applied sciences

MDPI

Systematic Review

A Systematic Review of Explainable Artificial Intelligence in Terms of Different Application Domains and Tasks

Mir Riyanul Islam *, Mobyen Uddin Ahmed, Shaibal Barua and Shahina Begum

Artificial Intelligence and Intelligent Systems Research Group, School of Innovation, Design and Engineering, Mälardalen University, Högskoleplan 1, 722 20 Västerås, Sweden; mobyen.ahmed@mdh.se (M.U.A.); shaibal.barua@mdh.se (S.B.); shahina.begum@mdh.se (S.B.)

* Correspondence: mir.riyanul.islam@mdh.se; Tel.: +46-21-10-3182

Abstract: Artificial intelligence (AI) and machine learning (ML) have recently been radically improved and are now being employed in almost every application domain to develop automated or semi-automated systems. To facilitate greater human acceptability of these systems, explainable artificial intelligence (XAI) has experienced significant growth over the last couple of years with the development of highly accurate models but with a paucity of explainability and interpretability. The literature shows evidence from numerous studies on the philosophy and methodologies of XAI. Nonetheless, there is an evident scarcity of secondary studies in connection with the application domains and tasks, let alone review studies following prescribed guidelines, that can enable researchers' understanding of the current trends in XAI, which could lead to future research for domain- and application-specific method development. Therefore, this paper presents a systematic literature review (SLR) on the recent developments of XAI methods and evaluation metrics concerning different application domains and tasks. This study considers 137 articles published in recent years and identified through the prominent bibliographic databases. This systematic synthesis of research articles resulted in several analytical findings: XAI methods are mostly developed for safety-critical domains worldwide, deep learning and ensemble models are being exploited more than other types of AI/ML models, visual explanations are more acceptable to end-users and robust evaluation metrics are being developed to assess the quality of explanations. Research studies have been performed on the addition of explanations to widely used AI/ML models for expert users. However, more attention is required to generate explanations for general users from sensitive domains such as finance and the judicial system.

Keywords: explainable artificial intelligence; explainability; evaluation metrics; systematic literature review

Citation: Islam, M.R.; Ahmed, M.U.; Barua, S.; Begum, S. A Systematic Review of Explainable Artificial Intelligence in Terms of Different Application Domains and Tasks. *Appl. Sci.* **2022**, *12*, 1353. https://doi.org/10.3390/app12031353

Academic Editor: Agostino Forestiero and Byung-Gyu Kim

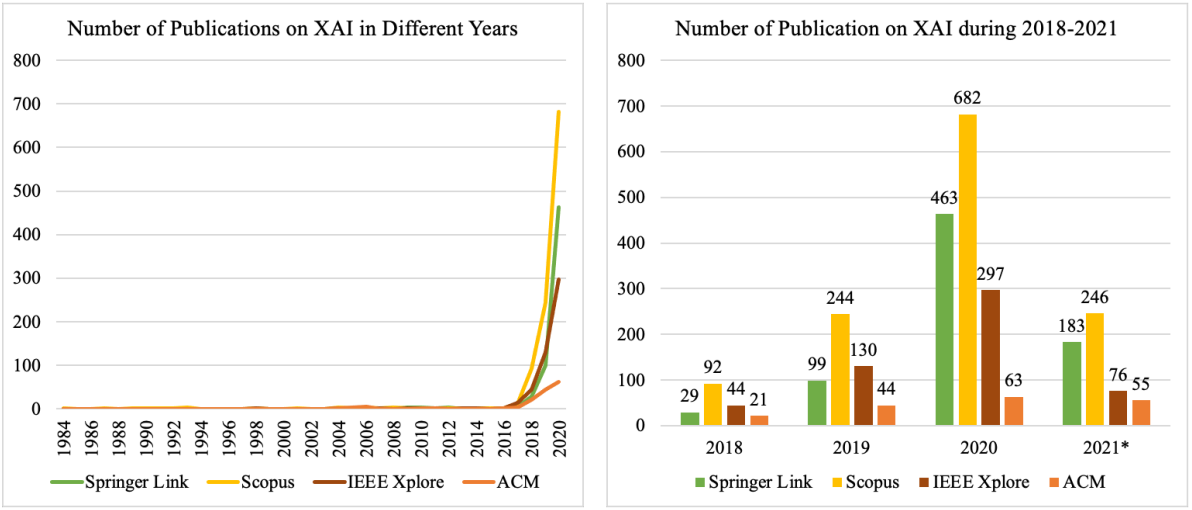
Received: 17 November 2021
Accepted: 25 January 2022
Published: 27 January 2022

1. Introduction

With the recent developments of artificial intelligence (AI) and machine learning (ML) algorithms, people from various application domains have shown increasing interest in taking advantage of these algorithms. As a result, AI and ML are being used today in many application domains. Different AI/ML algorithms are being employed to complement humans' decisions in various tasks from diverse domains, such as education, construction, health care, news and entertainment, travel and hospitality, logistics, manufacturing, law enforcement, and finance [1]. While these algorithms are meant to help users in their daily tasks, they still face acceptability issues. Users often remain doubtful about the proposed decisions. In worse cases, users oppose the AI/ML model's decision since their inference mechanisms are mostly opaque, unintuitive, and incomprehensible to humans. For example, today, deep learning (DL) models demonstrate convincing results with improved accuracy compared to established algorithms. DL models' outstanding performances hide one major drawback, i.e., the underlying inference mechanism remains

unknown to a user. In other words, the DL models function as a black-box [2]. In general, almost all the prevailing expert systems built with AI/ML models do not provide additional information to support the inference mechanism, which makes systems nontransparent. Thus, it has become a sine qua non to investigate how the inference mechanism or the decisions of AI/ML models can be made transparent to humans so that these intelligent systems can become more acceptable to users from different application domains [3].

Upon realising the need to explain AI/ML model-based intelligent systems, a few researchers started exploring and proposing methods long ago. The bibliographic databases contain the earliest published evidence on the association between expert systems and the term *explanation* from the mid-eighties [4]. Over time, the concept evolved to be an immense growing research domain of explainable artificial intelligence (XAI). However, researchers did not pay much attention to XAI until 2017/2018 which can be justified by the trend of publications per year with the keyword *explainable artificial intelligence* in titles or abstracts from different bibliographic databases illustrated in Figure 1a. The increased attention paid by researchers towards XAI from all the domains utilising systems developed with AI/ML models was caused by three major incidents. First of all, the funding of the "Explainable AI (XAI) Program" was funded in early 2017 by the Defense Advanced Research Projects Agency (DARPA) [5]. After a couple of months in mid-2017, the Chinese government released "The Development Plan for New Generation of Artificial Intelligence" to encourage the high and strong extensibility of AI [6]. Last but not least, in mid-2018, the European Union granted their citizens a "Right to Explanation" if they were affected by algorithmic decision making by publishing the "General Data Protection Regulation" (GDPR) [7]. The impact of these events is prominent among the researchers since the search results from the significant bibliographic databases depict a rapidly increasing number of publications related to XAI during recent years (Figure 1b). The bibliographic databases that were considered to assess the number of publications per year on XAI were found to be the main sources of the research articles from the AI domain.



(a) (b)

Figure 1. Number of published articles (y axis) on XAI made available through four bibliographic databases in recent decades (x axis): (**a**) Trend of the number of publications from 1984 to 2020. (**b**) Specific number of publications from 2018 to June 2021. The illustrated data were extracted on 1 July 2021 from four renowned bibliographic databases. The asterisk (*) with 2021 refers to the partial data on the number of publications on XAI until June.

The continuously increasing momentum of publications in the domain of XAI is producing an abundance of knowledge from various perspectives, e.g., philosophy, taxonomy, and development. Unfortunately, this scattered plentiful knowledge and the use of differ-

ent closely related taxonomies interchangeably demand the organisation and definition of boundaries through a systematic literature review (SLR), as it contains a structured procedure for conducting the review with provisions for assessing the outcome in terms of a predefined goal. Figure 2 presents the distribution of articles on XAI methods for various application domains and tasks. From Figure 2a, it is realisable that today, most of XAI methods are developed as domain agnostic. However, the most influential use of XAI is in the healthcare domain; this may be because of the demand for explanations from the end-user perspective. Obviously, in many application domains, AI and ML methods are used for decision support systems, and the need for XAI is high for decision support tasks, as can be seen in Figure 2b. Although there is an increasing number of publications, some challenges have not been considered, for example, user-centric and domain knowledge incorporating explanation. This article aimed to present the outcome of an SLR on the current developments and trends in XAI for different application domains by summarising the methods and evaluation metrics for explainable AI/ML models. Moreover, the aim of this SLR includes identifying the specific domains and applications in which XAI methods are exploited and that are to be further investigated. To achieve the aim of this study, three major objectives are highlighted:

- To investigate and present the application domains and tasks for which various XAI methods have been explored and exploited;
- To investigate and present the XAI methods, validation metrics and the type of explanations that can be generated to increase the acceptability of the expert systems to general users;
- To sort out the open issues and future research directions in terms of various domains and application tasks from the methodological perspective of XAI.

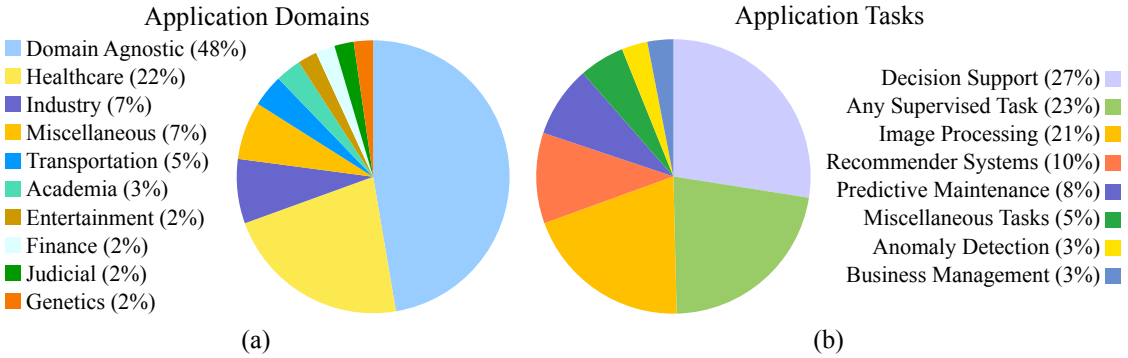


Figure 2. Percentage of the selected articles on different XAI methods for different application (**a**) domains and (**b**) tasks.

The remainder of this article is arranged as follows: relevant concepts of XAI from a technical point of view are presented in Section 2, followed by a discussion on prominent review studies previously conducted on XAI in Section 3. Section 4 contains the detailed workflow of this SLR, followed by the outcome of the performed analyses in Section 5. Finally, a discussion on the findings of this study and its limitations and conclusions are presented in Sections 6 and 7, respectively.

2. Theoretical Background

This section concisely presents the theoretical aspects of XAI from a technical point of view for a better understanding of the contents of this study. Emphatically, the philosophy and taxonomy of XAI have been excluded from this manuscript because they are out of the scope of this study. However, the term *explainability* is associated with the interface between decision makers and humans. This interface is synchronously comprehensible to humans and accurately represents the decision maker [2]. Specifically, in XAI, the interface between the models and the end-users is called explainability, through which an end-user

obtains clarification on the decisions that the AI/ML model provides them with. Based on the literature, the concepts of XAI within different application domains are categorised as stage, scope, input and output formats. This section includes a discussion on the most relevant aspects that seem necessary to make XAI efficiently and credibly work on different applications. Figure 3 summarises the prime concepts behind developing XAI applications which were adopted from the recent review studies by Vilone and Longo [8,9].

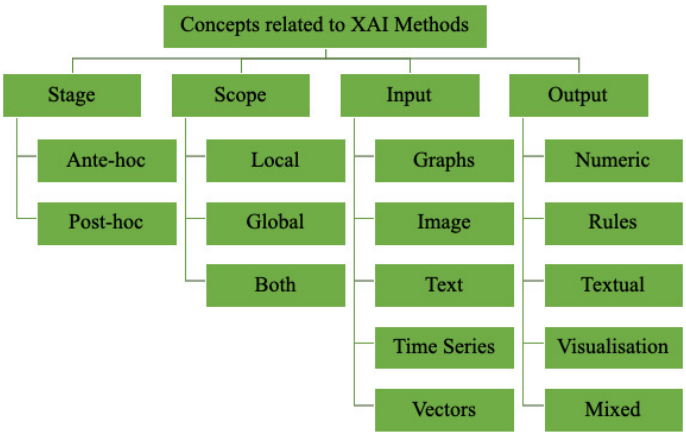


Figure 3. Overview of the different concepts on developing methodologies for XAI, adapted from the review studies by Vilone and Longo [8,9].

2.1. Stage of Explainability

The AI/ML models learn the fundamental characteristics of the supplied data and subsequently try to cluster, predict or classify unseen data. The stage of explainability refers to the period in the process mentioned above when a model generates the explanation for the decision it provides. According to Vilone and Longo, the stages are *ante hoc* and post hoc [8,9]. Brief descriptions of the stages are as follows:

- *Ante hoc* methods generally consider generating the explanation for the decision from the very beginning of the training on the data while aiming to achieve optimal performance. Mostly, explanations are generated using these methods for transparent models, such as fuzzy models and tree-based models;
- *Post hoc* methods comprise an external or surrogate model and the base model. The base model remains unchanged, and the external model mimics the base model's behaviour to generate an explanation for the users. Generally, these methods are associated with the models in which the inference mechanism remains unknown to users, e.g., support vector machines and neural networks. Moreover, the post hoc methods are again divided into two categories: model-agnostic and model-specific. The model-agnostic methods apply to any AI/ML model, whereas the model-specific methods are confined to particular models.

2.2. Scope of Explainability

The scope of explainability defines the extent of an explanation produced by some explainable methods. Two recent literature studies on more than 200 scientific articles published on XAI deduced that the scope of explainability can be either *global* or *local* [8,9]. With a *global* scope, the whole inferential technique of a model is made transparent or comprehensible to the user, for example, a decision tree. On the other hand, explanation with a *local* scope refers to explicitly explaining a single instance of inference to the user, e.g., for decision trees, a single branch can be termed as a local explanation.

2.3. Input and Output

Along with the core concepts, stages and scopes of explainability, input and output formats were also found to be significant in developing XAI methods [2,8,9]. The explainable

models' mechanisms unquestionably differ when learning different input data types, such as images, numbers, texts, etc. Including these basic forms of input, several others are found to be utilised in different studies, which are elaborately discussed in Section 5.3.1. Finally, the prime concern of XAI, the output format or the form of explanation varies following the solution to the prior problems. The different forms of explanation simultaneously vary concerning the circumstances and expertise of the end-users. The most common forms of explanations are numeric, rules, textual, visual and mixed. These forms of explanation are illustrated and briefly discussed in Section 5.3.4.

3. Related Studies

During the past couple of years, research on the developing theories, methodologies and tools of XAI has been very active, and over time, the popularity of XAI as a research domain has continued to increase. Before the massive attention of researchers towards XAI, the earliest review that could be found in the literature was that by Lacave and Diéz [10]. They reviewed the then prevailing explanation methods precisely for Bayesian networks. In the article, the authors referred to the level and methods of explanations followed by several techniques that were mostly probabilistic. Later, Ribeiro et al. reviewed the suggested interpretable models as a solution to the problem of adding explainability to AI/ML models, such as additive models, decision trees, attention-based networks, and sparse linear models [11]. Subsequently, they proposed a model-agnostic technique that involves the combined development of an interpretable model from the predictions of black-box and perturbing inputs to observe the reaction of black-box models [12].

With the remarkable implications of GDPR, an enormous number of works have been published in recent years. The initial works included the notion of explainability and its use from different points of view. Alonso et al. accumulated the bibliometric information on the XAI domain to understand the research trends, identify the potential research groups and locations, and discover possible research directions [13]. Gobel et al. discussed older concepts and linked them to newer concepts such as deep learning [14]. Black-box models were compared with the white-box models based on their advantages and disadvantages from a practical point of view [3]. Additionally, survey articles were published that advocated that explainable models replace black-box models for high-stakes decision-making tasks [1,15]. Surveys were also conducted on the methods of explainability and addressed the philosophy behind the usage from the perspective of different domains [16–18] and stakeholders [19]. Some works included the specific definitions of technical terms, possible applications, and challenges towards attaining responsible AI [6,20,21]. Adadi and Berrada and Guidotti et al. separately studied the available methods of explainability and clustered them in the form of explanations, e.g., textual, visual, and numeric [22,23]. However, the literature contains a good number of review studies on specific forms or methods of explaining AI/ML models. For example, Robnik-Sikonja and Bohanec conducted a literature review on the perturbation-based explanations for prediction models [24], Zhang et al. surveyed the techniques of providing visual explanations for deep learning models [25], and Daglarli reviewed the XAI approaches for deep meta-learning models [26].

Above all, several review studies were conducted by Vilone and Longo to gather and present the recent developments in XAI [8,9,27]. These studies presented extensive clustering of the XAI methods and evaluation metrics, which makes the studies more robust than the other review studies from the literature. However, none of these studies presented insights on the application domains and tasks that are facilitated with the developments of XAI. However, researchers from specific domains also surveyed the possibilities and challenges from their perspectives. The literature contains most of the works from the medical and health care domains [28–34]. However, there are review articles available in the literature from the domains of industry [35], software engineering [36], automotive [37], etc.

In the studies mentioned above, the authors reviewed and analysed the concepts and methodologies of XAI, challenges and possible actions to the solutions from the perspective of individual domains or without concerning the application domains and

tasks. However, to our knowledge, none of the studies exploited XAI methods considering different application domains and tasks as a whole. Moreover, a survey following an SLR guideline to review the methods and evaluation metrics for XAI to maintain a rigid objective throughout the study is still not present. Hence, in this article, an established guideline for SLR [38] was followed to gather and analyse the available methods of adding explainability to AI/ML models and the metrics of assessing the performance of the methods as well as the quality of the generated explanations. In addition, this survey study produced a general notion on the utilisation of XAI in different application domains based on the selected articles.

4. SLR Methodology

The methodology was designed according to the guidelines provided by Kitchenham and Charters for conducting an SLR [38]. The guidelines contain clear and robust steps for identifying and analysing potential research works intending to consider future research possibilities followed by the proper reporting of the SLR. The SLR methodology includes three stages: (i) *planning the review*; (ii) *conducting the review*; and (iii) *reporting the review*. The SLR methodology stages are briefly illustrated in Figure 4. The first two stages are broken down into major aspects and described in the following subsections, while the third stage, reporting the SLR, is self-explanatory.

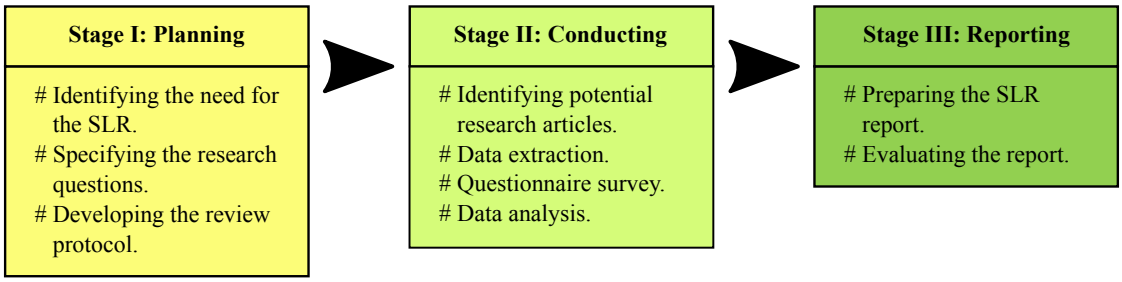


Figure 4. SLR methodology stages following the guidelines from Kitchenham and Charters [38].

4.1. Planning the SLR

The first stage involves creating a comprehensive research plan for the SLR. This stage includes identifying the need for conducting the SLR, outlining the research questions (RQs) and determining a detailed protocol for the research works to be accomplished.

4.1.1. Identifying the Need for Conducting the SLR

In a continuation of the discussion in Sections 1 and 3, with the increasing number of research works on XAI methodologies, the underlying knowledge becomes increasingly disorganised. However, very few secondary studies have been conducted solely to organise the profuse knowledge on the methodologies of XAI. In addition, no evidence of an SLR was found in the investigated bibliographic databases. Therefore, the need to conduct an SLR is stipulated to compile and analyse the primary publications on the methods and metrics of XAI and purposefully present an extensive and unbiased review.

4.1.2. Research Questions

Considering the urge to conduct an SLR of the exploited methods of providing explainability for AI/ML systems and their evaluations in different application domains and tasks, several RQs were formulated. Primarily, the questions were defined to investigate the prevailing approaches towards making AI/ML models explainable. This included the probe of explainable models by design, different structures of the generated explanation, and the significant application domains and tasks utilising the XAI methods. Furthermore, the means of validating the explainable models were also considered, followed by the open issues and future research directions. For convenience, the RQs for conducting this SLR are outlined as follows:

- *RQ1:* What are the application domains and tasks in which XAI is being explored and exploited?
  - *RQ1.1:* What are the XAI methods that have been used in the identified application domains and tasks?
  - *RQ1.2:* What are the different forms of providing explanations?
  - *RQ1.3:* What are the evaluation metrics for XAI methods used in different application domains and tasks?

4.1.3. SLR Protocol

The SLR protocol was designed to achieve the objective of this review by addressing the RQs outlined in Section 4.1.2. The protocol mainly contained the specification of each aspect of conducting the SLR. First, the identification of the potential bibliographic databases, the definition of the inclusion/exclusion criteria and quality assessment questions, and the selection of research articles are discussed elaborately in Section 4.2.1. In the second step, thorough scanning of each of the articles was performed, and relevant data were extracted and tabulated in a feature matrix. The feature set was defined from the knowledge of previous review studies mentioned in Section 3, motivated by the RQs outlined in Section 4.1.2. To support the feature extraction process, a survey was conducted in parallel which involved the corresponding/first authors of the selected articles. The survey responses were further used to obtain missing data, clarify any unclear data, and assess the extracted data quality. Upon completing feature extraction and the survey, an extensive analysis was performed to complement the defined RQs. Finally, to portray this SLR outcome, all the authors were involved in analysing the extracted features, and a detailed report was generated.

4.2. Conducting the SLR

This is the prime stage of an SLR. In this stage, most of the significant activities defined in the protocol were performed (Section 4.1.3), i.e., identifying potential research articles, conducting the author survey, extracting data and performing an extensive analysis.

4.2.1. Identifying Potential Research Articles

Inclusion and exclusion criteria were determined to identify potential research articles and are presented in Table 1. The criteria for inclusion in the SLR were peer-reviewed articles on XAI written in the English language and published in peer-reviewed international conference proceedings and journals. The criteria for exclusion from the SLR were articles that were related to the philosophy of XAI and articles that were not published in any peer-reviewed conference proceedings or journals. Throughout the article selection process, these inclusion and exclusion criteria were considered.

Table 1. Inclusion and exclusion criteria for the selection of research articles.

Inclusion Criteria	Exclusion Criteria
Describing the methods of XAI	Describing the methods in different contexts than AI
Peer reviewed	Describing the concept/philosophy of XAI
Published in conferences/journals	Preprints and duplicates
Published from 2018 to June 2021	Published in workshops
Written in English	Technical reports

To ensure the credibility of the selected articles, a checklist was designed. The list contained 10 questions that were adapted from the guidelines for conducting an SLR by Kitchenham and Charters and García-Holgado et al. [38,39]. Moreover, to facilitate the validation, the questions were categorised on the basis of design, conduct, analysis, and conclusion. The questions are outlined in Table 2.

Table 2. Questions for checking the validity of the selected articles.

Perspective	Quality Questions
Design	Are the aims clearly stated? If the study involves assessing a methodology, is the methodology clearly defined? Are the measures used in the study fully defined?
Conduct	Was outcome assessment blind to treatment group? If two methodologies are being compared, were they treated similarly within the study?
Analysis	Do the researchers explain the form of data (numbers, images, etc.)? Do the numbers add up across different tables and methodologies?
Conclusion	Are all study questions answered? How do results compare with previous reports? Do the researchers explain the consequences of any problems with the validity of their measures?

The process for identifying potential research articles included the identification, screening, eligibility, and sorting of the selected articles. A step-by-step flow diagram of this identification process is illustrated using the "Preferred Reporting Items for Systematic Reviews and Meta-Analyses" (PRISMA) diagram by Moher et al. [40] in Figure 5. The process started in June 2021. An initial search was conducted using Google Scholar (https://scholar.google.com/ accessed on 30 June 2021) with the keyword *explainable artificial intelligence* to assess the available sources of the research articles. The search results showed that most of the articles were extracted from SpringerLink (https://link.springer.com/ accessed on 30 June 2021), Scopus (https://www.scopus.com/ accessed on 30 June 2021), IEEE Xplore (https://ieeexplore.ieee.org/ accessed on 30 June 2021) and the ACM Digital Library (https://dl.acm.org/ accessed on 30 June 2021). Other similar sources were also present, but those were not considered since they primarily indexed data from the mentioned sources. Moreover, Google Scholar was not used for further article searches since it was observed that the results contained articles from diverse domains. In short, to narrow the search specifically to the AI domain, the mentioned databases were set to be the main sources of research articles for this review. Initially, 1709 articles were extracted from the bibliographic databases after searching with the keyword *explainable artificial intelligence*, as before. To focus this review on the recent research works, 113 articles were excluded because they were published before 2018. A total of 1596 articles were selected for screening, and after reviewing the titles or abstracts, more than half of the articles were excluded as they were not related to AI and XAI. From the 647 articles screened from the AI domain, 376 articles were excluded as they were duplicates or preprint versions of the articles. After evaluating the eligibility of the published articles, 277 articles were further considered, and 159 articles were excluded because they were notions or review articles. Specifically, a "yes" was provided for the selected articles for at least 7 out of the 10 quality questions mentioned in Table 2 following Dáu and Salim [41] and Genc-Nayebi and Abran [42]. Therefore, 118 articles were selected for a thorough review. During the process, 19 additional related articles were found from a complementary snowballing search [43], in simpler terms, a recursive reference search. Among the newly included articles, some were published prior to 2018 but were included in this study due to substantial contribution to the XAI domain. Finally, 137 articles were selected for the authors' survey, data/metrics extraction and analysis, among which 128 articles described different methodologies of XAI and 9 articles were solely related to the evaluation of the explanations or the methods to provide explanations.

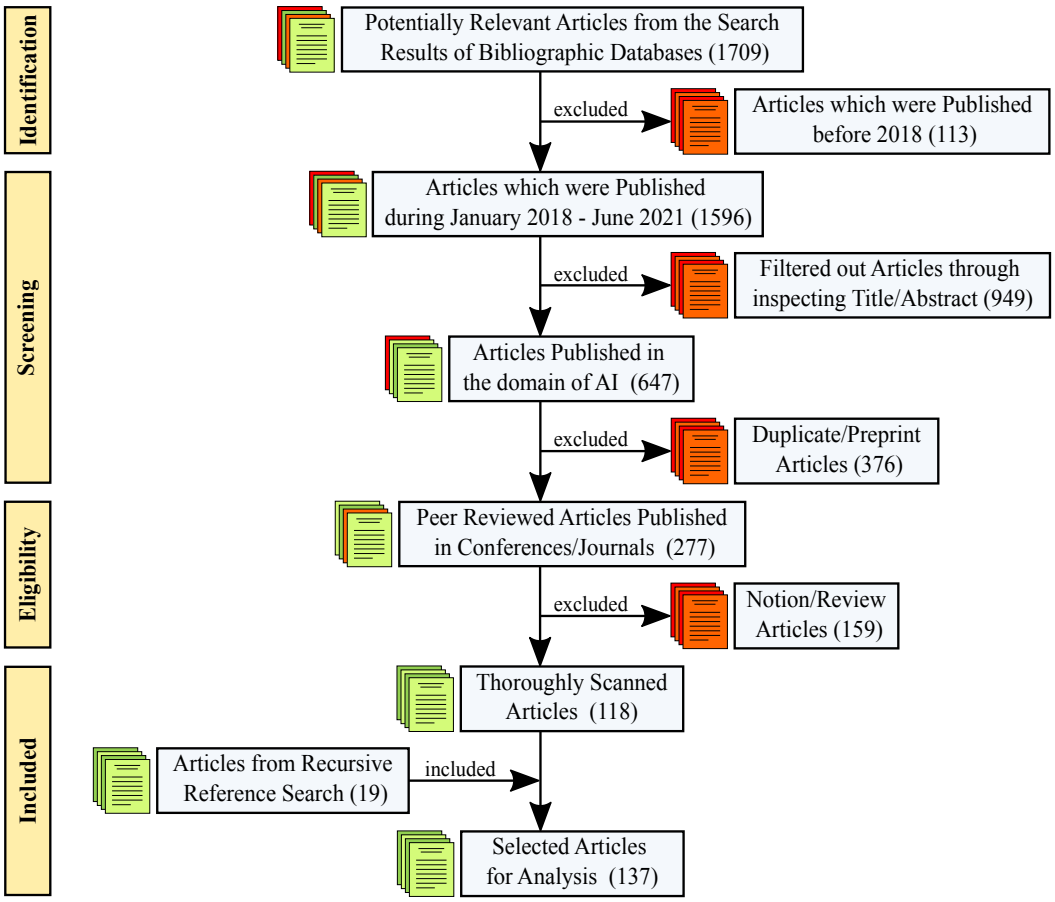


Figure 5. Flow diagram of the research article selection process adapted from the PRISMA flow chart by Moher et al. [40]. The number of articles obtained/included/excluded at different stages is presented in parentheses.

4.2.2. Data Collection

In this review study, the data collection was conducted in two parallel scenarios. Several features were extracted by reading the published article. Simultaneously, a questionnaire survey was distributed among the corresponding or first authors of the selected articles to gather their subjective remarks on the article and some features that were not clear from reading the articles. Each of the phases is elaborately described in the following paragraphs.

Feature Extraction

All the selected articles on the methodologies and evaluation of explainability were divided among the authors for thorough scanning to extract several features. The features were extracted from several viewpoints, namely *metadata*, *primary task*, *explainability*, *explanation*, and *evaluation*. The features extracted as metadata contained information regarding the dissemination of the selected study. Features from the viewpoint of the primary task were extracted to assess a general idea of the variety of AI/ML models that were deliberately used to perform classification or regression tasks prior to adding explanations to the models. The last three sets of features were extracted related to the concept of explainability, the explored or proposed method of making AI/ML models explainable and the evaluation of the methods and generated explanations, respectively. After extracting the features, a feature matrix was built to concentrate all the information for further analysis. The principal features from the feature matrix are concisely presented in Table 3.

Table 3. List of prominent features extracted from the selected articles.

Viewpoint	Feature	Description
Metadata	Source	Name of the conference/journal where the article was published.
	Keywords	Prominent words from the abstract and keywords sections that represents the concept of the article.
	Domain	The targeted domain for which the study was performed.
	Application	Specific application that was developed or enhanced.
Primary task	Data	The form of data that was used to develop a model, e.g., images, texts.
	Model	AI/ML model that was used for performing the primary task of classification/regression.
	Performance	The performance of the models for the defined tasks.
Explainability	Stage	The stage of generating explanation—during the training of a model (ante hoc) or after the training ends (post hoc).
	Scope	Whether the explanation is on the whole model, on a specific inference instance, i.e., global, local or both.
	Level	The level for which explanation is generated, i.e., feature, decision or both.
Explanation	Method	The procedure of generating explanations.
	Type	The form of explanations generated for the models or the outcomes.
Evaluation	Approach	The technique of evaluating the explanation and the method of generating explanation.
	Metrics	The criteria of measuring the quality of the explanations.

Questionnaire Survey

In parallel to the process of feature extraction through reading the articles, a questionnaire survey was conducted among the corresponding or first authors of the selected articles. The questionnaire was developed using Google Forms and distributed through separate emails to authors. The prime motivation behind the survey was to complement the feature extraction process by collecting authors' subjective remarks on their studies, curating the extracted features, and gathering specific information that was not present or unclear in the articles. The survey questionnaire contained queries on some of the features described in the previous section. In addition to that, queries on the experts' involvement, the use of third-party tools, potential stakeholders of this study etc., were also present in the questionnaire. In response to the invitation to the survey, approximately half of the invited authors submitted their remarks voluntarily, and these responses add value to the findings of this review.

4.2.3. Data Analysis

Following the completion of feature extraction from the selected articles and the questionnaire survey by the authors of the articles, the available data were analysed from multiple viewpoints, as presented in Table 3. From the metadata, sources were assessed to obtain an idea of the venues in which the works on XAI are published. Furthermore, the author-defined keywords and the abstracts were analysed by utilising natural language processing (NLP) techniques to assess the relevance of the articles to the XAI domain. Afterwards, the selected articles were clustered based on application domains and tasks to determine future research possibilities.

Before analysing the selected articles, clustering was performed in accordance with the primary tasks and input data mentioned in Section 2 and the method deployed to perform the primary task. Additionally, the proposed methods of explainability were clustered

based on scopes and stages. Finally, the evaluation methods were investigated. All the clustering and investigations performed in this review work were intended to summarise the methods of generating explanations along with the evaluation metrics and to present guidelines for the researchers devoted to exploiting the domain of XAI.

5. Results

The findings from the performed analysis of the selected articles and the questionnaire survey are presented concerning the viewpoints defined in Table 3. To facilitate a clear understanding, the subsections are titled with specific features, e.g., the results from the analysis on primary tasks are presented in separate sections. Again, the concepts of explainability are illustrated along with the methods to provide explanations in the corresponding sections.

5.1. Metadata

This section presents the results obtained from analysing the metadata extracted from the selected articles—primarily bibliometric data. Among the 137 selected articles, 83 were published in journals, and the rest were presented in conference proceedings. As per the inclusion criteria of this SLR, all the articles were peer reviewed prior to publication. In most of the articles, relevant keywords were the author-defined keywords, which facilitates the indexing of the article in bibliographic databases. The author-defined keywords were compared with the keywords extracted from the abstracts of the articles through a word cloud approach. Figure 6 illustrates the word cloud of the author-defined keywords and the prominent words extracted from the abstracts. The illustrated word clouds are expressed with varying font sizes. More often occurring words are presented in larger fonts [44] and different colours are used to differentiate words with the same frequencies.

Figure 7 presents the number of publications related to XAI from different countries of the world. Here, the countries were determined based on the affiliations of the first authors of the articles. The USA is the pioneer in the development of XAI topics and is still in the leading position. Similarly, several countries in Europe are following and have developed an increasing number of systems considering XAI. Based on the number of publications, Asian countries are apparently still quiescent in research and development on XAI.

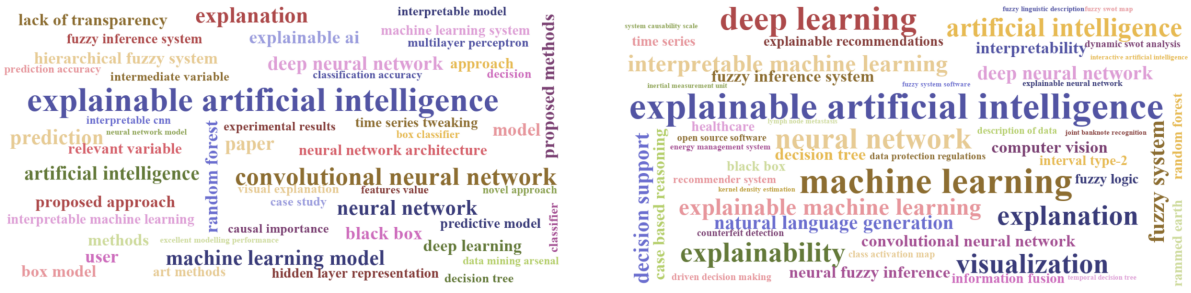


Figure 6. Word cloud of the (**a**) author-defined keywords and (**b**) keywords extracted from the abstracts through natural language processing. The font size is proportional to the number of occurrences of the terms and different colours are used to discriminate terms with equal font size. Both figures illustrate remarkable terms of XAI. However, the terms from keywords are more conceptual whereas the abstracts contained specific terms on the methods and tasks.

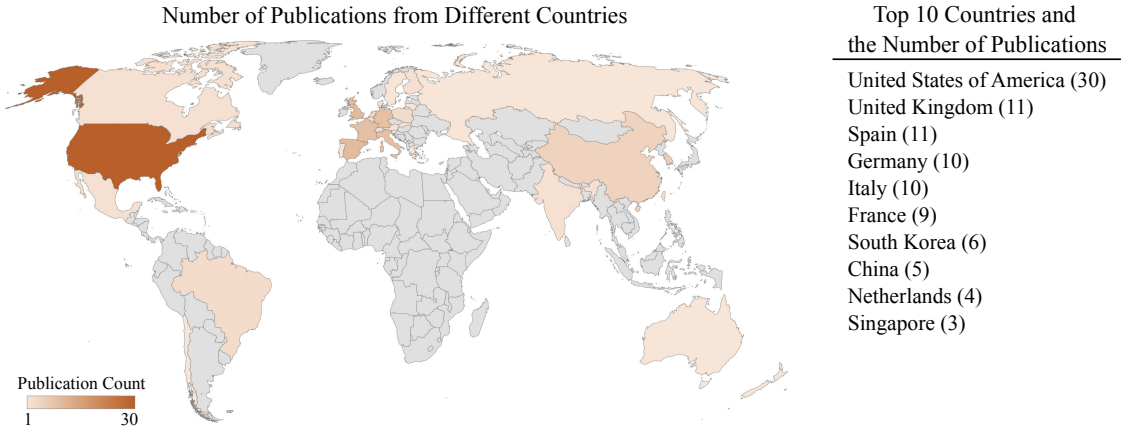


Figure 7. Number of publications proposing new methods of XAI from different countries of the world and the top 10 countries based on the publication count shown in parentheses, which is approximately 72% of the 137 articles selected for this SLR. The countries were determined from the affiliations of the first authors of the articles.

5.2. Application Domains and Tasks

To gain an idea of the research areas that have been enhanced with XAI, the application domains and tasks were scrutinised. The number of articles on different domains and tasks are illustrated in Figure 2. Among the selected articles, approximately 50% of the publications were domain-agnostic. Half of the remaining articles were published in the domain of healthcare. Other domains of interest among XAI researchers were found to be industry, transportation, the judicial system, entertainment, academia, etc. Table 4 presents the application domains and corresponding tasks on which the selected articles substantially contributed. It is evident from the content of the table that most of the published articles were not specific to one domain, and safety-critical domains, such as healthcare, industry, and transportation, received more attention from XAI researchers than domains, such as telecommunication and security. Some domains can be clustered together in a miscellaneous domain because of the small number of articles (as can be seen in Figure 2a). In the case of application tasks, most of the selected articles were published on supervised and decision-support tasks. A good number of works have been published on recommendation systems and systems developed on image processing tasks, e.g., object detection and facial recognition. Other noteworthy applications in the selected articles were predictive maintenance and anomaly detection. It was also observed that several articles presented works on supervised tasks, i.e., classification or prediction without specifying the application. Moreover, very few articles have been published on modelling gene relationships, business prediction, natural language processing, etc. Figure 8 presents a chord diagram [45] illustrating the distribution of the articles published from different application domains for various tasks. Most of the studies not specific to one domain were for decision support and image processing tasks.

Table 4. List of references to selected articles published on the methods of XAI from different application domains for the corresponding tasks.

Domain	Application/Task	Study Count	References
Domain agnostic	Supervised tasks	23	[46–68]
	Image processing	20	[25,69–87]
	Decision support	13	[7,12,23,88–97]
	Recommender system	4	[98–101]
	Anomaly detection	1	[102]
	Evaluation process	1	[103]
	Natural language processing	1	[104]
	Predictive maintenance	1	[105]
	Time series tweaking	1	[106]
Healthcare	Decision support	20	[107–126]
	Risk prediction	4	[127–130]
	Image processing	3	[131–133]
	Recommender system	2	[134,135]
	Anomaly detection	1	[136]
Industry	Predictive maintenance	5	[137–141]
	Business management	3	[142–144]
	Anomaly detection	1	[145]
	Modelling	1	[146]
Transportation	Image processing	4	[147–150]
	Assistance system	2	[151,152]
Academia	Evaluation	3	[153–155]
	Recommender system	1	[156]
Entertainment	Recommender system	3	[157–159]
Finance	Anomaly detection	1	[160]
	Business management	1	[161]
	Recommender system	1	[162]
Judicial system	Decision support	3	[163–165]
Genetics	Prediction	2	[166,167]
	Modelling gene relationship	1	[168]
Aviation	Automated manoeuvring	1	[169]
	Predictive maintenance	1	[170]
Architecture	Recommender system	1	[171]
Construction	Recommender system	1	[172]
Culture	Recommender system	1	[173]
Defence	Simulation	1	[174]
Geology	Recommender system	1	[175]
Network	Supervised tasks	1	[176]
Security	Facial recognition	1	[177]
Telecommunication	Goal-driven simulation	1	[178]

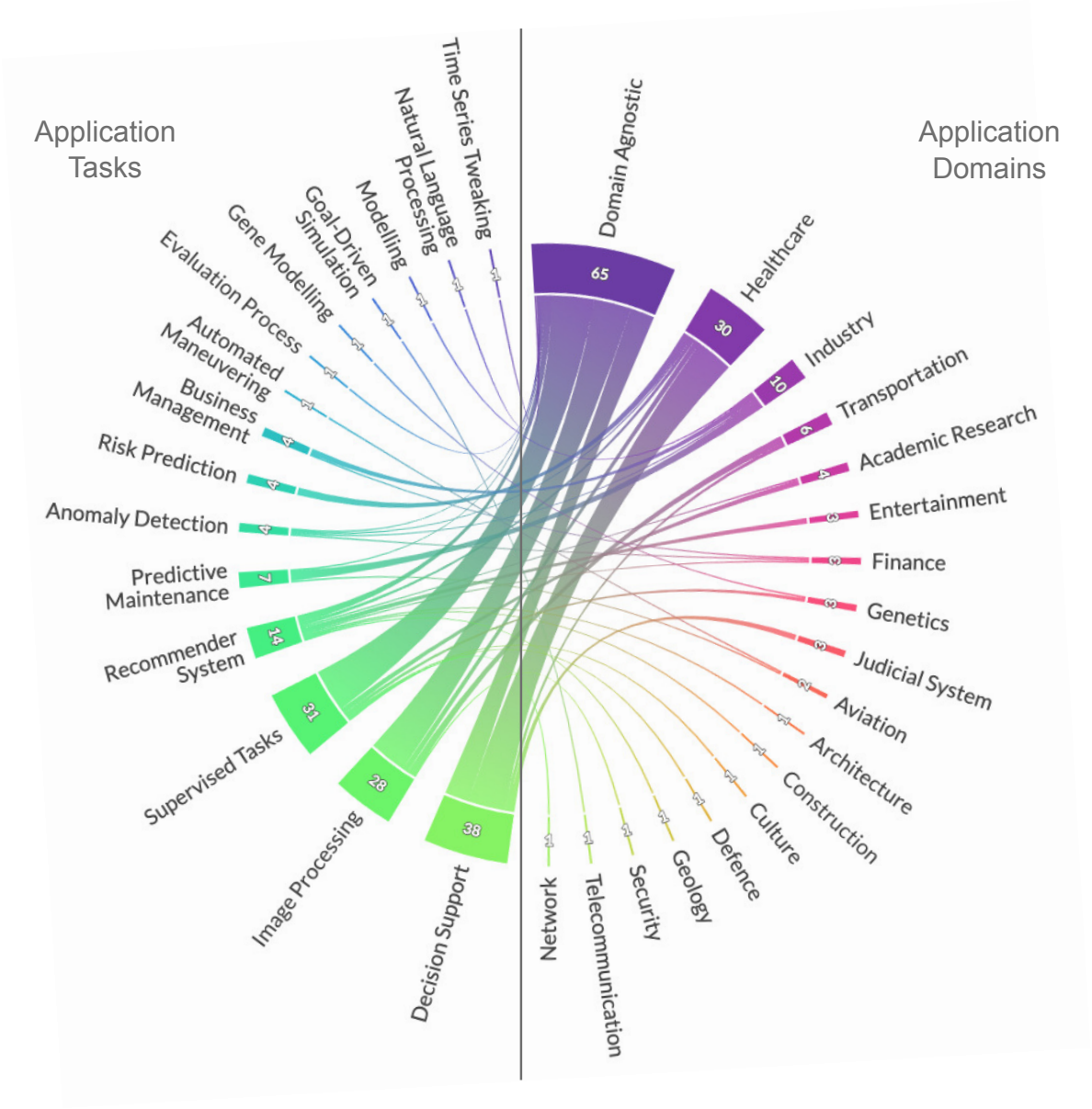


Figure 8. Chord diagram [45] presenting the number of selected articles published on the XAI methods and evaluation metrics from different application domains for the corresponding tasks.

5.3. Development of XAI in Different Application Domains

This section briefly describes the concepts of XAI stated in Section 2 from the perspective of different application domains. Figure 9 illustrates the number of articles selected from different application domains and further clustered the number of articles in terms of AI/ML model types, stage, scope, and form of explanations. In the following subsections, shreds of evidence of linkage between the application domains and concepts of XAI are presented.

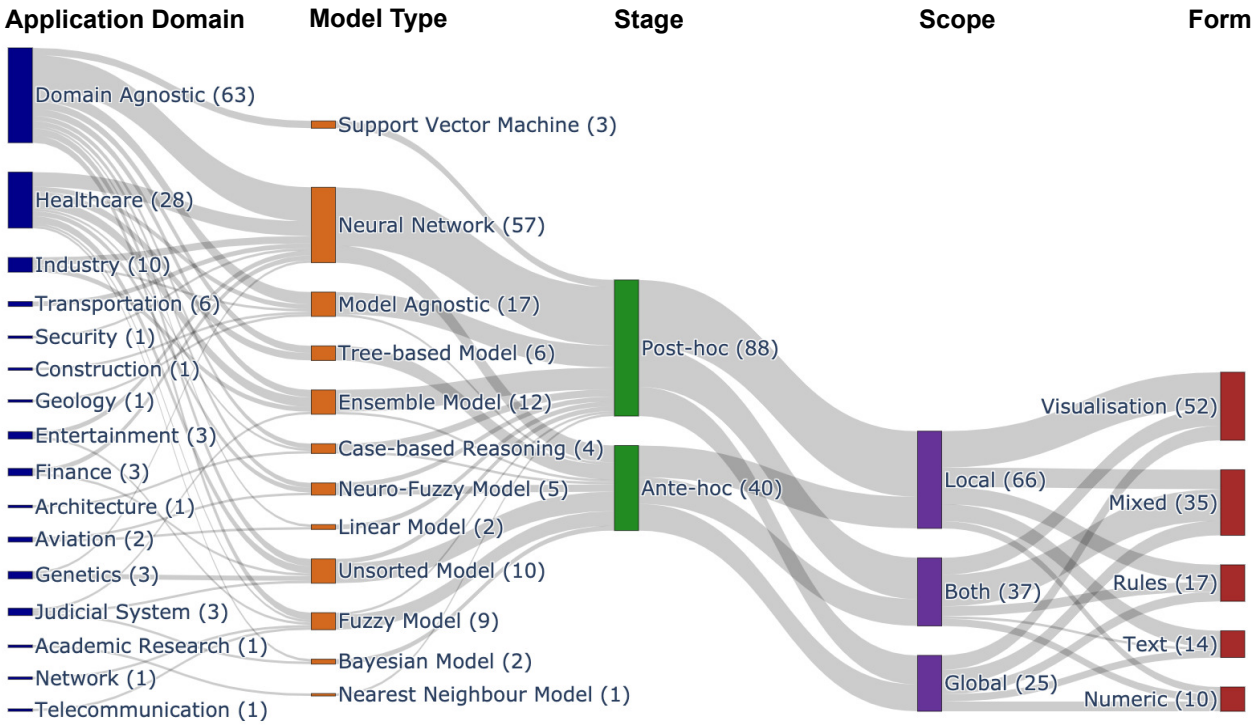


Figure 9. Number of the selected articles published from different application domains and clustered on the basis of AI/ML model type, stage, scope, and form of explanations. The number of articles with each of the properties is given in parentheses.

5.3.1. Input Data

The selected articles presented diverse XAI models that can train on different forms of input data corresponding to the primary tasks and application domain. Figure 10 illustrates the use of different input data types with a Venn diagram depicting the number of articles for each type. The basic types of input data used in the proposed methods were vectors containing numbers, images, and texts. However, the use of sensor signals and graphs were also observed but in low numbers. Some of the works considered diverse forms of data altogether, such as the works of Ribeiro et al. [96], Alonso et al. [52] and Lundberg et al. [59], who proposed methods that can deal with the input types, images, texts, and vectors. Another proposed method was developed to learn on graphs and vectors containing numbers [175]. In addition to the mentioned forms of input data, a specialised form of input data was observed, namely the logic scoring preference (LSP) criteria [103], which was later counted as numbers due to apparent similarity.

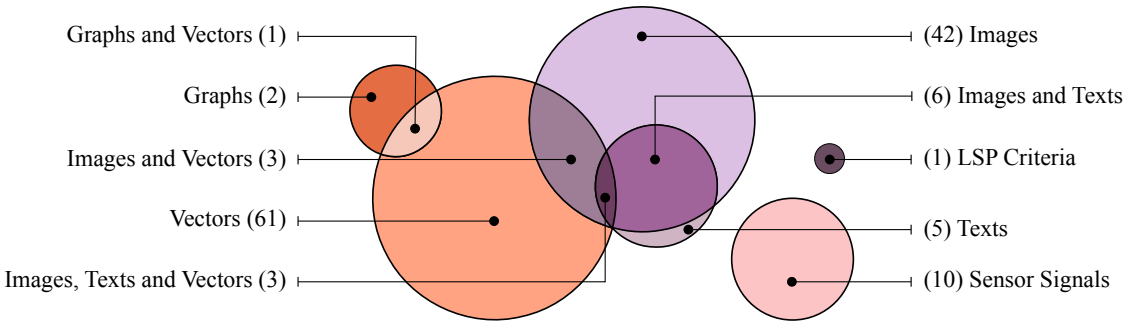


Figure 10. Venn diagram with the number of articles using different forms of data to assess the functional validity of the proposed XAI methodologies. The sizes of the circles are approximately proportional to the number of articles (shown within parentheses) that were observed in this review study.

5.3.2. Models for Primary Tasks

The majority of the applications built on the concepts of AI perform two basic types of tasks, i.e., supervised (classification and regression) and unsupervised (clustering) tasks which have undoubtedly remained unchanged in the XAI domain. The authors of the selected articles used different established AI/ML models depending on the tasks. The methods were clustered based on the basic type of the models, specifically, neural network (NN), ensemble model (EM), Bayesian model (BM), fuzzy model (FM), tree-based model (TM), linear model (LM), nearest neighbour model (NNM), support vector machine (SVM), neuro-fuzzy model (NFM), and case-based reasoning (CBR). Works related to these models were clustered on the basis of their types and are presented in Table 5. Moreover, the table contains the names of different variants of the AI/ML models references to the articles featuring the models, and the number of studies performed. It was observed that neural network-based models were exploited in most of the studies (63) from the selected articles. The second-highest number of studies (21) utilised the ensemble techniques for performing the primary supervised or unsupervised tasks. Based on the increased interest of researchers in neural networks and ensemble techniques, it can be inevitably assumed that these models were chosen to incorporate explainability because of their wide acceptability over various domains in terms of their performances. In addition to the renowned algorithms, there are some other algorithms, such as probabilistic soft logic (PSL) [100], LSP [103], sequential rule mining (SRM) [168], preference learning [113], Cartesian genetic programming (CGP) [122], Predomics [129], and TriRank [162]. The acronyms of the model types are further referenced in Table 6 to indicate their relation to the core AI/ML models.

Throughout this study, it was evident that most of the research works were domain-agnostic. For specific domains, healthcare, industry, and transportation were revealed to be more exploited than other domains. In these domains, as stated above, diverse forms of neural networks had been invoked to perform different tasks (see Figure 9) followed by other types of models, as listed in Table 5. The numbers associated with different model types stated in Figure 9 and Table 5 varied because the illustration presents the number of articles and the table lists the number of variations of the models. It was observed that in some articles, the authors presented theirs using different models of similar types.

Table 5. Different models used to solve the primary task of classification or regression and their study count.

Model Types	Models	Count	References
Neural Networks (NNs)	ApparentFlow-net; Convolutional Neural Network (CNN); Deep Neural Network (DNN); Deep Reinforcement Learning (DRL); Explainable Deep Neural Network (xDNN); Explainable Neural Network (ExNN); Global–Local Capsule Networks (GLCapsNet); GoogleLeNet; Gramian Angular Summation Field CNN (GASF-CNN); Hopfield Neural Networks (HNN); Knowledge-Aware Path Recurrent Network; Knowledge-Shot Learning (KSL); LeNet-5; Locally Guided Neural Networks (LGNN); Long/Short-Term Memory (LSTM); LVRV-net; MatConvNet; Multilayer Perceptrons (MLP); Nilpotent Neural Network (NNN); Recurrent Neural Network (RNN); Region-Based CNN (RCNN); RestNet; ROI-Net; Temporal Convolutional Netwrok (TCN); VGG-19; YOLO	63	[7,23,25,49,51,58,60,66,68–71,73,74,76,78–81,83–87,89,93,98,101,108–111,115,117,119,120,124,128,132,133,135,137,140,141,143,144,147,149–152,156,158–161,170,172,177]

Table 5. *Cont.*

Model Types	Models	Count	References
Ensemble Models (EMs)	Adaptive Boosting (AdaBoost); Explainable Unsupervised Decision Trees (eUD3.5); eXtreme Gradient Boosting (XGBoost); Gradient Boosting Machines (GBM); Isolation Forest (IF); Random Forest (RF); Random Shapelet Forest (RSF)	21	[23,47–50,55,63,65,102,106,111,112,114,123,130,139,142,145,163,170,172]
Tree-Based Models (TB)	Classification and Regression Tree (CART); Conditional Inference Tree (CTree); Decision Tree (DT); Fast and Frugal Trees (FFTs); Fuzzy Hoeffding Decision Tree (FHDT); J48; One-Class Tree (OCTree); Multi-Operator Temporal Decision Tree (MTDT); Recursive Partitioning and Regression Trees (RPART)	10	[12,52,54,88,105,108,127,128,136,172]
Fuzzy Models (FMs)	Big Bang–Big Crunch Interval Type-2 Fuzzy Logic System (BB-BC IT2FLS); Constrained Interval Type-2 Fuzzy System (CIT2FS); Cumulative Fuzzy Class Membership Criterion (CFCMC); Fuzzy Unordered Rule Induction Algorithm (FURIA); Hierarchical Fuzzy Systems (HFS); Multi-Objective Evolutionary Fuzzy Classifiers (MOEFC); Wang–Mendal Algorithm of Fuzzy Rule Generation (WM Algorithm)	09	[52,61,64,95,104,134,157,178]
Support Vector Machines (SVMs)	SVM with Linear and Radial Basis Function (RBF) Kernels	08	[12,23,47–49,63,87,128,139,170,176]
Unsorted Models (UMs)	Cartesian Genetic Programming (CGP); Computational Argumentation (CA); Logic Scoring of Preferences (LSP); Preference Learning (PL); Probabilistic Soft Logic (PSL); Sequential Rule Mining (SRM); TriRank	07	[100,103,113,122,162,164,168]
Linear Models (LMs)	Linear Discriminant Analysis (LDA); Logistic Regression (LgR); Linear Regression (LnR)	06	[12,99,124,128,172]
Nearest Neighbours Models (NNMs)	k-Nearest Neighbours (kNN); Distance-Weighted kNN (WkNN)	06	[12,106,128,139,156,170]
Neuro-Fuzzy Models (NFMs)	Adaptive Network-Based Fuzzy Inference System (ANFIS); Improved Choquet Integral Multilayer Perceptron (iChIMP); LeNet with Fuzzy Classifier; Mamdani Fuzzy Model; Sugeno-Type Fuzzy Inference System; Zero-Order Autonomous Learning Multiple-Model (ALMMo-0*)	05	[82,90,116,118,169]
Case-Based Reasoning (CBR)	CBR-kNN; CBR-WkNN; CBR-PRVC (Pattern Recognition, Validation and Contextualisation) Methodology	04	[92,97,121,171]
Bayesian Models (BM)	Bayesian Network (BN); Bayesian Rule List (BRL); Gaussian Naive Bayes Classifier/Regressor (GNBC/GNBR)	03	[126,139,165]

5.3.3. Methods for Explainability

The available methods for adding explainability to the existing and proposed AI/ML models were initially clustered on the basis of three properties: (i) the stage of generating an explanation; (ii) the scope of the explanation; and (iii) the form of the explanation. Figure 11

illustrates the number of articles presenting research works concerning each of the properties. The summary of the clustering is represented in Table 6 where model-specific methods are cross-referenced to the model types described in Section 5.3.2. A good number of model-agnostic (MA) methods were also deployed to provide explainability in the selected articles of this review, such as Anchors [96], Explain Like I'm Five (ELI5) [139], Local Interpretable Model-Agnostic Explanations (LIME) [12], and Model Agnostic Supervised Local Explanations (MAPLE) [65]. LIME was modified and proposed as SurvLIME by Kovalev et al. [179]. Afterwards, the authors incorporated well-known Kolmogorov–Smirnov bounds to SurvLIME and proposed SurvLIME-KS [57]. The authors also utilised feature importance to generate numeric explanations in several research works [67,128,144,172]. The Shapley Additive Explanations (SHAP) was proposed by Lundberg and Lee [84], and it was later used by several authors to generate mixed explanations containing numbers, texts, and visualisations [123,150]. However, another variant of SHAP, Deep-SHAP, was proposed to explicitly explain deep learning models. Two very recent studies proposed Cluster-Aided Space Transformation for Local Explanation (CASTLE) [47] and Pivot-Aided Space Transformation for Local Explanation (PASTLE) [48]. The authors claimed that a higher quality of local explanations can be generated with these methods than with the prevailing methods for unsupervised and supervised tasks, respectively.

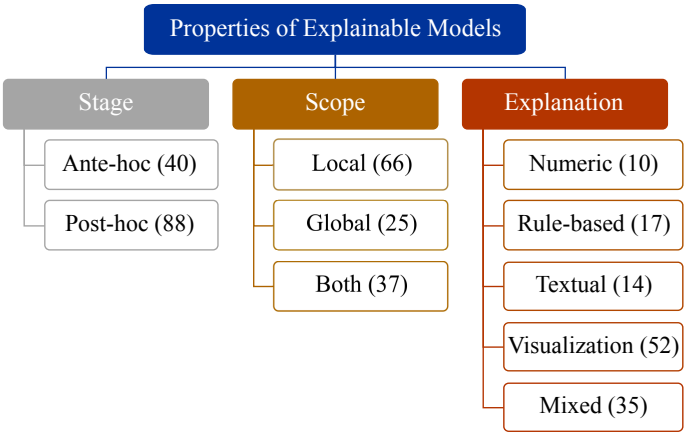


Figure 11. Distribution of the selected articles based on the stage, scope, and form of explanations. The number of articles with each of the properties is given in parentheses.

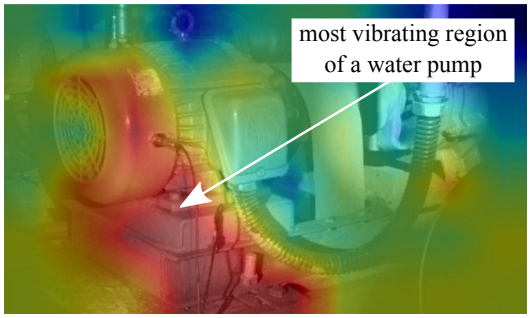
In terms of application domains, post hoc techniques are more developed for producing explanations at the local scope. One can see in the illustration of Figure 9 that the majority of the post hoc techniques were developed for complex models such as neural networks and ensemble models. On the other hand, most of the ante hoc techniques are associated with fuzzy and tree-based models across all the application domains.

5.3.4. Forms of Explanation

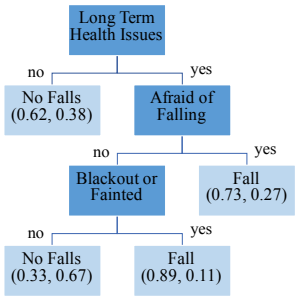
This section presents the different forms of explanations that have been added to different AI/ML models. From the selected articles, it was observed that mostly four different forms of explanations were generated to explain the decisions of the models as well as the process of deducing a decision. The forms of explanations are numeric, rules, textual, and visualisation. Figure 12 illustrates the basic forms of explanations. In some of the works, the authors used these forms in a combined fashion to make the explanation more understandable and user friendly. All the forms of explanation are discussed along with the references to key works with the corresponding forms in the subsequent paragraphs.

Text record: *"Where is Mile High Stadium?"*			
Prediction: *LOC:other*			
Explanation:			
Class: *LOC:other* Score: *2.555*		Class: *NUM:count* *Score: 0.666*	
Itemset	Confidence	Itemset	Confidence
<where>	0.888	<mile>	0.666
<stadium>	0.666		
<where>, <stadium>	1.0		

(a)



(b)



(c)

"if there were 9 more bare nucleus, the patient would be classified as malignant RATHER THAN benign"

"The message is classified as spam RATHER THAN ham because the word 'credit' is used twice as frequent as that of ham message"

(d)

Figure 12. Different forms of explanations: (**a**) numeric explanation of remaining life estimation in industry appliances [49]; (**b**) visual explanation for fault diagnosis of industrial equipment by Sun et al. [140]; (**c**) example of rule-based explanation in the form of a tree [127]; and (**d**) explanation text generated with GRACE, proposed by Le et al. [58].

Numeric Explanations

Numeric explanations are mostly generated by the models by measuring the contribution of the input variables for the model's outcome. The contribution is represented by various measures, such as the confidence measures of features [49] illustrated in Figure 12a, saliency, causal importance [68], feature importance [144,172], and mutual importance [99]. Islam et al. improvised the MLP with the Choquet integral to add numeric explanations within both the local and global scope [90]. Sarathy et al. computed and compared the quadratic mean among the instances to generate the decision with explanations [177]. Carletti et al. used depth-based isolation forest feature importance (DIFFI) to support the decisions from depth-based isolation forests (IFs) in anomaly detection for industrial applications [145], and the FDE measure was developed to add precise explainability for failure diagnosis in automated industries [170]. Moreover, several model-agnostic tools generate numeric explanations, e.g., Anchors [96], ELI5, LIME [139], SHAP [150], and LORE [123]. Moreover, Table 6 contains additional examples of numeric explanations, and the methods are clustered on the basis of stage and the scope of explanations. However, the numeric explanations demand high expertise in the corresponding domains as they are associated with the features. This assumption supports the low number of studies on numeric explanations, as shown in Figure 9.

Rule-Based Explanations

Rule-based explanations illustrate a model's decision-making process in the form of a tree or list. Figure 12c demonstrates an example of a rule-based explanation. Largely, the models producing rule-based explanations generate explanations with a global scope, i.e., of the whole model. De et al. proposed the existing TREPAN decision tree as a surrogate model with an FFNN to generate rules depicting the flow of information within the neural network [89]. Rutkowski et al. used the Wang–Mendal (WM) algorithm to generate fuzzy rules to support recommendations with explanations [157]. A novel neuro-fuzzy system,

ALMMo-0*, was proposed by Soares et al. [116]. In addition, model-specific methods have been proposed to generate rule-based explanations such as eUD3.5, an explainable version of UD3.5 [102] and Ada-WHIPS to support the AdaBoost ensemble method [112]. More methods generating rule-based explanations are listed in Table 6. The rule-based explanations are much simpler in nature than the numeric explanations that facilitate this type of explanation in supporting recommendation systems developed for general users from domains such as entertainment and finance.

Textual Explanations

The use of textual explanations is found to be least common among all forms of explanations due to their higher computational complexity which requires natural language processing. The textual explanations are mostly generated at the local scope, i.e., for an individual decision. In notable works, textual explanations were generated using counterfactual sets [7,55], template-based natural language generation [164], etc. Weber et al. proposed textual CBR (TCBR) utilising patterns of input-to-output relations in order to recommend citations for academic researchers through textual explanations [156]. Unlike TCBR, interpretable confidence measures were used by Waa et al. with CBR to generate textual explanations [92]. Le et al. proposed GRACE which can generate intuitive textual explanations along with the decision [58]. The textual explanations generated with GRACE were revealed to be more understandable by humans in synthetic and real experiments. Moreover, textual explanations are found to be generated at the local scope (see Figure 9) and these explanations are associated with academic research, judicial systems, etc. Table 6 lists several other proposed methods to generate textual explanations.

Visual Explanations

The most common form of explanation was found to be visualisations, as shown in Table 6. With respect to the stage of adding explanations, in the majority of the cases, visual explanations in both the local and global scopes were generated using post hoc techniques and the research studies were carried out as domain-agnostic and from the healthcare domain (see Figure 9). Common visualisation techniques are class activation maps (CAM) [140,141] and attention maps [79,152]. CAM was further extended with gradient weights, and Grad-CAM was proposed by Selvaraju et al. [80]. Brunese et al. used Grad-CAM to detect COVID-19 infection based on X-rays [109]. Han and Kim adopted another form pGrad-CAM to provide an explanation for banknote fraud detection [160]. Heatmaps of salient pixels were used by Graziani et al. as a complement to the concept-based explanation. They proposed a framework of concept attribution for deep learning to quantify the contribution of features of interest to the deep network's decision making [132]. In addition, several explanation techniques were proposed with attribution-based visualisations, such as Multi-Operator Temporal Decision Trees (MTDTs) [105], Layerwise Relevance Propagation (LRP) [87], Selective LRP (SLRP) [70], etc. The Rainbow Boxes-Inspired Algorithm (RBIA) was extensively used by Lamy et al. in different decision support tasks within the healthcare domain [113,121]. Specialised methodologies have also been developed by researchers from diverse domains to add visual explanations to the outcomes of different AI/ML models such as iNNvestigate [135], non-negative matrix factorisation (NMF) [83], candlestick plots [161], and sequential rule mining (SRM) [168]. In addition to the methodologies mentioned above, Table 6 contains additional methods to add visual explanations to different types of AI/ML models.

Table 6. Methods for explainability, stage (*Ah*: ante hoc; *Ph*: post hoc) and scope (*L*: local; *G*: global) of explainability, forms of explanations (*N*: numeric; *R*: rules; *T*: textual; *V*: visual) and the type of models used for performing the primary tasks (refer to Table 5 for the elaborations of the model types).

Methods for Explainability	References	Stage		Scope		Form				Models for Primary Task
		Ah	*Ph*	*L*	*G*	*N*	*R*	*T*	*V*	
Ada-WHIPS	[112]		✓	✓			✓			EM
ALMMo-0*	[116]	✓		✓	✓		✓			NFM
Anchors	[96]		✓	✓	✓	✓	✓	✓	✓	MA
ANFIS	[66,118,137,169]		✓	✓	✓	✓	✓	✓	✓	FM; GA; NN;
ApparentFlow-net	[124]	✓		✓					✓	NN
Attention Maps	[79,149,151,152]		✓	✓					✓	NN
BB-BC IT2FLS	[178]	✓		✓		✓	✓		✓	FM
BEN	[69]		✓	✓					✓	NN
BN	[71,165]	✓		✓				✓	✓	BM
BRL	[126]	✓			✓		✓			BM
CAM	[140,141]		✓		✓				✓	NN
Candlestick Plots	[161]		✓		✓				✓	NN
CART	[127]	✓		✓			✓			TM
CASTLE	[47]		✓	✓	✓	✓			✓	MA
Causal Importance	[68]		✓		✓	✓				NN
CFCMC	[61]	✓			✓			✓		FM
CGP	[122]	✓			✓	✓				UM
CIE	[49]		✓	✓	✓	✓		✓		EM; NN; SVM
CIT2FS	[134]	✓		✓				✓	✓	FM
Concept Attribution	[132]		✓	✓	✓				✓	NN
Counterfactual Sets	[7,55]		✓	✓				✓		EM; NN
CTree	[108,127]	✓		✓			✓			TM
DeconvNet	[83]		✓	✓	✓				✓	NN
Decision Tree	[54,75]		✓	✓	✓		✓			NN; TM
Deep-SHAP	[143]		✓	✓	✓	✓			✓	MA
DTD	[85]		✓	✓					✓	NN
DIFFI	[145]		✓		✓	✓				EM
ELI5	[139,142]		✓	✓	✓	✓			✓	MA
Encoder–Decoder	[133]	✓		✓					✓	NN
eUD3.5	[102]	✓		✓	✓		✓			EM
ExNN	[60]	✓		✓	✓				✓	NN
FACE	[78]		✓	✓					✓	NN
FDE	[170]		✓	✓	✓	✓				EM; NN; NNM; SVM
Feature Importance	[67,128,144,172]		✓	✓	✓	✓		✓	✓	MA
Feature Pattern	[163]		✓		✓		✓			EM
FFT	[127]	✓		✓			✓			TM
FINGRAM	[88]	✓		✓					✓	TM
FormuCaseViz	[97]		✓	✓					✓	CBR
FURIA	[52]	✓		✓			✓			FM
Fuzzy LeNet	[82]	✓		✓					✓	FM
Fuzzy Relations	[64,104]	✓		✓				✓		FM
gbt-HIPS	[50]	✓		✓	✓		✓			EM
Generation	[159]		✓	✓				✓		NN
GLAS	[77]		✓	✓					✓	MA
GRACE	[58]		✓	✓				✓		NN

Table 6. *Cont.*

Methods for Explainability	References	Stage		Scope		Form				Models for Primary Task
		Ah	*Ph*	*L*	*G*	*N*	*R*	*T*	*V*	
Grad-CAM	[53,72,80,109,117,137,146]		✓	✓					✓	NN
Growing Spheres	[63]		✓	✓	✓			✓		EM; SVM
HFS	[95]	✓		✓	✓				✓	FM
iChIMP	[90]	✓		✓	✓	✓				NFM
ICM	[92]		✓	✓				✓		CBR
iNNvestigate	[135]		✓	✓	✓				✓	NN
Interpretable Filters	[25]		✓	✓					✓	NN
J48	[52,127,166]	✓		✓			✓	✓		TM
Knowledge Graph	[158]		✓	✓	✓				✓	NN
KSL	[110]	✓	✓			✓	✓			NN
LEWIS	[46]	✓		✓	✓	✓	✓			MA
LGNN	[81]	✓			✓				✓	NN
LIME	[12,111,119,139,146]		✓	✓	✓	✓			✓	MA
LORE	[23,123]		✓	✓			✓			EM; NN; SVM
LPS	[76]		✓		✓				✓	NN
LRP	[87]		✓	✓					✓	NN; SVM
LRCN	[86]		✓		✓				✓	NN;
LSP	[103]	✓		✓	✓	✓				UM
MAPLE	[65]		✓	✓	✓	✓			✓	MA
MTDT	[105]	✓			✓				✓	TM
Mutual Importance	[99]		✓	✓		✓				LM
MWC, MWP	[93]		✓		✓			✓	✓	NN
Nilpotent Logic Operators	[98]	✓		✓	✓		✓		✓	NN
NLG	[51]		✓	✓	✓			✓	✓	NN
NMF	[25]		✓	✓					✓	NN
OC-Tree	[136]	✓			✓		✓			TM
Ontological Perturbation	[115]		✓	✓			✓			NN
PAES-RCS	[176]		✓	✓			✓			FM
PASTLE	[48]		✓	✓	✓	✓			✓	MA
pGrad-CAM	[160]		✓	✓					✓	NN
Prescience	[130]		✓	✓					✓	EM
PRVC	[171]	✓		✓	✓	✓		✓	✓	CBR
PSL	[100]	✓		✓				✓	✓	UM
QMC	[177]		✓		✓	✓				NN
QSAR	[167]	✓		✓		✓				NN
RAVA	[175]	✓			✓				✓	MA
RBIA	[113,121]		✓	✓	✓				✓	CBR
RetainVis	[120]		✓	✓	✓	✓		✓	✓	NN
RISE	[148]		✓	✓					✓	NN
RPART	[108]	✓		✓			✓			TM
RuleMatrix	[94]		✓		✓				✓	MA
Saliency	[68]		✓		✓	✓				NN
SHAP	[84,107,114,123,138,139,150]		✓	✓	✓	✓			✓	MA
Shapelet Tweaking	[106]		✓	✓					✓	EM
SLRP	[70]		✓	✓					✓	NN
SRM	[168]	✓		✓	✓				✓	UM
SurvLIME-KS	[57]		✓	✓	✓	✓			✓	MA
TCBR	[156]		✓	✓				✓		CBR

Table 6. *Cont.*

Methods for Explainability	References	Stage		Scope		Form				Models for Primary Task
		Ah	*Ph*	*L*	*G*	*N*	*R*	*T*	*V*	
Template-Based Natural Language Generation	[164]	✓		✓				✓		UM
Time-Varying Neighbourhood	[101]		✓	✓		✓			✓	NN
TreeExplainer	[84]		✓	✓	✓	✓			✓	MA
TREPAN	[89]		✓	✓			✓			NN
Tripartite Graph	[162]	✓		✓		✓			✓	UM
WM Algorithm	[157]	✓			✓		✓			FM
xDNN	[116]	✓			✓			✓		NN
XRAI	[147]		✓	✓					✓	NN

5.3.5. Evaluation of Explainability

The development of methodologies or definitions of metrics to evaluate the explanation generation techniques as well as to assess the quality of the generated explanations is comparatively lower than the extreme increase in research works devoted to exploring new methodologies of XAI. In this study, only nine articles among the selected articles were found to be fully intended for the evaluation OF and metrics for XAI. However, all the articles proposing new methods to add explainability considered one of the three techniques to assess their explainable model or the explanations generated by the models. These techniques were (i) user studies; (ii) synthetic experiments; and (iii) real experiments. The number of studies adopting each of the techniques Are illustrated in Figure 13. It was observed that most of the studies invoked user studies and synthetic experiments as standalone methods for evaluating the proposed explainable systems. Very few studies only used real experiments to evaluate their proposed systems. However, several studies conducted a combination of the user studies, real and synthetic experiments in the evaluation process as illustrated in the UpSet plot in Figure 13. User studies were mostly performed to evaluate the quality of the generated explanation in the form of case studies and questionnaire surveys. Generally, these cases are formulated by the researchers combining a real or synthetic scenario that is associated with some prediction/classification output and its explanation in any of the forms presented in Section 5.3.4. The surveys were observed to be conducted among the respective domain experts. They had to answer questions on the understandability and quality of the explanations from the presented case studies. To facilitate the user studies, Holzinger et al. proposed the System Causability Scale (SCS) to measure the quality of explanations [56]. In simpler terms, the SCS resembles the widely known Likert scale [180]. In earlier work, Chander and Srinivasan introduced the notion of the *cognitive value* of an explanation and related its function in generating significant explanations within a given setting [62]. Lage et al. proposed the methodology of a user study to measure the human-interpretability of logic-based explanations [125]. The prime metrics were the response time for understanding, the accuracy of understanding, and the subjective satisfaction of the users. Ribeiro et al. explicitly conducted a simulated user experiment to address the following questions [12]: (1) Aare the explanations faithful to the model? (2) Can the explanations aid users to ascertain trust in predictions? and (3) Are the explanations useful for evaluating the model as a whole? They also involved human subjects in evaluating the explanations generated by LIME and SP-LIME within the following situations [12]: (1) whether users can choose a better classifier in terms of generalisation; (2) whether the users can perform feature engineering to improve the model; and (3) whether the users are capable of pointing out the irregularities of a classifier by observing the explanations.

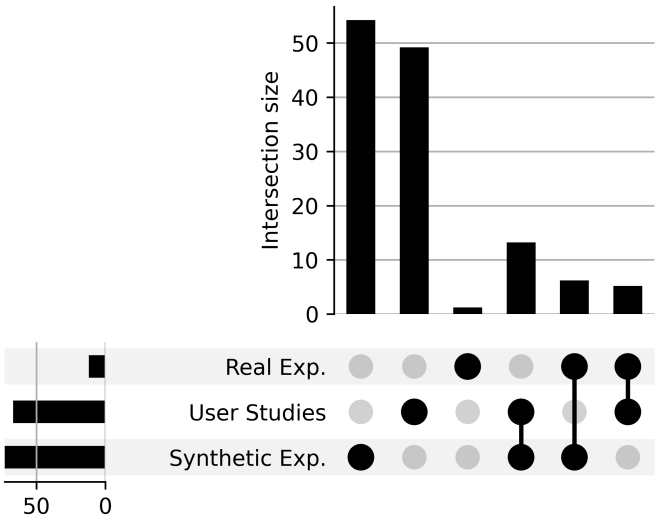


Figure 13. UpSet plot presenting the distribution of different methods of evaluating the explainable systems. The vertical bars in the bottom-left represents the number of studies conducting each of the methods. The single and connected black circles represents the combination of the evaluation methods and the horizontal bars illustrate their number of studies.

Different types of experiments with real and synthetic data were performed to quantify various metrics for the generated explanations to evaluate the quality of the explanations. Vilone and Longo proposed two types of evaluation methods for assessing the quality of the explanations; objective and human-centred [27]. Human-centred methods are mostly performed through user studies as discussed earlier. The prominent objective measures are briefly stated here. Guidotti et al. used fidelity, l-fidelity, and hit scores and proposed the use of the Jaccard measure of stability, the number of falsified conditions in counterfactual rules, the rate of the agreement of black-box and counterfactual decisions for counterfactual instances, F1-score of agreement of black box and counterfactual decisions, etc. [23]. In another work, *stability* was proposed as an objective function that acts as an inhibitor to include too many terms in the textual explanations [112]. To evaluate the visual explanations, Bach et al. proposed a pixel-flipping method that enables users to discriminate between two heatmaps [87]. Moreover, sentence evaluation metrics, such as METEOR and CIDEr were used to evaluate textual explanations associated with visualisations [86]. Samek et al. proposed the Area over the MoRF (Most Relevant First) Curve (AOPC)) to measure the impact on classification performance when generating a visual explanation [181]. In the proposition, the authors illustrated that a large AOPC value provides a good measure for a very informative heatmap. AOPC can assess the amount of information present in a visual explanation but it lacks in terms of being able to assess the quality of the understandability of the users. In another study, Rio-Torto et al. proposed Percentage of Meaningful Pixels Outside the Mask (POMPOM) as another measurable criterion of explanation quality [133]. POMPOM is defined as the ratio between the number of meaningful pixels outside the region of interest and the total number of pixels in the image. The authors have also conducted a comparative study with AOPC and POMPOM. They concluded that POMPOM generates superior results for the supervised approach whereas AOPC has the upper hand for the unsupervised approach. Significantly, Sokol and Flanch provided a comprehensive and representative taxonomy and associated descriptors in the form of a fact sheet with five dimensions that can help researchers develop and evaluate new explainability approaches [155].

The associations among the evaluation methods and different application domains and applications are illustrated in Figure 14. It can be easily observed that synthetic experiments and user studies were mostly used to evaluate proposed explainable systems from the domains of healthcare and industry. Moreover, a good number of domain-specific

studies also utilised the aforementioned evaluation methods. In terms of specific tasks, user studies were mostly conducted for evaluating recommender systems. Very few studies have conducted real experiments, which were found to be from healthcare and industry domains for decision support, image processing, and predictive maintenance.

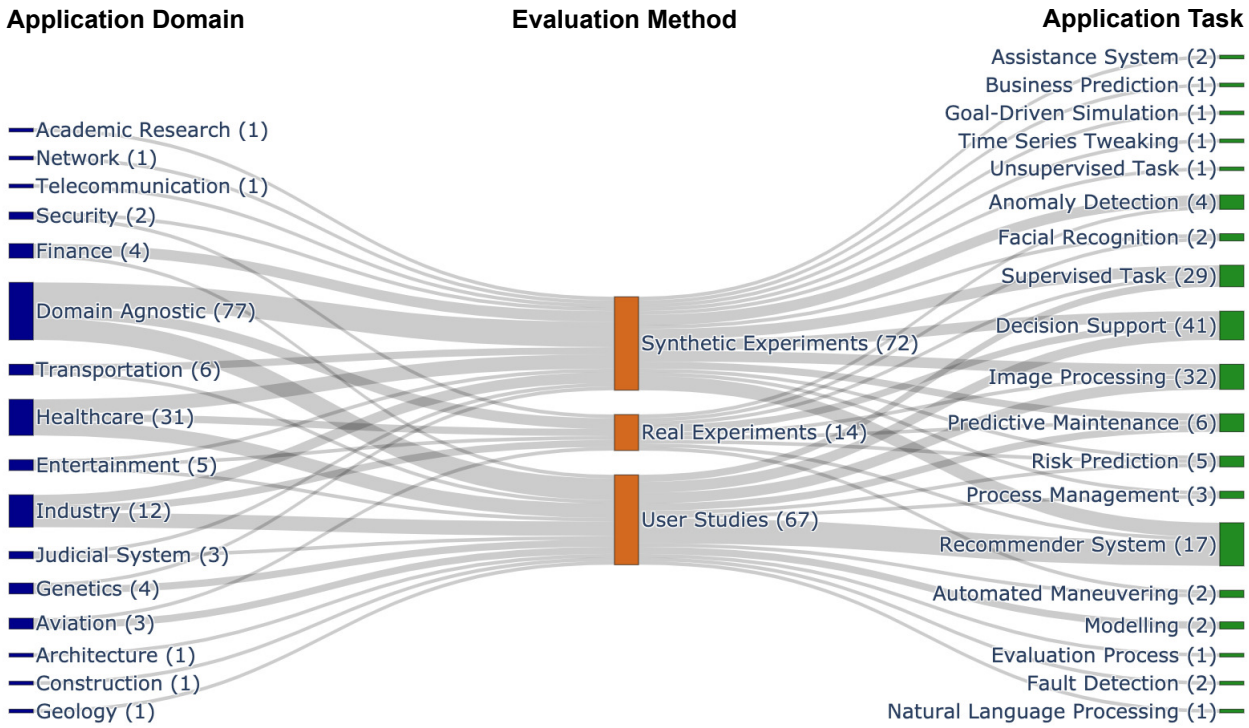


Figure 14. Different methods of evaluating explanations which were presented in the selected articles with the number of studies given in parentheses. Corresponding application domains and tasks of the performed evaluation methods are illustrated with links. The widths of the links are proportional to the number of studies. Some of the studies invoked a combination of different evaluation methods.

6. Discussion

The continuously growing interest in the research domain of XAI worldwide resulted in the publication of a large number of research articles containing diverse knowledge of explainability from different perspectives. In the published articles, it is often noticed that similar terms are used interchangeably [20], which is one of the major hurdles for a new researcher to initiate work on developing a new methodology of XAI. In addition, an "Explainable AI (XAI) Program" by DARPA [5], the Chinese Government's "The Development Plan for New Generation of Artificial Intelligence" [6] and the GDPR by the EU [7] escalated the number of research studies during the past couple of years, as demonstrated in Figure 1. The literature shows several review and survey studies on XAI philosophy, taxonomy, methodology, evaluation, etc. Nevertheless, to our knowledge, no study has been performed that has wholly focused on the XAI methodologies from the perspective of different application domains and tasks, let alone following some prescribed technique of conducting literature reviews. In contrast, this SLR followed a proper guideline [38] that precisely defines the methodology of surveying the recent developments in XAI techniques and evaluation criteria. One of the major advantages of an SLR is that the methodology contains a workflow for reviewing literature by defining and addressing specific RQs to restrict the subject matter of a study to the scope of the designated topic. Here, the RQs presented in Section 4.1.2 were purposefully designed to review the development and evaluation of XAI methodologies and were addressed with the presented outcomes of the study listed in Section 5.

This study started with the task of scanning more than a thousand peer-reviewed articles from different bibliographic databases. Following the process described in Section 4.2.1, 137 articles were thoroughly analysed to summarise the recent developments. Among the selected articles, 19 were added through the snowballing search, prescribed by Wohlin [43]. Here, the cited articles in the pre-selected articles were checked to identify more articles that met this study's inclusion criteria. While conducting the snowballing search, some of the articles meeting the inclusion criteria were found to be published prior to the defined period of 2018–2020 in the inclusion criteria (Table 1) but were apparently very significant in terms of content as they were cited in many of the pre-selected articles. Considering the impact of those articles in developing XAI methodologies, they were included in the study despite not completely meeting the inclusion criteria. Moreover, during the screening of articles, some of the articles were unintentionally overlooked due to the use of the specific keyword searched (*explainable artificial intelligence*) in the bibliographic databases. For example, this could be the article in which Spinner et al. presented a visual analytics framework for interactive and explainable machine learning [182]. For some unforeseen reason, the index terms of the article did not contain the aforementioned search keyword, but the abstract and keywords of the articles contained the term "Explainable AI". The interchangeable use of several closely related terms (e.g., interpretability, transparency, and explainability) in metadata impedes the proper acquisition of knowledge on XAI. As a result, a few potentially significant articles were overlooked during this review study. The absence of acquired knowledge from the neglected articles can be considered a limitation of this SLR.

The selected articles were analysed from five different viewpoints, i.e., metadata, primary task, explainability, the form of explanation, and the evaluation of methods and explanations. The prominent features from the respective viewpoints are summarised in Table 3. The features and possible alternatives were set in such a way that the result of the analysis can substantially address the RQs. Section 5 presents the outcomes of the analysis by identifying insights into the domains and applications in which XAI is developing, the prevailing methods of generating and evaluating explanations, etc. This information is thus readily available for prospective researchers from miscellaneous domains to instigate research projects on the methodological development of XAI. In addition, a questionnaire survey was designed and administered to the authors of the selected articles with several aims: to cure the extracted feature values from the articles, to assess the credibility of the definition of the features, etc. The questionnaire was distributed to the authors through email, and the response rate was approximately 50%. The responses were apparently similar to the information extracted from the articles, except in a few cases. For example, from the article, it was found that the input data for the method developed by Dujmovic were numeric [103]. In contrast, from the author's response, the input data were mentioned as LSP, and this information was incorporated in the analysis. This instance of curating, clarifying, and cross-checking the information extracted from the articles advocates the need for a questionnaire survey. This review study took advantage of the questionnaire survey to assess the credibility of the literature reviewer as well as clarify the information.

During the exploration of the contents of the sorted-out articles, the first step was to analyse the metadata. To determine the relevancy of the articles, keywords that were explicitly defined by the authors and keywords extracted from the abstracts were investigated in the form of word clouds following the methodology developed by Helbich et al. [44]. It was observed that the significant terms were *explainable artificial intelligence*, *deep learning*, *machine learning*, *explainability*, *visualisation* etc. These terms were considered significant due to their larger appearance in the word cloud, which resulted from repeated occurrences of the terms in the supplied texts. In addition, a higher number of occurrences of terms, such as *deep learning* or *visualisation*, aligns with the higher number of studies with concepts presented in Tables 5 and 6, indicating tunnel vision in XAI development. More attention towards less investigated models, such as SVM and neuro-fuzzy models and visualisation techniques would add more value and novelty towards XAI. Moreover, the

prominent terms are strongly related to the primary concept of this study, which increases the confidence in the selected articles that they are related. In addition, the terms from the author-defined keywords were more conceptual than the terms from the abstracts of the articles. On the other hand, the abstracts contained more specific terms based on the application tasks and AI/ML models. From the metadata, the countries of the authors' affiliations were evaluated, and it was found that the USA leads by a significant margin in terms of the number of publications. However, the collective publications from the countries belonging to the EU exceed the number of publications from the USA. This high number of publications indicated the immense impact of imposing various regulations and expressing interest through different programs from different governments. Although there was a development plan on XAI from the Government of China, the number of screened articles was lower, and they were published by the authors affiliated with the institutions in China. Overall, it can be stated that the number of research studies on XAI escalated in the regions where the government authorities put forward some programs or regulations. Concerning the recent regulatory developments, it is safe to assume that the government funding agencies have increased patronising this specific field which has resulted in a higher number of research publications, as shown in Figure 7.

In the subsequent sections, significant aspects of developing XAI methods are discussed, including addressing the RQs (defined in Section 4.1.2) with respect to the defined features and outcomes of the performed analyses.

6.1. Input Data and Models for Primary Task

Input data were stated to be an essential aspect to be considered for developing explainable systems by Vilone and Longo [8]. Therefore, the different forms of input data which were deliberately used in the studies of the selected articles were investigated in this review. It was observed that the vectors containing numeric values were used in most of the articles, followed by the use of images as input. With the growing variety of data forms, more concentration is required to explain models and decisions that can be derived from other forms of data, such as graphs and texts. However, from the findings of this study, it is apparent that some specific forms of data are already being exploited by the researchers of respective subjects in a limited margin; for example, graph structures are considered as input to XAI methodologies developed with fuzzy and neuro-fuzzy models. The uses of different input data types are illustrated in Figure 10 within the structure of a Venn diagram as many of the articles used multiple types of input data for their proposed models, and the Venn diagram has the capability of presenting combined relations in terms of frequencies.

While investigating the models that were designed or applied to solve primary tasks, it was observed that most of the studies were performed concerning neural networks. Specifically, out of 122 articles on XAI methods, 60 articles presented work with various neural networks. The reason behind this overwhelming interest of researchers towards making neural networks explainable is undoubtedly the performance of these types of models in various tasks from diverse domains. A good number of studies utilised ensemble methods, fuzzy models and tree-based models. Other significant types of models were found to be SVM, CBR and Bayesian models (Table 5).

6.2. Development of Explainable Models in Different Application Domains

This section addresses this review study's outcome within the scope of *RQ1: What are the application domains and tasks in which XAI is being explored and exploited?* The question was further split into three research sub-questions to more precisely analyse the subject.

6.2.1. Application Domains and Tasks

To generate insight into the possible fields of application of XAI methods, *RQ1.1* was raised. A broader idea of the concerned application domains and tasks was developed from the metadata analysis. As illustrated in Figure 2a, most of the articles were published

without targeting any specific domain, which extends the horizon for XAI researchers to utilise the concepts from the studies and further enhance them in a domain-specific or domain-agnostic way. In the case of the domain-specific publications on XAI, the healthcare domain has been being developed much more than the other domains. The reason behind this massive interest in XAI from the healthcare domain is unquestionably the involvement of machines in matters that deal with human lives. Simultaneously, it was observed from Figure 2b that most of the research studies were carried out to make decision support systems more explainable to users. Additionally, a good number of studies have been performed on image processing and recommender systems. All these application tasks can also be employed in the healthcare domain. From the distribution of the articles based on the application domain and tasks, it could be concluded that XAI has been profoundly exploited where humans are directly involved.

6.2.2. Explainable Models

RQ1.1 was proposed to investigate the models that are explainable by design. From the theoretical point of view, as discussed in Section 2, the inference mechanism of some models can be understood by humans provided they have a certain level of expertise. In reality, these models are often termed transparent models. Barredo Arrieta et al. categorised linear/logistic regression, decision trees, k-nearest neighbours, rule-based learners, general additive models and Bayesian models as transparent AI/ML models [20]. Concerning the stages of generating explanations, ante hoc methods are invoked for the transparent models where the explanations are generated based on their transparency by design. Table 6 presents the methods available for generating explanations. Similar shreds of evidence found that ante hoc methods were used for generating explanations from most of the transparent models used for solving the primary task of classification/regression or clustering. On the other hand, post hoc methods were observed in action for the simplification of ensemble models, neural networks, SVMs, etc. (Table 6). Generally, in the post hoc method, a surrogate model is developed to mimic the inference mechanism of the black-box models, which is comparatively simpler and less complex than ante hoc methods, where the explanation is generated during the inference process. It can be deduced from the thematic synthesis of the selected articles that post hoc methods are suitable for the established and running systems without manipulating the prevailing mechanism and performance of the systems. However, for new systems with the requirement of explaining model decisions, ante hoc methods are more appropriate. In addition, visualisation and feature relevance techniques were induced to generate explanations for users of different levels of expertise. As a result, several tools for post hoc methods, such as LIME, SHAP, Anchors, and ELI5 and their variations have evolved for advanced users. Researchers from different domains have utilised these tools and added explainability to the black-box AI/ML models.

6.2.3. Forms of Explanation

The outcome of an explainable model, i.e., the form of an explanation, was the prime concern of *RQ1.2*. Four basic types of explanations were observed, i.e., numeric, rule-based, visual and textual (Figure 12). In addition to that, some of the articles presented mixed explanations, which combined the four types. Generally, visualisations are mostly used, which humans can more easily interpret than other types of explanations. This type of explanation contains charts, trend-lines etc., and conventionally visual explanation is preferable for image processing tasks. Numeric explanations were deliberately adopted in the developed systems targeted by the experts to show the clarification of the decision of a model with respect to different attributes in terms of feature importance. Understanding the numbers associated with different attributes seems slightly more difficult than the visual or textual representation for a general end-user. For providing numeric explanations, ante hoc methods are very few compared to post hoc methods. Rule-based methods are generally produced from the tree-based or ensemble methods, and most of them are ante hoc methods.

In this type of explanation, the inference mechanisms of the models were presented in the form of a table containing all the rules and tree-like graphs depicting the decision process in short. Finally, the textual explanations are some statements presented in a human-understandable format, which are less common than the other forms of explanations. This type of explanation can be adopted for the interactive systems where general users are involved but it demands higher computational complexity due to NLP tasks. In summary, textual explanations in the form of natural language should be presented for the general users, rule-based explanations and visualisations are found to be appropriate for advanced users, and numeric explanations are mostly appropriate for experts.

6.3. Evaluation Metrics for Explainable Models

This section addresses *RQ1.3*, which was proposed to investigate the development of evaluation methods for the explainability of a model and the metrics for validating the generated explanations. Currently available methods of evaluating explainable AI/ML models are apparently not as substantial as those for state-of-the-art black-box models, let alone the evaluation metrics of the explanations. From the studied articles, it was observed that most of the articles adopted state-of-the-art performance metrics to validate the developed explainable models, such as accuracy, precision, and recall. In addition to these established metrics, several works have proposed and utilised novel metrics which are discussed in Section 5.3.5. On the other hand, it was found that researchers conducted user studies to validate the quality of the explanations. In most cases, user studies included a meagre number of participants. However, several researchers proposed effective means of measuring the quality of an explanation and developing proper explainable models. For example, Holzinger et al. proposed SCS to measure the causability of the explanations generated from a model [56]. In another article, Sokol and Flach developed an explainability fact sheet to be followed while developing XAI methodologies, which is a major takeaway of this review study [155]. However, further investigation is required to establish domain-, application-, and method-specific methodologies that keep humans in the loop, as users' level of expertise largely contributes to their understanding of the explanations.

6.4. Open Issues and Future Research Direction

One of the objectives of this study was to sort out the open issues on developing explainable models and propose future research directions for different application domains and tasks. On the basis of the studies presented in the selected articles for this SLR, it was observed that the proposed methodologies' major limitation lies with the evaluation of the explanations. The studies addressed this issue with different techniques of user studies and experiments. However, there is still an urgent need for a generic method for evaluating the explanations. Another observed issue was algorithm-specific approaches of adding explainability. It is an obstacle to making the established systems in action explainable. Additionally, there remain other open issues to be addressed. Based on the observed shortcomings of prevailing explainable models, several possible research directions are outlined below:

- It is evident in the findings of the study that safety-critical domains and associated tasks are most facilitated with the development of XAI. However, less investigation was performed for other sensitive domains, such as the judicial system, finance and academia, in contrast with the domains of healthcare and industry. Further exploitation of the methods can be performed for the less developed domains in terms of XAI;
- One of the promising research areas in the domain of networking is the Internet of Things (IoT). The literature indicates that several applications such as anomaly detection [183] and building information systems [184,185] for IoT have been facilitated by agent-based algorithms. These applications can be further associated with XAI methods to make them more acceptable to end-users;

- The impact of the dataset (particularly the effect of dataset imbalance, feature dimensionality, different types of bias problems in data acquisition and dataset, etc.) on developing an explainable model can be assessed through studies;
- It was observed that most of the works were performed done for neural networks and through post hoc methods, explanations were generated at the local scope. Similar cases were also observed for other models, such as SVM and ensemble models, since their inference mechanism remains unclear to users. Although several studies have shown approaches to produce explanations at a global scope by mimicking the models' behaviour, they lack performance accuracy. More investigations can be carried out to produce an explanation in a global scope without compromising the models' performance for the base task;
- The major challenge of evaluating an explanation is to develop a method that can deal with the different levels of expertise and understanding of users. Generally, these two characteristics of users vary from person to person. Substantial research is needed to establish a proper methodology for evaluating the explanations based on the intended users' expertise and capacity;
- User studies were invoked to validate explanations based on natural language, in short, textual explanations. Automated evaluation metrics for textual explanations are not yet prominent in the research works;
- Evaluating the quality of heatmaps as a form of visualisation is still undiscovered beyond the visual assessment technique. In addition to heatmaps, evaluation metrics for other visualisation techniques, e.g., saliency maps, are yet to be defined.

7. Conclusions

This paper presented a thematic synthesis of articles on the application domains of XAI methodologies and their evaluation metrics through an SLR. The significant contributions of this study are (1) lists of application domains and tasks that have been facilitated with the XAI methods; (2) currently available approaches for adding explanations to AI/ML models and their evaluation metrics; and (3) exploited mediums of explanations, such as numeric and rule-based explanations. References to the preliminary research studies could provide an example to assist prospective researchers from diverse domains to initiate research on developing new XAI methodologies. However, articles published after the mentioned period were not analysed during this study due to time constraints. Several articles were also excluded because of the specific search keywords used in the bibliographic databases. More comprehensive primary and secondary analyses on the methodological development of XAI are required across different application domains. We believe such studies could expedite the human acceptability of intelligent systems. Accommodating the varying levels of expertise will also help understand different user groups' needs. These studies would explicitly explore underlying characteristics of transparent models (fuzzy, CBR, etc.) deployed for respective tasks, carefully analyse the dataset's impact, and consider well-established metrics for evaluating all forms of explanations.

Author Contributions: Conceptualisation, M.R.I.; methodology, M.R.I.; software, M.R.I.; validation, M.R.I., M.U.A., and S.B. (Shaibal Barua); formal analysis, M.R.I.; investigation, M.R.I.; resources, M.R.I.; data curation, M.R.I., M.U.A., and S.B. (Shaibal Barua); writing—original draft preparation, M.R.I.; writing—review and editing, M.R.I., M.U.A., S.B. (Shaibal Barua), and S.B. (Shahina Begum); visualisation, M.R.I. and M.U.A.; supervision, M.U.A. and S.B. (Shahina Begum); project administration, M.U.A. and S.B. (Shahina Begum); funding acquisition, M.U.A. All authors have read and agreed to the published version of the manuscript.

Funding: This study was performed as part of the following projects: (i) SIMUSAFE, funded by the European Union's Horizon 2020 research and innovation programme under grant agreement N. 723386; (ii) BrainSafeDrive, co-funded by the Vetenskapsrådet - The Swedish Research Council and the Ministero dell'Istruzione dell'Università e della Ricerca della Repubblica Italiana under Italy–Sweden Cooperation Program; and (iii) ARTIMATION, funded by the SESAR Joint Undertaking under

the European Union's Horizon 2020 research and innovation programme under grant agreement N. 894238.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Rai, A. Explainable AI: From Black Box to Glass Box. *J. Acad. Mark. Sci.* **2020**, *48*, 137–141. [CrossRef]
2. Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Giannotti, F.; Pedreschi, D. A Survey of Methods for Explaining Black Box Models. *ACM Comput. Surv.* **2018**, *51*, 1–42. [CrossRef]
3. Loyola-González, O. Black-Box vs. White-Box: Understanding Their Advantages and Weaknesses From a Practical Point of View. *IEEE Access* **2019**, *7*, 154096–154113. [CrossRef]
4. Neches, R.; Swartout, W.; Moore, J. Enhanced Maintenance and Explanation of Expert Systems Through Explicit Models of Their Development. *IEEE Trans. Softw. Eng.* **1985**, *SE-11*, 1337–1351. [CrossRef]
5. Gunning, D.; Aha, D. DARPA's Explainable Artificial Intelligence (XAI) Program. *AI Mag.* **2019**, *40*, 44–58.
6. Xu, F.; Uszkoreit, H.; Du, Y.; Fan, W.; Zhao, D.; Zhu, J. Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges. In *Natural Language Processing and Chinese Computing*; Tang, J., Kan, M.Y., Zhao, D., Li, S., Zan, H., Eds.; Springer International Publishing: Berlin/Heidelberg, Germany, 2019; Volume 11839 LNAI, pp. 563–574._51. [CrossRef]
7. Wachter, S.; Mittelstadt, B.; Russell, C. Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. *Harv. J. Law Technol.* **2018**, *31*, 841–887. [CrossRef]
8. Vilone, G.; Longo, L. Explainable Artificial Intelligence: A Systematic Review. *arXiv* **2020**, arXiv:2006.00093.
9. Vilone, G.; Longo, L. Classification of Explainable Artificial Intelligence Methods through Their Output Formats. *Mach. Learn. Knowl. Extr.* **2021**, *3*, 615–661. [CrossRef]
10. Lacave, C.; Diéz, F.J. A Review of Explanation Methods for Bayesian Networks. *Knowl. Eng. Rev.* **2002**, *17*, 107–127. [CrossRef]
11. Ribeiro, M.T.; Singh, S.; Guestrin, C. Model-Agnostic Interpretability of Machine Learning. In Proceedings of the ICML Workshop on Human Interpretability in Machine Learning, New York, NY, USA, 23 June 2016. Available online: https://arxiv.org/abs/1606.05386 (accessed on 30 June 2021).
12. 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; ACM: New York, NY, USA, 2016; pp. 1135–1144. [CrossRef]
13. Alonso, J.M.; Castiello, C.; Mencar, C. A Bibliometric Analysis of the Explainable Artificial Intelligence Research Field. In *Information Processing and Management of Uncertainty in Knowledge-Based Systems. Theory and Foundations*; Medina, J., Ojeda-Aciego, M., Verdegay, J.L., Pelta, D.A., Cabrera, I.P., Bouchon-Meunier, B., Yager, R.R., Eds.; Springer International Publishing: Berlin/Heidelberg, Germany, 2018; Volume 853, pp. 3–15._1. [CrossRef]
14. Goebel, R.; Chander, A.; Holzinger, K.; Lecue, F.; Akata, Z.; Stumpf, S.; Kieseberg, P.; Holzinger, A. Explainable AI: The New 42? In *Machine Learning and Knowledge Extraction*; Holzinger, A., Kieseberg, P., Tjoa, A.M., Weippl, E., Eds.; Springer: Cham, Switzerland, 2018; Volume 11015 LNCS, pp. 295–303._21. [CrossRef]
15. Rudin, C. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nat. Mach. Intell.* **2019**, *1*, 206–215. [CrossRef]
16. Dosilovic, F.K.; Brcic, M.; Hlupic, N. Explainable Artificial Intelligence: A Survey. In Proceedings of the 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO 2018), Opatija, Croatia, 21–25 May 2018; pp. 0210–0215. [CrossRef]
17. Mittelstadt, B.; Russell, C.; Wachter, S. Explaining Explanations in AI. In Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* 2019), Atlanta, GA, USA, 29–31 January 2019; ACM Press: New York, NY, USA, 2019; pp. 279–288. [CrossRef]
18. Samek, W.; Müller, K.R. Towards Explainable Artificial Intelligence. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*; Samek, W., Montavon, G., Vedaldi, A., Hansen, L.K., Müller, K.R., Eds.; Springer International Publishing: Berlin/Heidelberg, Germany, 2019; Volume 1, pp. 5–22._1. [CrossRef]
19. Preece, A.; Harborne, D.; Braines, D.; Tomsett, R.; Chakraborty, S. Stakeholders in Explainable AI. In Proceedings of the AAAI FSS-18: Artificial Intelligence in Government and Public Sector, Arlington, VA, USA, 18–20 October 2018. Available online: https://arxiv.org/abs/1810.00184 (accessed on 30 June 2021).
20. 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 Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. *Inf. Fusion* **2020**, *58*, 82–115. [CrossRef]

21. Longo, L.; Goebel, R.; Lecue, F.; Kieseberg, P.; Holzinger, A. Explainable Artificial Intelligence: Concepts, Applications, Research Challenges and Visions. In *Machine Learning and Knowledge Extraction*; Holzinger, A., Kieseberg, P., Tjoa, A.M., Weippl, E., Eds.; Springer International Publishing: Berlin/Heidelberg, Germany, 2020; Volume 12279, pp. 1–16._1. [CrossRef]
22. Adadi, A.; Berrada, M. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access* **2018**, *6*, 52138–52160. [CrossRef]
23. Guidotti, R.; Monreale, A.; Giannotti, F.; Pedreschi, D.; Ruggieri, S.; Turini, F. Factual and Counterfactual Explanations for Black Box Decision Making. *IEEE Intell. Syst.* **2019**, *34*, 14–23. [CrossRef]
24. Robnik-Šikonja, M.; Bohanec, M. Perturbation-Based Explanations of Prediction Models. In *Human and Machine Learning*; Zhou, J., Chen, F., Eds.; Springer International Publishing: Berlin/Heidelberg, Germany, 2018; pp. 159–175._9. [CrossRef]
25. Zhang, Q.; Wu, Y.N.; Zhu, S.C. Interpretable Convolutional Neural Networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, June 19–21 2018; pp. 8827–8836. [CrossRef]
26. Dağlarli, E. Explainable Artificial Intelligence (xAI) Approaches and Deep Meta-Learning Models. In *Advances and Applications in Deep Learning*; Marco Antonio Aceves-Fernandez Eds.; InTechOpen: London, UK, 2020. [CrossRef]
27. Vilone, G.; Longo, L. Notions of explainability and evaluation approaches for explainable artificial intelligence. *Inf. Fusion* **2021**, *76*, 89–106. [CrossRef]
28. Holzinger, A.; Langs, G.; Denk, H.; Zatloukal, K.; Müller, H. Causability and Explainability of Artificial Intelligence in Medicine. *WIREs Data Min. Knowl. Discov.* **2019**, *9*, 1–13. [CrossRef]
29. Mathews, S.M. Explainable Artificial Intelligence Applications in NLP, Biomedical, and Malware Classification: A Literature Review. In *Intelligent Computing*; Arai, K., Bhatia, R., Kapoor, S., Eds.; Springer International Publishing: Berlin/Heidelberg, Germany, 2019; Volume 998, pp. 1269–1292._90. [CrossRef]
30. Fellous, J.M.; Sapiro, G.; Rossi, A.; Mayberg, H.; Ferrante, M. Explainable Artificial Intelligence for Neuroscience: Behavioral Neurostimulation. *Front. Neurosci.* **2019**, *13*, 1–14. [CrossRef]
31. Jiménez-Luna, J.; Grisoni, F.; Schneider, G. Drug Discovery with Explainable Artificial Intelligence. *Nat. Mach. Intell.* **2020**, *2*, 573–584. [CrossRef]
32. Payrovnaziri, S.N.; Chen, Z.; Rengifo-Moreno, P.; Miller, T.; Bian, J.; Chen, J.H.; 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]
33. Ahmed, M.U.; Barua, S.; Begum, S. Artificial Intelligence, Machine Learning and Reasoning in Health Informatics—Case Studies. In *Signal Processing Techniques for Computational Health Informatics. Intelligent Systems Reference Library*; Ahad, M.A.R., Ahmed, M.U., Eds.; Springer International Publishing: Berlin/Heidelberg, Germany, 2021; Volume 192, pp. 261–291._12. [CrossRef]
34. 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]
35. Gade, K.; Geyik, S.C.; Kenthapadi, K.; Mithal, V.; Taly, A. Explainable AI in Industry. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Anchorage, AK, USA, 4–8 August 2019; ACM: New York, NY, USA, 2019; pp. 3203–3204.
36. Dam, H.K.; Tran, T.; Ghose, A. Explainable Software Analytics. In Proceedings of the 40th International Conference on Software Engineering New Ideas and Emerging Results–ICSE-NIER '18, Gothenburg, Sweden, 27 May – 3 June 2018; ACM Press: New York, NY, USA, 2018; pp. 53–56.
37. Chaczko, Z.; Kulbacki, M.; Gudzbeler, G.; Alsawwaf, M.; Thai-Chyzhykau, I.; Wajs-Chaczko, P. Exploration of Explainable AI in Context of Human–Machine Interface for the Assistive Driving System. In *Intelligent Information and Database Systems*; Nguyen, N.T., Jearanaitanakij, K., Selamat, A., Trawiński, B., Chittayasothorn, S., Eds.; Springer International Publishing: Berlin/Heidelberg, Germany, 2020; Volume 12034, pp. 507–516._42. [CrossRef]
38. Kitchenham, B.; Charters, S. *Guidelines for Performing Systematic Literature Reviews in Software Engineering*; Technical Report; Keele University; Keele, UK; Durham University: Durham, UK, 2007.
39. García-Holgado, A.; Marcos-Pablos, S.; García-Peñalvo, F. Guidelines for Performing Systematic Research Projects Reviews. *Int. J. Interact. Multimed. Artif. Intell.* **2020**, *6*, 9. [CrossRef]
40. Moher, D.; Liberati, A.; Tetzlaff, J.; Altman, D.G.; The PRISMA Group. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PLoS Med.* **2009**, *6*, e1000097. [CrossRef] [PubMed]
41. Da'u, A.; Salim, N. Recommendation System based on Deep Learning Methods: A Systematic Review and New Directions. *Artif. Intell. Rev.* **2020**, *53*, 2709–2748. [CrossRef]
42. Genc-Nayebi, N.; Abran, A. A Systematic Literature Review: Opinion Mining Studies from Mobile App Store User Reviews. *J. Syst. Softw.* **2017**, *125*, 207–219. [CrossRef]
43. 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–EASE '14, London, UK, 13–14 May 2014; ACM Press: New York, NY, USA, 2014; pp. 1–10. [CrossRef]
44. Helbich, M.; Hagenauer, J.; Leitner, M.; Edwards, R. Exploration of Unstructured Nrrative Crime Reports: An Unsupervised Neural Network and Point Pattern Analysis Approach. *Cartogr. Geogr. Inf. Sci.* **2013**, *40*, 326–336. [CrossRef]

45. Tintarev, N.; Rostami, S.; Smyth, B. Knowing the unknown: Visualising consumption blind-spots in recommender systems. In Proceedings of the 33rd Annual ACM Symposium on Applied Computing, Pau, France, 9–13 April 2018; ACM: New York, NY, USA, 2018; SAC '18, pp. 1396–1399. [CrossRef]
46. Galhotra, S.; Pradhan, R.; Salimi, B. Explaining Black-Box Algorithms Using Probabilistic Contrastive Counterfactuals. In Proceedings of the 2021 International Conference on Management of Data, Virtual Event, 20–25 June 2021; pp. 577–590. [CrossRef]
47. La Gatta, V.; Moscato, V.; Postiglione, M.; Sperlì, G. CASTLE: Cluster-Aided Space Transformation for Local Explanations. *Expert Syst. Appl.* **2021**, *179*, 115045. [CrossRef]
48. La Gatta, V.; Moscato, V.; Postiglione, M.; Sperlì, G. PASTLE: Pivot-Aided Space Transformation for Local Explanations. *Pattern Recognit. Lett.* **2021**, *149*, 67–74. [CrossRef]
49. Moradi, M.; Samwald, M. Post-hoc Explanation of Black-box Classifiers using Confident Itemsets. *Expert Syst. Appl.* **2021**, *165*, 113941. [CrossRef]
50. Hatwell, J.; Gaber, M.M.; Muhammad Atif Azad, R. Gbt-hips: Explaining the classifications of gradient boosted tree ensembles. *Appl. Sci.* **2021**, *11*, 2511. [CrossRef]
51. Rubio-Manzano, C.; Segura-Navarrete, A.; Martinez-Araneda, C.; Vidal-Castro, C. Explainable hopfield neural networks using an automatic video-generation system. *Appl. Sci.* **2021**, *11*. [CrossRef]
52. Alonso, J.M.; Toja-Alamancos, J.; Bugarin, A. Experimental Study on Generating Multi-modal Explanations of Black-box Classifiers in terms of Gray-box Classifiers. In Proceedings of the 2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Glasgow, UK, 19–24 July 2020; pp. 1–8. [CrossRef]
53. Biswas, R.; Barz, M.; Sonntag, D. Towards Explanatory Interactive Image Captioning Using Top-Down and Bottom-Up Features, Beam Search and Re-ranking. *KI–Kunstl. Intell.* **2020**, *34*, 571–584. [CrossRef]
54. Cao, H.; Sarlin, R.; Jung, A. Learning Explainable Decision Rules via Maximum Satisfiability. *IEEE Access* **2020**, *8*, 218180–218185. [CrossRef]
55. Fernández, R.R.; Martín de Diego, I.; Aceña, V.; Fernández-Isabel, A.; Moguerza, J.M. Random Forest Explainability using Counterfactual Sets. *Inf. Fusion* **2020**, *63*, 196–207. [CrossRef]
56. Holzinger, A.; Carrington, A.; Müller, H. Measuring the Quality of Explanations: The System Causability Scale (SCS). *KI - Künstliche Intell.* **2020**, *34*, 193–198.
57. Kovalev, M.S.; Utkin, L.V. A Robust Algorithm for Explaining Unreliable Machine Learning Survival Models using Kolmogorov–Smirnov Bounds. *Neural Netw.* **2020**, *132*, 1–18.
58. Le, T.; Wang, S.; Lee, D. GRACE: Generating Concise and Informative Contrastive Sample to Explain Neural Network Model's Prediction. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Virtual Event, 23–27 August 2020; ACM: New York, NY, USA, 2020; pp. 238–248.
59. Lundberg, S.M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J.M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S.I. From Local Explanations to Global Understanding with Explanable AI for Trees. *Nat. Mach. Intell.* **2020**, *2*, 56–67.
60. Yang, Z.; Zhang, A.; Sudjianto, A. Enhancing Explainability of Neural Networks Through Architecture Constraints. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *6*, 2610–2621.
61. Sabol, P.; Sinčák, P.; Magyar, J.; Hartono, P. Semantically Explainable Fuzzy Classifier. *Int. J. Pattern Recognit. Artif. Intell.* **2019**, *33*, 2051006. [CrossRef]
62. Chander, A.; Srinivasan, R. Evaluating Explanations by Cognitive Value. In *Machine Learning and Knowledge Extraction*; Holzinger, A., Kieseberg, P., Tjoa, A.M., Weippl, E., Eds.; Springer International Publishing: Berlin/Heidelberg, Germany, 2018; pp. 314–328._23. [CrossRef]
63. Laugel, T.; Lesot, M.J.; Marsala, C.; Renard, X.; Detyniecki, M. Comparison-Based Inverse Classification for Interpretability in Machine Learning. In *Information Processing and Management of Uncertainty in Knowledge-Based Systems. Theory and Foundations*; Medina, J., Ojeda-Aciego, M., Verdegay, J.L., Pelta, D.A., Cabrera, I.P., Bouchon-Meunier, B., Yager, R.R., Eds.; Springer International Publishing: Berlin/Heidelberg, Germany, 2018; Volume 853, pp. 100–111._9. [CrossRef]
64. Pierrard, R.; Poli, J.P.; Hudelot, C. Learning Fuzzy Relations and Properties for Explainable Artificial Intelligence. In Proceedings of the 2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Rio de Janeiro, Brazil, 8–13 July 2018; pp. 1–8. [CrossRef]
65. Plumb, G.; Molitor, D.; Talwalkar, A. Model Agnostic Supervised Local Explanations. In Proceedings of the 32nd International Conference on Neural Information Processing Systems (NIPS '18), Montreal, Canada, 3–8 December 2018; pp. 2520–2529.
66. Bonanno, D.; Nock, K.; Smith, L.; Elmore, P.; Petry, F. An Approach to Explainable Deep Learning using Fuzzy Inference. In *Next-Generation Analyst V*; Hanratty, T.P., Llinas, J., Eds.; SPIE: Bellingham, WA, USA, 2017; Volume 10207. [CrossRef]
67. Štrumbelj, E.; Kononenko, I. Explaining Prediction Models and Individual Predictions with Feature Contributions. *Knowl. Inf. Syst.* **2014**, *41*, 647–665. [CrossRef]
68. Féraud, R.; Clérot, F. A Methodology to Explain Neural Network Classification. *Neural Netw.* **2002**, *15*, 237–246. [CrossRef]
69. Chandrasekaran, J.; Lei, Y.; Kacker, R.; Richard Kuhn, D. A Combinatorial Approach to Explaining Image Classifiers. In Proceedings of the 2021 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW), Virtual Event, 12–16 April 2021; pp. 35–43. [CrossRef]
70. Jung, Y.J.; Han, S.H.; Choi, H.J. Explaining CNN and RNN Using Selective Layer-Wise Relevance Propagation. *IEEE Access* **2021**, *9*, 18670–18681. [CrossRef]

71. Yang, S.H.; Vong, W.; Sojitra, R.; Folke, T.; Shafto, P. Mitigating Belief Projection in Explainable Artificial Intelligence via Bayesian Teaching. *Sci. Rep.* **2021**, *11*, 9863. [CrossRef] [PubMed]
72. Schorr, C.; Goodarzi, P.; Chen, F.; Dahmen, T. Neuroscope: An Explainable AI Toolbox for Semantic Segmentation and Image Classification of Convolutional Neural Nets. *Appl. Sci.* **2021**, *11*, 2199. [CrossRef]
73. Angelov, P.; Soares, E. Towards Explainable Deep Neural Networks (xDNN). *Neural Netw.* **2020**, *130*, 185–194. [CrossRef] [PubMed]
74. Apicella, A.; Isgrò, F.; Prevete, R.; Tamburrini, G. Middle-Level Features for the Explanation of Classification Systems by Sparse Dictionary Methods. *Int. J. Neural Syst.* **2020**, *30*, 2050040. [CrossRef] [PubMed]
75. Dutta, V.; Zielińska, T. An Adversarial Explainable Artificial Intelligence (XAI) based Approach for Action Forecasting. *J. Autom. Mob. Robot. Intell. Syst.* **2020**, *14*, 3–10. [CrossRef]
76. Murray, B.J.; Anderson, D.T.; Havens, T.C.; Wilkin, T.; Wilbik, A. Information Fusion-2-Text: Explainable Aggregation via Linguistic Protoforms. In *Information Processing and Management of Uncertainty in Knowledge-Based Systems*; Lesot, M.J., Vieira, S., Reformat, M.Z., Carvalho, J.P., Wilbik, A., Bouchon-Meunier, B., Yager, R.R., Eds.; Springer International Publishing: Berlin/Heidelberg, Germany, 2020; Volume 1239 CCIS, pp. 114–127._9. [CrossRef]
77. Oh, K.; Kim, S.; Oh, I.S. Salient Explanation for Fine-Grained Classification. *IEEE Access* **2020**, *8*, 61433–61441. [CrossRef]
78. Poyiadzi, R.; Sokol, K.; Santos-Rodriguez, R.; De Bie, T.; Flach, P. FACE: Feasible and Actionable Counterfactual Explanations. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, New York, NY, USA, 7–8 February 2020; ACM: New York, NY, USA, 2020; pp. 344–350.
79. Riquelme, F.; De Goyeneche, A.; Zhang, Y.; Niebles, J.C.; Soto, A. Explaining VQA Predictions using Visual Grounding and a Knowledge Base. *Image Vis. Comput.* **2020**, *101*, 103968. [CrossRef]
80. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *Int. J. Comput. Vis.* **2020**, *128*, 336–359.
81. Tan, R.; Khan, N.; Guan, L. Locality Guided Neural Networks for Explainable Artificial Intelligence. In Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 19–24 July 2020; pp. 1–8.
82. Yeganejou, M.; Dick, S.; Miller, J. Interpretable Deep Convolutional Fuzzy Classifier. *IEEE Trans. Fuzzy Syst.* **2020**, *28*, 1407–1419. [CrossRef]
83. Oramas M., J.; Wang, K.; Tuytelaars, T. Visual Explanation by Interpretation: Improving Visual Feedback Capabilities of Deep Neural Networks. In Proceedings of the International Conference on Learning Representations (ICLR 2019), New Orleans, LA, USA, 6–9 May 2019.
84. Lundberg, S.M.; Lee, S.I. A Unified Approach to Interpreting Model Predictions. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17), Long Beach, CA, USA, 4–9 December 2017; pp. 4766–4775.
85. Montavon, G.; Lapuschkin, S.; Binder, A.; Samek, W.; Müller, K.R. Explaining Nonlinear Classification Decisions with Deep Taylor Decomposition. *Pattern Recognit.* **2017**, *65*, 211–222.
86. Hendricks, L.A.; Akata, Z.; Rohrbach, M.; Donahue, J.; Schiele, B.; Darrell, T. Generating Visual Explanations. In *Computer Vision–ECCV 2016*; Leibe, B., Matas, J., Sebe, N., Welling, M., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2016; Volume 9908, pp. 3–19._1. [CrossRef]
87. Bach, S.; Binder, A.; Montavon, G.; Klauschen, F.; Müller, K.R.; Samek, W. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLoS ONE* **2015**, *10*, e0130140. [CrossRef] [PubMed]
88. Alonso, J.M.; Ducange, P.; Pecori, R.; Vilas, R. Building Explanations for Fuzzy Decision Trees with the ExpliClas Software. In Proceedings of the 2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Glasgow, UK, 19–24 July 2020; pp. 1–8. [CrossRef]
89. De, T.; Giri, P.; Mevawala, A.; Nemani, R.; Deo, A. Explainable AI: A Hybrid Approach to Generate Human-Interpretable Explanation for Deep Learning Prediction. *Procedia Comput. Sci.* **2020**, *168*, 40–48. [CrossRef]
90. Islam, M.A.; Anderson, D.T.; Pinar, A.; Havens, T.C.; Scott, G.; Keller, J.M. Enabling Explainable Fusion in Deep Learning with Fuzzy Integral Neural Networks. *IEEE Trans. Fuzzy Syst.* **2020**, *28*, 1291–1300,
91. Meskauskas, Z.; Jasinevicius, R.; Kazanavicius, E.; Petrauskas, V. XAI-Based Fuzzy SWOT Maps for Analysis of Complex Systems. In Proceedings of the 2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Glasgow, UK, 19–24 July 2020; pp. 1–8. [CrossRef]
92. Waa, J.V.D.; Schoonderwoerd, T.; Diggelen, J.V.; Neerincx, M. Interpretable Confidence Measures for Decision Support Systems. *Int. J. Hum.-Comput. Stud.* **2020**, *144*, 102493. [CrossRef]
93. Garcia-Magarino, I.; Muttukrishnan, R.; Lloret, J. Human-Centric AI for Trustworthy IoT Systems With Explainable Multilayer Perceptrons. *IEEE Access* **2019**, *7*, 125562–125574. [CrossRef]
94. Ming, Y.; Qu, H.; Bertini, E. RuleMatrix: Visualizing and Understanding Classifiers with Rules. *IEEE Trans. Vis. Comput. Graph.* **2019**, *25*, 342–352.
95. Magdalena, L. Designing Interpretable Hierarchical Fuzzy Systems. In Proceedings of the 2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Rio de Janeiro, Brazil, 8–13 July 2018; pp. 1–8. [CrossRef]
96. Ribeiro, M.T.; Singh, S.; Guestrin, C. Anchors: High-precision Model-agnostic Explanations. In Proceedings of the 32nd AAAI Conference on Artificial Intelligence, (AAAI-18), New Orleans, LA, USA, 2–7 February 2018; pp. 1527–1535.

97. Massie, S.; Craw, S.; Wiratunga, N. A Visualisation Tool to Explain Case-Base Reasoning Solutions for Tablet Formulation. In *Applications and Innovations in Intelligent Systems XII*; Springer: London, UK, 2004; pp. 222–234._16. [CrossRef]
98. Csiszár, O.; Csiszár, G.; Dombi, J. Interpretable Neural Networks based on Continuous-valued Logic and Multicriteria Decision Operators. *Knowl.-Based Syst.* **2020**, *199*, 105972. [CrossRef]
99. Jung, A.; Nardelli, P.H.J. An Information-Theoretic Approach to Personalized Explainable Machine Learning. *IEEE Signal Process. Lett.* **2020**, *27*, 825–829.
100. Kouki, P.; Schaffer, J.; Pujara, J.; O'Donovan, J.; Getoor, L. Generating and Understanding Personalized Explanations in Hybrid Recommender Systems. *ACM Trans. Interact. Intell. Syst.* **2020**, *10* 1–40. [CrossRef]
101. Bharadhwaj, H.; Joshi, S. Explanations for Temporal Recommendations. *KI–Künstliche Intell.* **2018**, *32*, 267–272.
102. Loyola-González, O.; Gutierrez-Rodriguez, A.E.; Medina-Perez, M.A.; Monroy, R.; Martinez-Trinidad, J.F.; Carrasco-Ochoa, J.A.; Garcia-Borroto, M. An Explainable Artificial Intelligence Model for Clustering Numerical Databases. *IEEE Access* **2020**, *8*, 52370–52384. [CrossRef]
103. Dujmovic, J. Interpretability and Explainability of LSP Evaluation Criteria. In Proceedings of the 2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Glasgow, UK, 19–24 July 2020; pp. 1–8. [CrossRef]
104. Ramos-Soto, A.; Pereira-Fariña, M. Reinterpreting Interpretability for Fuzzy Linguistic Descriptions of Data. In *Information Processing and Management of Uncertainty in Knowledge-Based Systems. Theory and Foundations.*; Medina, J., Ojeda-Aciego, M., Verdegay, J.L., Pelta, D.A., Cabrera, I.P., Bouchon-Meunier, B., Yager, R.R., Eds.; Springer: Cham, Switzerland, 2018; Volume 853, pp. 40–51._4. [CrossRef]
105. Shalaeva, V.; Alkhoury, S.; Marinescu, J.; Amblard, C.; Bisson, G. Multi-operator Decision Trees for Explainable Time-Series Classification. In *Information Processing and Management of Uncertainty in Knowledge-Based Systems. Theory and Foundations*; Medina, J., Ojeda-Aciego, M., Verdegay, J.L., Pelta, D.A., Cabrera, I.P., Bouchon-Meunier, B., Yager, R.R., Eds.; Springer: Cham, Switzerland, 2018; Volume 853, pp. 86–99. [CrossRef]
106. Karlsson, I.; Rebane, J.; Papapetrou, P.; Gionis, A. Locally and Globally Explainable Time Series Tweaking. *Knowl. Inf. Syst.* **2020**, *62*, 1671–1700. [CrossRef]
107. Hu, Z.; Beyeler, M. Explainable AI for Retinal Prostheses: Predicting Electrode Deactivation from Routine Clinical Measures. In Proceedings of the 10th International IEEE EMBS Conference on Neural Engineering (NER '21), Virtual Event, 4–6 May 2021; pp. 792–796. [CrossRef]
108. Porto, R.; Molina, J.M.; Berlanga, A.; Patricio, M.A. Minimum relevant features to obtain explainable systems for predicting cardiovascular disease using the statlog data set. *Appl. Sci.* **2021**, *11*, 1285. [CrossRef]
109. 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] [PubMed]
110. Chou, Y.h.; Hong, S.; Zhou, Y.; Shang, J.; Song, M.; Li, H. Knowledge-shot Learning: An Interpretable Deep Model For Classifying Imbalanced Electrocardiography Data. *Neurocomputing* **2020**, *417*, 64–73. [CrossRef]
111. Dindorf, C.; Teufl, W.; Taetz, B.; Bleser, G.; Fröhlich, M. Interpretability of Input Representations for Gait Classification in Patients after Total Hip Arthroplasty. *Sensors* **2020**, *20*, 4385. [CrossRef]
112. Hatwell, J.; Gaber, M.M.; Atif Azad, R.M. Ada-WHIPS: Explaining AdaBoost Classification with Applications in the Health Sciences. *BMC Med. Inform. Decis. Mak.* **2020**, *20*, 250. [CrossRef] [PubMed]
113. Lamy, J.B.; Sedki, K.; Tsopra, R. Explainable Decision Support through the Learning and Visualization of Preferences from a Formal Ontology of Antibiotic Treatments. *J. Biomed. Inform.* **2020**, *104*, 103407. [CrossRef]
114. Lin, Z.; Lyu, S.; Cao, H.; Xu, F.; Wei, Y.; Hui, P.; Samet, H.; Li, Y. HealthWalks: Sensing Fine-grained Individual Health Condition via Mobility Data. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **2020**, *4*, 26. [CrossRef]
115. Panigutti, C.; Perotti, A.; Pedreschi, D. Doctor XAI An Ontology-based Approach to Black-box Sequential Data Classification Explanations. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* 2020), Barcelona, Spain, 27–30 January 2020; pp. 629–639. [CrossRef]
116. Soares, E.; Angelov, P.; Gu, X. Autonomous Learning Multiple-Model Zero-order Classifier for Heart Sound Classification. *Appl. Soft Comput. J.* **2020**, *94*, 106449. [CrossRef]
117. Tabik, S.; Gomez-Rios, A.; Martin-Rodriguez, J.; Sevillano-Garcia, I.; Rey-Area, M.; Charte, D.; Guirado, E.; Suarez, J.; Luengo, J.; Valero-Gonzalez, M.; et al. COVIDGR Dataset and COVID-SDNet Methodology for Predicting COVID-19 Based on Chest X-Ray Images. *IEEE J. Biomed. Health Inform.* **2020**, *24*, 3595–3605. [CrossRef] [PubMed]
118. Aghamohammadi, M.; Madan, M.; Hong, J.K.; Watson, I. Predicting Heart Attack Through Explainable Artificial Intelligence. In *Computational Science–ICCS 2019*; Rodrigues, J.M.F., Cardoso, P.J.S., Monteiro, J., Lam, R., Krzhizhanovskaya, V.V., Lees, M.H., Dongarra, J.J., Sloot, P.M., Eds.; Springer International Publishing:Berlin/Heidelberg, Germany, 2019; Volume 11537 LNCS, pp. 633–645. [CrossRef]
119. Palatnik de Sousa, I.; Maria Bernardes Rebuzzi Vellasco, M.; Costa da Silva, E. Local Interpretable Model-Agnostic Explanations for Classification of Lymph Node Metastases. *Sensors* **2019**, *19*, 2969. [CrossRef] [PubMed]
120. Kwon, B.C.; Choi, M.J.; Kim, J.T.; Choi, E.; Kim, Y.B.; Kwon, S.; Sun, J.; Choo, J. RetainVis: Visual Analytics with Interpretable and Interactive Recurrent Neural Networks on Electronic Medical Records. *IEEE Trans. Vis. Comput. Graph.* **2019**, *25*, 299–309.
121. Lamy, J.B.; Sekar, B.; Guezennec, G.; Bouaud, J.; Séroussi, B. Explainable Artificial Intelligence for Breast Cancer: A Visual Case-Based Reasoning Approach. *Artif. Intell. Med.* **2019**, *94*, 42–53. [CrossRef] [PubMed]

122. Senatore, R.; Della Cioppa, A.; Marcelli, A. Automatic Diagnosis of Neurodegenerative Diseases: An Evolutionary Approach for Facing the Interpretability Problem. *Information* **2019**, *10*, 30. [CrossRef]
123. Wang, D.; Yang, Q.; Abdul, A.; Lim, B.Y. Designing Theory-Driven User-Centric Explainable AI. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI 2019), Glasgow, UK, 4–9 May 2019; ACM Press: New York, NY, USA, 2019; pp. 1–15. [CrossRef]
124. Zheng, Q.; Delingette, H.; Ayache, N. Explainable Cardiac Pathology Classification on Cine MRI with Motion Characterization by Semi-supervised Learning of Apparent Flow. *Med. Image Anal.* **2019**, *56*, 80–95.
125. Lage, I.; Chen, E.; He, J.; Narayanan, M.; Kim, B.; Gershman, S.J.; Doshi-Velez, F. An Evaluation of the Human-Interpretability of Explanation. In Proceedings of the 32st International Conference on Neural Information Processing Systems (NIPS'18), Montreal, Canada, 3–8 December 2018.
126. Letham, B.; Rudin, C.; McCormick, T.H.; Madigan, D. Interpretable Classifiers using Rules and Bayesian Analysis: Building a Better Stroke Prediction Model. *Ann. Appl. Stat.* **2015**, *9*, 1350–1371. [CrossRef]
127. Lindsay, L.; Coleman, S.; Kerr, D.; Taylor, B.; Moorhead, A. Explainable Artificial Intelligence for Falls Prediction. In *Advances in Computing and Data Sciences*; Communications in Computer and Information Science; Singh, M., Gupta, P.K., Tyagi, V., Flusser, J., Ören, T., Valentino, G., Eds.; Springer: Singapore, 2020; Volume 1244, pp. 76–84. [CrossRef]
128. Pintelas, E.; Liaskos, M.; Livieris, I.E.; Kotsiantis, S.; Pintelas, P. Explainable Machine Learning Framework for Image Classification Problems: Case Study on Glioma Cancer Prediction. *J. Imaging* **2020**, *6*, 37. [CrossRef]
129. Prifti, E.; Chevaleyre, Y.; Hanczar, B.; Belda, E.; Danchin, A.; Clément, K.; Zucker, J.D. Interpretable and Accurate Prediction Models for Metagenomics Data. *GigaScience* **2020**, *9*, giaa010. [CrossRef]
130. Lundberg, S.M.; Nair, B.; Vavilala, M.S.; Horibe, M.; Eisses, M.J.; Adams, T.; Liston, D.E.; Low, D.K.W.; Newman, S.F.; Kim, J.; Lee, S.I. Explainable Machine Learning Predictions to Help Anesthesiologists Prevent Hypoxemia During Surgery. *Nat. Biomed. Eng.* **2018**, *2*, 749–760. [CrossRef]
131. Muddamsetty, S.; Jahromi, M.; Moeslund, T. Expert Level Evaluations for Explainable AI (XAI) Methods in the Medical Domain. In Proceedings of the 25th International Conference on Pattern Recognition Workshops (ICPR 2020), Virtual Event, 10–15 January 2021; Springer Science and Business Media Deutschland GmbH: Berlin/Heidelberg, Germany, 2021; Volume 12663 LNCS,
132. Graziani, M.; Andrearczyk, V.; S., M.M.; Müller, H. Concept Attribution: Explaining CNN Decisions to Physicians. *Comput. Biol. Med.* **2020**, *123*, 103865. [CrossRef]
133. Rio-Torto, I.; Fernandes, K.; Teixeira, L.F. Understanding the Decisions of CNNs: An In-model Approach. *Pattern Recognit. Lett.* **2020**, *133*, 373–380. [CrossRef]
134. D'Alterio, P.; Garibaldi, J.M.; John, R.I. Constrained Interval Type-2 Fuzzy Classification Systems for Explainable AI (XAI). In Proceedings of the 2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Glasgow, UK, 19–24 July 2020; pp. 1–8. [CrossRef]
135. Lauritsen, S.M.; Kristensen, M.; Olsen, M.V.; Larsen, M.S.; Lauritsen, K.M.; Jørgensen, M.J.; Lange, J.; Thiesson, B. Explainable Artificial Intelligence Model to Predict Acute Critical Illness from Electronic Health Records. *Nat. Commun.* **2020**, *11*, 3852,
136. Itani, S.; Lecron, F.; Fortemps, P. A One-class Classification Decision Tree based on Kernel Density Estimation. *Appl. Soft Comput. J.* **2020**, *91*, 106250. [CrossRef]
137. Chen, H.Y.; Lee, C.H. Vibration Signals Analysis by Explainable Artificial Intelligence (XAI) Approach: Application on Bearing Faults Diagnosis. *IEEE Access* **2020**, *8*, 134246–134256. [CrossRef]
138. Hong, C.; Lee, C.; Lee, K.; Ko, M.S.; Kim, D.; Hur, K. Remaining Useful Life Prognosis for Turbofan Engine Using Explainable Deep Neural Networks with Dimensionality Reduction. *Sensors* **2020**, *20*, 6626. [CrossRef]
139. Serradilla, O.; Zugasti, E.; Cernuda, C.; Aranburu, A.; de Okariz, J.R.; Zurutuza, U. Interpreting Remaining Useful Life Estimations Combining Explainable Artificial Intelligence and Domain Knowledge in Industrial Machinery. In Proceedings of the 2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Glasgow, UK, 19–24 July 2020; pp. 1–8. [CrossRef]
140. Sun, K.H.; Huh, H.; Tama, B.A.; Lee, S.Y.; Jung, J.H.; Lee, S. Vision-Based Fault Diagnostics Using Explainable Deep Learning With Class Activation Maps. *IEEE Access* **2020**, *8*, 129169–129179. [CrossRef]
141. Assaf, R.; Schumann, A. Explainable Deep Neural Networks for Multivariate Time Series Predictions. In Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI 2019), Macao, 10–16 August 2019; Number 2, pp. 6488–6490. [CrossRef]
142. Sarp, S.; Knzlu, M.; Cali, U.; Elma, O.; Guler, O. An Interpretable Solar Photovoltaic Power Generation Forecasting Approach using an Explainable Artificial Intelligence Tool. In Proceedings of the 2021 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT), Virtual Event, 15–17 September 2021. [CrossRef]
143. Zhang, K.; Zhang, J.; Xu, P.; Gao, T.; Gao, D. Explainable AI in Deep Reinforcement Learning Models for Power System Emergency Control. *IEEE Trans. Comput. Soc. Syst.* **2021**, 1–9. [CrossRef]
144. Rehse, J.R.; Mehdiyev, N.; Fettke, P. Towards Explainable Process Predictions for Industry 4.0 in the DFKI-Smart-Lego-Factory. *KI–Künstliche Intell.* **2019**, *33*, 181–187. [CrossRef]
145. Carletti, M.; Masiero, C.; Beghi, A.; Susto, G.A. Explainable Machine Learning in Industry 4.0: Evaluating Feature Importance in Anomaly Detection to Enable Root Cause Analysis. In Proceedings of the 2019 IEEE International Conference on Systems, Man and Cybernetics (SMC), Bari, Italy, 6–9 Ocotber, 2019; pp. 21–26. [CrossRef]

146. Schönhof, R.; Werner, A.; Elstner, J.; Zopcsak, B.; Awad, R.; Huber, M. Feature Visualization within an Automated Design Assessment Leveraging Explainable Artificial Intelligence Methods. In *Procedia CIRP*; Elsevier B.V.: Amsterdam, The Netherlands, 2021; Volumr 100, pp. 331–336. [CrossRef]
147. Lorente, M.P.S.; Lopez, E.M.; Florez, L.A.; Espino, A.L.; Martínez, J.A.I.; de Miguel, A.S. Explaining deep learning-based driver models. *Appl. Sci.* **2021**, *11*, 3321. [CrossRef]
148. Li, Y.; Wang, H.; Dang, L.; Nguyen, T.; Han, D.; Lee, A.; Jang, I.; Moon, H. A Deep Learning-based Hybrid Framework for Object Detection and Recognition in Autonomous Driving. *IEEE Access* **2020**, *8*, 194228–194239. [CrossRef]
149. Martinez-Cebrian, J.; Fernandez-Torres, M.A.; Diaz-De-Maria, F. Interpretable Global-Local Dynamics for the Prediction of Eye Fixations in Autonomous Driving Scenarios. *IEEE Access* **2020**, *8*, 217068–217085. [CrossRef]
150. Ponn, T.; Kröger, T.; Diermeyer, F. Identification and Explanation of Challenging Conditions for Camera-Based Object Detection of Automated Vehicles. *Sensors* **2020**, *20*, 3699. [CrossRef] [PubMed]
151. Nowak, T.; Nowicki, M.R.; Cwian, K.; Skrzypczynski, P. How to Improve Object Detection in a Driver Assistance System Applying Explainable Deep Learning. In Proceedings of the 30th IEEE Intelligent Vehicles Symposium (IV19), Paris, France, 9–12 June 2019; pp. 226–231. [CrossRef]
152. Kim, J.; Canny, J. Interpretable Learning for Self-Driving Cars by Visualizing Causal Attention. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22-29 October 2017; pp. 2961–2969.
153. Amparore, E.; Perotti, A.; Bajardi, P. To trust or not to trust an explanation: Using LEAF to evaluate local linear XAI methods. *PeerJ Comput. Sci.* **2021**, *7*, 1–26. [CrossRef] [PubMed]
154. van der Waa, J.; Nieuwburg, E.; Cremers, A.; Neerincx, M. Evaluating XAI: A comparison of rule-based and example-based explanations. *Artif. Intell.* **2021**, *291*, 03404. [CrossRef]
155. Sokol, K.; Flach, P. Explainability Fact Sheets: A Framework for Systematic Assessment of Explainable Approaches. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* 2020), Barcelona, Spain, 27–30 January 2020; ACM: New York, NY, USA, 2020; pp. 56–67.
156. Weber, R.O.; Johs, A.J.; Li, J.; Huang, K. Investigating Textual Case-Based XAI. In *Case-Based Reasoning Research and Development*; Cox, M.T., Funk, P., Begum, S., Eds.; Springer International Publishing: Berlin/Heidelberg, Germany, 2018; Volume 11156 LNAI, pp. 431–447._29. [CrossRef]
157. Rutkowski, T.; Łapa, K.; Nielek, R. On Explainable Fuzzy Recommenders and their Performance Evaluation. *Int. J. Appl. Math. Comput. Sci.* **2019**, *29*, 595–610. [CrossRef]
158. 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, (AAAI-19), Honolulu, HI, USA, 27 January – 1 February 2019; Volume 33, pp. 5329–5336.
159. Zhao, G.; Fu, H.; Song, R.; Sakai, T.; Chen, Z.; Xie, X.; Qian, X. Personalized Reason Generation for Explainable Song Recommendation. *ACM Trans. Intell. Syst. Technol.* **2019**, *10*, 1–21. [CrossRef]
160. Han, M.; Kim, J. Joint Banknote Recognition and Counterfeit Detection Using Explainable Artificial Intelligence. *Sensors* **2019**, *19*, 3607. [CrossRef]
161. Chen, J.H.; Chen, S.Y.C.; Tsai, Y.C.; Shur, C.S. Explainable Deep Convolutional Candlestick Learner. In Proceedings of the Thirty Second International Conference on Software Engineering and Knowledge Engineering (SEKE 2020), Pittsburgh, PA, USA, 9–11 July 2020; Volume PartF16244, pp. 234–237.
162. He, X.; Chen, T.; Kan, M.Y.; Chen, X. TriRank: Review-aware Explainable Recommendation by Modeling Aspects. In Proceedings of the 24th ACM International Conference on Information and Knowledge Management (CIKM '15), Melbourne, Australia, 18–23 October 2015; ACM Press: New York, NY, USA, 2015; pp. 1661–1670. [CrossRef]
163. Loyola-González, O. Understanding the Criminal Behavior in Mexico City through an Explainable Artificial Intelligence Model. In *Advances in Soft Computing*; Martínez-Villaseñor, L., Batyrshin, I., Marín-Hernández, A., Eds.; Springer: Cham, Switzerland, 2019; Volume 11835, pp. 136–149._12. [CrossRef]
164. Zhong, Q.; Fan, X.; Luo, X.; Toni, F. An Explainable Multi-attribute Decision Model based on Argumentation. *Expert Syst. Appl.* **2019**, *117*, 42–61. [CrossRef]
165. Vlek, C.S.; Prakken, H.; Renooij, S.; Verheij, B. A Method for Explaining Bayesian Networks for Legal Evidence with Scenarios. *Artif. Intell. Law* **2016**, *24*, 285–324. [CrossRef]
166. Bonidia, R.; MacHida, J.; Negri, T.; Alves, W.; Kashiwabara, A.; Domingues, D.; De Carvalho, A.; Paschoal, A.; Sanches, D. A Novel Decomposing Model with Evolutionary Algorithms for Feature Selection in Long Non-coding RNAs. *IEEE Access* **2020**, *8*, 181683–181697. [CrossRef]
167. Huang, L.C.; Yeung, W.; Wang, Y.; Cheng, H.; Venkat, A.; Li, S.; Ma, P.; Rasheed, K.; Kannan, N. Quantitative Structure–Mutation–Activity Relationship Tests (QSMART) Model for Protein Kinase Inhibitor Response Prediction. *BMC Bioinform.* **2020**, *21*, 520. [CrossRef] [PubMed]
168. Anguita-Ruiz, A.; Segura-Delgado, A.; Alcalá, R.; Aguilera, C.M.; Alcalá-Fdez, J. eXplainable Artificial Intelligence (XAI) for the Identification of Biologically Relevant Gene Expression Patterns in Longitudinal Human Studies, Insights from Obesity Research. *PLoS Comput. Biol.* **2020**, *16*, e1007792. [CrossRef] [PubMed]
169. Keneni, B.M.; Kaur, D.; Al Bataineh, A.; Devabhaktuni, V.K.; Javaid, A.Y.; Zaientz, J.D.; Marinier, R.P. Evolving Rule-Based Explainable Artificial Intelligence for Unmanned Aerial Vehicles. *IEEE Access* **2019**, *7*, 17001–17016. [CrossRef]

170. Ten Zeldam, S.; De Jong, A.; Loendersloot, R.; Tinga, T.; ten Zeldam, S.; de Jong, A.; Loendersloot, R.; Tinga, T. Automated Failure Diagnosis in Aviation Maintenance Using Explainable Artificial Intelligence (XAI). In Proceedings of the 4th European Conference of the PHM Society (PHME 2018), Utrecht, Netherlands, 3–6 July 2018; pp. 1–11.
171. Eisenstadt, V.; Espinoza-Stapelfeld, C.; Mikyas, A.; Althoff, K.D. Explainable Distributed Case-Based Support Systems: Patterns for Enhancement and Validation of Design Recommendations. In *Case-Based Reasoning Research and Development*; Cox, M.T., Funk, P., Begum, S., Eds.; Springer International Publishing: Berlin/Heidelberg, Germany, 2018; Volume 11156 LNAI, pp. 78–94._6. [CrossRef]
172. Anysz, H.; Brzozowski, Ł.; Kretowicz, W.; Narloch, P. Feature Importance of Stabilised Rammed Earth Components Affecting the Compressive Strength Calculated with Explainable Artificial Intelligence Tools. *Materials* **2020**, *13*, 2317. [CrossRef]
173. Díaz-Rodríguez, N.; Pisoni, G. Accessible Cultural Heritage through Explainable Artificial Intelligence. In Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization (UMAP 2020), Genoa, Italy, 12-18 July 2020; ACM: New York, NY, USA, 2020; pp. 317–324. [CrossRef]
174. Van Lent, M.; Fisher, W.; Mancuso, M. An Explainable Artificial Intelligence System for Small-unit Tactical Behavior. In Proceedings of the National Conference on Artificial Intelligence, San Jose, CA, USA, 25–29 July 2004; pp. 900–907.
175. Segura, V.; Brandão, B.; Fucs, A.; Vital Brazil, E. Towards Explainable AI Using Similarity: An Analogues Visualization System. In *Design, User Experience, and Usability. User Experience in Advanced Technological Environments*; Marcus, A., Wang, W., Eds.; Springer Nature Switzerland: Orlando, FL, USA, 2019; pp. 389–399._28. [CrossRef]
176. Callegari, C.; Ducange, P.; Fazzolari, M.; Vecchio, M. Explainable internet traffic classification. *Appl. Sci.* **2021**, *11*, 4697. [CrossRef]
177. Sarathy, N.; Alsawwaf, M.; Chaczko, Z. Investigation of an Innovative Approach for Identifying Human Face-Profile Using Explainable Artificial Intelligence. In Proceedings of the 18th IEEE International Symposium on Intelligent Systems and Informatics (SISY 2020), Subotica, Serbia, 17–19 September 2020; IEEE: Subotica, Serbia, 2020; pp. 155–160. [CrossRef]
178. Ferreyra, E.; Hagras, H.; Kern, M.; Owusu, G. Depicting Decision-Making: A Type-2 Fuzzy Logic Based Explainable Artificial Intelligence System for Goal-Driven Simulation in the Workforce Allocation Domain. In Proceedings of the 2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), New Orleans, LA, USA, 23-26 June 2019; pp. 1–6. [CrossRef]
179. Kovalev, M.S.; Utkin, L.V.; Kasimov, E.M. SurvLIME: A Method for Explaining Machine Learning Survival Models. *Knowl.-Based Syst.* **2020**, *203*, 106164. [CrossRef]
180. Albaum, G. The Likert Scale Revisited. *Mark. Res. Soc J.* **1997**, *39*, 1–21. [CrossRef]
181. Samek, W.; Binder, A.; Montavon, G.; Lapuschkin, S.; Müller, K.R. Evaluating the Visualization of What a Deep Neural Network Has Learned. *IEEE Trans. Neural Netw. Learn. Syst.* **2017**, *28*, 2660–2673. [CrossRef]
182. Spinner, T.; Schlegel, U.; Schafer, H.; El-Assady, M. explAIner: A Visual Analytics Framework for Interactive and Explainable Machine Learning. *IEEE Trans. Vis. Comput. Graph.* **2019**, *26*, 1064–1074. [CrossRef]
183. Forestiero, A. Metaheuristic Algorithm for Anomaly Detection in Internet of Things leveraging on a Neural-driven Multiagent System. *Knowl.-Based Syst.* **2021**, *228*, 107241. [CrossRef]
184. Forestiero, A.; Mastroianni, C.; Spezzano, G. Reorganization and Discovery of Grid Information with Epidemic Tuning. *Future Gener. Comput. Syst.* **2008**, *24*, 788–797. [CrossRef]
185. Forestiero, A.; Papuzzo, G. Agents-Based Algorithm for a Distributed Information System in Internet of Things. *IEEE Internet Things J.* **2021**, *8*, 16548–16558. [CrossRef]