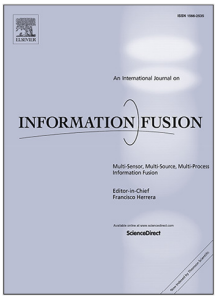
Journal Pre-proof



Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence

Sajid Ali, Tamer Abuhmed, Shaker El-Sappagh, Khan Muhammad, Jose M. Alonso-Moral, Roberto Confalonieri, Riccardo Guidotti, Javier Del Ser, Natalia Díaz-Rodríguez, Francisco Herrera

PII: S1566-2535(23)00114-8
DOI: https://doi.org/10.1016/j.inffus.2023.101805
Reference: INFFUS 101805

To appear in: *Information Fusion*

Received date : 19 January 2023
Revised date : 23 March 2023
Accepted date : 6 April 2023

Please cite this article as: S. Ali, T. Abuhmed, S. El-Sappagh et al., Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence, *Information Fusion* (2023), doi: https://doi.org/10.1016/j.inffus.2023.101805.

Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence

Sajid Ali[a], Tamer Abuhmed[b,*], Shaker El-Sappagh[b], Khan Muhammad[c,*], Jose M. Alonso-Moral[d], Roberto Confalonieri[e], Riccardo Guidotti[f], Javier Del Ser[g,h], Natalia Díaz-Rodríguez[i] and Francisco Herrera[i]

[a]*Information Laboratory (InfoLab), Department of Electrical and Computer Engineering, College of Information and Communication Engineering, Sungkyunkwan University, Suwon 16419, South Korea*

[b]*Information Laboratory (InfoLab), Department of Computer Science and Engineering, College of Computing and Informatics, Sungkyunkwan University, Suwon 16419, South Korea*

[c]*Visual Analytics for Knowledge Laboratory (VIS2KNOW Lab), Department of Applied Artificial Intelligence, College of Computing and Informatics, Sungkyunkwan University, Seoul 03063, South Korea*

[d]*Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela, Rúa de Jenaro de la Fuente Domínguez, s/n, 15782 Santiago de Compostela, A Coruña, Spain*

[e]*Department of Mathematics, University of Padua, Padova 35100, Italy*

[f]*Knowledge Discovery and Data Mining Laboratory (KDDLab), Department of Computer Science, University of Pisa, Pisa 56126, Italy*

[g]*TECNALIA, Basque Research and Technology Alliance (BRTA), 48160 Derio, Spain*

[h]*Department of Communications Engineering, University of the Basque Country (UPV/EHU), 48013 Bilbao, Spain*

[i]*Department of Computer Science and Artificial Intelligence, University of Granada, Granada 18071, Spain*

ARTICLE INFO

Keywords:
Explainable Artificial Intelligence
Interpretable Machine Learning
Trustworthy AI
AI principles
Post-hoc explainability
XAI assessment
Data Fusion
Deep Learning

ABSTRACT

Artificial intelligence (AI) is currently being utilized in a wide range of sophisticated applications, but the outcomes of many AI models are challenging to comprehend and trust due to their black-box nature. Usually, it is essential to understand the reasoning behind an AI model's decision-making. Thus, the need for eXplainable AI (XAI) methods for improving trust in AI models has arisen. XAI has become a popular research subject within the AI field in recent years. Existing survey papers have tackled the concepts of XAI, its general terms, and post-hoc explainability methods but there have not been any reviews that have looked at the assessment methods, available tools, XAI datasets, and so on. Therefore, in this comprehensive study, we provide readers with an overview of the current research and trends in this rapidly emerging area with a case study example. The review starts by explaining the background of XAI, common definitions, and summarizing recently proposed techniques in XAI for supervised machine learning. The review divides XAI techniques into four axes using a hierarchical categorization system: (i) data explainability, (ii) model explainability, (iii) post-hoc explainability, and (iv) assessment of explanations. We also introduce available evaluation metrics as well as open-source packages and datasets with future research directions. Then, the significance of explainability in terms of legal demands, user viewpoints, and application orientation is outlined, termed as XAI concerns. This paper advocates for tailoring explanation content to specific user types. An examination of XAI techniques and evaluation was conducted by looking at 410 critical articles, published between January 2016 and October 2022, in reputed journals and using a wide range of research databases as a source of information. The article is aimed at XAI researchers who are interested in making their AI models more trustworthy, as well as towards researchers from other disciplines who are looking for effective XAI methods to complete tasks with confidence while communicating meaning from data.

1. Introduction

Artificial Intelligence (AI) has become ingrained in our society as it assists various sectors in dealing with difficult issues and reforming outdated methods. AI models run in people's smartphones to do various tasks [1], in cars to avoid accidents [2], in banks to manage investment and loan decisions [3, 4], in hospitals to aid doctors diagnosing and detecting disease [5], in law enforcement to help officials recover evidence and make law enforcement easier [6], in the military of many countries [7], in insurance organizations to determine risk [8], etc. Moreover, many organizations are actively trying to integrate AI into their workflows due to its remarkable performance, which competes with human performance in a wide variety of tasks [9].

AI enables data-driven decision-making systems. In other words, a tremendous quantity of data is required to produce an accurate AI model. Primitive Machine learning (ML) models, such as linear regression, logistic regression, and Decision tree (DT), are less accurate due to the assumption of smooth linear/sub-linear data [10]. However, real-world data is highly non-linear and complex, this makes processing it to gain knowledge and insights a real challenge. Under these circumstances, Deep neural networks (DNNs) are exploited to extract information from highly complex datasets [11]. After using DNNs, scientists have realized that a deeper network is better for decision-making than a shallow network [12]. Moreover, to extract patterns from

*Corresponding authors at: College of Computing and Informatics, Sungkyunkwan University
E-mail addresses: tamer@skku.edu (Tamer Abuhmed), khanmuhammad@g.skku.edu (Khan Muhammad)

this kind of complex data, a sophisticated DNN must be trained on a large dataset. A collection of convolutional filters/kernels is used to cover all the differences that come from the non-linearity of the real-world data, this leads to high-performance AI models. However, by increasing the number of filters an AI model uses, strain is put on subsequent layers of the DNN. Thus, even a basic network may have several layers, with many filters, and neuronal units. DNNs for complex tasks often have millions or even billions of parameters. The underlying representations and data flow across the network's layers are difficult to examine, while the number of learnable variables increases as the networks' designs become increasingly complex [13].

Furthermore, the structural design of a DNN model is influenced by a number of factors, including the activation function, input type and size, number of layers, pooling operation, connectivity pattern, classifier mechanisms, and the results of compound learning techniques. The learning technique is further influenced by a number of additional functions, such as normalization/regularization, weight updating mechanisms, cost/loss functions, and the type of end classifier used. As a result, unlike other ML techniques such as DT, Fuzzy rule-based systems (FRBSs), or Bayesian networks (BNs), a decision from a DNN is difficult to comprehend and trust. Due to these hurdles, there is a problem, we are left with the aforementioned **black-box conundrum** [14]. The simpler ML models, such as DT, are easier to comprehend and self-interpret. In this case, interpret means to provide an explanation for the system's decisions or to portray them in a logical/reasonable manner [15]. Unlike in black-box systems, in the context of AI, a person may comprehend simpler ML models by glancing at the summary or parameters of the model without the need for an external model to provide an explanation. We refer as a *white-box* or a *glass-box* model. In the research community, there is also a concept known as the *gray-box* model which applies for example to FRBSs [16] or BNs [17], which are models that users can interpret at some degree if they are carefully designed. As it can be seen in Figure 1, the labels white-box, gray-box, and black-box refer to various levels of the internal component [18]. The following paragraph will go through more in-depth descriptions and solutions for the black-box problem.

1.1. The Black-box Issue and Solution

The AI community is more concerned about the black-box issue following the establishment of rules for trustworthy AIs that are safe to use. eXplainable Artificial Intelligence (XAI) techniques are aimed at producing ML models with a good interpretability-accuracy tradeoff via: (i) building white/gray-box ML models which are interpretable by design (at least at some degree) while achieving high accuracy or (ii) endowing black-box models with a minimum level of interpretability when white/gray-box models are not able to achieve an admissible level of accuracy. XAI techniques play a crucial role when dealing with DNN models and how to make their results comprehensible to humans [19].

Furthermore, there are two terminologies by which we can try to elucidate a DNN model: (i) interpretability and (ii) explainability. **Interpretability** enables developers to delve into the model's decision-making process, boosting their confidence in understanding where the model gets its results. Instead of a simple prediction, the *interpretation* technique provides an interface that gives additional information or explanations that are essential for interpreting an AI system's underlying functioning [20]. It aids in opening a door into the black-box model for users with the required knowledge and skills, e.g, developers. On the other hand, **explainability** provides insight into the DNN's decision to the end-user in order to build trust that the AI is making correct and non-biased decisions based on facts. Figure 1 depicts the distinction between white-box, gray-box, and black-box decision-making processes, as well as shows how XAI is applied to achieve a trustworthy model with a good interpretability-accuracy tradeoff.

The ML approaches, which include different mathematical methods for extracting and exploiting important information from huge collections of data, from a technical point of view are now dominated by AI models. The goal of XAI research is to make AI systems more comprehensible and transparent to humans without sacrificing performance [21, 20]. The ability to understand patterns hidden in complex data is both a strength and a weakness of automated decision-making systems: an AI model may discover complex structures in the data automatically, but the learned patterns are hidden knowledge without explicit rules or logical processes involved in finding them. Although AI algorithms are capable of extracting correlations across a wide range of complicated data, there is no assurance that these correlations are relevant or relate to real causal connections. Furthermore, the intricacy of the models used, particularly the cutting-edge DNNs, often hinders human operators from inspecting and controlling them in a straightforward manner. In this way, AI is both a source of innovation and one significant problem in terms of security, safety, privacy, and transparency. In the next paragraph, we will go through the entire list of goals behind XAI.

1.2. The Goal of XAI

The primary goal of XAI is to obtain human-interpretable models, especially for applications in sensitive sectors such as military, banking, and healthcare applications, since domain specialists need help solving problems more effectively, but they also want to be provided with meaningful output to understand and trust those solutions. It is not only beneficial for domain specialists to examine appropriate outputs, but it is also beneficial for developers if the outputs turn out to be incorrect as it prompts them to investigate the system. AI methods enable (i) the assessment of current knowledge, (ii) the advancement of knowledge, and (iii) the evolution of new assumptions/theories [22]. In addition, the goals behind XAI methods that researchers would like to

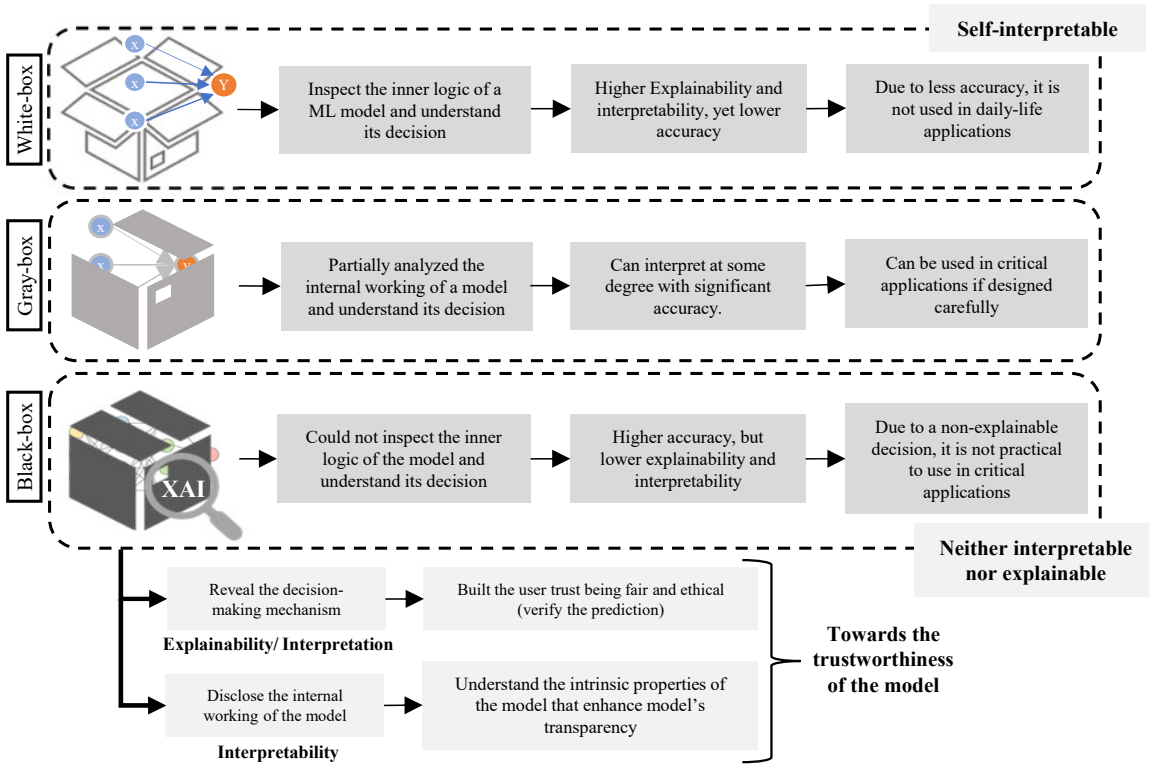


Figure 1: A comparison of white-box, gray-box, and black-box models. On the one hand, white-box models are interpretable by design thus making their outputs easier to understand but less accurate. In addition, gray-box models yield a good interpretability-accuracy tradeoff. On the other hand, black-box models are more accurate but less interpretable. More complex XAI techniques are required for creating trustworthy models.

accomplish with explainability are enhanced justification, control, improvement, and discovery [21]. The following list summarizes the benefits by opening a window into these black-box systems [18]:

- **To empower** individuals to combat any negative consequences of automated decision-making.
- **To assist** individuals in making more informed choices.
- **To expose** and protect security vulnerabilities.
- **To integrate** algorithms with human values is an important goal.
- **To enhance** industry standards for the development of AI-powered products, thus improving consumer and business confidence.
- **To enforce** the *Right of Explanation* policy.

For a model to be embraced by end-users and industries, it must be trustworthy [23]. Building a reliable model, however, is difficult. A few of the factors that contribute to the model's trustworthiness include fairness [24], robustness [25], interpretability [26], and explainability/interpretation [27]. Explainability is one of the most crucial aspects. Existing studies have focused solely on providing better explanations and insights for future research. Researchers have proposed different strategies to explain AI models qualitatively using comprehensible text [28], mathematics [29], or visualizations [15]. In the following subsection, we will discuss our motivation for conducting this study.

1.3. Motivation

Most existing research on XAI focuses on providing a comprehensive overview of approaches for either explaining black-box models or designing white-box models, as well as looking at general reasons why explainability is important. Some research concentrates on specific issues such as notions of explainability and interpretability, their benefits and drawbacks, and the necessity for explainability in critical fields like healthcare, banking, the military, etc.

It becomes essential to explain AI models' decisions once government regulations have been enforced. The field of XAI has evolved to comprehend AI systems better and is helping us move towards systems that can provide human-friendly explanations. However, no previous research has examined whether the availability of an ever-expanding range of methodologies and tools is sufficient for the XAI research field to crystallize and give practical support in the risky scenarios described by regulatory stakeholders. For example, does score-CAM [30] or Grad-CAM [31] guarantee that a DNN may be utilized for medical diagnosis? The answer

is **NO** since supervisory agencies have not prescribed risk-aware scenarios that may assist the research community in determining what is needed to implement XAI-supported AI-based models in real-world contexts. Therefore, society requires techniques in which XAI tools are an essential but insufficient step in determining whether or not an AI-based system can be trusted and employed for the task at hand.

We discuss more comprehensive XAI definitions and generally accepted terminologies in this study. In addition, we break down the XAI worries into three distinct perspectives: (i) user, (ii) application, and (iii) government. We focus on approaches for analyzing the four axes that make up the purposes behind an explanation in order to help us evaluate the results from intelligent systems more effectively. These four axes are: (i) data explainability, (ii) model explainability, (iii) post-hoc explainability, and (iv) assessment of explanations. Questions such as "What constitutes an acceptable explanation?" and "how to establish user trust in AI-powered systems?" are still unresolved. We also examined responsible principles in terms of *fairness*, *security*, *accountability*, *ethics*, and *privacy* in order to improve user trust in XAI.

The novel contributions of this paper can be summarized as follows:

- Explainability may be used at any point throughout the AI development process. We propose a four-axes methodology to diagnose the training process and to refine the model for robustness and trustworthiness. These four axes are: (i) data explainability, (ii) model explainability, (iii) post-hoc explainability, and (iv) assessment of explanations. We believe that explanations should be created by considering each axis in terms of a typical AI pipeline.

- Since our methodology has four axes, we formulated research questions for each axis and will address them in the following section. In addition, we introduce a taxonomy for each axis and discuss various techniques, including a case study example of a basic supervised binary classification task in which a model is created to distinguish whether employees earn an annual income over 50K.

- Furthermore, we present a comparison between different post-hoc methods including a discussion of their advantages, disadvantages, and underlying principles. A mathematical model and a simplified visualization of its working process are used to demonstrate each post-hoc technique.

- We propose another methodology to provide a roadmap on how to determine a given model and its explainability criteria. A list of XAI tools and open-source datasets for developers and end-users is also presented. We provide a summary of each tool in terms of the data types that are allowed, its explainability and the explanations that are offered, model type, and evaluation matrix.

- The XAI concept is defined using background research followed by a commonly accepted definition of a good explanation, explainability, and associated terms being given. We look at the explainability concept from three main points of view: (i) regulatory entities, (ii) various stakeholders and decision-makers, and (iii) combat applications. As a result, we suggest that explanations should be created with the kind of user and evaluation criterion in mind.

- XAI researchers are currently developing new techniques and tools for the exploration, debugging, and validation of AI models. Based on the literature, we highlight and discuss XAI's open challenges and future directions in terms of (i) XAI system design, (ii) generalization of XAI, (iii) user interactions with XAI, (iv) XAI ground truth evaluation, and (v) advanced XAI tools.

Organization: The outline of the article is as follows: Section 2 looks at previous XAI literature and related surveys. Section 3 begins with the XAI concepts, a set of novel definitions, and the balance between accuracy and interpretability. In section 4, a potential XAI model is discussed, along with the questions that may be addressed along each of the explainability axes. The general classification of XAI methods is enumerated in Section 5. In Sections 6, 7, and 8, a major part of the proposed taxonomies is discussed in terms of data explainability, model explainability, and post-hoc explainability, respectively. In Section 9, the techniques and metrics for assessing explanations of XAI algorithms are presented. The question of selecting an XAI model according to research direction is addressed in Section 10. Step-by-step guidance for future researchers starting in the emerging field of XAI is offered in Section 11. This is an important guide to a research area that has the potential to influence society, particularly those industries that have gradually adopted AI as one of their core technologies. The significance of XAI from the different stakeholders, government restrictions and policies, and application perspectives are discussed further in Section 12. Section 13 concludes our survey. Figure 2 depicts the organization of our survey tow help readers navigate through its content more easily.

2. Background Studies

Research interest in the field of XAI is resurgent. In 2019, Mueller et al. [32] published a systematic analysis of XAI's methods and explanation systems, these were classified into three generations: (i) First-generation systems attempted to describe the system's internal working process explicitly by integrating expert knowledge into the rules via transforming these rules into natural language expressions such as those used in expert systems from the early 1970s, (ii) Second-generation systems are human-computer systems that provide cognitive assistance by focusing on human knowledge and reasoning abilities from the early 2000s, and (iii) Third-generation systems seek to clarify the inner workings of the systems in the same way as the first generation. However, the third-generation systems became mostly black-box systems from about 2012. Due to improved computer technology,

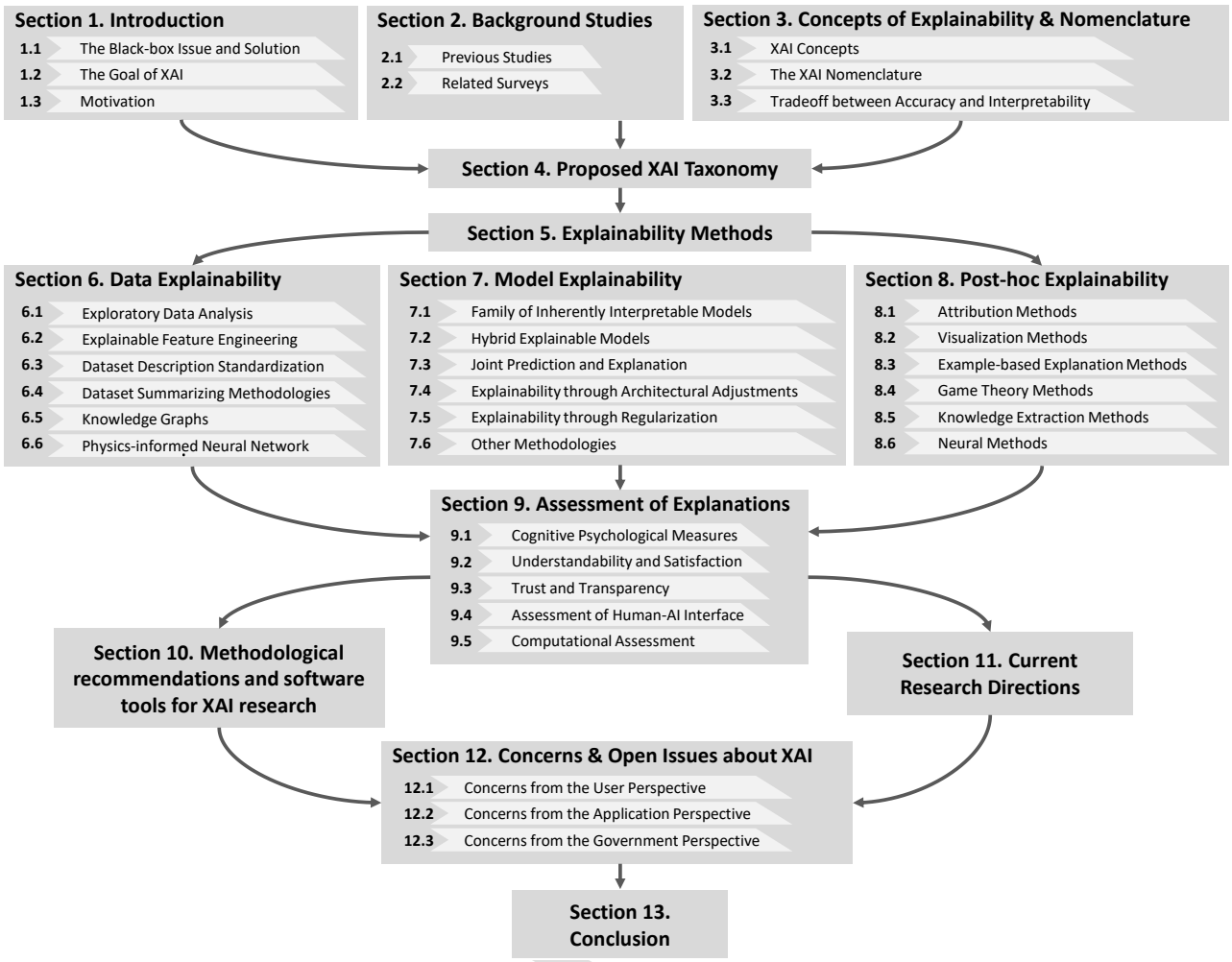


Figure 2: Detailed overview of the different sections and topics covered in the survey for easing its readability.

several novel concepts for producing explainable choices have become more feasible. These ideas have arisen from the need for mainly responsible, fair, and trustworthy processes and decisions. The three generations of intelligent systems will be discussed in detail in the following paragraphs.

First-Generation Systems. Researchers have been interested in understanding the underlying workings of AI since the early stages of AI systems. Chandrasekaran et al. [33], and Swartout et al. [34, 35] were among the first to describe the decision-making process of knowledge-based and expert systems. Expert systems but also Fuzzy Sets and System (FSS) [36, 37, 38] in the 1970s, Artificial Neural Networks (ANN) [39] and BNs [40] in the 1980s, as well as recommendation systems [41, 42] in the 2000s have all sparked interest in intelligent system explanations. Despite their mathematical correctness, these earlier works are inaccurate. However, they inspired subsequent research into understanding intelligent systems.

Second-Generation Systems. This generation saw a more powerful AI system being built. However, the models that were built became complex in nature. The decision-making process of the AI systems was one that humans, including domain experts, did not fully comprehend when it came to powerful ML classification models trained on large datasets with high-performance infrastructures [43]. Another risk associated with these techniques is that they may unintentionally make incorrect decisions as a result of biased artifacts or false correlations in the data. This is a particularly essential issue when it comes to using these systems in high-risk applications like self-driving vehicles and medicine, where a single incorrect decision may result in a person's death [44].

Third-Generation Systems. The current advances in AI, its application to diverse fields, concerns about unethical usage [45], lack of transparency [46], and unintended biases [47] in the models are the main reasons for the increased interest in XAI research. This has an effect not just on the amount of information that can guide ethical decisions but also on the accountability, safety, and industrial liabilities of these XAI systems. Furthermore, new regulations enacted

by various countries mean an even greater need for XAI research to establish trust in AI models [48]. The AI models used in sensitive areas of scientific research, such as in health, biology, and socioeconomic sciences, need to be able to provide an explanation of their results to allow scientific discovery and advances in research.

Existing XAI work can be categorized in various ways, including XAI applications, multidisciplinary method fusion, and explainability by internal functionality modification, among others. The following subsection goes through the reviewed literature.

2.1. Previous Studies

XAI has the potential to be extremely beneficial to the AI research community. This subsection explores several prospective areas of research where explainable models are currently being used.

Applications of XAI. Meske et al. [49] described the theoretical impact of explainability on AI trust and how XAI may be utilized in a medical setting. Their work used CNN and Multi-layer Perceptron (MLP) models to identify a sickness (malaria) from the input image data (thin blood smear slide images). In addition, the necessity for explainability, and prior research in XAI for information systems highlighted some of the risks of black-box AI [50]. Islam et al. [51] illustrated common XAI techniques using credit default prediction as the subject of a case study, the results were evaluated in terms of gaining a competitive advantage from both local and global perspectives, offering significant insight on how to assess explainability, and recommending routes toward responsible or human-centered AI.

Social Science and Argumentation in XAI. Miller [52] supported XAI while including articles from the social sciences. The author discusses how XAI incorporates ideas from philosophy, cognitive science, and social psychology in order to produce good explanations of its results. Similarly, when it comes to psychological theories of explanation, T. Mueller [32] first stresses *what a good explanation is*. Furthermore, argumentation and XAI have a lot in common in terms of explainability. Vassiliades et al. [53] examined the major techniques and research on the linked subjects of argumentation and XAI. The authors explored more interpretable prediction models that integrate ML and argumentation theory. Humans have also been involved in assessing XAI systems. Hussain et al. [54] used an engineering approach to illustrate the concepts behind XAI by giving mathematical outlines of the methods used.

Methods for Improving Explanations: Scientists have attempted to decipher the inner workings of black-box systems and create transparent counterparts. Liu et al. [55] described an interactive visualization method that aids in the diagnosis, comprehension, and refinement of an AI system and associated data mining issues. Zhang et al. [56] focused on the interpretability of CNNs' middle-layer representations. Ras et al. [57] investigated the explainability of certain systems in terms of dataset bias, which can result in biased models. Montavon et al.[58] provided a brief review of the interpretability problem and its potential applications. The authors also thoroughly discussed the Layer-wise Relevance Propagation (LRP) method.

XAI in Other Learning Methods: Explainability has been used in many studies on supervised ML, unsupervised ML, and Reinforcement Learning (RL) [59]. Puiutta et al. [60] published the first review on Explainable Reinforcement Learning (XRL). The authors provided an overview of the problem and definitions of key terms, they also gave their classification and assessment of certain XRL methods. Burkart and Huber [61] provided an overview of some Explainable Supervised ML (XSML) principles and techniques, as well as discussed other important concepts in the field. The authors focused on supervised learning and offered a taxonomy of interpretable model learning, surrogate models, explanation types, and data explainability. Gerlings et al. [62] identified four thematic arguments (motivating the need for XAI, completeness vs. interpretability dilemma, human explanations, and technologies producing XAI) by conducting a thorough study of the XAI literature on the subject. These arguments are essential to how XAI handles the black-box issue.

Many surveys on XAI have been published previously, these have looked at the necessity of XAI as well as at some related notions, methods, software tools, and challenges. For instance, Arrieta et al. [63] showed that the model's explainability is one of the most important elements to guarantee a system is able to provide good explanations inside its methodological framework. Many other reviews have looked at subjects such as post-hoc explanations [18, 26, 64]. The next subsection will briefly summarize the existing surveys.

2.2. Related Surveys

Despite the fact that the number of studies on XAI is quickly growing (see Table 1), there is still a lack of thorough surveys and a systematic classification of these studies, except [65]. There are numerous review articles on XAI, but the majority of these reviews concentrate on general XAI techniques, their significance, and evaluation approaches. For example, Doshi et al. [15] chart the path toward the definition and rigorous evaluation of interpretability. Their main contribution is a taxonomy for assessing interpretability. Consequently, the authors focused on just one element of explainability, i.e., interpretability and its related evaluation techniques. Abdul et al. [66] built a citation network from a vast corpus of explainable research based on 289 core articles and 12412 citing publications. However, their review was mainly concerned with Human-Computer Interface (HCI) research that emphasizes explainability. Adadi et al. [21] attempted to offer information on the idea, motives, and consequences that underpin XAI study in order to understand the important topics in XAI.

Furthermore, Guidotti et al. [18] investigated a variety of approaches to explaining large-scale black-box models, including looking at data mining and ML techniques involved. The authors presented a comprehensive taxonomy of explainability methods for systems that suffer from the

black-box problem. Their work comprehensively assessed ML models in terms of XAI; however, it only focused on the interpretability processes, leaving out other elements of explainability like evaluation. Consequently, despite a comprehensive technical overview of the approaches under consideration it was difficult to gain a general understanding of the XAI immediately. Samek et al. [67] described two methods for interpreting a model's output. In their approach, the sensitivity of the output is first calculated in relation to changes in the input. The second step is to break down the output decision into its input variables. Dosilovic et al. [64] highlighted recent developments in XAI to provide a fair comparison between interpretability and explainability in supervised ML. Lipton [28] defined model properties and techniques, as well as the notion of interpretability for supervised ML in terms of identity transparency to humans and of post-hoc explanations.

In recent years, Carvalho et al. [68] examined the interpretability of ML with a focus on the established techniques and metrics. Vilone et al. [69] divided popular XAI methods into four categories: review articles, theories and concepts, methodologies, and evaluation. Arrieta et al. [63] discussed the achievements of XAI in terms of effort and contributions. Two taxonomic approaches to explainability are discussed in their review: (i) ML model transparency, and (ii) post-hoc explainability. Linardatos et al. [70] carried out research that focused on ML interpretability techniques, particularly, literature analysis and the taxonomy of interpretability methods, as well as looked at links to programming implementations. In addition, Li et al. [26] described and defined two key XAI concepts: interpretations and interpretability. The authors used a novel taxonomy to describe the architecture of several interpretation algorithms and they also highlighted some interpretation research initiatives. In addition to simply attempting to comprehend any interpretation results, they went beyond to examine certain performance metrics for evaluating the interpretation algorithms. Langer et al. [71] looked at the main stakeholder groups that seek AI explainability, as well as their needs. Confalonieri et al. [72] provided a historical perspective of XAI, where they analyzed how the notion of explainability evolved from expert systems to machine learning and recommender systems, until neuro-symbolic AI.

To sum up, Table 1 provides a summary of existing review articles, and we can draw two main conclusions. First, the majority of the surveys addressed the research trend, the core concepts of (and terms related to) XAI, their concerns, and post-hoc explainability. Despite the fact that many researchers have concentrated on XAI concerns and terminologies associated with it, there are still evolving government regulations to impose explainability, as well as unacceptable and inconsistent definitions by the XAI community. Second, numerous researchers have identified three significant and ongoing challenges with XAI: (i) lack of XAI tools, (ii) different axes or dimensions of explainability, and (iii) highlight the need to take care seriously of XAI evaluation (both automatic metrics and human evaluation) in future directions.

3. Concepts of Explainability and Important Nomenclature

AI is a powerful technology with a wide range of applications. AI has attained great accuracy not just as a result of improved hardware performance, but also as a result of employing more sophisticated algorithms, such as those employed in cutting-edge DL methods. Due to the complex nature of the algorithms used, these modern AI systems are unable to explain their decisions in a straightforward manner, limiting their practical applicability [84]. As a result, AI must tackle this black-box issue, even if the developers have to sacrifice performance. The necessity to explain AI and encourage its adoption by many stakeholders has inspired the creation of XAI as a new field of research. This section is organized as follows: (i) the concept behind XAI is defined via background research, (ii) associated XAI terminology is explained, and (iii) the trade-off between accuracy and interpretability is explored.

3.1. XAI Concepts

Van Lent et al. [85] created the XAI concept in 2004 to characterize their system's capacity to explain the actions of AI-controlled units in simulation gaming applications. Academics and practitioners have recently rekindled their interest in the subject of XAI [86, 87]. Several research groups have investigated the notion of explainability in AI decision-making. Each research community, however, approaches the issue from a different angle and gives explanations with various meanings. The word *explainability*, in terms of AI concepts, means functional knowledge of the model, the purpose of which is to attempt to describe the model's black-box behavior [28]. It is often used interchangeably with the term *interpretability* in the literature. Explainability expresses what is occurring in the model by providing a human-readable explanation of the model's decision. However, it is difficult to come up with a precise description of what qualifies as an explanation. The following are some of the most widely recognized definitions of an explanation [53]:

- An attribution of causal responsibility is referred to as an explanation [88].
- An explanation is an act of describing something and providing the response to the question of why this description of something is correct [89].
- A process for finding or creating common meaning is known as an explanation [90].

A more recent widely embraced definition of explainable IA is the one given in [63], where the focus is on the receiver of the explanation: given an audience, an explainable Artificial Intelligence is one that produces details or reasons to make its functioning clear or easy to

Table 1
Details on existing surveys related to XAI and this review especially. The surveys in this table were considered for their significance to the main theme, published year, publisher reputation, and citation support from the relevant research community.

Reference	Published Year	Number of Reviewed Papers	Literature Coverage Range	Analyzed Research Trend	Existing Surveys are Reviewed	Concepts of XAI	Terms Related to XAI	Concern of XAI	Experimental Analysis	XAI Tools	Detailed Recommendation of Future Direction	Data Explainability	Model Explainability	Post-hoc Explainability	Assessment Methods	Main Theme
Ours	2022	410	2016-2022	■	■	■	■	■	■	■	■	■	■	■	■	Model's trustworthiness
[73]	2022	70	2006-2021	■		■							■	■		Natural Language Explanations
[74]	2022	53	2015-2020	■		■					■					Knowledge based XAI
[75]	2022	165	2016-2019	■	■	■	■									Counterfactual based XAI
[76]	2022	71	2016-2021	■									■	■		Introduction to XAI
[77]	2022	190	2017-2022	■	■		■		■					■	■	Counterfactual explanations
[78]	2022	182	2018-2022	■	■	■					■			■	■	XAI for time series
[79]	2022	168	2018-2021	■	■	■	■		■					■		XAI in healthcare
[80]	2021	113	1991-2020	■	■				■	■			■	■	■	Contrastive and Counterfactual XAI
[14]	2021	206	2015-2020	■	■	■	■								■	Evaluation approaches of XAI
[50]	2021	45	2017-2020			■	■									Black-box issue
[70]	2021	150	2016-2020	■		■				■			■			ML interpretability methods
[51]	2021	55	2017-2020		■	■			■					■		XAI methods classification
[53]	2021	120	2014-2020	■	■	■	■									Argumentation enabling XAI
[61]	2021	210	2015-2020			■	■	■				■		■	■	XAI methods classification
[62]	2021	121	2016-2020			■	■	■								Necessity of explainability
[54]	2021	40	2017-2020	■		■		■								User and their concerns
[71]	2021	111	2016-2021				■	■								User and their concerns
[26]	2021	123	2016-2020			■	■			■			■	■	■	XAI methods classification
[60]	2020	60	2016-2020			■	■	■						■		XAI in reinforcement learning
[69]	2020	196	2017-2019	■	■	■	■							■	■	XAI methods classification
[63]	2020	400	2012-2019	■	■	■	■	■			■			■		Responsible AI
[81]	2020	205	2015-2020	■	■									■	■	Explainable recommendation
[49]	2020	50	2017-2019	■		■		■	■					■		Impact of XAI on trust
[52]	2019	150	2014-2018	■		■	■									Social Sciences related to XAI
[68]	2019	140	2014-2019	■	■	■	■	■					■		■	ML interpretability
[82]	2019	57	2015-2019	■		■								■		Counterfactual in XAI
[32]	2019	350	2000-2018	■	■	■	■	■							■	Good explanation
[58]	2018	46	2016-2018			■	■		■					■		Introduction to XAI
[66]	2018	289	2010-2017	■		■					■					Accountable System
[18]	2018	130	2012-2017			■		■					■	■		Black-box issues
[21]	2018	381	2014-2018	■		■	■	■			■			■		Key aspects of XAI
[57]	2018	57	2016-2018					■								User and their concerns
[64]	2018	50	2015-2017			■								■		Introduction to XAI
[28]	2018	30	2013-2017			■	■							■		Interpretability and its desiderata
[67]	2018	35	2014-2017			■			■					■		Necessity of explainability
[56]	2018	30	2015-2017	■		■							■			Visual model interpretability
[83]	2018	24	2014-2018	■			■				■					XAI with human intelligence
[15]	2017	48	2014-2017			■	■								■	Definition of interpretability
[55]	2017	53	2012-2016	■							■	■	■			Visual model interpretability

understand. In relation with the specific audience, Miller [52] also considered presently available explanations to be excessively static. An ideal explanation is one in which the explainer and the explainee interact with each other. The author suggested that explanations are social and should be interactively communicated to users. In the same vein, Grice's created cooperative principles [91] and four maxims that must be followed by explanations:

❶ **Quality:** Ascertain that the explanation is of good quality with the following properties:

- Do not provide some random explanation that may not be true, and
- Do not provide an explanation that does not have enough supporting evidence.

❷ **Quantity:** Deliver the appropriate amount of information in an explanation that has the following properties:

- Explanation must be informative i.e., provide as much information as needed, and
- At the same time, not provide more information than is required.

❸ **Relation:** An explanation must contain only information relevant to the discussion. This maxim may be used to improve the quantity of the explanation.

❹ **Manner:** Rather than what is given, manner refers to how information is delivered. Grice [91] has divided this into a number of maxims:

- Avoid ambiguous language in the explanation.
- Avoid ambiguity in the explanation.
- Avoid prolixity with a concise explanation.
- Avoid information that is not in order.

❺ **Context-oriented:** Explanations for developers are different from those for regulators that are different from those for end-users [92].

Furthermore, XAI is not a unitary entity; it encompasses several interconnected principles. According to the research reviewed, there are various contributing concepts for explaining AI systems. While there may appear to be some overlap between these concepts, we believe they reflect the many motives for explainability. In the next paragraph, we will go through several concepts that outline the standard definition.

3.2. The XAI Nomenclature

The black-box issue in AI refers to a system's difficulty in offering a reasonable explanation for how the system arrived at a decision. The words black-box, gray-box, and white-box are used in computing science and engineering to refer to varying levels of closure of a system's internal component.

The principle of explainability is closely connected to that of interpretability. Interpretable methods are explainable if humans can comprehend their operations. Even though explainable is a keyword in the XAI nomenclature, the term interpretable is more commonly used in the ML community than explainable. The related terms are defined as follows:

Definition 3.2.1 (Explainability)**.** The process of elucidating or revealing the decision-making mechanisms of models. The user may see how inputs and outputs are mathematically interlinked. It relates to the ability to understand *why* AI models make their decisions. The capacity to make automatic interpretations and describe the inner workings of an AI system in human terms is referred to as explainability. An explainable technique summarizes the reasons for an AI model's decision. Furthermore, a model's "Post-hoc Explainability" refers to methods/algorithms that are used to explain AI model's decisions [21, 26, 27].

Definition 3.2.2 (Interpretability)**.** Understanding the underlying workings of the AI model is another issue with black-box models. The intrinsic properties of a DL model are disclosed through interpretability. This has to do with being able to comprehend *how* AI models make their decisions. AI systems that explain the internals of an AI model in a manner that humans can comprehend are known as model intrinsic techniques [26, 27, 21].

There are many supplementary criteria that may be added to an XAI method such as transparency, fairness, reliability, or robustness. These are aimed at enhancing trust in the model. These concepts are further explained in the following.

Definition 3.2.3 (Transparency)**.** This is developed using an intrinsic method that generates a human-readable explanation for the model's decision. Transparency is essential for assessing the quality of a model's decision and for fending off adversaries [26, 64, 93, 94, 95].

Definition 3.2.4 (Fairness)**.** Due to fundamental biases in some datasets and algorithms, some groups of individuals can be treated unfairly and discriminated against by AI systems. Fairness refers to a model's ability to make unbiased decisions without favoring any of the populations represented in the input data distribution [27]. Biases may affect AI systems in a variety of ways. Biases such as the location of birth, socioeconomic background, and skills should not be a factor in AI models [24, 95]. During the development of an AI system and after its deployment, special methods may be developed for collecting and implementing user input [96, 97].

Definition 3.2.5 (Robustness)**.** The sensitivity of the system's output to a change in the input is measured by robustness [98]. It assesses the model's capacity to function correctly in case of uncertainty. The behavior of the system should not be dramatically affected by small changes in input [99]. This attribute is obtained by subjecting the model to adversarial inputs and ensuring that the system's error rate is near to that during training [100, 25].

A perturbation in the input example will cause a change in the outcome. Causality [101, 102, 103] measures the change in the predicted output. A selection of important insights and common associated XAI terminologies are completeness [14, 104], informativeness [28, 50], justifiability [105, 106, 107], monotonicity [108], reversibility [109, 110], simplicity [52, 111], reliability [112, 113, 114], and transferability [21, 28, 115].

XAI is focused on demystifying black-box models; it is also compatible with responsible AI since it may assist in creating transparent models.

Definition 3.2.6 (Satisfaction)**.** The ability of an explainability technique to improve the usability and utility of the ML-based system [116].

Definition 3.2.7 (Stability)**.** The ability of a procedure to provide comparable explanations for inputs that are similar [117, 118, 119].

Definition 3.2.8 (Responsibility)**.** Building trust and transparency makes a model trustworthy; but, in order for it to be responsible, societal values, morals, and ethical considerations must be also taken into account. Thus, Transparency, Responsibility, Accountability [21, 95, 120], Fairness, and Ethics [95, 120, 121] are the pillars that support Responsible AI [105, 120, 122, 123, 124].

Furthermore, we list out some XAI approaches, looking at issues ranging from trustworthiness to privacy awareness

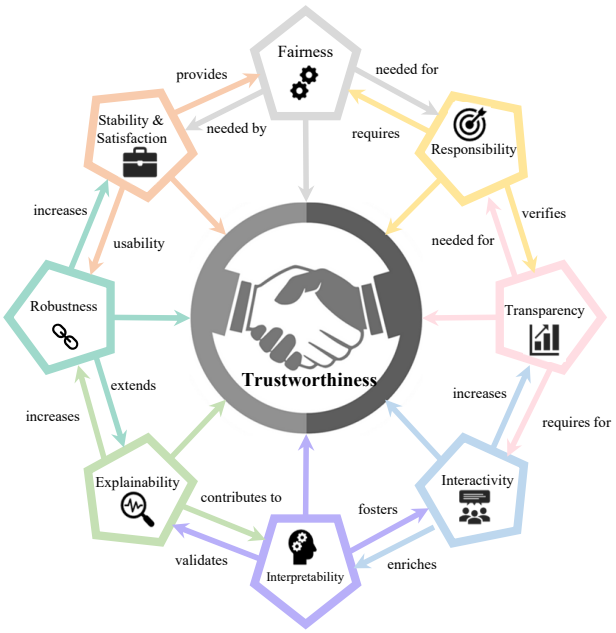


Figure 3: Relations among XAI concepts. The knowledge graph shows the interconnected potential uses of explainability concepts. The explainability concepts usually seek to accomplish one or more goals with the explanations that produce. The selection of the approach, the depth of the justifications, and the aims will be influenced by individual objectives. Inspired from [131].

[115, 125], and recognize the relevance of purpose and intended audience in data security [126, 127, 128] and safety [111, 129, 130].

Based on the previous nomenclature, we created a unified and organized perspective of the key concepts in the XAI area. Figure 3 depicts how such concepts are strongly interrelated. The explainability approaches always seek to accomplish one or more goals with the explanations they produce. Indeed, explainability is closely related to interpretability (which becomes a prerequisite for explainability) and robustness (which is increased by explainability). Similarly, robustness is related to (but it is not the same) stability; and both concepts have an important impact on satisfaction and reliability. In addition, confidence requires form interactivity while verifying reliability. Moreover, interactivity enriches interpretability while interpretability fosters interactivity. The selection of the XAI approach, the depth of the explanations, and their aims will all be influenced by all these concepts which have a direct impact on trustworthiness. Accordingly, we believe that studying the XAI concepts enables researchers to become familiar with the subject and its background rapidly. Additionally, knowing the primary search phrases in the area and the other terms that broadly allude to the same topics is a necessary prerequisite for conducting an insightful and compelling investigation.

3.3. Tradeoff between Accuracy and Interpretability

Researchers typically seek interpretable and high-performing models. Making the best model, however, usually increases model complexity, which tends to reduce interpretability. Understanding the tradeoff between accuracy and interpretability becomes critical for successful analytics as more corporations turn to AI models to spur development. The connection between accuracy and interpretability will be covered in this subsection.

Some experiments combine interpretable models to provide additional levels of insight; nevertheless, some interpretability may be sacrificed in order to get the most accurate model possible [52, 64]. The DT, for example, is quite interpretable; but, when it is repeated many times and combined with another model, such as Random Forest (RF), the interpretability suffers. As a result, we can say complicated models have become less interpretable in order to attain higher accuracy. Explainability enters the picture since an explanation entails comprehending the complicated system.

Figure 4 depicts the apparent balance between the ML model's performance and its ability to make explainable predictions in terms of the associated interpretability. For instance, a CNN is harder to understand than RF, and DT is easier to understand than RF. The crossover between interpretable and explainable models is represented by the gray region. This is because LR models with a few characteristics are simple to understand, but as the number of parameters increases, the model gets increasingly complex. A separable border between simple and complex models is difficult to define precisely. It is worth noting that according to some authors [132] "there is no scientific evidence for a general tradeoff between accuracy and interpretability": even if for many ML techniques the improvement of interpretability is at the cost of imposing constraints that in practice set an upper limit to the maximum accuracy to be achieved, a careful design can yield a good interpretability-accuracy tradeoff. Indeed, interpretability usually helps to understand how to improve a given model, so sometimes improving interpretability can also improve accuracy.

Moreover, DNNs are already capable of completing a wide range of tasks that before only a person could do, including classification, object detection, and recognition, as well as predictive maintenance tasks [107, 133]. However, humans may fool a DNN to categorize an input image incorrectly, despite the DNN having generally great performance with proper input and training. A tomato picture, for example, is altered by a human using random noise in order to deceive the DNN. When a model's formal goals (test-set prediction) and its real labels differ, explainability is required. The explanation is needed in order to acquire information, create a trustworthy connection between humans and AI systems, improve and learn from the system, as well as to comply with regulation. In addition, when comparing various models or architectures, model interpretability may be useful [58, 134]. The importance of models providing

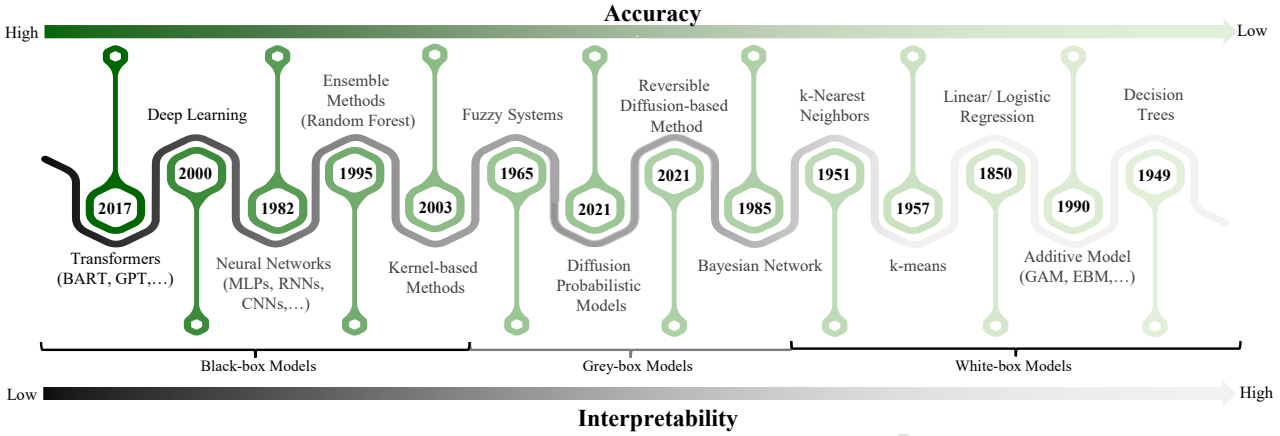


Figure 4: Illustration of the balance between accuracy and interpretability. The need for high model accuracy and interpretability is emphasized. Models with high accuracy need additional explainability, while models with low accuracy are simpler to comprehend but useless in many cases. The gray-box models represent the transition between black-box and white-box models. The number represents the year in which it first appeared in the field of AI research.

explanations for their decisions has been discussed in many ways in the literature [86, 135]. Accordingly, this article summarizes why explainability is important in terms of government regulations, user perspective, and application orientation. Explainability is not only a great academic interest but it will play a pivotal role in future AI systems that are expected to be used by millions of citizens, all around the world, in their everyday life.

4. Proposed XAI Taxonomy

How interpretable must the explanation be to satisfy the various user requirements? A plethora of factors may influence how an AI model works and produces its decisions, therefore a wide variety of explanations are needed. This is partly due to the absence of a universally accepted definition for XAI [18]. Also, it should be noted explainability focus on making explanations more user-friendly and trustworthy while avoiding making recommendations too strong without a basis existing in the data used throughout the AI development process in general. It is possible to conflict with the initial aim of gaining comprehension [136]. Explainability aims to understand the model and diagnose training processes that fail to converge and refine the model for robustness and better performance. We think explanations should be created by considering four axes in the typical AI pipeline: (i) data explainability, (ii) model explainability, (iii) post-hoc explainability, and (iv) assessment of explainability. Considering only one of the above may leave the potential audience with an inadequate understanding. As we can see in Figure 4, data scientists and developers are less concerned with post-hoc explanations but may gain more from knowing the internal workings of the model to improve the model's performance and comprehend how the data is applied to prevent overfitting. On the other hand, domain experts and end consumers are more interested in how and why a model generated a particular result and the key characteristics that led to that conclusion. As a result, this paper suggests that explanations must be created with the kind of user in mind.

As shown in Figure 4, this study uses a novel taxonomy that incorporates all four axes of explanation. There are two significant advantages to approaching explainability in this manner. (i) Since the goal of the explanation is more transparent and can be specified more precisely in a given axis, it makes the design and construction of explainable systems cheaper and easier. (ii) This approach will improve satisfaction among developers, researchers, domain specialists, and end consumers since they will get a more focused, easier-to-understand explanation compared to a broad general one for everyone. In addition, since metrics are unique to each axis, it will be simpler to assess which explanation is superior.

To what purpose does the explanation serve? Researchers have attempted to categorize the various explanations used to decipher the rationale in learning algorithms [137]. Explainability techniques should respond to many questions to create a comprehensive explanation. The most basic questions like *why* and *how* the model under investigation generates predictions and inferences have been addressed by researchers [42, 138, 139]. However, the research community has also recognized additional issues that could emerge and need other kinds of responses and, as a result, require different forms of explanation [140]. We formulated research questions based on a thorough examination of the literature on XAI research, which includes various research papers and previous surveys to ensure that the selected questions are aligned with the current state of XAI research. Our primary objective was to encompass a wide range of topics and factors that are pertinent to XAI, such as trustworthiness, ethics, interpretability, explainability, and human factors. By doing so, we ensured that the survey captures a comprehensive understanding of the subject matter, including

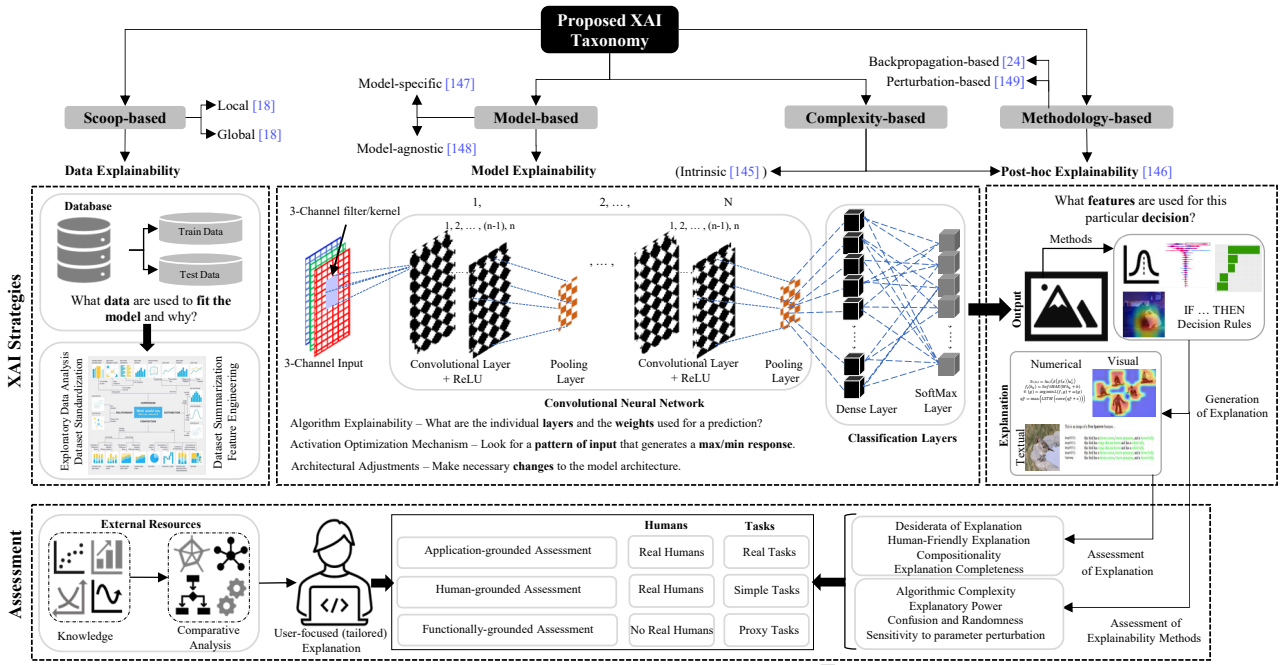


Figure 5: Proposed four-axes XAI methodology. At every level of the AI process, we present our explainability approach. Data explainability summarizes and analyzes data to offer insight into that data. A subsequent understanding of the data, feature engineering, and standardization can be achieved. Model explainability sheds light on the internal structure and running algorithm (notice that the picture depicts an example of DL but explainability applies also to other models). Post-hoc explainability elucidates significant features using several kinds of explanation. Several assessment approaches and their desiderata can be used to evaluate the explanations. The dotted lines define the four axes in the framework, whereas the solid lines differentiate between the entities of each axis.

Table 2
A list of research questions that address several levels/axes of explanation.

By Data Explainability	By Model Explainability	By Post-hoc Explainability
D1: What sort of information do we have in the database? **D2:** What can be inferred from this data? **D3:** What are the most important portions of the data? **D4:** How is the information distributed? **D5:** Is it possible to increase the model's performance by lowering the number of dimensions? **D6:** Can a better explanation be offered by using data summarizing techniques?	**M1:** What makes a parameter, objective, or action important to the system? **M2:** When did the system examine a parameter, objective, or action, and when did the model reject it? **M3:** What are the consequences of making a different decision or adjusting a parameter? **M4:** How does the system carry out a certain action? **M5:** How do these model parameters, objectives, or actions relate to one another? **M6:** What factors does the system take into account (or disregard) when making a decision? **M7:** In order to achieve a goal/inference, which techniques does the system utilize or avoid?	**P1:** What is the reason behind the model's prediction? **P2:** What was the reason for occurrence X? What would happen if Y was the cause of occurrence X? **P3:** What variables have the most influence on the user's decision? **P4:** What if the information is altered? **P5:** To keep current results, what criteria must be met? **P6:** Is there anything that can be done to have a different outcome? **P7:** Why is it essential to make a certain conclusion or decision?

various perspectives and dimensions that are crucial in XAI research. Table 2 summarizes the research questions for the first three axes of explainability. The fourth axis is distinct from the others yet depends on them, as such, we will examine this thoroughly in Section 9.

5. Explainability Methods

Explainability techniques come in a variety of shapes and sizes. The taxonomies covered in this section are summarized in Figure 4 (top portion). They can be divided into broad categories: scoop-based, model-based, complexity-based, and methodology-based. While there are many techniques for determining explainability, they can be discussed in detail in the following paragraphs. Since most papers

address explainability in ML algorithms, the word *interpretability* is often employed.

Scoop-based Explainers: Feature importance analysis is a common method for determining how model outputs relate to inputs either showing the entire model's behavior or a single prediction. Depending on the feature's importance, the kind of analysis performed can be categorized as either a local or global method. *Local explainers* only explain a specific decision or instance [21]. This implies that their decisions are limited to a single case with a single explanation. LIME is a seminal example of a local explanation [29]. On the other hand, *global explainers* are those that provide a rationale for the whole dataset [21]. These explanations remain true to overall observations. However, certain global explainers may offer localized explanations as well. For example, SHAP can provide local as well as global explanations [141].

Complexity-based Explainers: Interpretability is directly proportional to the complexity of the ML model. In general, the more complicated a model is, the more difficult it is to understand and explain. Interpretable ML algorithms can be classified as intrinsic or post-hoc interpretable depending on when interpretability is achieved. *Intrinsic interpretability* is accomplished by creating models that are self-explanatory and have interpretability built right in. To put it another way, intrinsic interpretable models have a simple structure [21]. In many cases, the simplicity and interpretability of the models, however, come at the expense of accuracy [142]. An alternative is to build a high-complexity, high-accuracy model and then utilize a different set of methods to give the necessary explanations without understanding how the original model works. Post-hoc explanations are provided by this class of techniques [143]. *Post-hoc* interpretability involves the development of a second model, usually as a surrogate of the original model (e.g., TREPAN [144]), in order to provide users with explanations.

Model-based Explainers: Model-agnostic or model-specific method is another way to categorize existing interpretability strategies [21]. A *model-specific* [145] method, as the name implies, is only applicable to particular kinds of models. By definition, intrinsic methods are model-specific. In contrast, *model-agnostic* [146] methods are independent of the kind of ML model used. Since model-agnostic interpretability techniques are model-free, there has been a recent increase in interest in them. Model-agnostic methods offer post-hoc interpretability; they are often used to interpret ANNs as either local or global explainers.

Methodology-based Explainers: XAI core algorithms may be classified in two ways depending on the implemented methodology: Backpropagation-based or perturbation-based methods. On the one hand, *Backpropagation-based* [27] methods may be used to backpropagate a significant signal from the output to the input. This begins with the output of the network and adds weight to each intermediate value calculated during the forward pass. To update the weights of each parameter and align the output to the ground truth, a gradient function distinguishes the network output with respect to each intermediate parameter. Thus, *Gradient-based* is another name for these techniques [27]. Saliency maps and class activation maps are other examples of this kind of method. On the other hand, *Perturbation-based* [147] algorithms use occlusion, partly replacing features via filling operations or generative algorithms, masking, conditional sampling, and other techniques to change the feature set of a given input instance and investigate the impact of these changes on the network output. Backpropagating gradients are not needed in this case since a single forward pass is enough to comprehend how the perturbed component in the input instance contributes to the network output [27].

Table 3
Publications in the literature regarding questions {D1, ..., D6} about model explainability, as described in Table 2.

Reference	Year	D1	D2	D3	D4	D5	D6
[148]	2021		■	■	■	■	■
[149]	2020	■	■		■	■	■
[150]	2020	■	■	■	■	■	■
[151]	2019	■		■		■	
[152]	2019	■	■		■		■
[153]	2018	■	■	■		■	
[154]	2018	■	■	■		■	
[155]	2017	■	■		■		
[156]	2017		■		■	■	
[157]	2016	■	■	■	■		■
[158]	2016	■	■		■		■
[159]	2011	■	■			■	■
[160]	2011	■			■	■	■
[161]	2008	■		■	■	■	
[162]	2008	■		■	■		■
[163]	2002	■			■	■	■

6. Data Explainability

Data explainability involves a group of techniques aimed at better comprehending the datasets used in the training and design of AI models. The fact that an AI model's behavior is heavily influenced by the training dataset makes this level of explainability very important. Therefore, many interactive data analysis tools have been developed to assist in understanding the input data. If data are not of high enough quality, it is impossible to create a model that will perform well. Data must be carefully examined after being collected.

The main publications related to data explainability are listed in Table 3. This table goes through each of the aspects of data explainability that make up the following subsections. As we will discuss below, data explainability may provide insights that can help AI systems (learned from data) become more explainable, efficient, and robust. The main aspects to consider are: Exploratory Data Analysis (EDA), explainable feature engineering, dataset description standardization, dataset summarizing methodologies, and knowledge graphs.

6.1. Exploratory Data Analysis

The goal of EDA is to compile a list of the most significant characteristics of a dataset, such as its dimensionality, mean, standard deviation, range, and missing samples. A powerful toolkit for extracting these characteristics rapidly from the dataset is *Google Facets* [155]. Consider a basic supervised binary classification task in which a model is created to distinguish whether an employee has an annual income over 50K or not. The UCI Census Income data [164] will be utilized throughout this manuscript as a case study. The dataset was obtained from the employment board and has a huge number of features. Further, assume that the classifier selected is less accurate. An EDA tool looks for biases in the dataset that may indicate an issue with class imbalance, such as having significantly fewer instances of adults with an income $>$ 50K than those with an income $<$ 50K, as depicted in Figure 6. After identifying a problem in the training dataset, a variety of remedies may be applied to fix it.

When it comes to evaluating datasets, however, relying only on statistical characteristics is seldom sufficient. For instance, Matejka et al. [156] have shown that datasets with similar statistical measurements may look different when plotted. As a result, data visualization techniques are a significant EDA tool. Data visualization provides a variety of charting options [165]. The best kind of chart to use draws on the dataset, application, and statistical characteristics that a data scientist wants to convey.

Real-world datasets are often complex and multidimensional, with a large number of variables. Visualizing such high-dimensional data may be challenging since humans only perceive three dimensions. To enable people to understand data with more than three dimensions, one approach is to use special charts, such as *Parallel Coordinate Plots* (PCP) [148]. These are utilized to figure out which features to keep and which ones to leave out as demonstrated in Figure 7. The high-dimensional dataset may also be projected onto a lower-dimensional form while keeping as much of the underlying structure as possible. Two well-known methods in this area are *Principal Component Analysis* (PCA) and *t-Distributed Stochastic Neighbor Embedding* [161] (t-SNE). Figure 8 illustrates the t-SNE case from the UCI Census Income dataset. If the underlying structure of a dataset is known to be mostly linear, then PCA is the best choice; otherwise, t-SNE is preferred. The *Embedding Projector toolbox* [157] facilitates the use of both techniques. Unfortunately, t-SNE is too slow when applied to large datasets. Dimensionality reduction approaches, such as Uniform Manifold Approximation and Projection (UMAP) [150], may be used in similar situations. It is claimed that UMAP is more accurate and scalable than t-SNE.

6.2. Explainable Feature Engineering

In addition to improving AI model performance, data explainability may aid in the development of explainable models and in the comprehension of post-hoc model explanations. Feature attribution, which involves evaluating the relative significance of input features in a particular model's decision, is a common kind of post-hoc explanation [151]. The related features should be explainable too, in the sense that developers should be able to intuitively assign a meaning to them and identify the most pertinent feature explanations for a specific end user. To put it another way, the accuracy of a model's predictions is limited by the characteristics that are employed to explain them [154].

The two most common approaches to explainable feature engineering are domain-specific and model-based methods [151]. Domain-specific methods rely on domain expert knowledge as well as insights gained via EDA to extract and identify significant features. Shi et al. [162], for example, utilized a binary classifier on satellite images to distinguish cloudy pixels from ice/snow pixels that looked quite similar. Model-based feature engineering, in contrast, makes use of a number of mathematical models to determine the underlying

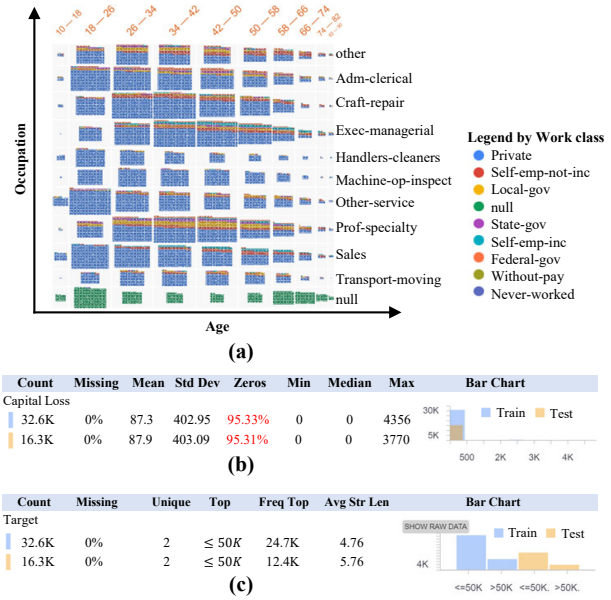


Figure 6: GoogleFacets: The UCI Census Income data was used to classify the income of an adult as over 50K a year or not. (a) Display all 16282 training data points that show the relationship between one feature (Age) and another feature (Occupation), then faceting is performed in a different dimension according to a discrete feature (Work class). (b) The table displays six integer-type statistical values from the UCI Census datasets. Non-uniformity is used to group the feature. For the sake of illustration, one feature, the capital loss is presented. Red numbers signify regions with the potential issue, in the capital loss numeric feature with a high proportion of values, are set to zero. The right-hand histograms compare the distributions of the training and test data. (c) The table displays one categorical (string) type feature out of the nine features in the UCI Census dataset. Distribution distance is used to group the features. The label values in the train and test sets are different, as shown in the right-hand histogram. A model trained and tested on such data would provide an incorrect assessment as a result of the label imbalance problem.

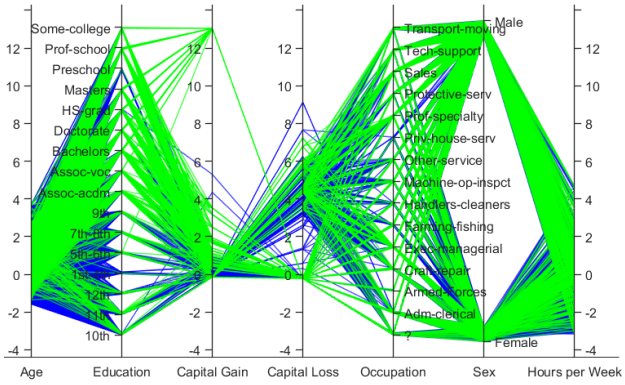


Figure 7: PCP: The UCI Census training data is presented in a single graph to reveal a 2D pattern. The z-score on the y-axis is plotted against each feature value. The graph can aid in deciphering feature correlations and identifying helpful predictors for class separation. It can be seen in the distinct cluster, the features of age and education have a significant role in determining a given class. In the class prediction, the capital gain, on the other hand, does not create separation boundaries. Thus, this feature may be left out of the classification task. The green line represents the target value > 50K, and the blue line denotes income value < 50K.

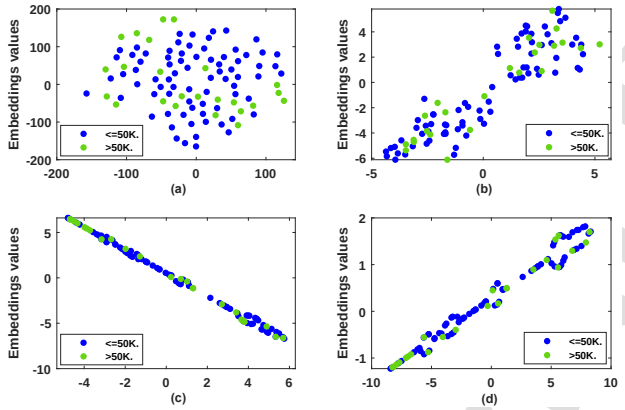


Figure 8: t-SNE: Produces a graph with well-defined clusters and a small number of integer data points. To get a better separation between the clusters of the UCI Census Income dataset, several distance measures are used: a) Mahalanobis, b) Cosine, c) Chebychev, and d) Euclidean. All these distance measures, except for Mahalanobis, provide reasonable separation between clusters in this scenario.

structure of a dataset. Clustering and dictionary learning are two examples of model-based methods [151]. Disentangled representation learning [153], which attempts to learn a representation of a dataset in which the generative latent variables are separated, is another important and related field of study. Latent variables may be thought of as explainable features of the dataset.

6.3. Dataset Description Standardization

Datasets are often released without adequate documentation. As such, standardization may solve issues such as systematic bias in AI models and data exploitation by enabling efficient communication between dataset creators and users. As a consequence of this, a number of suggestions for standard dataset descriptions have been made, including Datasheets for Datasets [166], Dataset Nutrition Labels [167], and Data Declarations for Natural Language Processing (NLP) [168]. These methods effectively offer various schemas for particular data connected to a dataset in order to track the datasets evolution, content, data collection method, legal/ethical problems, etc. For example, the nutrition label dataset approach [167] proposes a dataset document that includes information on many modules such as metadata, statistics, pair plots, the probabilistic model, provenance, and ground truth correlations. All these details are intended to be consistent with the nutrition information label on packaged foods. Similar to how customers can select their preferred food in a store regarding the nutrition information, AI experts may use the "nutrition labels" associated with a dataset as a reference to effectively identify the best dataset for their modeling objectives.

6.4. Dataset Summarizing Methodologies

Case-based reasoning [169] is a kind of explainable modeling technique that generates predictions for a given input and compares them to training samples/cases using a distance metric. Similar training samples, together with model predictions, can be provided to the end user as an explanation of the process. However, one significant drawback of this method is the need to retain the full training dataset, which may be prohibitively expensive or impractical for very large datasets, which have become more widely available. One solution to this problem is to save a portion of the training dataset that is nonetheless representative of the essence of that dataset. The goal of dataset summarization is to address this issue.

Document summarization [160], scene summarizing [170], and prototype selection [159] are some of the proposed techniques. To summarize a dataset, it is common to look for a small number of typical samples (known as *prototypes*) that provide a quick overview of the wider dataset. However, prototypes are insufficient for comprehending vast and complicated datasets; it is also important to include criticism with the prototypes. A *criticism* is an item of data that is relatively rare and is not properly represented by the prototype examples [149].

Kim et al. [158] presented an unsupervised learning technique for extracting both prototypes and criticisms from a dataset, they also performed additional testing by showing humans summarized datasets. Humans who were shown both prototype and critique images as their decision-making guide performed better than those who were just shown prototype images, according to the research. *Data squashing* is an another technique for data summarization [163]. The aim of data squashing is to create a smaller version of a dataset

that produces similar results. Unlike data summarization, weights are often assigned to samples in the smaller version of the dataset. Similar criteria for the initially stated data squashing are used in Bayesian learning, *Bayesian coresets* [152] have been highlighted as a data squashing method in a recent study.

6.5. Knowledge Graphs

Knowledge graphs provide a conceptualization of a given domain of application (e.g., finance, health, etc.) by modeling entities and their relationships by means of a directed, edge-labeled graph, often organizing them in an ontological schema. A knowledge graph enables us to determine which cues belong to ideas with similar semantic properties, factors that individuals may change, supporting data from the dataset, and other systems [171]. Several researchers envision using semantic technologies in the explainability field [74]. For example, Doctor XAI [172] creates an agnostic XAI approach for ontology-linked data classification. A Knowledge Graph is used by Gaur et al. [173] to feed DL models in order to increase their explainability. In addition to the methods listed above, we believe semantic technologies should (i) give background information, (ii) describe their properties, and (iii) give explanations in a context and language that are suitable.

In addition, ontologies offer a strong foundation for justifying predictions made by AI algorithms semantically. The Data Mining Ontology for Grid Programming (DAMON) [174] is a reference model for data mining approaches and existing tools. Another example, KDDONTO [175] emphasizes the development of data mining techniques. In addition, Panov et al. [176] created a heavy-weight ontology that offers ways to express data mining items and inductive queries. Confalonieri et al. [177] proposed an extension of Trepan [178] that integrates ontologies in the generation of explanations. In a user study, it was shown how explanations extracted using an ontology were perceived as more understandable than those extracted without the use of an ontology by human users.

6.6. Physics-Informed Neural Network

A class of neural networks known as physics-informed neural network incorporates physical rules and constraints into the architecture of the network. The integration of deep learning with physical modeling, process understanding, and domain knowledge enhances the interpretability and generalization of the models. Several methods have proposed incorporating physical equations and constraints into neural networks for modeling complex and non-linear processes. Earth system science (atmospheric and oceanic modeling, land surface processes, and cryospheric science) [179] and a two-step process to improve the spatio-temporal resolution of turbulent flows [180] are a few of these examples. Researchers also proposed new methods to improve this family of models. For instance, Seo et al. [181] proposed a method to control the behavior of neural networks using rule-based representations. Interested readers are encouraged to read the following survey paper [182].

Table 4
Publications in the literature regarding questions {M1, … , M7} about model explainability, as described in Table 2.

Reference	Year	M1	M2	M3	M4	M5	M6	M7
[183]	2020	■		■		■		■
[184]	2019	■	■	■		■		
[185]	2019		■			■	■	■
[186]	2019	■			■	■	■	■
[187]	2019	■	■	■	■		■	
[188]	2019	■	■		■			
[189]	2019		■		■	■		■
[190]	2019	■	■			■	■	■
[191]	2018	■			■	■	■	
[118]	2018	■	■	■	■	■	■	
[192]	2018		■	■	■	■	■	■
[193]	2018	■	■		■		■	■
[194]	2018		■		■	■	■	■
[195]	2018	■	■		■	■	■	■
[196]	2017	■	■		■			
[197]	2017	■	■		■		■	
[198]	2016	■		■	■		■	
[199]	2016	■	■	■		■		
[200]	2016	■		■	■	■		■
[201]	2015	■	■	■		■		■

7. Model Explainability

Even if data are clean and carefully prepared for training thanks to data explainability techniques like those discussed in the previous section, if the model lacks a clear understanding, then developers may still find it challenging to incorporate their own knowledge into the learning process with the aim of getting better results. Accordingly, in addition to data explainability, model explainability is of paramount importance. In many cases, just analyzing the outputs or taking a single input is insufficient to comprehend why a training procedure failed to provide the desired results. In such a case, the training procedure needs to be investigated. Model explainability aims to create models that are naturally more understandable. Limiting the selection of AI models to a particular family of models that are deemed intrinsically explainable is often considered identical to running explainable modeling. The debate, however, extends beyond the traditional explainable model families to cover more modern and innovative methods such as hybrid, joint prediction and explanation, and many other approaches. Whichever way we look at it, the most challenging part is still coming up with an explanatory mechanism that is firmly ingrained in the model.

Table 4 includes some of the most relevant papers regarding the model explainability issues to be discussed in the following subsections. The following subsections go through each of the aspects of model explainability, it includes information such as important parameters, the relationship between object and action, the internal working process, and infer the technique used to make decisions.

7.1. Family of Inherently Interpretable Models

The conventional method for building explainable models is to select the modeling technique from a set of techniques that are deemed interpretable (white-box models). Lipton [28] suggested three modeling phases to ensure interpretability: (i) Algorithmic transparency, (ii) Simulatability, and (iii) Decomposability. LR [202], DT [203], Decision sets [204], Rule sets [205], Case-based reasoning [169], Interpretable Fuzzy Systems [206], and Generalized Additive Models (GAMs) [207] are all examples of this family.

However, merely choosing a model from an interpretable family does not ensure explainability in practice. For example, it may not be possible to simulate an LR model using high-dimensional input data and therefore the model would be unexplainable [28]. To overcome this problem, one might use some kind of regularization, such as the *L1 norm*, to restrict the number of relevant input features while training the model. Furthermore, the coefficients calculated for the LR model may be unstable in the situation of feature multicollinearity (i.e., input features that are correlated). To solve this problem, further regularization, such as the *L2 norm*, may be used [208]. While there are particular techniques that mitigate these issues, interpretable model families are in general very basic, and therefore fall short of reflecting the complexity of some real-world situations. This results in the so-called interpretability-performance tradeoff, which states that the more performant a model is, the less interpretable it is, and vice versa [209]. However, by creating models that are both interpretable and performant, a number of academics have shown that the claimed interpretability versus performance tradeoff does not always hold true [132, 210]. When creating such a model, the primary issue is to make it simple enough for its target audience to understand while still being complex enough to properly match the underlying facts [151].

7.2. Hybrid Explainable Models

To develop a high-performance and explainable model, it may be feasible to combine an inherently interpretable modeling technique (like those cited in the previous section) with a sophisticated black-box method [211]. Hybrid explainable models are based on this idea. The *Deep k-Nearest Neighbors* (DkNN) [195] method uses kNN inference on the hidden representation of the training dataset that is learned via layers of a DNN, as shown in Figure 9. The conformal prediction approach is then used to integrate the kNN predictions for all layers. DkNNs have been demonstrated to be efficient and robust, with example-based explanations provided for its predictions in terms of the closest training samples utilized in each layer. On the other hand, DkNNs necessitate the storage of a hidden representation of the whole training dataset, this may be prohibitively expensive for big datasets. It can also provide neighborhood-based explanations, which make it easier to interpret the model's predictions. The synergy between robustness and explainability lies in the fact that DkNN can identify and handle non-conformal predictions that could potentially lead to the model's failure or poor performance, while also providing insights into how the model is making its predictions. It is worth noting that DkNN does not directly provide counterfactual explanations. However, it can be used in conjunction with counterfactual methods to improve model robustness and interpretability.

In terms of generating predictions within a conformal prediction framework, the *Deep Weighted Averaging Classifier* (DWAC) [187] technique is similar to DkNN models in that it relies on the labels of training examples that are comparable to the given input instance. However, the similarity is calculated only on the basis of the final layer's low-dimensional representation.

Self-Explaining Neural Networks (SENN) [118] are another example. The main concept behind SENN is to generalize a linear classifier by utilizing NNs to learn its features, their associated coefficients, and how the networks are aggregated into a prediction. Concept encoders, input-dependent parameterizer, and aggregators are used to describe three NNs. The resultant hybrid classifier is said to have a linear model's explainable structure but a black-box expressive capacity and flexibility. The *Contextual Explanation Networks* (CEN) [183] are related to SENNs in certain ways. The CEN presupposes a learning issue in which the input in a particular context has to be predicted. The aim is to utilize a complicated model to encode the context in a probabilistic way into the parameter space of an inherently interpretable model. The data is then entered into the CEN model to produce a prediction.

BagNets [188] are another example of hybrid explainable model. A BagNet is a bag-of-features model in which the features are learned using a DNN. This kind of model treats each input image as a bag of features when it comes to image classification. This bag-of-features representation is created by slicing an image into many segments and passing each segment through a DNN to get local class evidence as shown in Figure 10. All local evidence is then aggregated for each class and put through the SoftMax function to determine the overall probability.

Memory networks [212] also combine the learning capabilities of connectionist networks with a type of read- and write-able memory, as well as inference powers.

Neural-symbolic (NeSy) models [213] look at the application of connectionist mechanisms to symbolic computation principles and the logical characterization and analysis of sub-symbolic computation. These models can be used to explore, comprehend, visualize, and influence on the network complexity. We describe some of these models next.

To begin with, *Conceptors* [214] are a type of neuro-computational mechanism that can be coupled with Boolean logic to add a semantic interpretation component. Logic based concept induction (a formal logical reasoning over description logics) can also be used to explain data differentials over background knowledge; in the case in [215], from Wikipedia knowledge base. Also using logic as the symbolic component to attain explainable-by-design models

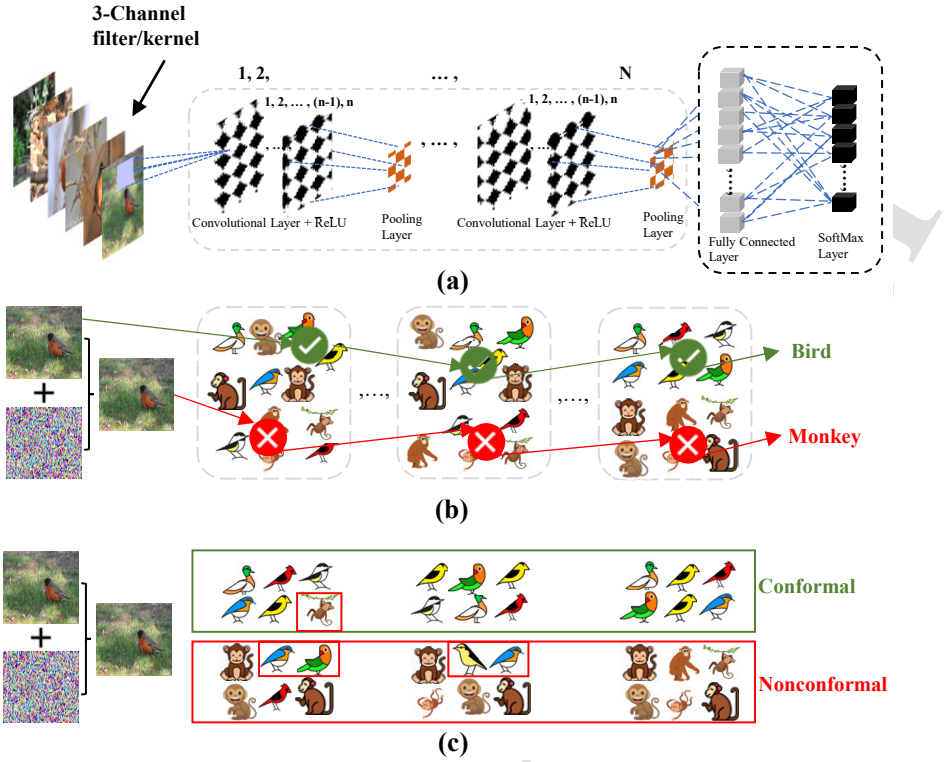


Figure 9: DkNN: (a) The DNN, (b) the output representations at each layer, and (c) the nearest neighbors at each layer. The mini ImageNet training points are shown by the bird and monkey icons. High-dimensional representation spaces are shown in 2D for clarity. When the nearest neighbor labels are homogeneous, such as in the case of the bird images, confidence is high. The nearest neighbors contribute to the interpretability of each layer's outcome. The term robustness refers to the ability to identify nonconformal predictions using nearest neighbor labels discovered for out-of-distribution inputs, such as an adversarial bird, across different hidden layers.

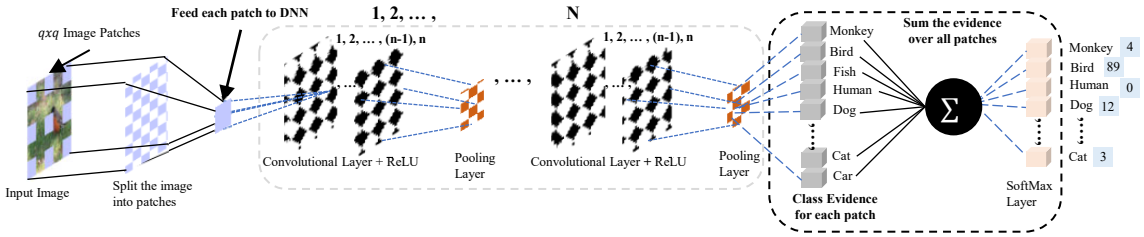


Figure 10: BagNets: The input is first split into $q \times q$ patches. Each patch is passed to the DNN to extract the evidence score. In the next stage, we take the sum of the class evidence scores overall patches to reach the final image classification decision.

are Logic Explained Networks [216], which allow to describe black-box model decisions as first order logic axioms in a logic-based simple, compositional and approachable readable manner.

Examples that facilitate the integration of expert knowledge into the black-box model for explainability within the NeSy paradigm consist of using ontologies or knowledge graphs as symbolic elements to encode common sense knowledge. For instance, to filter and refine opaque models (e.g., scene graph generators that describe images to facilitate machine scene understanding [217]), or to expose the functioning of a compositional model that can be first audited by verifying what object "parts" the model detected properly to draw a particular´´whole" object classification decision [218].

These architectures can even go further and, after identifying the misalignment in the expected explanation, correct it. One example of the latter approaches using XAI (via SHAP-backprop or alternatives) is the X-NeSyL (eXplainable Neural Symbolic Learning) methodology [219], which allows aligning machine explanations and domain expert explanations via knowledge graphs.

Apart from knowledge graphs, linguistic summaries aid to better align explanations with the most universal mean of symbolic understanding, i.e., using natural language [220]. *PLENARY translates* SHAP generated explanations

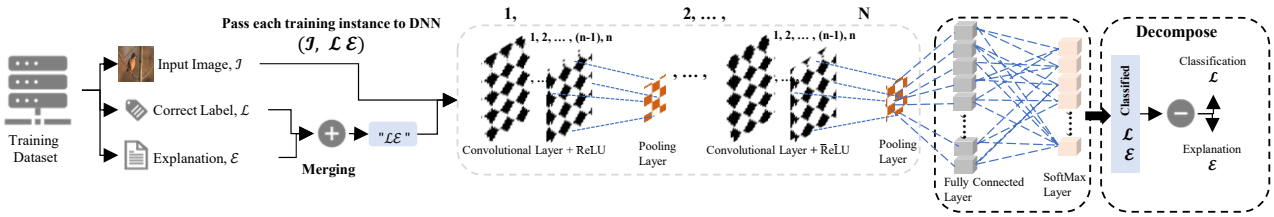


Figure 11: TED: The instance value $\mathcal{J}$, true label $\mathcal{L}$, and explanation $\mathcal{E}$ are all part of the training data. The label and explanation are encoded before sending them to the DNN. The encoded component and the input for prediction are fed to the DNN. Finally, the decoder will break down the label and explanation into their constituent parts.

of model outcomes into linguistic summaries so that experts can more easily validate complex XAI technique outputs. Finally, it is also shown that having prior knowledge about the environment in the form of a hierarchical knowledge graph, a significant speed up of contribution-based explainability analyses can be achieved. More concretely, Myerson values can be an efficient alternative to Shapley analysis in multiagent systems or RL [221].

Other symbolic elements showing to be a promising way to provide graphical clarity to interpret models include learning state spaces [222], or using Finite State Automata [223] to extract implicit knowledge learned by agents and discern *Aha!* or *Eureka* moments during learning. At last, causal graphs [224] can be a post-hoc manner to perform mediation analysis of causal effects of certain features by identifying confounding and mediating factors. These can act as the root of non-causal models motivating discriminatory policies.

7.3. Joint Prediction and Explanation

An explainable model may be trained to give both a prediction and an explanation. To put it another way, a complicated model may be explicitly trained to explain its predictions. This subsection will go through the methods that jointly explain a model's decision as well as their benefits and drawbacks.

To begin with, the *Teaching Explanations for Decisions* (TED) framework [189] is used to supplement the training dataset by including a collection of features, and output, as well as the user's reasoning for that decision, which is called an explanation, in each sample. When the model is trained, the provided output and its explanation are combined into a single label, as shown in Figure 11. The model's output is decoded at the time of the test, to give an output and a related explanation together at the end of the process. The TED framework offers a number of benefits, including the ability to provide explanations that meet the end user's requirements and the ability to be broadly used.

Park et al. [191] proposed a model explainability method for generating multimodal explanations. Their approach is comparable to the TED framework in that it needs a training dataset containing both textual and visual explanations. To test their method, the researchers utilized two new datasets relating to activity recognition and visual question-answering tasks, both of which were supplemented with multi-modal explanations. The authors suggested that incorporating multi-modal explanations improves prediction accuracy. Two major flaws exist in the techniques described above: (i) the authors presume that explanations are available in the training dataset, which is not always the case, and, (ii) the explanations produced by these techniques may not always represent how the model actually makes its predictions, but rather they might show what people want to perceive as the explanation.

Some approaches in this category do not need explanations for every prediction in the training dataset, this helps to overcome the limitations of the aforementioned methodologies. Hendricks et al. [199], for instance, suggested utilizing DNNs to provide visual explanations for the problem of object detection and recognition. In order to produce class-specific visual explanations of the input image predictions at the time of the test, their approach simply requires a textual explanation of images and their class labels at the time of training. Another example of a joint prediction and explanation approach is the *Rationalizing Neural Predictions* (RNP) model [200], which consist of two parts (both trained simultaneously): a generator and an encoder. In order to make a prediction, the generator uses the distribution of input text segments as potential explanations. The textual explanations are discovered through training, not explicitly given to the network. This is only accomplished by imposing two requirements: (i) the input text fragments must be brief and cohesive, and (ii) the model must be able to act as a replacement for the original content for the specified prediction task. As the encoder makes predictions based on the generator's rationale, it avoids two of the flaws stated previously. However, only providing rationale is insufficient to enable the end user to completely understand the prediction with confidence [192].

7.4. Explainability through Architectural Adjustments

By adjusting model architectures, it is also possible to improve model explainability. For example, Zhang et al. [193] created an explainable CNN that can push representations of upper-layer filters to be an object component rather than just a combination of patterns. This is accomplished by incorporating a particular loss function into the feature maps of an ordinary CNN. This loss function gives preference

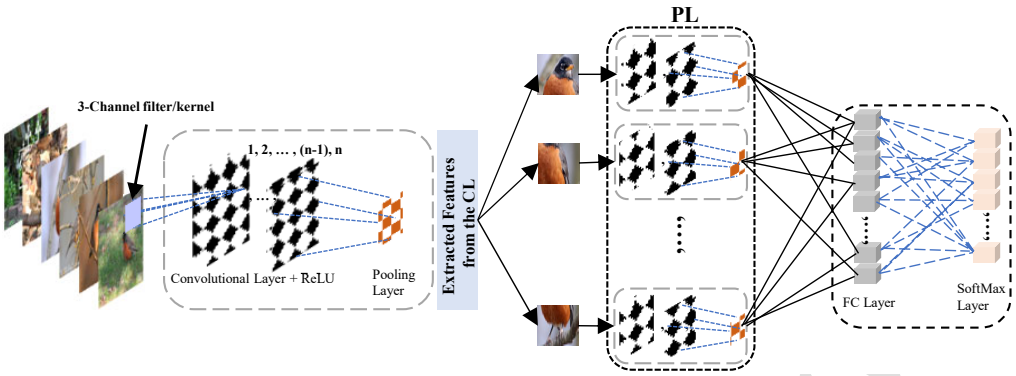


Figure 12: This Look Like That: CL extracts the valuable features to employ for making predictions with the given bird dataset. The $\mathcal{P}$ prototypes are learned by the network as representations of prototypical activation patterns in a region of the input image. As a result, each prototype may be seen as a hidden representation of a prototypical element of the given bird image.

to certain parts of an object inside a class category while remaining quiet on images from other classes. The key point to highlight is that this method does not need any object component annotation data. Explainable CNNs store more relevant information in their high-layer filters than traditionally trained CNNs do.

Chen et al. [185] proposed *This Looks Like That*, an Explainable Deep Network (EDN) architecture for image recognition. The motivation behind this architecture is based on how people explain classification reasoning in terms of different parts of an image being compared to a collection of learned image component prototypes. The authors suggested adding a Prototypes Layer (PL) between the Convolutional Layers (CL) and the Fully Connected (FC) layer to the standard CNN architecture as can be seen in Figure 12. For each class, the PL includes a certain number of image component prototypes. Each prototype is intended to contain the most important information for recognizing images within that class. Using a specific loss function, the PL and the CL layer parameters are learned simultaneously. A sparse convex optimization method is then used to learn the weights of the FC layer. The suggested EDN outperformed black-box DNNs in two image classification problems.

Furthermore, attention mechanisms [201] is said to provide some degree of explainability and they have altered the way how DL algorithms are used. There are many different kinds of attention mechanisms, interested readers are kindly referred to [225]. Concisely, attention-based models are widely employed in applications of NLP [186], computer vision [197] or time series modeling [198]. Their goal is to identify the parts of an input that are most relevant for performing well the specific task under consideration. Relevance is often defined by a collection of weights/scores given to the input components, referred to as the *Attention Map*. Explainable DNNs usually include some kind of attention mechanism. However, formal research on attention as an explainability mechanism claims that attention is not the same as explanation [190]. For example, attention maps are very weakly associated with gradient-based metrics of feature significance, according to a large collection of studies on different NLP tasks. Furthermore, extremely diverse sets of attention maps may provide the same predictions.

7.5. Explainability through Regularization

Regularization techniques are often used to enhance the prediction performance of AI models, and may also be used to increase model explainability. For example, *Tree Regularization* is presented by Wu et al. [192] to improve DNN explainability. The main concept is to encourage people to learn a model with a decision boundary that can be well approximated using a tiny DT, allowing humans to simulate the predictions. This is accomplished by introducing a new regularization term into the loss function that was used to train the model. Models built using this technique are more explainable without compromising predictive performance, according to their experimental findings for a variety of real-world applications.

In addition, a significant corpus of research is focused on utilizing regularization to explicitly limit the explanations of model predictions; ensuring this way that they are correct for suitable reasons. For instance, Ross et al. [196] proposed a method for constraining local model predictions during training to reflect domain knowledge. The authors considered input gradients as local first-order linear approximations that can be used to map model behavior, i.e., they are used as a first-order explanation for particular model input. The domain knowledge is stored as a binary mask matrix, with each feature indicating whether it should be utilized to forecast each input. The model loss function is then supplemented with a new term that penalizes input gradients that do not match the mask matrix. Models trained using this method generalize considerably better when training and testing on datasets with large differences. In another similar example, Ghaeini et al. [184] developed a technique called *Saliency learning*. In their method, expert annotations concentrate on important portions of the input rather than irrelevant parts, as well as having annotations at the word embedding level rather than at the input dimension level. The model extracts the event first by feeding the embeddings to two CNNs, as shown in Figure 13. After using max-pooling,

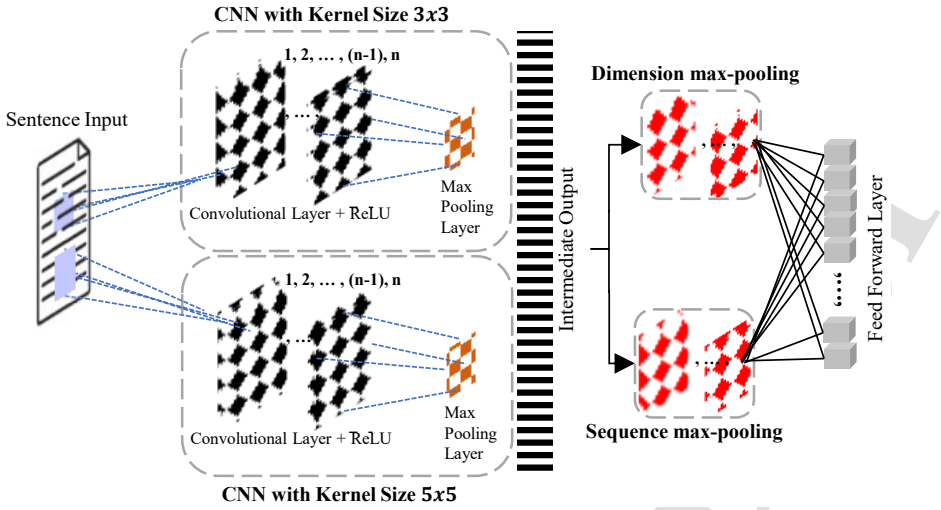


Figure 13: Saliency learning: A phrase is fed to two CNNs with kernel sizes of 3×3 and 5×5. The initial max-pooling operation produces an intermediate result. After that, the result is decomposed by performing the dimensional and sequential max-pooling operations. For the final prediction, the decoded output is concatenated and sent via a feed-forward layer.

an intermediate output is generated. Later, dimension-wise and sequence-wise max-pooling is used to get the final result. Experiments utilizing simulated explanations in a variety of tasks indicate that Saliency Learning produces more accurate and reliable results.

7.6. Other Methodologies

There are a few more notable model explainability methods worth mentioning. Angelino et al. developed the *Certifiable Optimum RulE ListS* (CORELS) method [194], which offers a solution for finding optimal rule lists for reducing the empirical risk of a given set of training data. Furthermore, the CORELS method has been shown to be quick and needs only simple software [226]. The fact that it can only deal with categorical data is its main disadvantage.

8. Post-hoc Explainability

After discussing data and model explainability issues, it is now time to go in-depth with post-hoc explainability issues. The various methods to deal with post-hoc explainability are grouped around six important features, as shown in Figure 14: (i) attribution methods, (ii) visualization methods, (iii) example-based explanation methods, (iv) game theory methods, (v) knowledge extraction methods, and (vi) neural methods. Let us start by formulating the problem.

Problem Formulation: In supervised ML, a model $h(x) = y$ maps a feature vector $x \in \mathcal{X}$ to a target $y \in \mathcal{Y}$. A training set $D = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ is utilized during fitting/training of the model. According to whether target y is a discrete value or a continuous value, supervised ML models can be categorized either as classification or regression problems. A black-box model can be formulated as $b : \mathcal{X} \rightarrow \mathcal{Y}$, $b(x) = y$ with $b \in \mathcal{B}$, where $\mathcal{B} \subset \mathcal{H}$ gives the model's hypothesis space. For instance, $\mathcal{B} = \{$NN with one hidden layer, two hidden layers$\}$. White or gray boxes can be formulated in a similar way. Let's suppose $\hat{w} : \mathcal{X} \rightarrow \mathcal{Y}$, $\hat{w}(x) = y$ with $\hat{w} \in \hat{\mathcal{W}}$, where $\hat{\mathcal{W}} \subset \mathcal{H}$ stands for the model's hypothesis space, is a white-box model. For instance, $\hat{\mathcal{W}} = \{$decision trees of depth 2, 3, 4$\}$.

The error measure $\mathcal{L} : \mathcal{Y} * \mathcal{Y} \rightarrow \mathbb{R}$ is used to evaluate the trained model's prediction in terms of its performance. A common example is the *hinge loss* from binary classification $\mathcal{L}\Big(h(x), y\Big) = max\{0, 1 - h(x) * y\}$ with $y \in \{-1, 1\}$. When the actual label y and the prediction $h(x)$ are identical, the loss is zero. The *squared deviation* $\mathcal{L}(h(x), y) = (h(x) - y)^2$ is a popular error metric used in regression tasks. Supervised ML is aimed at minimizing a given error metric:

$$h^* = \arg\min_{h \in \mathcal{H}} \frac{1}{|n|} \sum_{x \in \mathcal{X}}^{n} \mathcal{L}\Big(h(x_i), y_i\Big), \tag{1}$$

where h^* is the optimized model with the smallest loss. Table 5 contains the nomenclature, symbols, and variables used in this study.

Table 6 includes some of the most relevant papers regarding the post-hoc explainability issues to be discussed in the following subsections.

8.1. Attribution Methods

In the context of image processing, the majority of attribution methods depend on pixel associations to show which pixel of a training input image is relevant in terms of the model activating in a certain manner. Therefore, each pixel of the input image is given an attribution value known as its *relevance or contribution* [243]. Mathematically, a DNN, $\mathcal{B}$, takes an input image $\mathcal{I} = [i_1, i_2, ..., i_m] \in \mathbb{R}^N$, the output may be considered as $\mathcal{S}(\mathcal{I}) = [\mathcal{S}_1(\mathcal{I}), \mathcal{S}_2(\mathcal{I}), ..., \mathcal{S}_C(\mathcal{I})]$, where the total number of output neurons is denoted by C. An attribution method's purpose is to estimate the Relevance Score (RS) of each input pixel i_m to the output $\mathcal{S}_c$ when a particular target neuron c is specified. When all of the

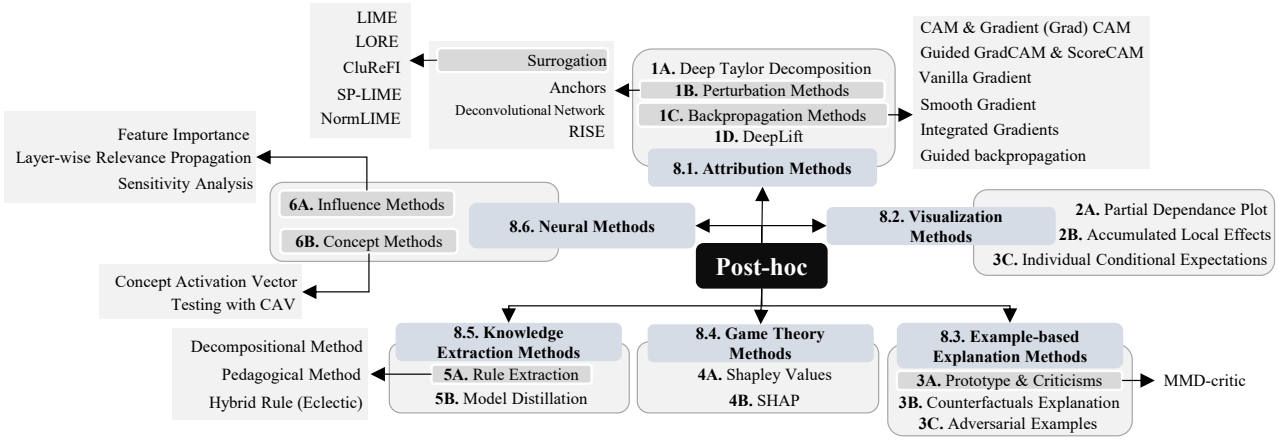


Figure 14: A proposed taxonomy for Post-hoc Explainability. Definition of acronyms: LIME - Locally Interpretable Model-Agnostic Explainer, LORE - Local Rule-based Explanation, CluReFI - Cluster Representatives with LIME, SP LIME - Submodular Pick LIME, RISE - Randomized Input Sampling to Provide Explanations, CAM - Class Activation Map, MMD - Maximum Mean Discrepancy, CAV - Concept Activation Vector. The numerical value and sub-index indicate the sequence and sub-sections in which these techniques are presented in the manuscript.

Table 5
Nomenclature in this manuscript.

Symbol	Description
D	Dataset with n instances
$\mathcal{X}$	Training set, a training instance x
$\mathcal{Y}$	Testing set, a testing instance y
$\mathcal{H}$	Hypothesis space, a hypothesis $h \in \mathcal{H}$
$\mathcal{B}$	black-box model family
$\hat{\mathcal{W}}$	White box model family
$\mathcal{F}(.)$	A ML model, either black-box or white box $F \in \mathcal{B}, \mathcal{W}$
$\mathcal{I}$	Input image of size $m \times m$ having pixel i
$\mathcal{S}(.)$	Output from b
$\mathcal{C}$	Number of Classes, a specific class represent by c
$\mathcal{L}$	Loss function
$\mathcal{RS}$	Relevance score or contribution
$\mathcal{N}_L$	Number of layers in a network
$\mathcal{T}$	Taylor series function
$\mathcal{BM}$	Binary mask
τ	Precision threshold
e	Number of neurons
w	Weight matrix
$\mathcal{M}$	Activation map
$\mathbb{P}$	Class probability
$\mathcal{G}$	Gaussian noise
σ	Standard deviation

Table 6
Publications in the literature regarding questions {P1, ..., P7} about post-hoc explainability, as described in Table 2.

Reference	Year	P1	P2	P3	P4	P5	P6	P7
[30]	2020	■	■		■		■	
[227]	2020	■	■		■	■	■	
[228]	2020	■		■		■		■
[229]	2018		■		■	■		■
[230]	2018		■		■	■		■
[231]	2018	■	■		■		■	■
[232]	2017	■			■	■	■	■
[141]	2017	■	■			■	■	
[233]	2017	■	■	■		■		■
[234]	2017	■		■	■		■	
[31]	2017	■	■	■		■		
[29]	2016	■	■			■	■	■
[235]	2016	■			■	■	■	■
[236]	2015	■	■		■			■
[237]	2015	■	■	■	■		■	
[238]	2014	■	■		■			■
[239]	2014	■		■	■	■		■
[240]	2013		■	■	■	■	■	■
[241]	2009	■	■	■	■	■	■	■
[242]	1988	■		■		■		

RSs have the same dimension as the input image, the two sets are merged to form an *Attribution Map* [244]. In the recent last couple of years, a number of novel attribution techniques have been developed. As you can see in Table 7 we distinguish four families of attribution methods: Deep Taylor Decomposition (DTD); Perturbation Methods; Back-propagation Methods; and DeepLift. They will be carefully described in the rest of this section.

1A. Deep Taylor Decomposition (DTD) was inspired by the SA method, decomposes the function value $\mathcal{F}(.)$ by summing the RSs to elucidate the model's behavior [236, 240]. Figure 15 shows the pixel-wise Taylor decomposition process. In the classification step; the input image $\mathcal{I}$ is used as a feature vector feed to the network. The network classifies the input into a specific class. In the next step; the classification output $\mathcal{S}(.)$ is decomposed into the RSs. The RSs are the terms in a first-order Taylor series expansion of the function $\mathcal{T}$ at an initial point $\overline{x}$ such that $\mathcal{T}(\overline{x}) = 0$. The initial point removes the information from the input for which $\mathcal{T}(x) > 0$. The following is a possible way to write this Taylor series

Table 7
A comprehensive overview of attribution-based XAI methods, highlighting advantages and disadvantages.

Method	Ref.	Advantages	Disadvantages	Concept
DTD	[240]	Training free method, may apply directly to any NNs.	i) Inconsistent in providing a unique solution, and slow computations [245]; ii) Partial explanation as higher order derivatives terms are set to zeros.	SA methods
LIME	[29]	i) Suitable to a very large number of explanatory variables, sparse explainer; ii) Same local interpretable model could be replaced [149]; iii) Selective and possibly contrastive explanations; iv) Provides local fidelity; v) Makes no assumptions about the model.	i) Incapable of explaining models with non-linear decision boundaries; ii) Incapable of explaining surrounding observations [149]; iii) Unsolved problem with tabular data.	Model agnostic local surrogate
LORE	[246]	i) Provide a counterfactual suggestion with the explanation; ii) Utilise a genetic algorithm that takes advantage of the black-box to generate examples; iii) Parameter-free method.	i) Based on assumption; ii) Cannot provide a global explanation; iii) Works for tabular data.	Local explanation
CluReFi	[247]	Provides local explanation to a cluster.	Representative of each cluster presents the explanation of important features.	Local explanation
SP-LIME	[29]	To check the entire model by extracting some data points. Aggregate the local models to form a global interpretation.	Less beneficial for high-level comprehension.	Model agnostic global surrogate
NormLIME	[227]	Provides finer-grained interpretation in a multi-class setting and add proper normalization to reduce the computation.	Aggregate many explanations for the class-specific explanation.	Local explanation
Anchors	[231]	i) Less computation than SHAP; ii) Better generalizability than LIME [227].	i) Requires discretization, highly configurable, and impactful setup; ii) Coverage drastically decreases with an increase in the number of feature predicates.	Perturbation-based model agnostic RL
DeconvNet	[238]	i) Highlights fine-grained details; ii) Dense feature representation with multi-layer.	Artifacts in the visualization [31]; ii) Training is difficult due to the large output space.	Pixel-space gradient visualization
RISE	[229]	i) Any architecture can be generalized; ii) Proposes causal metrics.	i) Inconsistent due to random mask; ii) Slow computation.	Pixel saliency
CAM	[235]	i) Identifies discriminative areas in an image classification task; ii) Fast and accurate.	i) Modify the network architecture that lends to complex model [31]; ii) Applicable to a specific type of CNN.	Regularization
Grad-CAM	[31]	i) Applies to a broad range of CNN model-families; ii) Robust to adversarial perturbations in an image classification task; iii) Help to achieve the model generalization by removing biases.	i) Lacks the ability to highlight fine-grained details; ii) Individual interpretations are difficult to aggregate for global knowledge.	Regularization
Guided Backpropagation	[248]	i) Highlights the fine-grained details and less noisy explanation [31]; ii) Provides more interpretable results than DeepLift.	i) Captures pixels detected by neurons, not the ones that suppress neurons [31]; ii) Less class-sensitive than the vanilla gradient.	Pixel-space Gradient Visualization
Guided Grad-CAM	[31]	i) Removes negative gradients and understand the model's decision; ii) Provides class descriptive and high-resolution maps.	i) Distinguishes an object of the same class; ii) Does not consider the entire class region.	Guided Back-propagation + Grad-CAM
ScoreCAM	[30]	i) Solves the dependency's problem on the gradients; ii) Achieves better visualization and fair interpretation.	i) Localization results are poor and lead to non-interpretability; ii) Smoothing generates inconsistent explanations.	CAM
Vanilla Gradient	[239]	i) Simple to implement based on backpropagation; ii) Pixel-wise features are important.	i) Makes undesirable changes with data pre-processing [249]; ii) Vulnerable to adversarial attacks [250]; iii) Decision-making process is unknown.	Backpropaga-tion interpretation
SmGrad	[233]	i) Denoising impact on the sensitivity map is achieved by training with noisy data; ii) Generates images with multiple levels of noise.	i) More effective with Large areas of the class object. ii) Degeneralizes to different networks.	Regularization [251]
Integrated Gradients (IG)	[252]	i) Very suitable for neural networks; ii) Optimizes the heatmap for faithful explanations.	i) Does not meet the Shapley values' axiom; ii) Frail mechanism to identify specific features and inconsistent to produce the explanation.	Shapley value
DeepLift	[234]	i) Gradient-free [227]; ii) Achieves the goal of completeness.	i) Depends on a reference point or baseline; ii) Produces inconsistent results due to redefining gradient.	Feature importance

expansion:

$$\mathcal{T}(x) = \sum_{i=1}^{d} \mathcal{RS}_i(x) + \mathcal{O}(xx^{\mathcal{T}}), \tag{2}$$

where higher-order derivative terms are denoted by $\mathcal{O}(.)$. The higher-order derivative terms are non-zero. In this way, a fraction of the explanation is generated. By neglecting $\mathcal{O}(.)$, the first-order terms are used to compute the $\mathcal{RS}$ as the

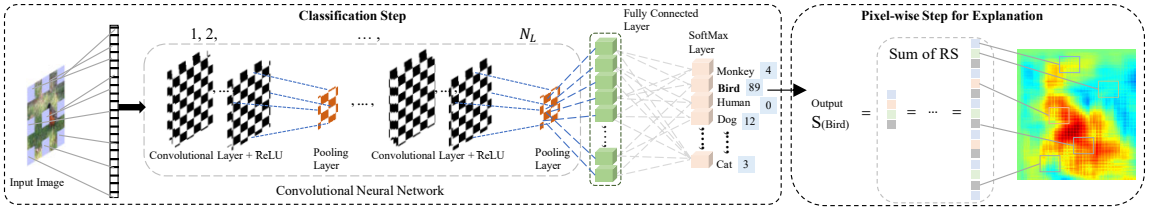


Figure 15: Deep Taylor Decomposition: The input image has been identified as a bird in the first step, while the model's reception of the features is shown as a heat map based on the relevance scores estimated from each hidden layer in the second step. The pixels surrounding the bird's location had a substantial impact on the outcome, as shown by the red regions that proved useful in the decision. In contrast, the blue regions were found not to be helpful in the decision.

partial explanation of $\mathcal{T}$:

$$\mathcal{RS}_i(x) = \frac{\partial \mathcal{T}}{\partial x_i}\Big|_{x=\bar{x}} \cdot (x_i - \bar{x}_i). \tag{3}$$

1B. Perturbation Methods are the second family of attribution methods under consideration in this survey. They calculate the attribution of a training instance feature directly by deleting, masking, or changing the input instance, then a forward pass on the modified input is executed before comparing the obtained results to the original output. While these approaches provide for direct measurement of a feature's marginal influence, the methods become very sluggish as the number of attributes to test increases [253].

Surrogation. A distinct model is created to explain the black-box decision either locally or globally [254], and the model created is intrinsically interpretable. Separating a black-box model from its explanation, according to Ribeiro et al. [255], provides better accuracy, flexibility, and usability. Surrogate models may be classified as local or global. By solving the following model equation, surrogate models are fit to the data:

$$\mathcal{F}^* = \arg\min_{w \in I} \frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} \mathcal{FS}\Big(\mathcal{F}(x), \mathcal{B}(x)\Big). \tag{4}$$

The function $\mathcal{FS}$ acts as the *fidelity score*, indicating how well the surrogate $\mathcal{F}^*$ approximates the black-box model $\mathcal{B}$. The *Global* scenario occurs, when the surrogate $\mathcal{F}^*$ uses the whole training dataset, while we define $\mathcal{X} = \{x_1, x_2, ..., x_n\}$ as a subset of the entire training dataset D. Often a subset $\mathcal{X}$ can represent the data distribution of the model $\mathcal{B}$ sufficiently well. The *Local* scenario occurs, when the surrogate $\mathcal{F}^*$ approximates $\mathcal{B}$ around a single test input x defined as $\mathcal{X} = \{x' | x' \in \mathcal{K}(x)\}$, where $\mathcal{K}$ is a neighborhood of x.

Locally Interpretable Model-Agnostic Explainer (LIME) is a proxy model and their derivatives are the best examples of both local and global surrogate methods [29]. This approach focuses on developing local surrogate models to explain individual predictions of black-box ML techniques. LIME investigates what happens to predictions when different types of data are fed into the ML model. LIME creates a new dataset using altered samples and the black-box model's predictions. LIME then uses the perturbed dataset to build an interpretable model that is weighted by the sampled instances' closeness to the instance of interest [29]. Any interpretable model, such as Least Absolute Shrinkage and Selection Operator (LASSO), LR, or DT, may be used. The learned model should be a good local approximation of the ML model's predictions but does not necessarily need to give a good global approximation. Local fidelity is another term for this level of precision.

The ideal way to acquire data variation depends on the type of data, which might be images, text, or tabular information. For images and text, turning single words or super-pixels on/off often is the best solution. LIME generates fresh samples from tabular data by perturbing each feature independently and drawing sample points from a normal distribution with the feature's mean and standard deviation [149]. The LIME model can be defined as:

$$\Theta(\phi) = \arg\min_{\mathcal{F} \in \mathcal{B}} \mathcal{L}(f, \mathcal{F}, \pi_\phi) + \omega(\mathcal{F}). \tag{5}$$

The obtained explanation $\Theta(\phi)$ interprets the target sample x, with linear weights when $\mathcal{F}$ is a linear model. A model $\mathcal{F} \in \mathcal{W}$, where $\mathcal{W}$ is a class of interpretable models; $\omega(\mathcal{F})$ is the complexity measure; $f : \mathbb{R}^d \to \mathbb{R}$ is the model being explained; and $\pi_\phi(x)$ is a proximity measure between the perturbed sample x and ϕ. The function $\mathcal{L}$ is a measure of the unfaithfulness of $\mathcal{F}$ in approximating f in the locality defined by π_ϕ. LIME is a model-agnostic method, which means that the obtained proxy model is suitable for use with any model [29].

Local Rule-based Explanation (LORE) [246] is an agnostic method capable of providing interpretable and trustworthy explanations. It constructs a simple, interpretable predictor by first using an ad-hoc "genetic algorithm" to generate a balanced set of neighbor instances of the given instance x, from which a decision tree classifier can be extracted. The resulting decision tree is then used to infer a local explanation e as follows:

$$e = \langle r = p \to y, \oint \rangle, \tag{6}$$

where the first part, $r = p \to y$ is a rule for making a decision y with a binary predictor p. The second part, $\oint$, is a set of counterfactual rules, which are the minimum changes to the feature x values that would cause the predictor to reverse its decision.

Cluster Representatives with LIME (CluReFI) [247] was created by extending LIME. First, LIME describes the representation of a cluster once the data has been clustered. After allocating an unknown data instance to the closest cluster, the explanation visualizes cluster assignments using a range of per-feature validity. Then, CluReFI visualizes each cluster's feature validity ranges for the most essential features contributing to the specified class. Unlike LIME, CluReFI shows the user the most significant aspects that contribute to the class for their representation.

Submodular Pick (SP) – LIME. Examining the model's predictions one by one can assist in deciding whether the model can be trusted as a whole. However, under typical conditions, it is impossible to examine all predictions. SP-LIME is a strategy for identifying must-see events. The number of events to examine is selected to be as large as possible so that the model can be understood. Cases with different features are also included [29]. By looking at the explanations for the subset chosen by SP-LIME, users will be able to choose whether or not to trust the model's general behavior. The pick set problem may be stated as a problem of choosing a set that leads to maximum coverage while remaining within a certain budget $\mathbb{B}$.

$$PickSet(\mathcal{RS}, \mathcal{I}) = \arg \max_{D, |D| \leq \mathbb{B}} \mathbb{C}(\mathcal{D}, \mathcal{RS}, \mathcal{I}) \tag{7}$$

where $\mathbb{C}$ is the coverage; the overall relevance of features that appear at least once in the examples from set $\mathcal{D}$ is given local importance $\mathcal{RS}$ for instances $\mathcal{I}$. As solving the given equation is NP-hard, a greedy approach is used.

NormLIME. LIME approximates a large NN on a small subset of the data manifold locally. Extraction of common explanations from many local approximations yields global explanations. However, the optimum way to integrate local approximations remains unclear. Based on local model explanations, NormLIME determines a feature's importance.

Formally, for a certain model $f : \mathcal{X} \rightarrow \mathcal{Y}$, it is possible to train an interpretable model $\mathcal{F}$ that is local to the region surrounding a certain input $x_0 \in \mathcal{X}$. A Gaussian probability distribution π_{x_0} is used to sample the data around x_0. Drawing x' from π_{x_0} and applying $f(.)$ repeatedly produces a new dataset $\mathcal{X}' = \{x', f(x')\}$. Thus, given the local dataset $\mathcal{X}'$, we develop a sparse LR $\mathcal{F}(x', x_0) = w_{x_0}^T x'$ by maximizing the following loss function using $\omega(.)$ as the degree of complexity.

$$\arg\min_{w_{x_0}} \mathcal{L}(f, \mathcal{F}, \pi_{x_0}) + \omega(w_{x_0}), \tag{8}$$

where the loss weight $\mathcal{L}(f, \mathcal{F}, \pi_{x_0}) = \mathbb{E}_{x' \approx \pi_{x_0}}(f(x' - \mathcal{F}(x', x_0))^2$. An upper limit $\mathcal{K}$ is set for the number of non-zero components in w_{x_0}, such that $\omega(w_{x_0}) = (||w_{x_0}||_0 > \mathcal{K})$. Although the optimization is difficult, it may be approached by choosing $\mathcal{K}$ features using LASSO regression and then carrying out regression exclusively on the top $\mathcal{K}$ features.

Anchors is another variant from LIME that looks for a decision rule that will explain individual predictions of any black-box classification model. To create local explanations for predictions made by black-box ML models, *Anchors* uses perturbations [231]. The resulting explanations are given as easy-to-understand IF-THEN rules termed anchors, this is in contrast to the surrogate models employed by LIME [29]. As LIME only uses a linear decision boundary that best approximates the model in a given perturbation space, its findings do not reflect how faithful the models are. *Anchors*, given an identical perturbation space, generates explanations in such a way that the coverage is customized to the model's behavior and clearly expresses the decision boundaries. As a result, *Anchors* is trustworthy by design and clearly identifies which scenarios it applies to.

As previously stated, the algorithm's conclusions or explanations are given in the form of anchors, which are decision rules. This approach overcomes the shortcomings of LIME. Refer to Table 7 for the downsides of each explainer. *Anchors* reduce the number of model calls by combining RL techniques with a graph search method [231]. We label an instance as x, the collection of predicates is $\mathcal{A}$, i.e., the resultant anchor or rule when $\mathcal{A}(x) = 1$ implies all of $\mathcal{A}'s$ feature predicates relate to the feature values of x. The following is a formal definition of an anchor $\mathcal{A}$:

$$\mathfrak{E}_{D_x(\mathcal{Z}|\mathcal{A})}[\mathbb{1}_{\mathcal{F}(x)=\mathcal{F}(z)}] \geq \tau; \mathcal{A}(x) = 1. \tag{9}$$

A rule or anchor $\mathcal{A}$ must be discovered for an instance x, and predicts a similar class to x for a fraction of at least τ (a precision threshold), i.e., only rules with at least τ local faithfulness are deemed valid, this based on, $\mathcal{D}_x(\mathcal{Z}|\mathcal{A})$ utilizing the given ML model (supplied by the indicator function $1_{\mathcal{F}(x)=\mathcal{F}(z)}$). Wherein $\mathcal{D}_x(.|\mathcal{A})$ represents the distribution of x's neighbors, which correspond to $\mathcal{A}$, and while the categorization model to be explained is denoted by $\mathcal{F}$.

Deconvolutional Network. Zeiler et al. [238] suggested exploring the intermediate layers of a Convolutional Network (ConvNet) to explain the model's decision. The authors visualized the activity of the intermediary layers to match the input pixel space by the use of a Deconvolutional Network (DeconvNet) [256]. A DeconvNet is similar to a ConvNet, it uses the same layer components such as pooling, regularization, and filtering in reverse order. A DeconvNet layer was connected to a ConvNet layer, as seen in Figure 16. The DeconvNet process, seen at the bottom of the figure, will use the layer underneath to rebuild an approximate replica of the ConvNet features.

To examine a given ConvNet activation map, all other maps in the network are set to zero. Only the non-zero activation map is passed to the DeconvNet layer. The DeconvNet layer performs three operations on the input map: (i) Unpooling - Despite the max pooling operation being non-invertible, an approximate inverse is recorded in a *Switches* variable. (ii) Rectification - The reconstructed features pass through the Rectifying Linear Unit (ReLU) non-linearity to ensure that the reconstructed maps are positive. (iii) Filtering - The DeconvNet takes the transpose of the learned filters. With the exception of the reverse of the ReLU layer, DeconvNet calculates its results in the same manner as *Vanilla*

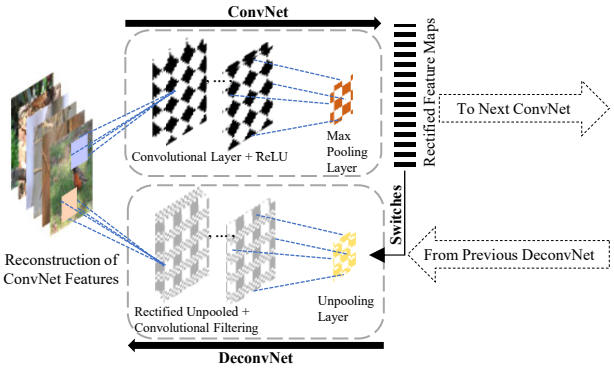


Figure 16: DeconvNet: Every layer in the ConvNet has a DeconvNet linked to it, allowing a continuous route back to the original input. The ConvNet receives an image and computes features across all layers. All the other activations in a layer are set to zero and feed the extracted feature maps into the attached DeconvNet layer to investigate ConvNet activations. DeconvNet can rebuild an approximate replica of the feature identified by ConvNet. During ConvNet's pooling operations, switches keep track of where the local maxima are located.

Gradient [257]. Vanilla Gradient may be thought of as a more generalized version of DeconvNet. When it comes to backpropagating the gradient via ReLU, DeconvNet takes a different approach:

$$\mathcal{RECON}_{\mathcal{N}_L} = \mathcal{RECON}_{\mathcal{N}_L+1}\mathbb{1}(\mathcal{RECON}_{\mathcal{N}_L+1} > 0), \tag{10}$$

where the reconstructions of the layers $\mathcal{N}_L$ and $\mathcal{N}_L + 1$ are $\mathcal{RECON}_{\mathcal{N}_L}$ and $\mathcal{RECON}_{\mathcal{N}_L+1}$ respectively. During backpropagation, the DeconvNet layer remembers which activation maps in the $\mathcal{N}_L$ layer have been set equal to zero in the forward pass and ensures they are unchanged in the $\mathcal{N}_L - 1$ layer.

Randomized Input Sampling to Provide Explanations (RISE). Petsiuk et al. [229] estimated the value of important pixels in an image by lowering the brightness of pixels to zero in random combinations. By multiplying an input image $\mathcal{I}$ elementwise with a randomly generated binary mask $\mathcal{BM}$, the authors were able to mimic this effect. Next, a confidence score is computed using the masked images by passing them to a DNN. A heatmap is produced by a linear combination of the masks, the confidence score is derived from the target class for the masked input. The authors further explain that this technique may be used to provide visual explanations for object detector predictions [258].

1C. Backpropagation Methods are another family of attribution methods. In one forward and one backward pass to the DNN, backpropagation methods calculate the attribution values for all the input features. Several of these passes may be required in certain cases, although this number does not rely on the number of input features and is often significantly less computationally expensive than perturbation approaches. Backpropagation approaches are often quicker than perturbation-based approaches, while their results are seldom directly tied to output variation [243]. The following summarizes the backpropagation approaches that are discussed in this article.

Gradient-only methods are only concerned with the gradient when determining if a change to a given pixel would affect the final prediction. Grad-CAM [31] and Vanilla Gradient [239] are two examples of such methods. The common idea behind gradient-only methods is that if a pixel in the input image is altered, the predicted probability of the class will either increase (positive gradient) or decrease (negative gradient). The greater the impact of an alteration to a pixel, the higher the absolute value of that gradient. The Class Activation Map (CAM) method and its variants will be discussed first, followed by Vanilla-based gradient approaches.

Class Activation Map (CAM). Lin et al. [259] utilized *Global Average Pooling* (GAP) as a structural regularizer in a CNN to reduce the number of parameters used while retaining exceptional performance. With little modification to the GAP method, Zhou et al. [235] discovered that the network could efficiently detect discriminative image areas in a single forward pass. The weighted activation map produced for each feature map is referred to as a CAM. Figure 17 depicts the process of creating a CAM. The GAP layer is positioned immediately before the last layer (SoftMax). The GAP takes the previously generated feature maps and calculates the spatial average. The SoftMax layer returns the class probability according to the weighted sum of the spatial average map values. The weight matrix is then passed back to the last convolutional layer, where it is used to calculate the weighted sum of the feature mappings and produce a CAM.

Let $\mathcal{M}_e(x, y)$ be the activation map of the e-th neuron from the last convolutional layer at a given location (x,y). The GAP spatial average may be calculated as follows:

$$\mathcal{GAP} = \sum_{x,y} \mathcal{M}_e(x, y). \tag{11}$$

Consider w_e^C to be the weight matrix that corresponds to the class C at the e-th neuron. Thus, for class C, the SoftMax takes an input of $\sum_e w_e^C \cdot \mathcal{GAP}$. The SoftMax layer will return the class probability as:

$$\mathbb{P}_C = \frac{exp\left(\sum_e w_e^C \cdot \mathcal{GAP}\right)}{\sum_C \left(exp\left(\sum_e w_e^C \cdot \mathcal{GAP}\right)\right)}. \tag{12}$$

The weight matrix w_e^C is passed back to the feature maps generated by the last convolutional layer. In this way, the obtained map is referred to as the CAM and defined by:

$$\mathcal{CAM}_C(x, y) = \sum_e w_e^C \cdot \mathcal{M}_e(x, y). \tag{13}$$

Gradient-weighted CAM (Grad-CAM) [31] provides visual explanations for any model in the CNN family without needing to go through architectural modifications or

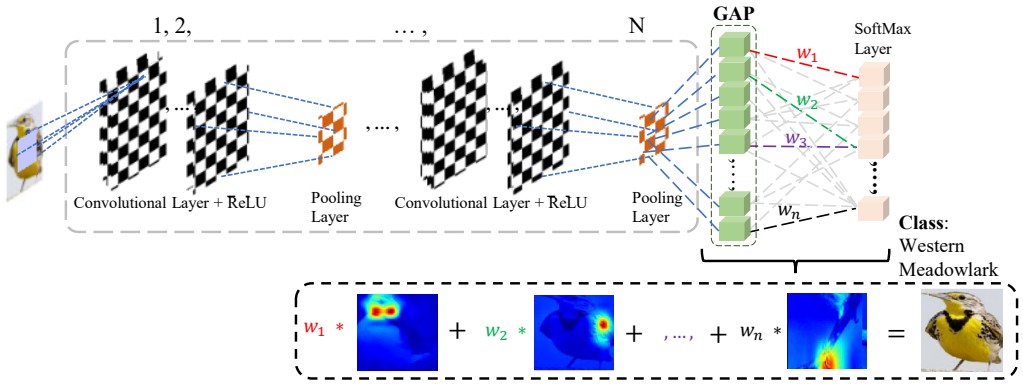


Figure 17: The spatial average of each unit's feature map, from the last possible CL, is generated by the GAP. The final result is generated using a weighted sum of the spatial data. The discriminative areas, distinct to each class, are highlighted in the CAM.

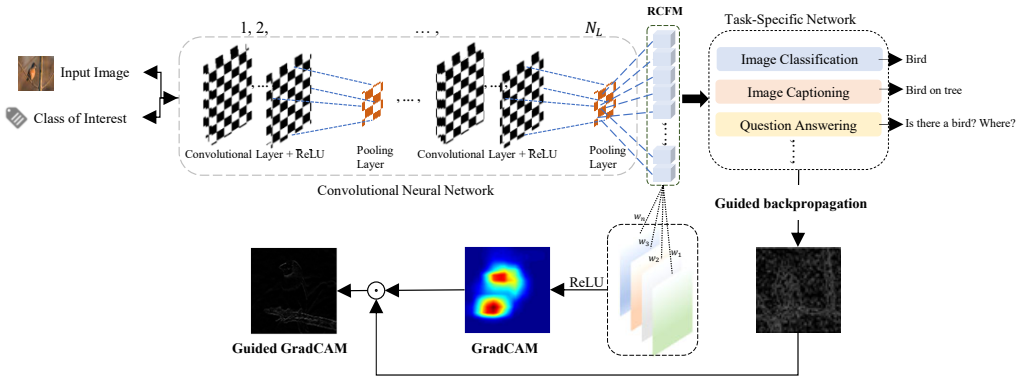


Figure 18: The input image and class of interest are fed into the DNN, which then performs task-specific calculations to provide a raw score for the class. Guided backpropagation is the output of the task-specific network. The Guided backpropagation result is passed to the RCFM in order to calculate the rough Grad-CAM localization, which reflects where the model must look in order to make specific decisions. Elementwise multiplication of the heatmap of the Grad-CAM with Guided backpropagation produces Guided Grad-CAM, which is concept-specific and has high resolution.

retraining, unlike regular CAM approaches. The CNN layers are well-known for capturing both spatial information and high-level semantics. With this foundation in place, the final CNN layer may have the optimal composition for extracting important data. Grad-CAM assigns significance ratings to each neuron for the given target class using the gradient information backpropagated to the final convolutional layer. An overview of Grad-CAM is shown in Figure 18. This model considers: (i) an input image, and (ii) a target class. To get a raw score for a particular category, the input image is passed via a CNN module and uses task-specific calculations. All the gradients are set equal to zero except for the target class. The non-zero signal is backpropagated to an interesting features map, these are referred to as *Rectified Convolutional Feature Maps* (RCFM) which are combined to produce the Grad-CAM map of the target class.

Let the Grad-CAM localization map be $\mathcal{M}^{C}_{Grad-CAM} \in \mathbb{R}^{bxt}$ with b representing width and t representing height for the class C. The class score is defined by S_C before going through the SoftMax function. Firstly, the gradient of S_C is computed with respect to the RCFM m_k; $\gamma_k = \frac{\partial S_C}{\partial m_k}$. The computed gradient γ_k is passed to the GAP layer to obtain the significant weights matrix for the neurons of the last convolutional layer as follows:

$$w_k^C = \frac{1}{\mathcal{Y}} \sum_p \sum_q \gamma_k(p,q) = \frac{1}{\mathcal{Y}} \sum_p \sum_q \frac{\partial S_C}{\partial m_k(p,q)}. \quad (14)$$

The weight matrix w_k^C is a partial linearization of the DNN that represents the significance of the k-th feature map for the class C. The weight matrix multiply with the RCFM m_k and passed to the ReLU layer to obtain the Grad-CAM map as:

$$\mathcal{M}^{C}_{Grad-CAM} = ReLU\Big(\sum_k w_k^C m_k\Big). \quad (15)$$

Guided Backpropagation [248] and deconvolution methods compute the gradient of the target output with respect to the input as shown in Figure 18. However, the backpropagation of ReLU functions is overridden so that only non-negative gradients are backpropagated. In guided backpropagation, the ReLU function is applied to the input gradients, and during deconvolution, the ReLU function is applied

to the output gradients and directly backpropagated. DeconvNet and guided backpropagation approaches generate imputed versions of the gradient rather than the true gradient [248].

Guided Grad-CAM is a class-discriminative method and locates target class areas, however, it lacks the capacity to emphasize fine-grained features that pixel-space gradient visualization techniques (e.g., DeconvNet [238]) or Guided Backpropagation [248] can provide. When backpropagating via ReLU layers, Guided Backpropagation illustrates gradients in relation to the input image, where the negative gradients are suppressed. This seems to be aimed at capturing pixels that are sensed by neurons rather than capturing those that inhibit neurons. Figure 18 shows how to combine (by means of element-wise multiplication) both Guided Backpropagation and Grad-CAM visualizations, thus producing Guided Grad-CAM.

Score-CAM [30] also incorporates gradient information, but a concept known as Increase of Confidence is used to provide priority for each activation map. Let $\mathcal{F} = \mathcal{B}(X)$ be a model that accepts $\mathcal{I}$ as an input image and produces logits $\mathcal{F}'$. $\mathcal{A}^i_{\mathcal{N}_L}$ denotes the i-th channel of the convolutional layer $\mathcal{N}_L$. The contribution of $\mathcal{A}^i_{\mathcal{N}_L}$ to $\mathcal{F}\prime$ with $\mathcal{I}_b$ as the baseline image for class category c is:

$$Cont(\mathcal{A}^i_{\mathcal{N}_L}) = \mathcal{F}^c(\mathcal{I} \cdot \mathbb{H}^i_l) - \mathcal{F}^c(\mathcal{I}_b), \tag{16}$$

where $\mathbb{H}^i_{\mathcal{N}_L} = s\left(Up\left(\mathcal{A}^i_{\mathcal{N}_L}\right)\right)$. The operator $Up(.)$ upsamples $\mathcal{A}^i_{\mathcal{N}_L}$ to the required input size and s normalizes each element to [0, 1]. Score-CAM can be represented as:

$$\mathcal{L}^c_{ScoreCAM} = ReLU\Big(\sum_i \beta^c_i \mathcal{A}^i_{\mathcal{N}_L}\Big), \tag{17}$$

where $\beta^c_i = Cont\left(\mathcal{A}^i_{\mathcal{N}_L}\right)$.

Vanilla Gradient. Simonyan et al. [239] named this approach as *Image-Specific Class Saliency*, or simply *Saliency Maps*. In this approach, the loss function's gradient is calculated with regard to the input pixels. Many XAI algorithms generate saliency maps, i.e., heatmaps that emphasize significant input with the largest impact on the prediction; with hotness denoting areas that have a significant effect on the model's ultimate decision [239]. Accordingly, saliency maps are a means of evaluating a CNN's prediction, but they have been criticized for concentrating on the input and failing to describe how the model actually makes its decision. The first technique offered by Zeiler et al. [238] used DeconvNet. A DeconvNet reconstructs the input from the activation of an intermediate layer of the network to distinguish the features (pixels) in the input that the particular intermediate layer of the network was looking for. The second, and simplest, method of obtaining a saliency map was proposed by Simonyan et al. [239]. This method computes the gradients of logits with respect to the network's input using the backpropagation technique. It highlights pixels of the input image via backpropagation based on the quantity of the gradient received, indicating the pixel contribution to the final relevance score. A guided backpropagation algorithm was presented as a third method of obtaining saliency maps by combining both techniques [248]. Instead of masking the importance signal based on negative input signal values in the forward pass or using negative reconstruction signal values (deconvolution), the authors mask the signal according to whether each of these situations occurs. This method excels at obtaining high-resolution, precise saliency maps. As the idea of the gradient is present in all NNs, this technique may be used with any ANN. As a result, this approach might be called a model-agnostic interpretation approach.

The formal definition of a saliency map is that an image $\mathcal{I}$ has a class C with $S_C(.)$ being the class relevance score, then the pixels of image $\mathcal{I}$, based on the score function for a linear model, can be represented as:

$$S_C(\mathcal{I}) = b_C + \mathcal{I}w, \tag{18}$$

where w denotes the network's weight vector and b_C denotes its bias. The significance of the pixels is determined by the magnitude of w. In the case of a DNN, however, the scoring function is quite nonlinear. As a result, the above equation may be expressed as:

$$S_C(\mathcal{I}) \approx b_C + \mathcal{I}w, \tag{19}$$

where w can be derived for an image $\mathcal{I}_0$:

$$w = \frac{\partial S_C}{\partial \mathcal{I}}\Big|_{\mathcal{I}_0}. \tag{20}$$

As non-linear units like ReLU return unsigned values, there is uncertainty about how the gradient will be calculated in the backward pass. The ReLU function is defined as $\mathcal{N}_{L+1}(l) = \max(0, \mathcal{N}_L)$ from layer $\mathcal{N}_L$ to layer $\mathcal{N}_{L-1}$. The following is how the uncertainty is resolved:

$$\frac{\partial \mathcal{F}}{\partial \mathcal{N}_L} = \frac{\partial \mathcal{F}}{\partial \mathcal{N}_{L+1}} \cdot \mathcal{I}(\mathcal{N}_L > 0). \tag{21}$$

Rearranging the components of w yields the saliency or sensitivity or pixel attribution map $\mathcal{M} \in \mathbb{R}^{mxn}$. The number of components in w equals the number of pixels for the grey-scale image $\mathcal{I}$. Thus, the saliency map is defined as:

$$\mathcal{M}_{ij} = \left|w_{idx(i,j)}\right|, \tag{22}$$

where $idx(i, j)$ represents the component of w that corresponds to the i-th row and j-th column. The saliency map for an RGB image $\mathcal{I}$ is derived as:

$$\mathcal{M}_{ij} = \max_{ch}\left|w_{idx(i,j,ch)}\right|, \tag{23}$$

where ch denotes the color channel of image $\mathcal{I}$. To create a single saliency map, the equation takes the maximum value from all the color channels.

SmoothGrad (SmGrad). In practice, sensitivity maps tend to be very noisy as the maps are based on the gradients of

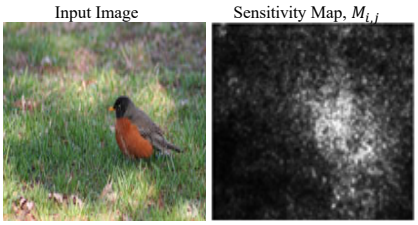


Figure 19: A noisy sensitivity map based on the gradient from an image classification network. Partial derivatives with greater absolute values are represented by brighter pixels.

the class score [239, 260]. This noise is due to the sharp fluctuations of the partial derivatives. Moreover, sensitivity maps do not show correlations between the highlighted pixels and the input label over the entire region, as shown in Figure 19. It is possible to smooth the gradient using a Gaussian kernel, instead of visualizing the gradient values directly. The gradients are smoothed by adding several forms of Gaussian noise to the input image before averaging the sensitivity maps. As a result, SmGrad has the following definition [233]:

$$\hat{\mathcal{M}}_{sm}(\mathcal{I}) = \frac{1}{n}\sum_{1}^{n} \mathcal{M}_n\big(\mathcal{I} + \mathcal{G}(0,\sigma^2)\big), \tag{24}$$

where the number of instances is n, the input image is $\mathcal{I}$, and $\mathcal{G}$ is the Gaussian noise with σ as the standard deviation.

Gradient-based sensitivity maps may be sharpened using two types of smoothing, according to previous studies reported in [233]. First, it seems that averaging maps that were created from numerous small perturbations of an input image have a substantial smoothing impact. Second, if the training data has been skewed with random noise, then the impact may be amplified even further.

Integrated Gradient. According to Sundarajan et al. [252] most gradient-based techniques miss key propositions that are desirable attributes. Intuitively, we understand the Integrated Gradient (IG) approach as combining Gradient Implementation Invariance (GII) with the sensitivity of LRP or DeepLift techniques. Let $\mathcal{F}$ be a DNN, $\mathcal{I}$ be an input image, and $\mathcal{I}'$ be the baseline image, which could represent a black image for image classification networks, or could be a vector of zeros for word embedding in text prediction models. The gradients along the inputs, which are on a straight line between the baseline image $\mathcal{I}'$ and the input image $\mathcal{I}$ are grouped together using an IG technique to suppress noise. As a result, the IGs along the k-th dimension are defined as:

$$IG_k(\mathcal{I}) = (\mathcal{I}_k - \mathcal{I}'_k) * \int_{\beta=0}^{1} \frac{\partial \mathcal{F}\big(\mathcal{I}' + \beta(\mathcal{I} - \mathcal{I}')\big)}{\partial \mathcal{I}_k} d\beta. \tag{25}$$

Notice that, a Riemann sum or Gauss Legendre quadrature rule can be used to approximate this integral.

1D. DeepLIFT [234] is the last family of attribution methods that assign significance ratings to input variables, in a similar way to pixel-wise decomposition. The fundamental premise in DeepLIFT is that it frames the topic of significance in terms of deviations from a reference condition $\hat{\mathcal{I}}$, which is selected by the user. At the layer $\mathcal{N}_L$, the contributions can be specified as follows:

$$\mathcal{RS}_i^{\mathcal{N}_L}(\mathcal{I}) = S_i(\mathcal{I}) - S_i(\hat{\mathcal{I}}), \tag{26}$$

where i is the interested neuron in the NN. For all others, $\mathcal{RS}_i^{\mathcal{N}_L}(\mathcal{I})$ is set to zero. The reference is often set to zero, just as it is in LRP. Running a forward pass determines all of the values in $\hat{\mathcal{I}}_{ij}$ for each hidden layer $\mathcal{N}_L$. The RS may be defined as follows:

$$\mathcal{RS}_i^{\mathcal{N}_L}(\mathcal{I}) = \sum_i \frac{\mathcal{I}_{ij} - \hat{\mathcal{I}}_{ij}}{\sum_j \mathcal{I}_{ij} - \sum_j \hat{\mathcal{I}}_{ij}} \mathcal{RS}_i^{\mathcal{N}_L+1}, \tag{27}$$

when a reference $\hat{\mathcal{I}}_{ij}$ is fed to the NN, the weighted activation is denoted as $\hat{\mathcal{I}}_{ij} = w_{ij}^{\mathcal{N}_L+1,\mathcal{N}_L}\hat{\mathcal{I}}_i^{\mathcal{N}_L}$ for a neuron i with respect to neuron j. This rule was included as part of the method's original development.

8.2. Visualization Methods

Understanding an AI model, by visualizing its representations to investigate the underlying patterns is a natural concept. Visualization methods are most often used with supervised learning models. Various visualization approaches will be covered in the following paragraphs and their strengths and weaknesses are summarized in Table 8.

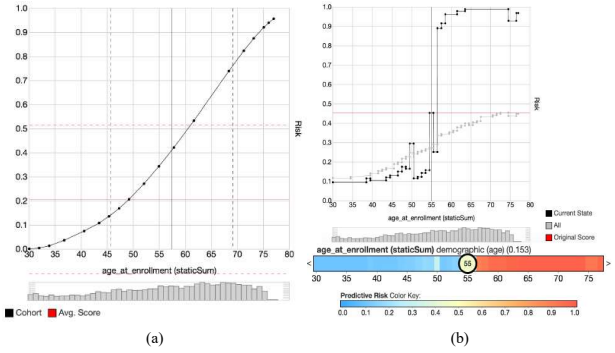


Figure 20: (a) The average projected risk, or probability of a certain event, is shown by the black curve. (b) The line plot (top) and partial dependency bar (middle) are shown. The color represents the outcome's predicted risk. At the bottom of the figure is a color map. Taken from [139].

2A. Partial Dependence Plot (PDP) [237]. When an individual feature is changed throughout its range, the PDP displays the black-box's average prediction. Partial dependency is a concept that attempts to demonstrate how a single feature influences the global model's prediction. In PDPs, the connection between an individual feature and the target is represented. As seen in Figure 20(a), for the original data, the red line depicts the average projected risk. The mean of the observed values is represented by a vertical line, and the

Table 8
An overview of visualization-based XAI methods, highlighting advantages and disadvantages.

Method	Ref.	Advantages	Disadvantages	Concept
PDP	[237]	i) Provides a clear interpretation; ii) Intuitive, easy to implement, and shows global effects.	i) Issue with the assumption of independence; ii) Heterogeneous effects are hidden.	Global technique Feature visualization
ICE	[241]	i) Potential to reveal heterogeneous relationships; ii) Fitted values vary over a wide range of relevant factors; iii) Reveals the potential locations and magnitude of variation.	i) Shows a single feature at a time; ii) Not easy to plot the average results; iii) Independence assumption for a single feature.	Global technique Feature visualization
ALE	[228]	i) Able to compute the plots more quickly; ii) The interpretation is extremely apparent; iii) Unbiased plots.	i) Interpretation is more challenging with closely correlated features; ii) Unsteady plots.	Global technique

distribution of observed values is represented by a histogram underneath the figure. A range of one standard deviation around the mean values is shown by dotted lines. Krause et al. [139] visualized how features influence a prediction using a PDP extension. A partial dependency bar has been added to the PDP, this displays a colored depiction of the prediction value across the range of input values that a feature may take, as shown in Figure 20(b). For regression, the partial dependency function is:

$$\hat{f}_{x_\mathcal{P}}(x_\mathcal{P}) = \mathbb{E}_{x_\mathcal{O}}\left[\hat{f}(x_\mathcal{P}, x_\mathcal{O})\right] = \int \hat{f}(x_\mathcal{P}, x_\mathcal{O}) d\mathbb{P}(x_\mathcal{O}), \quad (28)$$

where the set of $x_\mathcal{P}$ features and other features $x_\mathcal{O}$ are utilized in model $\hat{f}$, such that $x_\mathcal{P}, x_\mathcal{O} \subset \mathcal{X}$, the whole feature set. The feature set $\mathcal{P}$ contains one or two features for which the PDP is plotted to analyze their impact on the prediction.

A method for the average calculation of the training data, commonly known as the *Monte Carlo* technique, is used to estimate the partial function $\hat{f}_{x_\mathcal{P}}$:

$$\hat{f}_{x_\mathcal{P}}(x_\mathcal{P}) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}(x_\mathcal{P}, x_\mathcal{O}^{(i)}), \quad (29)$$

where the total number of samples in the dataset is n, and $x_\mathcal{O}^{(i)}$ represents the actual feature values which are not included.

2B. Individual Conditional Expectations (ICE). PDPs are extended to include ICE plots [241]. These plots show the connection between the target and a single feature, rather than the whole model. The difference between a feature's individual behavior and its average behavior may be seen when ICE plots and PDPs are displayed together in the same graph, as shown in Figure 21. The centered and derivative ICE plots are two further extensions of the standard ICE plots that may be used to identify heterogeneity and to investigate the existence of interacting effects [237]. In practice, ICE is defined as follows: for each example, be $x \in \left\{(x_\mathcal{P}^{(i)}, x_\mathcal{O}^{(i)})\right\}_{i=1}^{N}$, the ICE plot $\hat{f}_\mathcal{P}^{(i)}$ is drawn against $x_\mathcal{P}^{(i)}$, while $x_\mathcal{P}^{(i)}$ remains the same.

2C. Accumulated Local Effects (ALE) is a novel technique for visualization methods that does not rely on erroneous extrapolation with associated predictors [228]. The changes

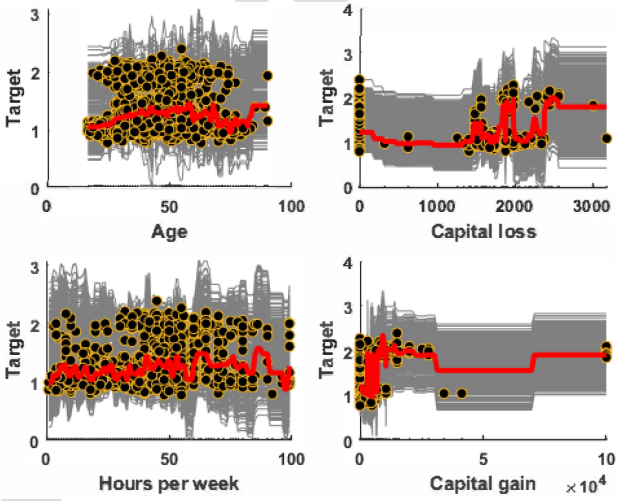


Figure 21: The figure shows income prediction (target variable on Y-axis) based on the employee's age, capital loss, hours per week, and capital gain. The red line shows the average behavior of all features (PDP), and the gray lines show the behavior of individual features (ICE). The selected features and response variables are also presented in a scatter plot (circle markers).

in the predictions are averaged and accumulated over the grid in graphs. It is defined as follows:

$$\hat{f}_{x_p} = \int \mathbb{E}_{x_\mathcal{O}\|x_p}\left[\hat{f}(x_p, x_\mathcal{O})|x_p = x_\mathcal{O}\right] dx_p - const,$$
$$= \int \left(\int \hat{f}(x_\mathcal{P}, x_\mathcal{O}) d\mathbb{P}(x_\mathcal{O}\|x_\mathcal{O} = x_p)\right) dx_p - const.$$

The formula shows three differences from PDP [149]. First, averaging prediction changes rather than the predictions themselves. Second, determining how a feature affects a prediction by adding up the local partial derivatives across the range of features in set $\mathcal{X}$. Third, subtracting a constant from the result such that the ALE plot is centered, i.e., the average effect across the data is 0.

8.3. Example-based Explanation Methods

Example-based explanations are also commonly known as case-based explanations. We have found in the literature the following methods for generating this kind of

Table 9
A comprehensive overview of example-based XAI methods, highlighting their advantages and disadvantages.

Method	Advantages	Disadvantages	Concept
Prototype and Criticisms [158]	Provides intuitive and interpretable explanations to end-users. Can help improve model accuracy and applied to various types of models	May fail to identify important features due to sampling of prototypes. Prone to high variance and bias.	Local technique
Counterfactuals [261]	Provides specific and actionable explanations for individual instances. Helps identify the causal effect of input features.	Computationally expensive and may not scale well with high-dimensional data. Generates explanations that may not be intuitive to end-users.	Generation-based method
Adversarial Example [149]	Can provide insights into the robustness of a model against malicious attacks. Helps identify model vulnerabilities and improve adversarial training.	May not provide meaningful insights into model behavior. Generated adversarial examples may not be representative of real-world data.	Attack-based method

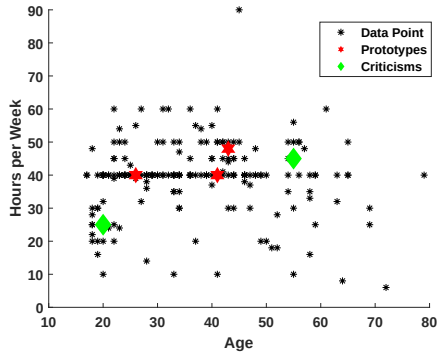


Figure 22: Prototypes and Criticisms for two variables, age and hours per week, from the UCI Income dataset presented with their data distribution.

explanation: prototypes and criticisms, counterfactuals, and adversarial examples. In the upcoming paragraphs, these techniques are discussed, and their respective advantages and drawbacks are outlined in Table 9.

3A. Prototype and Criticisms. *Prototypes* are single instances with the capability to represent the entire dataset. A *criticism* is a data instance that is not included in the collection of prototypes because it is distinct enough for representing complimentary insights [262]. For example, in Figure 22, the small black circles represent data points, prototypes (in red) are manually selected to encompass the data distribution's centers, while criticisms are green diamonds associated with clusters different from those of prototypes.

There is a number of ways for finding prototypes in data. K-medoids [263], a clustering method similar to k-means, is one of the oldest and most popular among them. However, most of these methods provide only prototypes without criticisms. Accordingly, one of the methods recently introduced by Kim et al. [158], called Maximum Mean Discrepancy (MMD-critic), has gained popularity. This method integrates prototypes and criticisms into a single framework. MMD-critic compares the data distribution with the distribution of selected prototypes. Firstly, the user defines the number of prototypes and criticisms to be identified. Then, prototypes and criticisms are discovered using a *greedy search technique*. Criticisms are selected where the distribution of prototypes and the distribution of data varies. For example, MMD-critic is applied to the ImageNet mini dataset to learn different bird breeds as prototypes along with criticisms (see Figure 23).

The following are the fundamental elements in the MMD-critic method: (i) a *kernel function* to analyze the data densities that determine the prototypes; (ii) a *witness function* to measure how the two distributions are different at specific data points in order to identify criticisms; and (iii) a greedy search strategy for prototype and criticism selection. The equation below is used to calculate the squared MMD measure:

$$\mathcal{MMD}^2 = \frac{1}{p^2}\sum_{i,j=1}^{p}\mathbf{k}(z_i, z_j) - \frac{2}{pn}\sum_{i,j=1}^{p,n}\mathbf{k}(z_i, x_j) + \frac{1}{n^2}\sum_{i,j=1}^{n}\mathbf{k}(x_i, x_j).$$

The kernel function is defined as $\mathbf{k}$, p is the number of prototypes z, and n is the number of data points x. The $\mathcal{MMD}^2$ measure is combined with the witness function to find criticisms. The *witness* estimator is defined as follows:

$$witness(x) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{k}(x, x_i) - \frac{1}{p}\sum_{j=1}^{p}\mathbf{k}(x, z_j). \tag{30}$$

There are three ways in which the MMD-critic may improve interpretability: (i) assist in a better understanding of data distributions; (ii) construct understandable models; and (iii) make black-box models understandable [257].

In this context, an interpretable model is defined as:

$$\hat{f}(x) = \arg\max_{i \in S}\mathbf{k}(x, x_i),$$

where the prototype i is selected from the set S that tends to the highest value of the kernel function. The explanation of the model prediction is the prototype itself.

3B. Counterfactuals are "contrary-to-fact" examples [264]. Unlike prototypes, counterfactuals do not need to match with actual training set instances; instead, they may be synthetically generated. Wachter et al. [261] introduced the

Figure 23: The MMD-critic approach learned two bird breeds from the ImageNet mini dataset.

idea of *counterfactual explanations* for a model's decision. The authors defined a loss function that takes an instance of interest x, a counterfactual x', and the desired outcome y'. The loss function is optimized to get the best counterfactual explanation as follows:

$$\mathcal{L}(x, x', y', \lambda) = \lambda.\Big(\hat{f}(x') - y'\Big)^2 + d(x, x') \tag{31}$$

where the factor λ balances between the first and second terms. When λ is high, then the priority are counterfactuals x' with predictions close to the desired outcome y'. On the contrary, when λ is small, the counterfactuals x' are very close to x. The first term in $\mathcal{L}$ represents the quadratic distance between the model prediction for x' and y'. The second term represents the *Manhattan distance* d between x and x', which is defined as:

$$d(x, x') = \sum_{j=1}^{n} \frac{|x_j - x'_j|}{MAD_j} \tag{32}$$

where MAD is the Median Absolute Deviation of feature j over the whole dataset, which is defined as:

$$MAD_j = \underset{i\in\{1,2,\ldots,n\}}{median} \Big|x_{i,j} - \underset{l\in\{1,2,\ldots,n\}}{median}(x_{l,j})\Big|. \tag{33}$$

Counterfactual explanations provide the minimal circumstances that would have led to an alternate conclusion. Dandl et al. [265] published the Multi-Objective Counterfactuals (MOC) approach that enables more detailed post-hoc explainability. To do this, the authors simultaneously minimize four objective losses ($\mathfrak{O}_1, \mathfrak{O}_2, \mathfrak{O}_3, \mathfrak{O}_4$):

$$\begin{aligned}\mathcal{L}(x, x', y', \mathfrak{X}^{obs}) = \Big(&\mathfrak{O}_1(\hat{f}(x'), y'), \mathfrak{O}_2(x, x'),\\ &\mathfrak{O}_3(x, x'), \mathfrak{O}_4(x', \mathfrak{X}^{obs})\Big)\end{aligned} \tag{34}$$

To know in detail, the meaning of each loss function and how they are calculated, the interested readers are kindly referred to [265].

Synthetically generated counterfactuals may not be realistic and therefore yield misleading explanations and jeopardize trustworthiness. To cope with this problem, Suffian et al. [266] proposed the generation of counterfactual explanations with user feedback. Accordingly, the user can set preferences and constraints (e.g., protected features, variation ranges, etc.) with the aim of enhancing the automated explanations which are better aligned with user expectations. Finally, in addition to numerical counterfactuals, it is also possible to generate linguistic counterfactuals as proposed by Stepin et al. [267]. Thanks to the ability of fuzzy sets and systems to compute with words and information granules, the generated counterfactuals can be verbalized in natural language.

3C. Adversarial Examples can be used to fool DNNs [149], but they can be also used for generating analogical and contrastive explanations. On the one hand, analogical explanations are supported by analogical reasoning, i.e., by searching for two explanatory evidences coming from familiar and unfamiliar domains [268]. On the other hand, contrastive explanations are supported by contrastive reasoning, i.e., by searching for two competing or opposite explanatory evidence [80].

8.4. Game Theory Methods

In 1953, Lloyd Shapley wanted to know how much each player in a coalition game contributes [242]. Afterward, researchers in the field of ML used this approach to investigate what is the link between interpretability and ML predictions. In this context, the "game" is a single instance of a dataset's prediction in a task. The "gain" is the difference between the actual prediction for the given prediction and the average of predictions for all instances in the dataset. The "players" are the instance's feature values who work together to obtain the gain, i.e., the Shapely value of a feature tells us how much it contributes to a particular prediction outcome.

4A. Shapley Values. The question is how each attribute influences a certain data point's prediction. Here is an example of how a linear model can do prediction for a given dataset:

$$\hat{F}(x) = w_0 + w_1x_1 + w_2x_2 + ... + w_nx_n, \tag{35}$$

where x_i is the $i-th$ instance/feature value from which the contributions are calculated. The weight for feature i is w_i for a total number of features n. The $i-th$ feature contribution, Φ_i can be computed as:

$$\Phi(\hat{F}) = w_ix_i - \mathbb{E}(w_i\mathcal{X}_i) = w_ix_i - w_i\mathbb{E}(x_i), \tag{36}$$

where $\mathbb{E}(w_i\mathcal{X}_i)$ is the estimated mean effect for feature i. The contribution is equal to the difference between the feature and the mean effect. When all feature contributions for one

Table 10
Game theory-based XAI methods, together with their advantages and disadvantages.

Method (Ref.)	Advantages	Disadvantages	Concept
Shapley Values [242]	i) Fairly Distributed; ii) Solid theoretical foundation; iii) Contrastive Explanations.	i) High computing time and misinterpretation; ii) Cannot be used for sparse explanations; iii) Suffers from the inclusion of unrealistic data instances [149].	Coalitional Game Theory
SHAP [141]	i) Computes many Shapley values; ii) All Shapley values advantages connect to LIME; iii) Fast implementation for tree-based models.	TreeSHAP produces unintuitive feature attributions; ii) Does not provide causality. A problem of misinterpretation; iii) KernelSHAP is slow and ignores feature dependence (TreeSHAP solves it) [149]	Optimal Shapley values Game Theory

instance are combined together, it results in:

$$\begin{aligned}\sum_{i=1}^{n}\Phi(\hat{\mathcal{F}}) &= \sum_{i=1}^{n}\Big(w_i x_i - \mathbb{E}(w_i x_i)\Big)\\ &= \Big(w_0 + \sum_{i=1}^{n} w_i x_i\Big) - \Big(w_0 + \sum_{i=1}^{n}\mathbb{E}(w_i x_i)\Big)\\ &= \hat{\mathcal{F}}(x) - \mathbb{E}\Big(\hat{\mathcal{F}}(x)\Big).\end{aligned} \tag{37}$$

The weighted, total contribution of all potential feature values is the Shapley value. A value function *val* of players in subset $\mathcal{S}$ is used to define the Shapley value as:

$$\Phi_i(val) = \sum_{\mathcal{S}\subseteq\{x_1,x_2,\ldots,x_n\}\{x_i\}} \frac{|\mathcal{S}|!(p-|\mathcal{S}|-1)!}{n!} \Big(val(\mathcal{S}\cup\{x_i\}) - val(\mathcal{S})\Big). \tag{38}$$

4B. Shapley Additive Explanation (SHAP) suggested by [141], is a unified way to understand the output of any ML model. SHAP is a technique for explaining individual predictions using the coalitional game's best Shapley values [141]. A player can be represented by a single feature value, such as in tabular data. A player can also be made up of a collection of feature values. For instance, pixels can be grouped into superpixels, and the information to make the prediction that describes the image is spread among them. The Shapley value explanation is an *Additive Feature Attribution approach*, a linear model, which is a step forward that **SHAP** brings to the table. According to SHAP, the explanation is given as follows:

$$g(\hat{\mathcal{Z}}) = \Phi_0 + \sum_{i=1}^{\mathcal{M}} \Phi_i \hat{\mathcal{Z}}_i, \tag{39}$$

where g stands for the explanatory model, the feature attribution for $i-th$ feature is $\Phi_i \in \mathbb{R}$, the maximum size of the coalition is $\mathcal{M}$, the coalition vector (the simplified features) is denoted by $\mathcal{Z} \in \{0,1\}^{\mathcal{M}}$. Where 1 in the coalition vector indicates that the relevant feature value is "present", whereas 0 indicates that the feature is "missing". SHAP has properties such as local accuracy, missingness, and consistency in addition to the Shapley value properties of efficiency, symmetry, dummy (Shapley value equal to 0), and additivity [149]. Table 10 compares Shapley values and SHAP techniques.

Moreover, KernelSHAP, an alternative kernel-based estimate strategy based on Shapley values inspired by local surrogate models, and TreeSHAP, an efficient estimation strategy for tree-based models, were proposed by the SHAP authors. SHAP values can be determined for any tree-based model, in contrast to other approaches that rely on surrogate models such as linear regression or logistic regression. Many global interpretation techniques based on aggregations of Shapley values are also included in the family of SHAP-based techniques.

The following is a synopsis of SHAP in terms of whether the approach is based on local or global interpretations.

- *Local Interpretability:* SHAP values are assigned to each observation, as a result, their transparency is substantially improved. This allows us to see why a case is predicted in terms of the predictors' contributions. Due to the local interpretability, the impacts of the components may be localized and compared.
- *Global Interpretability:* The combined SHAP values can show how much each predictor contributes to the target variable, either favorably or negatively. This is similar to the variable importance plot, except it can also display if each variable has a positive or negative connection with the target.

8.5. Knowledge Extraction Methods

It is challenging to describe how black-box ML models behave internally. For example, ANN algorithms may change the filter/kernel in the hidden layer, which can lead to intriguing internal representations of the whole network. The task of extracting explanations from an ANN implies retrieving the knowledge learned by an individual layer during training and encoding it in a human-understandable format. Several publications in the literature (see Tables 11 and 12) offer methods for extracting information from black-box models. These methods depend primarily on two techniques: Rule Extraction and Model Distillation.

5A. Rule Extraction According to Mark Craven [178], the rule extraction process produces an understandable but rough approximation of a network's predicted behavior from the training data and the trained ANN.

There are different types of rule extraction techniques, depending on the type of rule under consideration:

- *IF-THEN Rule:* It is the most generic form of a simple and comprehensible conditional statement:

$$\mathbb{IF} \quad x \in \mathcal{X} \quad \mathbb{THEN} \quad \mathcal{Y} = y(i) \tag{40}$$

Table 11
A family of Rule Extraction Systems (RULES). SETAV - SET of Attributes and Values, PRSET - Partial Rules SET, SRI - Scalable Rule Induction Algorithm, IS - Immune System inspired, TL - Transfer Learning, IT - Incremental Transfer, REX - Rule Extractor

RULES	Reference	Upgrade features
RULES-1	[269]	Extracting IF-THEN rule by considering all examples.
RULES-2	[270]	RULES-1 have been upgraded to provide individual example analysis.
RULES-3	[271]	New version of RULES-2 with more general features.
RULES-3+	[272]	RULES-3 has been extended to include two new features: 1) SETAV and 2) PRSET.
RULES-4	[271]	The first incremental learning system that updates and refines previously learned information in preparation for new examples.
RULES-5	[273]	The first version of RULES to deal with continuous attributes without discretizing them.
RULES-5+	[274]	A novel rule space representation method that improves performance.
RULES-6	[275]	It is an expansion of RULES-3 plus that makes a scalable version of the RULES family.
RULES-7	[276]	The RULES-6 extension that focuses on one seed at a time.
RULES-8	[277]	A new version that takes into account online continuous attributes.
RULES-F	[278]	RULES-5 are extended to accommodate both continuous characteristics and continuous classes.
RULES-F+	[274]	RULES-F included a new rule space representation technique.
RULES-SRI	[279]	Extension version of RULES-6 to enhance the scalability.
RULES-IS	[280]	An immune system-inspired incremental algorithm.
RULES-3EXT	[281]	An enhanced version of RULES-3.
RULES-TL	[282]	Another scalable method that has been suggested to improve speed and performance while also including more intelligent features.
RULES-IT	[283]	An incremental version based on the RULES-TL for dealing with big and incomplete problems incrementally.
REX-1	[284]	RULES-3, RULES-3+, and RULES-4 were improved to speed up the process and create simpler models with fewer rules.

The output will be tagged to a certain class if the condition is true, i.e., x is a member of $\mathcal{X}$. The expressive power of a rule extraction algorithm is directly related to the *if... then...else...* rule structure. For example, $\mathcal{Z}$ is medium if $\mathcal{X}$ is low and $\mathcal{Y}$ is high, where low, medium, and high are fuzzy sets with associated membership functions. The interested reader is referred to [285] for further details on how to deal properly with fuzzy rules.

- *M-of-N rules:* A Boolean expression is used to look for rules with this strategy. When $\mathcal{P}$ of $\mathcal{Q}$ sets are fulfilled, the expression is completed. This strategy is both effective and universal [286]. M-of-N rules are written as $\mathbb{IF}\ \mathcal{P}$ of $\{\mathcal{Q}\}\ \mathbb{THEN}\ \mathcal{Z}$.

Thus, two main categories have been chosen to represent the extracted rules: (i) Propositional/Boolean logic and (ii) Non-conventional logic. Notice that, rule extraction facilitates gaining insight into ML models. Rule extraction techniques include, among others, fuzzy modeling [287], genetic programming [288], boolean rule extraction [289], and the decomposition approach [290]. In addition, Andrew et al. [39] and Gopi [291] suggested multidimensional modalities for extracting rules.

Regarding the relation between the extracted rule and the trained NN architecture, there are three distinct types of methods: (a) *Decompositional* methods operate on the neuron level rather than over the whole NN design; (b) *Pedagogical* methods operate disregarding the NN architectural design; and (c) *Eclectic* methods are a combination of decompositional and pedagogical methods.

Decompositional Methods operate by breaking down a network into its constituent neurons. The results from each neuron are then combined to represent the whole network. After decomposing an ANN, it may be scrutinized and translated into rules that are viewed as composing a transparent model [39, 297]. A fundamental requirement for rule extraction methods that use this approach is that the extracted output from each neuron must be in the form of a consequential rule, i.e., a binary result (yes/1 or no/0). Thus, each hidden neuron can be thought of as a *step* function or a Boolean rule, this reduces the rule extraction problem to determining the instances in which the rule is true.

Pedagogical Methods consider rule extraction as a learning problem in which the learning task pay attention to the network parameters and input features [297]. Therefore, pedagogical methods are aimed at extracting rules that directly relate inputs to outputs. These methods are often employed in combination with a symbolic learning algorithm. The fundamental concept is to utilize the trained ANN to create instances for the learning algorithm. These methods include, among others, Valid Interval Analysis (VIA), reverse engineering, and sampling methods [305].

Eclectic methods include aspects of both decompositional and pedagogical rule extraction methods. On the one hand, a decompositional method is usually more transparent than a pedagogical one, but they do operate in layers. As a consequence, decompositional methods may be time-consuming and laborious. On the other hand, the pedagogical methods outperform the decompositional ones in terms of computing burden and execution time [306]. In terms of ANN architecture, pedagogical methods also offer the benefit of flexibility. Techniques that use knowledge of the trained ANN's internal architecture and weight vectors to supplement a symbolic learning method are classified as eclectic methods [307].

5B. Model Distillation is another approach that comes under the knowledge extraction category. Distillation implies transferring information (dark knowledge) from a teacher network (e.g., a DNN) to a student network (e.g., a shallow NN) via model compression [308, 309]. Model compression was first suggested to decrease a model's runtime computing cost, but it has subsequently been used to improve explainability. Tan et al. [310] explored how to translate complicated models into interpretable ones via model distillation. Che et al. [311] proposed *Interpretable Mimic Learning* as a method for learning phenotype features that are interpretable for generating robust predictions while imitating the performance of black-box DL models. *DarkSight*, a

Table 12
Rule-based methods for knowledge extraction from black-box models.

Techniques	Type of ANN	Method	Rule Extraction Approach	Drawback
DIFACON-miner [292]	MLP	Decompositional	IF-THEN	It is not an application to DNN.
CRED [293]	MLP	Decompositional	Decision Tree	Discretization is not used in this method and may not apply directly to DNN.
FERNN [294]	MLP	Decompositional	M-of-N split, IF-THEN	The relevance of DNN is not addressed.
KT [295]	MLP	Decompositional	IF-THEN	DNNs are ignored in the analysis.
Tsukimoto's algorithm [296]	MLP & RNN	Decompositional	IF-THEN	This method has polynomial computational complexity.
TREPAN [297]	MLP	Pedagogical	M-of-N split, Decision Tree	The hidden layer in NN is the only one.
HYPINV [298]	MLP	Pedagogical	Hyperplane Rule	DNN is not being considered.
BIO-RE [299]	MLP	Pedagogical	Binary Rule	A shallow MLP is used to test the algorithm.
KDRuleEX [300]	MLP	Pedagogical	Decision Tree	DNN is not being considered, and the design of ANN is not disclosed.
RxTEN [301]	MLP	Pedagogical	IF-THEN	a traditional feedforward neural network is employed.
ANN-DT [203]	MLP	Pedagogical	Binary and Decision Tree	–
RX [302]	MLP	Eclectic	IF-THEN	A shallow MLP is used to test the algorithm.
KAA [303]	MLP	Eclectic	IF-THEN	–
DeepRED [304]	DNN	Decompositional	IF-THEN	–

KAA: Kahramanli and Allahverdi's Algorithm

Table 13
Neural-based XAI methods, their advantages and disadvantages.

Method (Ref.)	Advantages	Disadvantages	Concept
SA [232]	i) Provides unique solution, training free process, and fast computation [245]; ii) Identifies weak and prominent features.	i) Inconsistent procedure; ii) Generates noisy explanation maps.	Input alteration
LRP [236]	i) Scalable and explainable to complicated DNNs; ii) Calculates the weights for each neuron to improve interpretability.	i) Usable with ReLU activation; ii) Compatible with backpropagation networks.	Propagation rules
TCAV [230]	i) Provides human-interpretable explanation of any neural network; ii) Works on high-level features vector.	i) Reduced effectiveness with strong correlations in the data; ii) Inappropriate with a random selection of input concepts [27].	Concept method

visualization technique for understanding the predictions of black-box models on datasets inspired by the concept of dark knowledge, was proposed in [312]. This approach integrates concepts from DNN visualization, knowledge distillation, and dimension reduction. For further information interested scholars may look at [310, 313].

8.6. Neural Methods

This section concentrates on neural network interpretation techniques. These techniques explain specific predictions, simplify neural networks, or visualize the features and concepts that a neural network has learned. Table 13 summarizes the strengths and weaknesses of the most relevant techniques under consideration.

6A. Influence Methods. By altering the input or internal elements and analyzing which ones (and how much) change model performance, these methods assess the significance of a feature [70]. Then, ML models can be debugged, while their behavior and prediction explanations can be improved by finding influential training examples. There are three different techniques in the literature for determining the significance of an input variable: (i) feature importance, (ii) Layer-wise Relevance Propagation (LRP), and (iii) Sensitivity Analysis (SA).

Feature Importance. A data instance with a significant impact on the trained model is an important feature. When a model is retrained with that specific instance removed from the training data, the model parameters or predictions vary to a large extent, indicating how important that instance is. In this way it is possible to assign a degree of significant value to each feature, this is especially useful when the selected instance has a significant impact on model performance. The significance value of an instance for the goal y determines whether it has an influence on the trained model. A useful example of the LR model may be seen in Figure 24.

Feature importance is calculated using the change in the model's error seen in the feature permutation process. As the model depends on features for its prediction, a feature is considered important if rearranging its values raises the model's error. A feature is irrelevant if rearranging its values has no effect on the model's error since the feature was disregarded for prediction in the input instance. For example, Lei et al. [314] proposed *Leave-One-Covariate-Out* (LOCO) inference that uses local feature importance. Fisher et al. [315] suggested *Model Class Reliance* (MCR) as a model-agnostic variant of feature significance based on this approach. The MCR algorithm has the following steps for finding feature importance:

1. Input - A model f, feature matrix $\mathcal{X}$ with target vector y, and error function $\mathcal{L}(y, f)$.
2. Measure the error of the original model using Mean Squared Error (MSE); $e^* = \mathcal{L}\Big(y, f(\mathcal{X})\Big)$
3. For each feature $i = 1, ..., p$:
   - Permute the feature i, and get the feature matrix $\mathcal{X}^{pre}$.
   - Calculate the permuted error; $\hat{e} = \mathcal{L}\Big(y, f(\mathcal{X}^{pre})\Big)$.
   - Estimate the permuted feature importance $\mathcal{FI}^i = \frac{\hat{e}}{e^*}$.

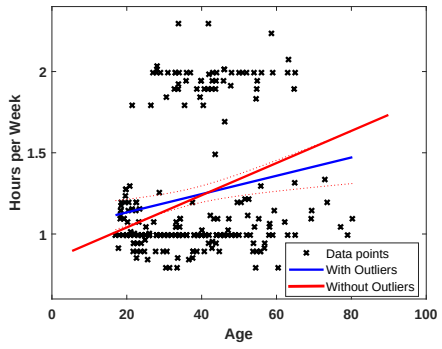


Figure 24: A linear model was trained on two cases, one with unimportant features and one without unimportant features. In the instance, without unimportant features, the slope produced by the model changes significantly in contrast to the instance with unimportant features.

Using Feature Importance on Training or Testing Data? When error estimates are based on the same training data on which the model was initially trained, the model error or performance measurements appear to be much better than they are in reality. Given that feature importance permutations rely on having accurate model errors, unseen data must be considered here [257]. Finding the importance of features using training data leads us to assume that certain features are significant for predictions, however, the model may be overfitting, so in practice, these features could be unimportant. Therefore, when it is necessary to know how much a model relies on each feature for making predictions, we use the feature importance approach in the training data. On the other hand, if it is required to know how much the feature contributes to the performance of the model on unseen data, we use the feature importance approach in the testing data. According to our review, there is no study in the literature on the topic of feature importance based on training vs. test data. To get a deeper understanding of this area, further research is required.

Layer-wise Relevance Propagation (LRP) [236] has proven to be widely applicable while performing very well in benchmark experiments [316, 317]. The LRP algorithm was suggested as another method for computing relevance. Starting from a network's output layer and backpropagating up to the input layer, LRP redistributes the prediction functions in their opposite order. Relevance conservation is a key feature of this redistribution procedure. It presupposes that the classifier, in its basic form, may be broken into multiple layers of computation. A typical forward pass is performed on a network, as illustrated in Figure 25, and the activations at every layer are recorded. Following that, using a specific set of rules, a score calculated at the output layer is backpropagated. To formulate the LRP problem, consider $\mathcal{RS}_k$ to be the k-th neuron's relevance, while j and k are the indices of two neurons in successive layers. The share of the relevance score, $\mathcal{RS}_k$ redistributed to neuron j, may be defined as $\mathcal{RS}_{j\leftarrow k}$. The following conservation property

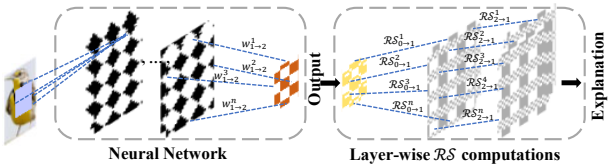


Figure 25: LRP: The rationale behind LRP is to decompose a model's prediction function into a sum of layer-by-layer relevance values. LRP can be thought of as the Deep Talyor Decomposition of a prediction when used with ReLU networks. $w^n_{1\to 2}$ denotes a weight propagating from layer 1 to layer 2 for the n-th neuron. A similar notation may be applied to $\mathcal{RS}$ to generate an explanation.

must hold:

$$\sum_j \mathcal{RS}_{j\leftarrow k} = \mathcal{RS}_k. \tag{41}$$

Similarly, the contribution to a neuron's relevance coming from the higher layer may be aggregated to produce the relevance in the lower layer:

$$\mathcal{RS}_j = \sum_k \mathcal{RS}_{j\leftarrow k}. \tag{42}$$

By combining these two equations, the relevance conservation property between two layers may be obtained. Therefore, the sequence of equalities for the whole network can be written as:

$$\sum_{i=1}^{d} \mathcal{RS}_i = \ldots = \sum_j \mathcal{RS}_j = \sum_k \mathcal{RS}_k = \ldots = \mathcal{F}(x), \tag{43}$$

where x is the input data, and $\mathcal{F}(.)$ is the function that encodes the concept at the output neuron.

Sensitivity Analysis (SA) is another approach for identifying the most relevant input features [318, 319]. The most important input features are those with the greatest impact on the output. This approach has already been used in applications such as mutagenicity predictions [260], medical diagnosis [320], or ecological modeling [321]. SA is increasingly utilized to explain results of image classification in specific terms [322, 323]. In the context of ML and DL, the effect of input and/or weight perturbations on the model output is referred to as its sensitivity [232]. In this approach, data is deliberately perturbed, and the resulting output from the model is used to check its behavior and the stability of the model outputs. As showing model stability as data changes over time improves confidence in ML results, visualizing the outcomes of SA is considered a model-agnostic explanation method. SA is defined formally in terms of a relevance score as follows, based on the local gradient x of model $\mathcal{F}$:

$$\mathcal{RS}_i(x) = \left(\frac{\partial \mathcal{F}}{\partial x_i}\right)^2. \tag{44}$$

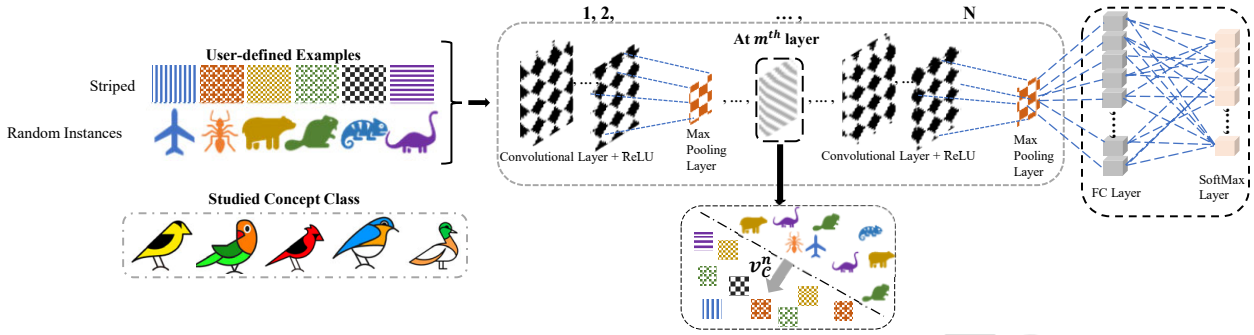


Figure 26: The DNN is supplied with a user-defined collection of striped samples and random instances. For the examined class (bird), labeled training data was also provided to the network. The sensitivity of the network to the concept behind the examined class may be quantified using Concept Activation Vectors (CAVs). CAVs are created by teaching a linear classifier to discriminate between the activation generated by a concept's instances and the activations caused by examples at the m-th layer. The vector orthogonal to the classification border is known as the CAV; v_C^n. The directional derivative $S_{C,\subsetneq}$ is used by the Testing CAV to evaluate the conceptual sensitivity of the studied class (bird).

The above relevance scores are decomposed into the gradient square norm as follows:

$$\sum_{i=1}^{d} \mathcal{R}_i(x) = ||\nabla \mathcal{F}(x)||^2. \tag{45}$$

It is worth noting that SA does not offer an explanation itself, but rather shows explanation variations. As a result, the goal of SA is rarely to explain any connections that have been discovered. However, SA is often used to check for model trustworthiness and stability, as a tool for identifying or removing irrelevant input features.

6B. Concept Methods. Concerns about bias in ML are valid, and the stakes are even higher when it comes to AI. The concept-based methods are introduced in order to make AI more trustworthy and transparent.

Concept Activation Vectors (CAVs) were proposed by Kim et al. [230]. This method provides human-friendly explanations of the internal states of NNs globally. Consider the model $\mathcal{F}(.)$ as a space $\mathcal{V}_m$ in the form of a vector with a basis vector v_m. The vector space $\mathcal{V}_h$ represents the space of human understanding with a basis vector v_h. Thus, in order to explain model decisions in a human-friendly way, an explanation function, $g : \mathcal{V}_m \rightarrow \mathcal{V}_h$, may produce the human-understandable concepts $\mathcal{C}$, as explanations.

A vector with n activations can be determined for a particular dataset in order to express a concept of human interest. Activations in the layer n, generated by a concept set instance against random examples, may be used to find such a vector. CAVs, as shown in Figure 26 (gray arrow), are orthogonal to a hyperplane that separates instances without a concept from instances with a concept in the layer activations [230]. A positive concept $\mathcal{P}_C$ denotes a vector heading to a set of concepts of human interest, whereas a negative concept $\mathcal{N}_C$ denotes that there is no concept of human interest (random inputs). This approach uses a binary classification task in which a classifier v_n^C distinguishes between the layer activation of two sets: $\mathcal{F}_n(x) : x \in \mathcal{P}_C$ and $\mathcal{F}_n(x) : x \in \mathcal{N}_C$.

In addition, the use of CAVs for testing AI models is known as Testing CAVs (or just TCAVs for short). TCAV utilizes directional derivatives to assess the sensitivity of a model, $\mathcal{F}$, in a similar way to gradient-based methods. The sensitivity of a model is determined by shifting the input in a direction toward the concept $\mathcal{C}$ for a particular layer n. Consider for an input x, $\mathcal{H}_c(x)$ is the gradients logit of layer n for class c, then the conceptual sensitivity $S_{C,c,n}$ of the class c to C can be computed as the directional derivative for a concept vector v_C^n:

$$\begin{aligned} S_{C,c,n} &= \lim_{\in \to 0} \frac{\mathcal{H}_{n,c}\left(\mathcal{F}_n(x) + \in v_C^n\right) - \mathcal{H}_{n,c}\left(\mathcal{F}_n(x)\right)}{\in} \\ &= \nabla \mathcal{H}_{n,c}\left(\mathcal{F}_n(x)\right) \cdot v_C^n. \end{aligned} \tag{46}$$

Moreover, the sensitivity of all classes of inputs can be computed with TCAV techniques. For the entire dataset $\mathcal{X}_c$ with class c, TCAV may be defined as:

$$\mathbb{TCAV}_{C,c,n} = \frac{\left|\{x \in \mathcal{X}_c : S_{C,c,n} > 0\}\right|}{|\mathcal{X}_c|}. \tag{47}$$

The CAV method has been built upon and enhanced further in numerous research articles where techniques such as Automatic Concept-based Explanations (ACE) [324], Causal Concept Effect (CaCE) [325], Ground truth CaCE (GT-CaCE) [325], Variational Auto Encoders based CaCE (VAE-CaCE) [325], and ConceptSHAP [326] have been put forward.

9. Assessment of Explanations

After revisiting different methods for dealing with data, models, and post-hoc explainability in the proposed XAI taxonomy, it is now time to go deeper with the fourth axis in our proposal. Accordingly, this section pays attention to the evaluation of explainability.

Table 14
Desirable qualities of explanation methods, individual explanations, and human-friendly explanations.

Type	Qualities	Description
Explanation Methods	Translucency	Expresses how deeply an explanation approach probes the model [328].
	Portability	Expresses how successfully an approach covers a wide variety of models [329].
	Explanatory Power	The number of events that can be explained using the explainability technique [328].
	Algorithmic Complexity	The computational complexity of explanation algorithms [329].
	Generalizability	To increase the utility because of the diversity of model architectures [288].
Individual Explanations	Fidelity	How closely the explanations match the prediction model's behavior [288].
	Consistency	To extent various models learned on the same problem give similar explanations [29].
	Accuracy	To generalize an explanation of a specific decision to previously unknown situations [52].
	Stability	The frequency with which identical explanations are offered for the same instances [288].
	Comprehensibility	The readability and length of explanations [288].
	Certainty	To a model decision's degree of certainty [330].
	Interpretability	The ease with which people can comprehend the model and/or the outcomes [52].
	Representativeness	How well the explanation depicts the most important aspects of the explanation.
Human-friendly Explanations	Explanation using contrastiveness	The ability to represent distinct properties between the instance being explained and a reference point [261].
	Specificity	Capability of providing particular reasons indicating which explanations are the key reasons for a prediction [52].
	Sociological	The social context and intended audience of the model should be considered while choosing the most applicable explanation [52].
	Abnormality	Identification of the odd circumstances that might have a substantial influence on the outcome [29].
	Factuality	Plausibility and relevant to other examples' predictions [29].
	Fairness	The predictions do not include any implicit or explicit bias against targeted users [52].
	Privacy	Assurance of the security of sensitive data [288].
	Reliability	Ensuring that minor input modifications do not have a significant influence on the model prediction [261].
	Causality	The identification of cause-and-effect relations between inputs and outputs in a given model [328, 331].

Furthermore, achieving progress in XAI research that measures the level of explainability for AI systems has gained importance after proposals for EU laws regulating AI, and after current standardization activities, that would transform AI systems breakthroughs into the *de facto* regulatory norm [327]. All regulatory actions agree on the need to assess carefully the goodness of automated explanations.

Desiderata of Explainability: When it comes to delivering an explanation, it is advantageous for a model to have certain desirable properties. They are defined in terms of explanation methods, individual explanation properties, and human-friendly explanation capabilities. Thus, Table 14 comprises a list of traits that every explanation should have, based on our review of the literature. These characteristics may be used to evaluate and compare various explanation approaches.

In this regard, Robnik et al. [328] specified certain characteristics that are desirable for high-quality explanations. Given that the recipients of these explanations are humans, analyzing what makes an explanation human-friendly is important. In addition, Miller [52] undertook a comprehensive study of explanatory articles in the humanities. The degree to which a model is explainable, along with its privacy and non-discrimination promises, has a great impact on how much human users will trust it. In addition, the degree of trust in a model increases when it is built in accordance with users' monotonicity constraints [332]. Usability is another aspect that raises a model's level of confidence [333]. Individuals are more inclined to trust a model which provides them with information that helps them understand how it completed its task. In this scenario, an interactive and questionable explanation is preferable to a printed and static one.

The majority of the existing approaches are built with ill-defined or very general explainability aims and often lack well-defined context-specific use cases. As a consequence, methodologies are created without a thorough grasp of the unique needs of a certain domain and use case, this results in poor adoption and sub-optimal outcomes. Nonetheless, it is normal to see one or more assessment settings proposed in pioneering publications about XAI approaches, most of which are centered on explanatory desiderata [231, 334].

User evaluation is sometimes only emulated [29] or even removed entirely from the assessment process [234]. The value of explanations is significantly influenced by how valuable these explanations prove to be for end-users in the decision-making process [87, 335]. Therefore, in this study, we consider human-in-the-loop approaches for evaluating automated explanations. Accordingly, end-users must be involved in the review process, preferably in a setting with real tasks and data. Furthermore, measurements should represent a user's performance, e.g., the accuracy or speed with which judgments are made.

It is worth noting that the focus of this section is on XAI assessment with end-users as the target audience. This is because the human user is generally the final decision-maker. We pay special attention to end-users who are in roles of responsibility, such as judges, doctors, or other domain experts. Key assessment algorithms are categorized based on their appearance in the literature for XAI systems, as illustrated in Figure 27. In addition, Table 15 contains an overview of XAI assessment methodologies.

Table 15
Summary of assessment methods for XAI.

Methods		References
Cognitive Psychological Theories	Failure and Output	[29, 231, 349, 337, 350]
	Model understanding	[336, 340, 351, 341, 230, 352, 204, 338, 353, 342]
Understandability and Satisfaction	Understandability	[354, 345, 348, 346]
	Satisfaction	[355, 345, 338, 346, 354, 348, 356, 357, 358, 359, 360]
Trust and Transparency	Trust	[361, 362, 363, 364]
	Transparency	[361, 365, 366, 367, 368]
Assessment by Human-AI Interface	Model Performance	[369, 359, 29, 370, 371, 343]
	User Performance	[346, 204, 356, 343, 372, 373, 374]
Computational Assessment	Explainer Fidelity	[29, 231, 375, 249, 376, 377, 378, 253]
	Model Trustworthiness	[238, 379, 141, 380, 381]

9.1. Cognitive Psychological Measures

In the domain of XAI, explanations aid users in developing a mental image of how the AI operates. Researchers in the field of HCI keep into account the mental state of humans to see how well they comprehend intelligent technologies in a variety of settings. For instance, how users comprehend a smart grid system was investigated in [336] and adjusted to uncertainty in ML in terms of time for predictions arrival [337]. Cognitive psychology theories may be used to describe a formal representation of how humans interpret a system. The efficiency of explanations in conveying a model's decision-making process may be verified by looking at the mental state of the human user. Furthermore, psychology research has also looked at the types [338], structure [89], and roles of explanations [339] in discovering the fundamental basis for good explanations in order to improve user comprehension of AI systems.

A user's understanding of AI systems may be investigated by questioning the related decision-making process. Accordingly, some researchers have looked at how users understand AI agents [340, 341] and algorithms [342] in order to determine what kind of explanation is preferred. During the design process for adding explainability to AI systems, users' attention and expectations should also be taken into account [343].

9.2. Understandability and Satisfaction

When assessing explainability, it is important to consider the users' understanding of and satisfaction with the explanations given. Despite the fact that there are implicit ways for measuring user satisfaction [344], a large portion of the research relies on qualitative assessments of user satisfaction such as surveys and interviews. For instance, Gedikli et al. [345] and Lim et al. [346] assessed distinct explanation formats based on user satisfaction ratings.

Researchers have utilized a variety of subjective and objective metrics for quantifying understandability and adequacy of sufficiency [52]. For instance, Curran et al. [347] ranked and coded user transcripts to determine how well users understood the explanations given in a computer vision challenge. Participants showed varying degrees of trust in the correctness of the explanations, this was based on the clarity and understandability of the explanations, despite the fact that they all came from the same model. According to Lage et al. [348], increasing the complexity of an explanation decreases satisfaction. The length and intricacy of explanations have an impact on both understandability and satisfaction, in addition to accuracy and response time. Confalonieri et al. [177] measured the perceived understandability of explanations by users through task performance, namely accuracy and time of response, and subjective measures such as confidence in their answers and explicit understandability provided in a Likert scale.

9.3. Trust and Transparency

When the decision-making process in a model is thoroughly understood, the model becomes transparent. Transparency promotes trust in the model. Trust is an emotive and cognitive component which determines how a system is perceived, either positively or negatively. Various types of trust, such as the initial trust of a user as well as the building of trust through time have been described in the following ways: (i) Swift trust [382], (ii) Default trust [383], and (iii) Suspicious trust [384].

Prior information and beliefs have a role in forming the initial state of trust; however, trust and confidence may evolve over time as the system is explored and experienced. Common variables used to assess and study trust include user knowledge, familiarity, technical competence, confidence, emotions, beliefs, faith, and personal attachments [385, 386]. These variables may be quantified by explicitly questioning users about their experiences with a system during and after usage. For instance, Yin et al. [368] and Nourani et al. [361] found that over time, both the declared

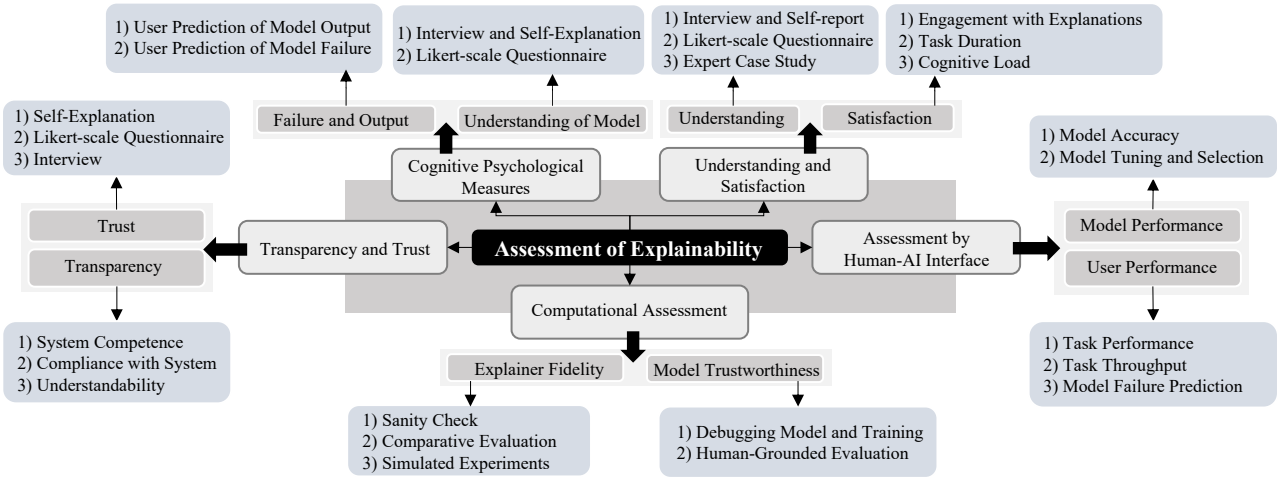


Figure 27: Relationship among assessment methods for XAI and their desiderata.

accuracy of a model and the accuracy perceived by the user influenced user trust.

In addition, multiple scales can be used to evaluate user perceptions of system predictability, dependability, and safety. Cahour and Forzy [364] proposed a thorough trust assessment setup that evaluates three ways a system presents itself to users by measuring user trust in terms of three different trust scales. Another study compared user trust to explanations for AI decisions in terms of transparency [366]. The authors used perceived understandability to assess user trust and found that clear explanations may help mitigate the negative consequences of trust loss. Bussone et al. [363] used a Likert scale and think-aloud to assess user trust in a clinical decision-support system and discovered that factual explanations resulted in increased user trust. In addition, Stepin et al. [387] used Likert scales to evaluate how humans appreciate trustworthiness of automated counterfactual explanations.

9.4. Assessment of Human-AI Interface

One of the main goals in the XAI research field is to assist end-users in becoming more effective in the use of AI decision-support systems. As a result, the human-AI interface can be judged by the performance of the human in the given task. For instance, to assess the influence of various forms of explanation, Lim et al. [346] examined the performance of human users in terms of task completion time and success rate while using AI systems with those various forms of explanation. Another benefit of assessing the human-AI interface is to assist in the verification of the model's output [373, 374] and in the debugging of interactive AI models designed for specific requirements [369, 372]. To achieve these aims, Myers et al. [388] created a framework in which users may ask *why* and *why not* questions while expecting an intelligent interface to respond reasonably to those questions.

Domain specialists may scrutinize models and change their hyper-parameters to facilitate the AI system's analysis. This process is guided by visualizing the internal structure of the model, its details, and the uncertainty in the model outputs. This corpus of work highlights how important it is to include user feedback in order to enhance model outcomes. TopicPanaroma [370] is an example of a text analysis visual analytical tool that was evaluated by two domain experts. FairSight [389] is another visual analytic solution that, through visualizing, evaluating, diagnosing, and reducing biases, aids fair data-driven decision-making.

In addition to domain experts, AI specialists and developers can employ visual analytics to discover defects in the architecture of their models. For example, LSTMVis [360] and RNNVis [390] are both tools that may be used to interpret neural models for NLP applications, with the aim of better understanding some training issues and, at the end, enhancing classification and prediction performance. DGMTracker [391] is another example of a tool that provides visual representations of training dynamics. All these tools assist users in visualizing the internal mechanisms of a given model.

9.5. Computational Assessment

Due to the preference of users for simpler explanations, only relying on human assessments of explanations may result in convincing explanations instead of transparent systems. As a result of this issue, Herman [136] argued that computational approaches, rather than human-subject investigations, should be used to assess the fidelity of explanations. The accuracy of an approach in creating genuine explanations, such as the correctness of a saliency map, is referred to as the fidelity of an explainer. As a consequence, a set of computational methods for assessing the validity of produced explanations, the uniformity of explanation outcomes, and the fidelity of explainability methodologies,

in relation to the original black-box model have been developed.

For example, Zeiler and Fergus [238] investigated the fidelity of a CNN visualization tool in terms of the validity of explanations in detecting model flaws; using this tool resulted in enhanced prediction outcomes. Other techniques of assessment include comparing the fidelity of an explanation to models that are intrinsically interpretable by design: explanations generated by the LIME explainer were compared to explanations from sparse LR and DT models [29]. Another approach for evaluating automated explanations relies on user-simulated assessments: by defining untrustworthy explanations, the user's trust and models are simulated for LIME [29] and Anchors [231] explainers.

In addition, Ross et al. [196] conducted empirical assessments and used the LIME approach as baseline to evaluate the consistency and computing cost of the proposed explanation. Schmidt and Biessmann [381] took an alternative route by evaluating the quality of explanations using human intuition: they proposed an explanation quality score.

Finally, through the generated explanations, explainability approaches can also offer quantitative measurements of model trustworthiness, in terms of domain-specific objectives: fair features (fairness), robust features (safety), and reliability. For instance, Zhang et al. [56] showed how explanations may be utilized to detect representation learning problems due to the biases induced in the training data.

10. Methodological recommendations and software tools for XAI research

This section covers XAI tools for model creation and exploration. The intended roadmap for how to determine model and explainability criteria is illustrated in Figure 28. The model structure is at the core of the taxonomy that is presented. In this paper, we consider three methods to providing XAI: (i) Interpretable-by-design methods, (ii) model-specific post-hoc methods, and (iii) model-agnostic post-hoc methods. Interpretable-by-design methods include approaches such as LR, DT, decision rules, or kNN models, among others. Careful model design facilitates explicitly explaining the behavior of a particular component in a given model. Unfortunately, in some cases, the model structure is so complicated that it can not be explained only in terms of individual model parameters and hyperparameters. Then, it is time to resort to alternative methods which are able to extracting information that is tailored to certain models. These are called **model-specific methods** and they assume complete access to the model structure in order to approximate the more complicated processes involved in reaching a decision. In contrast, **model-agnostic methods** are the most generic methods that allow us to analyze a model without having to know anything *a prior* about its underlying structure. Typically, this kind of analysis is based on the mixture of a series of model assessments using appropriately prepared perturbed input data. Open-source toolkits that aim to answer questions such as "what are the overall requirements to use each of the methods?" or "how to choose an explanation ?", will be explored below.

The number of tools available for analyzing predictive models is fast increasing, yet there is no agreed-upon definition of what constitutes an XAI tool. As a result, identifying and presenting all available XAI packages becomes a hard exercise. Table 16 provides a thorough comparison of the various pieces of available software. The comparison is based on the techniques that each package supports, the input data that each package accepts, and the type of explanation that each method provides including local, global, glass-box, or black-box approaches. In addition, the type of explainability provided by a package, based on the suggested taxonomy described in this article, along with the evaluation metrics that are utilized to assess the goodness of the automated explanations are also taken into account.

Arya et al. [394] evaluated new XAI tools in comparison to the most well-known packages available today. One of the most complete libraries in terms of number of methods implemented is OmniXAI, including feature analysis, feature selection methods, feature maps, prediction and bias metrics. This open source XAI library provides from 2 to 10 different methods for each input data type (tabular, image, text and time series) [407], while Shapash [410] facilitates interactive apps from SHAP and LIME in online interactive dashboards. The list of tools has been extended to include other related tools and commonly used R packages that also support XAI techniques. The Institute for Ethical AI and ML, for example, has provided an Ethical ML tool [411] based on the eight principles of Responsible ML. This tool covers three steps: 1) data analysis, 2) production monitoring, and 3) model assessment. Similarly, Wexler et al. [412] presented the What-If Tool, an interactive model-agnostic visualization tool to aid AI model comprehension. This tool was developed with the intent of identifying a wide range of user needs. The tool comes with the following features: (i) it can elucidate potential performance improvements for multiple models with minimal code, (ii) use visual representations to aid model comprehension, (iii) test hypotheses without knowing the internal workings of a model, and (iv) perform exploratory analysis of a model's performance. Another example of an XAI package is XPLIQUE [413], a TensorFlow-based tool for explaining NNs. The package contains attribution techniques, feature visualization methods, and concept-based approaches, among others.

Each of these packages provides comparable methods that can be utilized in similar ways when it comes to their core functionalities. DALEX offers a common wrapper for AI models that may be used with other XAI packages afterward. DALEX is built on the assumption that each explanation should be offered from the viewpoint of *Rashomon*. This implies that a single graph may include any number of explainers. AIX360 and *modelStudio* provide a wide

Table 16
Comprehensive overview of XAI software packages and their evaluation metrics

Packages	Supported Methods	Data Type			Explainability			Explanation		Model Type		Evaluation Metrics
		Tabular	Text	Image	Data	Model	Post-hoc	Global	Local	Glass box	black-box	
InterpretML [392]	Explainable Boosting	■				■	■	■	■		■	—
	Decision Tree	■				■	■	■	■		■	
	Decision Rule List	■				■	■	■	■		■	
	Linear/Logistic Regression	■				■	■	■	■		■	
	SHAP Kernel Explainer	■				■	■	■	■	■		
	LIME	■				■	■	■	■	■		
	Morris Sensitivity Analysis	■				■	■	■	■	■		
	Partial Dependence Plot (PDP)	■				■	■	■	■	■		
Alibi [393]	Accumulated Local Effects (ALE)	■					■	■				Trust Score Linearity Measure
	Anchors	■	■	■			■		■		■	
	Counterfactual Instances	■		■			■		■		■	
	Contrastive Explanation Method	■		■			■		■		■	
	Counterfactuals Guided by Prototypes	■		■			■		■		■	
	Integrated Gradients	■	■	■			■		■		■	
	Kernel SHAP	■					■	■	■		■	
	Tree SHAP	■					■	■	■	■		
AIX360 [394]	Boolean Decision Rules via Column Generation	■			■	■	■	■		■		Faithfulness Monotonicity
	Generalized Linear Rule Models	■			■	■	■	■		■		
	ProtoDash	■	■	■	■	■	■		■		■	
	ProfWeight	■	■	■	■	■	■	■			■	
	Teaching Explanation for Decisions	■	■	■	■	■	■		■		■	
	Contrastive Explanations Method	■		■	■	■	■		■		■	
	CEM with Monotonic Attribute Functions	■		■	■	■	■		■		■	
	Disentangled Inferred Prior Variational Autoencoder			■	■	■	■				■	
Skater [395]	Partial Dependence Plots (PDP)	■				■	■	■			■	Interpretability Transparency
	LIME					■	■		■		■	
	Feature Importance					■	■	■			■	
	Epsilon-LRP			■		■	■		■		■	
	Integrated Gradient		■	■		■	■		■		■	
	Scalable Bayesian Rule Lists					■	■	■	■		■	
	Tree Surrogates					■	■	■	■		■	
tf-explain [396]	Saliency Maps			■			■	■	■		■	—
	Activations Visualization			■			■	■	■		■	
	Vanilla Gradients			■			■	■	■		■	
	Gradients*Inputs			■			■	■	■		■	
	Occlusion Sensitivity			■			■	■	■		■	
	Grad CAM			■			■	■	■		■	
	SmoothGrad			■			■	■	■		■	
	Integrated Gradients			■			■	■	■		■	
Interpretable ML (IML) [397]	Partial Dependence Plots (PDP)	■		■		■	■	■	■		■	—
	Individual Conditional Expectation (ICE)	■		■		■	■				■	
	Feature Importance	■	■	■		■	■	■			■	
	Global Surrogate Tree			■		■	■	■			■	
	Local Surrogate Models	■		■		■	■		■		■	
	Shapley Value	■	■	■		■	■	■	■		■	
	Interaction Effects	■		■		■	■		■		■	
DALEX [398]	Partial Dependence Plots (PDP)	■	■				■	■			■	—
	Accumulated Local Effects Plot	■	■				■	■			■	
	Merging Path Plot	■	■				■	■	■		■	
	Shapley Values	■	■				■	■	■		■	
	LIME	■	■				■	■	■		■	
H2O [399]	Shapley Feature Importance	■				■	■	■	■		■	—
	Feature Importance	■				■	■	■	■		■	
	Partial Dependency Plots (PDP)	■				■	■		■		■	
	Individual Conditional Expectation (ICE)	■				■	■		■		■	
	Decision Tree	■				■	■	■	■		■	
	Local Linear Explanations	■				■	■		■		■	
	Global Interpretable Model	■				■	■	■			■	
ELI5 [400]	LIME	■		■		■	■		■		■	—
	Permutation Importance	■		■		■	■		■		■	
	Grad-CAM			■		■	■	■	■		■	
	TextExplainer		■			■	■	■	■		■	
iNNvestigate [401]	Gradient × Input			■			■		■		■	Perturbation Analysis (PixelFlipping) [317]
	SmoothGrad			■			■		■		■	
	Integrated Gradients			■			■		■		■	
	DeconvNet			■			■		■		■	
	Guided BackProp			■			■		■		■	
	PatternNet [402]			■			■		■		■	
	LRP			■			■		■		■	
	Shapley Value Sampling			■			■		■		■	
modelStudio [403]	Break Down Plot	■			■	■	■		■	■	■	—
	SHAP Values	■			■	■	■		■	■	■	
	Ceteris Paribus [404]	■			■	■	■		■	■	■	
	Feature Importance Plot	■			■	■	■	■		■	■	
	Partial Dependency Plot (PDP)	■			■	■	■	■		■	■	
	Accumulated Dependency Plot	■			■	■	■	■		■	■	
Captum [405]	Grad-CAM		■	■			■		■	■	■	Scalability Infidelity [406] Sensitivity [406]
	GuidedBackProp		■	■			■		■	■	■	
	Integrated Gradient		■	■			■		■	■	■	
	DeconvNet		■	■			■		■	■	■	
	DeepLift		■	■			■		■	■	■	
	SHAP		■	■			■		■	■	■	
	Occlusion		■	■			■		■	■	■	

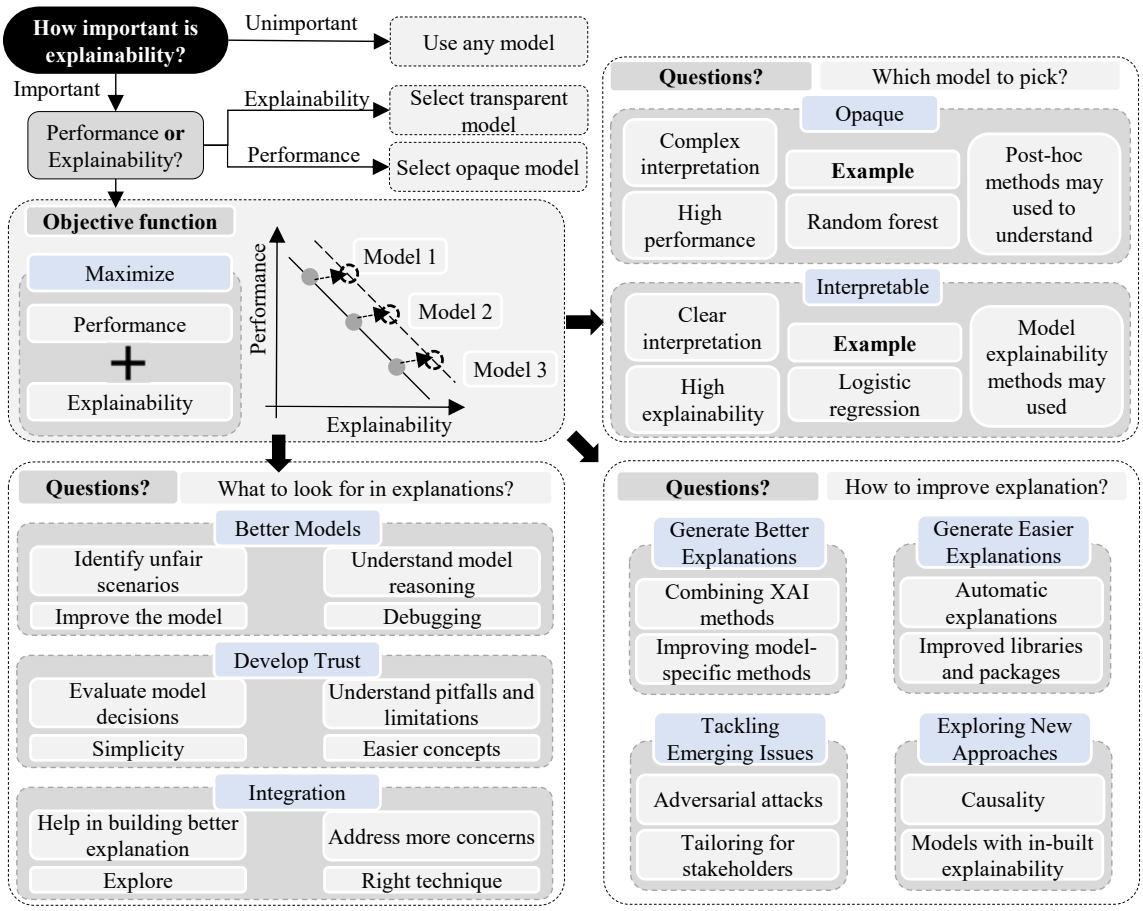


Figure 28: Step-by-step approach to the application of XAI using preferred selection criteria. It is recommended that an AI model is selected based on its performance and/or explainability. After a model is selected, it is advantageous to seek specific types of explanation and to use XAI to enhance the outcomes that can be achieved.

Packages	Supported Methods	Data Type			Explainability			Explanation		Model Type		Evaluation Metrics
		Tabular	Text	Image	Data	Model	Post-hoc	Global	Local	Glass box	black-box	
OmniXAI [407]	Grad-CAM, Grad-CAM++			■		■	■		■	■	■	—
	Score-CAM			■		■	■		■	■	■	
	LayerCAM [408]			■		■	■		■	■	■	
	Partial Dependency Plot (PDP)	■			■	■	■	■		■	■	
	GuidedBackProp			■			■		■	■	■	
	Integrated Gradient	■	■	■			■		■	■	■	
	Accumulated Local Effects (ALE)	■					■	■			■	
	Sensitivity Analysis	■					■	■		■	■	
	Counterfactual Explanations	■	■	■		■	■		■		■	
	Contrastive Explanations			■		■	■		■		■	
	SHAP	■	■	■			■	■	■	■	■	
	LIME	■	■	■		■	■		■	■	■	
	SmoothGrad			■		■	■		■	■	■	
	Layer-CAM			■		■	■		■	■	■	
	Learning to explain (L2E) [409]	■	■	■		■	■		■	■	■	
Shapash [410]	SHAP	■	■	■		■	■	■	■	■	■	Stability, Consistency Compacity
	LIME	■	■	■		■	■		■	■	■	

range of non-standard applications. Furthermore, *modelStudio* facilitates the generation of an interactive Javascript-based model exploration tool with a single command. Arenar [414], fairmodels [415], triplot [416], xai2shiny [417], auditor [418], and flashlight [419] are some other popular R packages. The interested readers are kindly referred to [420] for further details on other related R packages.

It is worth noting that all packages presented so far were designed to serve as XAI support for model developers and end-users. However, different tools are more or less suitable to use in the various stages of a model's development process. For example, the main target group of *flashlight* users is different than the target of *modelStudio* or *modelDown* tools. *flashlight* is a tool that serves mostly to developers, allowing them to fit models based on their experience, what is an important part of the model life cycle. In contrast, *modelStudio* is dedicated to end-users who usually get a model and want to explore its behavior rather than fitting a

new model for the same task. In the case of *modelDown*, it represents a new trend in the XAI toolkit research. This tool acts as a gateway to model exploration for those who lack expert knowledge about AI modeling but want to become familiar with the behavior of the model they use.

Additionally, other tools are available. For instance, *Quantus* [421] offers a list of more than 25 reference metrics to focus on the evaluation of explanations. Several businesses, developers, and researchers joined forces to develop more transparent and sociable AI systems. The What-If tool [412] is an endeavor to build a standard foundation for explainability of algorithms. An approach based on this tool for Fairness, Accountability, and Transparency (FAT-Forensics [422]) proposes inspecting all facets of the ML process. In addition, TensorFlow Extended [423] - a separate tool developed by the TensorFlow community, makes it easier to analyze the TensorFlow models.

Furthermore, PyCEbox [424] deals with explainability of algorithms. Another example of tool for algorithmic transparency is Yellowbrick [425]. Some tools with the focus on fairness analysis are BlackBoxAuditing [426], fairness-comparison [427], FairTest [428], FairML [429], and Fairlearn [430].

Finally, tools for assisting in the analysis of accountability (regarding also security and privacy) are the most difficult to find. Some examples are TensorFlow Privacy [431], DeepGame [432] (a deep neural network verification tool) or PyGrid [433]. In addition, there are some tools that pay special attention to the robustness of the model against adversarial attacks. For example, IBM's adversarial robustness tool [434], FoolBox [435] or CleverHans [436].

11. Current Research Directions

There is a scarcity of reliable and comprehensive systematic comparisons across available XAI methodologies [437]. Accordingly, concepts that reflect the range of opportunities, scope, and resources must be carefully organized to bridge the gap between the study and practice stages. Moreover, the development and regulation of trustworthy AI systems are ongoing work.

As the desire for XAI and the demand for trustworthy AI are so tightly linked, the importance of explainability in developing trustworthy AI is thoroughly examined in this work. Other related surveys are supplemented by this study, which provides a methodology with explicit suggestions for choosing XAI technologies. In addition, by reviewing metrics for quantitative assessment of XAI and proposing useful definitions, this paper provides additional contributions to the existing literature. This section discusses how XAI can pave the way towards building trustworthy AI. We also point out and discuss some open challenges and future directions.

Researchers in the XAI domain are currently developing tools for the exploration, debugging, and validation of AI models. These tools enable users to test models with a wide range of structures, allowing users to pick the best model for their task based on specific metrics. Namely, the requirements that XAI tools must fulfill to provide in-depth model analysis arise from a variety of scenarios:

- A model may make mistakes when dealing with some instances. In order to enhance the model, it is crucial to figure out what is causing such bad judgments. In certain situations, XAI tools may aid in the debugging of an AI model by identifying the causes of its inefficiency.
- Inquisitive individuals do not like to rely on model predictions without knowing extra justifications or the logic behind certain predictions that will gain the user's trust and confidence.
- It is conceivable that some hidden correlations in the data may be retrieved and understood by examining the AI model with XAI tools, what may help users to learn more about the problem under study.
- Increasingly not only decisions, but also arguments, explanations, and reasons for decisions, are expected to be produced automatically.
- If developers want to propose the adoption of a certain model for a given task, experts must first be able to grasp how it works. As a result, black-box models cannot be relied upon for important decisions requiring accountability, i.e., a more in-depth grasp of the decision-making model is demanded.

In addition, the following scenarios support and promote the intended design and assessment framework at multiple levels.

Evaluation Metrics and XAI System Design. When measuring the performance of XAI systems, it is critical to apply the right metrics. The use of the same evaluation metrics for diverse design objectives is a typical problem when selecting measurement methods for XAI systems. A basic solution to this problem is to use numerous scales to record distinct features in each assessment to discriminate between measurements. The idea of user trust, for example, is made up of numerous variables that may be examined using distinct scales in surveys and interviews [364]. To target certain explanation qualities, user satisfaction assessments might be established for variables such as explainability, usefulness, and sufficiency of information [438]. In iterative design processes, balancing diverse design approaches and assessment types is an effective strategy to connect design objectives with suitable evaluation metrics.

Overlap in Explanation Design Objectives. Four primary dimensions along which to place XAI systems are provided by our XAI classification axes: 1) data explainability, 2) model explainability, 3) post-hoc explainability, and 4) assessment of explanations. Across certain disciplines, there are overlaps in the axes. While the fundamental aims are similar (to produce better explanations), different explanation objectives should be explored in consideration of the various users, what results in a diverse collection of design parameters and implementation approaches. Designing

XAI systems for AI novices, for example, necessitates the development of human-centered XAI interfaces to convey the model explanations, but developing new interpretability approaches for AI specialists implies other requirements. Accordingly, XAI user groups may be considered as an additional dimension along which to arrange XAI goals in cross-disciplinary problems while emphasizing the integration of a variety of research aims in order to address the overlap between XAI goals across different research disciplines [439].

User Interactions in XAI. Another factor to consider when developing XAI systems is how to handle human interactions. Interactive visual tools enable AI and data specialists to enhance the performance of models. Moreover, interactive systems might also be beneficial to novices. A few papers have concentrated on the interactive design of AI systems [440, 441, 442, 443]. These studies demonstrated how interactive methods enable users to assess the effect of their actions and alter their subsequent queries for enhancing results. Expert users may utilize visual tools to comprehend the models they are using by interacting with the algorithms used. Allowing data scientists and model specialists to examine model representations interactively [444], assess model training processes [391], and discover learning biases [445], are just a few examples of the advantages of these XAI systems.

System and Ground Truth Evaluation. Taking user learning into consideration is a key part of assessing XAI systems. When conducting cognitive psychological experiments for assessing user understanding, satisfaction, and trust of XAI systems, their learnability becomes even more important [446]. With regular usage of the system, a user learns and becomes more comfortable with it. In terms of XAI assessments, this emphasizes the value of recurrent temporal data collection [447]. Moreover, the choice of the ground truth is a crucial aspect in interpreting XAI assessment outcomes but also in comparing results across many investigations. Controlled studies are often used to investigate the impact of model explanations on a control group compared to a baseline (no explanation required) group in human-subject investigations [361, 442].

Expansion and Generalization. Amershi et al. [448] provided 18 human-AI interaction design criteria. They systematically evaluated the recommendations from 20 AI-infused products via numerous rounds of assessments with 49 design practitioners. Their design guidelines give additional information inside the user interface design layer to assist in the creation of suitable end-user interactions with XAI systems. The framework suggested in this paper is expandable and consistent with current AI-infused interface design guidelines. Adaptive explainable models that provide context-aware explanations are also available [449].

Explainability in Dynamic Learning Scenarios. Large sample sizes can improve model generalization by preventing overfitting in individual cases, but they also increase the cost of model training. Moreover, when we add more data, the model often has to be fine-tuned or trained from the start using the extended dataset. Otherwise, learning numerous tasks in a row or from dynamic data might result in catastrophic forgetting [450]. One way to tame these problems is using continual learning [451] strategies, suitable in sequential data streams scenarios, or state representation learning [452, 453], as a way to intermediately learn the states of the problem space as an intermediate task to solve control problems involving deep learning. Explainable AI can contribute to a better refinement of the knowledge captured by a model from evolving data that is retained over time. For instance, a relevance-based neural freezing method was developed in [454] to reduce catastrophic forgetting. Unfortunately, it has been shown that the explanations produced by cutting-edge methodologies are inconsistent, unstable, and offer very little information regarding their accuracy and dependability [455]. These approaches also demand much hyperparameter adjustment and are computationally inefficient. Consequently, efficiently producing explanations suited to deal with the varying nature of data and/or learning tasks and improving the adaptation of the model to eventual changes by exploiting such explanations fall within a research niche that will surely attract the interest of the community in the future. Above all, standard protocols for continuous pipeline adaptation that produce explanations on-the-fly, cope with errors and correct them in continual learning settings are much in need nowadays.

Other usages of explanations. Explanations can supply the extra information required to boost a model's performance, convergence, robustness, efficiency, reasoning, and equality [456]. For instance, understanding relevant and irrelevant feature representations can cut training time and improve accuracy. Similarly, determining the most crucial neurons and filters in a neural architecture is essential to increase model effectiveness. The more stable, conservative learning process of augmented models allows for improved generalization [457]. Furthermore, in an active learning environment, explanatory interactive learning enables human users to correct a model's decision-making [458, 459].Deterioration in the model performance can be measured by the amount of Out of distribution (OoD) samples, since it is a signal that can be used to explain model failures. For instance, a clustering based on archetypical explanation saliency maps can detect OoD samples in settings with low intra-class variability [460]. It is also possible to use XAI for improving object counting and instance classification models, based on landmarks that assist heatmaps' sensitivity and uncertainty analyses, for more accurate and certain predictions [461].

Another approach that goes beyond exposing what models really learn consists of masking artifacts that may confuse models and their explainability (known as Clever Hans effects). This is a way towards the necessary but immature research line of model certification [462]. Issues beyond those carried out by explaining a model include the challenges involved in XAI-based model improvement, which can accumulate sequential errors in the explanation producing pipeline [456].

XAI can also be used to drive network improvement and compression, making DNNs low-bit and sparse. For instance, LRP XAI method can be used to preserve the highest information weights based on entropy, make most weights zero and in this way compress networks beyond 100 times their size, which can be useful for learning on the edge [463]. Within the same spirit of efficiency, explainability can also be used for pruning neural network layers as a criterion [464], or concept unlearning [465, 466].

Another usage of XAI, when using causal explanations, is facilitating accountability, providing algorithmic recourse to make explanation more actionable [467], or facilitating the explainability of the model. In [468], Bargal et al. demonstrate that increasing the explainability of a deep classifier can be used to improve its generalization, both to unseen domain samples and out of domain ones, fine grained predictions, and to be more efficient when using the network capacity, as well as robust to network compression. These and other desirable - but hard to quantify properties - of ML models that can be improved with XAI techniques are further unified in a theoretical framework in [456].

Finally, while XAI can be a tool to help conveying explanations in natural language, there is often a lack on datasets and benchmarks that include explanations' ground truth to fully validate models. One example of such datasets is CLEVR-X Visual Reasoning Dataset [469] for evaluating natural language explanations.

Explanation and model robustness. Given a domain, rbustness is understood as the ability of a system to maintain its performance quality under varying conditions (ISO/IEC 24029-2:20XX), and it needs to be monitored in all life cycle phases of the AI system (ISO/IEC DIS 22989:2021). There are several ways to evaluate the robustness of models and their respective explanations. Generic approaches to guarantee model robustness borrow inspiration from well developed disciplines within the field of Software Engineering, including Verification, and Validation (V&V) of the model. While the verification process confirms through the provision of objective evidence that the specified requirements (ISO/IEC 25000:2014 4.43, ISO/IEC 25030:2019(en), 3.22) have been met, the validation process confirms via objective evidence and testing that the verified software on real data does what it should and works as expected. In this area a broad family of formal methods are being extended to neural network models to prove whether they satisfy robustness properties (ISO/IEC 25000:2014, 4.41, ISO/IEC 25030:2019(en), 3.21).

Programmatic XAI metrics and methods that account for the variability of data/task are in much need. Some ways to perform verification are via program synthesis [470] (to learn policies in RL), or running *reality checks* [471] or checking for XAI technique pitfalls [472].

One approach that, apart from robustness, serves as well as solution to the above-mentioned uncertainty issues is to model uncertainty in black-box explanations. This requires ways to keep track of how certain the model is for different samples, or during its life cycle as it keeps learning in time, post-deployment, with new data. One example in this direction [473] consists of estimating epistemic and aleatoric uncertainty maps associated to segmentation maps produced by a Bayesian MultiResUNet.

Despite robustness being a desired and required property (Art. 15 of EU AI Act [474], Precision and Robustness and Cybersecurity of high-risk AI systems (HRAIs) and upcoming AI Act Sandbox), there is a lack of procedural ways to approach and certify model robustness. In addition, one of the most common limitations of formal methods for V&V approaches is the lack of scalability to highly dimensional inputs typical of deep learning. While local robustness properties can be documented defining a valid range for the input features, this may not be meaningful when individual features such as pixels have no semantic meaning. However, a list of recommendations towards institutional, legal and technical bodies towards auditing and verification of AI systems are being developed with paramount relevance to implement trustworthy AI [475].

XAI and late breaking models. Explaining generative models is an unexplored arena in XAI. Modern generative models (GPT-3 for text generation from a prompt, ChatGPT [476], DALL-e 2 [477] or stable diffusion methods for image generation from text) may, in occasions, require explaining their generated samples. This requirement depends stringently on the purpose for which such samples are created (e.g., as specified in the AI system requirements, or as anticipated by the compulsory requirements of e.g., HRAIs in the EU AI Act [474]).

The debate around the explainability of generative models departs from two main questions: 1) Do generative models require explainability? and 2) Is it even possible to get a satisfactory explanation of an output from the most modern large generative models? The first question can be answered around two cases:

- *No specific need for explanations*: in generative art, unless copyright creation out of the generated samples could be legally claimed, no issue requiring explanations may arise. However, who is to be compensated in posterior usages when, for instance, plagiarism is detected among generative art samples? Who is to blame when a generated resource exposes private content worth censoring, or belonging to some private data that can result into trouble if attempting against people's privacy, dignity or intimacy?

- *Explanations required*, as it occurs in HRAIS such as in data augmentation models for medical diagnosis where patients lives are at stake. If the model fails in its prediction for a given query, and post-hoc explanations reveal that the cause for the miss-prediction was indeed an augmented sample generated by these models, we need to have explanations indicating why the model generated it, unlearn [466, 465] this augmented sample from the trained model, and avoid that the generative model produces it again. The challenge resides in the creation of explanations from probability distribution learning models such as stable diffusion [478].

Another issue with modern generative models is that fact that language models can leak private information []. Thus, a remaining challenge is: how can XAI deal with explaining privacy enhancing technologies (PETs) or models having such complex and abstract blending capabilities (such as those exhibiting properties of style transfer [479] or image translation [480] models) that a human can hardly explain? A first challenge would be defining *what* constitutes an explanation in such models, so that XAI techniques for devising provenance and traceability of samples in generative modeling can be devised to gain trust in large generative models.

Towards an Ethical Code. The study of ethics in practical applications of AI is a complex and multi-faceted issue that requires interdisciplinary collaboration between experts in AI, ethics, law, and other related fields. One of the main challenges is the diversity of ethical issues that arise in the context of AI. These issues range from bias and fairness in decision-making to privacy and security concerns, and they can be approached technically in a variety of ways. For example, addressing bias in AI models may require data pre-processing techniques, algorithmic modifications, or human oversight [24]. Similarly, ensuring the robustness and reliability of AI systems may involve techniques such as adversarial training, uncertainty quantification, and fault-tolerant design [481]. In addition, ethical considerations may vary depending on the application domain and the stakeholders involved. For instance, medical AI systems raise unique ethical issues related to patients' safety, informed consent, and privacy, which may require different technical and legal frameworks compared to other domains such as finance or transportation [482]. Protocolizing the study of ethics in practical applications of AI requires a nuanced and context-specific approach that takes into account the complexity and diversity of ethical issues in different domains and applications. Recent works have proposed frameworks and guidelines for ethical AI design and deployment, such as the IEEE Global Initiative for Ethical Considerations in AI and Autonomous Systems [483] and the EU Ethics Guidelines for Trustworthy AI [484].

Advanced Tools in XAI. The state-of-the-art tools for explainability classification, qualification, and evaluation can be advanced by implementing new methods. For example, argumentation and XAI are intertwined, since argumentation has been utilized to provide explainability to AI in recent years. Vassiliades et al. [53] have demonstrated that argumentation can be used to explain how an AI system comes to a decision, how it achieves that decision in the face of ambiguity, and how it can solve problems when presented with contradicting data. The more sub-symbolic technique-based intelligent systems have saturated our daily lives, however, these systems are not comprehended well. As a result, symbolic techniques are gaining traction in a broader endeavor to make AI more understandable, explainable, and trustworthy. Calegari et al. [485] presented an overview of the most common symbolic/sub-symbolic integration approaches, with a special emphasis on those aimed toward XAI systems. The recent advancements in technology for the Internet of Things may aid in the transmission of explanations from Machine to Machine.

In addition, planning is a key aspect of AI that is employed in situations where learning is not possible. Incorporating explainability into the planning process entails converting the generated plan stages into a human-readable format. Furthermore, this process encourages economic interpretations that can handle concerns including cost estimates and variance, algorithm propriety, trade secret disclosure, as well as anticipating XAI market development. One potential approach to XAI in planning that can bring a fresh breeze to the current spectrum of XAI methods is the use of neural-symbolic learning for sequential decision making [486]. This can be done in two sequential steps, from symbolic to neural representations, or viceversa, and an interface in between, or b) end-to-end, being able to handle both symbolic and subsymolic formalisms at the same time. An example of model (of type a) with an interface in between modules) is in [487]. It combines the deep neural nets with symbolic components of planning and a symbolic descriptions. Common symbolic approaches for classical planning include the use of first order logic -FOL- or Planning Domain Definition Language -PDDL. And an example of b) that jointly processes symbolic, neural representation and inference, is DeepProbLog, based on neural probabilistic and deep learning [488].

It is worth noting that even if there are many XAI strategies, metrics and tools, as we will see in the next section, several questions still remain without an answer: which methodology delivers the best explanations? how should the quality of explanations be assessed?

12. Concerns and Open Issues about XAI

The more pervasive AI is in our daily life, the more concerns turn up. For example: (i) due to the size of AI systems' input and state spaces, exhaustive testing is impractical, (ii) most AI systems currently in use have complex internal structures that are difficult for humans to interpret, and (iii) most AI systems are highly dependent on the training data. We have identified three main categories of concerns (to be discussed in the rest of this section): user concerns, application concerns, and government concerns.

12.1. Concerns from the User Perspective

In this work, we have distinguished among data explainability, model explainability, and post-hoc explainability, as depicted by the internal border in the Figure 29. Within the outer boundary of the Figure, the various stakeholders and regulatory entities interested in AI system explanations are depicted. The text highlighted at the bottom of each group represents their motivations and desire for a property to be provided with explainability. While all organizations strive to get a correct answer, the degree of detail and intricacy involved in this may differ significantly.

Brandao et al. [489] claimed that most research on how to interpret and explain AI systems is mainly motivated by

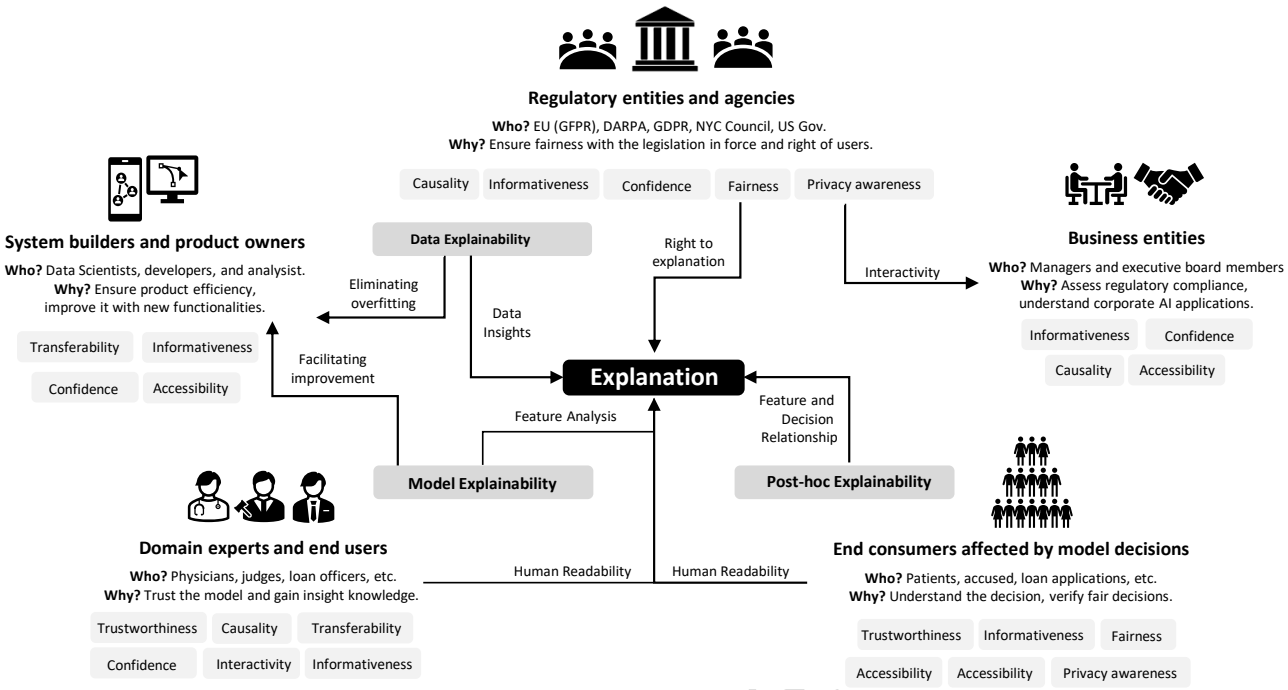


Figure 29: The aim of XAI is expressed to various stakeholders such as end-users, domain experts, developers, and government bodies. We also highlight how the explainability of various axes can benefit multiple stakeholders. Each stakeholder has a community of participants - outlined in **Who**, and objectives - outlined in **Why**. Additionally, each group promotes explainability in terms of a specific characteristic. Each targeted platform's characteristic is asserted in the gray box at the bottom.

the requirements of developers rather than users. Comprehensive research has confirmed and emphasized the need to validate AI systems with actual users to ensure transparency [95], accountability [490], and fairness [24]. Moreover, AI is becoming a keystone in the development of automated decision-making systems based on personal information, which may have substantial implications for people's basic rights. Notice that, people have the right to request an explanation and get a guarantee that AI systems will not negatively affect their lives under the GDPR policy [491].

In addition, the level of explanation given to specialists and ordinary users does not have to be the same. Accordingly, sociological studies have focused on how individuals react to explanations. For example, Miller [52] reviewed psychological research to determine what people perceive as a good explanation. He concluded that an explanation must be concise, socially acceptable, and contrastive.

Finally, Doshi-Velez and Kim [15] suggested three layers of testing to evaluate explainability: first, experiments with people carrying out real-world activities; second, basic experimental tasks with people; and lastly, proxy tasks and metrics that have been validated by previous research using the evaluation techniques mentioned above.

12.2. Concerns from the Application Perspective

Whatever application where an AI system is in charge of drawing autonomous decisions which may endanger human life, then trust emerges as the most important quality that the related intelligent system must possess. For instance, at the annual Neural Information Processing Systems conference in December 2018, two pictures were shown on a screen [492]: (i) a patient with a human surgeon and a caption showing the 15% risk of the patient dying during surgery; and (ii) a robotic arm with a 2% failure rate. Then, the audience was given the option to vote which surgeon was preferred. Everyone voted for the human surgeon except one. Thus, a 2% risk of dying should be preferable to a 15% chance of dying, but why did the audience not choose the more accurate model? Apart from accuracy issues, trust in AI systems is required in this kind of situations where human lives are involved. However, this result may change if AI systems were able to provide good explanations, which would increase trust by allowing individuals to understand how and why the system makes specific decisions. Medical domain experts can find a comprehensive overview of the state of XAI in healthcare, including applications, challenges, and future directions in the recent surveys [493, 494].

In detection and classification applications, a DL model can automatically explore, learn, and extract data representations. The capacity of models to describe their inner workings and decision-making processes is inevitably limited when trying to maximize data utilization and increase prediction accuracy. However, it is difficult to trust systems whose decisions are not easy to comprehend, such as those from CNNs and ensemble models. This is especially true in

applications like healthcare or autonomous vehicles, where fairness and moral issues naturally arise.

Finally, the need for reliable, fair, resilient, and high-performing models for real-world applications has been one of the triggers in the XAI field. The general trend depicted in Figure 30 (see the Appendix A) indicates that research publications on XAI have grown greatly during the previous decade. The large increases seen in recent years have been mirrored by an increase in studies on ethical issues within the same time span. Consequently, it is apparent that users require ethical concerns in addition to an explanation of decisions, i.e., XAI highlights the importance of safety, causation, security, transferability, privacy, informativeness, fairness, and ethical decision-making when it comes to AI systems making important decisions [495].

12.3. Concerns from the Government Perspective

In the USA, the Defense Advanced Research Projects Agency (DARPA), began its XAI initiative in 2017 with the goal of creating new methods for explaining intelligent systems [496]. The program comprised 19 projects and ran until 2021 [497]. DARPA's XAI initiative emphasizes the need for explainability in order to better understand, trust, and control the next generation of AI systems. This has an effect on accountability [498], safety [499], and industrial responsibility [500]. This is essential in high-risk applications like self-driving vehicles and medicine, where a single incorrect outcome may result in a person's death. As a result, across various business sectors and scientific fields, good explanations are at the core of responsible and open AI research. This necessitates an increase in investment by practitioners and industries to ensure the decision of AI systems is properly explained [32, 501].

Indeed, although AI systems are usually supervised by humans in reasonably controlled settings, AI is anticipated to be implemented on a much wider scale in the coming years, necessitating a response from regulatory authorities. To achieve this aim, the European Commission has committed to establishing guidelines for the trustworthy and safe use of AI in our digital society [502]. The provisions of the Cybersecurity Act aim to encourage an ecosystem surrounding AI technology to quickly develop and favor innovation while protecting basic rights [503]. This act was presented in 2017 and established an EU-wide certification structure for digital products, services, and processes.

Furthermore, one of the numerous responses to the new rules and the GDPR legislation has been a demand for XAI to give explanations not just to users, but to society at large [63]. Particularly, knowing the risks and responsibilities associated with AI systems is essential in healthcare, clinical, and judicial professions since human lives are involved. When responsibility is delegated to a single expert, risk avoidance occurs. Moreover, instances of minorities in employment procedures, recidivism in the COMPAS system, and overall fairness have all contributed to XAI literature's growth [63, 501, 504, 505]. Another element driving the need for XAI, according to Adadi and Berrada [21], is the development of algorithms that are not only fair and socially responsible, but also accountable and able to explain their output. The GDPR refers to the "Right to an Explanation", which has sparked a lot of interest in both academia and industry, pointing the people concerned to XAI as a potential compliance option [437, 506, 507]. There is widespread agreement that putting such a concept into practice is essential and that it is now a major open scientific issue. As a result, when it comes to understanding AI a primary emphasis is put on the audience for whom explainability is desired. In general, scholars concur on the need to develop user-friendly explanations.

Are the XAI techniques available today sufficient to resolve all the explainability concerns, even if several tools are used? The answer is **NO**. As described in the European AI Act (which is currently pending of final publication at the time of writing), AI systems for tackling real-world problems should be auditable and subject to regulatory pressure in terms of the criticality of the situation (such as safety) [508]. Furthermore, explainability alone does not suffice for realizing trustworthiness in AI-based systems: explanations need to be accompanied by robustness guarantees, causality studies, data governance, security, accountability or human in- or on-the-loop interfaces (depending on the level of risk of the AI system as per the AI Act), among other factors of relevance [509, 439].

Furthermore, there is a controversy when it comes to choosing between the best-performing model and the best XAI method. In reality, we are far from regulated top-performance models. Transparent models cannot handle sophisticated real-world applications. Many applications need the modeling complexity provided by black-box systems, but if authorities and agencies fail to recognize that not everything can be explained by existing technologies, regulation may become a threat to clamp down on unsafe systems. To define the requirements that should be satisfied by XAI tools, the idea of risk/criticality of the application should be identified beforehand and governed accordingly. Although specific audit procedures are already well established, others still call for more analysis and the creation of brand-new audit technologies and tools.

It is worth noting that end-user, application-oriented, and government-led efforts to audit AI systems are essential but insufficient to alleviate all concerns raised by the results and advancements made in explainability. Establishing the methodological criteria for guiding future endeavors to make AI systems acceptable in critical situations necessitates regulatory supervision surrounding crucial scenarios and applications. Additionally, in order to drive future research away from simply producing additional tools but toward more fruitful cases where XAI tools are used successfully, we require methodological measures, and assurances of transparency, among various other aspects, to confirm the credibility of explanations provided by XAI in specific critical situations.

Fortunately, to foster trustworthy AI, several regulatory efforts are in progress. For example, the European AI Act

[474] aims to ensure quality and trust to enhance industrial and research capabilities while upholding the fundamental rights of citizens [508]. This European regulation establishes 85 articles for which technical guides are being developed for HRAIs to ensure compliance with. Subsequently, in 2022, Spain established the first AI supervisory council in Europe. While admitting the threats that AI may pose to individuals' rights and freedoms, the Spanish government is also seeking to encourage and optimize the use of AI technology [510]. In order to monitor and erode any danger connected with AI technology, the government has taken the initial steps to establish a supervisory authority in 2023.

In order to modernize our laws for the 21st-century economy, American and European authorities have paved the way for a new era of technological collaboration. They agreed to work together to develop AI technologies that will strengthen privacy protections, explore cooperation on AI, and conduct an economic study looking at the effects of AI on the future of the workforce. AI systems must be innovative, trustworthy, and respectful of universal human rights while sharing democratic values [511]. Other countries have also shown interest and proposed several rules for AI. For instance, China's cyberspace authority recently announced algorithmic recommendations for internet information services to standardize Internet information services [512]. The Brazilian Congress enacted a law establishing a legal framework for AI in 2021 to lay forth broad guidelines for the advancement of AI and its regulation [513]. Moreover, other 58 countries proposed more than 700 rules in a legal framework for AI. For further details, interested readers should refer to [514]. Besides this, we still require global policies and regulatory frameworks to guarantee that cutting-edge technologies are advantageous to humanity. In order to achieve this, the 193 member nations of UNESCO released a global agreement on the ethics of AI in 2021 [515]. Despite several sub-national, national, regional, and global endeavors, the practical use of AI systems will hamper the lack of trust in terms of the sufficiency of the explanations produced for such systems.

13. Conclusion

The demand for Trustworthy AI will expand as the technology is used more often in practical applications, particularly when making automated decisions that might have adverse effects on human lives. In this work, we have analyzed the current state of XAI literature, standard definitions, XAI methods, and necessary concerns about the objectives of trustworthy AI. The study has looked at the four axes of explainability: data explainability, model explainability, post-hoc explainability, and assessment of explanations. The taxonomy is arranged to provide a high-level discussion on each method with good examples and insight into the related mathematical modeling. The proposed framework for end-to-end XAI system deployment integrates design objectives, including XAI concerns, with assessment methodologies. This approach encourages more conversation regarding the relationship between the design and assessment of XAI systems. Proper assessment metrics and properties for different user groups have been also addressed in the study.

In addition, we have considered the target audience for whom the explanation is required. To comprehend an AI system satisfactorily, each user needs a different level of explanation. The needed attributes for different groups of users have been identified and associated with the proposed explanation axes. According to the classification of explainability, research questions are addressed according to the various aims to achieve related to each axis. The proposed classification demonstrates the importance of multidisciplinary collaboration in designing and evaluating XAI systems. All interface and interaction design areas have been covered. This brings attention to complementary social science resources that might help expanding the scope of social and cognitive components of explanations. Standard terms related to XAI have also been specified to make intelligent systems trustworthy and ethically appropriate.

Moreover, we have presented two main contributions in this comprehensive survey. We have first elicited methodological suggestions using advanced XAI technologies. Second, we highlight the key issues and future directions for XAI research.

As AI technology advances, several technical (but also legal and ethical) issues are explored by academia, while the industry is also establishing new strategies. As a result, the creation of relevant standards, auditing plans, procedures, and tools is progressing quickly. While Trustworthiness is a matter of more aspects beyond explainability, explainability is a required ingredient. However, it alone is not sufficient. Application contexts where AI-based system are used can easily impose other restrictions that can affect the trustworthiness and actionability of AI system outputs, compromising fairness, data governance, privacy, accountability, sustainability, and robustness. This is why only in the confluence and provision of guarantees in these desiderata, we can provide AI pipelines we can trust.

To pose the general conclusion of the paper: we as a community have advanced notably in the explainability of AI models to date. However, we are progressing over only one of the requirements for trustworthiness. More work is done towards showing experiences and use cases accounting for more than the delivery of explanations for the knowledge and decisions elicited by AI models. Below there is a summary of what we have discovered about XAI and what is still required to attain truly trustworthy AI:

- Many methods are already available. Therefore, we do not simply need more methods. Instead, we need to pay more attention to critical situations so that regulatory bodies and supervisory authorities can enforce the necessity for explanations to be provided. Not just the AI model itself needs to be explainable; its decision-making procedures must also be explainable.
- Although we can elicit explanations of various kinds, the quantity of proposals without regard to whether they

satisfy the intended audience's needs has reached a saturation point.

- There have been many promises about creating transparent models, yet practical applications need complex modeling: How much performance can be sacrificed for transparency?
- Policies for AI Governance and supervisory regulations that are currently in design comply with seven identified requirements for developing Trustworthy AI: (i) robustness and safety, (ii) human agency and oversight, (iii) transparency, (iv) privacy and data governance, (v) diversity, non-discrimination, and fairness; (vi) accountability, and (vii) social and environmental well-being. In order to not advance only over only one of the requirements for trustworthiness, we advocate for simultaneously approaching the problem of reaching trustworthy AI from all these perspectives.

Furthermore, the importance of bridging the gap between legal clauses in trustworthy AI regulations and technical advances, tools, and practices in related fields is needed. This connection is crucial for developing risk-aware scenarios and increasing the number of cases in which trustworthiness is required over time. To achieve this goal, it is necessary to continuously learn from the initially approached methods and apply these lessons to future development. As such, there is a need for collaboration between legal and technical experts to establish a comprehensive framework for trustworthy AI. This enables the deployment of AI systems that are not only technically advanced, but also meet legal and ethical requirements. Addressing this gap facilitates the development of trustworthy AI and ensures that it is used for the benefit of the society.

This survey has shown that it is feasible to boost a model's epistemic confidence by taking advantage of the insights offered by several complementary explainable approaches. The goal of XAI is to learn more about AI systems, understand them, and establish trust in them. The XAI field, however, has more promise than merely promoting trustworthiness. Explainability may inspire the development of novel training methods and evaluation metrics that guarantee the trustworthiness and consistency of even the most complicated and abstract models. Due to techniques that primarily concentrate on technological aspects of AI systems, we are still far from having end-to-end XAI systems. The user or developer interactions required for an AI system to be trusted and employed are not taken into account by most XAI approaches. This is why interactive systems that offer explanations and feedback can be crucial for objectively and empirically demonstrating to users and decision-makers that AI systems can be trusted.

Acknowledgments

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT)(No. 2021R1A2C1011198), (Institute for Information & communications Technology Planning & Evaluation) (IITP) grant funded by the Korea government (MSIT) under the ICT Creative Consilience Program (IITP-2021-2020-0-01821), and AI Platform to Fully Adapt and Reflect Privacy-Policy Changes (No. 2022-0-00688).

Declaration of competing interest

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

A. Literature Search Strategy

This section describes the methodology followed to structure and filter the reviewed XAI papers included in the survey. It is hard to search and arrange the XAI literature due to its interdisciplinary nature. Therefore, some restrictions were imposed to exclude certain studies, as follows:

- Any XAI studies in fields other than AI and Computer Science, Mathematics, Philosophy, and Psychology.
- Studies with methods that were created only for the purpose of increasing model transparency but were not explicitly focused on explanation.
- Papers focused on concepts other than the explanation of Supervised ML models.
- Studies not written in the English language.
- Papers that are published before January 2016.

The keywords *explainable artificial intelligence*, *responsible artificial intelligence*, *explainable machine learning*, *trustworthy artificial intelligence*, *ethical artificial intelligence*, *understandable artificial intelligence* and *interpretable machine learning* were used to search for publications that discussed explainability using Google Scholar. The research articles discussing explainability and interpretability were selected from peer-reviewed journals, conferences, and workshops that were published between January 2016 and June 2022, and as shown in Figure 30, the trend in terms of the number of papers published in the field of XAI shows exponential growth over the study period. In addition, PubMed, ScienceDirect, Web of Science, SpringerLink, Nature, Scopus, and IEEE Xplore were used to perform a thorough literature search. The search list also included under-review papers from arXiv. These digital libraries were selected because they provide access to the most significant and current peer-reviewed full-text publications in the area of XAI.

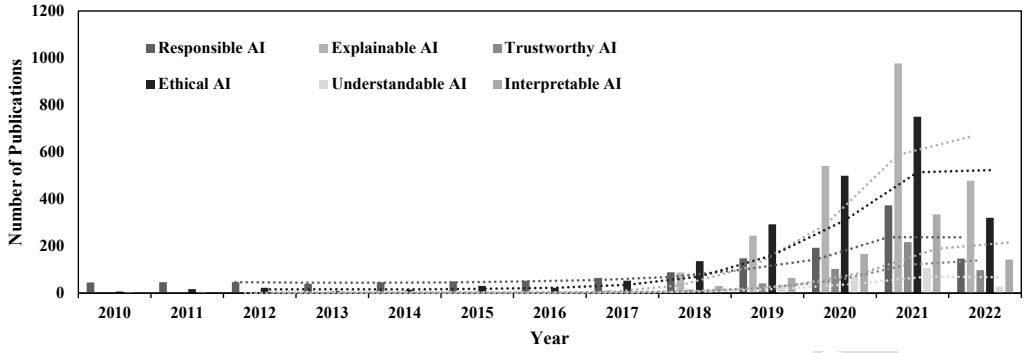


Figure 30: The evolution of the total number of publications on XAI over time. The dotted lines show the trend over the previous three years using a moving average. These statistics were retrieved from the Scopus database in June 2022.

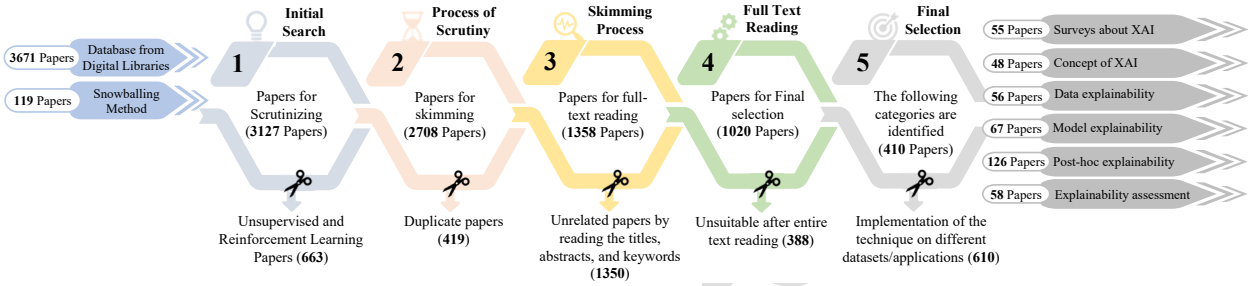


Figure 31: Search strategy for categorizing the selected XAI research papers on supervised learning.

In supervised ML, unsupervised ML, and reinforcement ML models, explainability techniques have been used. Although explainability is used with all three learning methods, the core of XAI research is focused on supervised learning. Therefore, this article discussed explainability in the context of supervised learning. For the purpose of completeness, the snowballing process was applied [516]. The Related Work Section of each article was briefly skimmed, and additional relevant papers were sought out. More papers from other sources, such as the European Conference on Computer Vision (ECCV), ACM Transactions on Intelligent Systems and Technology (ACM-TIST), Computational Visual Media (CVM), the Workshop on Human-In-the-Loop Data Analytics (HILDA), and IEEE Transactions on Big Data, were discovered using this procedure.

About 3790 peer-reviewed articles were identified as a result of this search strategy. The titles and abstracts of all these articles were scrutinized. After that, we excluded any articles that did not satisfy the criteria for inclusion by applying the restriction criteria mentioned above. Finally, a complete text analysis of the remaining articles was conducted in order to identify the most relevant papers. Furthermore, the reference lists of the shortlisted papers were manually searched to find out other relevant publications.

The following six major categories were identified after a comprehensive examination of all of the reviewed papers:

- *Surveys of explainability techniques* - This category contains systematic reviews in the area of XAI from the time period specified. Table 1 provides a comprehensive overview of the reviews that are currently accessible, along with open challenges.
- *Concepts related to explainability* - This category comprises research aimed at defining concepts linked to the idea of explainability, as well as determining the key features and requirements of a successful explanation.
- *Data explainability* - This category contains papers that suggest new and innovative approaches for improving explainability by interpreting training data.
- *Model explainability* - This category contains papers that suggest new and innovative approaches for improving explainability by understanding the inner working of AI models.
- *Post-hoc explainability* - This category contains papers that suggest new and innovative approaches for improving explainability by providing a human-understandable explanation of the AI model's decision.
- *Assessment of explainability* - This category contains articles that describe the findings of scientific research aimed at assessing the effectiveness of various explainability techniques.

We applied the limitation criteria step by step to obtain our set of distinct and specific papers for this study after gathering a list of research articles based on related keywords. Figure 31 depicts the whole procedure, with the number of

Table 17
Table of Abbreviations

ACE	Automatic Concept-based Explanations
AD	Alzheimer's Disease
AI	Artificial Intelligent
ALE	Accumulated Local Effect
ANN(s)	Artificial Neural Network(s)
BN(s)	Bayesian Network(s)
CaCE	Causal Concept Effect
CAM	Class Activation Map
CAV(s)	Concept Activation Vector(s)
CEM	Contrastive Explanations Method
CEN	Contextual Explanation Networks
CL	Convolutional Layers
CluReFI	Cluster Representatives with LIME
CNN(s)	Convolution Neural Network(s)
ConvNet	Convolutional Network

papers in brackets for each stage. It was feasible to create a map of the XAI literature using the proposed categorization method. The following concerns, in particular, are taken into account: (i) outliers - papers that concentrate on the unsupervised and RL models, (ii) typos - duplicate and unrelated papers, (iii) disparities - papers unsuitable for this study, and (iv) inconsistencies between various evaluations by checking and eliminating any obscure or misclassified papers. Note that since a paper may cover several dimensions, it may appear in numerous branches of this categorization.

B. Acronyms

The abbreviations used along the manuscript are summarized in Table 17.

References

[1] P. Georgiev, S. Bhattacharya, N. D. Lane, C. Mascolo, Low-resource multi-task audio sensing for mobile and embedded devices via shared deep neural network representations, Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 1 (3) (2017) 1–19.

[2] A. Jain, H. S. Koppula, B. Raghavan, S. Soh, A. Saxena, Car that knows before you do: Anticipating maneuvers via learning temporal driving models, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 3182–3190.

[3] E. Chong, C. Han, F. C. Park, Deep learning networks for stock market analysis and prediction: Methodology, data representations, and case studies, Expert Systems with Applications 83 (2017) 187–205.

[4] T. T. Pham, Y. Shen, A deep causal inference approach to measuring the effects of forming group loans in online non-profit microfinance platform, arXiv preprint arXiv:1706.02795 (2017).

[5] L. Nie, M. Wang, L. Zhang, S. Yan, B. Zhang, T.-S. Chua, Disease inference from health-related questions via sparse deep learning, IEEE Transactions on knowledge and Data Engineering 27 (8) (2015) 2107–2119.

[6] G. Goswami, R. Bhardwaj, R. Singh, M. Vatsa, Mdlface: Memorability augmented deep learning for video face recognition, in: IEEE International Joint Conference on Biometrics, IEEE, 2014, pp. 1–7.

[7] J. Lundén, V. Koivunen, Deep learning for hrrp-based target recognition in multistatic radar systems, in: 2016 IEEE Radar Conference (RadarConf), IEEE, 2016, pp. 1–6.

[8] W. Dong, J. Li, R. Yao, C. Li, T. Yuan, L. Wang, Characterizing driving styles with deep learning, arXiv preprint arXiv:1607.03611 (2016).

[9] I. M. Enholm, E. Papagiannidis, P. Mikalef, J. Krogstie, Artificial intelligence and business value: A literature review, Information Systems Frontiers (2021) 1–26.

[10] I. H. Sarker, Machine learning: Algorithms, real-world applications and research directions, SN Computer Science 2 (2021) 1–21.

[11] A. Saxe, S. Nelli, C. Summerfield, If deep learning is the answer, what is the question?, Nature Reviews Neuroscience 22 (2021) 55–67.

[12] F. Piccialli, V. Di Somma, F. Giampaolo, S. Cuomo, G. Fortino, A survey on deep learning in medicine: Why, how and when?, Information Fusion 66 (2021) 111–137.

[13] Z. Li, F. Liu, W. Yang, S. Peng, J. Zhou, A survey of convolutional neural networks: analysis, applications, and prospects, IEEE Transactions on Neural Networks and Learning Systems (2021).

[14] G. Vilone, L. Longo, Notions of explainability and evaluation approaches for explainable artificial intelligence, Information Fusion 76 (2021) 89–106.

[15] F. Doshi-Velez, B. Kim, Towards a rigorous science of interpretable machine learning, arXiv preprint arXiv:1702.08608 (2017).

[16] J. M. Alonso, C. Castiello, L. Magdalena, C. Mencar, Explainable Fuzzy Systems - Paving the Way from Interpretable Fuzzy Systems to Explainable AI Systems, Vol. 970, Springer International Publishing, 2021, http://dx.doi.org/10.1007/978-3-030-71098-9.

[17] J. Pearl, D. Mckenzie, The book of why: The New Science of Cause and Effect, Penguin, 2018.

[18] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, ACM computing surveys (CSUR) 51 (5) (2018) 1–42.

[19] L. Edwards, M. Veale, Slave to the algorithm: Why a right to an explanation is probably not the remedy you are looking for, Duke L. & Tech. Rev. 16 (2017) 18.

[20] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, G.-Z. Yang, XAI: Explainable artificial intelligence, Science Robotics 4 (37) (2019). doi:10.1126/scirobotics.aay7120.

[21] A. Adadi, M. Berrada, Peeking inside the black-box: a survey on explainable artificial intelligence (XAI), IEEE access 6 (2018) 52138–52160.

[22] T. Rieg, J. Frick, H. Baumgartl, R. Buettner, Demonstration of the potential of white-box machine learning approaches to gain insights from cardiovascular disease electrocardiograms, PloS one 15 (12) (2020) e0243615.

[23] C. Véliz, C. Prunkl, M. Phillips-Brown, T. M. Lechterman, We might be afraid of black-box algorithms, Journal of Medical Ethics 47 (2021) 339–340.

[24] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A survey on bias and fairness in machine learning, ACM Computing Surveys (CSUR) 54 (2021) 1–35.

[25] C. Finlay, A. M. Oberman, Scaleable input gradient regularization for adversarial robustness, Machine Learning with Applications 3 (2021) 100017.

[26] X. Li, H. Xiong, X. Li, X. Wu, X. Zhang, J. Liu, J. Bian, D. Dou, Interpretable deep learning: Interpretation, interpretability, trustworthiness, and beyond, arXiv preprint arXiv:2103.10689 (2021).

[27] A. Das, P. Rad, Opportunities and challenges in explainable artificial intelligence (XAI): A survey, arXiv preprint arXiv:2006.11371 (2020).

[28] Z. C. Lipton, The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery, Communications of the ACM (CACM) (2018) 31–57.

[29] M. T. Ribeiro, S. Singh, C. Guestrin, " 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, 2016, pp. 1135–1144.

[30] H. Wang, Z. Wang, M. Du, F. Yang, Z. Zhang, S. Ding, P. Mardziel, X. Hu, Score-cam: Score-weighted visual explanations for convolutional neural networks, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, 2020, pp. 24–25.

[31] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 618–626.

[32] S. T. Mueller, R. R. Hoffman, W. Clancey, A. Emrey, G. Klein, Explanation in human-ai systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable ai, arXiv preprint arXiv:1902.01876 (2019).

[33] B. Chandrasekaran, M. C. Tanner, J. R. Josephson, Explaining control strategies in problem solving, IEEE Intelligent Systems 4 (01) (1989) 9–15.

[34] W. R. Swartout, J. D. Moore, Explanation in second generation expert systems, in: Second generation expert systems, Springer, 1993, pp. 543–585.

[35] W. R. Swartout, Rule-based expert systems: The mycin experiments of the stanford heuristic programming project: Bg buchanan and eh shortliffe,(addison-wesley, reading, ma, 1984); 702 pages, 40.50 (1985).

[36] L. A. Zadeh, Fuzzy sets, Information and Control 8 (1965) 338–353.

[37] L. A. Zadeh, Outline of a new approach to the analysis of complex systems and decision processes, IEEE Transactions on Systems, Man, and Cybernetics 3 (1973) 28–44.

[38] L. A. Zadeh, The concept of a linguistic variable and its application to approximate reasoning, Information Sciences 8 (1975) 199–249.

[39] R. Andrews, J. Diederich, A. B. Tickle, Survey and critique of techniques for extracting rules from trained artificial neural networks, Knowledge-based systems 8 (6) (1995) 373–389.

[40] C. Lacave, F. J. Díez, A review of explanation methods for bayesian networks, The Knowledge Engineering Review 17 (2) (2002) 107–127.

[41] H. Cramer, V. Evers, S. Ramlal, M. Van Someren, L. Rutledge, N. Stash, L. Aroyo, B. Wielinga, The effects of transparency on trust in and acceptance of a content-based art recommender, User Modeling and User-adapted interaction 18 (5) (2008) 455.

[42] J. L. Herlocker, J. A. Konstan, J. Riedl, Explaining collaborative filtering recommendations, in: Proceedings of the 2000 ACM conference on Computer supported cooperative work, 2000, pp. 241–250.

[43] D. Doyle, A. Tsymbal, P. Cunningham, A review of explanation and explanation in case-based reasoning, Trinity College Dublin, Department of Computer Science (2003).

[44] H. Burns, C. A. Luckhardt, J. W. Parlett, C. L. Redfield, Intelligent tutoring systems: Evolutions in design, Psychology Press, 2014.

[45] W. Park, D. H. Yoo, J. Jaworski, J. Brzezicki, A. Gnylorybov, V. Kadinov, I. G. Sariego, C. Abud-Mendoza, W. J. O. Escalante, S. W. Kang, et al., Comparable long-term efficacy, as assessed by patient-reported outcomes, safety and pharmacokinetics, of ct-p13 and reference infliximab in patients with ankylosing spondylitis: 54-week results from the randomized, parallel-group planetas study, Arthritis research & therapy 18 (2016) 1–11.

[46] N. McCarty, K. T. Poole, H. Rosenthal, Polarized America: The dance of ideology and unequal riches, mit Press, 2016.

[47] R. Confalonieri, F. Lucchesi, G. Maffei, S. Catuara-Solarz, A unified framework for managing sex and gender bias in ai models for healthcare, in: D. Cirillo, S. Catuara-Solarz, E. Guney (Eds.), Sex and Gender Bias in Technology and Artificial Intelligence, Academic Press, 2022, pp. 179–204.

[48] B. Yun, M. Croitoru, P. Bisquert, S. Vesic, Graph theoretical properties of logic based argumentation frameworks, in: AAMAS: Autonomous Agents and Multiagent Systems, ACM, 2018, pp. 2148–2149.

[49] C. Meske, E. Bunde, Transparency and trust in human-ai-interaction: The role of model-agnostic explanations in computer vision-based decision support, in: International Conference on Human-Computer Interaction, Springer, 2020, pp. 54–69.

[50] C. Meske, E. Bunde, J. Schneider, M. Gersch, Explainable artificial intelligence: objectives, stakeholders, and future research opportunities, Information Systems Management 39 (1) (2022) 53–63.

[51] S. R. Islam, W. Eberle, S. K. Ghafoor, M. Ahmed, Explainable artificial intelligence approaches: A survey, CoRR (2021).

[52] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, Artificial intelligence 267 (2019) 1–38.

[53] A. Vassiliades, N. Bassiliades, T. Patkos, Argumentation and explainable artificial intelligence: a survey, The Knowledge Engineering Review 36 (2021).

[54] F. Hussain, R. Hussain, E. Hossain, Explainable artificial intelligence (XAI): An engineering perspective, arXiv preprint arXiv:2101.03613 (2021).

[55] S. Liu, X. Wang, M. Liu, J. Zhu, Towards better analysis of machine learning models: A visual analytics perspective, Visual Informatics 1 (1) (2017) 48–56.

[56] Q. Zhang, S.-C. Zhu, Visual interpretability for deep learning: a survey, Frontiers Inf Technol Electronic Eng 19 (2018) 27–39.

[57] G. Ras, M. van Gerven, P. Haselager, Explanation methods in deep learning: Users, values, concerns and challenges, in: Explainable and interpretable models in computer vision and machine learning, Springer, 2018, pp. 19–36.

[58] G. Montavon, W. Samek, K.-R. Müller, Methods for interpreting and understanding deep neural networks, Digital Signal Processing 73 (2018) 1–15.

[59] A. Heuillet, F. Couthouis, N. Díaz-Rodríguez, Explainability in deep reinforcement learning, Knowledge-Based Systems 214 (2021) 106685.

[60] E. Puiutta, E. M. Veith, Explainable reinforcement learning: A survey, in: International Cross-Domain Conference for Machine Learning and Knowledge Extraction, Springer, 2020, pp. 77–95.

[61] N. Burkart, M. F. Huber, A survey on the explainability of supervised machine learning, Journal of Artificial Intelligence Research 70 (2021) 245–317.

[62] J. Gerlings, A. Shollo, I. Constantiou, Reviewing the need for explainable artificial intelligence (XAI), HICSS (2021).

[63] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al., Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible ai, Information Fusion 58 (2020) 82–115.

[64] F. K. Došilović, M. Brčić, N. Hlupić, Explainable artificial intelligence: A survey, in: 2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO), IEEE, 2018, pp. 0210–0215.

[65] M. Nauta, J. Trienes, S. Pathak, E. Nguyen, M. Peters, Y. Schmitt, J. Schlötterer, M. van Keulen, C. Seifert, From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai, arXiv preprint arXiv:2201.08164 (2022).

[66] A. Abdul, J. Vermeulen, D. Wang, B. Y. Lim, M. Kankanhalli, Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda, in: Proceedings of the 2018 CHI conference on human factors in computing systems, 2018, pp. 1–18.

[67] W. Samek, T. Wiegand, K.-R. Müller, Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models, ITU Journal: ICT Discoveries (2018) 39–48.

[68] D. V. Carvalho, E. M. Pereira, J. S. Cardoso, Machine learning interpretability: A survey on methods and metrics, Electronics 8 (8) (2019) 832.

[69] G. Vilone, L. Longo, Explainable artificial intelligence: a systematic review, arXiv preprint arXiv:2006.00093 (2020).

[70] P. Linardatos, V. Papastefanopoulos, S. Kotsiantis, Explainable ai: A review of machine learning interpretability methods, Entropy 23 (1) (2021) 18.

[71] M. Langer, D. Oster, T. Speith, H. Hermanns, L. Kästner, E. Schmidt, A. Sesing, K. Baum, What do we want from explainable artificial intelligence (XAI)?–a stakeholder perspective on XAI and a

conceptual model guiding interdisciplinary XAI research, Artificial Intelligence 296 (2021) 103473.
[72] R. Confalonieri, L. Coba, B. Wagner, T. R. Besold, A historical perspective of explainable artificial intelligence, WIREs Data Mining and Knowledge Discovery 11 (1) (2021) e1391.
[73] E. Cambria, L. Malandri, F. Mercorio, M. Mezzanzanica, N. Nobani, A survey on xai and natural language explanations, Information Processing & Management 60 (1) (2023) 103111.
[74] I. Tiddi, S. Schlobach, Knowledge graphs as tools for explainable machine learning: A survey, Artificial Intelligence 302 (2022) 103627.
[75] Y.-L. Chou, C. Moreira, P. Bruza, C. Ouyang, J. Jorge, Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms, and applications, Information Fusion 81 (2022) 59–83.
[76] A. Holzinger, A. Saranti, C. Molnar, P. Biecek, W. Samek, Explainable ai methods-a brief overview, in: International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers, Springer, 2022, pp. 13–38.
[77] R. Guidotti, Counterfactual explanations and how to find them: literature review and benchmarking, Data Mining and Knowledge Discovery (2022) 1–55.
[78] A. Theissler, F. Spinnato, U. Schlegel, R. Guidotti, Explainable ai for time series classification: A review, taxonomy and research directions, IEEE Access (2022).
[79] G. Yang, Q. Ye, J. Xia, Unbox the black-box for the medical explainable ai via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond, Information Fusion 77 (2022) 29–52.
[80] I. Stepin, J. M. Alonso, A. Catala, M. Pereira-Farina, A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence, IEEE Access 9 (2021) 11974–12001. doi:10.1109/ACCESS.2021.3051315.
[81] Y. Zhang, X. Chen, et al., Explainable recommendation: A survey and new perspectives, Foundations and Trends® in Information Retrieval 14 (1) (2020) 1–101.
[82] R. M. Byrne, Counterfactuals in explainable artificial intelligence (XAI): Evidence from human reasoning, in: IJCAI, 2019, pp. 6276–6282.
[83] R. Goebel, A. Chander, K. Holzinger, F. Lecue, Z. Akata, S. Stumpf, P. Kieseberg, A. Holzinger, Explainable ai: the new 42?, in: International cross-domain conference for machine learning and knowledge extraction, Springer, 2018, pp. 295–303.
[84] M. G. Core, H. C. Lane, M. Van Lent, D. Gomboc, S. Solomon, M. Rosenberg, et al., Building explainable artificial intelligence systems, in: AAAI, 2006, pp. 1766–1773.
[85] M. Van Lent, W. Fisher, M. Mancuso, An explainable artificial intelligence system for small-unit tactical behavior, in: Proceedings of the national conference on artificial intelligence, Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2004, pp. 900–907.
[86] G. Alicioglu, B. Sun, A survey of visual analytics for explainable artificial intelligence methods, Computers & Graphics (2021).
[87] S. Mohseni, N. Zarei, E. D. Ragan, A multidisciplinary survey and framework for design and evaluation of explainable ai systems, ACM Transactions on Interactive Intelligent Systems (TiiS) 11 (3-4) (2021) 1–45.
[88] J. R. Josephson, S. G. Josephson, Abductive inference: Computation, philosophy, technology, Cambridge University Press, 1996.
[89] T. Lombrozo, The structure and function of explanations, Trends in cognitive sciences 10 (10) (2006) 464–470.
[90] B. F. Malle, How the mind explains behavior: Folk explanations, meaning, and social interaction, Mit Press, 2006.
[91] H. P. Grice, Logic and conversation, syntax and semantics, Speech Acts 3 (1975) 41–58.
[92] S. Anjomshoae, D. Omeiza, L. Jiang, Context-based image explanations for deep neural networks, Image and Vision Computing 116 (2021) 104310.
[93] E. Thelisson, Towards trust, transparency and liability in ai/as systems., in: IJCAI, 2017, pp. 5215–5216.
[94] S. Larsson, F. Heintz, Transparency in artificial intelligence, Internet Policy Review 9 (2020) 1–16.
[95] V. Bogina, A. Hartman, T. Kuflik, A. Shulner-Tal, Educating software and ai stakeholders about algorithmic fairness, accountability, transparency and ethics, International Journal of Artificial Intelligence in Education (2021) 1–26.
[96] T. Calders, E. Ntoutsi, M. Pechenizkiy, B. Rosenhahn, S. Ruggieri, Introduction to the special section on bias and fairness in ai, ACM SIGKDD Explorations Newsletter 23 (2021) 1–3.
[97] L. Lyu, J. Yu, K. Nandakumar, Y. Li, X. Ma, J. Jin, H. Yu, K. S. Ng, Towards fair and privacy-preserving federated deep models, IEEE Transactions on Parallel and Distributed Systems 31 (2020) 2524–2541.
[98] S. Khalilpourazari, S. Khalilpourazary, A. Özyüksel Çiftçioğlu, G.-W. Weber, Designing energy-efficient high-precision multi-pass turning processes via robust optimization and artificial intelligence, Journal of Intelligent Manufacturing 32 (2021) 1621–1647.
[99] A. Subbaswamy, R. Adams, S. Saria, Evaluating model robustness and stability to dataset shift, in: International Conference on Artificial Intelligence and Statistics, PMLR, 2021, pp. 2611–2619.
[100] M. Holland, Robustness and scalability under heavy tails, without strong convexity, in: International Conference on Artificial Intelligence and Statistics, PMLR, 2021, pp. 865–873.
[101] M. Naser, An engineer's guide to explainable artificial intelligence and interpretable machine learning: Navigating causality, forced goodness, and the false perception of inference, Automation in Construction 129 (2021) 103821.
[102] Y.-L. Chou, C. Moreira, P. Bruza, C. Ouyang, J. Jorge, Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms, and applications, arXiv preprint arXiv:2103.04244 (2021).
[103] R. Moraffah, M. Karami, R. Guo, A. Raglin, H. Liu, Causal interpretability for machine learning - problems, methods and evaluation, Association for Computing Machinery (2020).
[104] X. Cui, J. M. Lee, J. P.-A. Hsieh, An integrative 3c evaluation framework for explainable artificial intelligence, in: AMCIS, 2019, pp. 1–10.
[105] M. Coeckelbergh, Artificial intelligence, responsibility attribution, and a relational justification of explainability, Science and engineering ethics 26 (4) (2020) 2051–2068.
[106] J. M. Alonso Moral, C. Castiello, L. Magdalena, C. Mencar, Toward explainable artificial intelligence through fuzzy systems, in: Explainable Fuzzy Systems, Springer, 2021, pp. 1–23.
[107] O. Biran, C. Cotton, Explanation and justification in machine learning: A survey, in: IJCAI-17 workshop on explainable AI (XAI), Vol. 8, 2017, pp. 8–13.
[108] A. A. Freitas, Comprehensible classification models: A position paper, Association for Computing Machinery (2014).
[109] A. Kotriwala, B. Klöpper, M. Dix, G. Gopalakrishnan, D. Ziobro, A. Potschka, Xai for operations in the process industry-applications, theses, and research directions., in: AAAI Spring Symposium: Combining Machine Learning with Knowledge Engineering, 2021, pp. 1–12.
[110] C.-C. Chang, X. Wang, S. Chen, I. Echizen, V. Sanchez, C.-T. Li, Deep learning for reversible steganography: Principles and insights, arXiv preprint arXiv:2106.06924 (2021).
[111] M. Ghassemi, L. Oakden-Rayner, A. L. Beam, The false hope of current approaches to explainable artificial intelligence in health care, The Lancet Digital Health 3 (2021) e745–e750.
[112] A. Galli, S. Marrone, V. Moscato, C. Sansone, Reliability of explainable artificial intelligence in adversarial perturbation scenarios, in: International Conference on Pattern Recognition, Springer, 2021, pp. 243–256.
[113] M. Szczepański, M. Choraś, M. Pawlicki, A. Pawlicka, The methods and approaches of explainable artificial intelligence, in: International Conference on Computational Science, Springer, 2021, pp. 3–17.
[114] P. P. Angelov, E. A. Soares, R. Jiang, N. I. Arnold, P. M. Atkinson, Explainable artificial intelligence: an analytical review, Wiley

Interdisciplinary Reviews: Data Mining and Knowledge Discovery 11 (2021) e1424.
[115] R. Confalonieri, L. Coba, B. Wagner, T. R. Besold, A historical perspective of explainable artificial intelligence, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 11 (2021) e1391.
[116] J. H.-w. Hsiao, H. H. T. Ngai, L. Qiu, Y. Yang, C. C. Cao, Roadmap of designing cognitive metrics for explainable artificial intelligence (XAI), arXiv preprint arXiv:2108.01737 (2021).
[117] A. Rosenfeld, Better metrics for evaluating explainable artificial intelligence, in: Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems, 2021, pp. 45–50.
[118] D. Alvarez-Melis, T. S. Jaakkola, Towards robust interpretability with self-explaining neural networks, 32nd Conference on Neural Information Processing Systems (2018).
[119] S. El-Sappagh, J. M. Alonso, S. R. Islam, A. M. Sultan, K. S. Kwak, A multilayer multimodal detection and prediction model based on explainable artificial intelligence for alzheimer's disease, Scientific reports 11 (2021) 1–26.
[120] H. Smith, Clinical ai: opacity, accountability, responsibility and liability, AI & SOCIETY 36 (2) (2021) 535–545.
[121] B. Lepri, N. Oliver, A. Pentland, Ethical machines: the human-centric use of artificial intelligence, Iscience (2021) 102249.
[122] F. S. de Sio, G. Mecacci, Four responsibility gaps with artificial intelligence: Why they matter and how to address them, Philosophy & Technology (2021) 1–28.
[123] F. Santoni de Sio, The european commission report on ethics of connected and automated vehicles and the future of ethics of transportation, Ethics and Information Technology (2021) 1–14.
[124] P. Liu, M. Du, T. Li, Psychological consequences of legal responsibility misattribution associated with automated vehicles, Ethics and information technology (2021) 1–14.
[125] C. Zednik, Solving the black box problem: a normative framework for explainable artificial intelligence, Philosophy & Technology 34 (2021) 265–288.
[126] A. Bécue, I. Praça, J. Gama, Artificial intelligence, cyber-threats and industry 4.0: Challenges and opportunities, Artificial Intelligence Review 54 (2021) 3849–3886.
[127] S.-C. Fischer, A. Wenger, Artificial intelligence, forward-looking governance and the future of security, Swiss Political Science Review 27 (2021) 170–179.
[128] Z. Zhang, H. Ning, F. Shi, F. Farha, Y. Xu, J. Xu, F. Zhang, K.-K. R. Choo, Artificial intelligence in cyber security: research advances, challenges, and opportunities, Artificial Intelligence Review (2021) 1–25.
[129] H. Mankodiya, M. S. Obaidat, R. Gupta, S. Tanwar, Xai-av: Explainable artificial intelligence for trust management in autonomous vehicles, in: 2021 International Conference on Communications, Computing, Cybersecurity, and Informatics (CCCI), IEEE, 2021, pp. 1–5.
[130] R. Sheh, Explainable artificial intelligence requirements for safe, intelligent robots, in: 2021 IEEE International Conference on Intelligence and Safety for Robotics (ISR), IEEE, 2021, pp. 382–387.
[131] T. Rojat, R. Puget, D. Filliat, J. Del Ser, R. Gelin, N. Díaz-Rodríguez, Explainable artificial intelligence (xai) on timeseries data: A survey, arXiv preprint arXiv:2104.00950 (2021).
[132] C. Rudin, C. Chen, Z. Chen, H. Huang, L. Semenova, C. Zhong, Interpretable machine learning: Fundamental principles and 10 grand challenges, Statistics Surveys 16 (none) (2022) 1 – 85. doi:10.1214/21-SS133.
[133] D. Doran, S. Schulz, T. R. Besold, What does explainable ai really mean? a new conceptualization of perspectives, arXiv preprint arXiv:1710.00794 (2017).
[134] A. Carrington, P. Fieguth, H. Chen, Measures of model interpretability for model selection, in: International Cross-Domain Conference for Machine Learning and Knowledge Extraction, Springer, 2018, pp. 329–349.
[135] P. Biecek, T. Burzykowski, Explanatory model analysis: explore, explain, and examine predictive models, CRC Press, 2021.
[136] B. Herman, The promise and peril of human evaluation for model interpretability, arXiv preprint arXiv:1711.07414 (2017) 8.
[137] A. Preece, Asking 'why'in ai: Explainability of intelligent systems–perspectives and challenges, Intelligent Systems in Accounting, Finance and Management 25 (2) (2018) 63–72.
[138] W. Wang, I. Benbasat, Recommendation agents for electronic commerce: Effects of explanation facilities on trusting beliefs, Journal of Management Information Systems 23 (4) (2007) 217–246.
[139] J. Krause, A. Perer, K. Ng, Interacting with predictions: Visual inspection of black-box machine learning models, in: Proceedings of the 2016 CHI conference on human factors in computing systems, 2016, pp. 5686–5697.
[140] M. Ribera, A. Lapedriza, Can we do better explanations? a proposal of user-centered explainable ai., in: IUI Workshops, Vol. 2327, 2019, p. 38.
[141] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Proceedings of the 31st international conference on neural information processing systems, 2017, pp. 4768–4777.
[142] L. Breiman, Statistical modeling: The two cultures (with comments and a rejoinder by the author), Statistical science 16 (3) (2001) 199–231.
[143] Z. C. Lipton, The mythos of model interpretability, Communications of the ACM 61 (10) (2018) 36–43. doi:10.1145/3233231.
[144] M. W. Craven, J. W. Shavlik, Extracting tree-structured representations of trained networks, in: Proceedings of NIPS, 1995, pp. 24–30.
[145] Z. F. Hu, T. Kuflik, I. G. Mocanu, S. Najafian, A. Shulner Tal, Recent studies of XAI-review, in: Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization, 2021, pp. 421–431.
[146] M. R. Zafar, N. Khan, Deterministic local interpretable model-agnostic explanations for stable explainability, Machine Learning and Knowledge Extraction 3 (2021) 525–541.
[147] M. Ivanovs, R. Kadikis, K. Ozols, Perturbation-based methods for explaining deep neural networks: A survey, Pattern Recognition Letters (2021).
[148] S. Tilouche, V. Partovi Nia, S. Bassetto, Parallel coordinate order for high-dimensional data, Statistical Analysis and Data Mining: The ASA Data Science Journal 14 (5) (2021) 501–515.
[149] C. Molnar, Interpretable machine learning, Lulu. com, 2020.
[150] L. McInnes, J. Healy, J. Melville, Umap: Uniform manifold approximation and projection for dimension reduction, arXiv preprint arXiv:1802.03426 (2020).
[151] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, B. Yu, Definitions, methods, and applications in interpretable machine learning, Proceedings of the National Academy of Sciences 116 (44) (2019) 22071–22080.
[152] T. Campbell, T. Broderick, Automated scalable bayesian inference via hilbert coresets, The Journal of Machine Learning Research 20 (1) (2019) 551–588.
[153] I. Higgins, D. Amos, D. Pfau, S. Racaniere, L. Matthey, D. Rezende, A. Lerchner, Towards a definition of disentangled representations, arXiv preprint arXiv:1812.02230 (2018).
[154] M. Al-Shedivat, A. Dubey, E. P. Xing, The intriguing properties of model explanations, arXiv preprint arXiv:1801.09808 (2018).
[155] J. Wexler, Facets: An open source visualization tool for machine learning training data, Google Open Source Blog (2017).
[156] J. Matejka, G. Fitzmaurice, Same stats, different graphs: generating datasets with varied appearance and identical statistics through simulated annealing, in: Proceedings of the 2017 CHI conference on human factors in computing systems, 2017, pp. 1290–1294.
[157] D. Smilkov, N. Thorat, C. Nicholson, E. Reif, F. B. Viégas, M. Wattenberg, Embedding projector: Interactive visualization and interpretation of embeddings, arXiv preprint arXiv:1611.05469 (2016).
[158] B. Kim, R. Khanna, O. O. Koyejo, Examples are not enough, learn to criticize! criticism for interpretability, Advances in neural information processing systems 29 (2016).
[159] J. Bien, R. Tibshirani, Prototype selection for interpretable classification, The Annals of Applied Statistics 5 (4) (2011) 2403–2424.

[160] H. Lin, J. Bilmes, A class of submodular functions for document summarization, in: Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies, 2011, pp. 510–520.

[161] L. Van der Maaten, G. Hinton, Visualizing data using t-sne., Journal of machine learning research 9 (2008).

[162] T. Shi, B. Yu, E. E. Clothiaux, A. J. Braverman, Daytime arctic cloud detection based on multi-angle satellite data with case studies, Journal of the American Statistical Association 103 (482) (2008) 584–593.

[163] W. DuMouchel, Data squashing: constructing summary data sets, in: Handbook of Massive Data Sets, Springer, 2002, pp. 579–591.

[164] R. Kohavi, B. Becker, UCI Machine Learning Repository: Adult Data Set (1996).
URL https://archive.ics.uci.edu/ml/datasets/Adult

[165] R. Severino, The Data Visualisation Catalogue (2015).
URL https://datavizcatalogue.com

[166] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. Daumé III, K. Crawford, Datasheets for datasets, arXiv preprint arXiv:1803.09010 (2018).

[167] S. Holland, A. Hosny, S. Newman, J. Joseph, K. Chmielinski, The dataset nutrition label: A framework to drive higher data quality standards, arXiv preprint arXiv:1805.03677 (2018).

[168] E. M. Bender, B. Friedman, Data statements for natural language processing: Toward mitigating system bias and enabling better science, Transactions of the Association for Computational Linguistics 6 (2018) 587–604.

[169] R. Caruana, H. Kangarloo, J. D. Dionisio, U. Sinha, D. Johnson, Case-based explanation of non-case-based learning methods., in: Proceedings of the AMIA Symposium, American Medical Informatics Association, 1999, p. 212.

[170] I. Simon, N. Snavely, S. M. Seitz, Scene summarization for online image collections, in: 2007 IEEE 11th International Conference on Computer Vision, IEEE, 2007, pp. 1–8.

[171] J. M. Rožanec, B. Fortuna, D. Mladenić, Knowledge graph-based rich and confidentiality preserving explainable artificial intelligence (XAI), Information Fusion 81 (2022) 91–102.

[172] C. Panigutti, A. Perotti, D. Pedreschi, Doctor XAI: an ontology-based approach to black-box sequential data classification explanations, in: Proceedings of the 2020 conference on fairness, accountability, and transparency, 2020, pp. 629–639.

[173] M. Gaur, K. Faldu, A. Sheth, Semantics of the black-box: Can knowledge graphs help make deep learning systems more interpretable and explainable?, IEEE Internet Computing 25 (1) (2021) 51–59.

[174] M. Cannataro, C. Comito, A data mining ontology for grid programming, in: Proc. 1st Int. Workshop on Semantics in Peer-to-Peer and Grid Computing, Citeseer, 2003, pp. 113–134.

[175] C. Diamantini, D. Potena, E. Storti, Kddonto: An ontology for discovery and composition of kdd algorithms, Third Generation Data Mining: Towards Service-Oriented Knowledge Discovery (SoKD'09) (2009) 13–24.

[176] P. Panov, L. Soldatova, S. Džeroski, Ontology of core data mining entities, Data Mining and Knowledge Discovery 28 (5) (2014) 1222–1265.

[177] R. Confalonieri, T. Weyde, T. R. Besold, F. M. del Prado Martín, Using ontologies to enhance human understandability of global post-hoc explanations of black-box models, Artificial Intelligence 296 (2021). doi:https://doi.org/10.1016/j.artint.2021.103471.

[178] M. W. Craven, Extracting comprehensible models from trained neural networks, The University of Wisconsin-Madison, 1996.

[179] M. Reichstein, G. Camps-Valls, B. Stevens, M. Jung, J. Denzler, N. Carvalhais, Deep learning and process understanding for data-driven earth system science, Nature 566 (7743) (2019) 195–204.

[180] T. Bao, S. Chen, T. T. Johnson, P. Givi, S. Sammak, X. Jia, Physics guided neural networks for spatio-temporal super-resolution of turbulent flows, in: Uncertainty in Artificial Intelligence, PMLR, 2022, pp. 118–128.

[181] S. Seo, S. Arik, J. Yoon, X. Zhang, K. Sohn, T. Pfister, Controlling neural networks with rule representations, Advances in Neural Information Processing Systems 34 (2021) 11196–11207.

[182] R. Wang, R. Yu, Physics-guided deep learning for dynamical systems: A survey, arXiv preprint arXiv:2107.01272 (2021).

[183] M. Al-Shedivat, A. Dubey, E. P. Xing, Contextual explanation networks., J. Mach. Learn. Res. 21 (2020).

[184] R. Ghaeini, X. Z. Fern, H. Shahbazi, P. Tadepalli, Saliency learning: Teaching the model where to pay attention, in: Proceedings of NAACL-HLT 2019, 2019, pp. 4016–4025.

[185] C. Chen, O. Li, A. Barnett, J. K. Su, C. Rudin, This looks like that: deep learning for interpretable image recognition, in: Proceedings of the 33rd International Conference on Neural Information Processing Systems, 2019, pp. 1–12.

[186] D. Hu, An introductory survey on attention mechanisms in nlp problems, in: Proceedings of SAI Intelligent Systems Conference, Springer, 2019, pp. 432–448.

[187] D. Card, M. Zhang, N. A. Smith, Deep weighted averaging classifiers, in: Proceedings of the Conference on Fairness, Accountability, and Transparency, 2019, pp. 369–378.

[188] W. Brendel, M. Bethge, Approximating CNNs with bag-of-local-features models works surprisingly well on imagenet, in: International Conference on Learning Representations, 2019, pp. 1–15.

[189] M. Hind, D. Wei, M. Campbell, N. C. Codella, A. Dhurandhar, A. Mojsilović, K. Natesan Ramamurthy, K. R. Varshney, Ted: Teaching ai to explain its decisions, in: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, 2019, pp. 123–129.

[190] S. Jain, B. C. Wallace, Attention is not explanation, arXiv preprint arXiv:1902.10186 (2019).

[191] D. H. Park, L. A. Hendricks, Z. Akata, A. Rohrbach, B. Schiele, T. Darrell, M. Rohrbach, Multimodal explanations: Justifying decisions and pointing to the evidence, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8779–8788.

[192] M. Wu, M. C. Hughes, S. Parbhoo, M. Zazzi, V. Roth, F. Doshi-Velez, Beyond sparsity: Tree regularization of deep models for interpretability, in: Thirty-Second AAAI Conference on Artificial Intelligence, 2018, pp. 1670–1678.

[193] Q. Zhang, Y. N. Wu, S.-C. Zhu, Interpretable convolutional neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8827–8836.

[194] E. Angelino, N. Larus-Stone, D. Alabi, M. Seltzer, C. Rudin, Learning certifiably optimal rule lists for categorical data, J. Mach. Learn. Res. (2018).

[195] N. Papernot, P. McDaniel, Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning, arXiv preprint arXiv:1803.04765 (2018).

[196] A. S. Ross, M. C. Hughes, F. Doshi-Velez, Right for the right reasons: Training differentiable models by constraining their explanations, in: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17, 2017, pp. 2662–2670.

[197] Z. Wojna, A. N. Gorban, D.-S. Lee, K. Murphy, Q. Yu, Y. Li, J. Ibarz, Attention-based extraction of structured information from street view imagery, in: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), Vol. 1, IEEE, 2017, pp. 844–850.

[198] E. Choi, M. T. Bahadori, J. A. Kulas, A. Schuetz, W. F. Stewart, J. Sun, Retain: An interpretable predictive model for healthcare using reverse time attention mechanism, 30th Conference on Neural Information Processing Systems (NIPS) (2016).

[199] L. A. Hendricks, Z. Akata, M. Rohrbach, J. Donahue, B. Schiele, T. Darrell, Generating visual explanations, in: European conference on computer vision, Springer, 2016, pp. 3–19.

[200] T. Lei, R. Barzilay, T. Jaakkola, Rationalizing neural predictions, Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (2016).

[201] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, 3rd International Conference on Learning Representations (2015).

[202] B. Ustun, C. Rudin, Supersparse linear integer models for optimized medical scoring systems, Machine Learning 102 (3) (2016) 349–391.

[203] G. P. Schmitz, C. Aldrich, F. S. Gouws, Ann-dt: an algorithm for extraction of decision trees from artificial neural networks, IEEE Transactions on Neural Networks 10 (6) (1999) 1392–1401.

[204] H. Lakkaraju, S. H. Bach, J. Leskovec, Interpretable decision sets: A joint framework for description and prediction, in: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, pp. 1675–1684.

[205] J. Jung, C. Concannon, R. Shroff, S. Goel, D. G. Goldstein, Simple rules for complex decisions, Cognitive Social Science eJournal (2017).

[206] C. M. J.M. Alonso, C. Castiello, Interpretability of fuzzy systems: Current research trends and prospects, Springer Handbook of Computational Intelligence (2015) 219–237doi:10.1109/ACCESS.2021.3051315.

[207] Y. Lou, R. Caruana, J. Gehrke, G. Hooker, Accurate intelligible models with pairwise interactions, in: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, 2013, pp. 623–631.

[208] D. Schreiber-Gregory, Regulation techniques for multicollinearity: Lasso, ridge, and elastic nets, in: SAS Conference Proceedings: Western Users of SAS Software 2018, 2018, pp. 1–23.

[209] J. Wanner, L.-V. Herm, K. Heinrich, C. Janiesch, Stop ordering machine learning algorithms by their explainability! an empirical investigation of the tradeoff between performance and explainability, in: Conference on e-Business, e-Services and e-Society, Springer, 2021, pp. 245–258.

[210] S. Saisubramanian, S. Galhotra, S. Zilberstein, Balancing the tradeoff between clustering value and interpretability, in: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 2020, pp. 351–357.

[211] T.-N. Chou, An explainable hybrid model for bankruptcy prediction based on the decision tree and deep neural network, in: 2019 IEEE 2nd International Conference on Knowledge Innovation and Invention (ICKII), IEEE, 2019, pp. 122–125.

[212] S. Sukhbaatar, J. Weston, R. Fergus, et al., End-to-end memory networks, Advances in neural information processing systems 28 (2015).

[213] A. d. Garcez, S. Bader, H. Bowman, L. C. Lamb, L. de Penning, B. Illuminoo, H. Poon, C. Gerson Zaverucha, Neural-symbolic learning and reasoning: A survey and interpretation, Neuro-Symbolic Artificial Intelligence: The State of the Art 342 (2022) 1.

[214] H. Jaeger, Controlling recurrent neural networks by conceptors, arXiv preprint arXiv:1403.3369 (2014).

[215] C. Widmer, M. K. Sarker, S. Nadella, J. Fiechter, I. Juvina, B. Minnery, P. Hitzler, J. Schwartz, M. Raymer, Towards human-compatible xai: Explaining data differentials with concept induction over background knowledge, arXiv preprint arXiv:2209.13710 (2022).

[216] G. Ciravegna, P. Barbiero, F. Giannini, M. Gori, P. Lió, M. Maggini, S. Melacci, Logic explained networks, Artificial Intelligence 314 (2023) 103822.

[217] F. Amodeo, F. Caballero, N. Díaz-Rodríguez, L. Merino, Og-sgg: Ontology-guided scene graph generation. a case study in transfer learning for telepresence robotics, IEEE Access 10 (2022) 132564–132583. doi:10.1109/ACCESS.2022.3230590.
URL https://ieeexplore.ieee.org/document/9991965

[218] A. Bennetot, G. Franchi, J. Del Ser, R. Chatila, N. Díaz-Rodríguez, Greybox xai: a neural-symbolic learning framework to produce interpretable predictions for image classification, Knowledge-Based Systems 258 (2022) 109947.

[219] N. Díaz-Rodríguez, A. Lamas, J. Sanchez, G. Franchi, I. Donadello, S. Tabik, D. Filliat, P. Cruz, R. Montes, F. Herrera, Explainable neural-symbolic learning (x-nesyl) methodology to fuse deep learning representations with expert knowledge graphs: the monumai cultural heritage use case, Information Fusion 79 (2021) 58–83.

[220] K. Kaczmarek-Majer, G. Casalino, G. Castellano, M. Dominiak, O. Hryniewicz, O. Kamińska, G. Vessio, N. Díaz-Rodríguez, Plenary: Explaining black-box models in natural language through fuzzy linguistic summaries, Information Sciences 614 (2022) 374–399.

[221] G. Angelotti, N. Díaz-Rodríguez, Towards a more efficient computation of individual attribute and policy contribution for post-hoc explanation of cooperative multi-agent systems using myerson values, Knowledge-Based Systems 260 (2023) 110189.

[222] T. Lesort, N. Díaz-Rodríguez, J.-F. Goudou, D. Filliat, State representation learning for control: An overview, Neural Networks 108 (2018) 379–392.

[223] I. C. Kaadoud, A. Bennetot, B. Mawhin, V. Charisi, N. Díaz-Rodríguez, Explaining aha! moments in artificial agents through ike-xai: Implicit knowledge extraction for explainable ai, Neural Networks 155 (2022) 95–118.

[224] N. Díaz-Rodríguez, R. Binkytė, W. Bakkali, S. Bookseller, P. Tubaro, A. Bacevičius, S. Zhioua, R. Chatila, Gender and sex bias in covid-19 epidemiological data through the lenses of causality, Information Processing & Management (2023) 103276doi:https://doi.org/10.1016/j.ipm.2023.103276.
URL https://www.sciencedirect.com/science/article/pii/S0306457323000134

[225] L. Weng, Attention? Attention! (06 2018).
URL https://lilianweng.github.io/lil-log/

[226] E. Angelino, CORELS: Learning Certifiably Optimal RulE ListS (2018).
URL https://corels.eecs.harvard.edu/corels/install.html

[227] I. Ahern, A. Noack, L. Guzman-Nateras, D. Dou, B. Li, J. Huan, Normlime: A new feature importance metric for explaining deep neural networks, ICLR 2020 Conference (2020).

[228] D. W. Apley, J. Zhu, Visualizing the effects of predictor variables in black box supervised learning models, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 82 (4) (2020) 1059–1086.

[229] V. Petsiuk, A. Das, K. Saenko, Rise: Randomized input sampling for explanation of black-box models, in: Proceedings of the British Machine Vision Conference (BMVC), 2018, pp. 1–13.

[230] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, et al., Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav), in: International conference on machine learning, PMLR, 2018, pp. 2668–2677.

[231] M. T. Ribeiro, S. Singh, C. Guestrin, Anchors: High-precision model-agnostic explanations, in: Proceedings of the AAAI conference on artificial intelligence, 2018, pp. 1–9.

[232] Y. Zhang, B. Wallace, A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification, in: Proceedings of the The 8th International Joint Conference on Natural Language Processing, 2017, pp. 253–263.

[233] D. Smilkov, N. Thorat, B. Kim, F. Viégas, M. Wattenberg, Smoothgrad: removing noise by adding noise, Workshop on Visualization for Deep Learning, ICML (2017).

[234] A. Shrikumar, P. Greenside, A. Kundaje, Learning important features through propagating activation differences, in: International Conference on Machine Learning, PMLR, 2017, pp. 3145–3153.

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

[236] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, W. Samek, On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation, PloS one 10 (7) (2015) e0130140.

[237] A. Goldstein, A. Kapelner, J. Bleich, E. Pitkin, Peeking inside the black box: Visualizing statistical learning with plots of individual

conditional expectation, journal of Computational and Graphical Statistics 24 (1) (2015) 44–65.
[238] M. D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: European conference on computer vision, Springer, 2014, pp. 818–833.
[239] K. Simonyan, A. Vedaldi, A. Zisserman, Deep inside convolutional networks: Visualising image classification models and saliency maps, in: In Workshop at International Conference on Learning Representations, Citeseer, 2014, pp. 1–8.
[240] S. Bazen, X. Joutard, The taylor decomposition: A unified generalization of the oaxaca method to nonlinear models, Archive ouverte en Sciences de l'Homme et de la Société (2013) 101–121.
[241] A. Hyvärinen, J. Hurri, P. O. Hoyer, Independent component analysis, in: Natural Image Statistics, Springer, 2009, pp. 151–175.
[242] A. E. Roth, The Shapley value: essays in honor of Lloyd S. Shapley, Cambridge University Press, 1988.
[243] M. Ancona, E. Ceolini, C. Öztireli, M. Gross, Towards better understanding of gradient-based attribution methods for deep neural networks, ICLR 2018 Conference (2018).
[244] P. Sturmfels, S. Lundberg, S.-I. Lee, Visualizing the impact of feature attribution baselines, Distill 5 (2020) e22.
[245] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, K.-R. Müller, Explaining nonlinear classification decisions with deep taylor decomposition, Pattern Recognition 65 (2017) 211–222.
[246] R. Guidotti, A. Monreale, S. Ruggieri, D. Pedreschi, F. Turini, F. Giannotti, Local rule-based explanations of black box decision systems, arXiv preprint arXiv:1805.10820 (2018).
[247] N. El Bekri, J. Kling, M. F. Huber, A study on trust in black box models and post-hoc explanations, in: International Workshop on Soft Computing Models in Industrial and Environmental Applications, Springer, 2019, pp. 35–46.
[248] J. T. Springenberg, A. Dosovitskiy, T. Brox, M. Riedmiller, Striving for simplicity: The all convolutional net, arXiv preprint arXiv:1412.6806 (2014).
[249] P.-J. Kindermans, S. Hooker, J. Adebayo, M. Alber, K. T. Schütt, S. Dähne, D. Erhan, B. Kim, The (un) reliability of saliency methods, in: Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, Springer, 2019, pp. 267–280.
[250] A. Ghorbani, A. Abid, J. Zou, Interpretation of neural networks is fragile, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, 2019, pp. 3681–3688.
[251] C. M. Bishop, Training with noise is equivalent to tikhonov regularization, Neural computation 7 (1) (1995) 108–116.
[252] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, in: International Conference on Machine Learning, PMLR, 2017, pp. 3319–3328.
[253] L. M. Zintgraf, T. S. Cohen, T. Adel, M. Welling, Visualizing deep neural network decisions: Prediction difference analysis, ICLR Conference (2017).
[254] A. Henelius, K. Puolamäki, H. Boström, L. Asker, P. Papapetrou, A peek into the black box: exploring classifiers by randomization, Data mining and knowledge discovery 28 (5) (2014) 1503–1529.
[255] M. T. Ribeiro, S. Singh, C. Guestrin, Model-agnostic interpretability of machine learning, arXiv preprint arXiv:1606.05386 (2016).
[256] M. D. Zeiler, G. W. Taylor, R. Fergus, Adaptive deconvolutional networks for mid and high level feature learning, in: 2011 International Conference on Computer Vision, IEEE, 2011, pp. 2018–2025.
[257] C. Molnar, Interpretable Machine Learning (09 2021).
URL https://christophm.github.io/interpretable-ml-book/
[258] V. Petsiuk, R. Jain, V. Manjunatha, V. I. Morariu, A. Mehra, V. Ordonez, K. Saenko, Black-box explanation of object detectors via saliency maps, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 11443–11452.
[259] M. Lin, Q. Chen, S. Yan, Network in network, International Conference on Learning Representations (2013).
[260] D. Baehrens, T. Schroeter, S. Harmeling, M. Kawanabe, K. Hansen, K.-R. Müller, How to explain individual classification decisions, The Journal of Machine Learning Research 11 (2010) 1803–1831.
[261] S. Wachter, B. Mittelstadt, C. Russell, Counterfactual explanations without opening the black box: Automated decisions and the gdpr, Harv. JL & Tech. 31 (2017) 841.
[262] B. Kim, C. Rudin, J. A. Shah, The bayesian case model: A generative approach for case-based reasoning and prototype classification, in: Advances in neural information processing systems, 2014, pp. 1952–1960.
[263] H.-S. Park, C.-H. Jun, A simple and fast algorithm for k-medoids clustering, Expert systems with applications 36 (2) (2009) 3336–3341.
[264] N. Roese, Counterfactual thinking, Psychological Bulletin 121 (1997) 133–148. doi:10.1037/0033-2909.121.1.133.
[265] S. Dandl, C. Molnar, M. Binder, B. Bischl, Multi-objective counterfactual explanations, in: International Conference on Parallel Problem Solving from Nature, Springer, 2020, pp. 448–469.
[266] M. Suffian, P. Graziani, J. Alonso-Moral, A. Bogliolo, Fce: Feedback based counterfactual explanations for explainable ai, IEEE Access 10 (2022) 72363–72372. doi:10.1109/ACCESS.2022.3189432.
[267] I. Stepin, A. Catala, M. Pereira-Fariña, J. M. Alonso, Factual and Counterfactual Explanation of Fuzzy Information Granules, Springer-Verlag, 2021, pp. 153–185. doi:10.1007/978-3-030-64949-4_6.
[268] D. Gentner, Structure-mapping: A theoretical framework for analogy, Cognitive science 7 (1983) 155–170. doi:10.1109/ACCESS.2022.3189432.
[269] D. Pham, M. Aksoy, Rules: A simple rule extraction system, Expert Systems with Applications 8 (1) (1995) 59–65.
[270] D. T. Pham, M. Aksoy, An algorithm for automatic rule induction, Artificial Intelligence in Engineering 8 (4) (1993) 277–282.
[271] D. Pham, S. Dimov, An algorithm for incremental inductive learning, Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture 211 (3) (1997) 239–249.
[272] D. Pham, S. Dimov, The rules-3 plus inductive learning algorithm, in: In Proceedings of the Third World Congress on Expert Systems, 1996, pp. 917–924.
[273] D. T. Pham, S. Bigot, S. S. Dimov, Rules-5: a rule induction algorithm for classification problems involving continuous attributes, Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science 217 (12) (2003) 1273–1286.
[274] S. Bigot, A new rule space representation scheme for rule induction in classification and control applications, Proceedings of the Institution of Mechanical Engineers, Part I: Journal of Systems and Control Engineering 225 (7) (2011) 1018–1038.
[275] D. T. Pham, A. Afify, Rules-6: a simple rule induction algorithm for supporting decision making, in: 31st Annual Conference of IEEE Industrial Electronics Society, 2005. IECON 2005., IEEE, 2005, pp. 6–pp.
[276] K. Shehzad, Edisc: a class-tailored discretization technique for rule-based classification, IEEE Transactions on Knowledge and Data Engineering 24 (8) (2011) 1435–1447.
[277] D. Pham, A novel rule induction algorithm with improved handling of continuous valued attributes, Ph.D. thesis, Cardiff University (2012).
[278] D. T. Pham, S. Bigot, S. S. Dimov, Rules-f: A fuzzy inductive learning algorithm, Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science 220 (9) (2006) 1433–1447.
[279] D. Pham, A. Afify, Sri: a scalable rule induction algorithm, Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science 220 (4) (2006) 537–552.
[280] D. T. Pham, A. J. Soroka, An immune-network inspired rule generation algorithm (rules-is), in: Third Virtual International Conference on Innovative Production Machines and Systems, 2007, pp. 1–6.
[281] H. I. Mathkour, Rules3-ext improvements on rules-3 induction algorithm, Mathematical and Computational Applications 15 (3) (2010) 318–324.

[282] H. ElGibreen, M. S. Aksoy, Rules–tl: a simple and improved rules algorithm for incomplete and large data, Journal of Theoretical and Applied Information Technology 47 (1) (2013) 28–40.

[283] H. Elgibreen, M. S. Aksoy, Rules-it: incremental transfer learning with rules family, Frontiers of Computer Science 8 (4) (2014) 537–562.

[284] Ö. Akgöbek, Y. S. Aydin, E. Öztemel, M. S. Aksoy, A new algorithm for automatic knowledge acquisition in inductive learning, Knowledge-Based Systems 19 (6) (2006) 388–395.

[285] D. Dubois, H. Prade, What are fuzzy rules and how to use them, Fuzzy Sets and Systems 84 (1996) 169–185. doi:10.1016/0165-0114(96)00066-8.

[286] G. G. Towell, J. W. Shavlik, Extracting refined rules from knowledge-based neural networks, Machine learning 13 (1) (1993) 71–101.

[287] S. Altug, M. Chow, H. Trussell, Heuristic constraints enforcement for training of and rule extraction from a fuzzy/neural architecture, IEEE Transactions on Fuzzy Systems 7 (1999) 151–159. doi:10.1109/91.755397.

[288] U. Johansson, R. König, L. Niklasson, The truth is in there-rule extraction from opaque models using genetic programming., in: FLAIRS Conference, Miami Beach, FL, 2004, pp. 658–663.

[289] M. H. Aung, P. G. Lisboa, T. A. Etchells, A. C. Testa, B. Van Calster, S. Van Huffel, L. Valentin, D. Timmerman, Comparing analytical decision support models through boolean rule extraction: A case study of ovarian tumour malignancy, in: International Symposium on Neural Networks, Springer, 2007, pp. 1177–1186.

[290] R. Masuoka, N. Watanabe, A. Kawamura, Y. Owada, K. Asakawa, Neurofuzzy system-fuzzy inference using a structured neural network, in: Proceedings of the International Conference on Fuzzy Logic & Neural Networks, 1990, pp. 173–177.

[291] T. GopiKrishna, Evaluation of rule extraction algorithms, International Journal of Data Mining & Knowledge Management Process 4 (3) (2014) 9.

[292] L. Özbakır, A. Baykasoğlu, S. Kulluk, A soft computing-based approach for integrated training and rule extraction from artificial neural networks: Difaconn-miner, Applied Soft Computing 10 (1) (2010) 304–317.

[293] M. Sato, H. Tsukimoto, Rule extraction from neural networks via decision tree induction, in: IJCNN'01. International Joint Conference on Neural Networks. Proceedings (Cat. No. 01CH37222), Vol. 3, IEEE, 2001, pp. 1870–1875.

[294] R. Setiono, W. K. Leow, Fernn: An algorithm for fast extraction of rules from neural networks, Applied Intelligence 12 (1) (2000) 15–25.

[295] L. Fu, Rule generation from neural networks, IEEE Transactions on Systems, Man, and Cybernetics 24 (8) (1994) 1114–1124.

[296] H. Tsukimoto, Extracting rules from trained neural networks, IEEE Transactions on Neural networks 11 (2) (2000) 377–389.

[297] M. W. Craven, J. W. Shavlik, Using sampling and queries to extract rules from trained neural networks, in: Machine learning proceedings 1994, Elsevier, 1994, pp. 37–45.

[298] E. W. Saad, D. C. Wunsch II, Neural network explanation using inversion, Neural networks 20 (1) (2007) 78–93.

[299] I. A. Taha, J. Ghosh, Symbolic interpretation of artificial neural networks, IEEE Transactions on knowledge and data engineering 11 (3) (1999) 448–463.

[300] K. K. Sethi, D. K. Mishra, B. Mishra, Kdruleex: A novel approach for enhancing user comprehensibility using rule extraction, in: 2012 Third International Conference on Intelligent Systems Modelling and Simulation, IEEE, 2012, pp. 55–60.

[301] M. G. Augasta, T. Kathirvalavakumar, Reverse engineering the neural networks for rule extraction in classification problems, Neural processing letters 35 (2) (2012) 131–150.

[302] E. R. Hruschka, N. F. Ebecken, Extracting rules from multilayer perceptrons in classification problems: A clustering-based approach, Neurocomputing 70 (1-3) (2006) 384–397.

[303] H. Kahramanli, N. Allahverdi, Rule extraction from trained adaptive neural networks using artificial immune systems, Expert Systems with Applications 36 (2) (2009) 1513–1522.

[304] J. R. Zilke, E. L. Mencía, F. Janssen, Deepred–rule extraction from deep neural networks, in: International Conference on Discovery Science, Springer, 2016, pp. 457–473.

[305] S. Thrun, Extracting rules from artificial neural networks with distributed representations, Advances in neural information processing systems (1995) 505–512.

[306] M. G. Augasta, T. Kathirvalavakumar, Rule extraction from neural networks—a comparative study, in: International Conference on Pattern Recognition, Informatics and Medical Engineering (PRIME-2012), IEEE, 2012, pp. 404–408.

[307] T. Hailesilassie, Rule extraction algorithm for deep neural networks: A review, arXiv preprint arXiv:1610.05267 (2016).

[308] P. Sadowski, J. Collado, D. Whiteson, P. Baldi, Deep learning, dark knowledge, and dark matter, in: NIPS 2014 Workshop on High-energy Physics and Machine Learning, PMLR, 2015, pp. 81–87.

[309] G. Hinton, O. Vinyals, J. Dean, Distilling the knowledge in a neural network, arXiv preprint arXiv:1503.02531 (2015).

[310] S. Tan, R. Caruana, G. Hooker, Y. Lou, Distill-and-compare: Auditing black-box models using transparent model distillation, Association for Computing Machinery (2018).

[311] Z. Che, S. Purushotham, R. Khemani, Y. Liu, Distilling knowledge from deep networks with applications to healthcare domain, arXiv preprint arXiv:1512.03542 (2015).

[312] K. Xu, D. H. Park, C. Yi, C. Sutton, Interpreting deep classifier by visual distillation of dark knowledge, arXiv preprint arXiv:1803.04042 (2018).

[313] S. Tan, Interpretable approaches to detect bias in black-box models, in: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, 2018, pp. 382–383.

[314] J. Lei, M. G'Sell, A. Rinaldo, R. J. Tibshirani, L. Wasserman, Distribution-free predictive inference for regression, Journal of the American Statistical Association 113 (523) (2018) 1094–1111.

[315] A. Fisher, C. Rudin, F. Dominici, Model class reliance: Variable importance measures for any machine learning model class, from the rashomon, Perspective 68 (2018).

[316] S. Lapuschkin, A. Binder, G. Montavon, K.-R. Muller, W. Samek, Analyzing classifiers: Fisher vectors and deep neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2912–2920.

[317] W. Samek, A. Binder, G. Montavon, S. Lapuschkin, K.-R. Müller, Evaluating the visualization of what a deep neural network has learned, IEEE transactions on neural networks and learning systems 28 (11) (2016) 2660–2673.

[318] J. M. Zurada, A. Malinowski, I. Cloete, Sensitivity analysis for minimization of input data dimension for feedforward neural network, in: Proceedings of IEEE International Symposium on Circuits and Systems-ISCAS'94, Vol. 6, IEEE, 1994, pp. 447–450.

[319] A. Sung, Ranking importance of input parameters of neural networks, Expert systems with Applications 15 (3-4) (1998) 405–411.

[320] J. Khan, J. S. Wei, M. Ringner, L. H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. R. Antonescu, C. Peterson, et al., Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks, Nature medicine 7 (6) (2001) 673–679.

[321] M. Gevrey, I. Dimopoulos, S. Lek, Review and comparison of methods to study the contribution of variables in artificial neural network models, Ecological modelling 160 (3) (2003) 249–264.

[322] P. Cortez, M. J. Embrechts, Using sensitivity analysis and visualization techniques to open black box data mining models, Information Sciences 225 (2013) 1–17.

[323] P. Cortez, M. J. Embrechts, Opening black box data mining models using sensitivity analysis, in: 2011 IEEE Symposium on Computational Intelligence and Data Mining (CIDM), IEEE, 2011, pp. 341–348.

[324] A. Ghorbani, J. Wexler, J. Zou, B. Kim, Towards automatic concept-based explanations, 33rd Conference on Neural Information Processing Systems (NeurIPS 2019) (2019).

[325] Y. Goyal, A. Feder, U. Shalit, B. Kim, Explaining classifiers with causal concept effect (cace), arXiv preprint arXiv:1907.07165 (2019).

[326] C.-K. Yeh, B. Kim, S. Arik, C.-L. Li, P. Ravikumar, T. Pfister, On concept-based explanations in deep neural networks, ICLR 2020 Conference (2019) 1–17.

[327] F. Vitali, A survey on methods and metrics for the assessment of explainability under the proposed AI Act, in: Legal Knowledge and Information Systems: JURIX 2021: The Thirty-fourth Annual Conference, Vilnius, Lithuania, 8-10 December 2021, Vol. 346, IOS Press, 2022, p. 235.

[328] M. Robnik-Šikonja, M. Bohanec, Perturbation-based explanations of prediction models, in: Human and machine learning, Springer, 2018, pp. 159–175.

[329] E. Lughofer, R. Richter, U. Neissl, W. Heidl, C. Eitzinger, T. Radauer, Explaining classifier decisions linguistically for stimulating and improving operators labeling behavior, Information Sciences 420 (2017) 16–36.

[330] H. Jacobsson, Rule extraction from recurrent neural networks: Ataxonomy and review, Neural Computation 17 (2005) 1223–1263.

[331] A. Holzinger, A. Carrington, H. Müller, Measuring the quality of explanations: the system causability scale (scs) comparing human and machine explanations, KI-Künstliche Intelligenz 34 (2) (2020) 193–198.

[332] W. Verbeke, D. Martens, C. Mues, B. Baesens, Building comprehensible customer churn prediction models with advanced rule induction techniques, Expert systems with applications 38 (2011) 2354–2364.

[333] D. Martens, J. Vanthienen, W. Verbeke, B. Baesens, Performance of classification models from a user perspective, Decision Support Systems 51 (2011) 782–793.

[334] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, S.-I. Lee, From local explanations to global understanding with explainable ai for trees, Nature machine intelligence 2 (1) (2020) 56–67.

[335] K. Amarasinghe, K. Rodolfa, H. Lamba, R. Ghani, Explainable machine learning for public policy: Use cases, gaps, and research directions, arXiv preprint arXiv:2010.14374 (2020).

[336] E. Costanza, J. E. Fischer, J. A. Colley, T. Rodden, S. D. Ramchurn, N. R. Jennings, Doing the laundry with agents: a field trial of a future smart energy system in the home, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2014, pp. 813–822.

[337] M. Kay, T. Kola, J. R. Hullman, S. A. Munson, When (ish) is my bus? user-centered visualizations of uncertainty in everyday, mobile predictive systems, in: Proceedings of the 2016 chi conference on human factors in computing systems, 2016, pp. 5092–5103.

[338] B. Y. Lim, A. K. Dey, Assessing demand for intelligibility in context-aware applications, in: Proceedings of the 11th international conference on Ubiquitous computing, 2009, pp. 195–204.

[339] F. C. Keil, Explanation and understanding, Annu. Rev. Psychol. 57 (2006) 227–254.

[340] J. Dodge, S. Penney, A. Anderson, M. M. Burnett, What should be in an XAI explanation? what ift reveals., in: IUI Workshops, 2018, pp. 1–4.

[341] S. Penney, J. Dodge, C. Hilderbrand, A. Anderson, L. Simpson, M. Burnett, Toward foraging for understanding of starcraft agents: An empirical study, in: 23rd International Conference on Intelligent User Interfaces, 2018, pp. 225–237.

[342] E. Rader, R. Gray, Understanding user beliefs about algorithmic curation in the facebook news feed, in: Proceedings of the 33rd annual ACM conference on human factors in computing systems, 2015, pp. 173–182.

[343] S. Stumpf, S. Skrebe, G. Aymer, J. Hobson, Explaining smart heating systems to discourage fiddling with optimized behavior, in: CEUR Workshop Proceedings, Vol. 2068, 2018, pp. 1–5.

[344] R. R. Hoffman, Theory→ concepts→ measures but policies→ metrics, in: Macrocognition Metrics and Scenarios, CRC Press, 2018, pp. 3–10.

[345] F. Gedikli, D. Jannach, M. Ge, How should i explain? a comparison of different explanation types for recommender systems, International Journal of Human-Computer Studies 72 (4) (2014) 367–382.

[346] B. Y. Lim, A. K. Dey, D. Avrahami, Why and why not explanations improve the intelligibility of context-aware intelligent systems, in: Proceedings of the SIGCHI conference on human factors in computing systems, 2009, pp. 2119–2128.

[347] W. Curran, T. Moore, T. Kulesza, W.-K. Wong, S. Todorovic, S. Stumpf, R. White, M. Burnett, Towards recognizing" cool" can end users help computer vision recognize subjective attributes of objects in images?, in: Proceedings of the 2012 ACM international conference on Intelligent User Interfaces, 2012, pp. 285–288.

[348] I. Lage, E. Chen, J. He, M. Narayanan, B. Kim, S. J. Gershman, F. Doshi-Velez, Human evaluation of models built for interpretability, in: Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, Vol. 7, 2019, pp. 59–67.

[349] B. Nushi, E. Kamar, E. Horvitz, Towards accountable ai: Hybrid human-machine analyses for characterizing system failure, in: Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, Vol. 6, 2018, pp. 126–135.

[350] G. Bansal, B. Nushi, E. Kamar, W. S. Lasecki, D. S. Weld, E. Horvitz, Beyond accuracy: The role of mental models in human-ai team performance, in: Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, Vol. 7, 2019, pp. 2–11.

[351] R. Binns, M. Van Kleek, M. Veale, U. Lyngs, J. Zhao, N. Shadbolt, 'it's reducing a human being to a percentage' perceptions of justice in algorithmic decisions, in: Proceedings of the 2018 Chi conference on human factors in computing systems, 2018, pp. 1–14.

[352] T. Kulesza, S. Stumpf, M. Burnett, S. Yang, I. Kwan, W.-K. Wong, Too much, too little, or just right? ways explanations impact end users' mental models, in: 2013 IEEE Symposium on visual languages and human centric computing, IEEE, 2013, pp. 3–10.

[353] T. Lombrozo, Explanation and categorization: How "why?" informs "what?", Cognition 110 (2009) 248–253.

[354] S. Coppers, J. Van den Bergh, K. Luyten, K. Coninx, I. Van der Lek-Ciudin, T. Vanallemeersch, V. Vandeghinste, Intellingo: an intelligible translation environment, in: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, 2018, pp. 1–13.

[355] A. Bunt, M. Lount, C. Lauzon, Are explanations always important? a study of deployed, low-cost intelligent interactive systems, in: Proceedings of the 2012 ACM international conference on Intelligent User Interfaces, 2012, pp. 169–178.

[356] M. Kahng, P. Y. Andrews, A. Kalro, D. H. Chau, A cti v is: Visual exploration of industry-scale deep neural network models, IEEE transactions on visualization and computer graphics 24 (2017) 88–97.

[357] J. Krause, A. Perer, E. Bertini, Infuse: interactive feature selection for predictive modeling of high dimensional data, IEEE transactions on visualization and computer graphics 20 (2014) 1614–1623.

[358] M. Liu, S. Liu, X. Zhu, Q. Liao, F. Wei, S. Pan, An uncertainty-aware approach for exploratory microblog retrieval, IEEE transactions on visualization and computer graphics 22 (2015) 250–259.

[359] M. Liu, J. Shi, Z. Li, C. Li, J. Zhu, S. Liu, Towards better analysis of deep convolutional neural networks, IEEE transactions on visualization and computer graphics 23 (1) (2016) 91–100.

[360] H. Strobelt, S. Gehrmann, H. Pfister, A. M. Rush, Lstmvis: A tool for visual analysis of hidden state dynamics in recurrent neural networks, IEEE transactions on visualization and computer graphics 24 (2017) 667–676.

[361] M. Nourani, S. Kabir, S. Mohseni, E. D. Ragan, The effects of meaningful and meaningless explanations on trust and perceived system accuracy in intelligent systems, in: Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, Vol. 7,

2019, pp. 97–105.
[362] S. Berkovsky, R. Taib, D. Conway, How to recommend? user trust factors in movie recommender systems, in: Proceedings of the 22nd International Conference on Intelligent User Interfaces, 2017, pp. 287–300.
[363] A. Bussone, S. Stumpf, D. O'Sullivan, The role of explanations on trust and reliance in clinical decision support systems, in: 2015 international conference on healthcare informatics, IEEE, 2015, pp. 160–169.
[364] B. Cahour, J.-F. Forzy, Does projection into use improve trust and exploration? an example with a cruise control system, Safety science 47 (2009) 1260–1270.
[365] M. Eiband, D. Buschek, A. Kremer, H. Hussmann, The impact of placebic explanations on trust in intelligent systems, in: Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems, 2019, pp. 1–6.
[366] F. Nothdurft, F. Richter, W. Minker, Probabilistic human-computer trust handling, in: Proceedings of the 15th annual meeting of the special interest group on discourse and dialogue (SIGDIAL), 2014, pp. 51–59.
[367] P. Pu, L. Chen, Trust building with explanation interfaces, in: Proceedings of the 11th international conference on Intelligent user interfaces, 2006, pp. 93–100.
[368] M. Yin, J. Wortman Vaughan, H. Wallach, Understanding the effect of accuracy on trust in machine learning models, in: Proceedings of the 2019 chi conference on human factors in computing systems, 2019, pp. 1–12.
[369] T. Kulesza, M. Burnett, W.-K. Wong, S. Stumpf, Principles of explanatory debugging to personalize interactive machine learning, in: Proceedings of the 20th international conference on intelligent user interfaces, 2015, pp. 126–137.
[370] X. Wang, S. Liu, J. Liu, J. Chen, J. Zhu, B. Guo, Topicpanorama: A full picture of relevant topics, IEEE transactions on visualization and computer graphics 22 (2016) 2508–2521.
[371] N. Pezzotti, T. Höllt, J. Van Gemert, B. P. Lelieveldt, E. Eisemann, A. Vilanova, Deepeyes: Progressive visual analytics for designing deep neural networks, IEEE transactions on visualization and computer graphics 24 (2017) 98–108.
[372] T. Kulesza, S. Stumpf, M. Burnett, W.-K. Wong, Y. Riche, T. Moore, I. Oberst, A. Shinsel, K. McIntosh, Explanatory debugging: Supporting end-user debugging of machine-learned programs, in: 2010 IEEE Symposium on Visual Languages and Human-Centric Computing, IEEE, 2010, pp. 41–48.
[373] A. Groce, T. Kulesza, C. Zhang, S. Shamasunder, M. Burnett, W.-K. Wong, S. Stumpf, S. Das, A. Shinsel, F. Bice, et al., You are the only possible oracle: Effective test selection for end users of interactive machine learning systems, IEEE Transactions on Software Engineering 40 (2013) 307–323.
[374] J. Krause, A. Dasgupta, J. Swartz, Y. Aphinyanaphongs, E. Bertini, A workflow for visual diagnostics of binary classifiers using instance-level explanations, in: 2017 IEEE Conference on Visual Analytics Science and Technology (VAST), IEEE, 2017, pp. 162–172.
[375] C. Olah, A. Satyanarayan, I. Johnson, S. Carter, L. Schubert, K. Ye, A. Mordvintsev, The building blocks of interpretability, Distill 3 (2018) e10.
[376] A. S. Ross, F. Doshi-Velez, Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients, in: Thirty-second AAAI conference on artificial intelligence, 2018, pp. 1–10.
[377] J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, H. Lipson, Understanding neural networks through deep visualization, In ICML Deep Learning Workshop (2015).
[378] T. Zahavy, N. Ben-Zrihem, S. Mannor, Graying the black box: Understanding dqns, in: International Conference on Machine Learning, PMLR, 2016, pp. 1899–1908.
[379] A. Das, H. Agrawal, L. Zitnick, D. Parikh, D. Batra, Human attention in visual question answering: Do humans and deep networks look at the same regions?, Computer Vision and Image Understanding 163 (2017) 90–100.
[380] S. Mohseni, J. E. Block, E. D. Ragan, A human-grounded evaluation benchmark for local explanations of machine learning, arXiv preprint arXiv:1801.05075 (2018).
[381] P. Schmidt, F. Biessmann, Quantifying interpretability and trust in machine learning systems, arXiv preprint arXiv:1901.08558 (2019).
[382] D. Meyerson, K. E. Weick, R. M. Kramer, Swift trust and temporary group. trust in organisations, Frontiers of theory and research 166 (1996) 195.
[383] S. M. Merritt, H. Heimbaugh, J. LaChapell, D. Lee, I trust it, but i don't know why: Effects of implicit attitudes toward automation on trust in an automated system, Human factors 55 (2013) 520–534.
[384] P. Bobko, A. J. Barelka, L. M. Hirshfield, The construct of state-level suspicion: A model and research agenda for automated and information technology (it) contexts, Human Factors 56 (2014) 489–508.
[385] M. Madsen, S. Gregor, Measuring human-computer trust, in: 11th australasian conference on information systems, Vol. 53, Citeseer, 2000, pp. 6–8.
[386] J.-Y. Jian, A. M. Bisantz, C. G. Drury, Foundations for an empirically determined scale of trust in automated systems, International journal of cognitive ergonomics 4 (2000) 53–71.
[387] I. Stepin, J. M. Alonso-Moral, A. Catala, M. Pereira-Fariña, An empirical study on how humans appreciate automated counterfactual explanations which embrace imprecise information, Information Sciences 618 (2022) 379–399. doi:10.1016/j.ins.2022.10.098.
[388] B. A. Myers, D. A. Weitzman, A. J. Ko, D. H. Chau, Answering why and why not questions in user interfaces, in: Proceedings of the SIGCHI conference on Human Factors in computing systems, 2006, pp. 397–406.
[389] Y. Ahn, Y.-R. Lin, Fairsight: Visual analytics for fairness in decision making, IEEE transactions on visualization and computer graphics 26 (2019) 1086–1095.
[390] Y. Ming, S. Cao, R. Zhang, Z. Li, Y. Chen, Y. Song, H. Qu, Understanding hidden memories of recurrent neural networks, in: 2017 IEEE Conference on Visual Analytics Science and Technology (VAST), IEEE, 2017, pp. 13–24.
[391] M. Liu, J. Shi, K. Cao, J. Zhu, S. Liu, Analyzing the training processes of deep generative models, IEEE transactions on visualization and computer graphics 24 (2017) 77–87.
[392] H. Nori, S. Jenkins, P. Koch, R. Caruana, Interpretml: A unified framework for machine learning interpretability, arXiv preprint arXiv:1909.09223 (2019).
[393] J. Klaise, A. Van Looveren, G. Vacanti, A. Coca, Alibi explain: Algorithms for explaining machine learning models, Journal of Machine Learning Research 22 (181) (2021) 1–7.
[394] V. Arya, R. K. Bellamy, P.-Y. Chen, A. Dhurandhar, M. Hind, S. C. Hoffman, S. Houde, Q. V. Liao, R. Luss, A. Mojsilovic, et al., Ai explainability 360: An extensible toolkit for understanding data and machine learning models., J. Mach. Learn. Res. 21 (130) (2020) 1–6.
[395] oracle, GitHub - oracle/Skater: Python Library for Model Interpretation/Explanations (2018).
URL https://github.com/oracle/Skater
[396] S. Sicara, GitHub - sicara/tf-explain: Interpretability Methods for tf.keras models with Tensorflow 2.x (2019).
URL https://github.com/sicara/tf-explain
[397] C. Molnar, G. Casalicchio, B. Bischl, iml: An r package for interpretable machine learning, Journal of Open Source Software 3 (26) (2018) 786.
[398] P. Biecek, Dalex: explainers for complex predictive models in r, The Journal of Machine Learning Research 19 (1) (2018) 3245–3249.
[399] H. H2O, GitHub - h2oai/mli-resources: H2O.ai Machine Learning Interpretability Resources (2019).
URL https://github.com/h2oai/mli-resources
[400] T.-M. ELI5, GitHub - TeamHG-Memex/eli5: A library for debugging/inspecting machine learning classifiers and explaining their predictions (2019).

URL https://github.com/TeamHG-Memex/eli5

[401] M. Alber, S. Lapuschkin, P. Seegerer, M. Hägele, K. T. Schütt, G. Montavon, W. Samek, K.-R. Müller, S. Dähne, P.-J. Kindermans, innvestigate neural networks!, J. Mach. Learn. Res. 20 (93) (2019) 1–8.

[402] P.-J. Kindermans, K. T. Schütt, M. Alber, K.-R. Müller, D. Erhan, B. Kim, S. Dähne, Learning how to explain neural networks: Patternnet and patternattribution, 6th International Conference on Learning Representations, ICLR 2018 (2018).

[403] H. Baniecki, P. Biecek, modelstudio: Interactive studio with explanations for ml predictive models, Journal of Open Source Software 4 (43) (2019) 1798.

[404] P. Biecek, Ceterisparibus: Ceteris paribus profiles (2019).

[405] N. Kokhlikyan, V. Miglani, M. Martin, E. Wang, B. Alsallakh, J. Reynolds, A. Melnikov, N. Kliushkina, C. Araya, S. Yan, et al., Captum: A unified and generic model interpretability library for pytorch, ICLR 2021 workshop on Responsible AI: (2021).

[406] C.-K. Yeh, C.-Y. Hsieh, A. Suggala, D. I. Inouye, P. K. Ravikumar, On the (in) fidelity and sensitivity of explanations, Advances in Neural Information Processing Systems 32 (2019) 10967–10978.

[407] W. Yang, H. Le, S. Savarese, S. Hoi, Omnixai: A library for explainable ai (2022). arXiv:206.01612, doi:10.48550/ARXIV.2206.01612.
URL https://arxiv.org/abs/2206.01612

[408] P.-T. Jiang, C.-B. Zhang, Q. Hou, M.-M. Cheng, Y. Wei, Layercam: Exploring hierarchical class activation maps for localization, IEEE Transactions on Image Processing 30 (2021) 5875–5888.

[409] X. Situ, I. Zukerman, C. Paris, S. Maruf, G. Haffari, Learning to explain: Generating stable explanations fast, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021, pp. 5340–5355.

[410] MAIF, Shapash Python library for interpretable and understandable machine learning, https://github.com/MAIF/shapash, [Online; accessed 17-January-2023] (2023).

[411] E. EthicalML, GitHub - EthicalML/XAI: XAI - An eXplainability toolbox for machine learning (2018).
URL https://github.com/EthicalML/{XAI}

[412] J. Wexler, M. Pushkarna, T. Bolukbasi, M. Wattenberg, F. Viégas, J. Wilson, The what-if tool: Interactive probing of machine learning models, IEEE transactions on visualization and computer graphics 26 (1) (2019) 56–65.

[413] D.-A. thomas, GitHub - deel-ai/xplique: Xplique is a Neural Networks Explainability Toolbox (2019).
URL https://github.com/deel-ai/xplique

[414] P. Piatyszet, P. Biecek, Arena for the Exploration and Comparison of any ML Models (2020).
URL https://arenar.drwhy.ai/

[415] J. Wiśniewski, P. Biecek, fairmodels: A flexible tool for bias detection, visualization, and mitigation, arXiv preprint arXiv:2104.00507 (2021).

[416] M. Pekala, P. Biecek, GitHub - ModelOriented/triplot: Triplot: Instance- and data-level explanations for the groups of correlated features. (2020).
URL https://github.com/ModelOriented/triplot

[417] R. Adam, M. Polakowski, P. Biecek, Creates Shiny Application From A DALEX Explainer (2021).
URL https://modeloriented.github.io/{XAI}2shiny/

[418] A. Gosiewska, P. Biecek, auditor: an r package for model-agnostic visual validation and diagnostics, The R Journal 11 (2019) 85–98.

[419] M. Mayer, GitHub - mayer79/flashlight: Machine learning explanations (2020).
URL https://github.com/mayer79/flashlight

[420] S. Maksymiuk, A. Gosiewska, P. Biecek, Landscape of r packages for explainable artificial intelligence, arXiv preprint arXiv:2009.13248 (2021).

[421] A. Hedström, L. Weber, D. Bareeva, F. Motzkus, W. Samek, S. Lapuschkin, M. M.-C. Höhne, Quantus: an explainable ai toolkit for responsible evaluation of neural network explanations, arXiv preprint arXiv:2202.06861 (2022).

[422] K. Sokol, R. Santos-Rodriguez, P. Flach, Fat forensics: a python toolbox for algorithmic fairness, accountability and transparency, arXiv preprint arXiv:1909.05167 (2022).

[423] TensorFlow, Github - tensorflow/model-analysis: Model analysis tools for TensorFlow (may 19 2022).

[424] A. Rochford, Github - AustinRochford/PyCEbox: Python Individual Conditional Expectation Plot Toolbox (jan 25 2018).

[425] B. Bengfort, R. Bilbro, Yellowbrick: Visualizing the Scikit-Learn Model Selection Process, The Journal of Open Source Software 4 (35) (2019). doi:10.21105/joss.01075.
URL http://joss.theoj.org/papers/10.21105/joss.01075

[426] P. Adler, C. Falk, S. A. Friedler, T. Nix, G. Rybeck, C. Scheidegger, B. Smith, S. Venkatasubramanian, Auditing black-box models for indirect influence, Knowledge and Information Systems 54 (1) (2018) 95–122.

[427] S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton, D. Roth, A comparative study of fairness-enhancing interventions in machine learning, in: Proceedings of the conference on fairness, accountability, and transparency, 2019, pp. 329–338.

[428] F. Tramer, V. Atlidakis, R. Geambasu, D. Hsu, J.-P. Hubaux, M. Humbert, A. Juels, H. Lin, Fairtest: Discovering unwarranted associations in data-driven applications, in: 2017 IEEE European Symposium on Security and Privacy (EuroS&P), IEEE, 2017, pp. 401–416.

[429] Adebayoj, Github - adebayoj/fairml (mar 23 2017).

[430] A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, H. Wallach, A reductions approach to fair classification, in: International Conference on Machine Learning, PMLR, 2018, pp. 60–69.

[431] TensorFlow, Github - tensorflow/privacy: Library for training machine learning models with privacy for training data (feb 23 2022).

[432] M. Wu, M. Wicker, W. Ruan, X. Huang, M. Kwiatkowska, A game-based approximate verification of deep neural networks with provable guarantees, Theoretical Computer Science 807 (2020) 298–329.

[433] OpenMined, Github - OpenMined/PyGrid: A Peer-to-peer Platform for Secure, Privacy-preserving, Decentralized Data Science (jun 21 2021).

[434] Trusted-AI, Github - Trusted-AI/adversarial-robustness-toolbox: Adversarial Robustness Toolbox (ART) - Python Library for Machine Learning Security - Evasion, Poisoning, Extraction, Inference - Red and Blue Teams, [Online; accessed 2022-07-19] (jul 7 2022).

[435] J. Rauber, R. Zimmermann, M. Bethge, W. Brendel, Foolbox native: Fast adversarial attacks to benchmark the robustness of machine learning models in pytorch, tensorflow, and jax, Journal of Open Source Software 5 (53) (2020) 2607.

[436] N. Papernot, F. Faghri, N. Carlini, I. Goodfellow, R. Feinman, A. Kurakin, C. Xie, Y. Sharma, T. Brown, A. Roy, et al., Technical report on the cleverhans v2. 1.0 adversarial examples library. arxiv 2018, arXiv preprint arXiv:1610.00768 (2020).

[437] B. Goodman, S. Flaxman, European union regulations on algorithmic decision-making and a "right to explanation", AI magazine 38 (3) (2017) 50–57.

[438] R. R. Hoffman, S. T. Mueller, G. Klein, J. Litman, Metrics for explainable ai: Challenges and prospects, arXiv preprint arXiv:1812.04608 (2018).

[439] A. Holzinger, The next frontier: Ai we can really trust, in: Machine Learning and Principles and Practice of Knowledge Discovery in Databases: International Workshops of ECML PKDD 2021, Virtual Event, September 13-17, 2021, Proceedings, Part I, Springer, 2022, pp. 427–440.

[440] S. Amershi, M. Cakmak, W. B. Knox, T. Kulesza, Power to the people: The role of humans in interactive machine learning, Ai Magazine 35 (2014) 105–120.

[441] C. J. Cai, J. Jongejan, J. Holbrook, The effects of example-based explanations in a machine learning interface, in: Proceedings of the 24th international conference on intelligent user interfaces, 2019, pp. 258–262.

[442] R. Kocielnik, S. Amershi, P. N. Bennett, Will you accept an imperfect ai? exploring designs for adjusting end-user expectations of ai systems, in: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 2019, pp. 1–14.

[443] D. Wang, Q. Yang, A. Abdul, B. Y. Lim, Designing theory-driven user-centric explainable ai, in: Proceedings of the 2019 CHI conference on human factors in computing systems, 2019, pp. 1–15.

[444] F. Hohman, H. Park, C. Robinson, D. H. P. Chau, S ummit: Scaling deep learning interpretability by visualizing activation and attribution summarizations, IEEE transactions on visualization and computer graphics 26 (2019) 1096–1106.

[445] Á. A. Cabrera, W. Epperson, F. Hohman, M. Kahng, J. Morgenstern, D. H. Chau, Fairvis: Visual analytics for discovering intersectional bias in machine learning, in: 2019 IEEE Conference on Visual Analytics Science and Technology (VAST), IEEE, 2019, pp. 46–56.

[446] D. Holliday, S. Wilson, S. Stumpf, User trust in intelligent systems: A journey over time, in: Proceedings of the 21st international conference on intelligent user interfaces, 2016, pp. 164–168.

[447] J. K. Doyle, M. J. Radzicki, W. S. Trees, Measuring change in mental models of complex dynamic systems, in: Complex decision making, Springer, 2008, pp. 269–294.

[448] S. Amershi, D. Weld, M. Vorvoreanu, A. Fourney, B. Nushi, P. Collisson, J. Suh, S. Iqbal, P. N. Bennett, K. Inkpen, et al., Guidelines for human-ai interaction, in: Proceedings of the 2019 chi conference on human factors in computing systems, 2019, pp. 1–13.

[449] J. Chen, J. Vaughan, V. N. Nair, A. Sudjianto, Adaptive explainable neural networks (axnns), arXiv preprint arXiv:2004.02353 (2020).

[450] R. M. French, Catastrophic forgetting in connectionist networks, Trends in cognitive sciences 3 (4) (1999) 128–135.

[451] T. Lesort, V. Lomonaco, A. Stoian, D. Maltoni, D. Filliat, N. Díaz-Rodríguez, Continual learning for robotics: Definition, framework, learning strategies, opportunities and challenges, Information Fusion (2020).

[452] R. Traoré, H. Caselles-Dupré, T. Lesort, T. Sun, G. Cai, N. Díaz-Rodríguez, D. Filliat, Discorl: Continual reinforcement learning via policy distillation, in: ICML 2019 Workshop on Multi-task Reinforcement Learning. arXiv preprint arXiv:1907.05855, 2019.

[453] A. Raffin, A. Hill, R. Traoré, T. Lesort, N. Díaz-Rodríguez, D. Filliat, S-rl toolbox: Environments, datasets and evaluation metrics for state representation learning, in: NeurIPS workshop on Deep Reinforcement Learning, http://arxiv.org/abs/1809.09369, 2018.

[454] S. Ede, S. Baghdadlian, L. Weber, W. Samek, S. Lapuschkin, Explain to not forget: Defending against catastrophic forgetting with XAI, arXiv preprint arXiv:2205.01929 (2022).

[455] D. Slack, A. Hilgard, S. Singh, H. Lakkaraju, Reliable post hoc explanations: Modeling uncertainty in explainability, Advances in Neural Information Processing Systems 34 (2021) 9391–9404.

[456] L. Weber, S. Lapuschkin, A. Binder, W. Samek, Beyond explaining: Opportunities and challenges of XAI-based model improvement, arXiv preprint arXiv:2203.08008 (2022).

[457] S. Kwon, Y. Lee, Explainability-based mix-up approach for text data augmentation, ACM Transactions on Knowledge Discovery from Data (TKDD) (2022).

[458] S. Teso, K. Kersting, Explanatory interactive machine learning, in: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, 2019, pp. 239–245.

[459] P. Schramowski, W. Stammer, S. Teso, A. Brugger, F. Herbert, X. Shao, H.-G. Luigs, A.-K. Mahlein, K. Kersting, Making deep neural networks right for the right scientific reasons by interacting with their explanations, Nature Machine Intelligence 2 (8) (2020) 476–486.

[460] A. Martinez-Seras, J. Del Ser, P. Garcia-Bringas, Can post-hoc explanations effectively detect out-of-distribution samples?, in: 2022 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), IEEE, 2022, pp. 1–9.

[461] D. Marcos, J. Kierdorf, T. Cheeseman, D. Tuia, R. Roscher, A whale's tail-finding the right whale in an uncertain world, in: International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers, Springer, 2022, pp. 297–313.

[462] S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, K.-R. Müller, Unmasking clever hans predictors and assessing what machines really learn, Nature communications 10 (1) (2019) 1–8.

[463] D. Becking, M. Dreyer, W. Samek, K. Müller, S. Lapuschkin, Ecq: Explainability-driven quantization for low-bit and sparse dnns, in: International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers, Springer, 2022, pp. 271–296.

[464] S.-K. Yeom, P. Seegerer, S. Lapuschkin, A. Binder, S. Wiedemann, K.-R. Müller, W. Samek, Pruning by explaining: A novel criterion for deep neural network pruning, Pattern Recognition 115 (2021) 107899.

[465] C. J. Anders, D. Neumann, T. Marinc, W. Samek, K.-R. Müller, S. Lapuschkin, Xai for analyzing and unlearning spurious correlations in imagenet, in: ICML'20 Workshop on Extending Explainable AI Beyond Deep Models and Classifiers (XXAI), Vienna, Austria, 2020.

[466] L. Bourtoule, V. Chandrasekaran, C. A. Choquette-Choo, H. Jia, A. Travers, B. Zhang, D. Lie, N. Papernot, Machine unlearning, in: 2021 IEEE Symposium on Security and Privacy (SP), IEEE, 2021, pp. 141–159.

[467] A.-H. Karimi, J. von Kügelgen, B. Schölkopf, I. Valera, Towards causal algorithmic recourse, in: International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers, Springer, 2022, pp. 139–166.

[468] S. A. Bargal, A. Zunino, V. Petsiuk, J. Zhang, V. Murino, S. Sclaroff, K. Saenko, Beyond the visual analysis of deep model saliency, in: International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers, Springer, 2022, pp. 255–269.

[469] L. Salewski, A. Koepke, H. Lensch, Z. Akata, Clevr-x: A visual reasoning dataset for natural language explanations, in: International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers, Springer, 2022, pp. 69–88.

[470] O. Bastani, J. P. Inala, A. Solar-Lezama, Interpretable, verifiable, and robust reinforcement learning via program synthesis, in: International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers, Springer, 2022, pp. 207–228.

[471] C. Singh, W. Ha, B. Yu, Interpreting and improving deep-learning models with reality checks, in: International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers, Springer, 2022, pp. 229–254.

[472] C. Molnar, G. König, J. Herbinger, T. Freiesleben, S. Dandl, C. A. Scholbeck, G. Casalicchio, M. Grosse-Wentrup, B. Bischl, General pitfalls of model-agnostic interpretation methods for machine learning models, in: International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers, Springer, 2022, pp. 39–68.

[473] R. K. Singh, R. Gorantla, S. G. R. Allada, P. Narra, Skinet: A deep learning framework for skin lesion diagnosis with uncertainty estimation and explainability, Plos one 17 (10) (2022) e0276836.

[474] E. Commission, Proposal for a regulation of the european parliament and the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts, EUR-Lex-52021PC0206 (2021).
URL https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206

[475] M. Brundage, S. Avin, J. Wang, H. Belfield, G. Krueger, G. Hadfield, H. Khlaaf, J. Yang, H. Toner, R. Fong, et al., Toward trustworthy ai development: mechanisms for supporting verifiable claims, arXiv preprint arXiv:2004.07213 (2020).

[476] J. Schulman, B. Zoph, C. Kim, J. Hilton, J. Menick, J. Weng, J. Uribe, L. Fedus, L. Metz, M. Pokorny, et al., Chatgpt: Optimizing language models for dialogue (2022).

[477] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, M. Chen, Hierarchical text-conditional image generation with clip latents, arXiv preprint

arXiv:2204.06125 (2022).

[478] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-resolution image synthesis with latent diffusion models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10684–10695.

[479] L. A. Gatys, A. S. Ecker, M. Bethge, Image style transfer using convolutional neural networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2414–2423.

[480] J.-Y. Zhu, T. Park, P. Isola, A. A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: Computer Vision (ICCV), 2017 IEEE International Conference on, 2017.

[481] U. Gadiraju, J. Yang, What can crowd computing do for the next generation of ai systems?, in: CSW@ NeurIPS, 2020, pp. 7–13.

[482] D. S. Char, M. D. Abràmoff, C. Feudtner, Identifying ethical considerations for machine learning healthcare applications, The American Journal of Bioethics 20 (11) (2020) 7–17.

[483] R. Chatila, J. C. Havens, The ieee global initiative on ethics of autonomous and intelligent systems, Robotics and well-being (2019) 11–16.

[484] N. A. Smuha, The eu approach to ethics guidelines for trustworthy artificial intelligence, Computer Law Review International 20 (4) (2019) 97–106.

[485] R. Calegari, G. Ciatto, A. Omicini, On the integration of symbolic and sub-symbolic techniques for XAI: A survey, Intelligenza Artificiale 14 (1) (2020) 7–32.

[486] C. Núñez-Molina, Application of neurosymbolic ai to sequential decision making, in: L. D. Raedt (Ed.), Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22, International Joint Conferences on Artificial Intelligence Organization, 2022, pp. 5863–5864, doctoral Consortium. doi:10.24963/ijcai.2022/834.
URL https://doi.org/10.24963/ijcai.2022/834

[487] C. Núñez-Molina, J. Fernández-Olivares, R. Pérez, Learning to select goals in automated planning with deep-q learning, Expert Systems with Applications 202 (2022) 117265.

[488] R. Manhaeve, S. Dumancic, A. Kimmig, T. Demeester, L. De Raedt, Deepproblog: Neural probabilistic logic programming, Advances in Neural Information Processing Systems 31 (2018).

[489] R. Brandão, J. Carbonera, C. De Souza, J. Ferreira, B. Gonçalves, C. Leitão, Mediation challenges and socio-technical gaps for explainable deep learning applications, arXiv preprint arXiv:1907.07178 (2019).

[490] F. Gualdi, A. Cordella, Artificial intelligence and decision-making: The question of accountability, in: Proceedings of the 54th Hawaii International Conference on System Sciences, 2021, p. 2297.

[491] S. Liu, B. Zhao, R. Guo, G. Meng, F. Zhang, M. Zhang, Have you been properly notified? automatic compliance analysis of privacy policy text with gdpr article 13, in: Proceedings of the Web Conference 2021, 2021, pp. 2154–2164.

[492] C. Rudin, J. Radin, Why are we using black box models in ai when we don't need to? a lesson from an explainable ai competition, Harvard Data Science Review 1 (2) (2019).

[493] B. H. Van der Velden, H. J. Kuijf, K. G. Gilhuijs, M. A. Viergever, Explainable artificial intelligence (xai) in deep learning-based medical image analysis, Medical Image Analysis (2022) 102470.

[494] D. Saraswat, P. Bhattacharya, A. Verma, V. K. Prasad, S. Tanwar, G. Sharma, P. N. Bokoro, R. Sharma, Explainable ai for healthcare 5.0: opportunities and challenges, IEEE Access (2022).

[495] P. Kieseberg, E. Weippl, A. Holzinger, Trust for the doctor-in-the-loop, ERCIM news 104 (1) (2016) 32–33.

[496] D. Gunning, D. Aha, DARPA's explainable artificial intelligence (XAI) program, AI Magazine 40 (2) (2019) 44–58.

[497] D. Gunning, E. Vorm, Y. Wang, M. Turek, Darpa's explainable ai (XAI) program: A retrospective, Authorea Preprints (2021).

[498] J. A. Kroll, Accountable algorithms, Ph.D. thesis, Princeton University (2015).

[499] D. Danks, A. J. London, Regulating autonomous systems: Beyond standards, IEEE Intelligent Systems 32 (1) (2017) 88–91.

[500] J. K. Kingston, Artificial intelligence and legal liability, in: International Conference on Innovative Techniques and Applications of Artificial Intelligence, Springer, 2016, pp. 269–279.

[501] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, L. Kagal, Explaining explanations: An overview of interpretability of machine learning, in: 2018 IEEE 5th International Conference on data science and advanced analytics (DSAA), IEEE, 2018, pp. 80–89.

[502] K. Stöger, D. Schneeberger, A. Holzinger, Medical artificial intelligence: the european legal perspective, Communications of the ACM 64 (2021) 34–36.

[503] E. COMMISION, Regulation of the european parliament and of the council, EUROPIAN COMMISION, Brussels (2018).

[504] J. Zerilli, A. Knott, J. Maclaurin, C. Gavaghan, Transparency in algorithmic and human decision-making: is there a double standard?, Philosophy & Technology 32 (4) (2019) 661–683.

[505] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nature Machine Intelligence 1 (5) (2019) 206–215.

[506] J. M. Schoenborn, K.-D. Althoff, Recent trends in XAI: A broad overview on current approaches, methodologies and interactions., in: ICCBR Workshops, 2019, pp. 51–60.

[507] M. E. Kaminski, The right to explanation, explained, Berkeley Tech. LJ 34 (2019) 189.

[508] E. Commission, A European approach to artificial intelligence | Shaping Europe's digital future (apr 1 2021).

[509] A. Holzinger, B. Malle, A. Saranti, B. Pfeifer, Towards multi-modal causability with graph neural networks enabling information fusion for explainable ai, Information Fusion 71 (2021) 28–37.

[510] H. Lovells, Ai & Algorithms (Part 6): Spain to create Europe's first supervisory - Hogan Lovells Engage (jan 13 2022).

[511] T. Samp, Us and EU pledge to promote "innovative and trustworthy" AI | Insights | DLA Piper Global Law Firm (oct 19 2021).

[512] C. L. Translate, Provisions on the Management of Algorithmic Recommendations in Internet Information Services — (jan 4 2022).

[513] S. M. Santinato, Brazil: Proposed AI regulation (nov 15 2021).

[514] O. AI, Oecd AI's live repository of over 260 AI strategies & policies - OECD.AI (2021).

[515] U. , Recommendation on the ethics of artificial intelligence (feb 27 2020).

[516] C. Wohlin, 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, 2014, pp. 1–10.

CORELS	Certifiable Optimum RulE ListS
CP	Ceteris Paribus
DARPA	Defense Advanced Research Projects Agency
DeconvNet	Deconvolutional Network
DkNN	Deep k-Nearest Neighbors
DL	Deep Learning
DNN(s)	Deep Neural Network(s)
DT	Decision Tree
DWAC	Deep Weighted Averaging Classifier
EDA	Exploratory Data Analysis
EDN	Explainable Deep Network
FC	Fully Connected
FRBS(s)	Fuzzy Rule-based System(s)
FSS	Fuzzy Sets and System
GAM(s)	Generalized Additive Model(s)
GA^{2}M	GAMs with interaction
GAP	Global Average Pooling
GDPR	General Data Protection Regulation
Grad-CAM	Gradient-weighted Class Activation Mapping
HCI	Human-Computer Interaction
HLEG	High-Level Expert Group
HRAIs	High Risk AI Systems
ICE	Individual Conditional Expectation
IG(s)	Integrated Gradient(s)
IME	Interaction-based Method for Explanations
IML	Interpretable ML
kNN	k-Nearest Neighbours
LIME	Local Interpretable Model-agnostic Explanations
LOCO	Leave-One-Covariate-Out
LRP	Layer-wise Relevance Propagation
LSTM	Long Short-Term Memory
MCR	Model Class Reliance
ML	Machine Learning
MLP(s)	Multi-Layer Perceptron(s)
MMD-critic	Maximum Mean Discrepancy
MOC	Multi-Objective Counterfactuals
MSE	Mean Square Error
NC	Normal Cognitive
NeSy	Neural Symbolic
NN(s)	Neural Network(s)
NLP	Natural Language Processing
PCA	Principal Component Analysis
PDP	Partial Dependence Plot
PL	Prototypes Layer
RCFM	Rectified Convolutional Feature Maps
ReLU	Rectifying Linear Unit
RISE	Randomized Input Sampling for Explanation
RL	Reinforcement Learning
RML	Responsible ML
RNN(s)	Recurrent Neural Network(s)
RS	Relevance Score
SA	Sensitivity Analysis
SD	Standard Deviation
SENN(s)	Self-Explaining Neural Network(s)
SHAP	Shapley Additive Explanation
SL	Saliency Learning
SmGrad	Smooth Gradient
SML	Supervised Machine Learning
SP-LIME	Submodular Pick LIME
t-SNE	t-Distributed Stochastic Neighbor Embedding
TCAV(s)	Testing with CAV(s)
TED	Teaching Explanations for Decision
UMAP	Uniform Manifold Approximation and Projection
VAE-CaCE	Variational Auto Encoders based CaCE
VIA	Valid Interval Analysis
VQA	Visual Question Answering
WIT	What-If Tool
XAI	Explainable AI
XRL	Explainable Reinforcement Learning
XSML	Explainable Supervised ML

Highlights

- A novel four-axis framework to examine a model for robustness and explainability
- Formulation of research questions at each axis and its corresponding taxonomy
- Discussion of different explainability assessment methods
- A novel methodological workflow for determining the model and explainability criteria
- Revisited discussion on challenges and future directions of XAI and Trustworthy AI

Declaration of Generative AI and AI-assisted technologies in the writing process:

During the preparation of this work the author(s) have not used any such tools and all contents are prepared by the authors themselves. The authors take full responsibility for the content of the publication.

Declaration of interests

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

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: