

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/341817113>

# Explainable Artificial Intelligence: a Systematic Review

Preprint · May 2020

---

CITATIONS

0

READS

1,398

2 authors:



Giulia Vilone  
Technological University Dublin - City Campus  
8 PUBLICATIONS 22 CITATIONS

[SEE PROFILE](#)



Luca Longo  
Technological University Dublin - City Campus  
84 PUBLICATIONS 936 CITATIONS

[SEE PROFILE](#)

Some of the authors of this publication are also working on these related projects:



Mental workload modeling [View project](#)

# Explainable Artificial Intelligence: a Systematic Review

Giulia Vilone, Luca Longo

*School of Computer Science, College of Science and Health,  
Technological University Dublin, Dublin, Republic of Ireland*

## Abstract

Explainable Artificial Intelligence (**XAI**) has experienced a significant growth over the last few years. This is due to the widespread application of machine learning, particularly deep learning, that has led to the development of highly accurate models but lack explainability and interpretability. A **plethora of methods** to tackle this problem have been proposed, developed and tested. This systematic review contributes to the body of knowledge by clustering these methods with a hierarchical classification system with four main clusters: review articles, theories and notions, methods and their evaluation. It also summarises the state-of-the-art in XAI and recommends future research directions.

**Keywords:** Explainable artificial intelligence, method classification, survey, systematic literature review

## 1. Introduction

The number of scientific articles, conferences and symposia around the world in eXplainable Artificial Intelligence (XAI) has significantly increased over the last decade [1][2]. This has led to the development of a plethora of domain-dependent and context-specific methods for dealing with the interpretation of machine learning (ML) models and the formation of explanations for humans. Unfortunately, this trend is far from being over, with an abundance of knowledge in the field which is scattered and needs organisation. The goal of this article is to systematically review research works in the field of XAI and to try to define some boundaries in the field. From several hundreds of research articles focused on the concept of explainability, about 350 have been considered for review by using the following search methodology. In a first phase, Google Scholar was queried to find papers related to “explainable artificial intelligence”, “explainable machine learning” and “interpretable machine learning”. Subsequently, the bibliographic section of these articles was thoroughly examined to retrieve further relevant scientific studies. The first noticeable thing, as shown in figure 2(a), is the distribution of the publication dates of selected research articles: sporadic in the 70s and 80s, receiving preliminary attention in the 90s, showing raising interest in 2000 and becoming a recognised body of knowledge after 2010. The first research concerned the development of an explanation-based system and its integration in a computer program designed to help doctors make diagnoses [3]. Some of the more recent papers focus on work devoted to the clustering of methods for explainability, motivating the need for organising the XAI literature [4][5][6]. The upturn in the XAI research outputs of the last decade is prominently due to the fast increase in the popularity of ML and in particular of deep learning (DL), with many applications in several business areas, spanning from e-commerce [7] to games

[8] and including applications in criminal justice [9][10], healthcare [11], computer vision [10] and battlefield simulations [12], just to mention a few. Unfortunately, most of the models that have been built with ML and deep learning have been labelled ‘black-box’ by scholars because their underlying structures are complex, non-linear and extremely difficult to be interpreted and explained to laypeople. This opacity has created the need for XAI architectures that is motivated mainly by three reasons, as suggested by [12][13]: i) the demand to produce more transparent models; ii) the need of techniques that enable humans to interact with them; iii) the requirement of trustworthiness of their inferences. Additionally, as proposed by many scholars [13][14][15][16], models induced from data must be liable as liability will likely soon become a legal requirement. Article 22 of the General Data Protection Regulation (GDPR) sets out the rights and obligations of the use of automated decision making. Noticeably, it introduces the *right of explanation* by giving individuals the right to obtain an explanation of the inference/s automatically produced by a model, confront and challenge an associated recommendation, particularly when it might negatively affect an individual legally, financially, mentally or physically. By approving this GDPR article, the European Parliament attempted to tackle the problem related to the propagation of potentially biased inferences to society, that a computational model might have learnt from biased and unbalanced data.

Many authors surveyed scientific articles surrounding explainability within Artificial Intelligence (AI) in specific sub-domains, motivating the need for literature organisation. For instance, [17][18] respectively reviewed the methods for explanations with neural and bayesian networks while [19] clustered the scientific contributions devoted to extracting rules from models trained with Support Vector Machines (SVMs). The goal was, and in general is, to create rules highly interpretable by humans while maintaining a degree of accuracy offered by trained models. [20] carried out a literature review of all the methods focused on the production of visual representations of the inferential process of deep learning techniques, such as heat-maps. Only a few scholars attempted to make a more comprehensive survey and organization of the methods for explainability as a whole [1][21]. This paper builds on these efforts to organise the vast knowledge surrounding explanations and XAI as a discipline, and it aims at defining a classification system of a larger scope. The conceptual framework at the basis of the proposed system is represented in Figure 1. Most of the methods for explainability focus on interpreting and making the entire process of building an AI system transparent, from the inputs to the outputs via the application of a learning approach to generate a model. The outcome of these methods are explanations that can be of different formats, such as rules, numerical, textual or visual information, or a combination of the former ones. These explanations can be theoretically evaluated according to a set of notions that can be formalised as metrics, usually borrowed from the discipline of Human-Computer Interaction (HCI) [22].

The remainder of this paper is organised as it follows. Section 2 provides a detailed description of the research methods employed for searching for relevant research articles. Section 3 proposes a classification structure of XAI describing top branches while Sections 5-4 expand this structure. Eventually, section 8 concludes this systematic review by trying to define the boundaries of the discipline of XAI, as well as suggesting future research work and challenges.

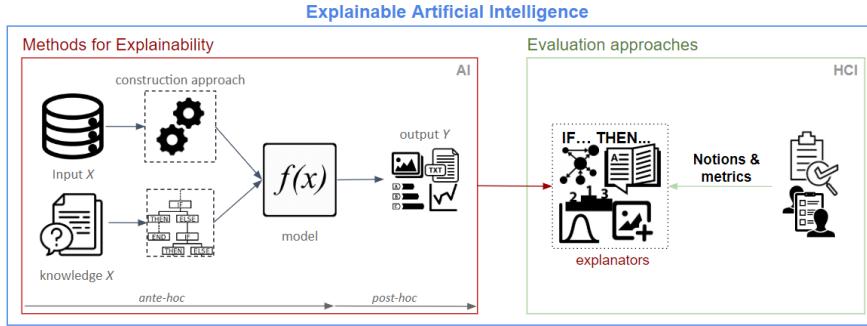


Figure 1: Diagrammatic view of Explainable Artificial Intelligence as a sub-field at the intersection of Artificial Intelligence and Human-Computer Interaction

## 2. Research methods

Organizing the literature of explainability within AI in a precise and indisputable way as well as setting clear boundaries is far from being an easy task. This is due to the multidisciplinarity surroundings of this new fascinating field of research spanning from Computer Science to Mathematics, from Psychology to Human Factors, from Philosophy to Ethics. The development of computational models from data belongs mainly to Computer Science, Statistics and Mathematics, whereas the study of explainability belongs more to Human Factors and Psychology since humans are involved. Reasoning over the notion of explainability touches Ethics and Philosophy. Therefore, some constraints had to be set, and the following publication types were excluded:

- scientific studies discussing the notion of explainability in different contexts than AI and Computer Science, such as Philosophy or Psychology;
- articles or technical reports that have not gone through a peer-review process;
- methods that could be employed for enhancing the explainability of AI techniques but that were not designed specifically for this purposes. For example, the scientific literature contains a considerable amount of articles related to methods designed for improving data visualization or feature selection. These methods can indeed help researchers to gain deeper insights into computational models, but they were not specifically designed for producing explanations. In other words, those methods developed only for enhancing model transparency but not directly focused on explanation were discarded.

Taking into account the above constraints, this systematic review was carried out in two phases:

1. papers discussing explainability were searched by using Google Scholar and the following terms: '*explainable artificial intelligence*', '*explainable machine learning*', '*interpretable machine learning*'. The queries returned several thousands of results, but it became immediately clear that only the first ten pages could contain relevant articles. Altogether, these searches provided a basis of almost two hundred peer-reviewed publications;
2. the bibliographic section of the articles found in phase one was checked thoroughly. This led to the selection of one hundred articles whose bibliographic section was recursively analysed. This process was iterated until it converged and no more articles were found.

### 3. Classification of scientific articles on explainability

After a thorough analysis of all the selected articles, four main categories were extracted as depicted in Fig. 2 and as listed below:

- reviews on methods for explainability - it includes either literature or systematic reviews of those methods devoted to the proposal and/or testing of solutions for the explainability of data- and knowledge-driven models;
- notions related to the concept of explainability - it includes studies focused on the definition of those notions related to the concept of explainability and on the determination of the main characteristics as well as the requirements of an effective explanation;
- development of new methods for explainability - it includes articles that propose novel and original methods for enhancing the explainability of data/knowledge-driven models;
- evaluation of methods for explainability - it includes articles reporting the results of scientific studies aiming at evaluating the performance of different methods for explainability.

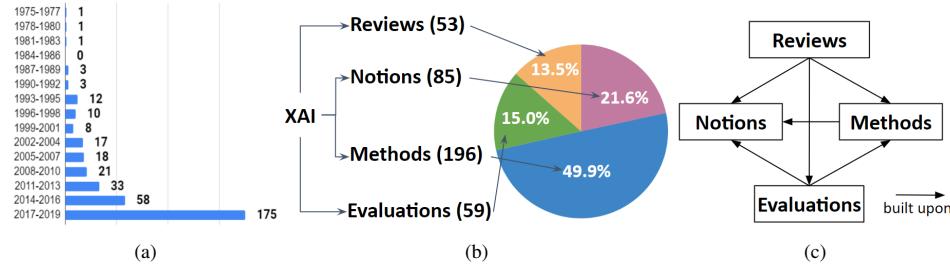


Figure 2: Proposed classification of the XAI literature with (a) the distribution of published scientific articles over time, (b) the root of our hierarchical classification system representing the main four categories and the percentage of articles in each , and (c) the salient relations between these categories that have emerged.

Following the proposed classification, it was possible to design a map of the XAI literature in form of a tree whose root contains the above four categories (figure 2, part b). This tree expands into branches of different depth where leaves represent scientific articles. Figure 2, part b, also shows the percentage of articles grouped by each category, clearly highlighting the distribution of the research efforts towards the development of methods for explainability. Note that, a paper might appear in multiple branches of this classification, as it might cover multiple dimensions. Figure 2, part c, depicts the dependencies of the main four categories. In general, scholars would not be able to carry out reviews of the XAI literature without the existence and consideration of relevant notions and methods for explainability as well as the approaches for evaluating the performances of these methods. Evaluation approaches naturally followed the creation of methods for explainability which have been engineered to meet as many requirements of an effective explanation as possible.

#### 4. Reviews of the XAI literature

This category contains literature and systematic reviews devoted to specific classes of solutions for explainability, such as systems generating textual explanations [23], or constrained to specific AI techniques as, for instance, neural networks [24] (summary in table A.2 and figure 3). These reviews provide an entry point for researchers to acquire information and get familiar with the key aspects of the rapidly growing body of research related to explainability. They also attempt to summarise the main techniques for explainability and to highlight their strengths and limitations. Seven clusters emerged based on distinct aspects of explainability covered by these reviews:

- **application fields** - reviews on methods for explainability in a specific field of application;
- **construction approaches** - reviews on methods for explainability specifically designed to explain the inferential process of models. This category has been further divided into:
  - **data-driven** approaches which focus on extracting new knowledge from trained models from data, but **without accounting for the prior knowledge** of domain experts.
  - **knowledge-driven** approaches focused on capturing an expert's knowledge and logic, often embedded in the notion of agent;
- **theories & concepts** - reviews of the notions related to the concept of explainability;
- **output formats** - reviews on methods for explainability focused on generating specific formats of explanations, such as visual or rules;
- **problem types** - review articles on methods designed to explain the logic of data and knowledge-driven models applied to a specific type of problem, namely regression or classification;
- **generic reviews** - generic reviews that cover a wide range of data/knowledge-driven models as well as their methods for explainability and cannot be placed within any other category.

In the application fields cluster, the assumption of the methods for explainability is that it is not possible to accept the inference made by a model without understanding its functioning because a decision, supported by a wrong prediction, can have a dramatic impact on people's lives [1]. The second cluster, construction data-driven approaches, contains reviews of methods for explainability for specific data-driven learning approaches, mainly neural networks [25] [20] [26] [27] [28] [29] [17] [30], bayesian networks [18] and SVMs [31] [19], not constrained to a specific type of input data for the approach or a particular output format for an explanation, such as images or texts. Other scholars instead focused on reviewing methods for knowledge-based approaches such as Expert Systems (ES) [32] and Intelligent Systems [33]. In particular, these surveys analysed what types and formats of explanations were tested on these systems and which ones work better than others. For instance, [34] showed that rich explanations, based on a combination of information regarding users, items and features, are very effective, while [33] claimed that explanations should be context-specific to be effective. The third cluster contains those reviews focused on objectively defining the concept of explainability and its set of related

## pause to review learning types

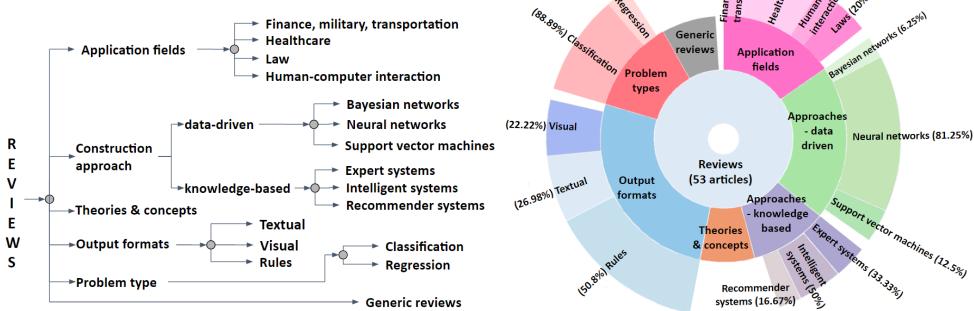


Figure 3: Hierarchical classification of the review articles on explainable artificial intelligence and machine learning interpretability (left) and distribution of the review articles across categories (right).

notions, which are discussed in depth in section 5.1. One of these studies presented an overview of different theories of explanation borrowed from the cognitive science and philosophy disciplines, contextualised within case-based reasoning [35]. In details, it is believed that, in order to be effective, an AI system should: (I) explain how it reached the answer and (II) why it is a good answer, (III) why a question is relevant or not, (IV) clarify the meaning of the terms used in the system that might not be understood by the users and, lastly, (V) teach the user about the domain. In short, the goals that an explanation must achieve depend on the domain under consideration, the underlying model and end-users. Similarly, [23] suggested that explanations should take into account the preferences and preconceptions of end-users. This can be achieved by incorporating more findings from the behavioural and social sciences into the newly emerging field of XAI. For example, people explain their behaviour based on their beliefs, desires and intentions hence these elements must be considered in an explanation. Eventually, explanations based on counterfactual examples should help end-users to understand the logic of an underlying model by leveraging on people's capability to infer general rules from a few examples. Counterfactuals add also something new to what is already known from the existing data and provide additional information on how a model behaves in novel, unseen situations [36]. The fourth cluster contains reviews of methods for explainability generating a specific output format for an explanation (further discussed in section 6). Methods generating textual explanations are surveyed in [32] and compared according to some requirements about the structure and content of the explanations to adapt them to the users' needs and knowledge. [37] focused on written explanations generated from fuzzy rules integrated with natural language generation tools. The underlying reasonable assumption is that the understandability of these rules cannot be given for granted. Researchers studied the capabilities of 'data-to-text' approaches that automatically create linguistic descriptions from a complex dataset by means of aggregation functions, implemented as fuzzy rules, that aggregate 'computational perceptions'. A computational perception is "a unit of meaning for the phenomenon under analysis and is identified by a set of linguistic expressions and their corresponding validity values given a situation." Some methods combine Logical AI and Statistical AI to generate textual explanations [38]. The former is concerned with 'formal languages' to represent and reason with qualitative specifications, while the latter is focused on learning quantitative specifications from data. However, the authors claimed that the search for

an effective way to learn representations of the inferential process of data-driven models is still open [38]. A body of literature focused on the visual explanation of deep learning models. Explanators generating salient masks were investigated in [20, 30] whilst [20, 39] reviewed methods that graphically represent the inner structure and functioning of neural networks with flow-charts or other explanatory graphs. An interesting alternative was proposed in [40] whereby methods based on nomograms, rule induction, fuzzy logic, graphical models and topographic mapping can be utilised to explain data-driven models and learning techniques. Similarly to textual explanation, the problem of visually inspecting data-driven models has not been resolved and there are still challenges and open questions to be answered. Some reviews summarised the methods for explainability that generate sets of rules from underlying trained models [41] by extracting frequent relations from a dataset using fuzzy logic and fuzzy rules [42, 43, 40], the integration of symbolic logic with the neural networks [44, 45] and, more generally, the usage of automated reasoning to shed a light over the inferential process of automatically constructed data-driven models [46]. The fifth cluster contains reviews that analysed the methods for explainability for either regression [47] or classification [31, 48, 49] problems. They have a broader scope than the previous reviews as they range over several fields, AI techniques and explanation types. Their goal was to summarise the important issues, still unresolved, of interpreting prediction models for both problem types and encouraging researchers to improve the existing or discover novel methods for explainability. Eventually, some reviews have a more generic scope. They are aimed at proposing a comprehensive way of organizing the several methods for explainability [1, 50, 21] or describing them [50, 51, 52]. A group of these reviews tried to evaluate the performances of various methods. This is done by comparing the explanations automatically produced by these methods [19] or by measuring how much they fulfil certain notions of explainability, such as completeness, through the use of either quantitative or qualitative metrics [53] (further discussed in section 7).

## 5. Notions related to the concept of explainability

Explaining a model induced from data by employing a specific learning technique is not a trivial goal. A body of literature focused on achieving such a goal by investigating and attempting to define the concept of *explainability*, leading to many types of explanation and the formation of several attributes and structures. To organise these, the specific following clusters are proposed:

- attributes of explainability - it contains criteria and characteristics used by scholars to try to define the construct of ‘explainability’;
- types of explanation - it includes the different ways scholars reported explanations for their ad-hoc applications, what pieces of information are included or left out;
- structure of an explanation - it contains the various components an explanation can be constructed on, such as causes, context, and consequences of a model’s prediction as well as their ordering.

### 5.1. Attributes of explainability

One of the principal reasons to produce an explanation is to gain the trust of users [54]. Trust is the main way to increase users’ confidence with a system [55] and to make them feel comfortable while controlling and using it [56]. Besides trust, researchers determined other

## **"explanation" not just one thing**

positive effects brought by explainability. According to [57], it is part of human nature to assign causal attribution of events. A system that provides a causal explanation on its inferential process is perceived more human-like by end-users as a consequence of the innate tendency of human psychology to anthropomorphism. Thus, several scholars spoke at length about causality which is considered a fundamental attribute of explainability [12, 58, 59, 56, 23]. Explanations must make the causal relationships between the inputs and the model's predictions explicit, especially when these relationships are not evident to end-users. Data-driven models are designed to discover and exploit associations in the data, but they cannot guarantee that there is a causal relationship in these associations. As pointed out in [56], the task of inferring causal relationships strongly depends on prior knowledge, but some associations might be completely unexpected and not explainable yet. Scientists can use these associations to generate hypotheses to be tested in scientific experiments; however, this is outside the scope of the methods for explainability. Other four reasons supporting the necessity to explain the logic of an inferential system or a learning algorithm were suggested in [1]:

- **explain to justify** - the decisions made by utilising an underlying model should be explained in order to increase their justifiability;
- **explain to control** - explanations should enhance the transparency of a model and its functioning, allowing its debugging and the identification of potential flaws;
- **explain to improve** - explanations should help scholars improve the accuracy and efficiency of their models;
- **explain to discover** - explanations should support the extraction of novel knowledge and the learning of relationships and patterns.

Despite the widely recognised importance of explainability, researchers are striving to determine universal, objective criteria on how to build and validate explanations [22]. Numerous notions underlying the effectiveness of explanations were proposed in the literature (as summarised in table 1). [22] surveyed 250 articles from the fields of Philosophy, Psychology and Cognitive Science to analyse in depth how people define, generate, select, evaluate and present explanations. The author also presented an interesting definition of XAI as a human-agent interaction problem where the agent reveals the underlying causes to its or another agent's decision process. In other words, XAI is believed to be a subset of the human-agent interaction field that can be defined as the intersection of AI, social science and HCI.

Two studies on *explainability* demonstrated that this concept is utilised in several fields, spanning from Mathematics, Physics, Computer Science to Engineering, Psychology, Medicine and Social sciences [63, 75]. Explainability is often replaced with the notion of *interpretability*, considered as synonyms within the general AI community, and in particular by those scholars in automated learning and reasoning, whereas it seems that the software engineering community prefers the term *understandability* [63]. Generally speaking, interpretability is often defined as the capacity to provide or bring out the meaning of an abstract concept and understandability as the capacity to make the model understandable by end-users (see table 1). However, other definitions are proposed in the literature. Explainability or interpretability is defined in [26] as "the degree to which a human observer can understand the reason behind a decision (or a prediction) made by the model". An interesting distinction between the concepts of *interpretation*

Table 1: Definition of the notions related to the concept of explainability

Notion	Description & Reference
Algorithmic transparency	The degree of confidence of a learning algorithm to behave ‘sensibly’ in general [26][2]
Actionability	The capacity of a learning algorithm to transfer new knowledge to end-users [60][61]
Causality	The capacity of a method for explainability to clarify the relationship between input and output [12][58][57][59][56][23]
Completeness	The extent to which an underlying inferential system is described by explanations [53][60][61]
Comprehensibility	The quality of the language used by a method for explainability [62][63][64][65][66][13][67][68][69]
Cognitive relief	The degree to which an explanation decreases the “surprise value” which measures the amount of cognitive dissonance between the explanandum and the user’s beliefs. The explanandum is something unexpected by the user that creates dissonance with his/her beliefs [58]
Correctability	The capacity of a method for explainability to allow end-users make technical adjustments to an underlying model [60][61]
Effectiveness	The capacity of a method for explainability to support good user decision-making [70][71][72]
Efficiency	The capacity of a method for explainability to support faster user decision-making [70][55][71]
Explicability	The degree of association between the expected behaviour of a robot to achieve assigned tasks or goals and its actual observed actions [73]
Explicitness	The capacity of a method to provide immediate and understandable explanations [74]
Faithfulness	The capacity of a method for explainability to select truly relevant features [74]
Intelligibility	The capacity to be apprehended by the intellect alone [75][76][5][77][78]
Interactivity	The capacity of an explanation system to reason about previous utterances both to interpret and answer users’ follow-up questions [79][80]
Interestingness	The capacity of a method for explainability to facilitate the discovery of novel knowledge and to engage user’s attention [64][81][67][65][82]
Interpretability	The capacity to provide or bring out the meaning of an abstract concept [64][50][83][66][13][22][29][84][85][4][6][86]
Informativeness	The capacity of a method for explainability to provide useful information to end-users [56]
Justifiability	The capacity of an expert to assess if a model is in line with the domain knowledge [1][64][50][33]
Mental Fit	The ability for a human to grasp and evaluate a model [64][87]
Monotonicity	The relationship between a numerical predictor and the predicted class that occurs when increasing the value of the predictor leads to either always increase or decrease the probability of an instance’s membership to the class [88]
Persuasiveness	The capacity of a method for explainability to convince users perform certain actions [70][55][71]
Predictability	The capacity to anticipate the sequence of consecutive actions in a plan [73]
Refinement	The capacity of a method to guide experts in improving the performance/robustness of a model [89]
Reversibility	The capacity to allow end-users to bring a ML-based system to an original state after it has been exposed to an harmful action that makes its predictions worse [60][61]
Robustness	The persistence of a method for explainability to withstand small perturbations of the input that do not change the prediction of the model [90][89]
Satisfaction	The capacity of a method to increase the ease of use and usefulness of a ML-based system [70][55][71]
Scrutability / diagnosis	The capacity of a method for explainability to inspect a training process that fails to converge or does not achieve an acceptable performance [89][70][55]
Security	The reliability of a model to perform to a safe standard across all reasonable contexts [91]
Selection / simplicity	The ability of a method for explainability to select only the causes that are necessary and sufficient to explain the prediction of an underlying model [23]
Sensitivity	The capacity of a method for explainability to reflect the sensitivity of the underlying model with respect to variations in the input feature space [92][93]
Simplification	The capacity to reduce the number of variables under consideration to a set of principal ones [94]
Soundness	The extent to which each component of an explanation’s content is truthful in describing an underlying system [60][61]
Stability	The consistency of a method to provide similar explanations for similar/neighboring inputs [74]
Transparency	The capacity of a method to explain how the system works even when it behaves unexpectedly [76][26][13][95][84][14][15][70][55][16][96][86]
Transferability	The capacity of a method to transfer prior knowledge to unfamiliar situations [56]
Understandability	The capacity of a method of explainability to make a model understandable [75][63][64][89][97]

and *explanation* was proposed in [29]. On one hand, an interpretation is the mapping of an abstract concept (as a predicted class) into a domain that the human can make sense of, such as, for instance, images or texts that can be inspected and classified by people. On the other hand, an explanation is the collection of features of an interpretable domain that contributed to produce a prediction for a given item. The authors of [29] did not specify how to determine this collection of features. The selection criteria are to be decided by researchers according to several factors like the type of input data and the degree of refinement in the explanation demanded by end-users. An expansion of the definition of interpretability through the determination of its main characteristics was presented in [74][22][85]. In detail, [74] suggested the following requirements: (I) *fidelity* - the representation of inputs and models in terms of concepts should preserve and present to end-users their relevant features and structures, (II) *diversity* - inputs and models should be representable with few non-overlapping concepts, and (III) *grounding* - concepts should have an immediate human-understandable interpretation. These requirements were further expanded in [22] by listing a set of characteristics that an explanation should possess:

- *contrastive nature of explanations* - people seek for an explanation when they are presented with counterfactual and/or counter-intuitive events;
- *selectivity of explanations* - people usually do not expect that an explanation contains the actual and complete list of the causes of an event, but only a selection of the few causes deemed to be necessary and sufficient to explain it. Authors point out the risk that this selection might be influenced by cognitive biases;
- *social nature of explanations* - explanations are part of a dialogue aiming at transferring knowledge, therefore, they are based on the beliefs of both the explainer and explainee;
- *irrelevance of probabilities to explanations* - referring to the occurrence probabilities of events or to the statistical relationships between causes and events does not produce a satisfactory and intuitive explanation. Explanations are more effective when they refer to the causes and not to their likelihood.

Four further requirements for enhancing the interpretability of visual explanations were added in [85]: i) *graphical integrity* - the representations should highlight the features that contribute the most to the final predictions and distinguish those with positive and negative attribution, ii) *coverage* - a large fraction of the most important features should be visible in the representation, iii) *morphological clarity* - the important features should be clearly displayed, their visualization cannot be ‘noisy’, and iv) *layer separation* - the representation cannot occlude the raw image which should be visible for human inspection. Other two notions strongly correlated with interpretability are *comprehensibility* [64] and *intelligibility* [75]. However, scholars highlighted some differences. [66] proposed to distinguish between *interpretable systems*, systems in which end-users can mathematically analyse algorithms, and *comprehensible systems* that “emit symbols enabling user-driven explanations of how a conclusion is reached”. Two studies [75][78] defined intelligibility as an attribute of user-centric reasoned explanations that are easily interpretable by end-users and that draws from foundational concepts of other disciplines such as Philosophy and Cognitive Psychology. Additionally, both studies recommended exploiting the experience and knowledge of the HCI community in making interfaces that empower people to assure that intelligibility will be one of the core requirements of the next generation of AI systems. Other authors focused on breaking some of the notions identified in table [1] into sub-notions

or on assigning further requirements. For example, three sub-notions related to *transparency* that should be achieved by any learning model were defined in [26] [56]:

- *simulatability* - the capacity of a model to allow a user to understand its structure and functioning entirely;
- *decomposability* - the degree to which a model can be decomposed into its individual components (input, parameters and output) and of their intuitive explainability;
- *algorithmic transparency* - the degree of confidence of a learning algorithm to behave 'sensibly' in general (see also table I).

However, according to [56], it is not possible to achieve algorithmic transparency in neural networks because of the current incapacity of experts to understand the inferential process of these models and to prove that they work correctly on new, unseen observations. Scholars attempted to overcome this shortcoming by finding methods to trace the predictions of a model to the most influential features of the input. Examples of these methods are heat-maps [98] which are created by back-propagating the predictions of a model to the input space and highlighting relevant pixels. Alternatively, [99] proposed a solution to satisfy the simulatability and decomposability properties by substituting black-box models with Generalized Additive Models (GAMs). GAMs are linear combinations of simple models trained on a single feature of an input dataset, thus allowing end-users to quantify the contribution of each feature to the outcome. However, transparency must be handled with caution because it can be dangerous under certain circumstances, as highlighted in [96]. Requiring that data and models are fully visible to end-users prevents the creation of intellectual properties; this can significantly slow down the development of new technologies. Moreover, data can contain sensitive or personal information which cannot be made public without affecting people's privacy. Finally, the displaying of more information might push a researcher to optimise a model on specific instance(s) but deteriorating its overall performance and degree of generalisability. Scholars extensively investigated *sensitivity* [92] [93]. In this context, sensitivity is considered as the sensibility of explanations to variations in the input features, model implementation and, subsequently, in the model's predictions. [92] introduced the requirement of *input invariance* meaning that a method for explainability must mirror the sensitivity of the underlying model with respect to transformations of the inputs in order to ensure a reliable interpretation of their contribution to each prediction. [93] focused on the sensitivity of methods for explainability specifically designed for neural networks, in particular those that quantify the contribution of input features to the predictions, such as DeepLift [100] and Layer-wise Relevance Propagation (LRP) [101]. In this case, a method for explainability satisfies the sensitivity requirement if it assigns a non-zero contribution to an input feature when two instances, in the input space, differ in that feature only but lead to different predictions. According to [93], methods for explainability must also fulfill the requirement of *implementation invariance*. This suggests that a method applied to functionally equivalent neural networks should assign identical contributions to the features of the input. Two neural networks are *functionally equivalent* if their predictions are equal for all inputs despite having different implementations and architectures. Finally, scholars identified various factors that might affect the *interestingness* of a model, in particular of the rule-based ones [81] [67]. First, *rule size* is the number of instances satisfied by a rule. Usually, small size rules are undesirable as they explain only a few instances. The main aim is to discover rules that cover a large portion of the input data. However, there are situations where small rules might capture exception occurring in the

data that can be of interest for scientists. Second, *imbalance of class distributions* occurs when the instances belonging to a class are more frequent than those of another class. It might be more difficult, hence more interesting, to discover those rules aimed at predicting the minority classes. *Attribute costs* represent the cost to get access to the actual value of an attribute of the data. For example, it is easy to assess the gender of a patient but the determination of some health-related attributes can require an expensive investigation. Rules that utilise only ‘cheap’ attributes are more interesting. Eventually, the interestingness of a rule must take into account the *miscalcification costs*. In some domain of application, the erroneous classification of an instance might have a significant impact, not only in terms of money. In case of medical diagnosis, classifying as healthy a patient affected by a lethal disease might lead to premature death. Interestingness was also examined for Reinforcement Learning (RL) agents which are designed to take actions in a specific environment with the aim to maximize a cumulative reward [82]. The authors proposed a framework to make the behaviour of these agents explainable by analysing their historical interactions with the environment and extracting a few *interestingness elements*. Examples of interesting elements of these interactions are the portion of environment observed by the agent, the frequency of certain types of interactions and the cost (in terms of a reward) of the interactions carried out.

**rules:**

**backwards chaining “why” = current goal**

**forwrd chaning “how” = rules used to conclude X**

### 5.2. *Types of explanations*

Researchers tried to create a classification system for the types of explanation suitable for interpreting the logic of learning algorithms. A method for explainability should answer several questions to form an exhaustive explanation. The two most common questions are *why* and *how* the model under scrutiny produces its predictions/inferences [102] [103] [2] [7]. However, scholars identified other questions that might arise and that require different answers, thus different types of explanations [104]. Additionally, as pointed out in [105] [106], distinct behaviours, distinct problems and distinct types of users require distinct explanations. This has led to many ad-hoc classifications that are domain-dependent and are hard to be merged into one. For example, [107] focused on the types of users of methods for explainability. They proposed a two-class system consisting of *traced-based explanations*, useful for system designers, that accurately reflects the reasoning implemented within a model, and *reconstructive explanations*, designed for end-users, based on an active, problem-solving approach. A reconstructive explanation tends to build a ‘story’ exposing the input features contributing to a prediction. For instance, an image of a bird was assigned to a certain class because of the colour of the bird. However, the model might have analysed other features that did not influence the final assessment, like the image’s background. These characteristics can be included in the traced-based explanations but excluded from the reconstructive explanations. The same scholars also developed Reconstructive EXplanation (REX) [107] [108], an explanatory tool capable of producing reconstructive textual explanations for expert systems. REX is built on a model that maps the execution of the expert system onto a textbook representation of the domain. A textbook representation presents the domain knowledge in human-understandable explanations, much of which comes from domain textbooks. The explanation consists of mapping over key elements from the execution trace and expanding on them using the more structured textbook knowledge, which is a collection of relationships between cues, hypotheses and goals as illustrated by this example: “The presence of damages to the drainage pipes is a sign that the cause of an excessive high uplift pressures on a concrete dam is internal erosion of soil under the dam. Erosion would lead to broken pipes, therefore slowing drainage and causing high uplift pressures”. The goal is to determine the cause of high uplift



pressure on a concrete dam, the cues consist of the presence of broken pipes and the hypothesis is the erosion of soil. Another classification of the types of explanations was proposed in [109] for intelligent systems which include intelligent agents, such as those AI assistants utilised in customer support chats, or other support decision systems like those for medical diagnoses. Here, traced-based explanations were defined as *mechanistic explanations* and correspond to the answer of the question “*How does it work?*”. Hence, they must offer insights into the causes and consequences of events and how these events and the different components of the intelligent systems interact to give rise to complex actions. Reconstructive explanations were instead called *ontological explanations* and describe the structural properties of the intelligent systems: its components, their attributes, and how they are related to each other. [109] also added a third category, referred to as *operational explanations* which respond to the question “*How do I use it?*” by relating goals to the mechanics designed to realise them. A more articulated classification of the types of explanations was introduced in [110] and it is based on five types of explanations that intelligent systems should produce. The first one, *teaching explanations*, aims at informing humans about the concepts learned by the system such as, for example, the presence of some physical constraints (walls or other obstacles) that can limit its actions. *Introspective tracing explanations* have the goal of finding the cause of and the solution to a fault whilst *introspective informative explanations* aim at explaining predictions based on the reasoning process to improve human-system interaction. The last two types of explanations, *post-hoc explanations* and *execution explanations*, are respectively focused on explaining the decisions and their execution without necessarily following the same reasoning process and directly linking them with the inputs. An example of post-hoc and execution explanation is a robot describing the path it wants to follow to go from point A to point B and all the movements it must do to cover that path. This explanation can mention the characteristics of the surrounding environment that have been considered while planning the path, but it does not mention that alternative paths were considered and discarded and the reasons beyond these decisions. Finally, [111] presented a classification of the types of knowledge intrinsically embedded in an explanation. Explanations based on *reasoning domain knowledge* focus on the domain knowledge needed to perform reasoning, including rules and terminology. *Communication domain knowledge* is instead about the domain knowledge needed to inform, clearly and comprehensively, end-users about the underlying domain, and it might include additional information not strictly necessary for reasoning. Eventually, *domain communication knowledge* focuses on how to communicate within a certain domain of application and it deals with practical aspects of the communication process, such as the language to be used, the most effective strategies for effective explanations and the communication medium. This knowledge must be tuned to the prior knowledge and cognitive state of the hearer.

### 5.3. *Structures of explanations*

The most effective way to structure explanations is still an open problem despite being tackled by several scholars. As highlighted in [112], two properties of the structure of an explanation can have a significant effect on learning, namely the capacity to “accommodate novel information in the context of prior beliefs and do so in a way that fosters generalization”. As prior beliefs greatly vary according to the application field and the domain knowledge of end-users, researchers examined and proposed different structures for explanations which are domain-dependent. The first studies on the most suitable and effective structures of textual explanations were carried out in the 80-90s and focused on interpreting the inferential process of expert models. Most of these explanations were planned as dialogues where end-users were allowed to ask a (limited) number of questions via an explanatory tool. Blah [113], an example of these tools, was primarily concerned

with structuring explanations so that they do not appear too complex. It was based on a series of psycho-linguistic studies that analyzed how human beings explain decisions, choices, and plans to one another. Different ways to structure a conversational explanation, or dialogue, to successfully transfer knowledge from an explainer to an explainee were listed in [114] [115] [116] [117] [80]. All these studies proposed to split a dialogue into three stages: opening, explanation and closing stage. Each stage has to obey a set of rules to ensure that the knowledge about the model's inferential process can be successfully transferred to end-users. On one hand, [80] grounded this three-stage formal protocol on the data collected from almost four hundred real dialogues which were examined to detect the key components of an explanation, the relationships between them and their order of occurrence. These main components can be synthesised by a set of questions (mainly how, why and what) and the relative arguments presented by an explainer to an explainee who, respectively, answer the questions and acknowledge the explanation or challenge it with counterfactual examples. On the other hand, [115] [116] [117] focused on the most effective set of rules to manage interactive dialogues with interruptions from the user while maintaining coherence between the different sections of an explanation. They also developed a tool, called EDGE, that generates dialogues based on these rules. EDGE updates assumptions about the user's knowledge based on his/her questions and uses this information to influence the further planning of the explanation. Other studies on interactive dialogues [118] [119] [120] [121] [122] focused on the structure, the language and main components (what pieces of information must be included) of these dialogues. Based in these early studies, [123] [124] [125] [126] proposed a modular architecture for explaining the behavior of simulated entities in military simulations. It consists of three modules: a reasoner, a natural language generator and a dialogue manager. The user can stop simulation and query about what happened at the current time point by selecting questions from a list. The dialogue manager orchestrates the system's response: firstly, by using the reasoner to retrieve the relevant information from a relational database, then producing English responses using the natural language generator. More recently, interactive dialogues were used as the explanation format of choice in knowledge-based systems other than expert systems. AutoTutor [127], designed to be integrated into tutoring systems, is grounded on learning theories and tutoring research. It simulates a human tutor by holding a conversation with the learner in natural language.

The explanations of task planning systems, according to [12] [128], must contain information on (I) why a planner choose an action, (II) why a planner did not choose another action, (III) why the decisions of a planner are the best among a set of possible alternatives, (IV) why certain actions cannot be executed and (V) why one needs or does not need to change the original plan. The criterion of *episodic memory* was added to the above list by [128], whereby an agent should remember all the factors that influenced the generation and execution of a plan such as "states, actions, and values considered during plan generation, traces of plan execution in the environment, and anomalous events that led to plan revision". A formal framework to generate *preferred explanations* of a plan was introduced in [129]. Preferences over explanations must be contextualized with respect to complex observational patterns. Actions might be affected by several causes and requires reflecting on the past, meaning that explanations must take into consideration previous events and information.

## 6. Development of new methods for explainability

More than 200 scientific articles were found that aim at developing new methods for explainability. Over time, researchers have tried to comprehend and unfurl the inner mechanics of data-driven, knowledge-driven models in various ways. From an examination of these articles, two main criteria exist for discriminating methods for explainability:

- scope - it refers to the scope of an explanation that can be either *global* or *local*. In the former case, the goal is to make the entire inferential process of a model transparent and comprehensible as a whole. In the latter case, the objective is to explicitly explain each inference of a model [130] [26] [56] [17];
- stage - it refers to the stage at which a method generates explanations. *Ante-hoc* methods are generally aimed at considering explainability of a model from the beginning and during training to make it naturally explainable whilst still trying to achieve optimal accuracy or minimal error [13] [99] [131]; *post-hoc* methods are aimed at keeping a trained model unchanged and mimic or explain its behaviour by using an external explainer at testing time [13] [56] [29] [97].

Taking into account the articles examined in this systematic review, and inspired by the classification system in [21], we propose additional criteria:

- problem type - methods for explainability can vary according to the underlying problem: *classification* or *regression*;
- input data - the mechanisms followed by a model to classify images can substantially differ from those used to classify textual documents, thus the input format of a model (*numerical/categorical, pictorial, textual or times series*) can play an important role in constructing a method for explainability;
- output format - similarly, different formats of explanations useful for different circumstances can be considered by a method for explainability: *numerical, rules, textual, visual* or *mixed*.

Figure 4 depicts the main branches of methods for explainability and shows the distribution of the articles across these branches. Each of the many methods for explainability retrieved from the scientific literature can be robustly described by using the five categories of figure 4 (stage, scope, problem type, input data and output format). Additionally, as it is possible to notice from Figure 4 the post-hoc methods are further divided into *model-agnostic* and *model-specific* methods [21]. The former methods do not consider the internal components of a model such as weights or structural information, therefore they can be applied to any black-box model. The latter methods are instead limited to specific classes of models. For example, the interpretation of the weights of a linear regression model is specific to the learning approach (linear regression). Similarly, methods that only work with the interpretation of neural networks are model-specific [25] [26] [30]. Model agnosticity and specificity do not usually apply to the class of ‘ante-hoc’ methods because their goal is to make the functioning of a model transparent, so almost all them are intrinsically model-specific [13]. Some post-hoc methods for explainability can be applied both at a global or local scope [132] and can work for either regression or classification problems [133].

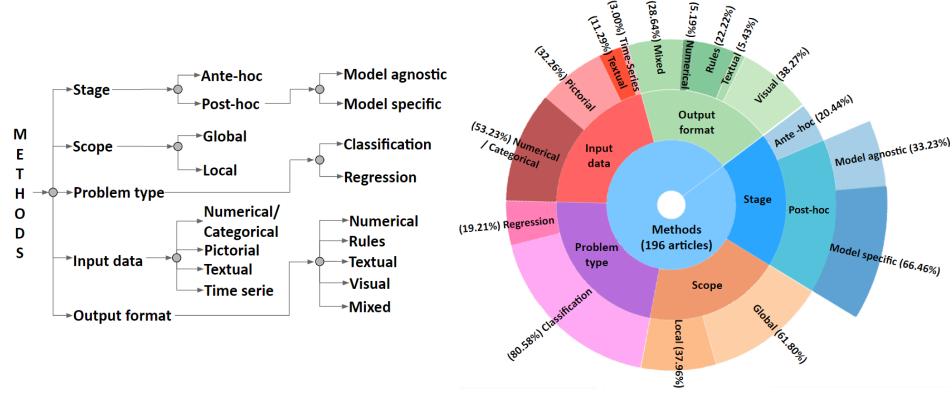


Figure 4: Classification of methods for explainability (left) and distribution of articles across categories (right).

The following sections try to succinctly describe the main classes of methods for explainability found during this systematic review, accompanied by tables for reporting their stage, scope, problem type, input data and output format and sorting them in alphabetic order. Given the large number of methods found, it was decided to group them into five thematic classes.

### 6.1. Output formats

*Visual explanations* are probably the most natural way of communicating things and a very appealing way to explain them. Visual explanations can also be used to illustrate the inner functioning of a model via graphical tools. For instance, heat-maps can highlight specific areas of an image or specific words of a text that mostly influence the inferential process of a model by using different colours [134] [135]. Similarly, a graphical representation can be employed to represent the inner structure of a model, such as the graphs proposed in [136] where each node is a layer of the network and the edges the connections between layers. Another intuitive form of explanation for humans are *textual explanations*, natural language statements that can be either written or orally uttered. An example is the phrase “This is a Brewer Blackbird because this is a blackbird with a white eye and long pointy black beak” shown by an explainer of an image classification model [137]. A schematic, logical format, more structured than visual and textual explanations but still intuitive for humans, are *rules* that can be used to explain the inferences produced by models induced from data. Rules can be in the form of ‘IF ... THEN’ statements with *AND/OR* operators and they are very useful for expressing combinations of input features and their activation values [138] [139]. Technically, rules of these type employ symbolic logic, a formalized system of primitive symbols and their combinations (example: ‘ $(Country = USA) \wedge (28 < Age \leq 37) \rightarrow (Salary > 50K)$ ’ [140]). The parts before and after the  $\rightarrow$  logical operator are respectively referred to as antecedent and consequent. Given this logic, rules can be implemented as fuzzy rules, linking one or more premises to a consequent that can be true to a degree, instead of being entirely true or false. This can be obtained by representing each antecedent and consequent as fuzzy sets [43]. Combining fuzzy rules with learning algorithms can become a powerful tool to perform reasoning and, for instance, explain the inner logic of neural networks [141]. Similarly, the combination of antecedents and consequent can be seen as an argument in the discipline of argumentation, and a set of arguments can

**anyone else  
bothered by  
arity=1  
explanations?**

be put together in a dialogical structure by employing attacks, the link between arguments that model conflictuality [142] [143]. Arguments and attacks form a complex structure but with high explanatory power, suitable for explaining the inner functioning of data-driven models. Explanations can also be constructed by only employing numerical formats as crisp values, vectors of numbers, matrices or tensors as in Probe [144] and Concept Activation Vectors (CAVs) [145], two methods for explainability. A Probe consists of a linear classifier fitted to the features, treated independently, learned by each layer of a neural network. Probes are engineered to better understand the roles and dynamics of the internal layers. The numerical explanations are the probability scores assigned by the probes to each class [144]. CAVs separates the activation values of a neural network's hidden layer relative to instances belonging to a class, forming a set, from those generated by the remaining part of the input dataset, forming a second set. Subsequently, a binary linear classifier is trained to distinguish the activation values of the two sets. Then, CAVs computes directional derivatives on this classifier to measure the sensitivity of the model to changes in inputs towards the class of interest. This is a scalar quantity, calculated for each class over the whole dataset, which quantifies how important a user-defined concept is to classify the input instances in the class under analysis. For example, CAVs measures how sensitive the class 'zebra' is to the presence of stripes in an input image. Eventually, the most powerful format of explanations are those that employ one or more of the formats described so far (visual, textual, rules, numeric). An example of a combination of visual and numerical explanation is utilized by Important Support Vectors and Border Classification [146] that provide insight into local classifications produced by a Support Vector Machine (SVM). The former method returns the support vectors which influence the most the final classification for a particular instance. The latter determines which features of a data point would need to be altered (and by how much) to be placed on the separating surface between two classes. The explanations are in the form of an interactive interface where the user can select a point and the tool shows the attributes that had the largest effect on classifying it and the closest border value. The user can modify the selected point's attributes to see how the SVM reclassifies it. Image Caption Generation with Attention Mechanism [147] is an example of visual and textual explanations jointly employed. It returns attention maps for a combination of a Convolutional Neural Network (CNN) and a Long-Short Term Memory (LSTM) network where the CNN performs object recognition in images and the LSTM generates their captions.

## 6.2. Model agnostic methods for explainability

Several methods for explainability were designed to work with any learning technique. However, this does not mean that they can be universally applied as they might be constrained to the types of inputs of the technical problem they try to solve and the explanation they try to provide.

### 6.2.1. Numeric explanations

A few model agnostic methods for explainability produce numerical explanations (see table A.3 and figure 5). Most of them focus on measuring the contribution of an input variable (or a group of them) with quantitative metrics. Distill-and-Compare [148] trains a transparent, simpler model, called student, on the output obtained from a large, complex model, considered as a teacher, to mimic its inferential process. In this study, the student model was constrained to be GAMs which allow to easily assess the contribution of each feature in a numerical format. Similarly, SHapley Additive exPlanations (SHAP) [149] utilizes additive feature attribution methods, basically linear combinations of the input features, to build a model which is an interpretable

approximation of the original model. Some methods for explainability are based on an ‘input perturbation’ approach and, generally speaking, they work by modifying the reported values of the variables of an input instance to cause a change in the model’s prediction. Explain and Ime [150] [151] assess respectively the contribution of a particular input variable or a set of variables. This is done by replacing the actual values of the variables describing each input instance with other values sampled from the same variable(s) and measuring the differences in the output probability scores. The assumption is that the larger the difference in the outcome, the more relevant the variable is for the prediction process. Similarly, the Global Sensitivity Analysis (GSA) method [152] [153] ranks input features by quantifying the effects on the predictions of a given model when they are varied through their range of values. [154] [155] [156] [157] [158] proposed a method to explain the prediction of a model at instance level also based on the contribution of each feature estimated by comparing the model output when all the features are known and when one or more of them are omitted. The contribution is positive for the features that lead to the prediction towards a class, negative for those that push the prediction against a class and zero when they don’t have influence. Four methods, Quantitative Input Influence (QII) functions [159], Gradient Feature Auditing (GFA) [160], Influence functions [161] and Monotone Influence Measures [162], utilize influence functions to assess the contribution of each feature to certain predictions. An influence function is a classic technique from statistics [161] measuring the sensitivity of a model to changes in the distributions of the independent variables. The perturbation of the input can be done in different ways such as applying a constant shift (Influence functions [161]), obscuring parts of the input (GFA [160]), rotating, reflecting or randomly assign labels to the input (Monotone Influence Measures [162]). Feature Importance [163] and Feature Perturbation [164] are also based on algorithms that modify subsets of the input features to find groups of interacting attributes used by different classifiers and to determine the extent to which a model exploits such interactions.

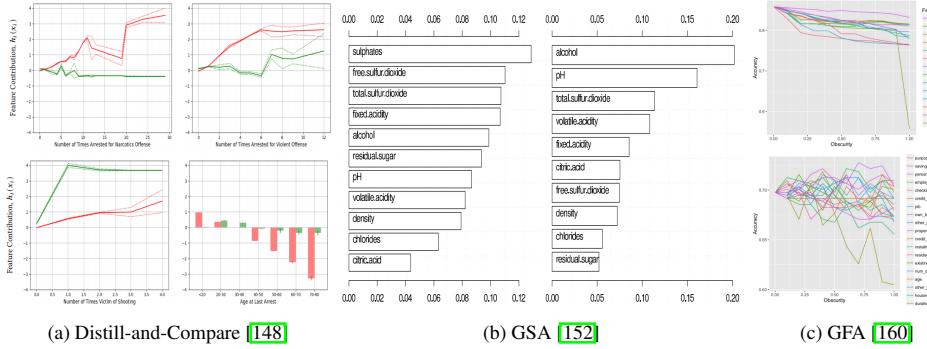


Figure 5: Examples of numerical explanations generated by model-agnostic methods for explainability.

**VEER:**  
problem with  
multi-tree  
multi-goal  
explanation

<https://arxiv.org/pdf/2106.02716.pdf>

### 6.2.2. Rule-based explanations

A few model-agnostic methods for explainability produce rule-based explanations by exploiting several rule-extraction techniques (see table A.5 and figure 6), such as automated reasoning-based approaches. The method presented in [165] extracts logical formulas as decision trees by combining split predicates along paths from inputs to predictions into logical conjunctions and all the paths related to an output class into logical disjunctions. These rules can be analyzed with logical reasoning techniques to extract information about the decision-making process. Similarly, Genetic Rule EXtraction (G-REX) [166, 167] employed genetic algorithms to generate IF-THEN rules with AND/OR operators. Anchor [140] uses two algorithms to extract IF-THEN rules which highlight the features of an input instance, called ‘anchors’, that are sufficient for a classifier to make a prediction. In an analogical manner, the words “not bad” are often used in sentences expressing a positive sentiment, and thus can be considered anchors in sentiment analyses. These two algorithms, a bottom-up formation of and a beam-search for anchors, identify the candidate rules with the highest estimated precision over a dataset where precision is equal to the fraction of correct predictions. The first algorithm starts from an empty set of rules and adds, at each iteration, a rule for each feature predicate. The second one instead starts from a set containing all the possible candidate rules and then selects the best ones in terms of precision. Model Extraction [168] and Partition Aware Local Model (PALM) [169] utilize decision trees (DTs) to approximate complex models with the assumption that, as long as the approximation quality is good, the statistical properties of the complex model are reflected in the interpretable ones. End-users have also the faculty to examine the DT’s structure and determine whether the rules match intuition. Model Extraction generates DTs by using the Classification And Regression Trees algorithm (CART) and trains them over a mixture of Gaussian distributions fitted to the input data using expectation maximization. PALM uses a two-part surrogate model: a meta-model, constrained to be a DT, that partitions the training data, and a set of sub-models fitting the patterns within each partition.

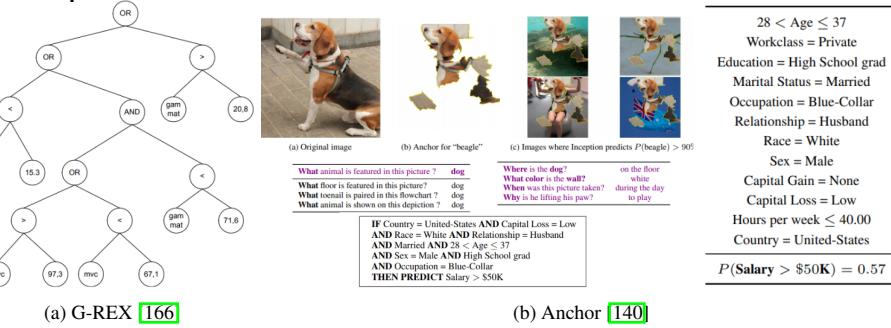


Figure 6: Examples of rule-based explanations generated by model-agnostic methods which can be visualized as (a) a decision tree (b) a list of rules accompanied by textual and visual examples.

### 6.2.3. Visual explanations

Visual explanations try to explain the inner functioning of a model via graphical aids and many model-agnostic methods exploit them (table A.6 and figure 7). One of the most widely used among these aids is represented by ‘salient masks’ that are efficient ways to point out what

parts of input, especially when images or texts are treated, most affect a model’s prediction by superimposing a mask highlighting them. Layer-Wise Relevance Propagation (LRP) [101] was developed as a model-agnostic solution to the problem of understanding image classification predictions by pixel-wise decomposition of nonlinear classifiers. In its general form, LRP assumes that the classifier can be decomposed into several layers of computation and it traces back contributions of each pixel to the final output, layer by layer, to attribute relevance to individual inputs. The pixel contributions can be visualized as heat-maps. Spectral Relevance Analysis (SpRAY) [8] consists of spectral clustering on a set of LRP explanations in order to identify typical and atypical decision behaviours of an underlying data-driven model. For example, to analyse the inferential process of a classifier trained on a dataset of images of animals, SpRAY produces an LRP heat-map for each image. Then, it checks if the heat-maps highlight the area representing the animal or if, for a specific animal, the classifier is focusing on other parts, such as the presence of a rider in case the animal is a horse. Image Perturbation [170] produces explanations in the forms of saliency maps by blurring different areas of the image and checking which ones most affect the prediction accuracy when perturbed. Similarly, the Restricted Support Region Set (RSRS) Detection method [171] visualizes a set of size-restricted and non-overlapping regions of an image that are critical to classification. This means that if any of them is removed, then the image is wrongly classified. The explanation consists of the original image with its critical regions determined by RSRS greyed out. The IVisClassifier [172] is based on linear discriminant analysis (LDA). It attempts at reducing the dimension of the input data and produces heat-maps that gives an overview of the relationship among clusters in terms of pairwise distances between cluster centroids both in the original and reduced dimensional spaces. The Saliency Detection method [173] utilizes a U-Net neural network trained to generate a saliency map, in a single forward pass, for any image and classifier received as inputs. The output map then highlights the parts of the image that are considered important by the classifier.

Some methods use other visual aids, like graphs and scatter-plots, to generate visual explanations. The Sensitivity Analysis method [174] generates explanations that correspond to local gradients. These gradients indicate how a data point must be moved to change its predicted label. The explanations can be either scatter-plots of the gradient vectors or heat-maps showing which parts of the inputs must be modified to change the predicted class. Individual Conditional Expectation (ICE) plots [175] are line charts graphing the functional relationship between a predicted response and a feature for each individual observation when keeping all the other features fixed and varying the value of the feature under analysis. [176] proposed two alternatives to ICE plots, called Partial Importance (PI) and Individual Conditional Importance (ICI) plots, which visualize the feature importance rather than its prediction. Both plots are aimed at showing how changes in a feature affect model performance. PI works at the global level by visualizing the point-wise average of all ICI curves across all observations, whereas ICI works at the local level by presenting changes for each observation. The importance of each feature is assessed using the Shapley Feature Importance measure which fairly distributes the model’s performance among them according to their marginal contribution. Explanation Graph [177] is based on the perturbations of the input features. It works by training a model on both the original and the perturbed data. Subsequently, a comparison of the original and perturbed input-output pairs is performed to infer causal dependencies between input and output. This method was tested across several word sequence generation tasks in Natural Language Processing (NLP) applications. The perturbed input contains statements that are semantically similar to the originals but differ in some elements (words and punctuation) and their order. The inferred dependencies are shown in graphs where

the nodes contain the words of the original and perturbed inputs and their relative outputs and the edges represent the connections between them. A Worst-Case Perturbation [178] corresponds instead to the smallest perturbation such that the perturbed input leads to an incorrect answer with high confidence. This method was applied only to images and the explanation consists of the perturbed images. Class Signatures [179] is a visual analytic interface that allows end-users to detect and interpret input-output relationships by presenting a mix of charts (line, bar charts and scatter plots) and tables organised in such a way that relationships become evident. Similarly, ExplainD [180] was designed to explain predictions made by classifiers that use additive evidence, such as linear SVMs and regressors. The graphs produced by this method represent the contribution of each feature to the prediction and how the prediction changes when the value of a feature varies across their value ranges. Manifold [181] and MLCube Explorer [182] are two visual analytical tools that provide comparative analysis for multiple models. They also enable end-users to define instance subsets using feature conditions, to identify instances that generate erroneous results so to explain potential reasons of these errors, and to iteratively refine the performance of a model by using different graphical aids such as scatter-plots, bar and line charts.

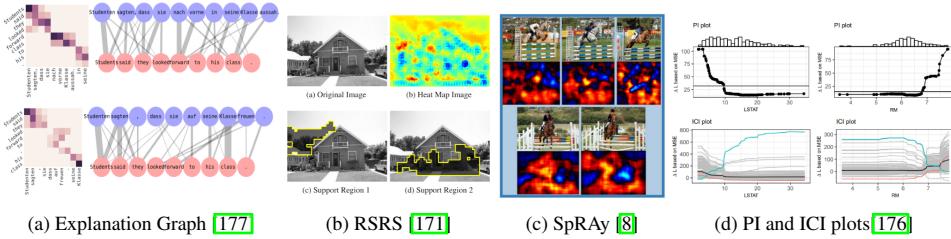


Figure 7: Examples of visual explanations generated by model-agnostic methods as (a) graphs, (b) restricted support regions, (c) heat-maps, or (e) plots.

#### 6.2.4. Mixed explanations

There are many methods for explainability that produce numerical explanations along with graphical representations to make them more interpretable for lay people (see table A.4 and figure 8). The Functional ANOVA decomposition [183] quantifies the influence of non-additive interactions within any set of input variables and depict them with Variable Interaction Network (VIN) graphs where the nodes represent the variables and the edges the interactions. The Justification Narratives method for explainability [184] consists of a simple model-agnostic mapping of the essential values underlying a classification (identified with any feature selection method) to a semantic space that automatically produces these narratives and realizes them visually (as bar-charts reporting the assessed relevance value of each variable) or textually. ExplAIner [133] and Rivelio [185] are two user interfaces showing mixes of numerical, visual and textual explanations. ExplAIner was designed to display visual and textual explanations of ML models which are the outcome of an iterative workflow of three stages: model understanding, diagnosis, and refinement. Using TensorBoard (a visualization tool developed by Google for machine learning) as a starting point, ExplAIner produces an interactive graph view of the model to be explained. The nodes of the graph represent the model’s components, such as inputs, parameters and outputs, accompanied by textual definitions, and the edges represent the relationships between the

components. There are also other visual explanatory tools in support of the model’s graph, such as line-charts of metrics, like loss and accuracy, and examples of input data together with their relative heat-maps generated with other visual methods for explainability. Rivelio works exclusively with binary classification problems and binary input features. It enables end-users to understand the causes behind predictions by interactively exploring a set of visual and textual instance-level explanations which lists the most relevant input features (words or image areas in a document/image), their frequency, number of instances with the feature with positive labels and are correctly/wrongly classified.

Other mixed explanations-based methods utilize a selection of prototypes, which are samples from the input that are correctly predicted by the model and can be considered as positive and iconic examples, or adversarial examples, which are samples misrepresented by the model and are used to generate contrastive explanations (see Section 5.1). This subset helps end-users understand the model by leveraging on the human ability to induce principles from a few examples. Being a subset of a training dataset, these explanations were classified as mixed as their format depends on the nature of the input data. The Bayesian Teaching methods for explainability [186] selects a small subset of prototypes that would lead the model to the correct inference as if trained on the overall dataset. [187] proposed to use Sequential Bayesian Quadrature (SBQ) in conjunction with Fisher kernels to select salient training data points. All the instances in a training dataset are firstly embedded in the space induced by the Fisher kernels. This provides a way to quantify the closeness of pairs of instances which, if close enough, should be treated similarly by a model. The embedded instances are inputted into SBQ, an importance-sampling-based algorithm that estimates the expected value of a function under a distribution using discrete samples drawn from it. Set Cover Optimization (SCO) [188] aims at selecting prototypes in such a way that they capture the full structure of the training examples in each class of the dataset, no points have a prototype of a different class in its neighbourhood and the prototypes are as few as possible. This leads to a set cover optimization problem that can be solved approximately with standard approaches such as, for instance, ‘linear program relaxation with randomized rounding’. Neighbourhood-Based Explanations [189] is based on a Case-Based Reasoning (CBR) approach. It presents to end-users the entries of a training dataset that are the most similar to the new input instance that needs to be explained. Similarity is measured through the Euclidean metrics applied to all the input features. Adversarial examples are instead used in Evasion-Prone Samples Selection [190], Maximum Mean Discrepancy (MMD)-critic [191] and Pertinent Negatives [192]. Evasion-Prone Samples Selection aims at detecting the instances closed to the classification boundaries that can be easily misclassified if slightly perturbed whereas MMD-critic utilizes the maximum mean discrepancy and an associated witness function to identify the portions of the input space most misrepresented by the underlying model. Pertinent Negatives highlights what should be minimally and necessarily absent to justify the classification of an instance. For example, the absence of glasses is a necessary condition to say if a person has a good sight. The input data are modified by removing some parts and the pertinent negatives are identified as those perturbations that maximise the prediction accuracy. eventually, some methods for explainability produce mixed explanations by approximating a black-box model with simpler, more comprehensible models that the end-users can inspect to assess the contribution of each feature. Local Interpretable Model-Agnostic Explanations (LIME) [193, 134] explains the prediction of any classifiers by learning a local self-interpretable model (such as linear models or decision trees), sometimes referred to as ‘white-box’ modes, trained on a new dataset which contains interpretable representations of the original data. These representations can be the binary vectors

representing the presence or absence of certain characteristics, such as words in texts or super-pixels (contiguous patch of similar pixels) in images. The black-box model can be explained through the weights of the white-box estimator which does not need to fully work globally, but it should approximate the black-box well in the vicinity of a single instance. However, the authors proposed the Sub-modular Pick (SP-LIME) to select, from an original dataset, a representative non-redundant explanation set of instances that is a global representation of the model.

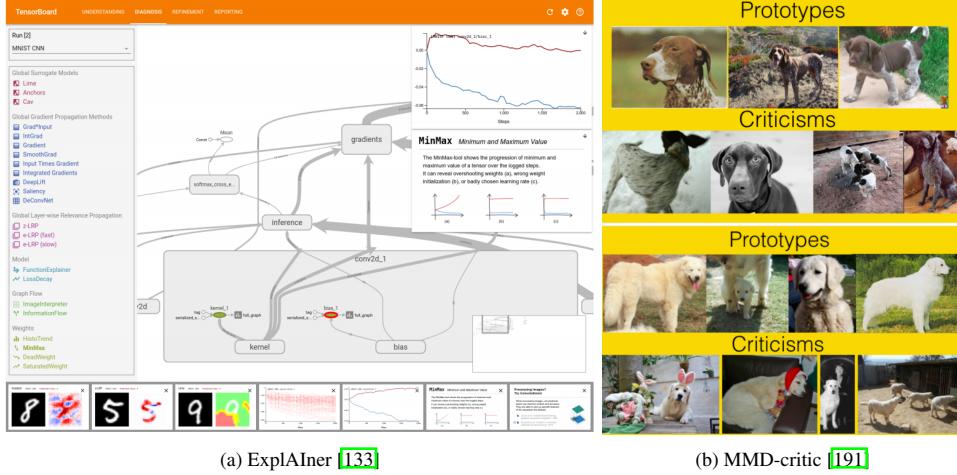


Figure 8: Examples of mixed explanations generated by model-agnostic methods for explainability which consists of a combination of visual and textual explanations in (a) interactive interfaces or (c) a selection of prototypes from inputs.

### 6.3. Model-specific methods for explainability based on neural networks

A considerable portion of the reviewed scientific articles about new methods for explainability is focused on interpreting deep neural networks (DNNs). This is not surprising giving the momentum of Deep Learning. Most of these methods produce visual explanations (table A.7), mostly in the form of salient masks and scatter-plots (figure 9), some as other visual aids (figure 10), rules (table A.8 and figure 11), textual and numerical explanations (table A.9 and figure 12) or a combination of them (table A.10 and figure 13).

#### 6.3.1. Visual explanations as salient masks

Class-Enhanced Attentive Response (CLEAR) [194] produces attention maps for image classification applications by back-propagating the activation values of the output layer. CLEAR was designed to return the attentive regions responsible for the prediction, along with their attentive levels to understand their influence and the dominant output class associated with these regions. DeepResolve [195] and GradCam [196] are two gradient ascent-based methods. DeepResolve computes and visualizes intermediate layer feature maps that summarize how a network combines elemental layer-specific features to predict a specific class. GradCam instead uses the gradients of any target concept (say ‘dog’ for instance) flowing into the final convolutional layer to generate a heat-map highlighting the influential regions in the image for predicting that concept. Heat-maps are generated by the last convolutional layer because the fully-connected layers

do not retain spatial information and it is expected that it has the best compromise between high-level semantics and detailed spatial information. Stacking with Auxiliary Features (SWAF) [197] utilizes heat-maps generated by GradCam to interpret and improve stacked ensembles for visual question answering (VQA) tasks. VQA includes answering a natural language question about the content of an image by returning, usually, a word or phrase or, in this case, a heat-map highlighting the relevant regions for a prediction. Guided BackProp and Occlusion [198] find what part of an input (pixels in images or words in questions) the VQA model focuses on while answering the question. Guided BackProp is another gradient-based technique to visualize the activation values of neurons in different layers of CNNs. It computes the gradients of the probability scores of predicted classes but restricts negative gradients from flowing back towards the input layer, resulting in sharper images showcasing the activation. Occlusion consists of masking, or occluding, subsets of an input (either a region of the image or a word of the question), then forward propagating it through the VQA model and computing the change in the probability of the answer predicted with the original input. A similar method, Occlusion Sensitivity [199], maps those features considered relevant in the intermediate layers of a DNN, by projecting the top nine activation values of each layer down to the input pixel space and masking the rest of the image. Net2Vec [200] maps instead semantic concepts to corresponding individual DNN filter responses. It returns images that are entirely greyed out except in the region related to a semantic concept, such as for instance the area representing a door of a building. The pixels of this region generate activation values that are above a threshold, corresponding to the 99.5th percentile of the distribution of all the activation values. Inverting Representations [201] inverts the representations of images produced by the inner layers and projects them on the input image as heat-maps. A representation can be thought of as a function of the image that characterise the image information. By reconstructing an approximate inverse function, it should be possible to reproduce the representations built by the layers. This method is based on the hypothesis that the layers consider only the relevant features and discard the irrelevant differences between images (such as, for instance, illumination or viewpoint) and consists of a reconstruction problem solved by optimizing an objective function with gradient descent.

Similarly, Guided Feature Inversion [202] generates an inversion image representation consisting of the weighted sum between the original image and another noisy background image, such as a grey-scale image with each pixel set to an average colour, a Gaussian white noise or a blurred image. The weights are calculated in such a way to highlight the smallest area that contains the most relevant features and to blur out everything else, especially things that might lead to an erroneous prediction, like objects belonging to other classes. SmoothGrad [203] was designed to sharpen in two ways gradient-based sensitivity maps, which are often visually noisy as they highlight pixels that, to a human, seem randomly selected. The first approach considers an image of interest along with sample similar images. The second approach generates a perturbed version of the image of interest by adding Gaussian white noise. Both approaches generate individual saliency maps with other methods for explainability such as GradCam, for instance, and take the average of the resulting maps. Deep Learning Important FeaTures (DeepLIFT) [100] computes the importance scores of features based on the difference between the activation of each neuron to a ‘reference activation’ value, computed by propagating a ‘reference input’ through the network. This represents a default or neutral input, such as a white image, chosen according to the problem at hand. According to the authors, this difference-from-reference approach has two advantages over the other methods producing saliency maps: (I) it can propagate importance signals even when the gradient is zero, avoiding artifacts caused by discontinuities in the gradient and (II)

it can reveal dependencies missed by other approaches because it can separately consider the effects of positive and negative contributions. Thus, the saliency maps produced by DeepLIFT contains all and only the important features that support or go against a certain prediction. Similarly, Integrated Gradients [93] attributes the prediction of a DNN to specific parts of the input. The attribution is measured as the cumulative sum of the gradients of the classification function representing the network calculated at all points along the straight-line path from a baseline input (a black image or an empty text, for example) to a specific input instance.

Feature Maps [204] and Prediction Difference Analysis [205] produce respectively feature- and heat-maps highlighting areas in an input image that gives evidence for or against a predicted class. Feature Maps utilizes a loss function that pushes each filter in a convolutional layer to encode a distinct and unique object part, exclusive of the object class under analysis. Prediction Difference Analysis instead is based on Explain [150], which was designed to evaluate the contribution of a feature at a time. In this case, a feature should correspond to a pixel of the image, but the authors proposed to consider patches of pixels. The assumption is that the value of each pixel is highly dependent on the surrounding pixels. The patches are overlapping so that, ultimately, an individual pixel's relevance is calculated as the average relevance of the different patches it was in. Two studies proposed variations of LRP, namely LRP with Relevance Conservation [206] and LRP with Local Renormalization Layers [207]. LRP was used in conjunction with the Pixel-wise Decomposition methods for explaining the automated image classification process of neural networks [101]. In both studies, the authors wanted to extend LRP to DNNs with non-linearities, such as LSTM models that have multiplicative interactions within their architecture [206] or networks with local renormalization layers [207]. [206] proposed a strategy to back-propagate the relevance of the neurons in the output layer back to the input layer through the two-way multiplicative interactions between lower-layer neurons of the LSTM. The algorithm sets to zero the relevance related to the gate neuron and propagate the relevance of the source neuron only. The extension of LRP proposed in [207] is based on first-Taylor expansion for non-linearities in the renormalization layers. [98] proposed to generate saliency maps by computing the first-order Taylor expansion of the function that links each pixel of an input image to the function, representing the neural network, that assigns a probability score to each output class.

Similarly, [208] analysed the use of Taylor decomposition for interpreting generic multi-layer DNNs by decomposing the network's output classification into the contributions of its input elements and back-propagating them from the output to the input layer, which are then visualized as heat-maps. Receptive Fields [209] focused on visualizing the input patterns, called precisely receptive fields, that are most strongly related to individual neurons by reconstructing these from the highest activation values of each layer. PatternNet and PatternAttribution [210] aim at measuring the contribution of the input 'signal' dimension, which is the part of the input that contains information about the output class, to the prediction as well as how good the network is at filtering out the 'distractor', which is the rest of the input (like the image background). PatterNet yields a layer-wise back-projection of the estimated signal to the input space whereas PatternAttribution produces explanations consisting of neuron-wise contributions of the estimated signal to the classification scores. Relevant Features Selection [211] automatically identifies the relevant internal features of a neural network via a two-step algorithm. First, a set of relevant layer/filter pairs are identified for every class of interest by finding those pairs that reduce at the minimum the differences between the predicted and the actual labels. This results in a relevance weight for

every filter-wise response computed internally by the network. Then, an image is pushed through the network producing the class prediction and it generates a heat-map by taking into account the internal responses and relevance weights for the predicted class. A combination of a Neural Network and Case Base Reasoning (CBR) Twin-systems was proposed in [212]. This method maps the features' weights from the DNN to the CBR system to find similar cases from a training dataset that explain the prediction of the network of a new instance. To extract the weights of features, the authors proposed the Contributions Oriented Local Explanations (COLE) technique which is based on the premise that the feature contributions to the model's predictions are the most sensible basis to inform CBR explanations. COLE uses saliency maps methods, such as LRP and DeepLift, to estimate these contributions. This was tested on image classification problems with explanations generated in the form of similar images whose discriminating features were highlighted by saliency maps. Compositionality [213] consists of building the meaning of a sentence from the meanings of single words and phrases. This method is designed for visualizing compositionality in neural models trained for NLP tasks by plotting the salience value of each word as saliency maps. The salience values indicate the contribution of the words to the sentence meaning. For instance, the word 'hate' and 'boring' in the phrase 'I hate the movie because the plot is boring' can be considered the two most relevant ones in a sentiment analysis problem. The OpenBox method [214] computes exact and consistent interpretations for the family of Piecewise Linear Neural Networks (PLNN) by transforming them into a mathematically equivalent set of linear classifiers. Subsequently, each linear classifier is interpreted by the features that dominate its prediction and the decision boundaries of each feature can be determined and visualized as scatter-plots (for numeric inputs) or heat-maps (for images).

#### 6.3.2. Visual explanations as scatter-plots

The Convolutional Neural Network Interpretation method (Cnn-Inte) [215] uses a two-level k-means clustering algorithm to split into clusters the activation values of the neurons of hidden layers relative to each input feature. Clusters might contain the activation values of instances belonging to different classes. A random forest algorithm is then trained on each cluster. The results are visually displayed using scatter plots to show how a specific test instance is classified. [216] instead presented a method based on Principal Component Analysis (PCA) for analyzing the variation of features generated by CNNs to scene factors that occur in images such as object style, colour and lighting configuration. It analyzes CNN feature responses (or activation values) in the different layers by decomposing them as a linear combination of uncorrelated components associated to the different factors of variation and visualizing them into scatter-plots by using PCA. t-Distributed Stochastic Neighbor Embedding (t-SNE) maps [217] analyzes Deep Q-networks (DQNs) in reinforcement learning applications, in particular for agents that autonomously learn, for instance how to play video-games. This method extracts the neural activation values of the last DQN layer and apply t-SNE for dimensionality reduction and for generating cluster plots where each dot correspond to a particular learning phase. Similarly, Hidden Activity Visualization [218] uses t-SNE to visualize the projections of the activation values of the hidden neurons as a 2D scatter-plot with points coloured according to the class of the instances originating them. The distribution of the points in the scatter-plot gives a graphical representation of the data distribution, relationships between neurons and the presence of clusters in the activation values. Finally, TreeView [219] consists of a scatter plot representation of a DNN via hierarchical partitioning of the feature space. Features are clustered according to the activation values of the hidden neurons in such a way that each cluster comprised of a set of neurons with similar distribution of activation values across the whole training set.

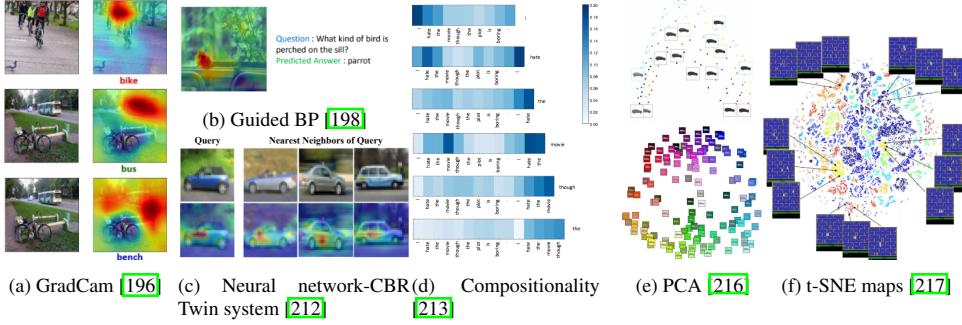


Figure 9: Examples of visual explanations, as salient masks (a-d) and scatter-plots (e-f).

### 6.3.3. Visual explanations - miscellaneous

A few other methods use alternative visualization tools. Generative Adversarial Network (GAN) Dissection [220] was designed to understand the inferential process of GANs at different levels of abstraction, from each neuron to each object, and the relationship between objects, by identifying units (or groups of units) that are related to semantic classes (doors, for example). This method intervenes on them by adding or removing these objects from the image and observing how the GAN network reacts to these changes. These reactions are represented as a new version of the input image where other objects or areas of the background are modified. For instance, if a door is intentionally removed from a building, the GAN might substitute it with a window or bricks. The Important Neurons and Patches method [221] analyzes the predictions of a DNN in terms of its internal features by inspecting information flow through the network. For instance, given a trained network and a test image, important neurons are selected according to two metrics, both measured over a set of perturbed images (each pixel is multiplied by a Gaussian noise): (I) the magnitude of the correlation between the neuron activation and the network output which approximates the influence of each neuron on the output, and (II) the precision of the activation of a neuron, which estimates the generalizability of the feature(s) encoded by it, by selecting those neurons whose activation values were not significantly affected by the perturbations. Given a rank of neurons, the top N are selected and their related image patches are determined by using a multi-layered deconvolutional network and enclosed in bounding boxes applied to the input image. [222] and [223] [224] proposed two similar methods, based on Activation Maximization, which modify the input images in such a way to maximise the activation of a given hidden neuron with respect to each pixel. These modified images should provide a good representation of what a neuron is doing. [225] instead presented a method to generate Activation maps which show what features activate the neurons in the penultimate layers. It is based on the idea that the final prediction of a DNN is dominated by the most highly-weighted neuron activations of this layer. Shifting from pictorial to textual inputs, Cell Activation Values [226] is a method of explainability for LSTMs and uses character-level language models as an interpretable test-bed for understanding the long-range dependencies learned by LSTMs by highlighting sequences of relevant characters.

A group of methods that produce visual explanations in the form of graphs. The method proposed in [136] generates data-flow graphs to visualize the structure of DNNs created and trained in Tensorflow. Similarly, Explanatory Graph [227] produces graphs from CNNs where each node represents a ‘part pattern’, which correspond to the peak activation in a layer related to a part of the input, and each edge connects two nodes in adjacent layers to encode co-activation relationships and spatial relationships between patterns. [228] instead added to CNNs a new Symbolic Graph Reasoning (SGR) layer which performs reasoning over a group of symbolic nodes whose outputs explicitly represent different properties of each semantic in a prior knowledge graph. To cooperate with local convolutions, each SGR is constituted by three modules: a) a primal local-to-semantic voting module where the features of all symbolic nodes are generated by voting from local representations; b) a graph reasoning module that propagates information over the knowledge graph to achieve global semantic coherency; c) a dual semantic-to-local mapping module that learns new associations of the evolved symbolic nodes with local representations, and accordingly enhances local features. Lastly, And-Or Graph (AOG) [229] is a method to grow a semantic AOG on a pretrained CNN. An AOG is a graphical representation of the reduction of problems (or goals) to conjunctions (AND) and disjunctions (OR) of sub-problems (or sub-goals). The AOG is used for parsing the part of the input images which corresponds to a semantic concept and the output explanation consists of the input image where the semantic part is included in a bounding box. Many scholars studied ways to exploit the visual explanatory tools, described so far, to create interactive interfaces for the lay audience. For example, [230] studied the usage of saliency maps as the building blocks of interactive interfaces to explain the inferential logic of CNNs. ActiVis [231] is an interactive visualization system for DNNs that unifies instance- and subset-level inspection by using flowcharts that show how neurons are activated by user-specified instances or instance subsets. Deep Visualization Toolbox [232] is based on two visualization tools. The first one depicts the activation values produced, while processing an image or video, on every layer of a trained CNN as heat-maps. The second tool modifies the input images via regularised optimization methods to enable a better visualization of the learned features by individual neurons at every layer. Deep View (DV) [233] measures the evolution of a DNN by using two metrics that evaluate the class-wise discriminability of the neurons in the final layer and the output feature maps. iNNvestigate [234] compares different methods for explainability, namely PatternNet, PatternAttribution and LRP. LSTMVis [135] is a visual analysis tool for recurrent neural networks, LSTM in particular, that facilitates the understanding of their hidden state dynamics. It is based on a set of interactive graphs and heat-maps of relevant words. A user can select a range of text in the heat-maps, which results in the selection of a subset of hidden states visualized in a parallel coordinate plot where each state is a data item, time-steps are the coordinates, and the tool then matches this selection to similar patterns in the dataset for further statistical analysis. Seq2seq-Vis [235] is similar to LSTMVis but it focuses on sequence-to-sequence models, also known as encoder-decoder models, for automatic translation of texts. Seq2seq-Vis allows interactions with trained models through each stage of the translation process intending to identify the learned pattern, detect errors and probe the model with counterfactual scenarios. Finally, N<sup>2</sup>VIS [236] is an interactive visualization tool for feed-forward neural networks trained with evolutionary computation which allows end-users to adjust training parameters during adaptation and to immediately see the results of this interaction. It considers graphs representing the network topology, connection weights and activation levels for specific inputs and weight volatility to facilitate the process of understanding the inferential process of a neural network and to improve its performances in terms of efficiency and prediction accuracy.

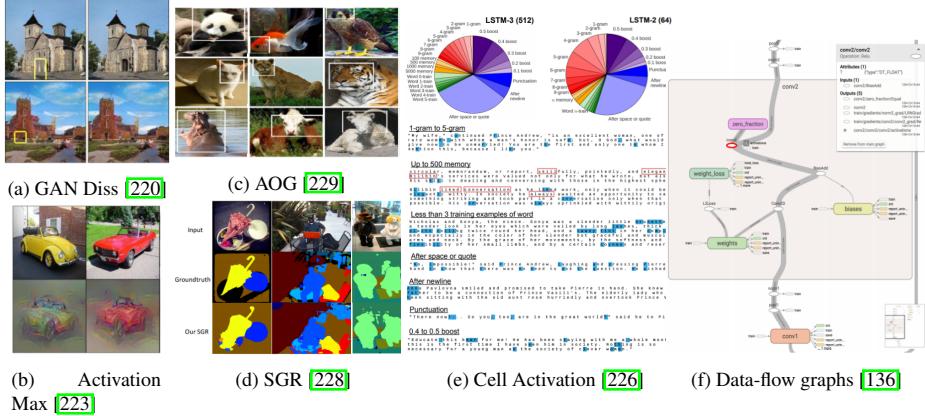


Figure 10: Examples of miscellaneous visual explanations generated by methods for explainability for neural networks.

#### 6.3.4. Rule-based explanations

Several methods for explainability are focused on rule-based explanations of the inferential process of neural networks, usually in the form of IF-THEN rules. Scholars divided these methods into three classes [41][132]: (I) *decompositional* methods work by extracting rules at the level of hidden and output neurons by analysing the values of their weights, (II) *pedagogical* methods treat an underlying neural network as a black-box and the rule extraction consists of mimicking the function computed by the network; weights are not subjected to analysis, and (III) *eclectic* methods that are a combination of the decompositional and pedagogical ones.

Regarding the decompositional methods, Discretizing Hidden Unit Activation Values by Clustering [237] generates IF-THEN rules by clustering the activation values of hidden neurons and replacing them with the cluster's average value. The rules are extracted by examining the possible combinations in the outputs of the discretised network. Similarly, Neural Network Knowledge eXtraction (NNKX) [238] produces binary decision trees from multi-layered feed-forward sigmoidal artificial neural networks by clustering the activation values of the last layer and propagating them back to the input to generate clusters. Interval Propagation [141] is an improved version of Validity Interval Analysis (VIA) [239] to extract IF-THEN crisp and fuzzy rules. VIA consists of finding a set of validity intervals for the activation range of each unit (or a subset of units) such that the activation values of a DNN lie within these intervals. The precondition of each extracted rule is given by a set of validity intervals and the output is a single target class. According to [141], VIA has two shortcomings: it fails sometimes to decide whether a rule is compatible or not with the network and the intervals are not always optimal. Interval Propagation overcomes these limitations by setting intervals to either the input or output and feed- or back-propagating them through the network. However, this method has still a drawback. Some neural networks require a big number of crisp rules to be approximated and to reach similar performances in terms of prediction accuracy. Then, [141] proposed to compact these crisp rules into fuzzy rules by using a fuzzy interactive operator which introduces the OR operators between rules. Discretized Interpretable Multi-Layer Perceptrons (DIMLP) [139, 240, 241] [132] returns symbolic rules from Interpretable Multi-Layer Perceptrons (IMLP) which are CNNs where each

neuron of the first hidden layer is connected to only an input neuron and its activation function is a step function while the remaining hidden layers are fully connected with a sigmoid activation function. In DIMPL, the step activation function becomes a staircase function that approximates the sigmoid one. The rule extraction is performed after a max-pool layer by determining the location of relevant discriminative hyperplanes, which are the boundaries between the output classes. Their relevance corresponds to the number of points passing through each hyperplane as they move to a different class. An example of a ruleset generated with DIMLP from a neural network with thirty neurons, represented as  $x_i$  with  $i = 1, \dots, 30$ , in a unique hidden layer and three output neurons is: Rule 1 -  $(\neg x_3) (\neg x_8) (x_{17} > 0.0061) (x_{19} < 0.151) (x_{21} > 0.065)$  *Class\_1*, Rule 2:  $(x_{17} > 0.0061) (x_{21} < 0.065)$  *Class\_2*, Default: *Class\_3*. Rule Extraction by Reverse Engineering (RxREN) [242] relies on a reverse engineering technique to trace back input neurons that cause the final result, whilst pruning the insignificant ones, and to determine the data ranges of each significant neuron in respective classes. The algorithm is recursive and generates hierarchical rules where conditions for discrete attributes are disjoint from the continuous ones. Rule Extraction from Neural Network using Classified and Misclassified data (RxNCM) [243] is a modification of RxREN. It incorporates also the input instances correctly classified in the range determination process, not only the misclassified ones as done by RxREN. Most of the rule-based methods for explainability are monotonic, that means they produce an increasing set of rules, thus the prepositions that can be derived. However, sometimes adding new rules might lead to the invalidation of some conclusion inferred by other rules, as in [244] where a method that captures non-monotonic symbolic rules coded in the network was presented. The rule extraction algorithm starts by partially ordering the vectors of a training dataset according to the activation values of the output layer. Then, it determines the minimum input point that activates an output neuron and creates a rule whose antecedents are based on the feature values of the selected instance. Thus, the expected set of rules has the following form:  $L_1, \dots, L_n, \sim L_{n+1}, \dots, \sim L_m \rightarrow L_{m+1}$  where  $L_i (1 \leq i \leq m)$  represents a neuron in the input layer,  $L_{m+1}$  represents a neuron in the output layer,  $\sim$  stands for default negation and  $\rightarrow$  means causal implication. Finally, [245] and [246] proposed two algorithms that extract DTs from the weights of a DNN. The former method produces a soft DT trained by stochastic gradient descent using the predictions of a neural network and its learned filters to make hierarchical decisions on where to split the data and how to create the paths from the root to the leaves. The latter, which is designed only for image classification tasks, aims at explaining an underlying CNN semantically, meaning that the nodes of the tree should correspond to parts of the objects that can be named. Nodes near the root should correspond to parts shared by many images (such as the presence of four legs in images showing animals) whereas the nodes close to the leaves should represent characteristics of minority images (a peculiar characteristic of each animal). To build such DTs, the network's filters are forced to represent object parts by a special modification of the loss function. The DT is then built on the part/filter pairs recursively on an image by image basis.

In regard to the pedagogical methods, Rule Extraction From Neural Network Ensemble (REFNE) [247] extracts symbolic rules from instances generated by neural network ensembles. The algorithm randomly selects a categorical attribute and checks if there is a value satisfying the condition that all the instances possessing such a value fall into the same class. If the condition is satisfied, a rule is created with the value as antecedent; otherwise, the algorithm selects another categorical attribute and examines all the combinations of the two attributes. When all the categorical attributes are analysed, continuous attributes are considered and the process terminates when no more rules can be created. Rules are limited to only three antecedents. Continuous

attributes are discretised and a fidelity evaluation mechanism checks that this process does not compromise the relationship between the attribute and the output classes. An alternative method to extract IF-THEN rules from neural network ensembles, called C4.5Rule-PANE [248], uses the C4.5 rule induction algorithm. After a neural network ensemble was trained on a dataset, the original labels of the training dataset are replaced by those predicted by the ensemble. Subsequently, C4.5Rule-PANE extracts a ruleset from the modified dataset to mimic the inferential process of the ensemble. The DecText method [249] also extracts high fidelity DTs from a DNN. It sorts input instances by increasing the value of a feature and split an input dataset by placing the cutpoint at the midpoint of the range. Then, four splitting algorithms check the partitions such created and choose the best ones according to four criteria. The first, SetZero, chooses the most discriminative features of the target variable. The other three, SSE, ClassDiff and Fidelity, respectively select the feature which maximizes the possibility that a single class dominating each partition is created, the quality of the partition and the fidelity between the DNN and the tree. According to the authors, these algorithms have comparable, if not better, prediction performance than a tree extraction technique based on entropy gain split. TREPAN [250][251] induces a DT that, like DecText, maintains a high level of fidelity to a DNN while being comprehensible and accurate. It queries an underlying network to determine the predicted class of each instance and selects the splits for each node of the tree by using the ‘gain ratio criterion’ and by considering the previously selected splits that lie on the path from the root to that node as constraints. Tree Regularization [252] consists of a penalty function of the parameters of a differentiable DNN which favours models whose decision boundaries can be approximated by small binary DTs. It finds a binary DT that accurately reproduces the network’s prediction and measures its complexity as the ‘average decision path length’. It then maps the parameter vector of each candidate network to an estimate of the average-path-length and chooses the shortest one. Word Importance Scores [253] visualizes the importance of specific inputs for determining the output of a LSTM. By searching for consistently important phrases and calculating their importance scores, the method extracts simple phrase patterns consisting of one to five words. To concretely validate these patterns, they are inputted to a rule-based classifier which approximates the performance of the original LSTM. Iterative Rule Knowledge Distillation [254] and Symbolic Logic Integration [255] are the only ante-hoc methods producing rule-based explanations for DNNs. The former combines DNNs with declarative first-logic rules to allow integrating human knowledge and intentions into the networks via an iterative distillation procedure that transfers the structured information of logic rules into the weights of DNNs. This is achieved by forcing the network to emulate the predictions of a rule-based teacher model and evolving both models throughout the training. The latter instead encodes symbolic knowledge in an unsupervised neural network by converting background knowledge, in the form of propositional IF-THEN rules and first-order logic formulas, into confidence rules which can be represented in a Restricted Boltzmann Machine.

#### 6.3.5. Textual and numerical explanations

Textual and numerical explanations are jointly less used than other methods for explainability. InterpNET [256] makes use of a DNN to generate natural language explanations of classifications done by external models and produces statements like “This is an image of a Red Winged Blackbird because...”. The explanations are built upon the activation values of this network. Most-Weighted-Path, Most-Weighted-Combination and Maximum-Frequency-Difference [257] are three methods for explainability that generates textual statements. Most-Weighted-Path starts from the output neuron and selects the corresponding input passing, layer-by-layer, via the

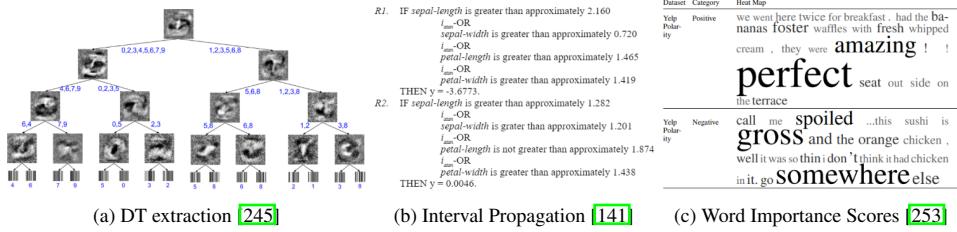


Figure 11: Examples of rule-based explanations generated by methods for explainability for neural networks and visualized as (a) decision trees, (b) list of rules or (c) by showing the most relevant input.

neuron connected with the highest weight. Then, it auto-generates a natural language explanation indicating the most relevant feature for predicting the output category. Most-Weighted-Combination works similarly, but it selects the two most-weighted input features. Maximum-Frequency-Difference retrieves from the training dataset the most similar cases for each instance, then perform the difference between the percentages of cases that have the same output and those with different output. The explanation is generated according to the input with the highest difference and it is a statement like “the smart kitchen estimates that you are sad because you are eating chocolate, which is 50% more frequent in this emotional state than people in other emotional states”. Rationales [258] is a method for justifying the predictions made by DNNs in natural language processing tasks, such as sentiment analysis, by extracting pieces of the input text as justifications, or rationales. These rationales must contain words that are interpretable and lead to the same prediction as to the entire input text. These words are selected via the combination of a ‘rationale generator’ function, which works as a tagging model assigning to each word a binary tag saying whether it must be included in the rationale, and an ‘encoder’ function that maps a string of words to a target class. Relevance and Discriminative Loss [259] [137] generates textual explanations for an image classifier like “The bird in the photo is a White Pelican because...”. It consists of a CNN that extracts visual features from the images, such as colours and object parts, and two LSTMs which produce a description of each image conditioned on visual features. The training process aims at reducing two loss functions, called respectively ‘Relevance’ and ‘Discriminative’, which assure that the generated sentences are both image relevant and category-specific.

A few methods for explainability produce pure numerical explanations. Concept Activation Vectors (CAVs) [145] separates the activation values of a hidden layer relative to the instances belonging to a class of interest from those generated by the remaining part of the dataset. Then it trains a binary linear classifier to distinguish the activation values of the two sets and compute directional derivatives on this classifier to measure the sensitivity of the model to changes in inputs towards the class of interest. This is a scalar quantity calculated over the whole dataset for each class. Probe [144] consists of a linear classifier fitted to a single feature learned by each layer of a DNN to predict the original classes. The numeric explanations consist of the predictions made by the Probes. Singular Vector Canonical Correlation Analysis (SVCCA) [260] returns the correlation matrix of the neurons’ activation vectors, calculated over the entire dataset, of two instances of a given DNN trained separately. The first network’s instance is obtained at the end of the training process, whereas the latter consists of multiple snapshots of the network during training. Every layer of the first instance is compared to every other layer

of the other instance to calculate correlation factors between pairs of layers. Causal Importance [261] is calculated by summing up the variations in the output when the values of a variable are perturbed instance by instance whilst all the other variables are kept fixed. Firstly, the algorithm suppresses the irrelevant variables by measuring their predictive importance via a metric based on the absolute difference between the predictions made by a DNN with the original and the perturbed variables. The network is then trained with the relevant variables only and data are clustered according to their hidden layer representation. This is done by training an unsupervised Kohonen map on the matrix containing the activation values of each pair neuron/input instance. Finally, causal importance is measured on a cluster by cluster basis, meaning that it is calculated only on the instances belonging to the cluster under analysis. [262] proposed to measure the Contextual Importance and Contextual Utility of input on the output variable. The former metric is the ratio between the range of the output values covered by varying a variable throughout its own range of values and the whole output space. For example, a neural network was trained to predict the price of a car over a set of variable, among which there is the engine size. By varying the engine size alone, the price varies only within a certain range. Contextual Utility represents the position of the actual output within Contextual Importance. So the price of cars with big engines, produced by the same manufacturer, are towards the upper end of the manufacturer's price range. The ideal values for these two metrics are domain-dependent. They are designed just to help end-users understand where each variable and instance lie within the input space, but they can be used to generate rule-based or textual explanations by structuring the domain in intermediate concepts which attach a positive or negative outlook to certain subsets of the output space. A certain price range can be deemed very good for a manufacturer and all its cars in that range can be considered as a deal. Finally, REcurrent LEXicon NETwork (RELEXNET) [263] combines the transparency of lexicon-based classifiers with the accuracy of recurrent neural networks. Lexicons are linguistic tools for classification and feature extraction which take the form of a list of terms weighted by their strength of association with a given class. RELEXNET uses lexicons as inputs of a naive gated recurrent neural network which returns the probability that the input belongs or not to a certain class.

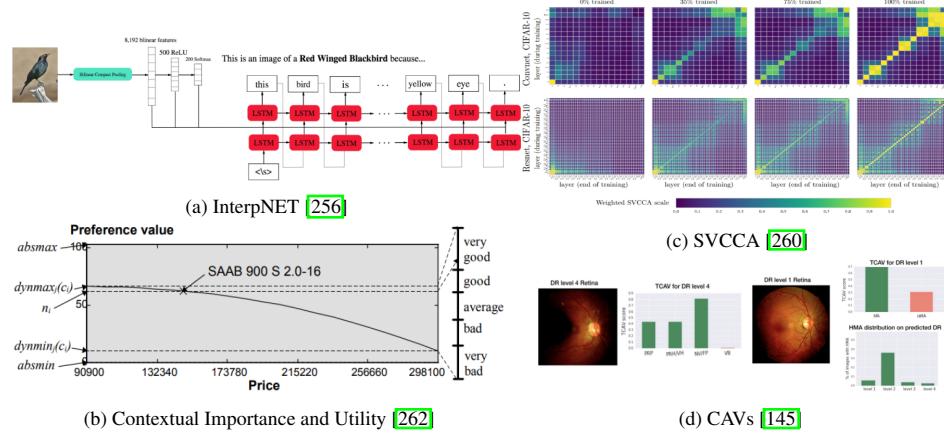


Figure 12: Examples of textual (a) and numerical (b-d) explanation generated by a method for explainability for neural networks.

### 6.3.6. Mixed explanations

To help end-user interpret a model, some scholars have proposed the use of multiple types of output formats for explanations. The Attention Alignment method [264] produces explanations in the form of attention maps. These maps highlight parts of a scene that mattered to a control DNN utilized in self-driving cars in combination with a perception DNN. The perception DNN combines the data received from cameras and other sensors, like radars and infrared, to ‘understand’ the environment and to generate manoeuvring commands like steering angles or braking. The control DNN is trained to identify the presence of specific objects, such as road signs, and obstacles like pedestrians and bikers, that influence the output of the perception network. Attention Alignment consists of an attention-based video-to-text algorithm that produces textual explanations of the model predictions such as “The car heads down the street because it is clear.” Similarly, Pointing and Justification Model (PJ-X) [265] and Image Caption Generation with Attention Mechanism [147] are two multi-modal methods for explainability, designed for VQA tasks, that provides joint textual rationale generation and attention-map visualization. The attention-maps are extracted from a CNN, which performs the object recognition in images, whereas the textual justifications are produced by an LSTM network as image captions. The word(s) in the caption related to the attention region is underlined. According to [265], the two explanations support each other in achieving high quality. The visual explanations help to generate better textual explanations which lead to better visual pointing. Image Caption Generation with Attention Mechanism is based on two algorithms: (I) a ‘soft’ deterministic attention mechanism trainable by standard back-propagation methods and (II) a ‘hard’ stochastic attention mechanism trainable by maximizing an approximate variational lower bound. The word(s) in the caption related to the attention region is/are underlined. [266] and [267] proposed two methods for replacing a DNN with a deterministic finite automaton that can be visualized as a graph where each node represent a cluster of values in the output space and the edges represent the presence of shared patterns in a network’s internal layers between these clusters. Lastly, the method in [268] uses prototypes to explain the prediction of a new instance. The prototypes are selected according to the activation values of hidden neurons related to training data. For a new observation, it is possible to foresee and justify its prediction by identifying the three most similar training samples based on cosine distance of their hidden activation values.

## 6.4. Model-specific methods for explainability related to rule-based systems

Explainable Artificial Intelligence was ignited by the interpretability problem of machine learning, in particular of the deep learning models. However, the problem of explainability exists even before the advent of neural networks. Many rules-based learning approaches already existed and the majority of these were interpreted with ante-hoc methods that act during the model training stage to make it naturally explainable (see table A.11 and figure 14). An Ant Colony Optimization (ACO) algorithm was proposed in [269] to create interpretable rules. It follows a sequential covering strategy, one-rule-at-a-time or also known as separate-and-conquer, to generate unordered sets of IF-THEN classification rules which can be inspected individually and independently from the others, thus they are easier to be interpreted. At each step, ACO creates a new unordered set of rules and compare it with those of previous iterations. If the new set contains fewer rules or has a better prediction accuracy, it replaces the previous one. Experiments run on thirty-two publicly available datasets showed that ACO gave the best results in terms of both predictive accuracy and model size, outperforming state-of-the-art rule

induction algorithms with statistically significant differences. Another method based on an ACO algorithm, called AntMinter+ [270], constructs monotonic rulesets by allowing the inclusion of domain knowledge via the definition of a directed acyclic graph representing the solution space. The nodes at the same depth in the graph represent the splitting values related to an input variable; the edges represent which values of the following variable can be reached from a node. AntMinter+ uses an iterative max-min ant system to construct a set of IF-THEN rules, starting from an empty set. A rule represents a path from the start to the end nodes of the graph. The algorithm stops adding rules when either a predefined percentage of training points is covered or when the addition of new rules does not improve the performance of the classifier. AntMinter+ can be combined with a non-linear SVM, in a method called Active Learning Based Approach (ALBA), to generate comprehensible and accurate rule-based models. Exception Directed Acyclic Graphs (EDAGs) [271] is an empirical induction tool that generates rules from the knowledge base of expert systems to create comprehensible knowledge structures in the form of graphs where nodes are premises, some of which have attached conclusions, leaves are conclusions and edges represent exceptions to some node. The ‘meaning’ of each node can be easily determined by following its path back to the root and by inspecting its child nodes, whilst the rest of the graph is irrelevant.

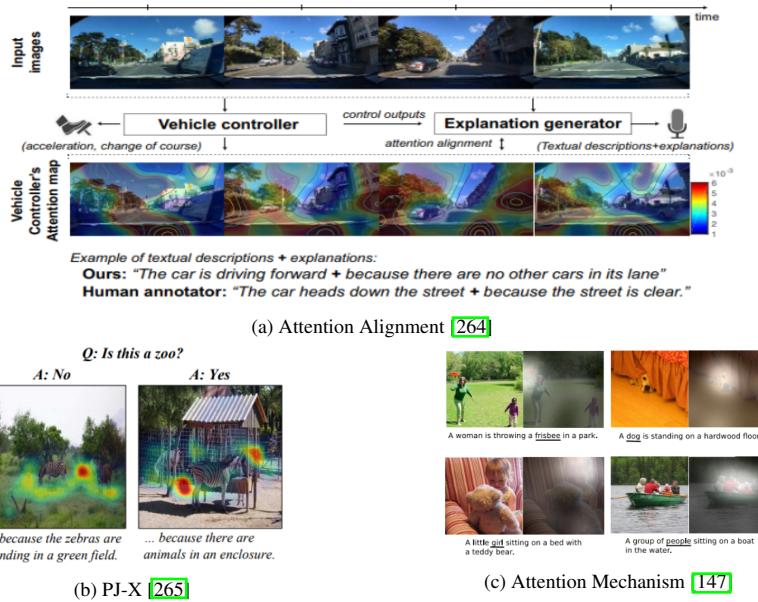


Figure 13: Examples of mixed explanations, consisting of combinations of images and texts, generated by the following methods for explainability for neural networks.

The Interpretable Decision Set [272] and the Bayesian Rule Lists (BRL) [273][274][275] are two methods that creates unordered sets of IF-THEN rules. Interpretable Decision Set is based on an objective function that simultaneously optimizes accuracy and interpretability by learning short and non-overlapping rules that cover the whole feature space and pay attention to small but important classes. BRL produces a posterior multinomial distribution over permutations of

rules, starting from a large set of possible rules, to assess the probability of predicting a certain label from the selected rules. The prior is the Dirichlet distribution and the permutation that maximises the posterior is included in the final decision set. The Bayesian Rule Sets (BRS) method [276] [277] is similar to BRL but it uses different probabilities, with the posterior as a Bernoulli distribution, and the prior a Beta distribution whose parameters can be adjusted by end-users to guide BRS toward more interpretable solutions by specifying the desired balance between size and length of rules. First Order Combined Learner (FOCL) [278] inductively constructs a set of rules in terms of predicates used to describe examples. Each clause body consists of a conjunction of predicates that cover some positive and no negative examples. The rules are displayed in a tree where the nodes are the predicates, the edge are the conjunctions and the leaves are the conclusions. Non-monotonic argumentation-based approaches for increasing explainability and dealing with conflictual information were proposed in [142] [143] [279]. They are based upon the concepts of defeasible arguments, in the form of rules, each composed by a set of premises, an inference rule and a conclusion as well as the notion of attacks between arguments to model conflictuality and the retraction of a final inference. Argumentation-based approaches are believed to have a higher explainability as the notions of arguments and conflictuality are common to the way human reason. Four methods based on fuzzy reasoning to generate interpretable sets of rules that clearly show the dependencies between inputs and outputs were presented in [280] [281] [282] [283]. Both [280] [283] examine the interpretability-accuracy tradeoff in fuzzy rule-based classifiers. A multiobjective fuzzy Genetics-Based Machine Learning (GBML) algorithm, analyzed by [280], is implemented in the framework of evolutionary multiobjective optimization (EMO) and consists of a hybrid version of Michigan and Pittsburgh approaches. Each fuzzy rule is represented by its antecedent fuzzy sets as an integer string of fixed length and the resulting fuzzy rule-based classifier, consisting of a set of fuzzy rules, is represented as a concatenated integer string of variable length. Multi-Objective Evolutionary Algorithms based Interpretable Fuzzy (MOEAIF) [283] consists instead of a fuzzy rule-based model engineered to classify gene expression data from microarray technologies. GBML and MOEAIF maximize the accuracy of rule sets, measured by the number of correctly classified training pattern, and minimize their complexity, measured by the number of fuzzy rules and/or the total number of antecedent conditions of fuzzy rules. The method in [281] is based on a five-step algorithm. First, it generates fuzzy rules that cover the extrema directly from data. Second, it checks rule similarity to delete the redundant and inconsistent rules. Third, it optimizes the rule structure using genetic algorithms based on a local performance index. Fourth, it performs further training of the rule parameters using gradient-based learning method and deletes the inactive rules. Last, it improves interpretability by using regularization. The method presented in [282] generates fuzzy rules by starting from a set of relations and properties, selected by an expert, of an input dataset. It then extracts the most relevant ones employing a frequent itemset mining algorithm. The authors do not provide a specific metric for evaluating the relevancy of a relation, but they suggest using “measures like the number of relations and properties in the antecedent or the value of their support”. Interpretable Classification Rule Mining (ICRM) [284] consists of a three-step evolutionary programming algorithm producing comprehensible IF-THEN classification rules, where comprehensibility is achieved by minimizing the number of rules and conditions. First, it creates a pool of rules composed of a single attribute-value comparison. Second, it utilizes evolutionary processes, designed to use only relevant attributes which are to discriminate a class from the others and improve the accuracy of the ruleset, based on the Iterative Rule Learning (IRL) genetic algorithm (also known as the Pittsburgh approach). IRL returns a rule per output class with the exception of one class which is set as default. The third step optimizes the accuracy of

the classifier by maximising the product of sensitivity and specificity. Linear Programming Relaxation [285][286] is a method for learning two-level Boolean rules in conjunctive normal form (AND-of-ORs) or disjunctive normal form (OR-of-ANDs) as a type of human-interpretable classification model. A first version uses a generalization of a linear programming relaxation from one level to two-level rules whose objective function is a weighted combination of the total number of errors and features used in the rule. In a second version, the 0-1 classification error is replaced with the Hamming distance between the current rule and the closest rule that correctly classifies a sample instance. The main advantages for explainability of the Hamming distance is that it avoids identical clauses in the ruleset, thus repetitions, by training each clause with a different subset of input instances.

All the above methods were intrinsically ante-hoc, but other methods exist for post-hoc explainability. For example, Mycin [3], probably the first method for explainability ever developed, is a backward chaining expert system based upon a knowledge-based of IF-THEN rules composed by an expert, a database of the context set of facts that satisfy the condition part of the rules, and an inference engine that interpret the rules. It also includes a natural language interface that allows end-users to interact with the system independently of the expert by asking English questions, and the system can respond to them by using its inference engine and performing the reasoning involved in composing an answer to them. In details, it searches for facts that match the condition part of the productions that match the action part of the question. This method allows the system to explain its reasoning and final inferences by using AND/OR trees created during the production system reasoning process, thus showing an element of explainability. Similarly, the Sugeno-type fuzzy inference system proposed in [287] consists of an explicit declarative knowledge representation of rules, which are fired at the same time by a given input, and produce a final inference. Besides this, the system includes an explanatory tool which shows a numerical representation of the input variables, the set of co-fired rules and an English statement exposing the reasoning process. In an example taken from the application to an Unmanned Aerial Vehicle (UAV) sent on a fight mission, is a statement like “UAV aborted the mission because the weather was a thunderstorm and the distance from the enemy was too close”. Another method, the Fuzzy Inference-Grams (Fingrams) [288] produces inference maps of sets of fuzzy rules which graphically depict the interaction between co-fired rules and gives support to detect redundant or inconsistent rules as well as it identifies the most significant ones, by using network scaling methods that simplify the maps while maintaining their most important relations. Fingrams also quantifies the comprehensibility of the ruleset, measured as the proportion of the co-fired rules. The assumption is that the larger the number of rules which are co-fired for a given input, the smaller the comprehensibility of the ruleset. ExpliClas [289] is a visual interface designed to explain, in an instance-based manner, rule-based classifiers (such as those algorithms extracting DTs from data, like C4.5 or CART, for instance) with visual and textual explanations. The rules are graphically displayed as DTs and a natural language generation approach returns textual explanations of the fired rules. ExpliClas was tested on in the context of recognising the role of basketball players from some physical characteristics and game statistics. An example of textual explanation produced in this case is “The player is a Point-Guard because he is medium-height and he has a small number of rebounds.” and it is accompanied by a graph of the DT structure and bar-charts of the player’s information.

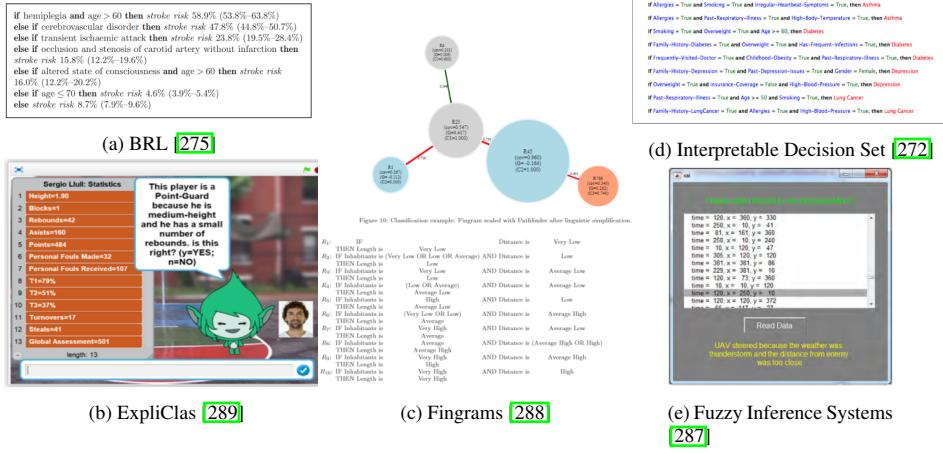


Figure 14: Examples of explanations generated by methods for explainability for rule-based inference systems.

### 6.5. Other model-specific methods for explainability

Scholars proposed other methods for explainability that are not strictly based on neural networks or rule-based classifiers (see table A.12 and figure 15).

#### 6.5.1. Ensembles.

Some methods were designed to interpret the logic followed by ensembles. [290] introduced an algorithm to tweak the input features to change the output of a tree ensemble classifier. It modifies a variable (or a set of variable) of an input instance by applying a linear shift, capped to a global tolerance value, until all the trees in the ensemble assign it to another target class. The delta between the original and the tweaked value represents the ‘effort’, or ‘tweaking cost’, required to move the instance into the target class and provides a measure of the sensitivity of the model to changes to that particular feature(s). This information can be used to rank the variables according to their tweaking costs and inform end-users on how a particular instance must be modified to change its output label and at what cost this can be achieved. Three methods for extracting a single DT from ensemble models, and for generating global explanations were presented in [291] [292] [293]. In detail, [292] uses the solution obtained from combining the several hypotheses (or models) of the ensemble as an oracle, and it selects the single hypothesis that is most similar to the oracle. The similarity is measured according to three formal metrics: ‘ $\theta$  – measure’ which determines the probability that both classifiers agree, ‘ $\kappa$  – measure’ which assesses the probability that two classifiers agree by chance and ‘ $Q$  – measure’ which assigns values between 0 and 1 to classifiers that correctly predict the same input instances and values between -1 and 0 to classifiers that commit errors on different instances. Instead, [293] proposed to generate a tree from the ensemble by using a divide-and-conquer algorithm analogous to C4.5. Similarly, [291] combined an opaque learning algorithm (random forest), with a more transparent and inherently interpretable algorithm (decision tree). On one hand, the opaque algorithm represents the ‘oracle’ which search for the most relevant output. On the other hand, a natural language generation approach is aimed at composing a textual explanation for this output which is the interpretation of the inference process carried out by the correspondent decision tree, if the

output of both the learning algorithms coincides. [294] proposed instead a method to efficiently merge a set DTs, each trained independently on distributed data, into a single tree to overcome the lack of interpretability of the distributed models. The algorithm consists of three steps. First, each DT is converted into a ruleset where each rule replicates a path from the root to a leaf and defines a region in the output space. All the regions are disjoint and they cover the entire space. In the second phase, the regions are combined by using a line sweep algorithm which sorts the limits of each region and merges overlapping and adjacent ones. Finally, a DT is extracted from the regions with an algorithm that mimics the C5.0 and uses the values in the regions as examples.

A similar approach is at the basis of the Factorized Asymptotic Bayesian (FAB) inference method [295] which consists of a bayesian model selection algorithm that simplified and optimized a tree ensemble. FAB estimates the model's parameters and the optimal number of regions of the input space (ensemble methods often splits the input space into a huge number of regions) to derive a simplified model with appropriate complexity and prediction accuracy. inTrees [296] extracts, prunes and selects rules from a tree ensemble. The algorithm starts from the set of all the rules in the ensemble and excludes those covering a small number of instances. At each iteration, the algorithm selects the rule with the minimum error and shorter condition, then it removes the instances satisfying this rule from the dataset and it updates the initial ruleset according to the instances left, by discarding rules that at this stage cover just a few, if not none, instances and recalculating their error. Discriminative Patterns [297] aims at interpreting a random forest model that classifies sentences according to their contents by extracting a ruleset that enables interpretation and gains insight of useful information in texts which corresponds to discriminative sequential patterns of words, or sentences that determine the predicted class. Tree Space Prototypes (TSP) [298] selects prototypes from a training dataset to explain the prediction made by ensembles of DTs and gradient boosted tree models on a new observation. To measure the similarity between the new instance and the prototypes, the authors proposed a metric based on the weighted average of the number of trees in the ensemble that assigns the points to the same output class to quantify the contribution of the predictions made by each DT to the overall prediction. By following the path root-to-leaf of the most relevant DT, it is possible to determine the values of the features deemed important by the tree for predicting the class of the new instance and select a prototype having the same values.

### 6.5.2. Support Vector Machines.

ExtractRule [138] converts hyperplane-based linear classifiers, like SVMs, into a set of non-overlapping symbolic rules which are human-understandable because they display, in a compact format, the inferential process of the underlying classifier. An example of a rule extracted from a classifier trained to distinguish between malign and benign tumors is “( $Cell\ Size \leq 3$ )  $\wedge$  ( $Bare\ Nuclei \leq 1$ )  $\wedge$  ( $Normal\ Nucleoli \leq 7$ )  $\implies$  *mass is benign*”. Each rule can be seen as a hypercube in the multidimensional space generated by the input variables with edges parallel to the axis. To define these hypercubes, each iteration of this algorithm is formulated as one of two possible optimization problems. The first formulation seeks to maximize the volume covered by each rule whereas the second formulation maximizes the number of samples covered. Important support vectors and Border classification [146] are two methods for providing insight into local classifications produced by a SVM. The former reports the most influential support vectors for the final classification of a particular data instance, thus determining the most similar instances to the test point that belong to the same class. As in the previous methods based on prototypes, presenting this subset helps users understand the model by leveraging on

the human ability to induce principles from a few examples. The latter determines which features of a testing data instance would need to be altered (and by how much) to be classified on the separating surface between the two classes, thus providing, as in the feature tweaking method, a measure of the cost required to change a model's prediction. SVM+Prototypes [299] is based on a clustering algorithm to detect the prototype vectors for each class, after the decision function is determined to employ a SVM. These vectors are combined with the support vectors using geometric methods to define ellipsoids in the input space, which are later transferred to IF-THEN rules as the mathematical equations that defined the ellipsoids, so a rule looks like "If  $AX_1^2 + BX_2^2 + CX_1X_2 + DX_1 + EX_2 + F \leq G$  Then *Class<sub>1</sub>*". Weighted Linear Classifier [300] generates weighted linear SVM classifiers or random hyperplanes to obtain models whose accuracy is comparable to that of a non-linear SVM classifier and whose results can be readily visualized by projecting them on separating hyperplanes and decision surfaces. These projections are considered as a sort of explanations. A method based on Self-Organizing Maps (SOM) used to visualise SVMs was proposed in [301]. It is a specific unsupervised network aimed at investigating a high-dimensional space of data for a cluster of points by projecting these clusters onto a 2-dimensional map, but trying to preserve their topologies. Thus, it allows visualising both data and the SVM models, providing an overall overview of the support vector decision surface which is not possible with other visualization approaches. [302] [303] [304] introduced a method for automatically generating nomograms as the graphical tool for visual explanations of the inferential mechanisms of SVM and naïve bayesian-driven models. A nomogram is a two-dimensional diagram designed to allow approximating graphical computation of mathematical functions by showing a set of scales, one for each variable (dependent and independent) in an equation. By drawing a line connecting specific values of all the scales related to the independent variables, it is possible to calculate the value of the dependent variable from the intersection point between the line and the variable's scale. The advantages for explainability of nomograms are simplicity of presentation and clear display of the effects of individual attribute values.

#### 6.5.3. Bayesian and hierarchical networks.

Explaining Bayesian network Inferences (EBI) [305] produces DT rules to show how the variables of a bayesian network interact to make predictions. In detail, EBI explains the value of a target node in terms of the influential nodes in the target's Markov blanket which include the target's parents, children and the children's other parents. Working backwards from the target node, EBI shows the derivation of each intermediate node and explains how missing and erroneous values are compensated by displaying these causal relationships in a DT hierarchy. [306] instead proposed an explanation method for understanding bayesian networks in terms of scenarios. Narrative approaches to reasoning with legal evidence, for instance, are based on the formulation of alternative scenarios which are subsequently compared according to two aspects: the relations with the evidence and the quality that depends on the completeness, internal consistency and plausibility of the scenario itself. The aim is to explain the content of the bayesian network by reporting which scenarios were modelled and evaluating their evidential support and quality. Probabilistically Supported Arguments (PSA) [307] is a two-phase method for extracting probabilistically explanatory supported arguments from a bayesian network. In the first phase, a support graph is constructed from a bayesian network representing the structure for a particular variable of interest. In the second phase, given a set of observations, arguments are built from that support graph. To do so, the algorithm defines a logical language and a set of rules built from the support graph by following its edges and nodes. The parents of a node are the rule conditions, the node itself is the rule's outcome. Only the parents supported by pieces of evidence are consid-

ered. Then, the ASPIC+ framework for structured argumentation is instantiated. Arguments can attack each other on the conclusion variable and defeat can be based on the inferential strength of the arguments which can be computed with two types of measures: ‘incremental measures’ which assign a number to the weight of the evidence (the Likelihood Ratio is an example of these measures) and ‘absolute measures’ which assign strength based on posterior probability, such as the posterior for instance. Such arguments can help interpret and explain the relationship between hypotheses and evidence that is modelled in the Bayesian network. Contribution propagation [308] is a per-instance method for hierarchical networks that explain which components of the input were responsible (and to what degree) for its classification. The central idea is that a node was important to the classification if it was important to its parents, and its parents were important to the classification. The contribution of each input component is visualized as heat-maps.

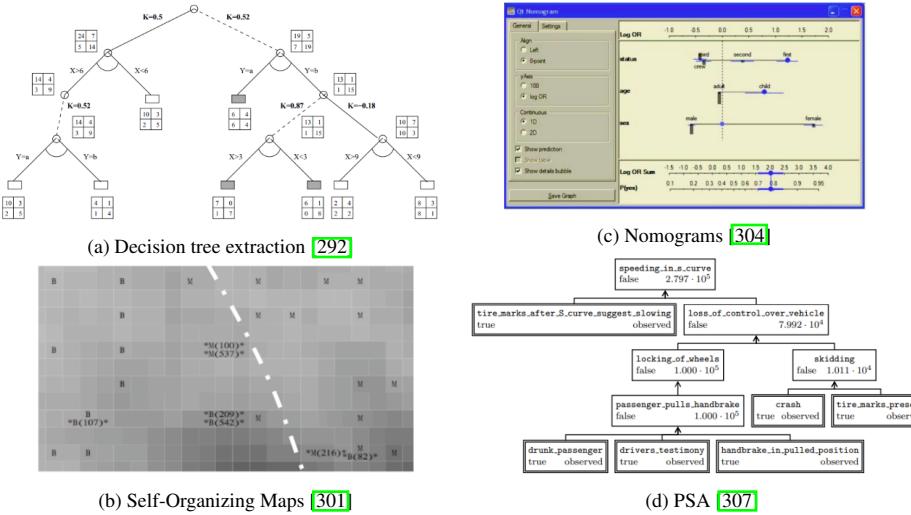


Figure 15: Examples of explanations generated by methods specifically designed to interpret ensembles (a), Support Vector Machines (b-c) and bayesian networks (d).

### 6.6. Self-explainable and interpretable methods

Naturally interpretable models, sometimes referred to as ‘white-box models’, are ‘ante-hoc’ (see table A.13 and figure 16). Their output format depends on their architecture and inputs format. Bayesian Case Model (BCM) [309] is a method for explainability for bayesian case-based reasoning, prototype classification and clustering. BCM learns prototypes, corresponding to the observations that best represent clusters in a dataset, by performing joint inference on cluster labels, prototypes and important features. Gaussian Process Regression (GPR) [310] is a powerful, but amenable to analysis, data-driven model. GPR is a non-parametric regression algorithm, meaning that it does not make any assumption on the estimator function as linear and logistic regression algorithms do, robust to missing data and interpretable because the weights assigned to each feature provide a measure of its relevance. Generalized Additive Models [99] and their extension with pairwise interactions (GA<sup>2</sup>Ms) [311] [31] are linear combinations of

simple models, called ‘shape functions’, trained on a single feature (GAMs) or up to two features (GA<sup>2</sup>Ms). Their simple structure allows end-user to easily understand the contribution of individual features to the predictions and to visualize them, together with the shape functions, with bar- and line-charts. Oblique Treed Sparse Additive Models (OT-SpAMs) [312] are instances of region-specific predictive models. They divide feature spaces into regions with sparse oblique tree splitting and assign local sparse additive experts to individual regions. Transparent Generalized Additive Model Tree (TGAMT) [313] was proposed as an explainable and transparent method that uses a CART-like greedy recursive search to grow the tree. Multi-Run Subtree Encapsulation, which comes from the genetic programming (GP) realm, was proposed in [314] as a way to generate simpler tree-based GP programs. If the tree contains sub-trees of different makeup but evaluating the same vector of results, they are to be considered as the same sub-tree. This reduces, according to the authors, the complexity of the entire tree structure and the resulting expressions, in favour of explainability.

Probabilistic Sentential Decision Diagrams (PSDD) [315] can be described as circuit representations where each parameter represents a conditional probability of deciding the input variables and each node is either a logical AND gate with two inputs or a logical OR gate with an arbitrary number of inputs. The PSDD structure can be visualized as an easily-interpretable binary tree. Mind the Gap Model (MGM) [316] is a method for interpretable feature extraction and selection. The goal is to split the observation into clusters while returning the list of dimensions that are important for distinguishing them. The results are presented as a mix of numbers, which are the relevance values of each dimension, texts and graphs that represent the dimensions themselves. For example, in a classification problem of images representing the four seasons, MGM returns samples of images belonging to each class (spring, summer, autumn and winter) together with the list of their relevant features (like snow, sun and flowers) and the relevance values of each feature per target class (snow has a high relevance value for the class ‘winter’). Supersparse Linear Integer Model (SLIM) [317] generates a scoring system from an input dataset by assigning a score to each variable that contributes to the prediction. These scores are multiplied by a set of coefficients inferred from the training dataset and then added, subtracted, and/or multiplied to make a prediction. Scores are generated by minimising the 0-1 loss to reach a high level of accuracy and to produce a classifier that is robust to outliers by applying a  $\ell_0$  – *penalty* to encourage a high level of sparsity and a set of interpretability constraints which restricts coefficients to a user-defined set of meaningful and intuitive values. Eventually, Unsupervised Interpretable Word Sense Disambiguation [318] produces interpretable word sense disambiguation models that create clusters, or inventories, of words. For example, an inventory can be the collection of all the words related to ‘furniture’ (such as table, chair and bed). The words are clustered according to their co-occurrence and relative position in a text, where close words are assumed to be highly correlated, and their syntactic dependency extracted from the Stanford Dependencies (representation of grammatical relations between words in a sentence.) The resulting word groups can be interpreted at three levels: (1) word sense inventory where each sense of the word under analysis is displayed as a separate network-graph where the nodes are the semantically related words and the edges represent their semantic relationships. For example, the word ‘table’ is connected to two networks corresponding to ‘furniture’ and ‘data’ senses; (2) sense feature representation characterized by a list of sparse features (which consists of words) ordered by relevance; (3) results of disambiguation in context by highlighting the most discriminative words that caused the prediction. Words like ‘cookie’ and ‘website’ indicate that ‘table’ refers to a collection of data and not as a piece of furniture.

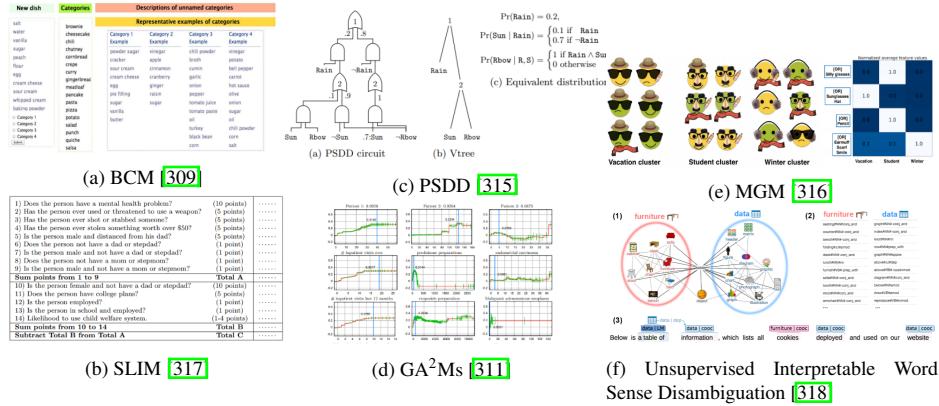


Figure 16: Examples of ante-hoc methods for explainability designed to generate self-explainable models.

## 7. Evaluation of methods for explainability

The proposal of many methods for explainability pushed authors to focus also on their evaluation. Different evaluation metrics were proposed and found in the literature as well as different types of evaluation were conducted, as summarised in figure 17. A thorough review of these studies led to the identification of two main ways to evaluate methods for explainability, as shown in figure 17.

- **objective evaluations** - it includes research studies that employed objective metrics and automated approaches to evaluate methods for explainability;
- **human-centred evaluations** - it contains those studies that evaluated methods for explainability with a human-in-the-loop approach by involving end-users and exploited their feedback or informed judgement.

The same dual categorisation system was suggested in [64], but they named the two classes *heuristic-based* and *user-based* metrics. The former includes quantitative measures which consist of mathematical entities such as, for example, the size of models [26] [88] [319] [320] [269]. This is a simple explainability indicator and it is based upon the assumption that the bigger the size of a model, the harder it becomes for the users to understand it. However, this assumption was proved false. One of the outcomes of the human-centred evaluation study conducted in [88] was that users found some larger models to be more comprehensible than some smaller models because larger models provided more classification-relevant information and users are unlikely to accept weaker, simpler models when the underlying modelled concept is believed to be complex. An alternative categorisation was presented in [2] with three classes: *application-grounded*, *functionally-grounded* and *human-grounded* evaluation metrics where functionally-grounded and human-grounded metrics respectively corresponds to the heuristic-based and user-based metrics proposed in [64]. Application-grounded metrics assess the quality of machine-produced explanations of data-driven models by comparing the increase in productivity of a few

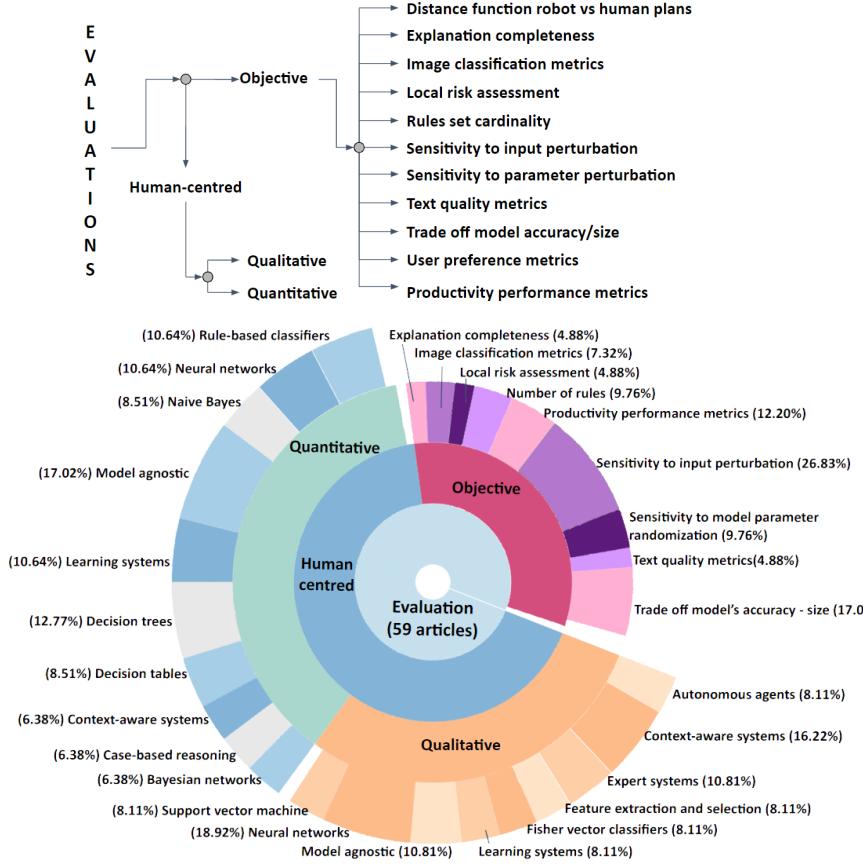


Figure 17: Classification of the approaches to evaluate methods for explainability (up) and distribution of the relative scientific studies across categories (down).

users of these models when following these explanations instead of those produced by human engineers, as done in [26] [2]. Because they involve humans, they are merged into the human-grounded metrics.

### 7.1. Objective evaluations

Scholars proposed several metrics to evaluate, formally and objectively, the methods for explainability, listed in table I. In the scientific literature, there is consensus that simpler learning techniques, such as linear regression and DTs, can lead to more transparent inferences than more complex techniques, such as neural network, as they are intrinsically self-interpretable [26]. However, these simpler techniques usually do not lead to the construction of models with the same level of accuracy than those induced by more complex learning techniques. The interpretability of these models depends on many factors such as the learning algorithm, the learning architecture and its configuration (hyper-parameters).

A few sale performance indicators were utilized as a formal metric to assess the increase in productivity of the sales department when two methods of explainability, Explain and Ime [150][151], were applied to a complex real-world business problem of business-to-business sales forecasting [321][322]. The system was tested for a long period in a real-world company and the sale performance indicators were monitored. The indicators showed that the forecasts based on the Explain and Ime explanations outperformed initial sales forecasts, which supported the hypothesis that data-driven explanations better facilitate unbiased decision-making than the mental models of sale experts based upon their previous experience. Two quantitative evaluation metrics to assess the interpretability of methods generating visual explanations of a neural network trained to classify images were presented in [30]. The first metric, *filter interpretability*, considers six types of semantics for CNN filters that must be annotated by humans on testing images at the pixel level: objects, parts, scenes, textures, materials, and colours. The metric measures the intersection areas between these annotations and the distribution of the activation values of a filter over a heat-map. If there are overlapping areas, it can be said that the filter represents these semantic concepts. The second metric, *location instability*, checks if a CNN locates the relevant parts of the same object, shown in different images, at an almost constant relative distance, as the distances between the parts of an object must be almost invariant. [256] proposed to use three automated quantitative metrics, designed to assess the quality of text documents, to evaluate textual explanations automatically generated by methods for explainability: *BiLingual Evaluation Understudy (BLEU)* that assesses the similarity of sentences based on the average percentage of n-gram matches, *Automatic NT Translation Metric (METEOR)* that evaluates semantically the similarity between words of sentences by using pre-trained word embeddings and *Consensus-based Image Description Evaluation (CIDEr)* that compares sentences generated by neural networks to reference explanations written by humans by counting ‘Term Frequency–Inverse Document Frequency’ weighted n-grams. A general evaluation metric for post-hoc methods for explainability was presented in [323] based on the risk of generating unjustified ‘counterfactual examples’ which are instances that do not depend on previous knowledge but are artefacts of the classifier. This might happen when a model must predict an area not covered by the training set. The algorithm that generates these examples applies the minimal perturbation that changes the predicted classes of observation in such a way that it is still connected to the input data and avoid the construction of examples representing situations that are neither feasible nor logical. The explanations of the predictions made by an underlying model for these observations would not make sense and would not help the understanding of the model’s logic.

Only a few scholars carried out formal comparisons, based on heuristic-based metrics, between different methods for explainability to evaluate their strengths and weaknesses. The methodologies utilized for the comparisons are (see also table A.15):

- **sensitivity to input perturbation** - some features of the input dataset are removed, masked or altered and the explanations generated by a method for explainability from the model trained on both the original and modified inputs are compared;
- **sensitivity to model parameter randomization** - the outputs a method for explainability generated from a trained model and another model of the same architecture with some or all parameters replaced by random values are compared;
- **explanation completeness** - these approaches check which method generates explanations that describe the inferential process of the underlying model to the highest extent.

This consists of capturing the highest number of features of the input that affect the decision process of the model.

The vast majority of these scientific articles compared methods for explainability designed to generate visual explanations of the logic followed by neural networks for the classification of either images or texts. All these methods produce maps, like heat-maps or feature maps, and the comparison is carried out by measuring the differences in these maps generated before and after the input or the model's parameters were perturbed. The complete list of the methods that were compared, along with the type of input data that were analysed in these comparisons, is shown in table A.16 [324] [325] compared the saliency maps generated by various methods for visual explainability to either a randomly initialised untrained network or from a copy of the dataset in which the labels were randomly permuted. The degree of correlation between the saliency maps was measured by calculating Spearman Rank Correlation coefficients. Similarly, [326] proposed to vary input images by occluding with zero-valued pixels their portions sharing the same relevance level, according to the saliency maps generated by four gradient-based attribution methods. The sensitivity of the four methods to this input perturbation was assessed with a formal metric, *Sensitivity-n*, which quantifies the variation in the output caused when features sharing the same relevance level are removed. The results of this analysis showed that Occlusion Sensitivity (see Section 9) is the method that identifies the few most important features, in respect with the other methods, because it suffers the faster variations in the output when the most relevant pixels are removed. Layer-wise Relevance Propagation (LRP) and Sensitivity Analysis were tested in [327] [328] [329] [330] by removing important words from input documents in text classification problems [327] [328] or replacing the most relevant pixels (in case of images) by randomly sampling new pixel values from a uniform distribution [329] [330]. The metric used in this study assesses the differences in the model's classification accuracy between the original and the perturbed input points when fed into the model. Both studies showed that LRP qualitatively and quantitatively provides a better explanation of what made a DNN reach a specific classification prediction. [331] used the same evaluation approach of [329] [330] to compare LRP with Occlusion Sensitivity and Sensitivity Analysis. LRP and Occlusion Sensitivity performed better than Sensitivity Analysis, seconding the findings of [329] [330]. Input perturbation was also used in [90] [332] [92] to test the robustness of several methods that generate visual explanations of the inferential process of DNNs applied to image classification problems. Robustness concerns variations of an explanation provided by a method with respect to changes in the input leading to that explanation. Intuitively, if the input being explained is modified slightly-subtly enough not to change the prediction of the model then the explanation must not change much either [90]. On the one hand, [90] applied a Gaussian noise to input images and measured the relative changes in the output with respect to these perturbations with the Local Lipschitz Continuity metric. On the other hand, [332] [92] added/subtracted a constant shift to the input images. Then, [332] used two metrics to assess the similarity between the heat-maps generated from the original and the perturbed images: Spearman Rank Correlation coefficients and Top- $\kappa$  intersection which measure the size of the intersection of the  $\kappa$  most important features before and after perturbation. [92] instead measure the differences in the model's predictions by checking whether the methods for explainability under analysis satisfy the requirement of 'input invariance' (see Section 5.1). These studies show that all tested methods (table A.16) are vulnerable even to small perturbations that do not affect the predictions of an underlying model but they significantly change the heat and saliency maps produced by the explainers. These do not satisfy the 'input invariance' requirement [92], that means they do not reflect the sensitivity of a model with respect to input

perturbations. Lastly, [333] compared the completeness of the explanations generated by seven methods for explainability that interpret the logic of DNNs by calculating the partial derivatives of the output according to the inputs, their perturbation and analysing the network's weights.

### 7.2. Human-centred evaluation

Explanations are effective when they help end-users to build a complete and correct mental representation of the inferential process of a given model. Many scientific articles focused on testing the degree of explainability of one or multiple methods, with a human-in-the-loop approach. These experiments involved human participants of two kinds. On the one hand, people randomly selected from the lay public and without any prior technical/domain knowledge who were asked to interact with one or more explanatory tools and give feedback by filling questionnaires. On the other hand, domain experts who were asked to give informed opinions on the explanations produced by these methods and verify the consistency of the explanations with the domain knowledge. Examined scientific articles can be clustered into two categories, depending on the nature of questions administered to people (see table A.14). *Qualitative studies* are based upon open-ended questions aimed at achieving deeper insights whereas *quantitative studies* make use of close-ended questions that can be easily analyzed statistically. The first methods for explainability that were tested by human users are those generating textual explanations for Expert Systems (ES) in the 90's [334][335]. These researches aimed at collecting pieces of evidence on whether and how explanations could enhance the user's confidence in the decisions proposed by an ES. Scholars carried out several human-centred evaluations over the years to assess the effects of textual explanations on end-users to other types of systems employing ML models, such as learning systems [336]. The impact of explanations on the reliability, trust and reliance of end-users on automated systems was explored in [54]. The participants were presented with photos of the Fort Sill terrain where the presence of a camouflaged soldier was indicated by an automated decision system. Initially, they considered the inference produced by the system to be trustworthy and reliable but, after observing the system making errors, they distrusted even reliable aids, unless an explanation was provided regarding why the system failed. In conclusion, explanations of these errors increased the trust of the participants in the automated system who were asked to estimate their perceived reliability on a 9-point Likert-format scale, ranging from 'not very well' to 'very well'. The influence of explanation factors over the capacity of end-users to understand and develop a mental representation of the internal reference process of a model was investigated in [337][338][339][340]. The explanations produced by the analysed methods for explainability consisted of graphical representation of the most relevant features. [337][338] showed that users of simulation-based training systems with virtual players prefer short explanations to long ones where length is defined by the number of elements in an explanation. An element can be a goal or an event of the training program. This was tested by showing to the participants four explanation alternatives of different length (they contained either one or two elements), in the form of DTs, for each action of the virtual players and asked to indicate which alternative they considered the most useful for increasing end-user understanding. The influence factors analysed in [339] were the number of independent variables of a trained linear regression model and the values of these variables for each instance. Some participants were randomly assigned to check either a model that uses only two features or a model that uses eight features. The coefficients of the linear regression model were also presented only to half of the participants. Then, the participants were asked to estimate what would be the output of the model and to correct it in case it was not accurate. As expected, users can easily simulate models with a small number of features whereas,

surprisingly, displaying model internal parameters can hamper their ability to notice unusual inputs and correct inaccurate predictions. The method for explainability tested in [340] listed the two most relevant features and the prediction of the model. The prediction was coded with a solid colour taken from a scale running from red, representing the most negative possible outcome, to green, the most positive one. Participants were asked to interact with the system for two weeks at the end of which they were interviewed to collect their feedback and to check whether they gained some insight on the logic of the model to be explained by the explanatory system under analysis. [103] studied the explainability of an interactive interface, called *Prospector*, containing a set of diagnostic tools that allows end-user, via visual and numerical representations, to understand how the features of a dataset affect the prediction of a model overall. Users can also inspect a specific instance to check its prediction and can weak feature values to see how the model responds. A team of data-scientists was asked to interact with this tool to debug a set of models designed to predict if a patient is at risk of developing diabetes by using a database of electronic medical records. The human experiment consists of interviewing, at the end of the experiment, the data scientists on whether they feel that it was beneficial for their work. Methods for evaluating the interpretability of data-driven models with a human-in-the-loop approach were proposed in [341][342][343][344]. The approach proposed in [341] identifies some *proxies* which consist of other, simpler models inherently more explainable. For example, a DT is inherently more interpretable than DNNs and the former methods can be used to explain the latter. The authors presented to participants a list of the coefficients of the features used by each proxy and a graphical depiction of its structure (in the form of a DT) and they asked them to identify the correct prediction. [344] proposed instead to assess the comprehensibility of DTs by asking the participants to perform the following tasks: (I) classify a data-point according to the classification tree, (II) explain the classification of a data-point by pointing out which attributes' values must be changed or retained to modify the instance's class, (III) validate a part of the classification tree by asking the participant to check whether a statement about the domain is confirmed/rejected by the tree, (IV) discover new knowledge by finding a property (attribute-value pair) that is unusual for instances from one class, (V) rate the classification tree by giving a subjective opinion on the comprehensibility of the tree and, lastly, (VI) compare two classification tree by saying which one is more comprehensible. The two studies presented in [342][343] analysed the interpretability of predictive models by asking participants to interact with them and fill self-reporting questionnaires. The surveys carried out in [342] aimed at comparing six supervised learning algorithms. These were ranked in order of preference based on the subjective quantification of understandability obtained from the self-reporting questionnaires filled by participants. Pairs of models trained on the same dataset were generated and participants were asked to rate them on a scale where one extreme is ‘the first model is the most understandable’ to the other extreme ‘the second model is the most understandable’ via increasingly positive grades. In the study carried out in [343], participants were required to evaluate the explanations of a credit model, trained to accept or reject loan applications, consisting of IF-THEN rules and displayed as a decision tree. They were asked to predict the model’s outcome on a new loan application, answer a few yes/no questions such as “Does the model accept all the people with an age above 60?” and rate, for each question, the degree of confidence in the answer on a scale from 1 (Totally not confident) to 5 (Very confident). The authors measured, besides the answer confidence, other two variables about task performance: accuracy, quantified as the percentage of correct answers, and the time in seconds spent to answer the questions. The effectiveness of why-oriented explanation systems in debugging a naïve Bayes learning model for text classification and in context-aware applications were respectively tested in [345][60] and [346][77]. [345][60] asked participants to

train a prototype system, based on a Multinomial naïve Bayes classifier, that can learn from users how to automatically classify emails by manually moving a few of them from the inbox into an appropriate folder. The system was subsequently run over a new set of messages, some of which were wrongly classified. The participants had to debug the system by asking ‘why’ questions via an interactive explanation tool producing textual answers and by giving two types of feedback: some participants could label the most relevant feature (words) whereas the others could only provide more labelled instances (moving more messages to the appropriate folders). At the end of the session, the participants filled a questionnaire to express their opinions on the prototype. In the experiment run in [346] [77], participants were invited to interact with a model that predicts whether a person is doing physical exercise or not, based on the body temperature and the pace. They were shown with some examples of inputs and outputs accompanied by graphical (in the form of decision trees) and textual explanations on the logic followed by the model. Half of the participants were presented with why explanations, such as “*Output classified as Not Exercising, because Body Temperature < 37 and Pace < 3*” whereas the other half were presented with why not explanations, such as “*Output not classified as Exercising, because Pace < 3, but not Body Temperature > 37*”. Subsequently, the participants had to fill two questionnaires to check their understanding by asking questions how the system works and to give feedback on the explanations in terms of understandability, trust and usefulness. Both questionnaires contained a mix of open and close questions, where the close ones consisted of a 7-point Likert scale. A qualitative evaluation of the interpretability of the Mind the Gap Model (MGM) method for explainability was gathered in [316]. MGM clustered the data of an input dataset according to the most relevant features. The participants were presented with the raw data and the data clustered with MGM and k-means and were asked to write a 2-3 sentence executive summary of each data representation within 5 minutes. They all found impossible to summarise the raw data, not being able to complete the task in the given amount of time, but they managed to do so on the data with clustered MGM and k-means. To test Bayesian Case Model (BCM), [309] asked the participants to assign sixteen recipes, described only by a set of ingredients, to the right category (so a recipe of cookies had to be classified under ‘cookie’) and then they counted how many of them were correctly classified. BCM was compared with Latent Dirichlet Allocation (LDA), a clustering approach based on extracting similar characteristics in the data. The need for XAI in Intelligent Tutoring Systems (ITS) was explored in [347]. The participants in the study were asked to use an ITS that provided tools to explore and explain an algorithm solving constraint satisfaction problems in an interactive simulation. The participants were instructed by the exploration tool with a textual hint message. They could select to have the hint explained by the explanatory tool which was also designed to solicit their suggestions on the explanations they would like to see for each hint by presenting them a checkbox list with the following options: ‘why the system gave this hint’, ‘how the system chose this hint’, ‘some other explanation about this hint’ (followed by an open-text field) and ‘I do not want an explanation for this hint.’

Many scholars proposed human-centred evaluation approaches for methods for explainability generating visual explanations. The participants selected in the study in [348] were presented with whole images misclassified by a DNN and the visual explanations generated by LIME and MMD-critic. For example, a photo of a Jeep with a red-cross was wrongly classified as an ambulance and the visual explanations show the red-cross with the rest of the image greyed out. Participants were asked to say whether the class predicted by the model was nonetheless relevant where the possible answers to questions like ‘Is the label Red-Cross relevant?’ were ‘yes’ and ‘no’. A similar experiment was carried out in [349] to test GAN Dissection. Participants were

presented with images reporting highlighted patches showing the most highly-activating regions for each unit at each intermediate convolutional layer of a DNN. Each layer was aligned with a semantic and were given labels across a range of objects, parts, scenes, textures, materials, and colours. For example, if a DNN was trained to recognize a list of object in input images, like flowers and cars, the semantic consists of this list and images showing flowers were labelled ‘flower’. The participants were asked to say if the highlighted patches were pertinent to the label by answering yes/no questions. The capacity of Anchors and LIME in helping end-users forecasting the predictions of an image classifier was tested in [140]. Participants were asked to predict the output class assigned by the classifier to ten random test instances before and ten instances after seeing two rounds of explanations generated by either Anchors or LIME. A few scholars conducted human-centred studies to test the interpretability of the heat-maps generated with LRP. [350] [351] applied it respectively to test models built with Fisher vector classifiers for object recognition in images and to SVMs, trained on videos, to understand and interpret action recognition and to check whether LRP allows identifying in which point of the video the important action happens. By visually inspecting the heat-maps, the authors of the two studies could show a possible weakness of the underlying classifiers by looking at the regions highlighted in the heat-maps and examining whether they were relevant for the recognition of an object (or at least part of it) in images and of the areas of video frames showing the actions performed in the video. Similarly, [352] employed LRP with DNNs for electroencephalography (EEG) data analysis. The predictions of the trained DNNs are transformed with LRP, for every trial, into heat-maps indicating the relevance of each data point. The relevance information can be plotted as a scalp topography that can be visually inspected by experts to check if there are neurophysiologically plausible patterns in the EEG data. [353] used LRP for computing the contribution of contextual words to arbitrary hidden states in the attention-based encoder-decoder framework of neural machine translation (NMT). As per the previous studies, the authors checked if the translation made by the NMT (Japanese-English) were right or wrong and what types of errors were made more frequently. The participants in [133] were asked to interact with explAIner which showed them explanations generated by LRP and Saliency Analysis of both a simple and a complex network trained on the MNIST dataset and, in the meantime, to communicate their thoughts and actions by ‘thinking aloud’. The sessions were audio-recorded and screen-captured. At the end of the sessions, participants were also interviewed to provide qualitative feedback on the overall experience. Another method for explainability producing visual explanations as maps, GradCam, was applied to multivariate time series from photovoltaic power plants that were fed into a neural network to forecast the energy production of these plants in different weather conditions [354]. GradCam was used to explain which features, such as environmental temperature, wind bearing or humidity, or any combinations of these features were responsible, at different time intervals, for a given prediction. The results showed that GradCam was able to visualise the network’s attention over the time dimension and the features of multivariate time series data. [355] compared other three methods, namely Activation Maximization, Sampling Unit and Linear Combination, designed to produce explanations as heat-maps of the most relevant features of the input images. Activation Maximization consists of selecting the input features that maximise the activation of a single hidden neuron. Sampling Unit consists of setting the value of a neuron to one and calculating the probability with which each sample is assigned to a class. Lastly, Linear Combination consists of choosing the largest weights of the connections between neurons of two adjacent layers. The authors did not use any objective measure to compare these methods but a qualitative visual inspection of the heat-maps and the connections between them.

## **8. Final remarks and recommendations**

Research on methods to explain the inner logic either of a learning algorithm, a model induced from it, or a knowledge-based approach for inference is now generally recognized as a core area of AI and is referred to as eXplainable Artificial Intelligence (XAI). Note that other common terms exist, such as ‘Interpretable Machine Learning’, but with XAI we would like to emphasise the wider applicability of this emerging and fascinating field of research. Several scientific studies are published every year, with many workshops and conferences organised around the world, to propose novel methods and disseminate findings. Although this has led to the production of an abundance of knowledge, unfortunately, this is very scattered. This systematic review attempted to fill this gap by organising this vast knowledge in a structured and hierarchical way. Some scholars already tried to classify scientific studies but, given a large amount of literature, they decided to focus only on a specific aspect of explainability. To the best of our knowledge, this is the first attempt to review a wider body of literature that has led to the definition of four clusters: (I) reviews focused on specific aspects of XAI, (II) the theories and notions related to the concept of explainability, (III) the methods aimed at explaining the inferential process of data-driven and knowledge-based modelling approaches, and (IV) the ways to evaluate the methods for explainability. Many studies within XAI have focused on improving the quality and widening the variety of explanations for several types of learning approaches with data. Since the early ’80s and ’90s, with research concerned only with textual explanations, to nowadays, scholars have been targeting new explanation formats such as visual aids, rules, numbers and different combinations of these. For each of these formats, scholars designed, deployed and tested several solutions, such as saliency masks, attention maps, heat-maps, feature maps, as well as graphs, rules sets, trees and dialogues. These advances were aimed at meeting the needs of different types of end-users operating in various fields, such as laypeople, doctors and lawyers, and adapting explanations to their domains of application. Additionally, scholars widened their research horizons by incorporating the knowledge developed in other fields, like Psychology, Philosophy and Social Sciences, into the design of the novel methods for explainability. The goal was to improve the structure, efficiency and efficacy on people’s understanding of automatically generated explanations. All this research has produced many definitions of explainability and identified several notions related to it, such as interpretability, understandability, comprehensibility and justifiability, just to mention a few. Coupled to these notions, different objective metrics have also been produced for their measurement. Despite the large number and variety of methods and metrics for explainability proposed so far, there are still important scientific issues that must be tackled. Firstly, there is no agreement among scholars on what an explanation exactly is and which are the salient properties that should be considered to make it effective and understandable for end-users, in particular non-experts. Secondly, the construct of explainability is a concept borrowed from Psychology, since it is strictly connected to humans, and it is also linked to other constructs such as trust, transparency and privacy. Thirdly, this concept has been invoked in several fields, such as Physics, Mathematics, Social Sciences and Medicine. All this make its formalisation and operationalisation a non-trivial research task. This holds true for every explanation format, either textual, visual, numerical. The same can be said for rule-based explanations, in particular for those that are generated after a model has been induced by employing deep-learning neural networks. In accordance with [1], we believe that scholars have produced enough material that can be used to construct a generally applicable framework for XAI. This would guide the advancement of novel, end-to-end methods for explainability, rather than keep creating isolated methods that remain only fragments of a broad solution which should also be flexible enough

to adapt to various contexts, fields of application and type of end-users. Additionally, the ultimate scope of an explanation is to help end-users build a complete and correct mental model of the inferential process of either a learning algorithm or a knowledge-based system and to promote trust for its outputs. An area for future research is the involvement of humans, as final users of artificial explanations, since their role has not been sufficiently studied in the creation and exploitation of existing explainability methods [1]. To support this research direction, we recommend exploiting knowledge and experiences belonging to the field of Human-Computer Interaction and its advances for the development of interactive explanatory interfaces [356]. This should always take into consideration the existing trade-off between the dimensions of models accuracy and their interpretability/explainability which are currently inversely correlated. One possible suggestion is the use of methods that take advantage of modern learning techniques, to maximise the former dimension, and reasoning approaches to optimise the latter dimension. The assumption is that integrating connectionist and symbolic paradigms is the most efficient way to produce meaningful and precise explanations. Advances on these two paradigms are immense, however, their intersection is under exploration. For example, on one hand, a school of thought suggests to firstly train accurate models from data and then wrap them with a reasoning layer [165]. This layer, for instance, can be produced by exploiting advances in defeasible reasoning and argumentation [143, 142, 279] making use of knowledge-bases constructed with a human-in-the-loop approach. On the other hand, another direction is to promote the use of neuro-symbolic learning and reasoning in parallel, each one informing the other at all stages of model construction and evaluation [45]. Eventually, another interesting, novel and under-explored direction for future scholars concerns the development of structured formats of explanations. These formats should consider all the elements and notions related to explainability, that can be trained with connectionist paradigms from data.

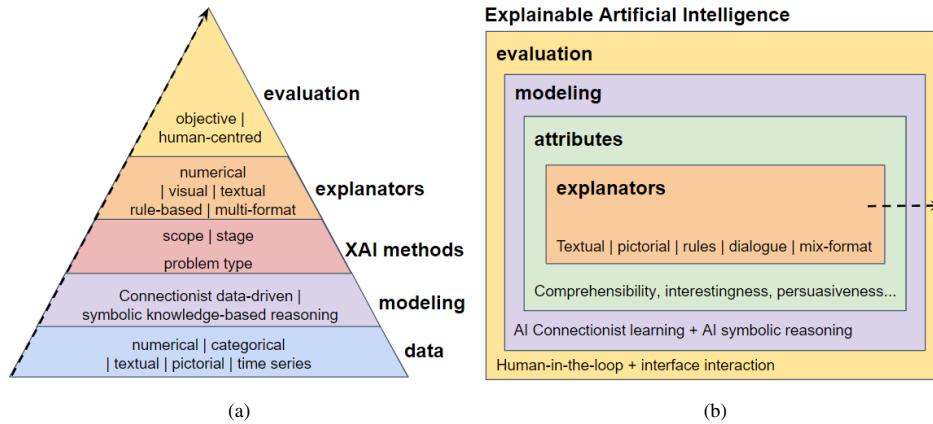


Figure 18: State of the art (a) and envisioned (b) frameworks for eXplainable Artificial Intelligence.

In summary, an high-level structure of a the current state-of-the-are in XAI is depicted in figure 18 (part a). On one hand, here, emphasis has been placed on the sequence of research activities currently and often performed by several scholars, their dependencies and order. This sequence usually starts from input data that is then used for modeling purposes, employing connectionist data-driven learning or symbolic reasoning knowledge-based paradigms. After a model has been formed, then an XAI methods is applied for its analysis, knowledge discovery, supporting its interpretability. This phase provide the end-users of these models with one or more explainers for the purpose of its explainability. Eventually, very few scholars have proposed approaches for evaluating such layer of explainability, either proposing formal, objective metrics or involving human-centred evaluation with model designers and end-users. On the other end, what we believe is an ideal framework for XAI is depicted in figure 18 (part b). Here, the main focus should be on the explainers, which is what end-users will ultimately interact with. The development of explainers should be designed by taking into account the multiple attributes that are linked to the psychological construct of explainability. Subsequently, scholars can focus on the modeling phase, preferably using both connectionist and symbolic paradigms from Artificial Intelligence. This will allow to develop models that are both robust in terms of accuracy but also intrinsically interpretable during all the stages of development. Eventually, the last phase should focus on the evaluation of explainability of such models with a human-in-the-loop approach, involving both designers and end-users, and the development of interactive interface for supporting model interpretability and inference explainability.

## References

- [1] Amina Adadi and Mohammed Berrada. Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, 6:52138–52160, 2018. doi: 10.1109/ACCESS.2018.2870052.
- [2] Alun Preece. Asking “why” in ai: Explainability of intelligent systems—perspectives and challenges. *Intelligent Systems in Accounting, Finance and Management*, 25(2):63–72, 2018. doi: 10.1002/isaf.1422.
- [3] Edward H Shortliffe, Randall Davis, Stanton G Axline, Bruce G Buchanan, C Cordell Green, and Stanley N Cohen. Computer-based consultations in clinical therapeutics: explanation and rule acquisition capabilities of the mycin system. *Computers and biomedical research*, 8(4):303–320, 1975. doi: 10.1016/0010-4809(75)90009-9.
- [4] Vanya Van Belle and Paulo Lisboa. Research directions in interpretable machine learning models. In *21st European Symposium on Artificial Neural Networks, ESANN*, pages 533–541, Bruges, Belgium, 2013. i6doc.
- [5] Jonathan Dodge, Sean Penney, Andrew Anderson, and Margaret M Burnett. What should be in an xai explanation? what ift reveals. In *IUI Workshop 7: Explainable Smart Systems - ExSS*, Tokyo, Japan, 2018. CEUR-WS.org.
- [6] Alfredo Vellido, José David Martín-Guerrero, and Paulo JG Lisboa. Making machine learning models interpretable. In *European Symposium on Artificial Neural Networks, ESANN*, volume 12, pages 163–172, Bruges, Belgium, 2012. i6doc.
- [7] Weiquan Wang and Izak Benbasat. Recommendation agents for electronic commerce: Effects of explanation facilities on trusting beliefs. *Journal of Management Information Systems*, 23(4):217–246, 2007. doi: 10.2753/mis0742-1222230410.
- [8] Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Unmasking clever hans predictors and assessing what machines really learn. *Nature communications*, 10(1):1096, 2019. doi: 10.1038/s41467-019-08987-4.
- [9] Cynthia Rudin. Algorithms for interpretable machine learning. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1519–1519, New York, New York, USA, 2014. ACM. doi: 10.1145/2623330.2630823.
- [10] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206, 2019. doi: 10.1038/s42256-019-0048-x.
- [11] Jean-Marc Fellous, Guillermo Sapiro, Andrew Rossi, Helen S Mayberg, and Michele Ferrante. Explainable artificial intelligence for neuroscience: Behavioral neurostimulation. *Frontiers in Neuroscience*, 13:1346, 2019. doi: 10.3389/fnins.2019.01346.
- [12] Maria Fox, Derek Long, and Daniele Magazzeni. Explainable planning. In *IJCAI 2017 Workshop on Explainable Artificial Intelligence (XAI)*, pages 24–30, Melbourne, Australia, 2017. International Joint Conferences on Artificial Intelligence, Inc.
- [13] Filip Karlo Došilović, Mario Brčić, and Nikica Hlupić. Explainable artificial intelligence: A survey. In *41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 0210–0215, Opatija, Croatia, 2018. IEEE. doi: 10.23919/mipro.2018.8400040.
- [14] Eva Thelisson, Kirtan Padh, and L Elisa Celis. Regulatory mechanisms and algorithms towards trust in ai/ml. In *IJCAI Workshop on Explainable AI (XAI)*, pages 53–57, Melbourne, Australia, 2017. International Joint Conferences on Artificial Intelligence, Inc.
- [15] Eva Thelisson. Towards trust, transparency, and liability in ai/as systems. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 5215–5216, Melbourne, Australia, 2017. International Joint Conferences on Artificial Intelligence, Inc. doi: 10.24963/IJCAI.2017/767.
- [16] Sandra Wachter, Brent Mittelstadt, and Luciano Floridi. Transparent, explainable, and accountable ai for robotics. *Science Robotics*, 2(6), 2017. doi: 10.1126/scirobotics.aan6080.
- [17] Wojciech Samek and Klaus-Robert Müller. Towards explainable artificial intelligence. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pages 5–22. Springer, Cham, Switzerland, 2019. doi: 10.1007/978-3-030-28954-6\1.
- [18] Carmen Lacave and Francisco J Díez. A review of explanation methods for bayesian networks. *The Knowledge Engineering Review*, 17(2):107–127, 2002. doi: 10.1017/s026988890200019x.
- [19] David Martens, Bart Baesens, Tony Van Gestel, and Jan Vanthienen. Comprehensible credit scoring models using rule extraction from support vector machines. *European journal of operational research*, 183(3):1466–1476, 2007. doi: 10.1016/j.ejor.2006.04.051.
- [20] Jaegul Choo and Shixia Liu. Visual analytics for explainable deep learning. *IEEE Computer Graphics and Applications*, 38(4):84–92, 2018. doi: 10.1109/mcg.2018.042731661.
- [21] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):93:1–93:42, 2018. doi: 10.1145/3236009.
- [22] Tim Miller. Explanation in artificial intelligence: insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019. doi: 10.1016/j.artint.2018.07.007.

- [23] Tim Miller, Piers Howe, and Liz Sonenberg. Explainable ai: Beware of inmates running the asylum or: How i learnt to stop worrying and love the social and behavioural sciences. In *IJCAI Workshop on Explainable AI (XAI)*, pages 36–42, Melbourne, Australia, 2017. International Joint Conferences on Artificial Intelligence, Inc.
- [24] Henrik Jacobsson. Rule extraction from recurrent neural networks: A taxonomy and review. *Neural Computation*, 17(6):1223–1263, 2005. doi: 10.1162/0899766053630350.
- [25] Leila Arras, Ahmed Osman, Klaus-Robert Müller, and Wojciech Samek. Evaluating recurrent neural network explanations. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 113–126, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4813.
- [26] Hoa Khanh Dam, Truyen Tran, and Aditya Ghose. Explainable software analytics. In *Proceedings of the 40th International Conference on Software Engineering: New Ideas and Emerging Results*, pages 53–56, Gothenburg, Sweden, 2018. ACM. doi: 10.1145/3183399.3183424.
- [27] Robert R Hoffman, Gary Klein, and Shane T Mueller. Explaining explanation for “explainable ai”. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 62, pages 197–201, Philadelphia, Pennsylvania, USA, 2018. SAGE Publications Sage CA: Los Angeles, CA. doi: 10.1177/1541931218621047.
- [28] Sherin Mary Mathews. Explainable artificial intelligence applications in nlp, biomedical, and malware classification: A literature review. In *Intelligent Computing-Proceedings of the Computing Conference*, pages 1269–1292, London, United Kingdom, 2019. Springer. doi: 10.1007/978-3-030-22868-2\\_90.
- [29] Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73:1–15, 2017. doi: 10.1016/j.dsp.2017.10.011.
- [30] Quan-shi Zhang and Song-Chun Zhu. Visual interpretability for deep learning: a survey. *Frontiers of Information Technology & Electronic Engineering*, 19(1):27–39, 2018.
- [31] Andreas Backhaus and Udo Seiffert. Classification in high-dimensional spectral data: Accuracy vs. interpretability vs. model size. *Neurocomputing*, 131:15–22, 2014. doi: 10.1016/j.neucom.2013.09.048.
- [32] William R Swartout and Johanna D Moore. Explanation in second generation expert systems. In *Second generation expert systems*, pages 543–585. Springer, Berlin, Germany, 1993. doi: 10.1007/978-3-642-77927-5\\_\\_24.
- [33] Shirley Gregor and Izak Benbasat. Explanations from intelligent systems: Theoretical foundations and implications for practice. *MIS quarterly*, 23(4):497–530, 1999. doi: 10.2307/249487.
- [34] Alexis Papadimitriou, Panagiotis Symeonidis, and Yannis Manolopoulos. A generalized taxonomy of explanations styles for traditional and social recommender systems. *Data Mining and Knowledge Discovery*, 24(3):555–583, 2012. doi: 10.1007/s10618-011-0215-0.
- [35] Frode Sørmo, Jörg Cassens, and Agnar Aamodt. Explanation in case-based reasoning—perspectives and goals. *Artificial Intelligence Review*, 24(2):109–143, 2005. doi: 10.1007/S10462-005-4607-7.
- [36] R Byrne. Counterfactuals in explainable artificial intelligence (xai): evidence from human reasoning. In *International Joint Conference on AI (IJCAI)*, pages 1438–1444, Macao, China, 2019. International Joint Conferences on Artificial Intelligence, Inc. doi: 10.24963/ijcai.2019/199.
- [37] Corrado Mencar and José M Alonso. Paving the way to explainable artificial intelligence with fuzzy modeling. In *International Workshop on Fuzzy Logic and Applications*, pages 215–227, Santa Margherita Ligure, Italy, 2018. Springer. doi: 10.1007/978-3-030-12544-8\\_\\_17.
- [38] Vaishak Belle. Logic meets probability: towards explainable ai systems for uncertain worlds. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pages 19–25, Melbourne, Australia, 2017. International Joint Conferences on Artificial Intelligence, Inc. doi: 10.24963/ijcai.2017/733.
- [39] Mark W Craven and Jude W Shavlik. Visualizing learning and computation in artificial neural networks. *International journal on artificial intelligence tools*, 1(03):399–425, 1992. doi: 10.1142/s0218213092000260.
- [40] Paulo JG Lisboa. Interpretability in machine learning—principles and practice. In *International Workshop on Fuzzy Logic and Applications*, pages 15–21, Genoa, Italy, 2013. Springer. doi: 10.1007/978-3-319-03200-9\\_\\_2.
- [41] Tameru Hailesilassie. Rule extraction algorithm for deep neural networks: A review. *(IJCSIS) International Journal of Computer Science and Information Security*, 14(7):376–381, 2016.
- [42] Alberto Fernandez, Francisco Herrera, Oscar Cordon, Maria Jose del Jesus, and Francesco Marcelloni. Evolutionary fuzzy systems for explainable artificial intelligence: Why, when, what for, and where to? *IEEE Computational Intelligence Magazine*, 14(1):69–81, 2019.
- [43] Serge Guillaume. Designing fuzzy inference systems from data: An interpretability-oriented review. *IEEE transactions on fuzzy systems*, 9(3):426–443, 2001. doi: 10.1109/91.928739.
- [44] Tarek R Besold and Kai-Uwe Kühnberger. Towards integrated neural-symbolic systems for human-level ai: Two research programs helping to bridge the gaps. *Biologically Inspired Cognitive Architectures*, 14:97–110, 2015. doi: 10.1016/j.bica.2015.09.003.
- [45] Artur d’Avila Garcez, Tarek R Besold, Luc De Raedt, Peter Földiak, Pascal Hitzler, Thomas Icard, Kai-Uwe Kühnberger, Luis C Lamb, Risto Miikkulainen, and Daniel L Silver. Neural-symbolic learning and reasoning: contributions and challenges. In *AAAI Spring Symposium Series*, pages 20–23, Palo Alto, California, USA, 2015.

- AAAI Press.
- [46] Maria Paola Bonacina. Automated reasoning for explainable artificial intelligence. In *The First International ARCADE (Automated Reasoning: Challenges, Applications, Directions, Exemplary Achievements) Workshop (in association with CADE-26)*, pages 24–28, Gothenburg, Sweden, 2017. ARCADE@ CADE. doi: 10.29007/4b7h.
  - [47] Clemens Otte. Safe and interpretable machine learning: a methodological review. In *Computational intelligence in intelligent data analysis*, pages 111–122. Springer, Berlin, Germany, 2013. doi: 10.1007/978-3-642-32378-2\\_8.
  - [48] Alex A Freitas, Daniela C Wieser, and Rolf Apweiler. On the importance of comprehensible classification models for protein function prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 7(1):172–182, 2010. doi: 10.1109/tcbb.2008.47.
  - [49] David Martens, Jan Vanthienen, Wouter Verbeke, and Bart Baesens. Performance of classification models from a user perspective. *Decision Support Systems*, 51(4):782–793, 2011. doi: 10.1016/j.dss.2011.01.013.
  - [50] Or Biran and Courtenay Cotton. Explanation and justification in machine learning: A survey. In *IJCAI 2017 Workshop on Explainable Artificial Intelligence (XAI)*, pages 8–13, Melbourne, Australia, 2017. International Joint Conferences on Artificial Intelligence, Inc.
  - [51] Krishna Gade, Sahin Cem Geyik, Krishnaram Kenthapadi, Varun Mithal, and Ankur Taly. Explainable ai in industry. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3203–3204, Anchorage, Alaska, USA, 2019. ACM.
  - [52] Feiyu Xu, Hans Uszkoreit, Yangzhou Du, Wei Fan, Dongyan Zhao, and Jun Zhu. Explainable ai: a brief survey on history, research areas, approaches and challenges. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 563–574, Switzerland, 2019. Springer, Cham. doi: 10.1007/978-3-030-32236-6\\_.51.
  - [53] Xiaocong Cui, Jung Min Lee, and J Hsieh. An integrative 3c evaluation framework for explainable artificial intelligence. In *AI and semantic technologies for intelligent information systems (SIGODIS)*, pages 1–10, Cancún, Mexico, 2019. AIS eLibrary.
  - [54] Mary T Dzindolet, Scott A Peterson, Regina A Pomranky, Linda G Pierce, and Hall P Beck. The role of trust in automation reliance. *International journal of human-computer studies*, 58(6):697–718, 2003. doi: 10.1016/s1071-5819(03)00038-7.
  - [55] Nava Tintarev and Judith Masthoff. A survey of explanations in recommender systems. In *IEEE 23rd international conference on data engineering workshop*, pages 801–810, Istanbul, Turkey, 2007. IEEE. doi: 10.1109/icdew.2007.4401070.
  - [56] Zachary C Lipton. The mythos of model interpretability. *Commun. ACM*, 61(10):36–43, 2018.
  - [57] Taehyun Ha, Sangwon Lee, and Sangyeon Kim. Designing explainability of an artificial intelligence system. In *Proceedings of the Technology, Mind, and Society*, page 14:1, Washington, District of Columbia, USA, 2018. ACM. doi: 10.1145/3183654.3183683.
  - [58] Urszula Chajewska and Joseph Y Halpern. Defining explanation in probabilistic systems. In *Proceedings of the Thirteenth conference on Uncertainty in artificial intelligence*, pages 62–71, Providence, Rhode Island, USA, 1997. Morgan Kaufmann Publishers Inc.
  - [59] Andreas Holzinger, Georg Langs, Helmut Denk, Kurt Zatloukal, and Heimo Müller. Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9: e1312, 2019. doi: 10.1002/widm.1312.
  - [60] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th international conference on intelligent user interfaces*, pages 126–137, Atlanta, Georgia, USA, 2015. ACM. doi: 10.1145/2678025.2701399.
  - [61] Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. Too much, too little, or just right? ways explanations impact end users’ mental models. In *IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*, pages 3–10, Raleigh, NC, USA, 2013. IEEE. doi: 10.1109/vlhcc.2013.6645235.
  - [62] Irit Aszira-Gelman. Knowledge discovery: comprehensibility of the results. In *Proceedings of the Thirty-First Hawaii International Conference on System Sciences*, volume 5, pages 247–255, Hawaii, 1998. IEEE. doi: 10.1109/hicss.1998.648319.
  - [63] Jose M Alonso, Ciro Castiello, and Corrado Mencar. A bibliometric analysis of the explainable artificial intelligence research field. In *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 3–15, Cádiz, Spain, 2018. Springer.
  - [64] Adrien Bibal and Benoît Frénay. Interpretability of machine learning models and representations: an introduction. In *Proceedings on the European Symposium on Artificial Neural Networks, ESANN*, pages 77–82, Bruges, Belgium, 2016. i6doc.
  - [65] Ivan Bratko. Machine learning: Between accuracy and interpretability. In *Learning, networks and statistics*, pages 163–177. Springer, Vienna, Austria, 1997. doi: 10.1007/978-3-7091-2668-4\\_.10.
  - [66] Derek Doran, Sarah Schulz, and Tarek R Besold. What does explainable ai really mean? a new conceptualiza-

- tion of perspectives. In *16th International Conference of the Italian Association of Artificial Intelligence, 2017. Workshop on Comprehensibility and Explanation in AI and ML*, pages 1–8, Bari, Italy, 2017. Cex, University of Bremen, Germany.
- [67] Alex A Freitas. Are we really discovering interesting knowledge from data. *Expert Update (the BCS-SGAI magazine)*, 9(1):41–47, 2006.
  - [68] Randy Goebel, Ajay Chander, Katharina Holzinger, Freddy Lecue, Zeynep Akata, Simone Stumpf, Peter Kieseberg, and Andreas Holzinger. Explainable ai: the new 42? In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, pages 295–303, Hamburg, Germany, 2018. Springer. doi: 10.1007/978-3-319-99740-7\\_21.
  - [69] David S Watson, Jenny Krutzinna, Ian N Bruce, Christopher EM Griffiths, Iain B McInnes, Michael R Barnes, and Luciano Floridi. Clinical applications of machine learning algorithms: beyond the black box. *BMJ*, 364:l886, 2019. doi: 10.1136/bmj.l886.
  - [70] Nava Tintarev and Judith Masthoff. Designing and evaluating explanations for recommender systems. In *Recommender systems handbook*, pages 479–510. Springer, Boston, Massachusetts, USA, 2011. doi: 10.1007/978-0-387-85820-3\\_15.
  - [71] Nava Tintarev and Judith Masthoff. Explaining recommendations: Design and evaluation. In *Recommender systems handbook*, pages 353–382. Springer, Boston, Massachusetts, USA, 2015. doi: 10.1007/978-1-4899-7637-6\\_10.
  - [72] Ajay Chander and Ramya Srinivasan. Evaluating explanations by cognitive value. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, pages 314–328, Hamburg, Germany, 2018. Springer.
  - [73] Yu Zhang, Sarath Sreedharan, Anatha Kulkarni, Tathagata Chakraborti, Hankz Hankui Zhuo, and Subbarao Kambhampati. Plan explicability and predictability for robot task planning. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 1313–1320, Singapore, 2017. IEEE. doi: 10.1109/icra.2017.7989155.
  - [74] David Alvarez-Melis and Tommi S. Jaakkola. Towards robust interpretability with self-explaining neural networks. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 7786–7795, Montréal, Canada, 2018. Neural Information Processing Systems Foundation, Inc.
  - [75] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y Lim, and Mohan Kankanhalli. Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 582–599, Montréal, Canada, 2018. ACM. doi: 10.1145/3173574.3174156.
  - [76] Michael Chromik, Malin Eiband, Sarah Theres Völkel, and Daniel Buschek. Dark patterns of explainability, transparency, and user control for intelligent systems. In *Joint Proceedings of the ACM IUI 2019 Workshops co-located with the 24th ACM Conference on Intelligent User Interfaces (ACM IUI)*, volume 2327, Los Angeles, California, USA, 2019. CEUR-WS.org.
  - [77] Brian Y Lim, Anind K Dey, and Daniel 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*, pages 2119–2128, Boston, Massachusetts, USA, 2009. ACM. doi: 10.1145/1518701.1519023.
  - [78] Brian Y Lim, Qian Yang, Ashraf M Abdul, and Danding Wang. Why these explanations? selecting intelligibility types for explanation goals. In *Joint Proceedings of the ACM IUI 2019 Workshops co-located with the 24th ACM Conference on Intelligent User Interfaces (ACM IUI)*, volume 2327, Los Angeles, California, USA, 2019. CEUR-WS.org.
  - [79] Johanna D Moore and Cécile L Paris. Planning text for advisory dialogues: Capturing intentional and rhetorical information. *Computational linguistics*, 19(4):651–694, 1993.
  - [80] Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere. A grounded interaction protocol for explainable artificial intelligence. In *Proceedings of the 18th International Conference on Autonomous Agents and Multi-Agent Systems*, pages 1033–1041, Montréal, Canada, 2019. International Foundation for Autonomous Agents and Multiagent Systems.
  - [81] Alex A Freitas. On rule interestingness measures. In *Research and Development in Expert Systems XV*, pages 147–158. Springer, United Kingdom, 1999. doi: 10.1016/s0950-7051(99)00019-2.
  - [82] Pedro Sequeira, Eric Yeh, and Melinda T Gervasio. Interestingness elements for explainable reinforcement learning through introspection. In *Joint Proceedings of the ACM IUI 2019 Workshops co-located with the 24th ACM Conference on Intelligent User Interfaces (ACM IUI)*, volume 2327, Los Angeles, California, USA, 2019. CEUR-WS.org.
  - [83] André Carrington, Paul Fieguth, and Helen Chen. Measures of model interpretability for model selection. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, pages 329–349, Hamburg, Germany, 2018. Springer.
  - [84] Isabel Sassoon, Nadin Köcryan, Elizabeth Sklar, and Simon Parsons. Explainable argumentation for wellness

- consultation. In *International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems*, pages 186–202, Switzerland, 2019. Springer, Cham.
- [85] Mukund Sundararajan, Jinhua Xu, Ankur Taly, Rory Sayres, and Amir Najmi. Exploring principled visualizations for deep network attributions. In *Joint Proceedings of the ACM IUI 2019 Workshops co-located with the 24th ACM Conference on Intelligent User Interfaces (ACM IUI)*, volume 2327, Los Angeles, California, USA, 2019. CEUR-WS.org.
- [86] Shang-Ming Zhou and John Q Gan. Low-level interpretability and high-level interpretability: a unified view of data-driven interpretable fuzzy system modelling. *Fuzzy Sets and Systems*, 159(23):3091–3131, 2008. doi: 10.1016/j.fss.2008.05.016.
- [87] Claus Weihs and UM Sondhauss. Combining mental fit and data fit for classification rule selection. In *Exploratory Data Analysis in Empirical Research*, pages 188–203. Springer, Munich, Germany, 2003. doi: 10.1007/978-3-642-55721-7\_21.
- [88] Alex A Freitas. Comprehensible classification models: a position paper. *ACM SIGKDD explorations newsletter*, 15(1):1–10, 2014. doi: 10.1145/2594473.2594475.
- [89] Shixia Liu, Xiting Wang, Mengchen Liu, and Jun Zhu. Towards better analysis of machine learning models: A visual analytics perspective. *Visual Informatics*, 1(1):48–56, 2017. doi: 10.1016/j.visinf.2017.01.006.
- [90] David Alvarez-Melis and Tommi S. Jaakkola. On the robustness of interpretability methods. In *Proceedings of the 2018 ICML Workshop in Human Interpretability in Machine Learning*, pages 66–71, Stockholm, Sweden, 2018. ICML.
- [91] Rowan McAllister, Yarin Gal, Alex Kendall, Mark Van Der Wilk, Amar Shah, Roberto Cipolla, and Adrian Vivian Weller. Concrete problems for autonomous vehicle safety: Advantages of bayesian deep learning. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pages 4745–4753, Melbourne, Australia, 2017. International Joint Conferences on Artificial Intelligence, Inc. doi: 10.24963/ijcai.2017/661.
- [92] Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. The (un)reliability of saliency methods. In *NIPS workshop on Explaining and Visualizing Deep Learning*, pages 93–101, Long Beach, California, USA, 2017. NIPS.
- [93] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3319–3328, Sydney, Australia, 2017. JMLR.org.
- [94] Fabian Offert. “i know it when i see it”. visualization and intuitive interpretability. In *NIPS Symposium on Interpretable Machine Learning*, pages 43–46, Long Beach, California, USA, 2017. NIPS.
- [95] Joseph B Lyons. Being transparent about transparency. In *AAAI Spring Symposium*, pages 48–53, Palo Alto, California, USA, 2013. AAAI Press.
- [96] Adrian Weller. Challenges for transparency. In *Proceedings of the ICML Workshop on Human Interpretability in Machine Learning*, pages 55–62, Sydney, Australia, 2017. ICML.
- [97] Andrés Páez. The pragmatic turn in explainable artificial intelligence (xai). *Minds and Machines*, 29:1–19, 2019. doi: 10.1007/s11023-019-09502-w.
- [98] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *Proceedings of ICLR Workshop*, Banff, Canada, 2014. ICLR.
- [99] Yin Lou, Rich Caruana, and Johannes Gehrke. Intelligible models for classification and regression. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 150–158, Beijing, China, 2012. ACM. doi: 10.1145/2339530.2339556.
- [100] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3145–3153, Sydney, Australia, 2017. JMLR.org.
- [101] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.
- [102] Jonathan L Herlocker, Joseph A Konstan, and John Riedl. Explaining collaborative filtering recommendations. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work*, pages 241–250, Philadelphia, Pennsylvania, USA, 2000. ACM. doi: 10.1145/358916.358995.
- [103] Josua Krause, Adam Perer, and Kenney 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*, pages 5686–5697, San Jose, California, USA, 2016. ACM. doi: 10.1145/2858036.2858529.
- [104] Mireia Ribera and Agata Lapedriza. Can we do better explanations? a proposal of user-centered explainable ai. In *Joint Proceedings of the ACM IUI 2019 Workshops co-located with the 24th ACM Conference on Intelligent User Interfaces (ACM IUI)*, volume 2327, Los Angeles, California, USA, 2019. CEUR-WS.org.
- [105] MM de Graaf and Bertram F Malle. How people explain action (and autonomous intelligent systems should too). In *AAAI Fall Symposium on Artificial Intelligence for Human-Robot Interaction*, pages 19–26, Arlington,

- Virginia, USA, 2017. AAAI Press.
- [106] Maaike Harbers, Karel van den Bosch, and John-Jules Ch Meyer. A study into preferred explanations of virtual agent behavior. In *International Workshop on Intelligent Virtual Agents*, pages 132–145, Amsterdam, Netherlands, 2009. Springer. doi: 10.1007/978-3-642-04380-2\\_17.
- [107] Michael R Wick and William B Thompson. Reconstructive explanation: Explanation as complex problem solving. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, pages 135–140, Detroit, Michigan, USA, 1989. International Joint Conferences on Artificial Intelligence, Inc.
- [108] Michael R Wick. Second generation expert system explanation. In *Second Generation Expert Systems*, pages 614–640. Springer, Berlin, Germany, 1993. doi: 10.1007/978-3-642-77927-5\\_26.
- [109] Steven R Haynes, Mark A Cohen, and Frank E Ritter. Designs for explaining intelligent agents. *International Journal of Human-Computer Studies*, 67(1):90–110, 2009. doi: 10.1016/j.ijhcs.2008.09.008.
- [110] Raymond Sheh and Isaac Monteath. Introspectively assessing failures through explainable artificial intelligence. In *IROS Workshop on Introspective Methods for Reliable Autonomy*, pages 40–47, Vancouver, Canada, 2017. iliad-project.eu.
- [111] Regina Barzilay, Daryl McCullough, Owen Rambow, Jonathan DeCristofaro, Tanya Korelsky, and Benoit Lavoie. A new approach to expert system explanations. In *Proceedings of the Ninth International Workshop on Natural Language Generation*, pages 78–87, Niagara-on-the-Lake, Ontario, Canada, 1998. Association for Computational Linguistics.
- [112] Tania Lombrozo. The structure and function of explanations. *Trends in cognitive sciences*, 10(10):464–470, 2006. doi: 10.1016/j.tics.2006.08.004.
- [113] JL Weiner. Blah, a system which explains its reasoning. *Artificial intelligence*, 15(1-2):19–48, 1980. doi: 10.1016/0004-3702(80)90021-1.
- [114] Douglas Walton. A dialogue system specification for explanation. *Synthese*, 182(3):349–374, 2011.
- [115] Alison Cawsey. Generating interactive explanations. In *Proceedings of the 9th National Conference on Artificial Intelligence*, volume 1, pages 86–91, Anaheim, California, USA, 1991. Citeseer.
- [116] Alison Cawsey. Planning interactive explanations. *International Journal of Man-Machine Studies*, 38(2):169–199, 1993. doi: 10.1006/imms.1993.1009.
- [117] Alison Cawsey. User modelling in interactive explanations. *User Modeling and User-Adapted Interaction*, 3(3): 221–247, 1993. doi: 10.1007/bf01257890.
- [118] Martha E Pollack, Julia Hirschberg, and Bonnie Webber. User participation in the reasoning processes of expert systems. In *Proceedings of Second National Conference Artificial Intelligence*, pages 358–361, Menlo Park, California, USA, 1982. MIT Press, Cambridge, Massachusetts.
- [119] Hilary Johnson and Peter Johnson. Explanation facilities and interactive systems. In *Proceedings of the 1st international conference on Intelligent user interfaces*, pages 159–166, Orlando, Florida, USA, 1993. ACM. doi: 10.1145/169891.169951.
- [120] Johanna D Moore and Cecile L Paris. Planning text for advisory dialogues. In *Proceedings of the 27th annual meeting on Association for Computational Linguistics*, pages 203–211, Vancouver, British Columbia, Canada, 1989. Association for Computational Linguistics. doi: 10.3115/981623.981648.
- [121] Johanna D Moore and William R Swartout. A reactive approach to explanation. In *IJCAI*, pages 1504–1510, Detroit, Michigan, USA, 1989. International Joint Conferences on Artificial Intelligence, Inc.
- [122] Johanna D Moore and William R Swartout. A reactive approach to explanation: Taking the user's feedback into account. In *Natural language generation in artificial intelligence and computational linguistics*, pages 3–48. Springer, USA, 1991.
- [123] Mark G Core, H Chad Lane, Michael Van Lent, Dave Gomboc, Steve Solomon, and Milton Rosenberg. Building explainable artificial intelligence systems. In *Proceedings of the 21st national conference on Artificial intelligence*, pages 1766–1773, Boston, Massachusetts, USA, 2006. AAAI Press. doi: 10.21236/ada459166.
- [124] Dave Gomboc, Steve Solomon, Mark G Core, H Chad Lane, and Michael Van Lent. Design recommendations to support automated explanation and tutoring. In *Proceedings of the Fourteenth Conference on Behavior Representation in Modelling and Simulation*, pages 331–340, Universal City, California, USA, 2005. Simulation Interoperability Standards Organization (SISO).
- [125] H Chad Lane, Mark G Core, Michael Van Lent, Steve Solomon, and Dave Gomboc. Explainable artificial intelligence for training and tutoring. In *Proceedings of the 12th International Conference on Artificial Intelligence in Education, AIED*, pages 762–764, Amsterdam, The Netherlands, 2005. IOS Press.
- [126] Michael Van Lent, William Fisher, and Michael Mancuso. An explainable artificial intelligence system for small-unit tactical behavior. In *Proceedings of the National Conference on Artificial Intelligence*, pages 900–907, San Jose, California, USA, 2004. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.
- [127] Arthur C Graesser, Patrick Chipman, Brian C Haynes, and Andrew Olney. Autotutor: An intelligent tutoring system with mixed-initiative dialogue. *IEEE Transactions on Education*, 48(4):612–618, 2005. doi: 10.1109/te.2005.856149.

- [128] Pat Langley, Ben Meadows, Mohan Sridharan, and Dongkyu Choi. Explainable agency for intelligent autonomous systems. In *Proceedings of the Thirty-First Conference on Artificial Intelligence*, pages 4762–4764, San Francisco, California, USA, 2017. AAAI Press.
- [129] Shirin Sohrabi, Jorge A Baier, and Sheila A McIlraith. Preferred explanations: Theory and generation via planning. In *Proceedings of the Twenty-Fifth Conference on Artificial Intelligence*, pages 261–267, San Francisco, California, USA, 2011. AAAI Press.
- [130] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013. doi: 10.1109/tpami.2013.50.
- [131] Yin Lou, Rich Caruana, Johannes Gehrke, and Giles Hooker. Accurate intelligible models with pairwise interactions. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 623–631, Chicago, Illinois, USA, 2013. ACM. doi: 10.1145/2487575.2487579.
- [132] Guido Bologna and Yoichi Hayashi. A comparison study on rule extraction from neural network ensembles, boosted shallow trees, and svms. *Applied Computational Intelligence and Soft Computing*, 2018, 2018. doi: 10.1155/2018/4084850.
- [133] Thilo Spinner, Udo Schlegel, Hanna Schäfer, and Mennatallah El-Assady. explainer: A visual analytics framework for interactive and explainable machine learning. *IEEE transactions on visualization and computer graphics*, 26: 1064–1074, 2019. doi: 10.1109/TVCG.2019.2934629.
- [134] Marco Túlio Ribeiro, Sameer Singh, and Carlos 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*, pages 1135–1144, San Francisco, CA, USA, 2016. ACM. doi: 10.1145/2939672.2939778.
- [135] Hendrik Strobelt, Sebastian Gehrmann, Hanspeter Pfister, and Alexander M Rush. Lstmvis: A tool for visual analysis of hidden state dynamics in recurrent neural networks. *IEEE transactions on visualization and computer graphics*, 24(1):667–676, 2018. doi: 10.1109/tvcg.2017.2744158.
- [136] Kanit Wongsuphasawat, Daniel Smilkov, James Wexler, Jimbo Wilson, Dandelion Mané, Doug Fritz, Dilip Krishnan, Fernanda B Viégas, and Martin Wattenberg. Visualizing dataflow graphs of deep learning models in tensorflow. *IEEE transactions on visualization and computer graphics*, 24(1):1–12, 2018. doi: 10.1109/tvcg.2017.2744878.
- [137] Lisa Anne Hendricks, Ronghang Hu, Trevor Darrell, and Zeynep Akata. Grounding visual explanations. In *Computer Vision - ECCV 2018 - 15th European Conference, Proceedings, Part II*, pages 269–286, Munich, Germany, 2018. Springer. doi: 10.1007/978-3-030-01216-8\\_\\_17.
- [138] Glenn Fung, Sathyakama Sandilya, and R Bharat Rao. Rule extraction from linear support vector machines. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 32–40, Chicago, Illinois, USA, 2005. ACM. doi: 10.1145/1081870.1081878.
- [139] Guido Bologna and Yoichi Hayashi. Characterization of symbolic rules embedded in deep dimlp networks: a challenge to transparency of deep learning. *Journal of Artificial Intelligence and Soft Computing Research*, 7(4): 265–286, 2017. doi: 10.1515/jaiscr-2017-0019.
- [140] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In *Thirty-Second AAAI Conference on Artificial Intelligence*, pages 1527–1535, New Orleans, Louisiana, USA, 2018. AAAI Press.
- [141] Vasile Palade, Daniel-Ciprian Neagu, and Ron J Patton. Interpretation of trained neural networks by rule extraction. In *International Conference on Computational Intelligence*, pages 152–161, Dortmund, Germany, 2001. Springer. doi: 10.1007/3-540-45493-4\\_\\_20.
- [142] Lucas Rizzo and Luca Longo. Inferential models of mental workload with defeasible argumentation and non-monotonic fuzzy reasoning: a comparative study. In *2nd Workshop on Advances In Argumentation In Artificial Intelligence*, pages 11–26, Trento, Italy, 2018. CEUR-WS.org.
- [143] Lucas Rizzo and Luca Longo. A qualitative investigation of the explainability of defeasible argumentation and non-monotonic fuzzy reasoning. In *Proceedings for the 26th AAI Irish Conference on Artificial Intelligence and Cognitive Science Trinity College Dublin*, pages 138–149, Dublin, Ireland, 2018. CEUR-WS.org.
- [144] Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *5th International Conference on Learning Representations, Workshop Track Proceedings*, page 68, Toulon, France, 2017. ICLR.
- [145] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International Conference on Machine Learning*, pages 2673–2682, Stockholm, Sweden, 2018. ICML.
- [146] David Barbella, Sami Benzaid, Janara M Christensen, Bret Jackson, X Victor Qin, and David R Musicant. Understanding support vector machine classifications via a recommender system-like approach. In *DMIN*, pages 305–311, Las Vegas, Nevada, USA, 2009. CSREA Press.
- [147] K Xu, J Ba, R Kiros, A Courville, R Salakhutdinov, R Zemel, and Y Bengio. Show, attend and tell: Neural image

- caption generation with visual attention. *Proceedings of the International Conference on Machine Learning*, 2048:2057–2088, 2015.
- [148] Sarah Tan, Rich Caruana, Giles Hooker, and Yin Lou. Distill-and-compare: Auditing black-box models using transparent model distillation. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 303–310, New Orleans, Louisiana, USA, 2018. ACM. doi: 10.1145/3278721.3278725.
  - [149] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, pages 4765–4774, Long Beach, California, USA, 2017. Neural Information Processing Systems Foundation, Inc.
  - [150] Marko Robnik-Šikonja and Igor Kononenko. Explaining classifications for individual instances. *IEEE Transactions on Knowledge and Data Engineering*, 20(5):589–600, 2008. doi: 10.1109/tkde.2007.190734.
  - [151] Marko Robnik-Šikonja. Explanation of prediction models with explain prediction. *Informatica*, 42(1):13–22, 2018.
  - [152] Paulo Cortez and Mark J Embrechts. Opening black box data mining models using sensitivity analysis. In *IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, pages 341–348, Paris, France, 2011. IEEE. doi: 10.1109/cidm.2011.5949423.
  - [153] Paulo Cortez and Mark J Embrechts. Using sensitivity analysis and visualization techniques to open black box data mining models. *Information Sciences*, 225:1–17, 2013. doi: 10.1016/j.ins.2012.10.039.
  - [154] Erik Štrumbelj and Igor Kononenko. An efficient explanation of individual classifications using game theory. *Journal of Machine Learning Research*, 11:1–18, March 2010. ISSN 1532-4435.
  - [155] Igor Kononenko, Erik Štrumbelj, Zoran Bosnić, Darko Pevec, Matjaž Kukar, and Marko Robnik-Šikonja. Explanation and reliability of individual predictions. *Informatica*, 37(1):41–48, 2013.
  - [156] Erik Štrumbelj, Igor Kononenko, and M Robnik Šikonja. Explaining instance classifications with interactions of subsets of feature values. *Data & Knowledge Engineering*, 68(10):886–904, 2009. doi: 10.1016/j.datak.2009.01.004.
  - [157] Erik Štrumbelj and Igor Kononenko. Towards a model independent method for explaining classification for individual instances. In *International Conference on Data Warehousing and Knowledge Discovery*, pages 273–282, Turin, Italy, 2008. Springer. doi: 10.1007/978-3-540-85836-2\\_26.
  - [158] Erik Štrumbelj, Zoran Bosnić, Igor Kononenko, Branko Zaktotnik, and Cvetka Grašić Kuhar. Explanation and reliability of prediction models: the case of breast cancer recurrence. *Knowledge and information systems*, 24(2):305–324, 2010. doi: 10.1007/s10115-009-0244-9.
  - [159] Anupam Datta, Shayak Sen, and Yair Zick. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *IEEE symposium on security and privacy (SP)*, pages 598–617, San Jose, California, USA, 2016. IEEE. doi: 10.1109/sp.2016.42.
  - [160] Philip Adler, Casey Falk, Sorelle A Friedler, Tionney Nix, Gabriel Rybeck, Carlos Scheidegger, Brandon Smith, and Suresh Venkatasubramanian. Auditing black-box models for indirect influence. *Knowledge and Information Systems*, 54(1):95–122, 2018. doi: 10.1007/s10115-017-1116-3.
  - [161] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1885–1894, Sydney, Australia, 2017. JMLR.org.
  - [162] Jakub Słowiński, Martin Strobel, and Yair Zick. A characterization of monotone influence measures for data classification. In *IJCAI-17 Workshop on Explainable AI (XAI)*, pages 48–52, Melbourne, Australia, 2017. International Joint Conferences on Artificial Intelligence, Inc.
  - [163] Andreas Henelius, Kai Puolamäki, Henrik Boström, Lars Asker, and Panagiotis Papapetrou. A peek into the black box: exploring classifiers by randomization. *Data mining and knowledge discovery*, 28(5-6):1503–1529, 2014. doi: 10.1007/s10618-014-0368-8.
  - [164] Erik Štrumbelj and Igor Kononenko. Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems*, 41(3):647–665, 2014. doi: 10.1007/s10115-013-0679-x.
  - [165] Hadrien Bride, Jie Dong, Jin Song Dong, and Zhé Hóu. Towards dependable and explainable machine learning using automated reasoning. In *International Conference on Formal Engineering Methods*, pages 412–416, Gold Coast, Australia, 2018. Springer. doi: 10.1007/978-3-030-02450-5\\_25.
  - [166] Ulf Johansson, Lars Niklasson, and Rikard König. Accuracy vs. comprehensibility in data mining models. In *Proceedings of the seventh international conference on information fusion*, volume 1, pages 295–300, Stockholm, Sweden, 2004. Elsevier.
  - [167] Ulf Johansson, Rikard König, and Lars Niklasson. The truth is in there-rule extraction from opaque models using genetic programming. In *FLAIRS Conference*, pages 658–663, Miami Beach, Florida, USA, 2004. AAAI Press.
  - [168] Osbert Bastani, Carolyn Kim, and Hamsa Bastani. Interpretability via model extraction. In *Fairness, Accountability, and Transparency in Machine Learning Workshop*, pages 57–61, Halifax, Nova Scotia, Canada, 2017. FAT/ML.
  - [169] Sanjay Krishnan and Eugene Wu. Palm: Machine learning explanations for iterative debugging. In *Proceedings*

- of the 2nd Workshop on Human-In-the-Loop Data Analytics*, page 4, Chicago, Illinois, USA, 2017. ACM. doi: 10.1145/3077257.3077271.
- [170] Ruth C Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3429–3437, Venice, Italy, 2017. IEEE. doi: 10.1109/iccv.2017.371.
  - [171] Lingqiao Liu and Lei Wang. What has my classifier learned? visualizing the classification rules of bag-of-feature model by support region detection. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3586–3593, Providence, Rhode Island, USA, 2012. IEEE. doi: 10.1109/cvpr.2012.6248103.
  - [172] Jaegul Choo, Hanseung Lee, Jaeyeon Kihm, and Haesun Park. ivisclassifier: An interactive visual analytics system for classification based on supervised dimension reduction. In *IEEE Symposium on visual analytics; Science and Technology*, pages 27–34, Salt Lake City, Utah, USA, 2010. IEEE. doi: 10.1109/vast.2010.5652443.
  - [173] Piotr Dabkowski and Yarin Gal. Real time image saliency for black box classifiers. In *Advances in Neural Information Processing Systems*, pages 6967–6976, Long Beach, California, USA, 2017. Neural Information Processing Systems Foundation, Inc.
  - [174] David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert MÃžller. How to explain individual classification decisions. *Journal of Machine Learning Research*, 11(Jun): 1803–1831, 2010.
  - [175] Alex Goldstein, Adam Kapelner, Justin Bleich, and Emil Pitkin. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics*, 24(1):44–65, 2015. doi: 10.1080/10618600.2014.907095.
  - [176] Giuseppe Casalicchio, Christoph Molnar, and Bernd Bischl. Visualizing the feature importance for black box models. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 655–670, Dublin, Ireland, 2018. Springer. doi: 10.1007/978-3-030-10925-7\_40.
  - [177] David Alvarez-Melis and Tommi S Jaakkola. A causal framework for explaining the predictions of black-box sequence-to-sequence models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 412–421, Copenhagen, Denmark, 2017. Association for Computational Linguistics. doi: 10.18653/v1/d17-1042.
  - [178] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *Proceedings of the 3rd International Conference on Learning Representations*, San Diego, California, USA, 2015. ICLR.
  - [179] Josua Krause, Adam Perer, and Enrico Bertini. Using visual analytics to interpret predictive machine learning models. In *ICML Workshop on Human Interpretability in Machine Learning*, pages 106–110, New York City, New York, USA, 2016. ICML.
  - [180] Brett Poulin, Roman Eisner, Duane Szafron, Paul Lu, Russell Greiner, David S Wishart, Alona Fyshe, Brandon Pearcy, Cam MacDonell, and John Anvik. Visual explanation of evidence with additive classifiers. In *Proceedings Of The National Conference On Artificial Intelligence*, volume 21, pages 1822–1829, Boston, Massachusetts, USA, 2006. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.
  - [181] Jiawei Zhang, Yang Wang, Piero Molino, Lezhi Li, and David S Ebert. Manifold: A model-agnostic framework for interpretation and diagnosis of machine learning models. *IEEE transactions on visualization and computer graphics*, 25(1):364–373, 2019.
  - [182] Minsuk Kahng, Dezhong Fang, and Duen Horng Polo Chau. Visual exploration of machine learning results using data cube analysis. In *Proceedings of the Workshop on Human-In-the-Loop Data Analytics*, page 1, San Francisco, California, USA, 2016. ACM. doi: 10.1145/2939502.2939503.
  - [183] Giles Hooker. Discovering additive structure in black box functions. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 575–580, Seattle, Washington, USA, 2004. ACM. doi: 10.1145/1014052.1014122.
  - [184] Or Biran and Kathleen McKeown. Justification narratives for individual classifications. In *Proceedings of the AutoML workshop at ICML*, volume 2014, pages 1–7, Beijing, China, 2014. ICML.
  - [185] Paolo Tamagnini, Josua Krause, Aritra Dasgupta, and Enrico Bertini. Interpreting black-box classifiers using instance-level visual explanations. In *Proceedings of the 2nd Workshop on Human-In-the-Loop Data Analytics*, pages 6:1–6:6, San Francisco, California, USA, 2017. ACM. doi: 10.1145/3077257.3077260.
  - [186] Scott Cheng-Hsin Yang and Patrick Shafto. Explainable artificial intelligence via bayesian teaching. In *NIPS 2017 workshop on Teaching Machines, Robots, and Humans*, pages 127–137, Long Beach, California, USA, 2017. NIPS.
  - [187] Rajiv Khanna, Been Kim, Joydeep Ghosh, and Sanmi Koyejo. Interpreting black box predictions using fisher kernels. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *The 22nd International Conference on Artificial Intelligence and Statistics*, volume 89, pages 3382–3390, Naha, Okinawa, Japan, 16–18 Apr 2019. Proceedings of Machine Learning Research.
  - [188] Jacob Bien, Robert Tibshirani, et al. Prototype selection for interpretable classification. *The Annals of Applied*

- Statistics*, 5(4):2403–2424, 2011. doi: 10.1214/11-aos495.
- [189] Rich Caruana, Hooshang Kangarloo, JD Dionisio, Usha Sinha, and David Johnson. Case-based explanation of non-case-based learning methods. In *Proceedings of the AMIA Symposium*, page 212, Washington, District of Columbia, USA, 1999. American Medical Informatics Association.
  - [190] Ninghao Liu, Hongxia Yang, and Xia Hu. Adversarial detection with model interpretation. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1803–1811, London, United Kingdom, 2018. ACM. doi: 10.1145/3219819.3220027.
  - [191] Been Kim, Rajiv Khanna, and Oluwasanmi O Koyejo. Examples are not enough, learn to criticize! criticism for interpretability. In *Advances in Neural Information Processing Systems*, pages 2280–2288, Long Beach, California, USA, 2016. Neural Information Processing Systems Foundation, Inc.
  - [192] Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Paishun Ting, Karthikeyan Shanmugam, and Payel Das. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. In *Advances in Neural Information Processing Systems 31 (NIPS)*, pages 592–603, Montréal, Canada, 2018. Curran Associates, Inc.
  - [193] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Model-agnostic interpretability of machine learning. In *Proceedings of the 22nd International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, San Francisco, California, USA, 2016. ACM.
  - [194] Devinder Kumar, Alexander Wong, and Graham W Taylor. Explaining the unexplained: A class-enhanced attentive response (clear) approach to understanding deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 36–44, Honolulu, Hawaii, USA, 2017. IEEE. doi: 10.1109/cvprw.2017.215.
  - [195] Ge Liu and David Gifford. Visualizing feature maps in deep neural networks using deepresolve. a genomics case study. In *International Conference on Machine Learning 2017 - Workshop on Visualization for Deep Learning*, pages 32–41, Sydney, Australia, 2017. ICML.
  - [196] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626, Venice, Italy, 2017. IEEE. doi: 10.1109/iccv.2017.74.
  - [197] Nazneen Fatema Rajani and Raymond J Mooney. Using explanations to improve ensembling of visual question answering systems. *Training*, 82:248–349, 2017.
  - [198] Yash Goyal, Akrit Mohapatra, Devi Parikh, and Dhruv Batra. Towards transparent ai systems: Interpreting visual question answering models. In *ICML Workshop on Visualization for Deep Learning*, New York City, New York, USA, 2016. ICML.
  - [199] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833, Zurich, Switzerland, 2014. Springer.
  - [200] Ruth Fong and Andrea Vedaldi. Net2vec: Quantifying and explaining how concepts are encoded by filters in deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8730–8738, Salt Lake City, Utah, USA, 2018. IEEE. doi: 10.1109/CVPR.2018.00910.
  - [201] Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5188–5196, Boston, Massachusetts, USA, 2015. IEEE. doi: 10.1109/cvpr.2015.7299155.
  - [202] Mengnan Du, Ninghao Liu, Qingquan Song, and Xia Hu. Towards explanation of dnn-based prediction with guided feature inversion. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1358–1367, London, United Kingdom, 2018. ACM. doi: 10.1145/3219819.3220099.
  - [203] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. In *International Conference on Machine Learning 2017 - Workshop on Visualization for Deep Learning*, pages 15–24, Sydney, Australia, 2017. ICML.
  - [204] Quanshi Zhang, Ying Nian Wu, and Song-Chun Zhu. Interpretable convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8827–8836, Salt Lake City, Utah, USA, 2018. IEEE. doi: 10.1109/cvpr.2018.00920.
  - [205] Luisa M Zintgraf, Taco S Cohen, Tameem Adel, and Max Welling. Visualizing deep neural network decisions: Prediction difference analysis. In *5th International Conference on Learning Representations*, Toulon, France, 2017. ICLR.
  - [206] Leila Arras, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. Explaining recurrent neural network predictions in sentiment analysis. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 159–168, Copenhagen, Denmark, 2017. Association for Computational Linguistics. doi: 10.18653/v1/w17-5221.
  - [207] Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, Klaus-Robert Müller, and Wojciech Samek. Layer-

- wise relevance propagation for neural networks with local renormalization layers. In *International Conference on Artificial Neural Networks*, pages 63–71, Barcelona, Spain, 2016. Springer. doi: 10.1007/978-3-319-44781-0\_8.
- [208] Grégoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recognition*, 65(May):211–222, 2017.
- [209] Sen He and Nicolas Pugeault. Deep saliency: What is learnt by a deep network about saliency? In *International Conference on Machine Learning - Workshop on Visualization for Deep Learning*, pages 1–5, Sydney, Australia, 2017. ICML.
- [210] Pieter-Jan Kindermans, Kristof T. Schütt, Maximilian Alber, Klaus-Robert Müller, Dumitru Erhan, Been Kim, and Sven Dähne. Learning how to explain neural networks: Patternnet and patternattribution. In *6th International Conference on Learning Representations*, Vancouver, Canada, 2018. ICLR.
- [211] Oramas Mogrovejo, José Antonio, Kaili Wang, and Tinne Tuytelaars. Visual explanation by interpretation: Improving visual feedback capabilities of deep neural networks. In *7th International Conference on Learning Representations*, New Orleans, Louisiana, USA, 2019. ICLR.
- [212] Eoin M Kenny and Mark T Keane. Twin-systems to explain artificial neural networks using case-based reasoning: comparative tests of feature-weighting methods in ann-cbr twins for xai. In *Twenty-Eighth International Joint Conferences on Artifical Intelligence (IJCAI)*, pages 2708–2715, Macao, China, 2019. International Joint Conferences on Artificial Intelligence Organization. doi: 10.24963/ijcai.2019/376.
- [213] Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. Visualizing and understanding neural models in nlp. In *The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 681–691, San Diego California, USA, 2016. The Association for Computational Linguistics. doi: 10.18653/v1/n16-1082.
- [214] Lingyang Chu, Xia Hu, Juhua Hu, Lanjun Wang, and Jian Pei. Exact and consistent interpretation for piecewise linear neural networks: A closed form solution. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1244–1253, London, United Kingdom, 2018. ACM. doi: 10.1145/3219819.3220063.
- [215] Xuan Liu, Xiaoguang Wang, and Stan Matwin. Interpretable deep convolutional neural networks via meta-learning. In *International Joint Conference on Neural Networks (IJCNN)*, pages 1–9, Rio de Janeiro, Brazil, 2018. IEEE. doi: 10.1109/ijcnn.2018.8489172.
- [216] Mathieu Aubry and Bryan C Russell. Understanding deep features with computer-generated imagery. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2875–2883, Santiago, Chile, 2015. IEEE. doi: 10.1109/iccv.2015.329.
- [217] Tom Zahavy, Nir Ben-Zrihem, and Shie Mannor. Graying the black box: Understanding dqns. In *International Conference on Machine Learning*, pages 1899–1908, New York City, New York, USA, 2016. PMLR.
- [218] Paulo E Rauber, Samuel G Fadel, Alexandre X Falcao, and Alexandru C Telea. Visualizing the hidden activity of artificial neural networks. *IEEE transactions on visualization and computer graphics*, 23(1):101–110, 2017. doi: 10.1109/tvcg.2016.2598838.
- [219] Jayaraman J Thiagarajan, Bhavya Kailkhura, Prasanna Sattigeri, and Karthikeyan Natesan Ramamurthy. Tree-view: Peeking into deep neural networks via feature-space partitioning. In *NIPS Interpretability Workshop*, Barcelona, Spain, 2016. NIPS.
- [220] David Bau, Jun-Yan Zhu, Hendrik Strobelt, Zhou Bolei, Joshua B. Tenenbaum, William T. Freeman, and Antonio Torralba. Gan dissection: Visualizing and understanding generative adversarial networks. In *Proceedings of the International Conference on Learning Representation*, New Orleans, Louisiana, USA, 2019. ICLR.
- [221] Benjamin J Lengerich, Sandeep Konam, Eric P Xing, Stephanie Rosenthal, and Manuela Veloso. Towards visual explanations for convolutional neural networks via input resampling. In *International Conference on Machine Learning 2017 - Workshop on Visualization for Deep Learning*, pages 25–31, Sydney, Australia, 2017. ICML.
- [222] Dumitru Erhan, Aaron Courville, and Yoshua Bengio. Understanding representations learned in deep architectures. *Department d'Informatique et Recherche Opérationnelle, University of Montreal, QC, Canada, Tech. Rep*, 1355:1, 2010.
- [223] Anh Nguyen, Jason Yosinski, and Jeff Clune. Multifaceted feature visualization: Uncovering the different types of features learned by each neuron in deep neural networks. In *Visualization for Deep Learning workshop. International Conference on Machine Learning*, New York City, New York, USA, 2016. ICML.
- [224] Anh Nguyen, Alexey Dosovitskiy, Jason Yosinski, Thomas Brox, and Jeff Clune. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. In *Advances in Neural Information Processing Systems*, pages 3387–3395, Barcelona, Spain, 2016. Neural Information Processing Systems Foundation, Inc.
- [225] Mandana Hamidi-Haines, Zhongang Qi, Alan Fern, Fuxin Li, and Prasad Tadepalli. Interactive naming for explaining deep neural networks: A formative study. In *Joint Proceedings of the ACM IUI 2019 Workshops co-located with the 24th ACM Conference on Intelligent User Interfaces (ACM IUI)*, volume 2327, Los Angeles, California, USA, 2019. CEUR-WS.org.

- [226] Andrej Karpathy, Justin Johnson, and Li Fei-Fei. Visualizing and understanding recurrent networks. In *ICLR Workshops*, San Juan, Puerto Rico, 2016. ICLR.
- [227] Quanshi Zhang, Ruiming Cao, Feng Shi, Ying Nian Wu, and Song-Chun Zhu. Interpreting cnn knowledge via an explanatory graph. In *Thirty-Second AAAI Conference on Artificial Intelligence*, pages 2124–2132, New Orleans, Louisiana, USA, 2018. AAAI Press.
- [228] Xiaodan Liang, Zhiting Hu, Hao Zhang, Liang Lin, and Eric P Xing. Symbolic graph reasoning meets convolutions. In *Advances in Neural Information Processing Systems*, pages 1853–1863, Montréal, Canada, 2018. Neural Information Processing Systems Foundation, Inc.
- [229] Quanshi Zhang, Ruiming Cao, Ying Nian Wu, and Song-Chun Zhu. Growing interpretable part graphs on convnets via multi-shot learning. In *Thirty-First AAAI Conference on Artificial Intelligence*, pages 2898–2906, San Francisco, California, USA, 2017. AAAI Press.
- [230] Chris Olah, Arvind Satyanarayanan, Ian Johnson, Shan Carter, Ludwig Schubert, Katherine Ye, and Alexander Mordvintsev. The building blocks of interpretability. *Distill*, 3(3):e10, 2018. doi: 10.23915/distill.00010.
- [231] Minsuk Kahng, Pierre Y Andrews, Aditya Kalro, and Duen Horng Polo Chau. A cti v is: Visual exploration of industry-scale deep neural network models. *IEEE transactions on visualization and computer graphics*, 24(1): 88–97, 2018.
- [232] Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. Understanding neural networks through deep visualization. In *In ICML Workshop on Deep Learning, poster presentation*, Lille, France, 2015. ICML.
- [233] Wen Zhong, Cong Xie, Yuan Zhong, Yang Wang, Wei Xu, Shenghui Cheng, and Klaus Mueller. Evolutionary visual analysis of deep neural networks. In *International Conference on Machine Learning 2017 - Workshop on Visualization for Deep Learning*, pages 6–14, Sydney, Australia, 2017. ICML.
- [234] Maximilian Alber, Sebastian Lapuschkin, Philipp Seegerer, Miriam Hägele, Kristof T. Schütt, Grégoire Montavon, Wojciech Samek, Klaus-Robert Müller, Sven Dähne, and Pieter-Jan Kindermans. innvestigate neural networks. *Journal of Machine Learning Research*, 20:1–8, 2019.
- [235] Hendrik Strobelt, Sebastian Gehrmann, Michael Behrisch, Adam Perer, Hanspeter Pfister, and Alexander M Rush. Seq2seq-vis: A visual debugging tool for sequence-to-sequence models. *IEEE transactions on visualization and computer graphics*, 25(1):353–363, 2018. doi: 10.1109/TVCG.2018.2865044.
- [236] Matthew J Streeter, Matthew O Ward, and Sergio A Alvarez. Nvis: An interactive visualization tool for neural networks. In *Visual Data Exploration and Analysis VIII*, volume 4302, pages 234–242, San Jose, California, USA, 2001. International Society for Optics and Photonics.
- [237] Rudy Setiono and Huan Liu. Understanding neural networks via rule extraction. In *IJCAI*, volume 1, pages 480–485. Montréal, Québec, 1995. International Joint Conferences on Artificial Intelligence, Inc. doi: 10.1142/S0129065797000380.
- [238] Andrey Bondarenko, Ludmila Aleksejeva, Vilen Jumutc, and Arkady Borisov. Classification tree extraction from trained artificial neural networks. *Procedia Computer Science*, 104:556–563, 2017. doi: 10.1016/j.procs.2017.01.172.
- [239] Sebastian Thrun. Extracting rules from artificial neural networks with distributed representations. In *Advances in neural information processing systems*, pages 505–512, Denver, Colorado, USA, 1995. MIT Press.
- [240] Guido Bologna. Symbolic rule extraction from the dimlp neural network. In *International Workshop on Hybrid Neural Systems*, pages 240–254, Denver, Colorado, USA, 1998. Springer. doi: 10.1007/10719871\\_\\_17.
- [241] Guido Bologna. A rule extraction study based on a convolutional neural network. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, pages 304–313, Hamburg, Germany, 2018. Springer. doi: 10.1007/978-3-319-99740-7\\_\\_22.
- [242] M Gethsiyal Augasta and Thangairulappan Kathirvalavakumar. Reverse engineering the neural networks for rule extraction in classification problems. *Neural processing letters*, 35(2):131–150, 2012. doi: 10.1007/s11063-011-9207-8.
- [243] Saroj Kumar Biswas, Manomita Chakraborty, Biswajit Purkayastha, Pinki Roy, and Dalton Meitei Thounaojam. Rule extraction from training data using neural network. *International Journal on Artificial Intelligence Tools*, 26 (03):1750006, 2017. doi: 10.1142/S0218213017500063.
- [244] Arthur d’Avila Garcez, Krysia Broda, and Dov M Gabbay. Symbolic knowledge extraction from trained neural networks: A sound approach. *Artificial Intelligence*, 125(1-2):155–207, 2001. doi: 10.1016/s0004-3702(00)00077-1.
- [245] Nicholas Frosst and Geoffrey Hinton. Distilling a neural network into a soft decision tree. In *16th International Conference of the Italian Association of Artificial Intelligence, 2017. Workshop on Comprehensibility and Explanation in AI and ML*, pages 1–8, Bari, Italy, 2017. Cex, University of Bremen, Germany.
- [246] Quanshi Zhang, Yu Yang, Haotian Ma, and Ying Nian Wu. Interpreting cnns via decision trees. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6261–6270, Long Beach, California, USA, 2019. IEEE.
- [247] Zhi-Hua Zhou, Yuan Jiang, and Shi-Fu Chen. Extracting symbolic rules from trained neural network ensembles.

- Ai Communications*, 16(1):3–15, 2003.
- [248] Zhi-Hua Zhou and Yuan Jiang. Medical diagnosis with c4. 5 rule preceded by artificial neural network ensemble. *IEEE Transactions on information Technology in Biomedicine*, 7(1):37–42, 2003. doi: 10.1109/tib.2003.808498.
- [249] Olcay Boz. Extracting decision trees from trained neural networks. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 456–461, Edmonton, Alberta, Canada, 2002. ACM. doi: 10.1145/775047.775113.
- [250] Mark W Craven and Jude W Shavlik. Using sampling and queries to extract rules from trained neural networks. In *Machine Learning Proceedings*, pages 37–45. Elsevier, New Brunswick, New Jersey, USA, 1994. doi: 10.1016/b978-1-55860-335-6.50013-1.
- [251] Mark Craven and Jude W Shavlik. Extracting tree-structured representations of trained networks. In *Advances in neural information processing systems*, pages 24–30, Denver, Colorado, USA, 1996. MIT Press.
- [252] Mike Wu, Michael C Hughes, Sonali Parbhoo, Maurizio Zazzi, Volker Roth, and Finale Doshi-Velez. Beyond sparsity: Tree regularization of deep models for interpretability. In *Thirty-Second AAAI Conference on Artificial Intelligence*, pages 1670–1678, New Orleans, Louisiana, USA, 2018. AAAI Press.
- [253] W James Murdoch and Arthur Szlam. Automatic rule extraction from long short term memory networks. In *5th International Conference on Learning Representations, Conference Track Proceedings*, Toulon, France, 2017. ICLR.
- [254] Zhiting Hu, Xuezhe Ma, Zhengzhong Liu, Eduard Hovy, and Eric Xing. Harnessing deep neural networks with logic rules. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL, Volume 1: Long Papers*, pages 2410–2420, Berlin, Germany, 2016. Association for Computational Linguistics. doi: 10.18653/v1/p16-1228.
- [255] Son N Tran. Unsupervised neural-symbolic integration. In *IJCAI-17 Workshop on Explainable AI (XAI)*, pages 58–62, Melbourne, Australia, 2017. International Joint Conferences on Artificial Intelligence, Inc.
- [256] Shane Barratt. Interpet: Neural introspection for interpretable deep learning. In *NIPS Symposium on Interpretable Machine Learning*, pages 47–53, Long Beach, California, USA, 2017. NIPS.
- [257] Iván García-Magariño, Rajarajan Muttukrishnan, and Jaime Lloret. Human-centric ai for trustworthy iot systems with explainable multilayer perceptrons. *IEEE Access*, 7:125562–125574, 2019. doi: 10.1109/ACCESS.2019.2937521.
- [258] Tao Lei, Regina Barzilay, and Tommi Jaakkola. Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 107–117, Austin, Texas, USA, 2016. Association for Computational Linguistics.
- [259] Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, and Trevor Darrell. Generating visual explanations. In *European Conference on Computer Vision*, pages 3–19, Amsterdam, The Netherlands, 2016. Springer. doi: 10.1007/978-3-319-46493-0\\_\\_1.
- [260] Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems*, pages 6076–6085, Long Beach, California, USA, 2017. Neural Information Processing Systems Foundation, Inc.
- [261] Raphael Féraud and Fabrice Clérot. A methodology to explain neural network classification. *Neural Networks*, 15(2):237–246, 2002. doi: 10.1016/s0893-6080(01)00127-7.
- [262] Kary Främling. Explaining results of neural networks by contextual importance and utility. In *Proceedings of Rule Extraction From Trained Artificial Neural Networks Workshop*, pages 43–56, Brighton, England, 1996. Citeseer.
- [263] Jérémie Clos, Nirmalie Wiratunga, and Stewart Massie. Towards explainable text classification by jointly learning lexicon and modifier terms. In *IJCAI-17 Workshop on Explainable AI (XAI)*, pages 19–23, Melbourne, Australia, 2017. International Joint Conferences on Artificial Intelligence, Inc.
- [264] Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John Canny, and Zeynep Akata. Textual explanations for self-driving vehicles. In *Proceedings of the European conference on computer vision (ECCV)*, pages 563–578, Munich, Germany, 2018. ECCV.
- [265] Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. Multimodal explanations: Justifying decisions and pointing to the evidence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8779–8788, Salt Lake City, Utah, USA, 2018. IEEE.
- [266] Franz Mayr and Sergio Yovine. Regular inference on artificial neural networks. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, pages 350–369, Hamburg, Germany, 2018. Springer. doi: 10.1007/978-3-319-99740-7\\_\\_25.
- [267] Christian W Omlin and C Lee Giles. Extraction of rules from discrete-time recurrent neural networks. *Neural networks*, 9(1):41–52, 1996. doi: 10.1016/0893-6080(95)00086-0.
- [268] Martin Tamajka, Wanda Benesova, and Matej Kompanek. Transforming convolutional neural network to an interpretable classifier. In *International Conference on Systems, Signals and Image Processing (IWSSIP)*, pages 255–259, Osijek, Croatia, 2019. IEEE.

- [269] Fernando EB Otero and Alex A Freitas. Improving the interpretability of classification rules discovered by an ant colony algorithm: Extended results. *Evolutionary computation*, 24(3):385–409, 2016. doi: 10.1162/evco\\_a\\_00155.
- [270] Wouter Verbeke, David Martens, Christophe Mues, and Bart Baesens. Building comprehensible customer churn prediction models with advanced rule induction techniques. *Expert systems with applications*, 38(3):2354–2364, 2011. doi: 10.1016/j.eswa.2010.08.023.
- [271] Brian R. Gaines. Transforming rules and trees into comprehensible knowledge structures. In Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, pages 205–226. American Association for Artificial Intelligence, Menlo Park, CA, USA, 1996. ISBN 0-262-56097-6.
- [272] Himabindu Lakkaraju, Stephen H Bach, and Jure 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*, pages 1675–1684, San Francisco, California, USA, 2016. ACM. doi: 10.1145/2939672.2939874.
- [273] Benjamin Letham, Cynthia Rudin, Tyler H McCormick, and David Madigan. Building interpretable classifiers with rules using bayesian analysis. *Department of Statistics Technical Report tr609, University of Washington*, 9(3):1350–1371, 2012.
- [274] Benjamin Letham, Cynthia Rudin, Tyler H. McCormick, and David Madigan. An interpretable stroke prediction model using rules and bayesian analysis. In *Proceedings of the 17th AAAI Conference on Late-Breaking Developments in the Field of Artificial Intelligence, AAAIWS’13-17*, pages 65–67, Palo Alto, California, USA, 2013. AAAI Press.
- [275] Benjamin Letham, Cynthia Rudin, Tyler H McCormick, David Madigan, et al. Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. *The Annals of Applied Statistics*, 9(3):1350–1371, 2015. doi: 10.1214/15-aos848.
- [276] Tong Wang, Cynthia Rudin, Finale Velez-Doshi, Yimin Liu, Erica Klampfl, and Perry MacNeille. Bayesian rule sets for interpretable classification. In *IEEE 16th International Conference on Data Mining (ICDM)*, pages 1269–1274, Barcelona, Spain, 2016. IEEE. doi: 10.1109/icdm.2016.0171.
- [277] Tong Wang, Cynthia Rudin, Finale Doshi-Velez, Yimin Liu, Erica Klampfl, and Perry MacNeille. A bayesian framework for learning rule sets for interpretable classification. *The Journal of Machine Learning Research*, 18(1):2357–2393, 2017.
- [278] M Pazzani. Comprehensible knowledge discovery: gaining insight from data. In *First Federal Data Mining Conference and Exposition*, pages 73–82, London, United Kingdom, 1997. ACM.
- [279] Zhiwei Zeng, Chunyan Miao, Cyril Leung, and Jing Jih Chin. Building more explainable artificial intelligence with argumentation. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th Symposium on Educational Advances in Artificial Intelligence (EAAI-18)*, pages 8044–8046, New Orleans, Louisiana, USA, 2018. AAAI Press.
- [280] Hisao Ishibuchi and Yusuke Nojima. Analysis of interpretability-accuracy tradeoff of fuzzy systems by multi-objective fuzzy genetics-based machine learning. *International Journal of Approximate Reasoning*, 44(1):4–31, 2007. doi: 10.1016/j.ijar.2006.01.004.
- [281] Yaochu Jin. Fuzzy modeling of high-dimensional systems: complexity reduction and interpretability improvement. *IEEE Transactions on Fuzzy Systems*, 8(2):212–221, 2000. doi: 10.1109/91.842154.
- [282] Régis Pierrard, Jean-Philippe Poli, and Céline Hudelot. Learning fuzzy relations and properties for explainable artificial intelligence. In *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–8, Rio de Janeiro, Brazil, 2018. IEEE. doi: 10.1109/FUZZ-IEEE.2018.8491538.
- [283] Zhenyu Wang and Vasile Palade. Building interpretable fuzzy models for high dimensional data analysis in cancer diagnosis. *BMC genomics*, 12(2):S5:1–S5:11, 2011. doi: 10.1186/1471-2164-12-s2-s5.
- [284] Alberto Cano, Amelia Zafra, and Sebastián Ventura. An interpretable classification rule mining algorithm. *Information Sciences*, 240:1–20, 2013. doi: 10.1016/j.ins.2013.03.038.
- [285] Dmitry M Malioutov, Kush R Varshney, Amin Emad, and Sanjeeb Dash. Learning interpretable classification rules with boolean compressed sensing. In *Transparent Data Mining for Big and Small Data*, pages 95–121. Springer, Cham, Switzerland, 2017. doi: 10.1007/978-3-319-54024-5\\_5.
- [286] Guolong Su, Dennis Wei, Kush R Varshney, and Dmitry M Malioutov. Interpretable two-level boolean rule learning for classification. In *Proceedings of ICML Workshop Human Interpretability in Machine Learning*, pages 66–70, New York City, New York, USA, 2016. ICML.
- [287] Blen M Keneni, Devinder Kaur, Ali Al Bataineh, Vijaya K Devabhaktuni, Ahmad Y Javaid, Jack D Zaientz, and Robert P Marinier. Evolving rule-based explainable artificial intelligence for unmanned aerial vehicles. *IEEE Access*, 7:17001–17016, 2019. doi: 10.1109/ACCESS.2019.2893141.
- [288] David P Pancho, Jose M Alonso, Oscar Cordón, Arnaud Quirin, and Luis Magdalena. Fingrams: visual repre-

- sentations of fuzzy rule-based inference for expert analysis of comprehensibility. *IEEE Transactions on Fuzzy Systems*, 21(6):1133–1149, 2013. doi: 10.1109/tfuzz.2013.2245130.
- [289] Jose M Alonso. Explainable artificial intelligence for kids. In *Conference of the International Fuzzy Systems Association and the European Society for Fuzzy Logic and Technology (EUSFLAT)*, Prague, Czech Republic, 2019. Atlantis Press. doi: 10.2991/eusflat-19.2019.21.
- [290] Gabriele Tolomei, Fabrizio Silvestri, Andrew Haines, and Mounia Lalmas. Interpretable predictions of tree-based ensembles via actionable feature tweaking. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 465–474, Halifax, Nova Scotia, Canada, 2017. ACM. doi: 10.1145/3097983.3098039.
- [291] José M Alonso, Alejandro Ramos-Soto, Ciro Castiello, and Corrado Mencar. Explainable ai beer style classifier. In *Proceedings of the SICSA Workshop on Reasoning, Learning and Explainability*, pages 1–5, Aberdeen, Scotland, UK, 2018. CEUR-WS.
- [292] César Ferri, José Hernández-Orallo, and M José Ramírez-Quintana. From ensemble methods to comprehensible models. In *International Conference on Discovery Science*, pages 165–177, Lübeck, Germany, 2002. Springer. doi: 10.1007/3-540-36182-0\\_16.
- [293] Anneleen Van Assche and Hendrik Blockeel. Seeing the forest through the trees: Learning a comprehensible model from an ensemble. In *European Conference on Machine Learning*, pages 418–429, Warsaw, Poland, 2007. Springer. doi: 10.1007/978-3-540-74958-5\\_\\_39.
- [294] Artur Andrzejak, Felix Langner, and Silvestre Zabala. Interpretable models from distributed data via merging of decision trees. In *IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, pages 1–9, Singapore, 2013. IEEE. doi: 10.1109/cidm.2013.6597210.
- [295] Satoshi Hara and Kohei Hayashi. Making tree ensembles interpretable: A bayesian model selection approach. In *International Conference on Artificial Intelligence and Statistics, AISTATS*, pages 77–85, Playa Blanca, Lanzarote, Canary Islands, Spain, 2018. PMLR.
- [296] Houtao Deng. Interpreting tree ensembles with intrees. *International Journal of Data Science and Analytics*, 7 (4):1–11, 2018. doi: 10.1007/s41060-018-0144-8.
- [297] Jun Gao, Ninghao Liu, Mark Lawley, and Xia Hu. An interpretable classification framework for information extraction from online healthcare forums. *Journal of healthcare engineering*, 2017:798–809, 2017. doi: 10.1155/2017/2460174.
- [298] Hui Fen Tan, Giles Hooker, and Martin T Wells. Tree space prototypes: Another look at making tree ensembles interpretable. In *NIPS Interpretability Workshop*, Barcelona, Spain, 2016. NIPS.
- [299] Haydemar Núñez, Cecilio Angulo, and Andreu Català. Rule extraction from support vector machines. In *European Symposium on Artificial Neural Networks, ESANN*, pages 107–112, Bruges, Belgium, 2002. i6doc.
- [300] Doina Caragea, Dianne Cook, and Vasant Honavar. Towards simple, easy-to-understand, yet accurate classifiers. In *Third IEEE International Conference on Data Mining*, pages 497–500, San Francisco, California, USA, 2003. IEEE. doi: 10.1109/icdm.2003.1250961.
- [301] Lutz Hamel. Visualization of support vector machines with unsupervised learning. In *IEEE Symposium on Computational Intelligence and Bioinformatics and Computational Biology*, pages 1–8, Toronto, Ontario, Canada, 2006. IEEE. doi: 10.1109/cibcb.2006.330984.
- [302] Aleks Jakulin, Martin Možina, Janez Demšar, Ivan Bratko, and Blaž Zupan. Nomograms for visualizing support vector machines. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 108–117, Chicago, Illinois, USA, 2005. ACM. doi: 10.1145/1081870.1081886.
- [303] Baek Hwan Cho, Hwanjo Yu, Jongshill Lee, Young Joon Chee, In Young Kim, and Sun I Kim. Nonlinear support vector machine visualization for risk factor analysis using nomograms and localized radial basis function kernels. *IEEE Transactions on Information Technology in Biomedicine*, 12(2):247–256, 2008. doi: 10.1109/titb.2007.902300.
- [304] Martin Možina, Janez Demšar, Michael Kattan, and Blaž Zupan. Nomograms for visualization of naive bayesian classifier. In *European Conference on Principles of Data Mining and Knowledge Discovery*, pages 337–348, Pisa, Italy, 2004. Springer. doi: 10.1007/978-3-540-30116-5\\_\\_32.
- [305] Ghim-Eng Yap, Ah-Hwee Tan, and Hwee-Hwa Pang. Explaining inferences in bayesian networks. *Applied Intelligence*, 29(3):263–278, 2008. doi: 10.1007/s10489-007-0093-8.
- [306] Charlotte S Vlek, Henry Prakken, Silja Renooij, and Bart Verheij. A method for explaining bayesian networks for legal evidence with scenarios. *Artificial Intelligence and Law*, 24(3):285–324, 2016. doi: 10.1007/s10506-016-9183-4.
- [307] Sjoerd T Timmer, John-Jules Ch Meyer, Henry Prakken, Silja Renooij, and Bart Verheij. A two-phase method for extracting explanatory arguments from bayesian networks. *International Journal of Approximate Reasoning*, 80: 475–494, 2017. doi: 10.1016/j.ijar.2016.09.002.
- [308] Will Landecker, Michael D. Thomure, Luis M. A. Bettencourt, Melanie Mitchell, Garrett T Kenyon, and Steven P. Brumby. Interpreting individual classifications of hierarchical networks. In *IEEE Symposium on Computational*

- Intelligence and Data Mining (CIDM)*, volume 165, pages 32–38, Singapore, apr 2013. IEEE. ISBN 978-1-4673-5895-8. doi: 10.1109/CIDM.2013.6597214.
- [309] Been Kim, Cynthia Rudin, and Julie A Shah. The bayesian case model: A generative approach for case-based reasoning and prototype classification. In *Advances in Neural Information Processing Systems*, pages 1952–1960, Montréal, Québec, Canada, 2014. Neural Information Processing Systems Foundation, Inc.
  - [310] Matthew S Caywood, Daniel M Roberts, Jeffrey B Colombe, Hal S Greenwald, and Monica Z Weiland. Gaussian process regression for predictive but interpretable machine learning models: An example of predicting mental workload across tasks. *Frontiers in human neuroscience*, 10:647–665, 2017. doi: 10.3389/fnhum.2016.00647.
  - [311] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1721–1730, Sydney, Australia, 2015. ACM. doi: 10.1145/2783258.2788613.
  - [312] Jialei Wang, Ryohei Fujimaki, and Yosuke Motohashi. Trading interpretability for accuracy: Oblique treed sparse additive models. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1245–1254, Sydney, Australia, 2015. ACM. doi: 10.1145/2783258.2783407.
  - [313] Gerald Fahner. Developing transparent credit risk scorecards more effectively: An explainable artificial intelligence approach. In *The Seventh International Conference on Data Analytics*, pages 17–24, Athens, Greece, 2018. Data Analytics.
  - [314] Daniel Howard and Mark A Edwards. Explainable ai: The promise of genetic programming multi-run subtree encapsulation. In *International Conference on Machine Learning and Data Engineering (iCMLDE)*, pages 158–159, Dallas, Texas, USA, 2018. IEEE. doi: 10.1109/iCMLDE.2018.00037.
  - [315] Yitao Liang and Guy Van den Broeck. Towards compact interpretable models: Shrinking of learned probabilistic sentential decision diagrams. In *IJCAI-17 Workshop on Explainable AI (XAI)*, pages 31–35, Melbourne, Australia, 2017. International Joint Conferences on Artificial Intelligence, Inc.
  - [316] Been Kim, Julie A Shah, and Finale Doshi-Velez. Mind the gap: A generative approach to interpretable feature selection and extraction. In *Advances in Neural Information Processing Systems*, pages 2260–2268, Montréal, Québec, Canada, 2015. Neural Information Processing Systems Foundation, Inc.
  - [317] Berk Ustun, Stefano Traca, and Cynthia Rudin. Supersparse linear integer models for interpretable classification. *Stat*, 1050:11–47, 2014.
  - [318] Alexander Panchenko, Eugen Ruppert, Stefano Faralli, Simone Paolo Ponzetto, and Chris Biemann. Unsupervised does not mean uninterpretable: The case for word sense induction and disambiguation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, volume 1, pages 86–98, Valencia, Spain, 2017. Association for Computational Linguistics.
  - [319] Maria Jose Gacto, Rafael Alcalá, and Francisco Herrera. Interpretability of linguistic fuzzy rule-based systems: An overview of interpretability measures. *Information Sciences*, 181(20):4340–4360, 2011. doi: 10.1016/j.ins.2011.02.021.
  - [320] Salvador García, Alberto Fernández, Julián Luengo, and Francisco Herrera. A study of statistical techniques and performance measures for genetics-based machine learning: accuracy and interpretability. *Soft Computing*, 13(10):959, 2009.
  - [321] Marko Bohanec, Marko Robnik-Šikonja, and Mirjana Kljajić Borštnar. Decision-making framework with double-loop learning through interpretable black-box machine learning models. *Industrial Management & Data Systems*, 117(7):1389–1406, 2017. doi: 10.1108/imds-09-2016-0409.
  - [322] Marko Bohanec, Mirjana Kljajić Borštnar, and Marko Robnik-Šikonja. Explaining machine learning models in sales predictions. *Expert Systems with Applications*, 71:416–428, 2017. doi: 10.1016/j.eswa.2016.11.010.
  - [323] Thibault Laugel, Marie-Jeanne Lesot, Christophe Marsala, Xavier Renard, and Marcin Detyniecki. The dangers of post-hoc interpretability: Unjustified counterfactual explanations. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2801–2807, Macao, China, 2019. International Joint Conferences on Artificial Intelligence Organization. doi: 10.24963/ijcai.2019/388.
  - [324] Julius Adebayo, Justin Gilmer, Ian Goodfellow, and Been Kim. Local explanation methods for deep neural networks lack sensitivity to parameter values. In *6th International Conference on Learning Representations*, Vancouver, Canada, 2018. ICLR.
  - [325] Julius Adebayo, Justin Gilmer, Michael Muellly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS*, pages 9505–9515, Montréal, Canada, 2018. Neural Information Processing Systems Foundation, Inc.
  - [326] Marco Ancona, Enea Ceolini, Cengiz Oztireli, and Markus Gross. Towards better understanding of gradient-based attribution methods for deep neural networks. In *6th International Conference on Learning Representations*, Vancouver, Canada, 2018. ICLR.
  - [327] Leila Arras, Franziska Horn, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. Explaining predic-

- tions of non-linear classifiers in nlp. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 1–7, Berlin, Germany, 2016. Association for Computational Linguistics. doi: 10.18653/v1/w16-1601.
- [328] Leila Arras, Franziska Horn, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. “what is relevant in a text document?”: An interpretable machine learning approach. *PloS one*, 12(8):e0181142, 2017. doi: 10.1371/journal.pone.0181142.
- [329] Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. Evaluating the visualization of what a deep neural network has learned. *IEEE transactions on neural networks and learning systems*, 28(11):2660–2673, 2017. doi: 10.1109/tnnls.2016.2599820.
- [330] Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *ITU Journal: ICT Discoveries*, 1:1–10, 2017.
- [331] Alexander Binder, Wojciech Samek, Grégoire Montavon, Sebastian Bach, and Klaus-Robert Müller. Analyzing and validating neural networks predictions. In *Proceedings of the ICML Workshop on Visualization for Deep Learning*, pages 118–121, New York City, New York, USA, 2016. ICML.
- [332] Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of neural networks is fragile. In *NIPS workshop on Machine Deception*, Long Beach, California, USA, 2017. NIPS.
- [333] Muriel Gevrey, Ioannis Dimopoulos, and Sovan Lek. Review and comparison of methods to study the contribution of variables in artificial neural network models. *Ecological modelling*, 160(3):249–264, 2003. doi: 10.1016/s0304-3800(02)00257-0.
- [334] Henri J Suermondt and Gregory F Cooper. An evaluation of explanations of probabilistic inference. *Computers and Biomedical Research*, 26(3):242–254, 1993. doi: 10.1006/cbmr.1993.1017.
- [335] L Richard Ye and Paul E Johnson. The impact of explanation facilities on user acceptance of expert systems advice. *Mis Quarterly*, 19(2):157–172, 1995. doi: 10.2307/249686.
- [336] Vincent AWMM Alevin and Kenneth R Koedinger. An effective metacognitive strategy: Learning by doing and explaining with a computer-based cognitive tutor. *Cognitive science*, 26(2):147–179, 2002. doi: 10.1016/s0364-0213(02)00061-7.
- [337] Maaike Harbers, Joost Broekens, Karel Van Den Bosch, and John-Jules Meyer. Guidelines for developing explainable cognitive models. In *Proceedings of ICCM*, pages 85–90, Berlin, Germany, 2010. Citeseer.
- [338] Maaike Harbers, Karel van den Bosch, and John-Jules Meyer. Design and evaluation of explainable bdi agents. In *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, volume 2, pages 125–132, Toronto, Canada, 2010. IEEE. doi: 10.1109/wi-iat.2010.115.
- [339] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. Manipulating and measuring model interpretability. In *NIPS Women in Machine Learning Workshop*, Long Beach, California, USA, 2017. NIPS.
- [340] Joe Tullio, Anind K Dey, Jason Chalecki, and James Fogarty. How it works: a field study of non-technical users interacting with an intelligent system. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 31–40, San Jose, California, USA, 2007. ACM. doi: 10.1145/1240624.1240630.
- [341] Isaac Lage, Andrew Ross, Samuel J Gershman, Been Kim, and Finale Doshi-Velez. Human-in-the-loop interpretability prior. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS*, pages 10180–10189, Montréal, Canada, 2018. Neural Information Processing Systems Foundation, Inc.
- [342] Hiva Allahyari and Niklas Lavesson. User-oriented assessment of classification model understandability. In *11th scandinavian conference on Artificial intelligence*, pages 11–19, Trondheim, Norway, 2011. IOS Press. doi: 10.3233/978-1-60750-754-3-11.
- [343] Johan Huysmans, Karel Dejaeger, Christophe Mues, Jan Vanthienen, and Bart Baesens. An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models. *Decision Support Systems*, 51(1):141–154, 2011. doi: 10.1016/j.dss.2010.12.003.
- [344] Mitja Luštrek, Matjaž Gams, Sanda Martinčić-Ipšić, et al. Comprehensibility of classification trees—survey design validation. In *6th International Conference on Information Technologies and Information Society-ITIS2014*, pages 46–61, Šmarješke toplice, Slovenia, 2014. ITIS.
- [345] Todd Kulesza, Simone Stumpf, Weng-Keen Wong, Margaret M Burnett, Stephen Perona, Andrew Ko, and Ian Oberst. Why-oriented end-user debugging of naive bayes text classification. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 1(1):2:1–2:31, 2011. doi: 10.1145/2030365.2030367.
- [346] Brian Y Lim and Anind K Dey. Assessing demand for intelligibility in context-aware applications. In *Proceedings of the 11th international conference on Ubiquitous computing*, pages 195–204, Orlando, Florida, USA, 2009. ACM. doi: 10.1145/1620545.1620576.
- [347] Vanessa Putnam and Cristina Conati. Exploring the need for explainable artificial intelligence (xai) in intelligent tutoring systems (its). In *Joint Proceedings of the ACM IUI 2019 Workshops co-located with the 24th ACM Conference on Intelligent User Interfaces (ACM IUI)*, volume 2327, Los Angeles, California, USA, 2019. CEUR-WS.org.

- [348] Pierre Stock and Moustapha Cisse. Convnets and imagenet beyond accuracy: Understanding mistakes and uncovering biases. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 498–512, Munich, Germany, 2018. Springer. doi: 10.1007/978-3-030-01231-1\\_31.
- [349] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6541–6549, Honolulu, Hawaii, USA, 2017. IEEE. doi: 10.1109/cvpr.2017.354.
- [350] Sebastian Lapuschkin, Alexander Binder, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. Analyzing classifiers: Fisher vectors and deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2912–2920, Las Vegas, Nevada, USA, 2016. IEEE. doi: 10.1109/cvpr.2016.318.
- [351] Vignesh Srinivasan, Sebastian Lapuschkin, Cornelius Hellge, Klaus-Robert Müller, and Wojciech Samek. Interpretable human action recognition in compressed domain. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1692–1696, New Orleans, Louisiana, USA, 2017. IEEE. doi: 10.1109/ICASSP.2017.7952445.
- [352] Irene Sturm, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller. Interpretable deep neural networks for single-trial eeg classification. *Journal of neuroscience methods*, 274:141–145, 2016. doi: 10.1016/j.jneumeth.2016.10.008.
- [353] Yanzhuo Ding, Yang Liu, Huanbo Luan, and Maosong Sun. Visualizing and understanding neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1150–1159, Vancouver, Canada, 2017. Association for Computational Linguistics. doi: 10.18653/v1/p17-1106.
- [354] Roy Assaf and Anika Schumann. Explainable deep neural networks for multivariate time series predictions. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 6488–6490, Macao, China, 2019. International Joint Conferences on Artificial Intelligence Organization. doi: 10.24963/ijcai.2019/932.
- [355] Dumitru Erhan, Yoshua Bengio, Aaron Courville, and Pascal Vincent. Visualizing higher-layer features of a deep network. *University of Montreal*, 1341(3):1, 2009.
- [356] William F Lawless, Ranjeev Mittu, Donald Sofge, and Laura Hiatt. Artificial intelligence, autonomy, and human-machine teams: Interdependence, context, and explainable ai. *AI Magazine*, 40(3):5–13, 2019.
- [357] Lauren Gordon, Teodor Grantcharov, and Frank Rudzicz. Explainable artificial intelligence for safe intraoperative decision support. *JAMA surgery*, 154(11):1064–1065, 2019. doi: 10.1001/jamasurg.2019.2821.
- [358] Andreas Holzinger. Interactive machine learning for health informatics: when do we need the human-in-the-loop? *Brain Informatics*, 3(2):119–131, 2016. doi: 10.1007/s40708-016-0042-6.
- [359] S Ram. The importance of ensuring artificial intelligence and machine learning can be understood at the human level: Sudha ram. *European Journal of Public Health*, 29(Supplement.4):ckz185–259, 2019. doi: 10.1093/eurpub/ckz185.259.
- [360] Robert Andrews, Joachim Diederich, and Alan B Tickle. Survey and critique of techniques for extracting rules from trained artificial neural networks. *Knowledge-based systems*, 8(6):373–389, 1995. doi: 10.1016/0950-7051(96)81920-4.
- [361] Katja Hansen, David Baehrens, Timon Schroeter, Matthias Rupp, and Klaus-Robert Müller. Visual interpretation of kernel-based prediction models. *Molecular Informatics*, 30(9):817–826, 2011. doi: 10.1002/minf.201100059.

## Appendix A. Appendix

Table A.2: Classification of the systematic and literature review articles on explainable artificial intelligence and machine learning interpretability

Category	Subcategory	Reference
Application fields	Finance, military and transportation	[1]
Application fields	Law	[1] [9] [10]
Application fields	Healthcare	[1] [48] [357] [358] [359] [9] [10]
Application fields	Human-computer interaction	[356]
Approaches - data-driven	Bayesian networks	[18]
Approaches - data-driven	Neural networks	[360] [25] [130] [20] [39] [26] [41] [27] [24] [28] [29] [17] [30]
Approaches - data-driven	Support vector machines	[31] [19]
Approaches - knowledge-driven	Expert systems	[32]
Approaches - knowledge-driven	Intelligent systems	[33]
Approaches - knowledge-driven	Recommender systems	[34]
Theories and concepts	See expanded subcategories in Sec. 5	[75] [64] [36] [66] [13] [56] [23] [97] [35]
Output formats	Rule	[360] [44] [46] [42] [45] [43] [41] [24] [40] [19]
Output formats	Textual	[38] [37] [32]
Output formats	Visual	[20] [39] [40] [30]
Problem Types	Classification	[31] [20] [48] [43] [41] [49] [47] [30]
Problem Types	Regression	[47]
Generic reviews	-	[1] [50] [53] [51] [21] [52]

Table A.3: Model agnostic methods for explainability generating numerical explanations, classified according to the output format, stage (AH: Ante-hoc, PH: Post-hoc), type of problems (C: Classification, R: Regression), scope (G: Global, L: Local) and input data (NC: Numerical / Categorical, P: Pictorials, T: Textual, TS: Time Series) of the model.

Method for explainability	Authors	Ref	Year	Stage	Scope	Problem	Input
Distill-and-Compare	Tan et al.	[148]	2018	PH	G	C / R	NC
Explain and Ime	Robnik-Sikonja and Kononenko	[150]	2008	PH	L	C	NC
	Robnik-Sikonja	[151]	2018				
Feature contribution	Strumbelj and Kononenko	[154]	2010	PH	L	C / R	NC
	Strumbelj	[155]	2013				
	Kononenko et al.	[156]	2009				
Feature contribution	Strumbelj and Kononenko	[157]	2008	PH	G	C / R	NC
	Strumbelj et al.	[158]	2010				
Feature Importance	Henelius et al.	[163]	2014	PH	G	C	NC
Feature perturbation	Strumbelj and Kononenko	[164]	2014	PH	G	C / R	NC
Global Sensitivity Analysis (GSA)	Cortez and Embrechts	[152]	2011	PH	G	C / R	NC
	Cortez and Embrechts	[153]	2013				
Gradient Feature Auditing (GFA)	Adler et al.	[160]	2018	PH	G	C / R	NC
Influence functions	Koh and Liang	[161]	2017	PH	G	C	P
Monotone Influence Measures	Sliwinski et al.	[162]	2017	PH	L	C	P
Quantitative Input Influence (QII) functions	Datta et al.	[159]	2016	PH	G	C	NC
SHapley Additive exPlanations (SHAP)	Lundberg and Lee	[149]	2017	PH	G	C	P

Table A.4: Model agnostic methods for explainability generating mixed explanations, classified according to the output format, stage (AH: Ante-hoc, PH: Post-hoc), type of problems (C: Classification, R: Regression), scope (G: Global, L: Local) and input data (NC: Numerical / Categorical, P: Pictorials, T: Textual, TS: Time Series) of the underlying model.

Method for explainability	Authors	Ref	Year	Stage	Scope	Problem	Input
Bayesian Teaching	Yang and Shafto	[186]	2017	PH	L	C / R	NC
Evasion-Prone Samples Selection	Liu et al.	[190]	2018	PH	G	C	T
ExplAIner	Spinner et al.	[133]	2019	PH	G	C / R	P; NC; TS
Functional ANOVA decomposition, Variable Interaction Network graph	Hooker	[183]	2004	PH	G	C / R	NC
Justification Narratives	Biran and McKeown	[184]	2014	PH	G	C	NC
Local Interpretable Model-Agnostic Explanations (LIME)	Ribeiro et al.	[193]	2016	PH	L	C	P; T
Maximum Mean Discrepancy (MMD)-critic	Kim et al.	[191]	2016	PH	L	C	P
Neighborhood-Based Explanations	Caruana et al.	[189]	1999	PH	L	C	NC
Pertinent negatives	Dhurandhar et al.	[192]	2018	PH	L	C	P; NC
Rivelo	Tamagnini et al.	[185]	2017	PH	L	C	T
Sequential Bayesian Quadrature	Khanna et al.	[187]	2019	PH	L	C	P; NC
Set Cover Optimization (SCO)	Bien et al.	[188]	2011	PH	L	C	P; NC

Table A.5: Model agnostic methods for explainability generating rule-based explanations, classified according to the output format, stage (AH: Ante-hoc, PH: Post-hoc), type of problems (C: Classification, R: Regression), the scope (G: Global, L: Local) and input data (NC: Numerical/Categorical, P: Pictorials, T: Textual, TS: Time Series) of the model.

Method for explainability	Authors	Ref	Year	Stage	Scope	Problem	Input
Anchors	Ribeiro et al.	[140]	2018	PH	G / L	C	T
Automated Reasoning	Bride et al.	[165]	2018	PH	G	C	NC
Genetic Rule EXtraction (G-REX)	Johansson et al.	[166]	2004	PH	G	C / R	NC
Model Extraction	Bastani et al.	[168]	2017	PH	G	C / R	NC
Partition Aware Local Model (PALM)	Krishnan and Wu	[169]	2017	PH	G	C / R	NC

Table A.6: Model agnostic methods for explainability generating visual explanations, classified according to output format, stage (AH: Ante-hoc, PH: Post-hoc), type of problems (C: Classification, R: Regression), the scope (G: Global, L: Local) and input data (NC: Numerical/Categorical, P: Pictorials, T: Textual, TS: Time Series) of the underlying model.

Method for explainability	Authors	Ref	Year	Stage	Scope	Problem	Input
Class Signatures	Krause et al.	[179]	2016	PH	G	C / R	NC
ExplainD	Poulin et al.	[180]	2006	PH	G	C	NC
Explanation Graph based on perturbed input element order	Alvarez-Melis and Jaakkola	[177]	2017	PH	L	C	T
Image Perturbation	Fong and Vedaldi	[170]	2017	PH	L	C	P
Individual Conditional Expectation	Goldstein et al.	[175]	2015	PH	G	C / R	NC
iVisClassifier	Choo et al.	[172]	2010	PH	G	C	NC
Layer-wise Relevance Propagation	Bach et al.	[101]	2015	PH	L	C	P
Manifold	Zhang et al.	[181]	2019	PH	G	C / R	NC
MLCube Explorer	Kahng et al.	[182]	2016	PH	G	C	NC
Partial Importance and Individual Conditional Importance plots based on Shapley feature importance	Casalicchio et al.	[176]	2018	PH	G	C / R	NC
Restricted Support Region Set (RSRS) Detection	Liu and Wang	[171]	2012	PH	L	C	T
Saliency Detection	Dabkowski and Gal	[173]	2017	PH	L	C	P
Sensitivity analysis	Bachrens et al.	[174]	2010	PH	L	C	P; NC
Spectral Relevance Analysis (SpRAY)	Lapuschkin et al.	[8]	2019	PH	G	C	P
Worst-case perturbations	Goodfellow et al.	[178]	2015	PH	L	C	P

Table A.7: Methods for explainability for neural networks generating visual explanations, classified according to the stage (AH: Ante-hoc, PH: Post-hoc), type of problems (C: Classification, R: Regression), scope (G: Global, L: Local) and input data (NC: Numerical/Categorical, P: Pictorials, T: Textual, TS: Time Series).

Method for explainability	Authors	Ref	Year	Stage	Scope	Problem	Input
Activation maps	Hamidi-Haines et al.	[225]	2019	PH	L	C	P
		[222]					
Activation Maximization	Erhan et al., Nguyen et al.	[223], [224]	2010, 2016	PH	L	C	P
ActiVis	Kahng et al.	[231]	2018	PH	G	C / R	NC
And-Or Graph (AOG)	Zhang et al.	[229]	2017	PH	G	C	P
Cell Activation Values	Karpathy et al.	[226]	2016	PH	G / L	C	T
CLass-Enhanced Attentive Response (CLEAR)	Kumar et al.	[194]	2017	PH	L	C	NC
Cnn-Inte	Liu et al.	[215]	2018	PH	G	C	P
Compositionality	Li et al.	[213]	2016	PH	L	C	T
Data-flow graphs	Wongsuphasawat et al.	[136]	2018	PH	G	C / R	P; NC; T
Deep Learning Important Features (DeepLIFT)	Shrikumar et al.	[100]	2017	PH	L	C	P; NC
Deep View (DV)	Zhong et al.	[233]	2017	PH	G	C / R	P
Deep Visualization Toolbox	Yosinski et al.	[232]	2015	PH	G	C	P
Deep-Taylor Decomposition	Montavon et al.	[208]	2017	PH	G	C	P
DeepResolve	Liu and Gifford	[195]	2017	PH	G	C	NC
Explanatory Graph	Zhang et al.	[227]	2018	PH	G	C	P
Feature maps	Zhang et al.	[204]	2018	AH	L	C	P
Generative Adversarial Network (GAN) Dissection	Bau et al.	[220]	2019	PH	L	C	P
GradCam	Selvaraju et al.	[196]	2017	PH	L	C	P
Guided BackProp and Occlusion	Goyal et al.	[198]	2016	PH	L	C	P
Guided Feature Inversion	Du et al.	[202]	2018	PH	L	C	P
Hidden Activity Visualization	Rauber et al.	[218]	2017	PH	G	C	P
Important Neurons and Patches	Lengerich et al.	[221]	2017	PH	G	C	P
iNNvestigate	Alber et al.	[234]	2019	PH	L	C	P
Integrated Gradients	Sundararajan et al.	[93]	2017	PH	L	C	P
Inverting Representations	Mahendran and Vedaldi	[201]	2015	PH	L	C	P
LRP w/ Relevance Conservation	Arras et al.	[206]	2017	PH	L	C	T
LRP w/ Local Renormalization Layers	Binder et al.	[207]	2016	PH	L	C	P
LSTMVis	Strobelt et al.	[135]	2018	PH	G / L	C	T
N <sup>2</sup> VIS	Streeter et al.	[236]	2001	PH	G	C / R	NC
Net2Vec	Fong and Vedaldi	[200]	2018	PH	G	C	P
Neural Network and CBR Twin-systems	Kenny and Keane	[212]	2019	PH	L	C	P
OcclusionSensitivity	Zeiler and Fergus	[199]	2014	PH	G	C	P
OpenBox	Chu et al.	[214]	2018	PH	G	C	P; NC
PatternNet, PatternAttribution	Kindermans et al.	[210]	2018	PH	L	C	P
Prediction Difference Analysis	Zintgraf et al.	[205]	2017	PH	L	C	P
Principal Component Analysis	Aubry and Russell	[216]	2015	PH	G	C	P
Receptive Fields	He and Pugeault	[209]	2017	PH	G	C	P
Relevant Features Selection	Mogrojejo et al.	[211]	2019	PH	L	C	P
Saliency maps	Oláh et al.	[230]	2018	PH	G / L	C	P
Saliency maps	Simonyan et al.	[98]	2014	PH	L	C	P
Seq2seq-Vis	Strobelt et al.	[235]	2018	PH	L	C	T
SmoothGrad	Smilkov et al.	[203]	2017	PH	L	C	P
Stacking w/ Auxiliary Features (SWAF)	Rajani and Mooney	[197]	2017	PH	L	C	P
Symbolic Graph Reasoning (SGR)	Liang et al.	[228]	2018	AH	G	C / R	P
t-SNE maps	Zahavy et al.	[217]	2016	PH	G	C	NC
TreeView	Thiagarajan et al.	[219]	2016	PH	G	C	P

Table A.8: Methods for explainability for neural networks generating rule-based explanations, classified according to the stage (AH: Ante-hoc, PH: Post-hoc), type of problems (C: Classification, R: Regression), scope (G: Global, L: Local) and input data (NC: Numerical/Categorical, P: Pictorials, T: Textual, TS: Time Series) of the underlying model.

Method for explainability	Authors	Ref	Year	Stage	Scope	Problem	Input
C4.5Rule-PANE	Zhou and Jiang	[248]	2003	PH	L	C / R	NC
DecText	Boz	[249]	2002	PH	G	C	NC
		[139]					
Discretized Interpretable Multi Layer Perceptrons (DIMLP)	Bologna and Hayashi	[240]	2017	PH	G / L	C	P; NC; T
	Bologna	[241]	1998				
		[132]	2018				
Discretizing Hidden Unit Activation Values by Clustering	Setiono and Liu	[237]	1995	PH	G	C	NC
DT extraction	Frosst and Hinton	[245]	2017	PH	G	C	P
	Zhang et al.	[246]	2019				
Interval Propagation	Palade et al.	[141]	2001	PH	G	C	NC
Iterative Rule Knowledge Distillation	Hu et al.	[254]	2016	AH	G	C	T
Neural Network Knowledge eXtraction (NNKX)	Bondarenko et al.	[238]	2017	PH	G	C	NC
Rule Extraction From Neural network Ensemble (REFNE)	Zhou et al.	[247]	2003	PH	G	C / R	NC
Rule Extraction from Neural Network using Classified and Misclassified data (RxNCM)	Biswas et al.	[243]	2017	PH	G	C	NC
Rule Extraction by Reverse Engineering (RxREN)	Augasta and Kathirvalavakumar	[242]	2012	AH	G	C / R	NC
Symbolic logic integration	Tran	[255]	2017	AH	G	C / R	NC
Symbolic rules	Garcez et al.	[244]	2001	PH	G	C / R	NC
Tree Regularization	Wu et al.	[252]	2018	AH	G	C	NC
Trepan	Craven and Shavlik	[250]	1994	PH	G	C / R	NC
	Craven and Shavlik	[251]	1996				
Validity Interval Analysis (VIA)	Thrun	[239]	1995	PH	G	C / R	TS
Word Importance Scores	Murdoch and Szlam	[253]	2017	PH	G	C	T

Table A.9: Methods for explainability for neural networks generating textual and numerical explanations, classified according to the output format (N: Numerical, T: Textual), stage (AH: Ante-hoc, PH: Post-hoc), type of problems (C: Classification, R: Regression), scope (G: Global, L: Local) and input data (NC: Numerical/Categorical, P: Pictorials, T: Textual, TS: Time Series) of the underlying model.

Method for explainability	Authors	Ref	Year	Output Format	Stage	Scope	Problem	Input
Causal Importance	Féraud and Clérot	[261]	2002	N	PH	G	C	NC
Concept Activation Vectors	Kim et al.	[145]	2018	N	PH	L	C	P
Contextual Importance, Utility	Främling	[262]	1996	N	PH	G / L	C	NC
InterpNET	Barratt	[256]	2017	T	PH	L	C	P
Most-Weighted-Path, Most-Weighted-Combination, Maximum-Frequency-Difference	García-Magariño et al.	[257]	2019	T	PH	L	C	TS
Probes	Alain and Bengio	[144]	2017	N	PH	G	C	P
Rationales	Lei et al.	[258]	2016	T	PH	L	C	T
REcurrent LEXicon NETwork (RELEXNET)	Clos et al.	[263]	2017	N	AH	G	C	T
Relevance, Discriminative Loss	Hendricks et al.	[259]	2016	T	PH	L	C	P
Singular Vector Canonical Correlation Analysis (SVCCA)	Raghuram et al.	[137]	2018	N	PH	G	C / R	P

Table A.10: Methods for explainability for neural networks generating mixed explanations, classified according to the stage (AH: Ante-hoc, PH: Post-hoc), type of problems (C: Classification, R: Regression), scope (G: Global, L: Local) and input data (NC: Numerical/Categorical, P: Pictorials, T: Textual, TS: Time Series) of the underlying model.

Method for explainability	Authors	Ref	Year	Stage	Scope	Problem	Input
Activation values of hidden neurons	Tamajka et al.	[268]	2019	AH	L	C	P
Attention Alignment	Kim et al.	[264]	2018	PH	L	C	P
Deterministic Finite Automaton	Mayr and Yovine	[266]	2018	PH	L	C	NC
Deterministic Finite-state Automata (DFAs)	Omlík and Giles	[267]	1996	PH	G	C	NC
Image Caption Generation with Attention Mechanism	Xu et al.	[147]	2015	PH	L	C	P
Pointing and Justification Model (PJ-X)	Park et al.	[265]	2018	AH	L	C	P

Table A.11: Methods for explainability for rule-based construction approaches, classified according to the output format (M: Mixed, N: Numerical, R: Rules, T: Textual, N: Numerical), stage (AH: Ante-hoc, PH: Post-hoc), type of problems (C: Classification, R: Regression), scope (G: Global, L: Local) and input data (NC: Numerical/Categorical, P: Pictorials, T: Textual, TS: Time Series) of the underlying model.

Method for explainability	Authors	Ref	Year	Output Format	Stage	Scope	Problem	Input
Ant Colony Optimization (ACO)	Otero and Freitas	[269]	2016	R	AH	G	C	NC
AntMiner+ and ALBA	Verbeke et al.	[270]	2011	R	AH	G	C	NC
Argumentation	Rizzo and Longo	[142]	2018	R	AH	G	C	NC
Argumentation	Zeng et al.	[143]	2018	R	AH	G	C / R	P
Bayesian Rule Lists (BRL)	Letham et al.	[274]	2013	R	AH	G	C	NC
		[275]	2015					
Bayesian Rule Sets (BRS)	Wang et al.	[276]	2016	R	AH	G	C	NC
		[277]	2017					
Interpretable Decision Set	Lakkaraju et al.	[272]	2016	R	AH	G	C	NC
Exception Directed Acyclic Graphs (EDAGs)	Gaines	[271]	1996	M	AH	G	C	NC
ExpliClas	Alonso	[289]	2019	M	PH	L	C	NC
First Order Combined Learner (FOCL)	Pazzani	[278]	1997	R	AH	G	C	NC
Fuzzy logic	Pierrard et al.	[282]	2018	R	AH	L	C	NC
Fuzzy system	Jin	[281]	2000	R	AH	G	C	NC
Fuzzy Inference-Grams (Fingrams)	Pancho et al.	[288]	2013	V	PH	G	C	NC
Fuzzy Inference Systems	Keneni et al.	[287]	2019	T	PH	L	C	TS
Genetics-Based Machine Learning (GBML)	Ishibuchi and Nojima	[280]	2007	R	AH	G	C	NC
Interpretable Classification Rule Mining (ICRM)	Cano et al.	[284]	2013	R	AH	G	C	NC
Linear Programming Relaxation	Malioutov et al.	[285]	2017	R	AH	G	C	NC
	Su et al.	[286]	2016					
Multi-Objective Evolutionary Algorithms based Interpretable Fuzzy (MOEAF)	Wang and Palade	[283]	2011	R	AH	G	C	NC
Mycin	Shortliffe et al.	[3]	1975	T	PH	L	C	NC

Table A.12: Methods for explainability for data-driven approaches, classified according to the output format (M: Mixed, N: Numerical, R: Rules, T: Textual, N: Numerical), stage (AH: Ante-hoc, PH: Post-hoc), type of problems (C: Classification, R: Regression), scope (G: Global, L: Local) and input data (NC: Numerical/Categorical, P: Pictorial, T: Textual, TS: Time Series).

Method for explainability	Authors	Ref	Year	Construction approach	Output Format	Stage	Scope	Problem	Input
Contribution Propagation	Landecker et al.	[308]	2013	Hierarchical networks	V	PH	L	C	P
DT extraction	Andrzejak et al.	[294]	2013	Distributed DTs	R	PH	G	C / R	NC
DT extraction	Ferré et al.	[292]	2002						
	Van Assche and Blockeel	[293]	2007	Ensembles	R	PH	G	C	NC
DT extraction	Alonso et al.	[291]	2018	Ensembles	T	PH	L	C	NC
Discriminative Patterns	Gao et al.	[297]	2017	Ensembles	T	PH	G	C	T
Explaining Bayesian network Inferences (EBI)	Yap et al.	[305]	2008	Bayesian networks	R	PH	G	C	NC
ExtractRule	Fung et al.	[138]	2005	Hyperplane-Based Linear Classifiers	R	PH	G	C	P; NC
Factorized Asymptotic Bayesian (FAB) inference	Hara and Hayashi	[295]	2018	Ensembles	R	PH	G	C	NC
Feature Tweaking	Tolomei et al.	[290]	2017	Ensembles	N	PH	L	C	NC
Important Support Vectors and Border Classification inTrees	Barbelli et al.	[146]	2009	SVM	N	PH	L	C	NC
Nomograms	Deng	[296]	2018	Ensembles	R	PH	G	C / R	NC
Nomograms	Jakulin et al.	[302]	2005	SVM	V	PH	G	C	NC
Nomograms	Možina et al.	[304]	2004	Naïve Bayes	V	PH	G	C	NC
Probabilistically Supported Arguments	Timmer et al.	[307]	2017	Bayesian networks	M	PH	G	C	NC
Scenarios	Vlek et al.	[306]	2016	Bayesian networks	T	PH	L	C	NC
Self-Organizing Maps	Hamel	[301]	2006	SVM	V	PH	G	C	NC
SVM+Prototypes Tree Space Prototypes	Núñez et al.	[299]	2002	SVM	M	PH	G	C	NC
Tree Space Prototypes	Tan et al.	[298]	2016	Ensembles	M	PH	L	C	NC
Visualization for Risk Factor Analysis (VRIFA)	Cho et al.	[303]	2008	SVM	V	PH	G	C	NC
Weighted Linear Classifier	Caragea et al.	[300]	2003	SVM	N	PH	G	C	NC

Table A.13: Methods for explainability generating white-box models, classified according to the output format (M: Mixed, N: Numerical, R: Rules, T: Textual, N: Numerical), stage (AH: Ante-hoc, PH: Post-hoc), type of problems (C: Classification, R: Regression), scope (G: Global, L: Local) and input data (NC: Numerical/Categorical, P: Pictorial, T: Textual, TS: Time Series) of the underlying model.

Method for explainability	Authors	Ref	Year	Construction approach	Output Format	Stage	Scope	Problem	Input
Bayesian Case Model (BCM)	Kim et al.	[309]	2014	Bayesian case-based reasoning	M	AH	L	C	P; T
Gaussian Process Regression (GPR)	Caywood et al.	[310]	2017	Gaussian Process Regression	N	AH	G	R	TS
General Additive Models (GAMs)	Lou et al.	[99]		Additive models	M	AH	G	C / R	NC
GAMs with pairwise interactions (GA <sup>2</sup> Ms)	Lou et al.	[131]	2012	Additive models	M	AH	G	C / R	NC
Mind the Gap Model (MGM)	Caruana et al.	[311]	2015	Mind the Gap Model (MGM)	M	AH	G	C	P; NC; T
Multi-Run Subtree Encapsulation	Howard and Edwards	[314]	2018	Tree-based Genetic Programming	M	AH	G	C	NC
Oblique Treed Sparse Additive Models (OT-SpAMs)	Wang et al.	[312]	2015	Additive models	N	AH	G	C	NC
Probabilistic Sentential Decision Diagrams (PSDD)	Liang and Van den Broeck	[315]	2017	Probabilistic sentential decision diagrams	R	AH	G	C / R	NC
Supersparse Linear Integer Model (SLIM)	Ustun et al.	[317]	2014	Scoring systems	N	AH	G	C	NC
Unsupervised Interpretable Word Sense Disambiguation	Panchenko et al.	[318]	2017	UIWSD	V	AH	G	C	T
Transparent Generalized Additive Model Tree (TGAMT)	Fahner	[313]	2018	Additive models	R	AH	G	C	NC

Table A.14: List and classification of the scientific articles proposing human-centered approaches to evaluate methods for explainability, classified according to the construction approach, the type of measurement employed (qualitative or quantitative), and the format of their output explanation.

Measure type	Construction approach	Output format	Reference
Qualitative	Autonomous agents	Textual	[54]
Qualitative	Context-aware systems	Textual	[346] [77]
Qualitative	Context-aware systems	Visual	[346]
Qualitative	Data clustering	Mixed	[316]
Qualitative	Expert Systems	Textual	[334] [335]
Qualitative	Learning Systems	Textual	[347]
Qualitative	ML (Model agnostic)	Visual	[103] [340]
Qualitative	Fisher vector classifiers	Visual	[350]
Qualitative	Support Vector Machine	Visual	[351]
Qualitative	Neural networks	Visual	[354] [355] [353] [352]
Quantitative	Case-based Reasoning	Mixed	[309]
Quantitative	Context-aware systems	Textual	[77]
Quantitative	Neural networks	Visual	[348]
Quantitative	Neural networks	Visual	[349]
Quantitative	Decision trees	Mixed	[344]
Quantitative	Kernel-based neural networks	Visual	[361]
Quantitative	Learning Systems	Textual	[336] [337] [338]
Quantitative	ML (Model agnostic)	Mixed	[341]
Quantitative	ML (Model agnostic)	Rules	[140]
Quantitative	ML (Model agnostic)	Mixed	[339]
Quantitative	ML (Model agnostic)	Visual	[133]
Quantitative	Naïve Bayes	Textual	[345] [60]
Quantitative	Rule-based classifier	Mixed	[342] [343]
Quantitative	Decision trees	Mixed	[342] [343]
Quantitative	Decision tables	Mixed	[343]

Table A.15: Classification of the scientific articles proposing comparative approaches to evaluate methods for explainability, classified according to the methodology followed to carry out the comparison task.

Comparison approach	Reference
Sensitivity to input perturbation	[324] [325] [90] [326] [327] [331] [332] [92] [330] [329]
Sensitivity to model parameter randomization	[324] [325] [355]
Explanation completeness	[333]

Table A.16: List of the methods for explainability generating visual explanations, such as heat-maps, whose degree of explainability is evaluated by comparison (listed in the fourth column). These comparisons were carried out over different types of input data (listed in the third column).

Method for explainability (references)	Acronym	Input type	Compared with (references)
Deep-Taylor Decomposition [208]	DTD	Pictorial	GBP, IG, SA, PM in [92]
DeepLIFT [100]	DLT	Pictorial	IG, LRP in [326] IG, SA in [332]
Gradient*Input [100]	GI	Pictorial	GC, GBP and SG [324] GC, GBP, IG, SG [325] IG, LIME, OS, SM, SHAP in [90]
GrandCam [196]	GC	Pictorial	GI, GBP, SG [324] GI, GBP, IG, SG [325]
Guided BackProp [198]	GBP	Pictorial	GI, GC, SG [324] GI, GC, IG, SG [325] DTD, IG, SA, PM in [92]
Integrated Gradients [93]	IG	Pictorial	GI, GC, GBP, SG [325] LRP, LIME, OS, SM, SHAP [90] DLT and LRP in [326] DLT, SA in [332] DTD, GBP, SA, PM in [92]
Layer-wise Relevance Propagation [101]	LRP	Pictorial	IG, LIME, OS, SM, SHAP in [90] DLT, IG in [326] SA in [331] [329] [330] OS in [331]
Local Interpretable Model-Agnostic Explanations [134]	LIME	Pictorial Numerical / Categorical	GI, IG, OS, SM, SHAP in [90]
Occlusion Sensitivity [199]	OS	Pictorial	GI, IG, LIME, SM, SHAP in [90] LRP in [331]
Sensitivity Analysis [174]	SA	Pictorial	LRP in [331] DLT, IG in [332] DTD, GBP, IG, PM in [92] LRP in [329] [330]
PatternNet [210]	PM	Pictorial	DTD, GBP, IG, SA in [92]
Saliency Maps [98]	SM	Pictorial	GI, IG, LIME, OS, SHAP in [90]
SHapley Additive exPlanations [149]	SHAP	Pictorial	GI, IG, LIME, OS, SM in [90]
SmoothGrad [203]	SG	Pictorial	GI, GC, GBP [324] GI, GC, GBP and IG [325]
Layer-wise Relevance Propagation [101]	LRP	Textual	SA in [327] [330]
Sensitivity Analysis methods [174]	SA	Textual	LRP in [327] [330]