen et al. [60], European Commission [5]
Salahuddin et al. [70]	Chen et al. [60], Li et al. [33]
Verma et al. [96]	Chou et al. [28], Darias et al. [40], Guidotti [39], Salahuddin et al. [70], Yuan and Bertini [64]
Mueller et al. [97]	Seeliger et al. [86]
Yang et al. [98]	Groen et al. [14], Nazar et al. [74], Salahuddin et al. [70], Weber et al. [89]
Anjomshoae et al. [85]	Chazette et al. [75], Guidotti [39], Haque et al. [59], Stepin et al. [77], Wells and Bednarz [73]
Barakat and Bradley [99]	Hakkoum et al. [65]
Wells and Bednarz [73]	Haque et al. [59]
Laato et al. [100]	Haque et al. [59]
Gerlings et al. [101]	Haque et al. [59]
Payrovnaziri et al. [78]	European Commission [5], Loh et al. [61], Nazar et al. [74], Weber et al. [89]
Lacave and Diez [102]	European Commission [5], Vilone and Longo [72]
Vilone and Longo [103]	European Commission [5], Vilone and Longo [13], Weber et al. [89]
Alam and Mueller [104]	Li et al. [33]
Antoniadi et al. [36]	Li et al. [33], Weber et al. [89]
Fuhrman et al. [105]	Li et al. [33]
Nazar et al. [74]	Li et al. [33], Okolo et al. [63]
Joshi et al. [106]	Nauta et al. [19]
Kulakova and Nieuwland [107]	Stepin et al. [77]
Kulakova and Nieuwland [107]	Vilone and Longo [13]
Minh et al. [108]	Weber et al. [89]

References

1. Miller, T. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.* **2019**, *267*, 1–38. [CrossRef]
2. Lee, M.S.A.; Floridi, L. Algorithmic Fairness in Mortgage Lending: From Absolute Conditions to Relational Trade-offs. *Minds Mach.* **2021**, *31*, 165–191. [CrossRef]
3. Amann, J.; Blasimme, A.; Vayena, E.; Frey, D.; Madai, V.I.; consortium, t.P. Explainability for artificial intelligence in healthcare: A multidisciplinary perspective. *BMC Med. Inform. Decis. Mak.* **2020**, *20*, 310. [CrossRef]
4. Seyyed-Kalantari, L.; Zhang, H.; McDermott, M.B.; Chen, I.Y.; Ghassemi, M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nat. Med.* **2021**, *27*, 2176–2182. [CrossRef] [PubMed]
5. European Commission. Proposal for a Regulation of the European Parliament and of the Council on Artificial Intelligence and Amending Certain Union Legislative Acts. *EUR-Lex* **2021**, *COM(2021) 206 final*. Available online: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206 (accessed on 27 August 2024).
6. Gianfagna, L.; Di Cecco, A. *Explainable AI with Python*; Springer: Cham, Switzerland, 2021.
7. Islam, M.; Ahmed, M.; Barua, S.; Begum, S. A Systematic Review of Explainable Artificial Intelligence in Terms of Different Application Domains and Tasks. *Appl. Sci.* **2022**, *12*, 1353. [CrossRef]
8. Kitchenham, B.; Charters, S. *Guidelines for Performing Systematic Literature Reviews in Software Engineering*; Technical Report EBSE-2007-01; University of Durham: Durham, UK, 2007.
9. Fusar-Poli, P.; Radua, J. Ten simple rules for conducting umbrella reviews. *Evid. Based Ment. Health* **2018**, *21*, 95–100. [CrossRef] [PubMed]
10. Saeed, W.; Omlin, C. Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities. *Knowl.-Based Syst.* **2023**, *263*, 110273. [CrossRef]
11. Mongeon, P.; Paul-Hus, A. The journal coverage of Web of Science and Scopus: A comparative analysis. *Scientometrics* **2016**, *106*, 213–228. [CrossRef]
12. Wohlin, C. Guidelines for snowballing in systematic literature studies and a replication in software engineering. In Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering, London, UK, 13–14 May 2014; pp. 1–10.
13. Vilone, G.; Longo, L. Classification of Explainable Artificial Intelligence Methods through Their Output Formats. *Mach. Learn. Knowl. Extr.* **2021**, *3*, 615–661. [CrossRef]

14. Groen, A.; Kraan, R.; Amirkhan, S.; Daams, J.; Maas, M. A systematic review on the use of explainability in deep learning systems for computer aided diagnosis in radiology: Limited use of explainable AI? *Eur. J. Radiol.* **2022**, *157*, 110592. [CrossRef]

15. Saleem, R.; Yuan, B.; Kurugollu, F.; Anjum, A.; Liu, L. Explaining deep neural networks: A survey on the global interpretation methods. *Neurocomputing* **2022**, *513*, 165–180. [CrossRef]

16. Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Pedreschi, D.; Giannotti, F. A Survey Of Methods For Explaining Black Box Models. *arXiv* **2018**, arXiv:1802.01933. [CrossRef]

17. Burkart, N.; Huber, M.F. A Survey on the Explainability of Supervised Machine Learning. *J. Artif. Intell. Res.* **2020**, *70*, 245–317. [CrossRef]

18. Sarker, I.H. Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Comput. Sci.* **2021**, *2*, 160. [CrossRef]

19. Nauta, M.; Trienes, J.; Pathak, S.; Nguyen, E.; Peters, M.; Schmitt, Y.; Schlötterer, J.; Van Keulen, M.; Seifert, C. From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI. *ACM Comput. Surv.* **2023**, *55*, 1–42. [CrossRef]

20. Wang, X.; Wang, D.; Xu, C.; He, X.; Cao, Y.; Chua, T.S. Explainable reasoning over knowledge graphs for recommendation. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 5329–5336.

21. Ribeiro, M.T.; Singh, S.; Guestrin, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *arXiv* **2016**, arXiv:1602.04938. [CrossRef]

22. Lundberg, S.M.; Lee, S.I. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*; Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30.

23. 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.

24. Macha, D.; Kozielski, M.; Wróbel, Ł.; Sikora, M. RuleXAI—A package for rule-based explanations of machine learning model. *SoftwareX* **2022**, *20*, 101209. [CrossRef]

25. Craven, M.; Shavlik, J. Extracting tree-structured representations of trained networks. *Adv. Neural Inf. Process. Syst.* **1995**, *8*, 24–30.

26. Chen, X.; Duan, Y.; Houthooft, R.; Schulman, J.; Sutskever, I.; Abbeel, P. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *Adv. Neural Inf. Process. Syst.* **2016**, *29*, 1–14.

27. Molnar, C. *Interpretable Machine Learning*; Lulu.com: Raleigh, NC, USA, 2020.

28. Chou, Y.L.; Moreira, C.; Bruza, P.; Ouyang, C.; Jorge, J. Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms, and applications. *Inf. Fusion* **2022**, *81*, 59–83. [CrossRef]

29. Kim, B.; Khanna, R.; Koyejo, O.O. Examples are not enough, learn to criticize! Criticism for interpretability. *Adv. Neural Inf. Process. Syst.* **2016**, *29*, 1–16.

30. Rajani, N.F.; McCann, B.; Xiong, C.; Socher, R. Explain Yourself! Leveraging Language Models for Commonsense Reasoning. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 4932–4942.

31. Olah, C.; Mordvintsev, A.; Schubert, L. Feature visualization. *Distill* **2017**, *2*, e7. [CrossRef]

32. Tiddi, I.; Schlobach, S. Knowledge graphs as tools for explainable machine learning: A survey. *Artif. Intell.* **2022**, *302*, 103627. [CrossRef]

33. Li, X.H.; Cao, C.; Shi, Y.; Bai, W.; Gao, H.; Qiu, L.; Wang, C.; Gao, Y.; Zhang, S.; Xue, X.; et al. A Survey of Data-Driven and Knowledge-Aware eXplainable AI. *IEEE Trans. Knowl. Data Eng.* **2022**, *34*, 29–49. [CrossRef]

34. Askr, H.; Elgeldawi, E.; Aboul Ella, H.; Elshaier, Y.; Gomaa, M.; Hassanien, A. Deep learning in drug discovery: An integrative review and future challenges. *Artif. Intell. Rev.* **2023**, *56*, 5975–6037. [CrossRef]

35. Ahmed, S.; Solis-Oba, R.; Ilie, L. Explainable-AI in Automated Medical Report Generation Using Chest X-ray Images. *Appl. Sci.* **2022**, *12*, 11750. [CrossRef]

36. Antoniadi, A.; Du, Y.; Guendouz, Y.; Wei, L.; Mazo, C.; Becker, B.; Mooney, C. Current challenges and future opportunities for XAI in machine learning-based clinical decision support systems: A systematic review. *Appl. Sci.* **2021**, *11*, 5088. [CrossRef]

37. Adak, A.; Pradhan, B.; Shukla, N. Sentiment Analysis of Customer Reviews of Food Delivery Services Using Deep Learning and Explainable Artificial Intelligence: Systematic Review. *Foods* **2022**, *11*, 1500. [CrossRef]

38. Chakrobartty, S.; El-Gayar, O. Explainable artificial intelligence in the medical domain: A systematic review. In Proceedings of the AMCIS 2021, Virtual, 9–13 August 2021.

39. Guidotti, R. Counterfactual explanations and how to find them: Literature review and benchmarking. *Data Min. Knowl. Discov.* **2022**. [CrossRef]

40. Darias, J.; DÃaz-Agudo, B.; Recio-Garcia, J. A systematic review on model-agnostic XAI libraries. *ICCBR Work.* **2021**, *3017*, 28–29.

41. Adadi, A.; Berrada, M. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access* **2018**, *6*, 52138–52160. [CrossRef]

42. Pisner, D.A.; Schnyer, D.M. Chapter 6—Support vector machine. In *Machine Learning*; Mechelli, A., Vieira, S., Eds.; Academic Press: San Diego, CA, USA, 2020; pp. 101–121. [CrossRef]

43. Maças, C.; Campos, J.R.; Lourenço, N. Understanding the Forest: A Visualization Tool to Support Decision Tree Analysis. In Proceedings of the 2023 27th International Conference Information Visualisation (IV), Tampere, Finland, 25–28 July 2023; pp. 223–229.

44. Rajabi, E.; Etminani, K. Knowledge-graph-based explainable AI: A systematic review. *J. Inf. Sci.* **2022**, *50*, 1019–1029. [CrossRef]

45. Hauser, K.; Kurz, A.; Haggenmüller, S.; Maron, R.; von Kalle, C.; Utikal, J.; Meier, F.; Hobelsberger, S.; Gellrich, F.; Sergon, M.; et al. Explainable artificial intelligence in skin cancer recognition: A systematic review. *Eur. J. Cancer* **2022**, *167*, 54–69. [CrossRef]

46. Szymanski, M.; Verbert, K.; Vanden Abeele, V. Designing and evaluating explainable AI for non-AI experts: Challenges and opportunities. In Proceedings of the 16th ACM Conference on Recommender Systems, Seattle, WA, USA, 18–23 September 2022; pp. 735–736.

47. Das, A.; Rad, P. Opportunities and challenges in explainable artificial intelligence (xai): A survey. *arXiv* **2020**, arXiv:2006.11371.

48. Abbasi, R.; Bashir, A.K.; Alyamani, H.J.; Amin, F.; Doh, J.; Chen, J. Lidar Point Cloud Compression, Processing and Learning for Autonomous Driving. *IEEE Trans. Intell. Transp. Syst.* **2023**, *24*, 962–979. [CrossRef]

49. Dekker, B.; Ton, B.; Meijer, J.; Bouali, N.; Linssen, J.; Ahmed, F. Point Cloud Analysis of Railway Infrastructure: A Systematic Literature Review. *IEEE Access* **2023**, *11*, 134355–134373. [CrossRef]

50. Tronchin, L.; Cordelli, E.; Celsi, L.R.; Maccagnola, D.; Natale, M.; Soda, P.; Sicilia, R. Translating Image XAI to Multivariate Time Series. *IEEE Access* **2024**, *12*, 27484–27500. [CrossRef]

51. Atik, M.E.; Duran, Z.; Seker, D.Z. Explainable Artificial Intelligence for Machine Learning-Based Photogrammetric Point Cloud Classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.* **2024**, *17*, 5834–5846. [CrossRef]

52. Zhang, M.; You, H.; Kadam, P.; Liu, S.; Kuo, C.C.J. Pointhop: An explainable machine learning method for point cloud classification. *IEEE Trans. Multimed.* **2020**, *22*, 1744–1755. [CrossRef]

53. Tsang, W.K.; Benoit, D.F. Interpretability and Explainability in Machine Learning. In *Living beyond Data: Toward Sustainable Value Creation*; Ohsawa, Y., Ed.; Springer International Publishing: Cham, Switzerland, 2023; pp. 89–100. [CrossRef]

54. Watson, D.S. Conceptual challenges for interpretable machine learning. *Synthese* **2022**, *200*, 65. [CrossRef]

55. Kumar, I.; Scheidegger, C.; Venkatasubramanian, S.; Friedler, S. Shapley Residuals: Quantifying the limits of the Shapley value for explanations. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 26598–26608.

56. Nematzadeh, H.; García-Nieto, J.; Navas-Delgado, I.; Aldana-Montes, J.F. Ensemble-based genetic algorithm explainer with automized image segmentation: A case study on melanoma detection dataset. *Comput. Biol. Med.* **2023**, *155*, 106613. [CrossRef] [PubMed]

57. Budgen, D.; Brereton, P.; Drummond, S.; Williams, N. Reporting systematic reviews: Some lessons from a tertiary study. *Inf. Softw. Technol.* **2018**, *95*, 62–74. [CrossRef]

58. Holzinger, A.; Saranti, A.; Molnar, C.; Biecek, P.; Samek, W. Explainable AI methods—A brief overview. In Proceedings of the International Workshop on Extending Explainable AI beyond Deep Models and Classifiers, Vienna, Austria, 18 July 2020; Springer: Cham, Switzerland, 2022; pp. 13–38.

59. Haque, A.; Islam, A.; Mikalef, P. Explainable Artificial Intelligence (XAI) from a user perspective: A synthesis of prior literature and problematizing avenues for future research. *Technol. Forecast. Soc. Chang.* **2023**, *186*, 122120. [CrossRef]

60. Chen, H.; Gomez, C.; Huang, C.M.; Unberath, M. Explainable medical imaging AI needs human-centered design: Guidelines and evidence from a systematic review. *Npj Digit. Med.* **2022**, *5*, 156. [CrossRef] [PubMed]

61. Loh, H.; Ooi, C.; Seoni, S.; Barua, P.; Molinari, F.; Acharya, U. Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022). *Comput. Methods Programs Biomed.* **2022**, *226*, 107161. [CrossRef] [PubMed]

62. Doh, R.; Zhou, C.; Arthur, J.; Tawiah, I.; Doh, B. A Systematic Review of Deep Knowledge Graph-Based Recommender Systems, with Focus on Explainable Embeddings. *Data* **2022**, *7*, 94. [CrossRef]

63. Okolo, C.; Dell, N.; Vashistha, A. Making AI Explainable in the Global South: A Systematic Review. In Proceedings of the 5th ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies, Seattle, WA, USA, 29 June–1 July 2022; pp. 439–452. [CrossRef]

64. Yuan, J.; Bertini, E. Context Sight: Model Understanding and Debugging via Interpretable Context. In Proceedings of the Workshop on Human-in-the-Loop Data Analytics, Philadelphia, PA, USA, 12 June 2022. [CrossRef]

65. Hakkoum, H.; Abnane, I.; Idri, A. Interpretability in the medical field: A systematic mapping and review study. *Appl. Soft Comput.* **2022**, *117*, 108391. [CrossRef]

66. Nimmy, S.; Hussain, O.; Chakrabortty, R.; Hussain, F.; Saberi, M. Explainability in supply chain operational risk management: A systematic literature review. *Knowl.-Based Syst.* **2022**, *235*, 107587. [CrossRef]

67. Engelmann, D.; Damasio, J.; Panisson, A.; Mascardi, V.; Bordini, R. Argumentation as a Method for Explainable AI: A Systematic Literature Review. In Proceedings of the 2022 17th Iberian Conference on Information Systems and Technologies (CISTI), Madrid, Spain, 22–25 June 2022. [CrossRef]

68. Giuste, F.; Shi, W.; Zhu, Y.; Naren, T.; Isgut, M.; Sha, Y.; Tong, L.; Gupte, M.; Wang, M.D. Explainable Artificial Intelligence Methods in Combating Pandemics: A Systematic Review. *IEEE Rev. Biomed. Eng.* **2022**, *16*, 5–21. [CrossRef] [PubMed]

69. Vo, T.; Nguyen, N.; Kha, Q.; Le, N. On the road to explainable AI in drug-drug interactions prediction: A systematic review. *Comput. Struct. Biotechnol. J.* **2022**, *20*, 2112–2123. [CrossRef]

70. Salahuddin, Z.; Woodruff, H.; Chatterjee, A.; Lambin, P. Transparency of deep neural networks for medical image analysis: A review of interpretability methods. *Comput. Biol. Med.* **2022**, *140*, 105111. [CrossRef]

71. Nor, A.; Pedapati, S.; Muhammad, M.; Leiva, V. Overview of explainable artificial intelligence for prognostic and health management of industrial assets based on preferred reporting items for systematic reviews and meta-analyses. *Sensors* **2021**, *21*, 8020. [CrossRef]

72. Vilone, G.; Longo, L. Notions of explainability and evaluation approaches for explainable artificial intelligence. *Inf. Fusion* **2021**, *76*, 89–106. [CrossRef]

73. Wells, L.; Bednarz, T. Explainable AI and Reinforcement Learning—A Systematic Review of Current Approaches and Trends. *Front. Artif. Intell.* **2021**, *4*, 550030. [CrossRef]

74. Nazar, M.; Alam, M.; Yafi, E.; Su'Ud, M. A Systematic Review of Human-Computer Interaction and Explainable Artificial Intelligence in Healthcare with Artificial Intelligence Techniques. *IEEE Access* **2021**, *9*, 153316–153348. [CrossRef]

75. Chazette, L.; Brunotte, W.; Speith, T. Exploring Explainability: A Definition, a Model, and a Knowledge Catalogue. In Proceedings of the 2021 IEEE 29th International Requirements Engineering Conference (RE), Notre Dame, IN, USA, 20–24 September 2021; pp. 197–208. [CrossRef]

76. Alamri, R.; Alharbi, B. Explainable Student Performance Prediction Models: A Systematic Review. *IEEE Access* **2021**, *9*, 33132–33143. [CrossRef]

77. Stepin, I.; Alonso, J.; Catala, A.; Pereira-Farina, M. A Survey of Contrastive and Counterfactual Explanation Generation Methods for Explainable Artificial Intelligence. *IEEE Access* **2021**, *9*, 11974–12001. [CrossRef]

78. Payrovnaziri, S.; Chen, Z.; Rengifo-Moreno, P.; Miller, T.; Bian, J.; Chen, J.; Liu, X.; He, Z. Explainable artificial intelligence models using real-world electronic health record data: A systematic scoping review. *J. Am. Med. Inform. Assoc.* **2020**, *27*, 1173–1185. [CrossRef]

79. Chatzimparmpas, A.; Martins, R.; Jusufi, I.; Kerren, A. A survey of surveys on the use of visualization for interpreting machine learning models. *Inf. Vis.* **2020**, *19*, 207–233. [CrossRef]

80. Grossberg, S. A Path Toward Explainable AI and Autonomous Adaptive Intelligence: Deep Learning, Adaptive Resonance, and Models of Perception, Emotion, and Action. *Front. Neurorobot.* **2020**, *14*, 36. [CrossRef]

81. Kadam, K.; Ahirrao, S. Bibliometric analysis of passive image forgery detection and explainable AI. *Libr. Philos. Pract.* **2020**, *2020*, 10959.

82. Cirqueira, D.; Nedbal, D.; Helfert, M.; Bezbradica, M. Scenario-Based Requirements Elicitation for User-Centric Explainable AI: A Case in Fraud Detection. In *Machine Learning and Knowledge Extraction*; Springer: Cham, Switzerland, 2020; Volume 12279 LNCS, pp. 321–341. [CrossRef]

83. Naiseh, M.; Jiang, N.; Ma, J.; Ali, R. Explainable Recommendations in Intelligent Systems: Delivery Methods, Modalities and Risks. In *Research Challenges in Information Science*; Springer: Cham, Switzerland, 2020; Volume 385 LNBIP, pp. 212–228. [CrossRef]

84. Chromik, M.; Schuessler, M. A taxonomy for human subject evaluation of black-box explanations in XAI. In Proceedings of the ExSS-ATEC'20, Cagliari, Italy, 17 March 2020; Volume 2582.

85. Anjomshoae, S.; Calvaresi, D.; Najjar, A.; Framling, K. Explainable agents and robots: Results from a systematic literature review. In Proceedings of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, QC, Canada, 13–17 May 2019; Volume 2, pp. 1078–1088.

86. Seeliger, A.; Pfaff, M.; Krcmar, H. Semantic web technologies for explainable machine learning models: A literature review. *CEUR Workshop Proc.* **2019**, *2465*, 30–45.

87. Neto, A.V.S.; Camargo, J.B.; Almeida, J.R.; Cugnasca, P.S. Safety Assurance of Artificial Intelligence-Based Systems: A Systematic Literature Review on the State of the Art and Guidelines for Future Work. *IEEE Access* **2022**, *10*, 130733–130770. [CrossRef]

88. Emaminejad, N.; Akhavian, R. Trustworthy AI and robotics: Implications for the AEC industry. *Autom. Constr.* **2022**, *139*, 104298. [CrossRef]

89. Weber, L.; Lapuschkin, S.; Binder, A.; Samek, W. Beyond explaining: Opportunities and challenges of XAI-based model improvement. *Inf. Fusion* **2023**, *92*, 154–176. [CrossRef]

90. Siering, M.; Deokar, A.V.; Janze, C. Disentangling consumer recommendations: Explaining and predicting airline recommendations based on online reviews. *Decis. Support Syst.* **2018**, *107*, 52–63. [CrossRef]

91. Heckler, W.F.; de Carvalho, J.V.; Barbosa, J.L.V. Machine learning for suicidal ideation identification: A systematic literature review. *Comput. Hum. Behav.* **2022**, *128*, 107095. [CrossRef]

92. Nunes, I.; Jannach, D. A systematic review and taxonomy of explanations in decision support and recommender systems. *User Model. User-Adapt. Interact.* **2017**, *27*, 393–444. [CrossRef]

93. Mary, S. Explainable Artificial Intelligence Applications in NLP, Biomedical, and Malware Classification: A Literature Review. In *Intelligent Computing*; Springer: Cham, Switzerland, 2019; pp. 1269–1292. [CrossRef]

94. Linardatos, P.; Papastefanopoulos, V.; Kotsiantis, S. Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy* **2021**, *23*, 18. [CrossRef]

95. Gulum, M.A.; Trombley, C.M.; Kantardzic, M. A Review of Explainable Deep Learning Cancer Detection Models in Medical Imaging. *Appl. Sci.* **2021**, *11*, 4573. [CrossRef]

96. Verma, S.; Boonsanong, V.; Hoang, M.; Hines, K.E.; Dickerson, J.P.; Shah, C. Counterfactual Explanations and Algorithmic Recourses for Machine Learning: A Review. *arXiv* **2020**, arXiv:2010.10596. [CrossRef]

97. Mueller, S.T.; Hoffman, R.R.; Clancey, W.; Emrey, A.; Klein, G. Explanation in Human-AI Systems: A Literature Meta-Review, Synopsis of Key Ideas and Publications, and Bibliography for Explainable AI. *arXiv* **2019**, arXiv:1902.01876. [CrossRef]

98. Yang, G.; Ye, Q.; Xia, J. Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond. *Inf. Fusion* **2022**, *77*, 29–52. [CrossRef] [PubMed]
99. Barakat, N.; Bradley, A.P. Rule extraction from support vector machines: A review. *Neurocomputing* **2010**, *74*, 178–190. [CrossRef]
100. Laato, S.; Tiainen, M.; Najmul Islam, A.K.M.; Mäntymäki, M. How to explain AI systems to end users: A systematic literature review and research agenda. *Internet Res.* **2022**, *32*, 1–31. [CrossRef]
101. Gerlings, J.; Shollo, A.; Constantiou, I. Reviewing the Need for Explainable Artificial Intelligence (xAI). *arXiv* **2020**, arXiv:2012.01007. [CrossRef]
102. Lacave, C.; Diez, F.J. A review of explanation methods for Bayesian networks. *Knowl. Eng. Rev.* **2002**, *17*, 107–127. [CrossRef]
103. Vilone, G.; Longo, L. Explainable Artificial Intelligence: A Systematic Review. *arXiv* **2020**, arXiv:2006.00093. [CrossRef]
104. Alam, L.; Mueller, S. Examining Physicians' Explanatory Reasoning in Re-Diagnosis Scenarios for Improving AI Diagnostic Systems. *J. Cogn. Eng. Decis. Mak.* **2022**, *16*, 155534342210851. [CrossRef]
105. Fuhrman, J.D.; Gorre, N.; Hu, Q.; Li, H.; El Naqa, I.; Giger, M.L. A review of explainable and interpretable AI with applications in COVID-19 imaging. *Med. Phys* **2021**, *49*, 1–14. [CrossRef] [PubMed]
106. Joshi, G.; Walambe, R.; Kotecha, K. A Review on Explainability in Multimodal Deep Neural Nets. *IEEE Access* **2021**, *9*, 59800–59821. [CrossRef]
107. Kulakova, E.; Nieuwland, M.S. Understanding Counterfactuality: A Review of Experimental Evidence for the Dual Meaning of Counterfactuals. *Lang Linguist Compass* **2016**, *10*, 49–65. [CrossRef]
108. Minh, D.; Wang, H.X.; Li, Y.F.; Nguyen, T.N. Explainable artificial intelligence: A comprehensive review. *Artif. Intell. Rev.* **2022**, *55*, 3503–3568. [CrossRef]