Information Sciences 615 (2022) 238–292

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Explainability of artificial intelligence methods, applications and challenges: A comprehensive survey

Weiping Ding [a,*], Mohamed Abdel-Basset [b], Hossam Hawash [b], Ahmed M. Ali [b]

[a] School of Information Science and Technology, Nantong University, Nantong 226019, China
[b] Faculty of Computers and Informatics, Zagazig University, Zagazig, Sharqiyah 44519, Egypt

ARTICLE INFO

Article history:
Received 24 February 2022
Received in revised form 31 August 2022
Accepted 3 October 2022
Available online 10 October 2022

Keywords:
Black-box
White-box
Explainable AI
Responsible AI
Machine learning
Deep learning

ABSTRACT

The continuous advancement of Artificial Intelligence (AI) has been revolutionizing the strategy of decision-making in different life domains. Regardless of this achievement, AI algorithms have been built as *Black-Boxes*, that is as they hide their internal rationality and learning methodology from the human leaving many unanswered questions about *how* and *why* the AI decisions are made. The absence of explanation results in a sensible and ethical challenge. Explainable Artificial Intelligence (XAI) is an evolving subfield of AI that emphasizes developing a plethora of tools and techniques for unboxing the Black-Box AI solutions by generating human-comprehensible, insightful, and transparent explanations of AI decisions. This study begins by discussing the primary principles of XAI research, Black-Box problems, the targeted audience, and the related notion of explainability over the historical timeline of the XAI studies and accordingly establishes an innovative definition of explainability that addresses the earlier theoretical proposals. According to an extensive analysis of the literature, this study contributes to the body of knowledge by driving a fine-grained, multi-level, and multi-dimension taxonomy for insightful categorization of XAI studies with the main aim to shed light on the variations and commonalities of existing algorithms paving the way for extra methodological developments. Then, an experimental comparative analysis is presented for the explanation generated by common XAI algorithms applied to different categories of data to highlight their properties, advantages, and flaws. Followingly, this study discusses and categorizes the evaluation metrics for the XAI-generated explanation and the findings show that there is no common consensus on how an explanation must be expressed, and how its quality and dependability should be evaluated. The findings show that XAI can contribute to realizing responsible and trustworthy AI, however, the advantages of interpretability should be technically demonstrated, and complementary procedures and regulations are required to give actionable information that can empower decision-making in real-world applications. Finally, the tutorial is crowned by discussing the open research questions, challenges, and future directions that serve as a roadmap for the AI community to advance the research in XAI and to inspire specialists and practitioners to take the advantage of XAI in different disciplines.

* Corresponding author.
E-mail addresses: dwp9988@163.com (W. Ding), mohamedbasset@ieee.org (M. Abdel-Basset), hossamreda@zu.edu.eg (H. Hawash), aabdelmonem@fci.zu.edu.eg (A.M. Ali).

https://doi.org/10.1016/j.ins.2022.10.013
0020-0255/

1. Introduction

Artificial Intelligence (AI) is a branch of computer science that has revolutionized the way people perform day-to-day tasks by utilizing machines with minimal human intervention to promote automated and intelligent behavior. Politicians all over the world have declared its use to be an achievable goal. Industry sees it as a booming driver, whereas medicine sees it as a great opportunity for solving medical problems, providing new insights, and improving decision support quality. Many systems will and are already being evolved by AI and machine learning. As a result, society has recently experienced a rise in the number of success stories and use cases in which AI has played a significant role in realizing unprecedented strides in equaling, if not surpassing, human performance in many tasks such as categorization, recommendation, game playing, and even medical decision-making. Machine Learning (ML) and Deep Learning (DL) are rapidly evolving sub-fields of AI that achieve extraordinary levels of performance when learning to progressively solve complex computational data-driven or data-free problems, making them critical for human civilization's future development. The complexity of AI solutions has lately advanced to the point in which no human intervention is required in their development and deployment. As a result, the user can make an informed decision with potentially far-reaching consequences and avoid costly mistakes. Such explanations improve AI models, increase trust in the system, and aid in the detection of errors and performance issues [17].

Moreover, AI solutions are being deployed in extremely sensitive policy areas (i.e., recidivism forecast in the criminal justice system, or face recognition in the police system), and in fields with a variety of societal and political authorities. Hence, now, AI solutions are integrated into a broad range of decision-making activities in almost all life sectors. Accordingly, the attention of science and policy communities has been positioned toward the extent to which the AI community, or the community influenced by AI decisions, can comprehend how the resultant decision-making process works, and the justification for a particular decision [278].

From the perspective of human decision-makers, the standard AI algorithms leave an essential question unanswered: in what way the human decision-makers can trust the outcomes of AI algorithms and justify their usage? Reaching confidence and discovering validation can scarcely be accomplished if the human is unable to gain access to a reasonable explanation for the inner computations that drive the AI decisions. For instance, given a hypothetical situation, where the therapeutic system that some AI algorithm to make forecasts about whether a patient has a life-threatening disease (e.g., COVID-19). It would be appropriate if the AI algorithms give a justification for their forecasts so that the doctors can trust them. Trustworthiness in AI decision can be gained upon a justification that is: 1) straightforwardly understandable for normal individuals or specialists; 2) relevant to the human; 3) relate the decision with contextual information regarding the choice or to the individual's preceding knowledge; and 4) echoes the midway rational of the individuals in accomplishing the decision. Considering the qualitative landscape of these attributes, a remarkable variation in the definitions, methods, and practices applied by the research community to deliver a justification for the AI decisions. This variation is further combined with the truth that the manner of a justification often fits a researcher's individual belief of what forms an "explanation" [100].

Unfortunately, the majority of ML and DL algorithms have been labeled as 'black boxes' by academics because their fundamental constructions are complex, quasi, and difficult to explain and justify to human beings. One such opacity has fashioned a need for explainable AI (XAI) algorithms, which is primarily motivated by three parameters: 1) the need to innovate and build quite translucent learning algorithm; 2) the need for methods that allow human individuals to interact and collaborate with them; and 3) the need for faith as well as trustworthiness in AI decision making. Likewise, as indicated by [283], AI systems based on data should be accountable, as accountability is assumed to become an official requirement pretty shortly. Article 22 of the General Data Protection Regulation (GDPR) [42] defines the rights and obligations associated with the use of automated decision-making. It exemplifies the "right of explanation" by granting individuals the right to obtain an explanation of the outcomes automatically generated by an AI solution, as well as to challenge and evoke a pertaining reference, particularly when it may adversely impact a human legally, financially, physiologically, or mentally. With the adoption of the GDPR article, the European Parliament battled to settle the quandary related to the spread of potentially unconscionable inferences to the community, which a computational algorithm might have managed to learn from inappropriate and misrepresentative data [197].

1.1. Related surveys

A small number of survey studies on XAI have been published in the literature. This section provides an analytical discussion of these surveys and shed the light on the contributions that distinguish this tutorial. For example, ADADI et al [2] presented an entry point survey on XAI that promotes the multidisciplinary landscape of the investigated XAI literature according to the key aspects and categories of explainability from various viewpoints. They also argued the potential of XAI in different application domains and pointed out the promising research directions in this era. In addition, Arrieta et al [21] provided a definition of explainability by placing the audience as a fundamental factor to be deemed when explaining AI algorithms. Then, they carefully analyzed and classified the XAI studies into two distinct taxonomies: 1) taxonomy that discourses the XAI algorithms exploitation the early made distinction between post-hoc explainability and transparency, including models that are transparent by themselves; 2) taxonomy considers XAI algorithms appropriate for the explaining the DL models based on the classification standards. In [100], GUIDOTTI et al studied four distinct types of problems of BlackBox models and the method to deliver an explanation by surveying group the methods for unboxing black box AI algo-

rithms according to the problem they are fronting, the type of explainability solutions projected, and the type of training, and the type of model. Vilone et al [281] surveyed the XAI studies by clustering the notions related to explainability from a theoretical standpoint, by exhaustively analyzing the previous explanation methods and explainability evaluation methods and metrics. They also discussed limitations and open research gaps that could help envision the future of XAI. In another direction, the authors of [248] emphasized surveying the XAI algorithms in the context of tabular data and argued different explainability aspects in this direction. They categorized the literature studies into four classes of techniques including Feature importance, Feature interaction, simplified models, and counterfactuals.

Owing to the criticality of decisions in medical domains, many studies had emphasized surveying the XAI from the perspective of medical application. For example, Tjoa et al [278] surveyed the interpretability of AI algorithms and taxonomized and applied them in the context of the medical domain to offer the clinical community a view on the usage of a variety of interpretable algorithms. The study focused on delivering insights about interpretability by considering the medical practices based on mathematical and technical foundations. Besides, Jiménez-Luna et al [133] reviewed the literature on XAI research by highlighting the advantages, shortcomings, and opportunities of XAI for drug discovery by restructuring the existing methods into theoretical categories. They also pointed out some possible XAI applications in the drug discovery field as well as the potential methodological advancements necessary to promote the functional pertinence of XAI in pharmaceutical research. Moreover, the authors of [306] provided a short survey of the literature on XAI algorithms for advancing the reliability of healthcare applications [61].

Thanks to the rapid and great advances made by deep learning, some survey studies had just emphasized debating and analyzing the explainability and interpretability of DL algorithms. For example, Das et al [67] provided a systematic overview of explainable DL algorithms and coarse-grained taxonomy of the XAI approaches according to three broad criteria (i.e., scope, methodology, and level of explanation) to enhance simplicity and convenience. In another way, Zhang et al [323] comprehensively defined neural network interpretability and its importance based on three factual requirements namely reliability requirements, ethical requirements, and scientific usage. Also, they proposed a three-dimensional taxonomy of literature studies based on the activeness of the approaches, the format of interpretations, and the locality of interpretability. Moreover, Liang et al [172] overviewed the representative studies and the attributes of the state-of-the-art techniques for local interpretability of DL models. They also reproduced and analyzed the experimental results obtained from different interpretability techniques to further validate their efficiency for interpreting DL models. Apart from the above studies have focused on surveying the counterfactual and plausible explanation methods for developing and unboxing the Black-Box AI solutions [61,116].

In [236], the authors emphasize that causability is not a synonym for causality in the way of measuring and guaranteeing the superiority of explanations by measuring the extent to which an XAI explanation is able to achieve a stated level of causal interpretation. The authors argued for adopting Graph Neural Networks for empowering the information fusion for multimodal capability.

Table 1 outlines the existing survey studies in the field of XAI based on different inclusion criteria in an analytical way that enables the readers to grasp the current progress and how our tutorial is more comprehensive and offers a contribution to these studies [118]. The analysis of the previous surveys brings us many findings that shape motivation for this work. Firstly, the existing surveys emphasize only some popular challenges of XAI methods, applications, evaluation metrics, etc. however, some other significant challenges are still unconsidered. Secondly, the absence of a fine-grained multi-dimensional taxonomy of explainability methods, where the taxonomies presented in the earlier surveys are depending on only some common criteria such as the scope, and methodology of explaining AI models. Thirdly, the absence of clear and detailed directives for researchers to aid them to apply XAI in real-world applications and develop future XAI solutions.

1.2. Novelty and contributions

To sum up, this study contributes to the body of knowledge as follows:

(1) First, this paper presents a contemporary and comprehensive survey that motivates the significance of XAI as an essential paradigm shift toward enabling responsible and trustworthy intelligence. Then, it delivers to the readers a concise definition of XAI drivers, goals, stockholders, and a set of helpful open-source frameworks that pave the way for further development in XAI and its applications.
(2) This study derives a novel fine-grained taxonomy for clustering the existing XAI solutions according to various categorization criteria (such as scope, methodology, data type, etc.) to give the reader a holistic view of XAI.
(3) we analytically study the recent metrics and methods (quantitative and qualitative) that serve as crucial criteria for evaluating the quality and expressiveness of generated explanations in different XAI applications. Further, we derive a taxonomy for these methods according to different categorization factors.
(4) After that, this study introduces a holistic analysis of the potential of XAI in a variety of real-world applications such as smart transportation, smart healthcare, cybersecurity, economy, justice system, and networking and communications.
(5) Finally, this study points out the state-of-the-art research challenges, open issues, and possible solutions; then collates potentially interesting and promising future research directions that we think are effective for improving the explainability and responsibility of AI solutions.

Table 1
Comparison for the important related XAI surveys through explainability dimensions.

Ref	Year	Focus area	Tax	Eval	ORC	Apps.	#CP
[2]	*2018*	entry point to XAI	✓ (3 criteria)	✓	×	✓ (5)	180
[100]	*2018*	Explainability	×	×	×	×	143
[67]	*2020*	Explainability	✓ (Two-level)	×	×	×	119
[323]	*2021*	Interpretability of NN	✓ (Three-level)	✓	×	×	150
[316]	*2020*	Explainability of GNN	✓ (Three-level)	✓	×	×	85
[70]	*2021*	Explainability of RL	×	×	×	×	249
[272]	*2021*	counterfactual explanation	✓ (Two-level)	×	×	×	166
[21]	*2020*	Explanation methods of AI	✓ (Multi-level)	×	✓	×	426
[248]	2021	Explainability of AI for tabular data	×	×	✓	✓ (4)	228
[84]	2021	Interpretability of NN	✓ (Two-level)	×	×	✓ (2)	212
[172]	2021	local interpretation of NN	✓ (Two-level)	×	×	×	113
[133]	2020	XAI for drug discovery	×	×	✓	✓	172
[17]	2021	XAI for pattern recognition	×	×	×	×	133
[278]	2020	Medical XAI	×	×	×	×	169
[281]	2021	Explainability and evaluation of XAI	✓	✓	×	×	189
[306]	2022	Medical XAI with multimodality	×	×	×	✓ medical	188
[197]	2021	Multidisciplinary study of XAI	×	✓	×	×	219
[61]	2022	Counterfactuals &Causability	✓	×	✓	×	164
[116]	2022	Counterfactuals &Causability	×	×	✓	✓ medical	237
[236]	2022	Generic XAI	✓	×	✓	×	/
[118]	2021	multi-modal causability	×	×	✓	×	91

Ref = Reference, RS = Related Sections, RI = Related Improvements, #CP = number of Citepapers, Tax = Taxonomy, ORC = Open Research Challenges, App = Applications (number of app).

In the nutshell, this tutorial offers a holistic study of the different aspects of XAI criteria with the main aim to assist 1) readers to simply visualize the existing challenges meeting XAI along with the legacy and advanced explainability methods that are proposed to discourse them; 2) AI community, in academia and industry, to partake a vibrant roadmap on how to develop and apply forthcoming solutions according to a group of explicit and definite taxonomy; and 3) newcomers in the AI to effortlessly interpret the main concepts of XAI and to be on the viewpoint for the up-to-date trends in this evolving field (Table 2).

1.3. Survey methodology

This tutorial covers the state-of-the-art studies carefully chosen from papers published between 2000 and 2022 in refereed conferences, journals, and preprints, resulting in a total of about 310 reviewed articles. To the best of the authors' knowledge, this study covered the majority of the reputable studies that tackled the challenges of explainability of AI. The methodology adopted to select these studies include three steps. First, different search engines (i.e., google scholar, IEEE explore, Sciencedirect) are used to search for the XAI-related keywords including "Explainable Artificial Intelligence", "Explainability", and "Interpretable Machine Learning", "Interpretability", and "White-Box Machine Learning". Second, the citations of the selected studies are tracked to assure that we cover the studies that might not be retrieved in the outcomes of the search engines. Third, the overall results are carefully filtered out by the authors to emphasize the most reputable studies having robust contributions.

1.4. Study organization

The rest part of this survey is structured as follows: Section 2 presents some foundations and historical literature on XAI. Then, Section 3 establishes a formal definition for XAI and related concepts. Section 4 provides a fine-grained taxonomy for categorizing explainability methods along with their technical and theoretical analysis. After that, we describe and taxonomize the explanation evaluation methods and metrics, in Section 5. Section 6 explains the opportunities of XAI in different application domains. Further, Section 7 figures out the current research challenges, open topics, and prospects. Section 8 chat out the roadmap of research in responsible AI. Finally, Section 9 concludes the main findings of this work.

Table 2
An overview of open sourceframework for XAI methods.

Frameworks /Tools	Language	Counterfactuals	Data Types	Deployment options	Explicitly Interpretable	Local	Global	Persona-related	Evaluation measures
Alibi [152]	Python	✓	✓	✓	×	✓	✓	×	×
H2O	Python, R	×	✓	×	✓	✓	✓	×	×
Skater [7]	Python	×	×	×	✓	✓	✓	×	×
tf-explain	Python	×	×	×	×	✓	✓	×	×
EthicalML-XAI [3]	Python	×	×	×	×	×	✓	×	×
iNNvestigate		×	×	×	×	×		×	×
Captum [154]	Python	×	✓	×	×	×	✓	×	✓
InterpretML[204]	Python	×	×	×	✓	✓	✓	×	×
DALEX [19]	Python, R	×	×	×	×	✓	✓	×	✓
XAITK	Python	✓	✓	×		✓	✓	×	✓
AIX360 [13,12]	Python	✓	✓	✓	✓	✓	✓	✓	✓

2. Background & foundations

Recently, the field of AI has been continuing to revolutionize our world and adjust the methodology in which the digital world act and behaves. This rapid development in AI offers great academia and industry a wide variety of techniques, each customized to address a particular range of problems. In this context, machine learning was presented as a subfield of AI that can enable designing automated decision-making and analytical solutions by training intelligent algorithms to learn inherent patterns from data. However, with the emergence of big data, ML becomes unable to obtain insightful analytics due to the complexity and high dimensionality of data. As a remedy, deep learning has evolved as a branch of ML interested in developing different categories of complex and deep neural networks (NN) to detect and learn inherent representations from such big-size data [50]. Unfortunately, with the increased complexity of DL solutions dominating the majority of computer science applications, they turned out to be opaque and impossible to understand for human stakeholders or even domain experts. This in turn give the rise to many questions about fairness, accountability, trust, and responsibility.

Explaining how ML/DL algorithms work is the part of having transparency for the reasoning about its outcomes and decisions, which can explain their behavior in human-understandable ways through developing interpretable models, methods, and interfaces. The interpretation needs to provide either transparency of the approach structure and forecasts, a visualization of discrimination rules of the model, hints, message feedback, or chatbots [1]. Several XAI methods tackle the issue of lacking transparency and interpretability in black box AI algorithms. This is because of the complicated internal structure of the models and the lack of offering interpretability, along with high performances. The consideration of the opaque nature of complicated models has obstructed their future applications in generating critical decisions such as industrial control and autonomous cars, which can put the life and health of humans in danger. The reasoning behind predictions of interpretable ML systems gives users explanations for accepting or rejecting forecasts and recommendations [2]. The question of why developing XAI systems is complicated because it crosses the boundary of computer science, psychology, and social science. A collection of academics working under trust and faithful in numerous ML, law, and cognitive science applications [6]. The Defense Advanced Research Projects Agency is a civilian and military research organization financed by the United States Department of Defense (DARPA), and the European Union approved a data protection law [121]. This involves ‘‘the right to an explanation”, in which the GDPR sought to address the issue of possibly biased conclusions being propagated to society as a result of a computer model learning from skewed and imbalanced data.

It is worth mentioning that the GDPR comprises confident rules, however incomplete and vague, which could be stated as a commitment to deliver explanations of the operative of AI system. thus, these rules are still not approved in law because of the existence of such a right in just specific scope. Moreover, the emergence of the draft European Artificial Intelligence Act (AIA) spots the first chief authority to propose definite AI regulation [78,82,271]. Whereas the AIA is effusive around high-risk AI, it is almost silent on any notion of ‘‘XAI”, favoring an emphasis on the less definite concept of ‘‘Trustworthy AI”. Given this AIA, the EU seems to depart from the concept of explainable AI [240]. In all, the European AIC aims to collate the standardization of the European trustworthy AI paradigm, which necessitates the AI to be lawfully, ethically, and strictly robust, while appreciating human rights, democratic values, and regulations. However, Act 13 of the European AIA stated that the design of high-risk AI solutions should be adequately transparent to allow users to interpret the AI decisions. This, in turn, further explain the role of explainability as a driving force to achieve trustworthy AI.

2.1. The drivers for XAI

For industrial advantages, ethical matters, and supervisory factors, XAI is vital when humans are to comprehend, properly trust, and successfully control the outcomes of AI algorithms. According to the investigated literature, the enthusiasm for the

explanation of AI algorithms can originate from at least four motives (See Fig. 1). It is worth noting that although it may appear that there is a commonality between these motives, they portray the various motives for XAI.

2.1.1. Explain to discover

Questioning for explanations is a valuable means to learn inherent information to acquire knowledge about the underlying task. Only explainability can provide a potent tool for verifying and obtaining new perceptions into the dilemma at hand to lead to a more reliable AI solution. Thus, it is expected that, in the future, XAI algorithms can help domain experts to discover unseen theories and laws in chemistry, physics, biology, medicine, and nanoscience. The knowledge gained by XAI can be further leveraged to understand the behavior of AI algorithms under different learning strategies, data formats, architectural designs, parameters, etc.

2.1.1.1. Explain to justify. The literature has been witnessing manifold debates regarding AI solutions generating subjective or unfair outcomes. This indicates a growing demand for explanations to make certain that the decisions made by AI algorithms are not flawed. When it comes to explaining the AI-driven results, it typically implies the necessity for motives or rationalizations for that specific conclusion, instead of describing the internal mechanisms or the thinking logic at the back of the decision-making method. In this context, XAI can be leveraged to offer the essential information to excuse outcomes, especially in case of unanticipated decisions. It also guarantees that there is an auditable and demonstrable practice to support decisions as being reasonable and fair, leading to gaining trust. Moreover, henceforward, AI algorithms require to afford justifications to be in accordance with statute, i.e., "right to explanation", as a part of the GDPR.

2.1.1.2. Explain to control. Unlike the common of thought, the explanation is not just for justifying the outcomes of AI algorithms. However, it can also assist inhibit entities from going incorrect. Actually, knowing more regarding the conduct of the system at hand offers a superior view over unidentified susceptibilities and faults, and facilitates promptly distinguishing and adjusting mistakes in low cruciality circumstances. This knowledge can be leveraged to establish a set of control rules that help improve the management of the AI-based system.

2.1.1.3. Explain to improve. One more motive for explaining AI algorithms is the requirement to constantly enhance their performance. it is broadly identified that XAI algorithms can be more definitely enhanced. Since users understand the reason behind the generation of particular productions, they will also realize the best way to make it brighter. Therefore, XAI can be the basis for continuing advancement between machines and humans. This in turn will pave the way for enhancing learning performance in dynamic and evolving environments.

2.2. 2.2 Tools & frameworks

With the increased importance of XAI every day, the development of a set of tools and frameworks comes to be of great importance for the research community to facilitate implementing and reproducing different explainability methods for different levels of stockholders. Accordingly, this part of the manuscript reports the common open-source frameworks for XAI (See Fig. 2). Table 1 outline and compare different characteristics of existing open-source XAI toolkits.

(1) AIX360[1]: It is an open-source toolkit presented in [12,13] to offers a cohesive, obliging, and simple programming interface and related software architecture to implement different explainability methods needed by a variety of stakeholders. The AIX360 presented a couple of quantitative metrics to act as surrogates of how "excellent" a specific local explanation is.
(2) H2O: It represents another interesting toolkit for explaining AI algorithms, which provides intrinsic local explanations as well as rule-based global explanations.
(3) Skater[2]: It is an integrated open-source python framework to implement model Interpretation for different kinds of ML modes in real-world problems. It allows the local and global demystification of the learned structures of a black box model in either regression or classification tasks.
(4) Tf-explain[3]: A framework for explaining the TensorFlow models using activation maps and gradient maps.
(5) EthicalML-XAI[4]: An ML tool for developing fundamental explainability in AI solutions through data analysis and model analysis.
(6) CaptumAI: An open-source framework for interpretability of PyTorch models [154]. It decomposes the explanation methods into three groups of methods namely primary attribution, layer attribution, and neuron attribution. It can be easily extended with new features and algorithms and can deal with different data modalities.

[1] https://github.com/Trusted-AI/AIX360
[2] https://github.com/oracle/Skater
[3] https://github.com/sicara/tf-explain
[4] https://github.com/EthicalML/xai

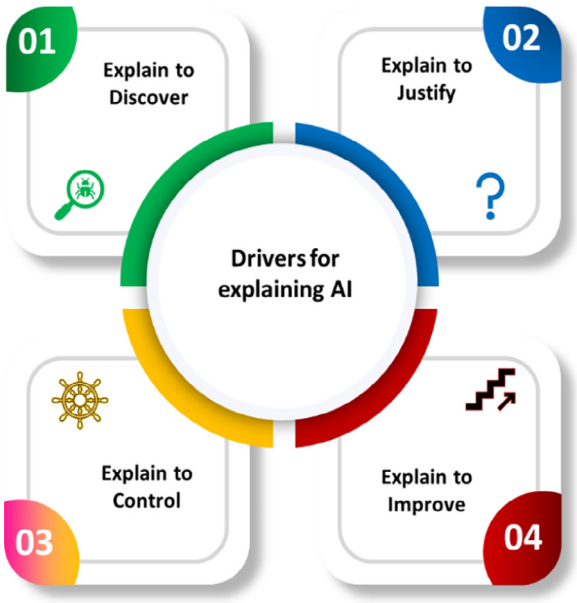

Fig. 1. Drivers of explaining AI.

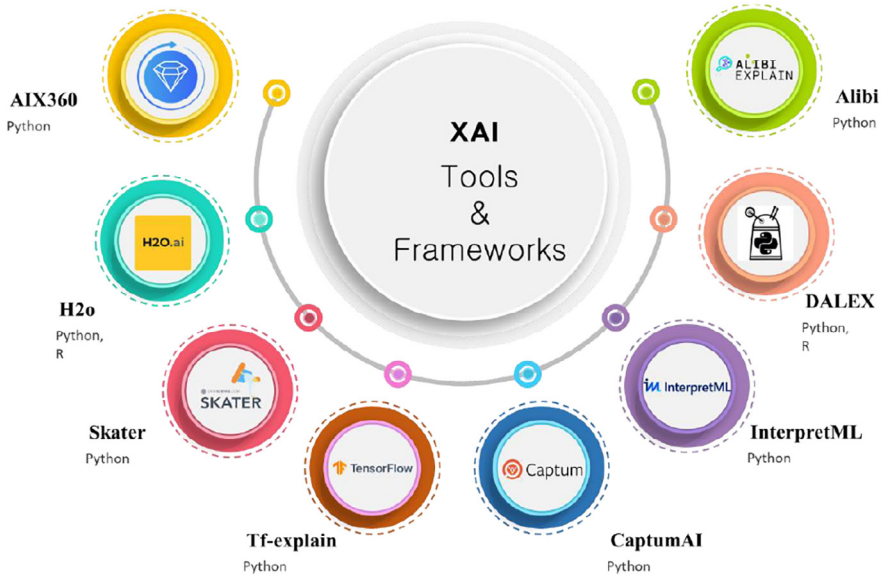

Fig. 2. Common tools and frameworks for explainable artificial intelligence.

(7) InterpretML[5]: It is a free python package [204] that provides two categories of interpretability methods namely glass-box ML models designed for interpretability, as well as Black-Box explainability methods for clarifying present AI solutions.
(8) DALEX[6]: It is presented by [19,26] to enable the development of responsible AI solutions that combine different approaches and fill the current gap splitting Black-Box models from XAI. It offers an interactive model explanation based on a model-agnostic interface and multiple fairness metrics. DALEX is available in two versions one for R and the other for Python programming. It is customized for tabular data and can afford a variety of visualization plots.
(9) Alibi[7]: an open-source Python framework presented in [152] to offer high-quality implementations of local and global XAI methods and is tailored to deal with different categories of data (i.e., tabular, text, and images) under both regression

[5] https://interpret.ml/
[6] https://dalex.drwhy.ai/
[7] https://github.com/SeldonIO/alibi

and classification tasks. It is integrated into deployment platforms Seldon Core, KFServing, and Ray to enable deploying XAI in real production under distributed and centralized scenarios.

(10) The Explainable AI Toolkit (XAITK)[8]**:** An open-source toolkit developed by the DARPA XAI program to assist the users, researchers, technicians, to interpret complex AI solutions. The toolkit joins a searchable repository of autonomous charities and a more cohesive, popular software framework.

(11) What-If Tool[9]: A framework for delivering a simple graphical interface from which it is feasible to understand opaque regression or classification AI model. It allows users to perform inference on a significant number of samples and directly visualize the findings in different ways.

3. Explainable AI: Preliminaries and definitions

The concept of explainability often has one or many drivers and targets that may vary according to the targeted audiences or underlying applications for which the explanations are generated. With the increased community of XAI and related objectives, this section introduced, to the reader, a standard and concise definition of explainability and related concepts in the field of XAI [61].

Definition 1 Black-Box AI. An AI algorithm is declared a **Black-Box** M^b if and only if its construction, internal functions, logic, and parameters are unreachable for humans and hence they are opaque. Thus, opaque AI can be declared as synonymous with Black-Box AI and both terms can be used interchangeably.

Definition 2 White-Box AI. An AI algorithm is declared a **White-Box** M^w if and only if the details of its construction, internal functions, logic, and parameters are easily obtainable and transparent for human practitioners, hence it is transparent for them. Thus, Transparent AI can be considered synonymous with "White-Box" AI and both terms can be used interchangeably.

Given the above definition, it is worth mentioning that transparency can be inherent in the algorithm by its own design or can be achieved by applying some peripheral explainability methods.

Provided the "Black-Box" AI algorithm, the challenge in explaining the Black-Box model resides in offering an explainable model that can simulate the conduct of the black box and also keep comprehensible to humans. Specifically, the explainable model resembling the black box must be generally understandable. As a result, the challenge of explaining the Black-Box model can be formulated as follows:

Definition 3(Model Explainability). Provided a Black-Box algorithm M^b, and dataset $D = \{X, Y\}_i^N$ containing N observations. A model explanation can be formulated as a problem of finding a mapping function $f := M^b, D \rightarrow M^w$ that takes M^b and D as input, and return the White-Box model M^w; where M^w can behave like M^b and come up with an explainability function $f := M^w, D \rightarrow E$ that enable generating a set of human interpretable explanations $E = \{e^1 \cdots e^E\}$.

On the other hand**,** having a Black-Box algorithm, the challenge in explaining decisions of a Black-Box model resides in delivering an interpretable result by designing a technique for justifying the reasons behind the output of a Black-Box model. In particular, it is unnecessary to explain the entire rationality of the Black-Box model but only the justifications for the decision made for a specific instance. Therefore, the dilemma of explaining a Black-Box decision is formally defined as follows:

Definition 4**(Decision Explainability**). Given Black-Box model M^b, and dataset $D = \{X, Y\}_i^N$ containing N observations. The model explanation can be formulated as a problem of finding a mapping function $f := M^b, D \rightarrow M^w$ that takes M^b and D as input, and return the White-Box model M^w; where M^w can behave like M^b but has decision explanation function $f := M^w, X_i \rightarrow e^i$ that provide a justification e^i for the decision made for each inputX_i.

Further, having a Black-Box model, the challenge in Black-Box inspection resides in generating a representation (visual or textual) for interpreting the working methodology Black-Box or the justifications for its decision. Thus, the challenge of inspecting the Black-Box model can be formally defined as follows:

Definition 5**(Model Inspection)**. Provided a Black-Box algorithm M^b, and dataset $D = \{X, Y\}_i^N$ containing N observations. Model inspection can be framed as a problem of finding a function $f := M^b, D \rightarrow V$ that takes M^b and D as input and generate a set of visual or textual representations for explaining the behavior M^b.

Definition 6 **(Explainability)**. Explainability of AI is any kind of activities and processes that can be performed on any part of the ML/DL model to unbox the model's internal affairs as well as justify its generated decisions in easy to comprehend manner.

Definition 7 (**Interpretability)**. It represents the ability of AI solutions to extract the significant sub-symbolic information about learned representations to allow the human observer to understand in what way the inputs are mathematically and statistically charted into outputs.

Definition 8 **(Causality).** Causality is the relationship between cause and effect. It connotes lawlike necessity, whereas probabilities connote exceptionality, doubt, and lack of regularity.There are two compelling reasons for starting with, and in fact stressing, probabilistic analysis of causality; one is fairly strsighforward, the other more subtle[215].

[8] https://xaitk.org/

[9] https://github.com/pair-code/what-if-tool

Defintion 9: (Causability). Causability is the measurable extent to which an explanation - resulting from an explainable AI method - to a human expert achieves a specified level of causal understanding. This can be measured e.g. with the System Causability Scale [119]. Causability can refer to a human model [117,118]. A good understandable example for Explainability and Causability in Medical AI is given in references [120], where a clinical case of lung cancer serves as a practical example.

Definition 10 **(Trustworthiness).** The degree of confidence to which the AI solution will behave as expected when encountering real-world problems. Eexaplainability can be used to achieve the trust in AI applications by enabling users and stakeholders to be more confidently using an AI model they understand.

Definition 11 **(Interactive AI).** Interactivity is a key goal of XAI, and it refers to the characteristics and abilities that make AI solutions easy to deal with, for users and stakeholders.

Definition 12 (**Stable AI).** An AI solution can be declared stable if and only if it cannot be deceived by some perturbations that often exist in real-world applications.

Definition 13 (**Robust AI).** An AI solution can be declared robust if and only if it can resist intentionally generated perturbations, especially adversarial attacks.

Definition 14 **(Reproducible AI).** An AI solution can be declared as reproducible if and only if it continually makes similar decisions each time, it receives the same data instance. Reproducibility can be considered a major goal of XAI, especially in applied sciences.

Definition 15 **(Bias).** The property of AI solution that makes it weigh, prejudge, prefer, or incline toward subgroups of data over others because of the intrinsic predispositions in human data aggregation and shortcomings in the learning model.

Learning the model performance by making use of XAI methods for a variety of input distributions is expected to enhance the human's perception of the predispositions and skewness in the training data. Thus, XAI can pave the way for a deeper understanding of the input space in such a way that can assist developers and practitioners to design and study innovative bias alleviation approaches and promote brighter AI solutions.

Definition 16 (**fairness**). An AI solution can be declared fair if and only if it has the ability to make unbiased decisions without including any preference for any of the populations represented in the input space.

4. Explainable AI: Taxonomic view

This section presents a new taxonomy for classifying explainability methods into multiple groups and subgroups according to different splitting criteria as shown in Fig. 3.

4.1. Scope

In this dimension, the XAI can be classified into three main (Fig. 4) categories. First, global explanation indicates the ability to grasp the whole decision rationality of the ML/DL model. Second, a local explanation concentrates on explaining the decision for a particular input instance. Third, a semi-local explanation, which emphasizes generating an explanation for a group of similar samples.

A. Global Explanations.

Global explanation aims to provide a global understanding of the whole logic of learning models and intuitively explore learned representations to humans [76]. The global model provides an understanding of the predicted outcome distribution (how does the model reach results), so the global model is very complicated to generate [49]. The global explanation is more refined when used to diagnose the trained model and extract knowledge from the model [304].

B. Local Explanations.

Local explanation techniques include feature importance and LIME [49], where Mollas et al. [198] provide"LioNets" explanations in a local scope that delicate feature importance changes, Botari et al. [32] provide local estimators instance and specific output. Much recent research in the domain of XAI systems widely focused on local explanations and justifications of deep learning systems like the Grad-CAM++ method that provide visual explanations of convolutional neural network (CNN) model predictions through object localization and explaining multiple object instances in a single picture.

C. Semi-Local Explanations.

Wang et al. [290] improve the interpretability of the XAI system, which (Fig. 5) use the SHAP method and combines local and global explanations, where the local explanations provide reasoning ‘‘why approach made an individual prediction on the specific case" and the global model gives crucial model characteristics from the model ‘‘how the model made predictions." Hohman et al. [115] show the TELEGRAM system, which includes the Global Model interface for describing and displaying a feature's overall influence on model decisions and predictions, and the Local Instance interface for describing each feature's contribution to the final forecast.

4.2. 4.2 Methodology

This criterion denotes the fundamental algorithmic theory followed to explain the AI model, where XAI can commonly be classified into two main categories according to the implementation policy.

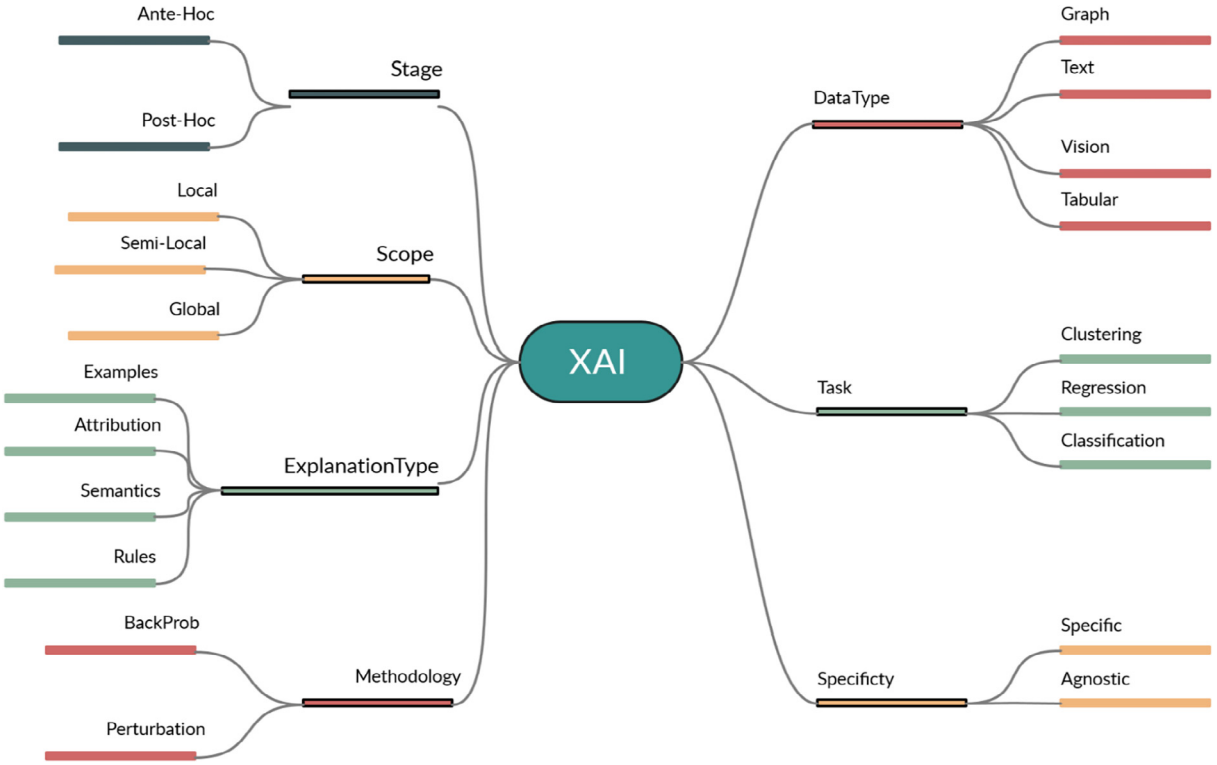

Fig. 3. An illustration for the proposed taxonomy for categorizing the XAI methods based on different classification criteria.

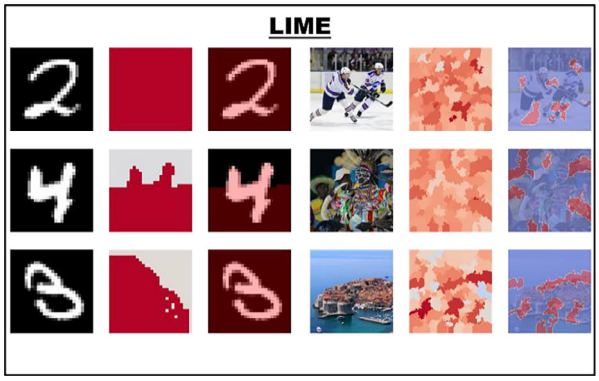

Fig. 4. The explanation results obtained from LIME method on MNIST and Imagenet datasets.

A. Backpropagation-based methods.

Back-propagation-based techniques, on the other hand, compute the score of the input's characteristics at a specific forward and backward run through the model structure. Their results are seldom immediately connected to a difference in the outcome. Back-propagation-based methods are divided into propagation mechanisms and gradient-based methods.

B. Perturbation-based methods.

In perturbation-based methods, the score of an input feature (or collection of features) is directly assessed by deleting, masking, or altering them and performing a forward pass on the new input, comparing the difference with the original output. In the realm of image classification, a perturbation-based approach has been used to CNNs, displaying the likelihood of the appropriate class as a function of the position of a grey patch occluding part of the picture. While perturbation-based techniques allow a direct evaluation of a feature's marginal effect, they are notoriously slow as the number of characteristics to be evaluated rises (up to hours for a single picture) [41].

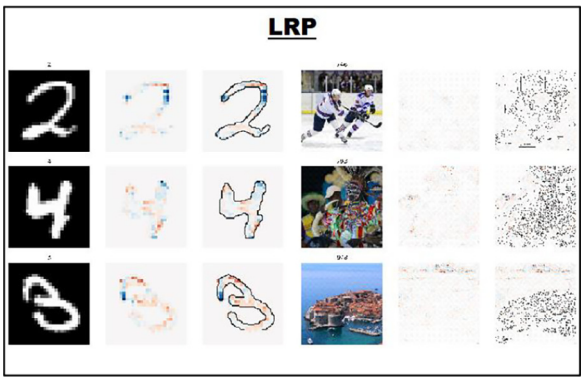

Fig. 5. The explanation results obtained from LRP method on MNIST and Imagenet datasets.

4.3. Stage

This criterion characterizes the stage at which the explanation method is provided. This can be either implanted into the ML/DL model itself (intrinsic or ant-hoc) or applied as an outer algorithm for explanation (post-doc).

A. Ant-hoc methods.

Intrinsic explanations are obtained by building self-explanatory systems with interpretability directly at their inherent structures [76].

Rule-based Models: Models that create and extract rules to characterize researched phenomena are referred to as rule-based learning. To explain their predictions and construct their knowledge [21]. Rule-based systems have the advantage of transparency and comprehensibility. Rule-based methods have been extensively utilized in knowledge-based and expert systems for knowledge discovery and representation [8]. Evolutionary optimization algorithms have been mostly applied to build rule-based systems. Cano et al. [45] developed the interpretable classification rule mining (ICRM) algorithm. Their system provides interpretable and comprehensible IF-THEN classification rules through an efficient evolutionary programming algorithm for solving classification problems. ICRM is designed to maximize the comprehensibility of the classifier by minimizing the count of rules and the count of criteria and conditions per each rule.

Fuzzy Systems: A fuzzy rule can be defined as a conditional statement in the shape of if-then fuzzy rules in which variables are linguistic values that have linear and complimentary membership functions of the fuzzy sets. Fuzzy logic-based approaches have been used mostly to deal with uncertainty and vagueness-defined problems. The initial goal of fuzzy modeling was to create knowledge-based algorithms capable of capturing highly non-linear relationships between inputs and outputs while still providing an understandable picture of such relationships by the use of a just simplified natural language [5,103]. In [189], the authors explained the reasons that make IF-THEN rules invalid to explain the output of Takagi-Sugeno-Kang fuzzy systems (Type-1) and also argued the reason makes the usage of association of the antecedents a valid way to explain the output of such fuzzy systems. they also studied how the careful selection of antecedent membership function is more serious for contemporary XAI and discussed how linguistic approximation and similarity contribute to estimating the quality of the explanations in XAI. In [46], the authors developed a hybrid model that integrates the interval Type-2 fuzzy rough neural network gene expression programming aiming to generate more expressive fuzzy rules through numerous logical operators. the training of their model was formulated as a multi-objective dilemma targeting precision, generalizability, and explainability. the fuzzy rules were minimized and applied to empower the model explainability with little complexity as possible. Moreover, the work [130] demonstrated that the fuzzy Choquet integral (ChI) could be embodied as a NN known as ChIMP enabling optimization of stochastic-gradient-descent under the exponential number of ChI inequality restraints while affording generation of explainable outputs. Many studies emphasized explaining CNNs. For example, in [309], the authors applied some CNN as employs a feature extractor, where the generated feature maps are passed to a fuzzy clustering algorithm, after which the Rocchio's algorithm was applied to categorize the data element. In [170], the authors developed a fuzzy decision tree regularization approach to generate produce decision rules for explaining the respiratory sound analysis built based on an ensemble knowledge distillation method is applied to learn distributed data and attains precise respiratory sound analysis. In [273], the authors defined the factual and counterfactual explanations in the context of fuzzy decision trees, wherever many twigs could be passionate simultaneously demonstrating that factual explanations can comprise many rules rather than a single rule.

Decision Trees: A decision tree is an ML model that distributes data into subsets and predicts outcomes based on decision rules (if-then-else rules). The partitioning of data begins with a binary split and continues until it cannot be split any further. Decision trees are considered the simplest and more interpretable White-Box ML algorithms for the human decision-making process, they can be used to provide helpful visual explanations for their results. Decision trees provide an interpretable graphical representation of the problems being modeled where internal nodes correspond to attribute tests and leaf nodes correspond to predicted outcomes [33].

Regression Models: Regression is a method for determining the functional connection between a dependent variable and a collection of independent variables. Model trees and regression rules are the most expressive models for regression and have a high degree of interpretability. Regression models provide a good reputation in terms of interpretability, but the accuracy is low compared with Black-Box models. Interpretability is achieved by using symbolic regression. However, this comes with the price of reduced accuracy. Logistic regression is known as a classification technique in statistical and ML and it can be treated as a part of a generalized linear model, where it predicts the binary response variables and explain the relationship between predictors and guarantee the reduced subspace with an easy-to-interpret relationship to the functional predictor [58].

K-Nearest Neighbors: Learning by analogy may be described as K-Nearest Neighbors. Models are learned by comparing a given example to a group of comparable training instances using either the Euclidean distance or the cosine similarity. The sample is then categorized according to the class of its nearest neighbor.). *K* denotes the neighbor count considered when constructing the class. In terms of system interpretability, it is critical to emphasize that KNN estimated values are based on the notion of similarity between instances and distance, which may be tailored to the unique scenario at hand. Interpretability of the outcome of each layer is provided by the nearest neighbors whereas KNN is more interpretable by nature because the nearest neighbors themselves have explanations that are readily understood by humans because they lie in the input domain [142].

Bayesian Models: The Bayesian Rule List (BRL) generates a posterior distribution across if-then rules permutations. Starting with a vast, pre-mined collection of potential rules, the list gives risk factor explanations that may be utilized by both human experts and ML. A novel sparsity-inducing hierarchical prior structure overrules permutations resulting in interpretability. The prior favors a shortlist of judgments with a small number of total rules, and the rules on the left side contain few words. BRL results in models inspired by traditional human-built decision-making algorithms and, thus, identical to them [307].

Attention Methods: It is worth noting that there is a recent trend in literature studies for an intrinsic explanation of DL models using either attention layers or attention visualization, which can be added to include an extra explainability operation to the basic learning operation, then mutually train to optimize both operations iteratively [279,320].

B. Post-hoc Methods.

Instead of describing the whole system's behavior, post-hoc systems attempt to offer local models for a single prediction and enable it repeatable [117]. LIME [237], for example, utilizes an intrinsic model like a linear model for approximating the opaque approach [65]. Because algorithmic transparency is thought to be unattainable for these systems, post-hoc explanations have gained popularity as a method in deep learning applications. Post-hoc methods are generally classified as either (1) model-specific or (2) model-agnostic. model-specific techniques are confined to specific model classes because each method is dependent on the internals of a given model. Model-specific methods embed interpretability restrictions into the fundamental structure and learning processes of algorithm models. Model-agnostic approaches, on the other hand, may be applied to any AI model and are utilized after the training has been completed: (1) post-hoc and (2) post-model [14], where model-agnostic utilize the black box models' inputs and predictions to create explanations.

Visualizations: Visualizations provide an effective way of increasing the transparency of AI systems that focus on specific points within the workflow of AI systems [294]. Visualization was the most common approach for post-hoc explanation in our review study. This is attributed to either because the underlying AI systems focus on image analysis (i.e., computer vision), or because they use visualization techniques to convey large amounts of information in a small amount of space. In this subsection, we focus on the most popular visualization approaches used in the literature, especially for visualizing opaque models (i.e., DNN). Gradient-weighted Class Activation Mapping (Grad-CAM) is a common approach for visualizing the outcomes of deep learning models. Grad-CAM visualizes the input regions that become crucial for forecasts from convolutional models, in which it utilizes class-specific gradient flow of information to the arrival at the conclusion layer of CNN to develop a crude localization map of the significant areas in the picture.

Feature Influence (Relevance) Methods: Feature relevance explanation methods, also known as feature-level interpretations, feature attributions, or salience maps, clear the internal workings of the approach by generating a relevance score to its variables for post-hoc explainability. These scores quantify how sensitive a feature is to the model's output. A scores comparison of several variables reveals the weight given by the approach to each of these variables while creating its outcome [21]. Perturbation-based methods and backpropagation-based techniques are the two types of feature-relevance approaches. In perturbation-based methods, the score of input features is directly assessed by deleting, masking, or altering them and performing a forward pass on the new input, comparing the difference with the original output. They also allow a direct evaluation of a feature's marginal effect; they are notoriously slow as the number of characteristics to be evaluated rises. Backpropagation-based techniques, compute the score of the input's characteristics at a specific forward and backward run through the model structure.

Feature Importance Explanations: One of the XAI approaches that may offer the value of each feature for a prediction is a feature importance-based explanation. Given the local linear model $f_z(x) = w_0 + w_1x_1 + w_2x_2 + \cdots + w_dx_d$ and an instance $z = [z_1, \cdots, z_d]$, the importance of each feature z_i can be calculated as $\mathrm{abs}(w_i * z_i)$, and provided to the user as an explanation of which features the original model deems as relevant for its decision on z. For example, it interprets the DNN findings to the rank order of the variables according to how essential they are for executing the classification task in distinguishing between positive and negative samples. LIME [237] is one of the major important and popular techniques

for providing local explanations. LIME is a model-agnostic methodology that gives a human interpretable representation of an explanation for the model's output. It provides a local model explanation by linearly approaching the ML model's decision boundary in the vicinity of the test sample. LIME attempts to explain why observation is made for a certain prediction and what the prediction supports, and conflicts are. It is compatible with any algorithms for supervised learning. The general principle is to develop a surrogate model through interpretable representations. Fan and Xiao [83] developed a methodology based on the conceptual framework which presents the general steps of LIME as follows: (1) the opaque model generates a collection of permutated samples and uses them to obtain predictions; (2) its proximity to permutated samples are measured for a given observation and used as weights to represent the relative value of each permutated sample; (3) the permutated data are converted into interpretable representations, e.g. transforming numeric values into categorical values; (4) Interpretable representations are recorded and used to show the locally stable relationship with the effects of the prediction; and (5) Explanations are obtained by interpreting the established local surrogate model, such as linear regression model coefficients.

Gradient-based Explanations: Gradient-based explanation methods handle explanation by analyzing the contribution of each neuron in the input space of a neural network to exploit the gradient of the latent nodes about the inserted data features by calculating the corresponding gradients. This approach considers neural networks as a function and relies on their gradients to generate explanations. These explanations are computed as the product of gradients and input activation.

Sensitivity Analysis: Simonyan et al. [266] consider two visualization techniques that discuss the visualization of image classification models learned using CNN based on calculating the gradient of the class rating in relation to the given picture. The first one produces an image captured by a CNN that maximizes the class ranking, thus visualizing the notion of the class. The second includes computing a class saliency map, unique to a given image and class where Sensitivity Analysis (SA) describes an outcome based on the locally assessed gradient of the model (partial derivative). SA measures the local variation of the function along each input dimension, not the function value itself which represents and measures the relationship between the change of one input variable within a range that can influence the output while controlling other input variables constant, So the contributions of the input variable can be assessed by calculating the change of output [65]. The most relevant input characteristics are assumed to be the most sensitive to the gradient of the model's output in a sensitivity analysis. Sensitivity analysis is widely used in explaining DNN models.

Counterfactual Explanations: This explanation method is a type of example-based explanation, Counterfactual explanations are contrastive explanations that illustrate contextually important model outcomes [194]. They describe the bare minimum of circumstances that might have resulted in a different conclusion. The reviewed literature advocates a group of features to be contented to guarantee to attain a high-quality counterfactual:

Variability: a feature early presented [139,201], and refers to the discovery of the adjoining points of a particular example depending on some distance function, which could result in very analogous counterfactual candidates exhibiting minor variances between them. Variability was presented as the method of engendering a group of different counterfactual explanations for given data observation x. This way, the generated explanations are more explainable and more comprehensible to humans.

Feasibility: a feature early presented in [225] to give a response to the disagreement that discovering the closest counterfactual to a data observation did not essentially result in a viable variation in the structures. Given y as the closest counterfactual to the data instance x, and lies in the decision plan of the Black-Box classifier, hence, it is not so confident about its class, which might result in unfair counterfactual explanations. As a remedy, the authors of [56] debated that counterfactual is an improved one for the reason that it lies in a definite area of the decision plane and also resembles the element that exhibits the shortest path tox. This way, it is possible to generate human-interpretable counterfactuals with the least possible feature changes.

Proximity: it refers to computing the counterfactual distance from the input data element during the generation of a counterfactual explanation, [280]. In this regard, numerous dissimilar distance functions (i.e., $L_0 - norm$, $L_1 - norm$, $L_2 - norm$, $L_\infty - norm$) could be adopted to estimate proximity leading to counterfactual explanations having various characteristics. In some literature studies, other forms of distances are taken into account to quantify the proximity including cosine similarity [144], Nearest Neighbor Search, and ranking methods [61].

Plausibility: A feature that accentuates that the created counterfactuals must real, while the searching procedure must guarantee rationally practical consequences. Sometimes referred to as Reasonability and Actionability [280]. The definition of plausibility implies that a necessary counterfactual must never alter the absolute attributes of data samples. When it comes to explaining a counterfactual, it is not possible to have elucidations saying "if the gender is man, then the loan will be granted", as this demonstrates intrinsic unfairness in the generated explanation. Variable attributes, like salary, cloud be changed instead to find good counterfactuals.

Sparsity: a feature that pertains to the approaches adopted to professionally bargain the least features that require to be transformed to attain a counterfactual [141]. In intellectual fields, counterfactuals were adopted for an operation of envisaging a theoretical situation opposing an actual incident that had taken place and thinking about its outcomes [217]. Sparsity is the favorite property of counterfactuals such that the least conceivable variations in their attributes result in more operative, human-comprehensible, and explainable counterfactuals. The work [201] demonstrated that sparsity is evaluating the number of attributes a human may require to alter to change over to the counterfactual class. In contrast, the work [280] debated that sparsity could be perceived as a trade-off between the count of attributes and the overall quantity of alteration

completed to attain the counterfactual. Similarly, the work [284] follows the same notion and declared that tracking the "closest possible world", or the minimum alteration to the world that could be applied to achieve the wanted consequence.

Knowledge Extraction: Knowledge extraction must be expressed in a way that encourages inference, where the extraction task is to find the links and relationships between the concepts of input and output in a trained network, in the sense that a specific output is induced by certain inputs. Knowledge extraction applies the most popular XAI methods on an ML model to extract knowledge that is stored within a rule-based system, in which knowledge is represented in symbolic form. Several works suggest techniques to extract the knowledge, which rely primarily on two methods: Rule Extraction and Knowledge Graph.

A. Rule Extraction.

The goal of rule extraction is to identify rules that define approaches and are intelligible to people, in the shape of (If Then Else) statements, in order to replicate Black-Box models. Producing decision trees is one of the most frequent ways of extracting rules from non-rule-based classifiers.

B. Knowledge Graph.

Their underlying semantic graph ontologies are used in the context of symbolic AI systems, which provide a solution for the explainability including structural or sequential relations such as social relationships, and provide knowledge representation to evaluate reasoning and estimation techniques [90]. The knowledge graph approach converts different types of data to a uniform graph format that can be used as an integrated knowledge base. Knowledge Graph also offers a strong information-sharing approach which helps robots share their awareness for improved prediction [52]. Also, knowledge graphs can expose a human-like explanation through the recognition of an object of any class in a knowledge graph representation.

Dialogue Models: The dialogue system of explanation was first introduced by Walton [286] to provide the automated justification and explanation of decisions made by ML systems, through a natural and human-oriented interface. Users are then asked to offer comments to the system for enhancing its explanation service Like Chat-bot can provide explanations and corrective fine-grained feedback messages. Causal chains are relevant in causality and interpretation, which causal chain is a path of causes between a series of events in which a cause means that C must occur before E, from event C to event E, in which any event without cause are root causes and causal explanations take the form of conversation where the explainer presents an explanation to an explained in a conversation.

Visual Question Answering (VQA) is considered an instance of a difficult job that necessarily requires reasoning and understanding of explanatory explanations. The aim of the VQA scene is to provide an accurate response to the subject, where the existence of these concerns may be about the relationships, or lack of specific items in the image, amongst these items, or the probable output of the execution of specific acts in the scene on objects [241]. VQA framework takes an image and a freeform, open-ended, as an input, natural language question about the picture in natural language and generates as the output, the natural language response.

4.4. Specificity

This criterion provides another way for classifying the explainability methods to be model-specific or model-agnostic, where the former category is applicable to all ML/DL algorithms, and the latter one can be applied only to a specific AI algorithm. Model-specific explainability methods are restricted to specialized model classes which imply limited that the explainability surface is restricted to the generating model that usually emphasizes improving its predictive modeling performance. Consequently, research interest is always attracted to investigate and use model-agnostic methods. On the other hand, model-agnostic methods are independent of the underlying ML/DL model which makes them easy to apply, develop, and scale.

4.5. Data type

This criterion categorizes the explanation methods according to the type of data on which the AI algorithm is performed. This includes four data formats namely vision data (i.e., video, image, etc.), tabular data, textual data, and graph data. In the following subsections, we overview and outline different explainability methods applicable for ML/DL trained on different data formats.

4.5.1. Explainability in vision data

In this section, we review and discuss XAI solutions for decisions made by AI systems behaving on visual data. following the proposed taxonomy, this section considers four categories of exploitability methods including Concept Attribution (CA), Counterfactuals, Prototypes, and Saliency Maps. In this regard, the research studies belonging to each category are analytically overviewed and contrasted based on different aspects as shown in Table 1 for experimental analysis, three popular datasets are considered as case studies for image recognition tasks namely MNIST, Stanford Dogs, and ImageNet data. These case studies are carefully selected due to their broad usage, the inclusion of different categories of classes, and variety in data distribution. In these case studies, a customized version of the VGG classifier is trained and finetuned to the data and then used for explainability analysis. Table 3 summarize and overview the XAI methods for vision data.

A. Saliency Map.

Table 3
An overview of explainability methods for vision-based AI solution.

Class	Method	YEAR	Ref.	Data Type	AH / PH	G / L	A / S
SM	SHAP	2007	[181]	ANY	PH	L	A
	LIME	2016	[237]	ANY	PH	L	A
	$\in$-LRP	2015	[16]	ANY	PH	L	S
	INTGRAD	2017	[276]	ANY	PH	L	S
	DEEPLIFT	2017	[265]	ANY	PH	L	S
	SMOOTHGRAD	2017	[269]	IMG	PH	L	S
	XRAI	2019	[138]	ANY	PH	L	S
	GRADCAM	2017	[255]	IMG	PH	L	S
	GRADCAM++	2018	[53]	IMG	PH	L	S
	CAM	2016	[328]	IMG	PH	L	S
	IS-CAM	2020	[203]	IMG	PH	L	S
	Score-CAM	2020	[288]	IMG	PH	L	S
	SSCAM	2020	[287]	IMG	PH	L	S
	LayerCAM	2021	[132]	IMG	PH	L	S
	XGrad-CAM [87]	2020	[88]	IMG	PH	L	S
	Smooth Grad-CAM pp	2019	[210]	IMG	PH	L	S
	RISE	2018	[219]	IMG	PH	L	S
CA	TCAV	2018	[146]	IMG	PH	L	A
	ACE	2019	[94]	IMG	PH	G	A
	CONCEPTSHAP	2020	[310]	IMG	PH	G	A
	CACE	2019	[96]	IMG	AH	G	A
CF	CEM	2018	[73]	IMG	PH	L	A
	ABELE	2020	[98]	IMG	PH	L	A
	L2X	2018	[56]	ANY	PH	L	A
	GUIDED PROTO	2019	[180]	IMG	PH	L	A
PR	MMD-CRITIC	2016	[145]	ANY	IN	G	A
	–	2017	[153]	ANY	PH	L	A
	PROTONET	2019	[55]	IMG	AH	G	S

L = Local; G = Global; A = Agnsotic; S = specific, AH = ant-hoc; PH = post-hoc, IMG = IMAGE.

Saliency maps (SMs) were defined as an image, where the brightness of each pixel (i,j) characterizes the corresponding saliency degree $s_{i,j}$. Officially, the SM is shaped with a matrix S with dimensions that are equal to the dimensions of the input image upon which the explanation is required. Increasing the value of $s_{i,j}$ imply increasing the equivalent pixels' saliency. The SMs can be visualized by leveraging different color maps. For instance, the red color implies the positive contribution of pixel (i,j) to the decision of the AI algorithm, blue color implies a negative contribution of pixel (i,j) to the decision of AI. The SMs can be created using two strategies. First, assigning a saliency degree $s_{i,j}$ to every pixel. Second, segmentation of mage into multiple pixel parties and then allocate a saliency degree $s_{i,j}$ for each party.

Local Interpretable Model-agnostic Explanations (LIME) could be exploited to recover SM for vision-based classifiers. For images, segmentation is applied to perform perturbation. In other words, LIME separates the input into pieces termed superpixels. So, it generates the vicinity by arbitrarily replacing the super-pixels with a regular, probably unbiased, color. This vicinity is later passed to the black box, where a sparse linear network is trained ahead. The localization of image or frame element division is important to attain a fine interpretation. In the case of a limited receptive field, the LIME's localization is ineffective at the outside of the box, leading the model to choose the whole image like a super-pixel. To acquire a reasonable outcome, the individual should adjust the localization factors. Recent literature has been exerting many efforts in improving and customizing LIME explanations for different computer vision tasks. Fig 0.4 shows the LIME-generated explanation for samples from MNIST and imagenet datasets.

Epsilon-Layer-wise Relevance Propagation ($\epsilon-$ LRP) [17] was identified as a model-specific explainability technique for generating local post-hoc explanations for every data format. $\epsilon-$LRP uses decomposition to explain and justify the decisions of the visual classifier. The $\epsilon-$LRP restructuring procedure is launched for FNN [12]. Scientifically, it reorganizes the projection y back utilizing local reallocation regulations till each pixel (i,j) is allocated a relevant grade G_i. Given a_i as the activations of $i-th$ neuron per layer l, G_i as the relevance grades related to the $j-th$ neuron in the next layer (i.e., $l+1$) and $w_{i,j}$ denoting the weight between the $i-th$ and $j-th$ neuron. The standard $\epsilon-$LRP redistribution of relevance from $l+1$ layer to l layer could be defined as:

$$G_i = \sum_j \frac{a_i \cdot w_{i,j}}{\sum_i a_i \cdot w_{i,j} + \epsilon} G_j, \tag{1}$$

where ϵ denotes a small equilibrium factor to avoid division by zero. Instinctively, this formula reallocates relevance proportionately from $l+1$ layer to l to every neuron in l layer depending on the network weights. Hence, the decisive explanation denotes the input layer's relevance. An extended version of $\epsilon-$LRP is presented in [81] which creates a spectral grouping over the local instance-founded $\epsilon-$LRP. In a similar way, the authors of [83] started applying $\epsilon-$LRP to the data samples,

and then the ϵ – LRP designation is obtained for specific data of concern. Fig 0.5 shows the ϵ – LRP generated explanation for samples from MNIST and imagenet datasets.

Integrated Gradient (INTEGRAD) [116], belong to model-specific techniques and is designed to generate local post-hoc explanations for the different modality of visual data by exploiting the model's gradient information as well as the sensitivity methods such as ϵ – LRP. So, it is mostly utilized with deferential AI algorithms. Officially, INTGRAD builds a route from the reference input $\tilde{x}$ to x and calculates the gradients of points through this route. In the case of images, the points are brought by intersecting $\tilde{x}$ and x and steadily changing the opaqueness of x, and the integrated gradients are acquired by gathering the gradient information. Mathematically, the integrated gradient over the $i-th$ dimension of the reference $\tilde{x}$ input x is as follows:

$$i-th\,gradient = \frac{\partial B(x)}{x_i} \tag{2}$$

The calculation of relevant scores is defined as follows:

$$e_i(x) = \left(x_i - \tilde{x}_i\right) \int_{\propto=0}^{1} \frac{\partial B\left(\tilde{x} + \left(\left(x - \tilde{x}\right)\right)\right)}{\partial x_i} .d \propto, \tag{3}$$

It is worth mentioning that the random selection of baselines might trigger some issues in the case of using a black image as a baseline, as it enables the reduction of the prominence of blackened pixels in the received input image/frame. This dilemma is because of the diversity among pixel values and their reference $\left(x_i - \tilde{x}_i\right)$ appear in the integral calculation. As a possible remedy, the authors of [46] presented Expected Gradients to take the average INTGRAD over multiple references.

Class Activation Mapping (CAM) [328]**:** visual explanation technique was designated for image recognition by capturing the output of the global average pooling layer usually attached at the tail of the feature extraction network. The generated CAM for a particular class is a direct weighted integration of feature maps of the model's layers, which represents the discriminatory part of the image or video exploited by the deep learning model to distinguish that class. In particular, this method exploits the pooling operation that precedes the final decision layer (e.g., SoftMax for categorization) to localize the significance of the input regions by mapping backward the final layer's parameters across convolutional maps of each layer in the model. Hence, the calculation of the localization map is formulated as:

$$L^{(c)}_{CAM}(x, y) = RELU\left(\sum_n w^{(c)}_n \sum_{x,y} f_n(x,y)\right), \tag{4}$$

where in $f_n(x,y)$ is the output of unit n at the network's decision layer with respect to the coordinate (x,y). $w^{(c)}_n$ is the weight equivalent to class c for unit n.

Score-CAM [288]**:** it is a gradient-free visual explanation technique that alleviates the reliance on gradients by acquiring the weight of each activation map out of its forward propagation score on a particular class, thereby the ultimate outcome can be calculated as a linear combination of saliency maps as well as weights. Hence, the localization map is computed as follows:

$$L^{(c)}_{Score-CAM}(x, y) = ReLU\left(\sum_n w^{(c)}_n f_n(x,y)\right), \tag{5}$$

whereas the factor $w^{(c)}_n$ is computed as:

$$w^{(c)}_n = softmax\left(Y^{(c)}(M_n) - Y^{(c)}(X_b)\right), \tag{6}$$

where $Y^{(c)}(X)$ denote the probability that output of the model belongs to given class c, X_b denote the baseline image, and M_n denote the normalization process and is specified as:

$$M_k = \frac{U(f_n) - \min_m U(f_m)}{\max_m U(f_m) - \min_m U(f_m)} \odot X, \tag{7}$$

where $\odot$ denotes the Hadamard product and U represents the up-sampling process.

Smoothed Score-CAM (SS-CAM) [287]: it extends the Score-CAM to improve the sharpness of visual explanation to generate post-hoc concentrated localization of target regions and features in input using two smoothing procedures namely input smoothing and activation map smoothing. Hence, the saliency map is calculated as follows:

$$L^{(c)}_{SS-CAM}(x, y) = ReLU\left(\sum_n w^{(c)}_n f_n(x,y)\right), \tag{8}$$

$$w_n^{(c)} = \frac{1}{N}\sum_{1}^{N} Softmax\Big(Y^{(c)}(M_n) - Y^{(c)}(X_b)\Big), \tag{9}$$

where M_n represent then normalization operation defined eq ().

Integrated Score-CAM (IS-CAM) [203]**:** it is an axiomatic-based visual explanation approach based on the combination of INTGRAD and score-CAM to generate sharper attribution maps, which can be defined as follows:

$$L_{ISS-CAM}^{(c)}(x, y) = ReLU\left(\sum_k w_k^{(c)} f_k(x, y)\right), \tag{10}$$

$$w_n^{(c)} = \sum_{i=1}^{N} \frac{i}{N} Softmax\Big(Y^{(c)}(M_n) - Y^{(c)}(X_b)\Big), \tag{11}$$

DEEPLIFT [111], is defined as data-agnostic and model-specific methods for generating local post-hoc explanations by calculating SMs backwardly as with ϵ – LRP, however, it utilizes a reference similar to the INTGRAD. Rather than using the gradients, DEEPLIFT utilizes the slope to explain in what way the outcome $y = B(x)$ adjusts when the input x varies from a reference $\tilde{x}$. In the same way as ϵ – LRP, a provenance score r is given to each neuron i in the model getting backward from the outcome y. This provenance characterizes the comparative impact of the neuron stimulated at the source input x matched with the activation at reference $\tilde{x}$. DEEPLIFT calculates the initial estimates of the final layer L based on the variation between the generated outcome from particular input and reference $\tilde{x}$. Next, it utilizes a recursive function to calculate the provenance scores of layers l based on the provenances of the next layer (i.e., l + 1). This can be formulated as follow:

$$r_i^l = \sum_j \frac{a_{j,i} - \tilde{a}_{j,i}}{\sum_i a_{j,i} - \sum_i \tilde{a}_{j,i}} r_j^{l+1}, \tag{12}$$

$$a_{i,j} = w_{j,i}^{(l+1,l)} \cdot x_i^{(l)}, \tag{13}$$

$$\tilde{a}_{j,i} = w_{j,i}^{(l+1,l)} \cdot \tilde{x}_i^{(l)}, \tag{14}$$

where $w_{j,i}^{(l+1,l)}$ denote the weights between l and $l+1$, layers and a represent the neural activations. Similar to INTGRAD, selecting a reference is not a simple task and could entail domain specialists. Fig. 6 shows the DeepLift generated explanation for samples from MNIST and IMAGENET datasets.

SMOOTHGRAD [113] is another model-specific method for generating post-hoc explanations, where SM has a tendency to be noisy and fuzzy. SMOOTHGRAD struggles to address this dilemma by ironing out the disruptiveness per the generated localization map. Generally, a saliency map is designed explicitly using the gradient of the outcomes of the model. Gaussian noise kernel is leveraged to augment this practice by ironing the gradient information. Specifically, it receives an input x, injects Gaussian noise to it, and recovers the localization map for each disturbed input by averaging the gradient information across layers. Strictly, provided a s as SM generated by saliency method $f(x)$, the relevant smoothed variant $\tilde{f}$ could be articulated as follow:

$$\tilde{f} = \frac{1}{n}\sum_{j}^{n} f\left(x + \mathcal{N}(0, \sigma^2)\right), \tag{15}$$

where n denotes the total of visual instances and $\mathcal{N}(0, \sigma^2)$ represent the Gaussian disturbance. Agreeing with [4,5], the SMOOTHGRAD still suffer from some limitations: human usually have a tendency to assess SMs based on what they are anticipated to get. For instance, in a dog image, one is expected to look at the appearance of a dog. Nevertheless, this did not imply that the network has the same vision. Fig. 7 shows up this dilemma by comparatively visualizing the SMs by taking the gradient of the input, output, and SMOOTHGRAD. It is notable that the SMs absolutely altered their comportment, shifting toward the subject.

Gradient-weighted Class Activation Mapping (Grad-CAM) [254] is a common approach for visualizing the outcomes of deep learning models. Grad-CAM visualizes the input regions that become crucial for forecasts from.

Convolutional Neural Network (CNN)-based models, in which it utilizes class-specific gradient flow of information into the arrive at the conclusion layer of CNN to develop a crude localization map of the significant areas in the picture. Grad-CAM is computed by.

$$L_{Grad-CAM}^{(c)}(x, y) = ReLU\left(\sum_n w_n^{(c)} f_n(x, y)\right), \tag{16}$$

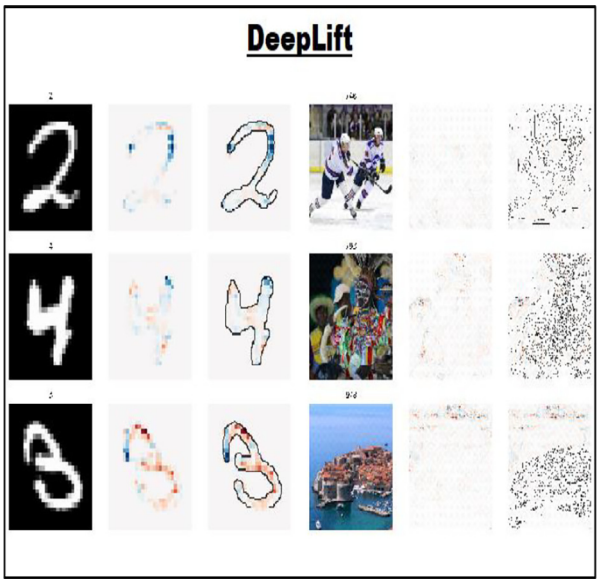

Fig. 6. The explanation results obtained from DeepLift method on MNIST and Imagenet datasets.

$$w_k^{(c)} = \frac{1}{H \,.\, w} \sum_{i=1}^{H} \sum_{j=1}^{W} \frac{\partial Y^{(c)}}{\partial f_n(i,j)}, \tag{17}$$

On the other hand, Grad-CAM++[53] was proposed as an extension to Grad-CAM to enable improved localization of items and visually explain the incidences of various entities of a class per one image. It calculates a subjective combination of the positive partial derivatives of the final feature maps pertaining to a certain class probability as weights to engender a visual clarification for the underlying class label.

$$L^{(c)}_{Grad-CAM++}(x,y) = ReLU\left(\sum_n w_n^{(c)} f_n(x,y)\right), \tag{18}$$

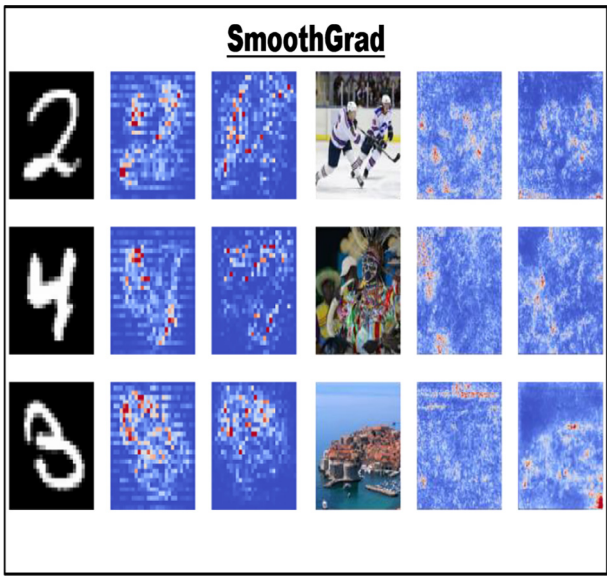

Fig. 7. The explanation results obtained from SmoothGrad method on MNIST and Imagenet datasets.

$$w_k^{(c)} = \sum_{i=1}^{H}\sum_{j=1}^{W} \alpha_n^{(c)}(i,j).ReLU\left(\frac{\partial Y^{(c)}}{\partial f_n(i,j)}\right), \tag{19}$$

where $\alpha_n^{(c)}(i,j)$ is defined as follows:

$$\alpha_n^{(c)}(i,j) = \frac{\frac{\partial^2 Y^{(c)}}{(\partial f_n(i,j))^2}}{2\cdot\frac{\partial^2 Y^{(c)}}{(\partial f_n(i,j))^2} + \sum_{a,b} f_n(a,b)\cdot\frac{\partial^3 Y^{(c)}}{(\partial f_n(i,j))^3}}, \tag{20}$$

$if\ \frac{\partial Y^{(c)}}{\partial A_k(i,j)} = 1 else 0,$

In a similar way, different versions and extensions of Grad-CAM have been developed to improve the quality of the generated visual explanations via smoothing and sharping operations as in Smooth Grad-CAM++[210], XGrad-CAM [88], and LayerCAM [132]. Fig. 8, Fig. 9, and Fig. 10 respectively show the explanation generated by XRAI, Grad-CAM, and RISE for samples from MNIST, and IMAGENET datasets amongst them XRAI shows the worst quality explanation. To further assess the quality of different explanations, we visually compare the non-interactive explanation generated by different methods on Stanford datasets as shown in Fig. 11. The interactive explanation obtained from the SHAP method is separately illustrated in Fig. 12.

To further analyze the performance of explainability techniques, we calculate the insertion and deletion metric, in which the pixels of an image are substituted to compute significance grades provided with the explainability. In case of insertion, the image gets blurred and then inserts the pixel value while replacing the black pixels for deletion. For each replacement, our black box queries the image to evaluate its correctness, and the final performance is computed with the area under the curve (AUC) as a function of the proportion of eliminated pixels. The results presented in Table 4 show the AUC for both insertion and deletion operations with respect to different explainability methods.

Meanwhile in classification and diagnostic tests, ROC and the area under the ROC curve (AUC) describe how an adjustable threshold causes changes in two types of errors: false positives and false negatives. However, the ROC curve and AUC are only partially meaningful when used with unbalanced data. Only part of the ROC curve and AUC are informative however when they are emplyed in imbalanced data. Hence, alternatives to the AUC (the partial AUC and the area under the precision-recall curve) have been put forward. But, these alternatives cannot be as fully interpreted as the AUC, mainly because they ignore some information about actual negatives. In [47] Carrington et al. presents a novel concordant partial AUC and partial c statistic for ROC data, which can be used for foundational measures and methods to explain parts of the ROC plot and AUC.

4.5.2. Explainability in tabular and textual data

This section emphasizes reviewing the methods and techniques for explaining the decisions made by AI algorithms trained on tabular data. In this context, this study taxonomizes the explanation methods into different categories including Features importance (FI), Rules, Prototype, and Counterfactual. Table 5 summarizes the XAI methods for tabular data.

A. Feature Importance.

Based on the techniques of the local explanation, the importance of features is considered the most common kind of expiations. In the results and analysis of prediction, the importance of features takes value by the explainer to decide how many features are important. There are many parameters in the calculations of the importance of the feature. a is a record, $x(a)$ is a model of a black box, $p(.)$ is an explainer, $v = \{v_1, v_1, v_1, \cdots\cdots\cdots .. v_y\}$, where the importance value is $v_l \in v$ of l^{th} the feature made by $x(a)$ to make a decision. the sign and magnitude are vital factors in deciding the value importance of features v_y. the sign factor denotes the importance of features based on Eq. (21):

$$\begin{cases} if\ v_l > 0, negativeimportancecontribution \\ if\ v_l < 0, nepositiveimportancecontribution \\ if\ v_l = 0, noimportancecontribution \end{cases} \tag{21}$$

The magnitude characterizes how excellent the influence of the feature is on the ultimate forecast u. An example $v = \{tall = 0, age = -0.2, fitness = 0.6\}, u = play$, these are valued according to the explanation feature. Fitness is the most positive important, age has small negative importance, and tall has no affecting on the result.

LIME is an explainability method for generating explanations in the form of feature significance vectors. Lime's idea stems from the fact that explanation can be designed nearby from records shaped at random in the vicinity of the case to be explained. There are two cases of samples to extract the locality, low weight, and high weight was closed to the maximum parameter value and abusing $\aleph_a$. The a and x refer to the case to be explained and the black box respectively. The LIME described the behavior of local x using $\aleph_a$ by drawing weighted samples. By random uniform, cases of samples nearly x by drawing elements grater of less than zero. This provides LIME an altered sample of cases $\{s \in \mathbb{R}^f\}$, which it can feed into the model x to get $x(s)$. On the altered data, explainable algorithm $m(.)$ is trained with these samples as a sparse linear solution. The linear model's weights make up the local feature importance explanation. A lot of studies look at how to get over LIME's constraints by presenting several alternatives. DLIME [318] is a deterministic variation in which agglomerative hierarchical clustering is used to choose neighbors from the training data. ILIME [81] uses weighted instances to construct the

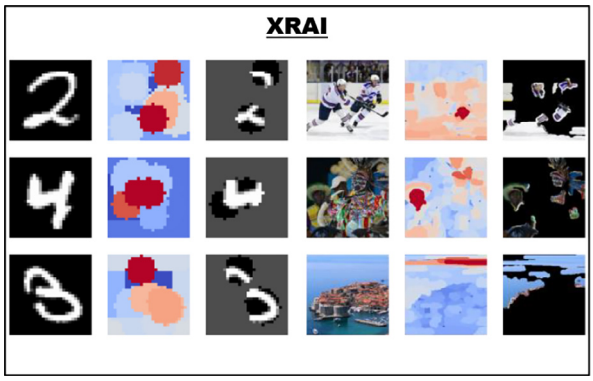

Fig. 8. The explanation results obtained from XRAI method on MNIST and Imagenet datasets.

synthetic neighborhood at random. At training time, ALIME [259] only conducts the random data creation once. To describe Bayesian prediction models, KL-LIME [216] integrated Kullback-Leibler divergence into traditional LIME calculation.

Shapley Additive exPlanations (SHAP) [181], is a local independent explanation approach that may generate a variety of models. All these methods share the same logic of determining the feature importance by computing the SHAP values according to a collaborative game assumption. The explanation generated by different SHAP methods varies in what way they compute SHAP values. Additive feature attribution is an alternative name of SHAP's explainability methods that are commonly used in the literature and is defined as:

$$g(s') = \varnothing_0 + \sum_{j=1}^{Y} \varnothing_j s'_j, \tag{22}$$

where $s' \in [0, 1]$, and $\varnothing_j \in \mathbb{R}$ represents the impact values allocated to every feature in the input, and Y represents the total of streamlined features. In general, SHAP methods have three characteristics: 1) local performance, which implies that $g(s')$ matches $b(s')$; 2) missingness, implying that SHAP values are not attributively affected by a particular feature $x_j = 0$; and 3) stability, implying that when the underlying model gets changed or replaced, the feature marginal contribution and the estimated SAHP values will exhibit no change. The way SHAP values are built allows them to be used at a local scale (e.g., the SHAP values are assigned to each input sample) and at a global scale by making use of mutual SHAP values. GradientExplainer, TreeExplainer, LinearExplainer, KernelExplainer, and DeepExplainer are the five methodologies for computing SHAP's values. The KernelExplainer, for example, is a model-agnostic technique, whereas the others are customized to different types of ML models.

DALEX [19] afford agnostic post-hoc explanation at both global and local scales. In terms of local explanations, DALEX has a flexible attribution technique implemented. It is made up of a breakdown of the model's predictions, each of which might be realized as a kind of local gradient information that can be leveraged to define the role of different features. Furthermore, DALEX supports a What-if explanation based on ceteris-paribus summaries to assess the impact of each feature independently while keeping the remaining features fixed. In terms of global explanations, DALEX includes a variety of exploratory tools, including variable significance measures, performance metrics, residue analyses, as well as dependency charts. Moreover, Neural Additive Models (NAM) [4,308] is a variety of GAM. The goal of this strategy is to merge the capability of robust AI solutions like DL possessing the intrinsic intelligibility of generalized additive models. As a consequence, a model capable

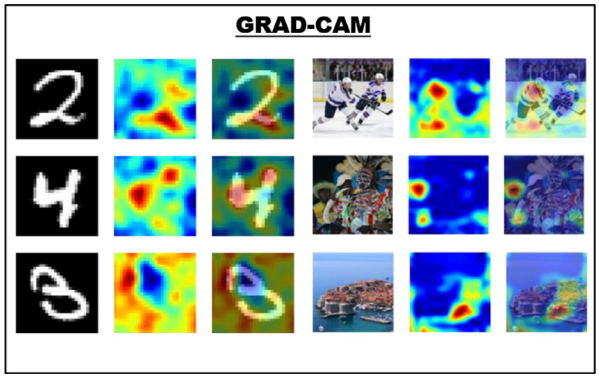

Fig. 9. The explanation results obtained from GRAD-CAM method on MNIST and Imagenet datasets.

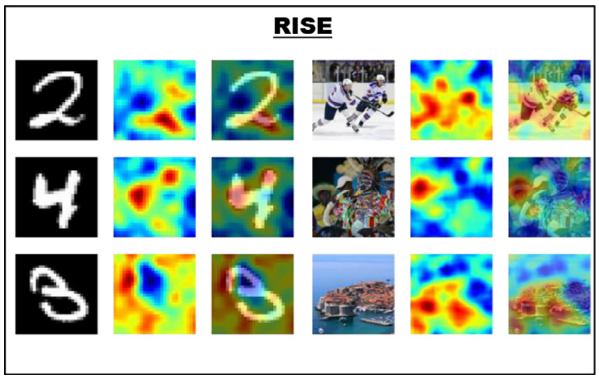

Fig. 10. The explanation results obtained from RISE method on MNIST and Imagenet datasets.

of learning graphs that define how the prediction is computed has been developed. This in turn enables additive training of many DL models, where each one of them emphasizes a specific attribute of the input. Furthermore, Contextual Importance and Utility (CIU) [10,11,86] stands for, and it's a local, agnostic explanation technique. The underlying principle of CIU has been that the context, or the group of input data getting evaluated, is a critical factor in giving proper explanations, which is why it is named as such. This method has found that basis that the significant characteristic in one situation may be irrelevant in another. Both the Contextual Importance (CI) and Contextual Utility (CU) techniques are designed to explain predictive performance in such a way that makes sense to both specialist and beginner audiences. The CI and CU are numeric that could be represented visually or in textual form in order to provide explanations for certain cases. When presenting explanations for particular instances, CI and CU are numerical numbers that could be conveyed in both visual and verbal formats, respectively. If more than one input, or perhaps all inputs, are taken into consideration, CI and CU could be determined. This allows for the inclusion of unlimited deeper notions that are composites of more than one insight in explanations. Technically, CIU uses Monte Carlo simulations to calculate the values for CI and CU. It is worth mentioning that this strategy does not necessitate the creation of a simpler model to use in order to derive the explanations.

B. Rule-based Explanation.

Local Rule-based Explainer **(LORE)** [233,257]**,** is a location-independent approach to delivering precise interpretations in the form of policies and counterintuitive regulations. LORE was created specifically for tabular data. To be more precise, it creates the record's neighborhood using an evolutionary algorithm. A neighborhood like this offers a more accurate and denser picture of a's location in relation to W.r.t LIME. LORE uses a genetic algorithm to construct a synthetic set S of neighbors given a black box x and an instance a, with $x(a) = u$, which is later declared to train a decision tree classifier $m(s)$. It retrieves an explanation from m that includes two parts: first, an objective decision directive that relates to the route taken by instance a on the decision tree to achieve the choice u, and (B) a collection of counterfactual rules that have a different categorization in relation to u. This collection of counterfactual rules illustrates the constraints that may be changed on a to alter the output decision. On the other side, DecText is a neural network-specific global model-specific explanation [22]. Dectext's goal is to locate the most important characteristics. Dectext is similar to trepan in that it evaluates four distinct splitting ways to attain this purpose. It also explores a fidelity-based pruning method to lower the size of the final explanation tree. Dectext can optimize fidelity while keeping the model basic in this approach.

Decision rules provide an explanation to the end-user of the factors that led to the final forecast. Because decision rules are human-readable, this category contains the minority of explanation techniques for tabular data. A decision ruler, also known as a factual or logic rule, has the form$c \rightarrow u$, where c is the premise, which is made up of a Boolean restriction on the value of the feature, and u is the ruling's outcome. Specifically, c is the conjunction of split conditions of the type $a_l \in [t_l^{(o)}, t_l^{(w)}]$, where a_l is a feature and in the domain of a_l the bound value of lower and upper are $t_l^{(o)}$, $t_l^{(w)}$ of a_l extended with $\pm\infty$. If all of c's Boolean criteria evaluate to true for a, then the instance a fulfills r, or r covers a. The rule $c \rightarrow u$ represents a proposed explanation of the decision $m(a) = u$ if the instance a to explain meets c. In addition, if the interpretable predictor behaves similarly to the black box in the vicinity of a, we may argue that the directive is a possible local interpretation for $x(a) = m(a) = u$. We emphasize the existence of so-called counterfactual rules in the context of rules. The main distinction between counterfactual rules and decision rules lies in the result of the rule u is distinct when compared to $x(a) = u$. It's critical to give an explanation to the human observers of what has to be altered to have a distinct result. $h = \{tall > 1.5, age < 35, fitness \geq efficiency\}, u = play$ is an example of a rule explanation. The record $\{tall = 1.7, age = 25, fitness = efficiency\}$ meets the condition above in this situation. Instead, a counterfactual rule may be $h = tall < 1.2, efficiency > fitness, u = notplay$. ANCHOR is a model-agnostic system that generates rules in the form of explanations. The output rules, known as anchors, give this technique its name. The concept is that adjustments in the remainder of the values of a feature of samples have no effect on the outcome of choices when the anchor holds. If $h(a) = x$, then h is an anchor given a record a. Anchor modifies example $\times$ in order to generate a group of artificial data suitable for extracting

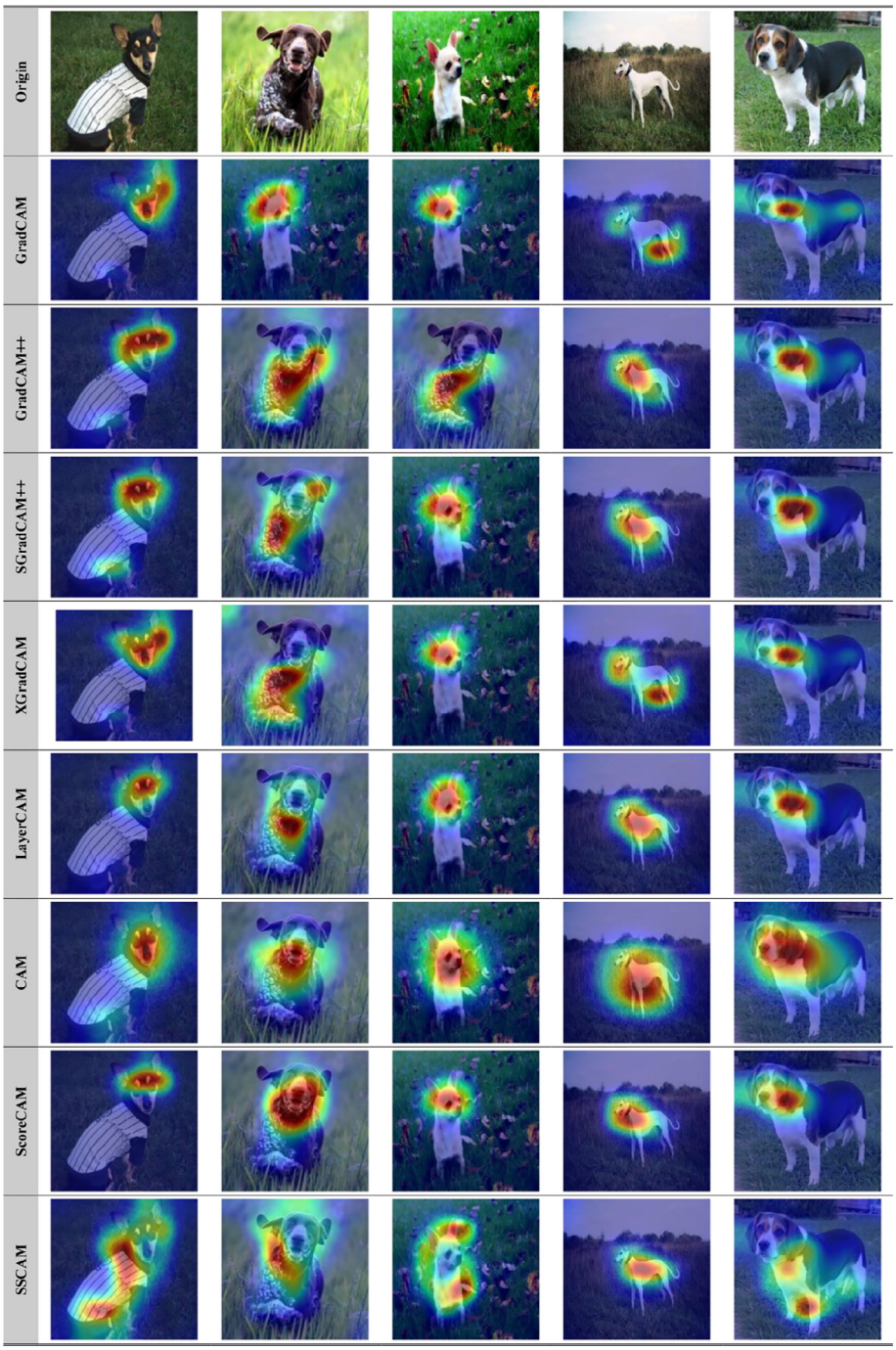

Fig. 11. An illustration of comparison between saliency maps generated by different explanation methods on using sample from Stanford Dogs Dataset.

anchors with accuracy greater than a user-defined limit. To begin, anchor employs a multi-armed bandit methodology so the data synthesis might generate a high number of samples. Besides that, because the number of potential anchors is exponential, anchors employ a bottom-up technique and beam search.

The Combined Multiple Model process (CMM) is a tree ensemble-specific global post-hoc explanation approach. Data enrichment is the most important aspect of CMM. In reality, CMM alters an input dataset A, z times before using it. It trains a black box on the z versions of the dataset. Then, on the black boxes, random records are created and tagged utilizing a bagging approach. This way, the amount of the data can be increased to allow training of final AI solutions. In another way, MSFT [31] is a post-hoc method for generating global explanations that use random forests to generate decision trees. It is based on

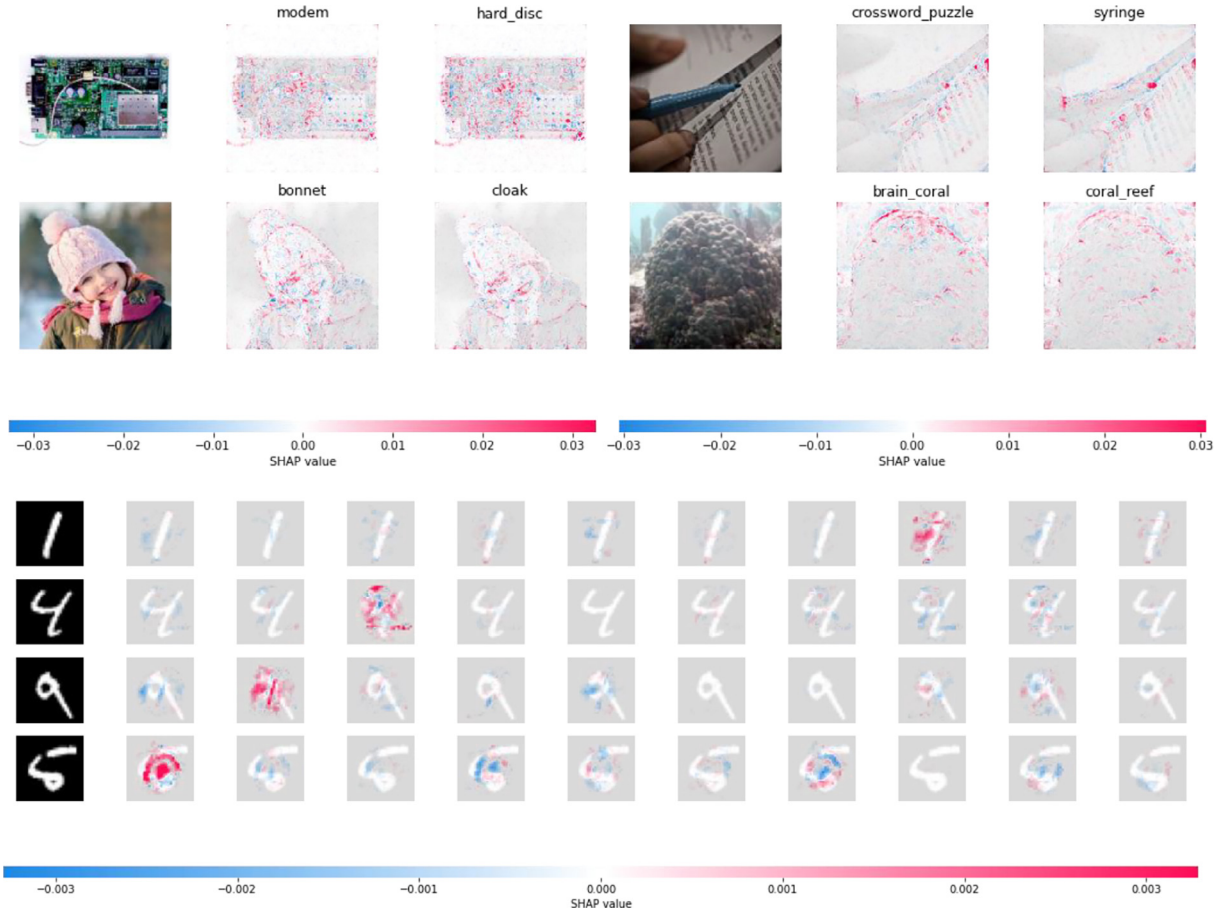

Fig. 12. An illustration for the results of SHAP method for explaining the results of the proposed model on MNIST and ImageNet datasets.

Table 4
The empirical results of different ambiguity methods on different.

	MNIST DATASET			DOGS vS CATS DATASET			IMAGE-NET DATASET		
METHODS	M1	M2	M3	M1	M2	M3	M1	M2	M3
Lime	96.56	0.189	1	95.15	0.284	20	42.27	0.028	67
Lrp	96.18	0.144	2	94.93	0.184	2	42.86	0.040	6
Deeplift	95.83	0.152	3	94.18	0.180	2	42.60	0.0162	7
Xrai	97.90	0.108	1	95.56	0.120	5	43.58	0.0328	24
Rise	96.42	0.130	1	94.18	0.160	6	42.53	0.0265	25
TCAV	97.82	0.179	67	96.43	0.209	111	44.33	0.0132	406
INTGRAD	96.43	0.173	5	95.51	0.210	3	43.57	0.0440	14
CAM	95.35	0.193	1	94.70	0.218	8	43.07	0.0602	2
SCORE-CAM	96.40	0.115	2	95.89	0.140	2	44.34	0.0128	4
SSCAM	96.88	0.199	2	95.61	0.214	7	43.00	0.0176	19
ISCAM	97.48	0.148	8	96.02	0.173	16	43.94	0.0495	30
GRAD-CAM	95.39	0.197	3	93.87	0.228	3	43.07	0.0462	4
GRAD-CAMpp	96.91	0.179	1	96.26	0.205	1	44.02	0.0132	3
SmoothGradCAMpp	98.16	0.109	1	95.27	0.151	1	43.62	0.0568	2
XGradCAM	98.95	0.132	1	95.95	0.180	15	43.39	0.0316	36
LayerCAM	97.37	0.167	2	95.42	0.197	3	43.61	0.0495	5

M1 = Insertion AUC, M2 = Deletion AUC, M3 = Time.

the fact that, despite the presence of hundreds of distinct trees, random forests are relatively similar, differing only in a few nodes. As a result, the authors recommend adopting a clustering approach to summarize the random forest trees using dissimilarity measures. After that, an archetype is obtained as an explanation for each cluster.

Table 5
An overview of explainability methods for solution based on tabular data.

Type	Name	Ref.	YEAR	Data Type	AH / PH	G / L	A / S
FI	SHAP	[181]	2007	ANY	PH	G / L	A
	LIME	[237]	2016	ANY	PH	L	A
	LRP	[16]	2015	ANY	PH	L	A
	DALEX	[27]	2020	ANY	PH	L / G	A
	NAM	[4]	2020	TAB	PH	L	S
	CIU	[11]	2020	TAB	PH	L	A
	MAPLE	[222]	2018	TAB	PH / IN	L	A
RB	ANCHOR	[238]	2018	TAB	PH	L / G	A
	LORE	[99]	2018	TAB	PH	L	A
	LRI	[295]	2000	TAB	AH	L	S
	MLRULE	[72]	2008	TAB	AH	G / L	S
	RULEFIT	[87]	2008	TAB	AH	G / L	S
	SCALABLE-PRL	[307]	2017	TAB	AH	G / L	A
	RULEMATRIX	[193]	2018	TAB	PH	G / L	A
	IDS	[164]	2016	TAB	AH	G / L	S
	DECTEXT	[33]	2002	TAB	PH	G	S
	STA	[329]	2016	TAB	PH	G	S
	SKOPERULE	[87]	2020	TAB	PH	L / G	A
PR	GLOCALX	[257]	2019	TAB	PH	L / G	A
	MMD-CRITIC	[145]	2016	TAB	AH	G	S
	PROTODASH	[104]	2019	TAB	AH	G	A
	TSP	[277]	2020	TAB	PH	L	S
CF	PS	[28]	2011	TAB	IN	G / L	S
	CEM	[73]	2018	ANY	PH	L	S
	DICE	[201]	2020	ANY	PH	L	A
	FACE	[225]	2020	ANY	PH	L	A
	CFX	[7]	2020	TAB	PH	L	S

L = Local; G = Global; A = Agnostic; S = specific, AH = ant-hoc; PH = post-hoc, TAB = Tabular.

The rule-based explanation technique GLocalX [258] uses a revolutionary methodology called the local to the global paradigm. The goal is to combine local logical rules into a global explanation. Starting with an array of factual rules, GLocalX integrates rules covering comparable data and expressing the same conditions in a hierarchical bottom-up approach. GLocalX determines the smallest collection of rules that are: 1) broad, meaning they should be valid to a huge fraction of the data, and 2) accurate. A collection of rules is the ultimate explanation proposed to the end-user. The model was validated in constrained environments with limited or no access to data or explanations from the natural ecosystem. Another simpler version of GLocalX is described wherein the eventual regulations are selected by a scoring system that emphasizes the applicability, completeness, and accuracy of the rules.

C. Prototypes.

A prototype, archetype, or artifact all refer to the same thing: a physical representation of a collection of similar records. 1) A sample from the training examples that are adjacent to input data x; 2) a center of the cluster that contains data in both of the first and second possibilities; 3) a record that was generated in an ad hoc manner, using a synthetic recording procedure. According to the explanatory method used, several conceptions and requirements for identifying a prototype are analyzed. In another way, look at entries that are similar to your own to get an idea of how the model thinks.

In the stream of post-hoc methods, the privacy-sustaining explanation is presented in [21] to produce local explanations in form of prototypes as well as shallow trees. It is the first approach to explainability that takes into account the idea of privacy by creating privacy-guarded descriptions. To establish a suitable trade-off between confidentiality and interpretability in the course of a model explanation, it used micro aggregation to design the explainer. The authors were able to create a collection of clusters, each with a representative record r_i, where i denotes the $i-th$ cluster. A shallow decision tree is extracted from each cluster to offer a complete explanation while maintaining strong comprehensibility due to the trees' modest depth. A representative record and its related shallow tree are chosen when a new record a comes. Specifically, based on the Black-Box choice, the representation r_i closest to a is chosen from m. At the same time, the prototype selection procedure navigates the search space for a set of prototypes that well represent the underlying training data. On the other hand, MMD-CRITIC was proposed by [100] as a pre-hoc technique, in that it merely looks at the distribution of the dataset being analyzed. It uses Maximum Mean Discrepancy to provide prototypes and critiques as explanations for a dataset (MMD). The first ones describe the dataset's overall behavior, whereas the latter ones reflect points that the prototypes don't adequately explain. MMD-CRITIC identifies prototypes by comparing the distribution of the instances to the distribution of the instances throughout the whole dataset. The set of cases nearest to the data distribution is termed prototypes, while the farthest is called critiques. Only minor data points that deviate significantly from the prototype but belong in the same category are shown by MMD-CRITIC. MMD-CRITIC chooses comments from sections of the dataset where the prototypes are underrepresented, with an extra restriction to guarantee that the criticisms are varied. In addition, as an extension to MMD-CRITIC, ProtoDash was proposed to generate an explanation that explains the input dataset using archetypal examples

and critiques. ProtoDash, on the other hand, correlates non-negative weights with the MMD-CRITIC, indicating the significance of each prototype. This allows it to reflect even the most complex constructions. Beyond and above, a model-specific explanation can also be derived from Tree Space Prototype (TSP) method at the local to help interpret different variants of decision trees algorithms [117]. The purpose is to locate prototypes in the tree ensemble x's tree space. TSP may extract prototypes for each class given a concept of tree proximity, with variants based on the kind of ensemble. For each class, multiple versions are proposed to allow for the option of a varied number of prototypes.

D Counterfactuals.

Counterfactuals describe the Black-Box model's reliance on external information to make a certain judgment. It concentrates on the discrepancies in order to arrive at the opposite conclusion in relation to $x(a) = u$. Counterfactuals are frequently referred to be the polar opposite of prototypes. The general form of a counterfactual explanation is formalized in [159]: Because variables of a have values $a_1, a_2, ..., a_z$, $x(a) = u$ was returned. If a had values of $a_{\frac{1}{1}}, a_{\frac{1}{2}}, ..., a_{\frac{1}{z}}$ and all other variables stayed constant, $x(\bar{a}) =\rightharpoonup u$ would have been returned, where $\bar{a}$ is the record a with the recommended adjustments. In an ideal counterfactual, the variable values should be changed as little as possible to find the closest configuration in which u is returned instead of u. Counterfactual explainers are classified as follows: 1) exogenous explainer for synthetic creation counterfactual; 2) endogenous explainer to gathers counterfactual from a total sample and thus provides a feasible sample in larger number compared to exogenous competent; 3) instance-based explainer for defining decision boundary based some distance function. Some of the desired outcomes in this context are efficiency, resilience, variety, actionability, and plausibility [9]. We recommend that the interested reader visit [152] gain a better understanding of the complex surroundings and the numerous choices available. [25] reports on a study on the understandability of factual and counterfactual explanations. The authors investigated the mental model supposition, which shows that individuals construct models that mimic the assertions made. They conducted research on a group of people and discovered that people prefer to reason using mental models to find likelihood, arithmetic, and rationality hard to follow. We will only briefly outline the most delegated strategies in this sector because there are published articles in this line of work.

Another post-hoc method is MAPLE [100], which is designed to generate a local agnostic explanation approach that can also be used as a translucent model due to its inner core. By integrating the decision tree model and feature engineering methods, it generates interpretations depending on feature importance. In maple, two strategies are used: SILO and DStump. Using random forest leaves, SILO is utilized to construct a local training distribution. DStump, in contrast, and, highlights the qualities. Maple utilizes the best k features from DStump to tackle a weighted regression analysis problem. Throughout this, the parameter of the local linear model, namely, the projected local contribution of every variable, includes an insight. Additionally, the Contrastive Explanations Approach (CEM) [40] is a contrastive local explanation method for the DL model that is post-hoc and model-specific by design. It is divided into two parts: relevant pluses, which could be believed of as prototypes which are the negligible and satisfactory variables that have to be available to get the output u, and relevant minuses, which are counterfactual considerations that need to be prevalent to get the output u. Also, it addresses the optimization challenge with a discrepancy that uses an autoencoder to assess a_1's in the proximity of the data.

The authors of [92] presented an agnostic approach called Diverse Counterfactual Explanations (DICE) is for delivering local explainability by ensuring feasibility and variety when returning counterfactuals by solving an optimization problem with many constraints. In the context of counterfactual reasoning, feasibility is crucial since it allows for the avoidance of unworkable scenarios. Take the example of an AI model that determines if to issue debts or not. When a model refuses a debt to a complainant with a limited income, the rationale for the rejection could be the lack of revenue. Nevertheless, a counterfactual like "You quadruple your pay" might well be unrealistic and so inadequate to explain the phenomenon. The closeness constraint from [103], the sparse constraint, and the z user-defined constraints are used to make the optimization problem feasible. Aside from practicality, variety is an important element since it allows for multiple approaches to change the result class. More, CFX [7] was developed as a model-specific approach for producing counterfactual explanations for bayesian DL, where the explanations are based on relationships of impact between factors, which reveal the classification's rationale. The primary accomplishment of this approach is that it may identify important components for the classification work, which, if eliminated, would result in a different categorization. Further, Feasible and Actionable Counterfactual Explanations (FACE) are presented in [225] as an agnostic method for generating local explanations that concentrate on generating feasible counterfactuals. In fact, it reveals "feasible routes" for creating counterfactuals. The shortest path lengths specified by density-weighted metrics make up these viable pathways. It could generate consistent counterfactuals matching with the distribution of input samples. The user may pick the prediction, density, weights, and conditions function before the face constructs a graph over the data points.

E. Transparent methods.

We show several transparent solutions for tabular data in this section. We offer various models that output feature importance first, followed by approaches that produce rules.

TED [66] is a transparent technique that takes a training dataset as input and links each record with its explanation. Explanations might be of any form, such as regulations or the relevance of a certain feature. The framework permits any AI solution that can deal with multi-class classification to be used in the training phase. As a result, the model can categorize the input record and link it to its explanation. The generation of explanations to feed throughout the training phase might be a potential restriction of this strategy. In aix360, ted is implemented. Extreme Boosting Machine (EBM) [94] is one of the variants of the Generalized Additive Model that use a boosting procedure that follows a round-robin strategy to recycle over

the features while mitigating the impacts of co-linearity. The model learns the optimal collection of feature tasks in this way, which may be used to deduce how each feature contributes to the final prediction.

The transparent rule learner SLIPPER [34] is built on a modified version of Adaboost. By enforcing limits on the rule builder, it produces concise and clear rules. On the other hand, LRI [126] is a translucent rule learning approach that provides interpretable rules as explanations while achieving high performance. Each training class in LRI is represented by a collection of rules that are not ordered. The rules are derived using an inductive technique that adaptively influences the accumulative miscalculation with no trimming. As soon as different data is counted, it is subjected to all of the available regulations. The output class for the record under analysis is the one with the most satisfactory set of rules. Moreover, MlRules [39] is a translucent rule induction technique that uses probability estimates to solve classification problems. Although boosting procedures are used for rule induction, maximum likelihood estimation is used for rule creation. RuleFit [49] is a translucent rule model that makes use of a tree ensemble. It starts by utilizing gradient boosting to generate an ensemble model. The ensemble's regulations are later obtained as a branch in the tree. Following the abstraction of the rules, they are prejudiced using an optimization dilemma according to L1 regularization. Furthermore, Interpretable Decision Sets (IDS) [79] is a clear and very precise decision-making approach. Decision sets are collections of if-then rules that are independent, short, precise, and non-overlapping. As a result, they can be used separately.

Without loss of generality, the explainability of ML/DL models on textual data can be performed with either tabular or vision explanation methods with few modifications. In Table 6, we review the state-of-the-art methods for explaining text-based AI solutions. The findings show that a wide span of the discussed methods can be applied to the text-related task. In this regard, we experimentally applied some of these methods to explain the transformer network trained for sentiment classification in the IMDB dataset. The interactive explanation obtained from the SHAP method can be observed in Fig. 13, while the visual comparative analysis between non-interactive explanations is presented in Fig. 14.

4.5.3. Explainability in graph data

The proliferation of graph-structured data brings the AI community with different categories of DL solutions called graph neural networks (GNNs) that can learn the complex relation embedded into graph data. As with other DL solutions, GNNs is designed as a black box as they are difficult to understand and debug, hence explaining these models become a vital requirement to enable further development in this era. This part of our tutorial will provide a deep analysis of explainability techniques that concentrate on distinct features of graph models and give alternative perspectives on how to comprehend them. They usually respond to a few problems, such as the most essential input links, input vertex, or vertex feature (VF), and the most graph pattern that has the ability to increase the prediction accuracy. To make things clearer, we present new taxonomy for categorizing explainability methods for graph machine learning solutions as indicated in Fig. 15. For experimental analysis, we evaluate different explainability methods on a case study of solubility prediction using 1000 organic molecules encoded in form of SMILES strings and the corresponding solubility [77]. A simple graph convolutional network (GCN) is trained and evaluated on this dataset and later leveraged to test different explanation methods.

The graph explanation methods can be broadly categorized into instance-centered and model-centered explanations [316]. The instance-based explanation interprets the GNN by finding relevant input features for prediction given an input network. The instance-based is classified into subcategories as gradient/feature(G/F) approaches, perturbations approaches, decomposition approaches, and surrogate approaches. The G/F -based approaches, in particular, use the G/F information to represent the feature relevance in input space. Furthermore, perturbation approaches analyze input importance scores by observing the change in prediction with regard to diverse input perturbations. The forecasting odds are first decomposed over the units of a final fully connected layer using the decomposition approaches. After that, these scores are propagated to each layer till reaching the input space and employ the decomposed scores as important values. On the other hand, surrogate-based approaches emphasize sampling the input data based on the neighborhood of a specific input sample. Furthermore, these approaches adopt a basic and understandable solution to the collected dataset, such as a decision tree. The surrogate model's interpretations are then utilized to explain the actual predictions. Following, model-based techniques describe GNNs in separation from any certain input example. The specific input justification is more general interpretation level and less explored compared with instance-based. Only XGNN is an approach that exists in model-based categories that depend on graph generation (GG). It creates graph examples in order to optimize the probability of prediction for a given class and then utilizes these graph patterns to interpret that class. In general, these two classifications describe DGM from various perspectives. Model-based approaches give high observations and a universal grasp of the function of DGM, whereas instance-based approaches convey example-specific explanations. Human supervision is required to review the explanations to validate and confide DGM. More human oversight is required for instance-based approaches, as specialists must investigate the interpretations for various input graphs. Because the explanations are global, model-based approaches require less human mo. In addition, the interpretations of instance-based techniques are based on real-world input examples to make them more understandable. Moreover, because the resulting graph patterns may not occur in the actual situations, the explanations for model-based procedures may not be easy to understand. Generally, these two types of categories may be integrated to better comprehend GNN.

4.5.4. Instance-based approaches

Herein, we examine the instance-centered methods for GNN explanations, involving G/F, perturbation, decomposition, and surrogate techniques. Furthermore, the goals, methodologies, benefits, and constraints of each approach will be clarified.

Table 6
An overview of explainability methods for text-based AI solution.

Type	Name	Ref.	YEAR	Data Type	AH / PH	G / L	A / S
SM	LIME	[237]	2016	ANY	PH	L	A
	INTGRAD	[276]	2017	ANY	PH	L	S
	L2X	[56]	2018	ANY	PH	L	A
	DEEPLIFT	[265]	2017	ANY	PH	L	S
	LIONETS	[199]	2019	ANY	PH	L	S
CA	–	[296]	2014	TXT	PH	L	S
	EXBERT	[124]	2019	TXT	PH	L	S
	–	[279]	2017	TXT	PH	L	S
PR	ANCHOR	[238]	2018	TXT	PH	L	A
	QUINT	[1]	2017	TXT	PH	L	S
	CRIAGE	[220]	2019	TXT	PH	L	S
	LASTS	[101]	2020	TXT	PH	L	S
	XSPELLS	[165]	2020	TXT	PH	L	S
	–	[232]	2019	TXT	PH	L	S
	DOCTORXAI	[213]	2020	ANY	PH	L	S

L = Local; G = Global; A = Agnsotic; S = specific, AH = ant-hoc; PH = post-hoc, TXT = TEXT.

Fig. 13. Illustration of the generated explanations from Deep-SHAP methods applied for transformer network on multi-class emotion classification dataset.

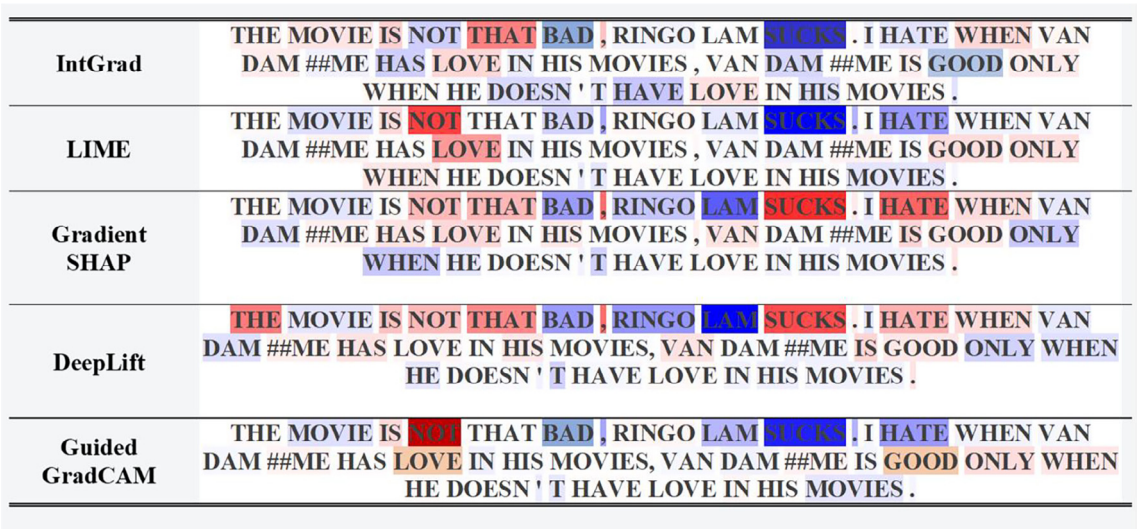

Fig. 14. Example of explanation generated for transformer network trained for sentimental text analysis from IMDB dataset.

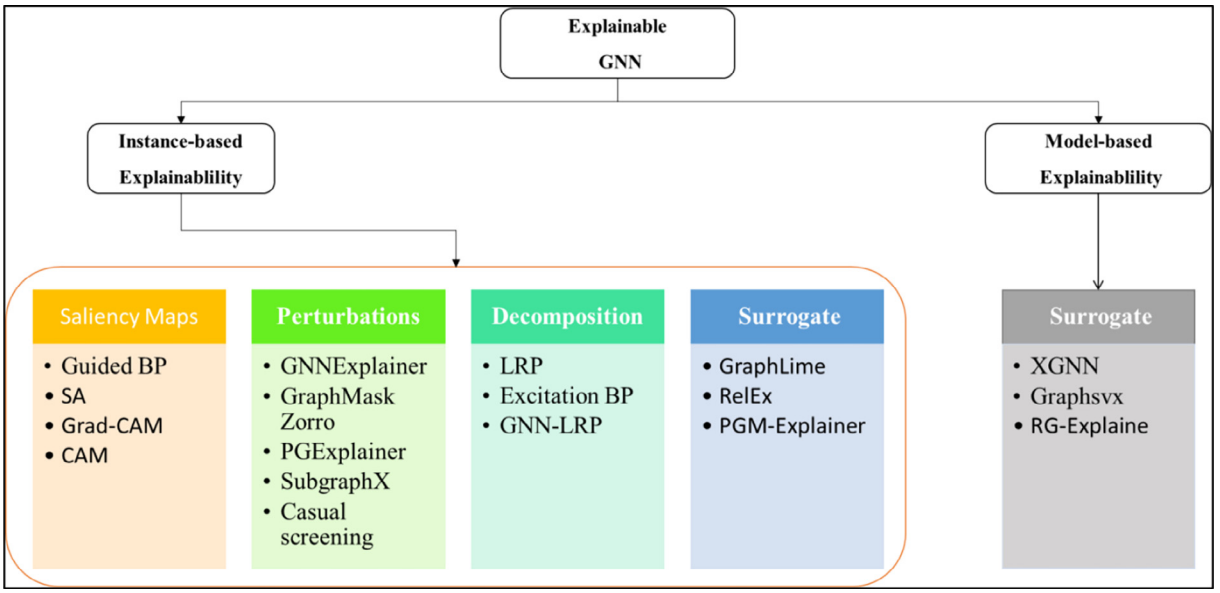

Fig. 15. Illustration of the proposed Taxonomy for categorizing different methods for explaining graph neural networks.

A. Saliency approach.

The simplest technique, which is extensively used in image and text operations, is to employ G/F to describe the GNN. The essential concept is to estimate input significance using feature maps and gradients information. Gradient-based algorithms use back-propagation (BP) to calculate the gradients of objective forecasts with regard to input features. However, features-based approaches use interpolation to map hidden patterns to the learned features and estimate importance scores. Larger G/F values, overall, imply more significance in such techniques. Because internal feature maps and gradient information are significantly connected to learning parameters, such explanations can represent the model's information. These approaches can be simply extended to the graph domain since they are straightforward and generic. Various approaches, such as sensitivity analysis, GBP, CAM, and Grad-CAM, have lately been used to describe GNNs. The mechanism for gradient BP and the way of integrating the various hidden feature maps are the main differences between these approaches.

Sensitivity analysis**:** obtain the importance score of various features by directly using the squared values of gradients. It can be directly calculated by BP, which is similar to the learning process, however, the objective is input space rather than weight space. The input features may be graph vertex, links, or VF. If the absolute gradient has superior values, it implies that comparable input spaces are more significant. Even though it is straightforward and effective, it has a number of drawbacks.

To begin with, SA can only indicate the thoughtfulness of input and production, which does not accurately indicate the significance. Furthermore, it struggles with infiltration issues. The gradients can rarely indicate the contributions of inputs in the model's saturation areas when the model output varies slightly in regard to any input change.

Guided BP (GBP): is based on the same concept as SA, but it alters the BP gradient's operation. Given the difficulty of explaining negative gradients, GBP just flows the positive gradients in the backward direction, limiting nonpositive gradient information to zeros. The importance of various input features is then measured using solely positive gradients. It's worth noting that Guided BP has the same drawbacks as SA. On the other hand, **CAM** aims to declare an important vertex by transferring the VF in the last layer to the input space. To derive importance scores for input vertex, CAM uses the final vertex embeddings and weighted summations to merge multiple feature maps. The weights come from the target prediction's last FC layer. This method is likewise incredibly easy and effective, but it has some significant drawbacks. CAM has a unique necessity on the GNN architecture, which restricts its use and generalization. Also, it makes the heuristic assumption that the end vertex embeddings can represent the input importance, which may be wrong. Moreover, it can only be used to interpret graph classification (GC) models and not to solve node classification (NC) operations.

By eliminating the GAP layer restriction, **Grad-CAM** [54] adapts the CAM to generalized GC models. It also measures node importance by transferring the final vertex embeddings to the input space. Rather than utilizing weights to integrate the GAP and FC outputs, it uses gradients as the weights to mix distinct feature maps. It begins by computing the target prediction's gradients with respect to the last vertex embeddings. The weight for every feature map is then calculated by averaging such gradients. Grad-CAM, unlike CAM, does not necessitate the use of a GAP layer before the last FC layer in the GNN model. It is dependent on heuristic premises, and it is unable to explain NC models.

B. Perturbation approach.

As with image data, the fundamental goal is to investigate output changes regarding various input perturbations. logically, the predictions should be equivalent to the actual predictions if important input information is kept. To explain GNN for images, recent approaches train an originator to produce a filter to choose significant input regions. On the other hand, such an approach cannot be straitly appended to graph models. In contrast with images, graphs are displayed as vertex and links, and thus can't be scaled such that the same node and link numbers are used. Furthermore, structural information is important for graphs and can influence their role.

SubgraphX, ZORRO, GNNExplainer, GraphMask, Causal Screening [57], and PGExplainer [45] are some of the perturbation-based approaches suggested to interpret DGM. As demonstrated in Fig. 16, they have the same global structure. First, filters are created from the input graph to represent important input features. Remarkably that various masks are created based on the interpretation operations, such as vertex filter, link filter, and VF filters. The created filters are then integrated with the input graph to produce a new graph including critical input data. Lastly, the newly created graph is fed.

to GNNs, which are used to assess the filters and renew the filter generating procedures. Logically, the critical input features collected by filters have to convey the key semantic definition, resulting in a prediction that is comparable to the original. The three key differences between these approaches are the mask creation method, the mask type, and the objective function.

Different categories of masks are presented in this literature including discrete, soft, and estimated discrete masks. Soft masks for VF, discrete masks for links, and approximated discrete masks for vertices as demonstrated in Fig. 16. The soft masks values lay in the range $[0; 1]$, and BP may directly update the mask generating process. Soft masks, on the other hand, struggle with the introduced evidence (IE) issue [316], which means that any value other than one and zero in the mask might bring additional semantic definition or distortion to the input graph, which affects the interpretation outcomes. To avoid the IE issue, the discrete masks only hold the values of zeros and ones and do not add a new numerical value. Discrete masks, on the other hand, usually include non-differentiable procedures like sampling. The policy gradient approach [107,315] is a prominent way to address it. Moreover, current works suggest using reparameterization techniques to estimate the discrete masks, such as Gumbel-Softmax approximation and scant reductions. Surprisingly, the output mask is not accurately distinct however provides a useful approximation, allowing GBP while also greatly reducing the problem of IE.

GNNExplainer [312] uses mask optimization to learn soft masks for links and vertex characteristics in order to interpret the forecasts. Soft masks are initialized at random and regarded as trainable variables. The masks are then combined into actual graph structure via duplications by GNNExplainer. The filters are then adjusted by augmenting the shared knowledge between the actual graph's forecasts and the newly acquired graph's forecasts. Even when different regularization techniques, including element-wise entropy, are applied to enhance the discontinuity of optimum masks, yielding soft masks, the IE problem is inescapable in GNNExplainer. Also, the customization of masks according to the input makes the interpretations prone to miss a high perspective.

To interpret the predictions, **PGExplainer** [182] learns estimated discontinuous masks for links. To predict link masks, it trains a parameterized mask predictor. It starts by combining vertex embeddings to get the embeddings for each link in an input graph. The predictor then employs the link embeddings to forecast the likelihood of every edge being chosen, which can be regarded as the signature mark. The reparameterization trick is then used to sample the estimated discontinuous masks. Lastly, by extending the joint knowledge between the actual and recent predictions, the mask predictor is trained. However, if the reparameterization approach is used, the resultant masks are not precisely discrete, but they can significantly reduce the problem of IE. Furthermore, because all links in the data have a single forecaster, the explanations can give a broad perspective of GNNs.

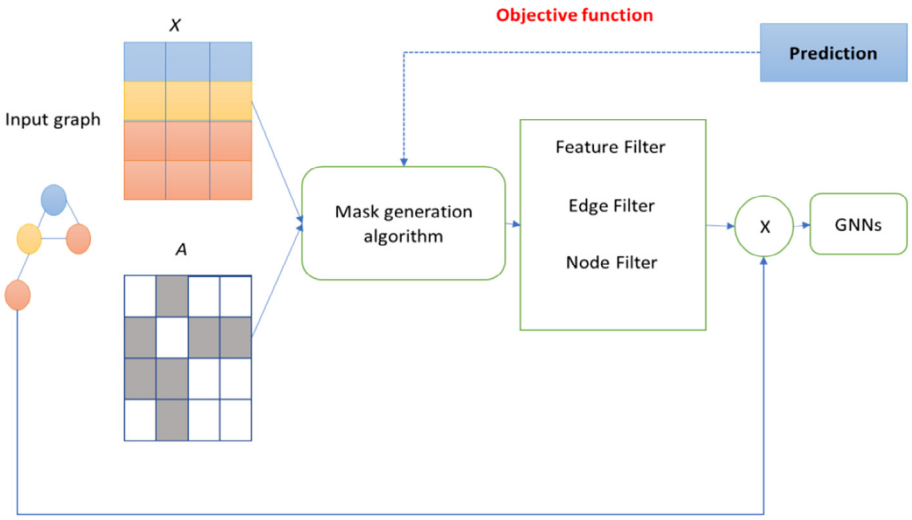

Fig.16. Illustration of generic workflow of Perturbation approaches for explaining graph machine learning.

SubgraphX [317] investigates DGM partial graph interpretations. It uses the Monte Carlo Tree Search (MCTS) method [37] to quickly interpret various subsets of the graph through vertex trimming and choose the very significant subgraph from the search tree's leaves as the prediction's interpretation. Furthermore, it uses the Shapley value [69] as the MCTS recompense to determine the significance of parts of graphs and presents an effective rough calculation of the Shapley value just by evaluating info exchanges in a particular transmission range. On the other hand, it does not investigate masks explicitly as the MCTS method may be thought of as a mask production technique, and its vertex pruning behaviors can be thought of as distinct masks for obtaining subgraphs. Furthermore, the Shapley values may be used to adjust the mask production algorithm's objective function. Its produced subgraphs are more individual-understandable and suited for graph data than previous perturbation-based approaches. Nevertheless, the complexity is higher since the MCTS method is used to interpret distinct subgraphs.

Causal Screening [95] investigates the attribution of causality to various links in the input graph. It locates the explanatory subgraph's link mask. The main concept behind causal attribution is to look at how predictions alter during a link is added to the present explanatory subgraph, which is termed the causal effect. It examines the causal consequences of many links at each step and chooses one to append to the subgraph. It selects links using the individual causal effect (ICE), which evaluates the change in joint info (among the forecasts of actual graphs and the interpretive subgraphs) when various links are added to the subgraph. Causal Screening, like ZORRO, is a greedy algorithm that generates discontinuous masks with no training. As a result, it does not emerge the issues of IE, but it may miss a global understanding and be bound to local optimum interpretations.

GraphMask [250] is a post-hoc approach for interpreting the significance of links in the graph layer. For estimating if a link may be eliminated without changing the actual predictions, it uses a classifier like the PGExplainer. GraphMask, gets a link mask for every layer of GNN, whereas PGExplainer solitary concentrates the input space. Also, the deleted links are changed with learnable baseline links, which are arrays with identical sizes as vertex embeddings, in order to prevent modifying graph structures. To estimate discontinuous masks, the binary Concrete distribution and reparameterization method are used. Also, the classifier is trained by reducing a variance, which evaluates the variance among network predictions, through all the datasets. It may significantly relieve the IE issues and find a global comprehension of the trained GNNs, like the PGExplainer.

To determine important input vertices and vertex attributes, **ZORRO** [15] uses discrete masks. A greedy method is used to choose vertices or vertex attributes gradually from an input network. For each phase, ZORRO selects the vertex or VF with the greatest quality rating. The fitness function, fidelity rating, assesses the method of matching the novel projections with the network's actual projections by preserving selected vertices and replacing the rest with random noise values. Because no training process is used, the non-differentiable constraint of separate masks is overcome. Furthermore, ZORRO solves the issue of IE by utilizing hard masks. The greedy mask chosen process, on the other hand, may result in local optimum interpretations. Furthermore, because masks are created for every graph separately, the interpretations may miss a global understanding.

C. Surrogate approaches.

Thanks to the complicated and irregular connections linking the input and output forecasts, DNN is difficult to interpret. The surrogate technique is a common approach to offering instance-based explanations for image networks. The basic concept is to use a simple and explainable surrogate model to approximate the complicated DNN's predictions for the input

example's vicinity regions. These strategies are based on the assumption that the connections in the input example's nearby regions are less complicated and may be adequately described by a simpler surrogate model. The explainable surrogate model's interpretations are then utilized to explore the actual prediction. Surrogate techniques in the graph domain are difficult to use because of the discrete nature of graph data with topological design. Then it's unclear how to determine the input graph's surrounding areas and which explainable surrogate models are appropriate.

In this context, a few of the surrogate approaches introduced lately to interpret GNN such as PGM-Explainer [282], Relix [324], and GraphLime [128]. Fig. 17 demonstrates the overall process for various technologies. They initially acquire a local dataset including numerous surrounding data items and their predictions in order to explore the prediction of a particular input graph. Then they fitted an explainable model to the local dataset to learn it. Eventually, the explainable model's interpretations are treated as the actual model's interpretations for the input graph. While all techniques have a similar global concept, the significant differences are in the way that the local dataset is obtained and which explainable surrogate model is used.

GraphLime [128] applies the LIME algorithm to GNN and investigates the role of various vertex attributes in vertex classification tasks. GraphLime considers a goal vertex in the input graph's *N*-hop surrounding vertices and their forecasts obtained from the corresponding data, where *N* is an acceptable depth of underlying GNNs. On the other hand, the interpretability of vertex categorization is also investigated in **RelEx** [324]. Surrogate models are created by merging the principles of surrogate approaches and approaches based on perturbation. Because of a goal vertex and a set of parameters, it basically generates its computational graph (*N*-hop vicinity), a local dataset by picking linked subgraphs at random and sampling these subgraphs from the computational graph to the GNNs that have been trained to make predictions. In the beginning, the target vertex chooses surrounding vertices at random in a BFS approach. The GCN model is then used as a surrogate model to fit the local datasets. It's worth noting that, unlike GraphLime, LIME and RelEx's surrogate model is not explainable. It then uses the previous perturbation-based approaches to interpret the predictions, such as producing a particular mask, after training. It can give interpretations for critical vertices in comparison to GraphLime. It involves various process of guesstimates, such as utilizing the surrogate approach to estimate regional interactions and masks for approaching link reputation, providing less persuasive and trustworthy interpretations. Moreover, because perturbation-based approaches may be used straight to describe actual DGM, another non-explainable DNN does not need to be built as a surrogate model to interpret. It's also unclear how it may be used to solve GC problems.

PGM-Explainer [282] was published as a probabilistic method for an instance-wise explanation of GNNs. Random NF perturbation is used to create the local dataset. PGM-Explainer, given an input graph, perturbs the vertex features of multiple random vertices inside the computational graph at random intervals. PGM-Explainer then preserves a random variable for every vertex in the computational graph, reflecting if its features have been modified and their impact on GNN predictions. A local dataset is created by repeating such methods several times. The PGM-local Explainer's dataset comprises vertex variables rather than distinct surrounding graph samples. The Grow-Shrink (GS) technique is then used to choose the top dependent variables in order to shrink the local dataset. PGM-Explainer can describe graph vertices, but it ignores graph links, which include important graph topological information. Moreover, unlike the above methods, the PGM-Explainer can describe both vertex classification and GC operations.

D. Decomposition approaches.

Decomposition approaches, which assess the significance of features in input space by decaying the actual model predictions into many terms, are another common way to describe DGM image classifiers. The significance grades of the relevant input features are then assigned to those terms. These approaches look at the model parameters straight to see how the features in the input space relate to the output predictions. The total of the decayed terms must equal the actual prediction score for these approaches to be accurate. But, because graphs contain vertices, links, and NF, applying such approaches directly to the graph domain is difficult. providing scores to various links is difficult because graph links carry important topological information that cannot be neglected.

Excitation BP [54], LRP [18], and GNN-LRP [223] are some of the latest decomposition approaches suggested to describe deep GNN. These algorithms work on the idea of creating record decomposition regulations to spread forecast records for all input features**.** Fig. 18 demonstrates the whole process for various approaches. These approaches backpropagate the prediction score in each layer until it reaches the input layer. The forecast obtained from GNN is used as the first goal mark beginning at the output layer. based on the decomposition principles, the record is decayed and allocated to the units in the preceding layer. They can acquire importance scores for vertex characteristics by repeating similar operations until input space is reached, which can then be concatenated to indicate link importance, vertex importance, and the importance of walking. Every of these approaches neglect the activation functions in GNN. The primary distinction between both strategies is the score decomposition rules and explanation goals.

LRP [18,253] is an extension to the traditional LRP method. The result prediction score is decayed into multiple vertex importance scores. The hidden features and weights are used to build the score decomposition rule. The score of a goal neuron is a linear estimation of the neuron scores from the preceding layer. In theory, the neuron that contributes the most to target neuron activity should obtain a bigger fraction of the goal neuron score. Because LRP is created directly from model parameters, the outcomes of its interpretation are more precise. Unfortunately, it could just investigate the significance of the various vertices and unable to be used for partial graphs and walks. Furthermore, such an approach necessitates overall comprehend of the model architecture, that restricting non-expert users' applications like multidisciplinary researchers. In addition, **Excitation BP** [54] is depending on the principle of absolute likelihood and has the same idea as the LRP proce-

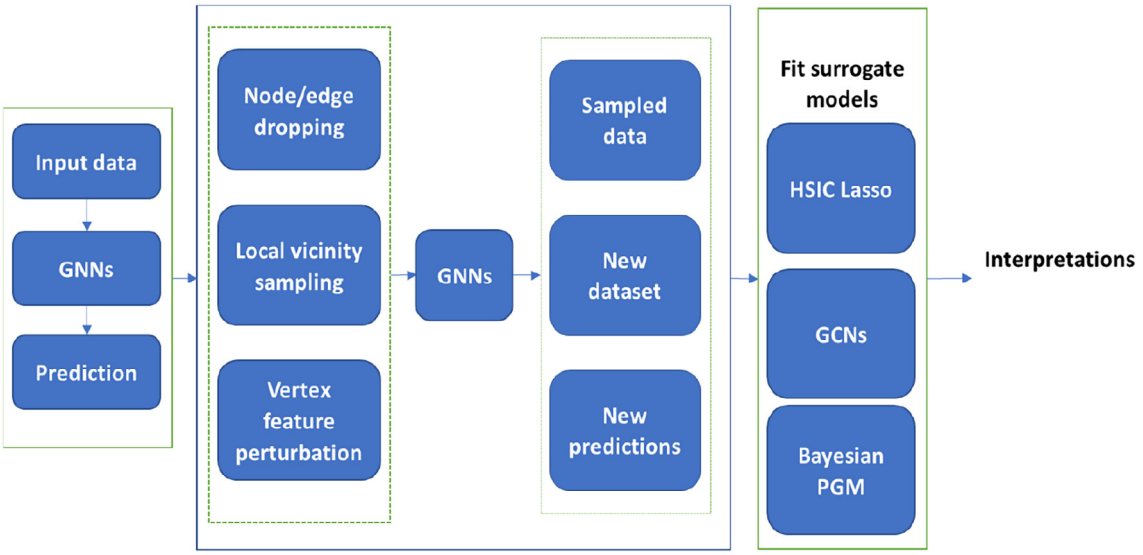

Fig. 17. Illustration of generic workflow of surrogate approaches for explaining graph machine learning solution by sampling the training data to signify. the interactions around the objective data.

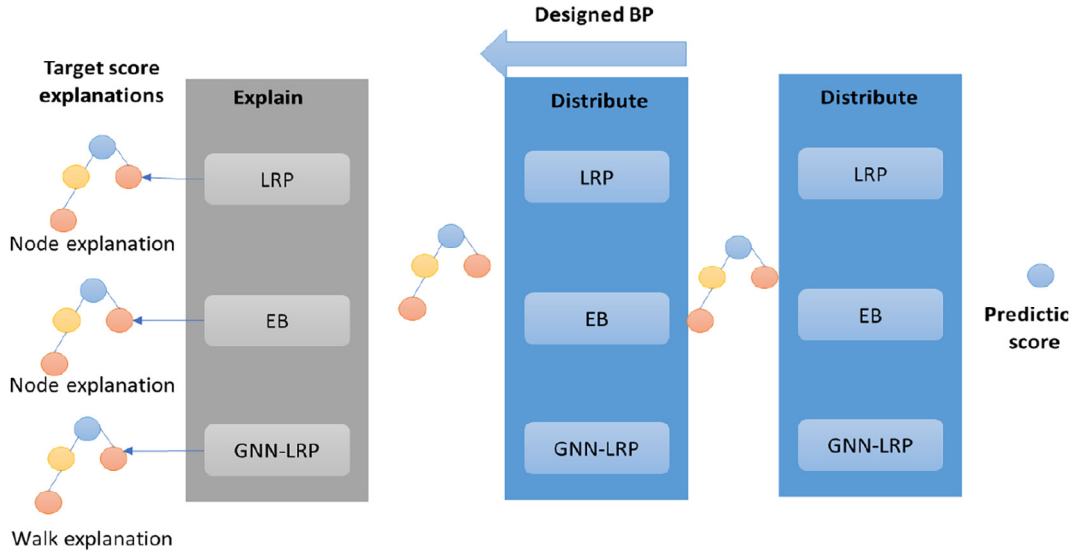

Fig. 18. Illustration for generic workflow of decomposition explainability methods for GNN, and they dispense the forecast record to input space to signify the significance of input.

dure. It states that a neuron's probability in a given layer equals the sum of the odds it delivers to all linked units in the preceding layer. The record decomposition regime may thus be thought of as breaking down the goal probability into a few conditional probability provisions. As a result, it has the same benefits and drawbacks as the LRP algorithm. Moreover, **GNN-LRP** [223] investigates the significance of the various graph strides when performing vicinity information gathering, it is more logical to use deep GNN as graph walks correlate to communication flows. It keeps the path of the distribution operations as they go to each layer, which is treated as separate walks, and the scores are derived from the vertices that correspond to them. While GNN-LRP has a strong theoretical foundation, its calculations may be inaccurate due to approximations. Furthermore, because each walk is analyzed independently, the computational cost is significant. It is also difficult for non-experts to utilize, particularly in transdisciplinary domains. In Fig. 19, we illustrate the explanation generated from LRP on the solubility prediction case.

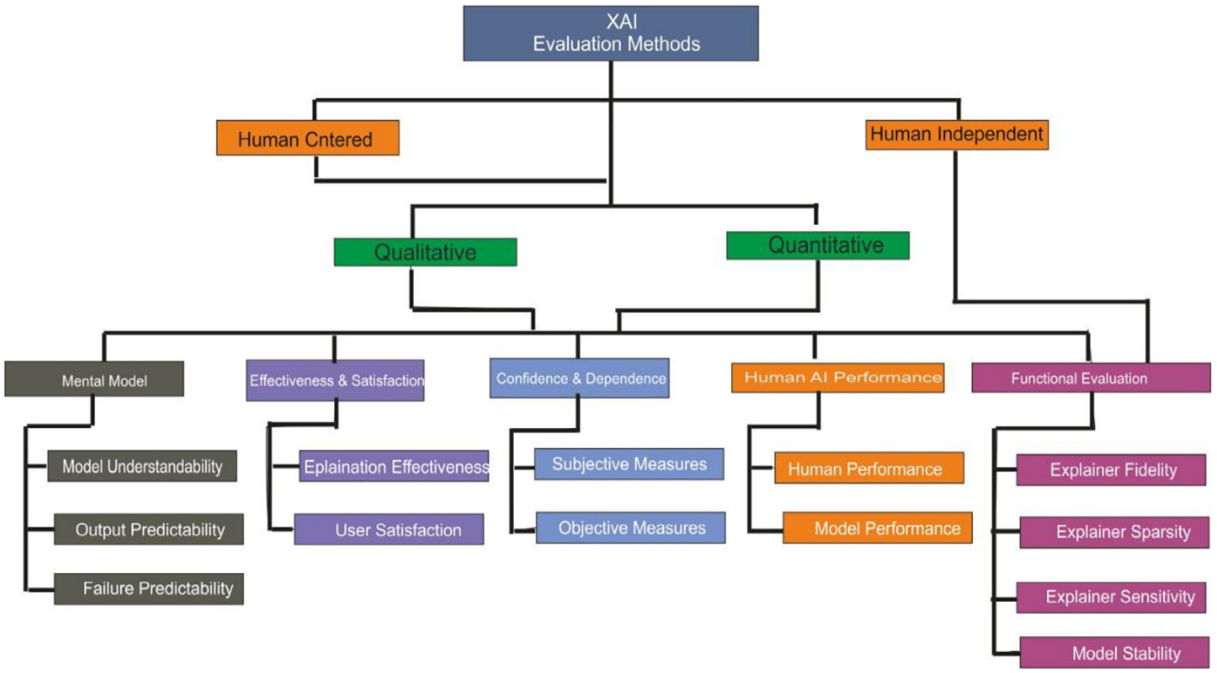

Fig. 19. systematic illustration of the taxonomy of exploitability evaluation measures in XAI.

4.5.5. Model-based approaches

Model-based approaches, unlike instance-based methods, seek to provide broad perspectives and global comprehension to interpret DGM. They investigate which graph representation could result in particular GNN behavior, such as optimizing goal forecasting. To generate model-based interpretations for image classifiers, input optimization is a preferred approach. Due to the discontinuous graph structure information, it cannot be easily appended to graph models, making interpreting GNNs at the model-based more difficult. As a result, it remains an essential yet understudied issue. Till now, XGNN [315] is the only model-based technique for describing graph neural networks that exist. In particular, **XGNN** [315] was introduced to interpret GNNs through GG. It trains a graph generator rather than directly optimizing the input graph such that the produced graphs can optimize an objective graph calculation. The produced graphs are then used as interpretations for the goal prediction, and discriminative graph patterns are assumed to be present. The GG issue is formulated as a reinforcement learning (RL) issue in XGNN. They are supposed to have discriminative graph patterns and are used as interpretations for the goal prediction. The GG issue is represented as an RL issue in XGNN. The generator predicts the way to add a link to the existing graph at every step. The resulting graphs are then fed to underlying GNNs to gain feedback for policy gradient training of the generator. In addition, additional graph principles are used to ensure that the interpretations are both correct and intelligible to humans. The XGNN is a generic framework for providing model-based interpretations that may be used with any appropriate GG technique. Furthermore, the explanations are broad in scope and provide a comprehensive interpretation of the underlying GNNs.

In this part, we provide a thorough examination of several interpretation strategies. Table 7 summarizes the characteristics of several models. We concentrate on the following factors for each approach.

(1) **Type:** This parameter specifies the kind of interpretations given by each explanation approach, such as instance-based and model-based interpretations. Model-based explanations are more global and broader, whereas instance-based interpretations are input-dependent and more specialized.
(2) **Learning:** This refers to whether the interpretation approach includes any kind of learning process. In general, the learning technique with interpretation technique conveys the connections between inputs and predictions well. When the learning techniques include more black boxes, they become less trustworthy. The feature/gradient-based approaches and decomposition approaches, which are stated in Table 7, have no learning procedure.
(3) **Task:** It specifies the tasks each technique can be used for. Only NC and GC problems are considered here. It illustrates the generalizability of several methodologies.
(4) **Target**: It displays the method's intended interpretation, such as vertex, link, or graph walk importance. V stands for vertices, L for Links, VF for Vertex feature, and walk for graph walks.
(5) **Black-Box** refers to how the interpretation approach deal with the trained GNNs. XGNN, GraphLime, PGMExplainer, GNNExplainer, ZORRO, Causal Screening, RelEx, or just require gain access to the output and input to interpret GNNs.

Table 7
An overview of explainability methods for graph-based AI solution.

Technique	Category	Learning	Task	Target	Black box	Direction	Design
Sensitivity analysis [18223]	Instance-based	No	GC/NC	N/E/NF	No	Backward	✗
GBP [18]		No		N/E/NF	No		✗
CAM [223]		No	GC	N	No		✗
Grad-CAM [223]		No		N	No		✗
GNNExplainer [312]		Yes	GC/NC	E/NF	Yes	Forward	Yes
PGExplainer [182]		Yes	G	E	No		Yes
GraphMask [250]		Yes		E	No		Yes
ZORRO [15]		No		N/NF	Yes		Yes
Causal Screening [95]		No		E	Yes		Yes
SubgraphX [317]		Yes		Subgraph	Yes		Yes
LRP [18253]		No		N	No	Backward	No
Excitation BP [54]		No		N	No	Backward	No
GNN-LRP [223]		No		Walk	No	Backward	Yes
GraphLime [128]		Yes	NC	NF	Yes	Forward	No
RelEx [324]		Yes		N / E	Yes	Forward	Yes
PGM-Explainer [282]		Yes	GC/NC	N	Yes	Forward	Yes
XGNN [315]	Model-level	Yes	GC	Subgraph	Yes	Forward	Yes

(6) **Direction:** The calculation flow in terms of forwarding or backward computations.
(7) **Design:** This factor indicates if an interpretation approach has a special design for graph data or is just a variation of the picture domain. Given the unique nature of graph data and the importance of topology/structural information, it is crucial to explicitly address such information while providing interpretations.

5. Explainable AI: Evaluation methods & metrics

With the expanding research on different categories and subcategories of techniques for explaining AI algorithms, it is extremely important to investigate and explore the methods and metrics for evaluating the XAI algorithms. Analytical findings indicate that when AI undergoes collaboration with human individuals, the individuals can usually make excellent decisions while the AI algorithms deliver an accurate explanation. In contrast, a flawed explanation is can usually lead to serious consequences. Also, the specialists ought to know if they could trust the generated explanation from XAI algorithms. It is broadly identified that explanation outcomes are prone to misunderstanding, specifically the visual ones. By extensive analysis of the XAI literature, it is worth noting, till now, that there is no generally accepted metric for evaluating the superiority of the generated explanations. This can be attributed to the absence of ground truth for most situations, the wide variety in the nature of the generated explanations, and the variability of input data (i.e., tabular, visual, graphical). Additionally, the satisfactory evaluation metric might differ a lot in accordance with the particular evaluation target and oriented audiences. The literature contains numerous regulated in-lab and real-time crowdsourced studies for the evaluation of explanations from algorithms. Likewise, case studies seek to gather expert users' feedback while executing high-level intellectual tasks leveraging a variety of analytical tools. In this context, Doshi-Velez et al [75] offer three types for the evaluation of XAI models. First, application-grounded evaluation - this approach uses humans to evaluate every result of the explainable Ai model within real applications and decide if this model works well or not. Where the domain expert evaluates the explanation of the application. This approach is simple, very slow, and needs high cost because it needs experts in that domain [137]. Second, human-grounded evaluation-this approach is similar to the application-grounded approach, but tasks are evaluated by non-experts rather than domain experts. Which experiment tasks are simple, reasonable, and understandable by average people? Third, in function-grounded evaluation humans are not involved in the evaluation of explainable AI models in this approach. The goal is to utilize structured proxy concerns to assess explainability, and which models can be evaluated. Moreover, the authors of [21] classified the evaluation of explanations into quantitative and qualitative approaches [9]. It is worth mentioning that some of the literature studies did not execute any evaluation for the explanations obtained from the XAI algorithms, these studies could be assessed by domain specialists. Comprehensive explanations given by, for example, Grad-CAM might be evaluated by professionals by just studying the global feature maps calculated by every class of gradient activation map indicating each class, instead of the local attribution maps one at a time. In the following, provide a comprehensive taxonomy for categorizing the different measures and the relevant methods for evaluation of XAI algorithms according to different criteria and perspectives (see Fig. 20). Table 8 overview the literature studies for evaluating explanation in XAI.

5.1. Mental models

Based on rational psychology principles, the mental model represents the way in which human individuals interpret AI decisions. Mental models have been broadly used in the literature to define their interpretation of intelligent systems in a variety of technical and scientific domains. In the spectrum of XAI, the generated explanations provide a useful source for

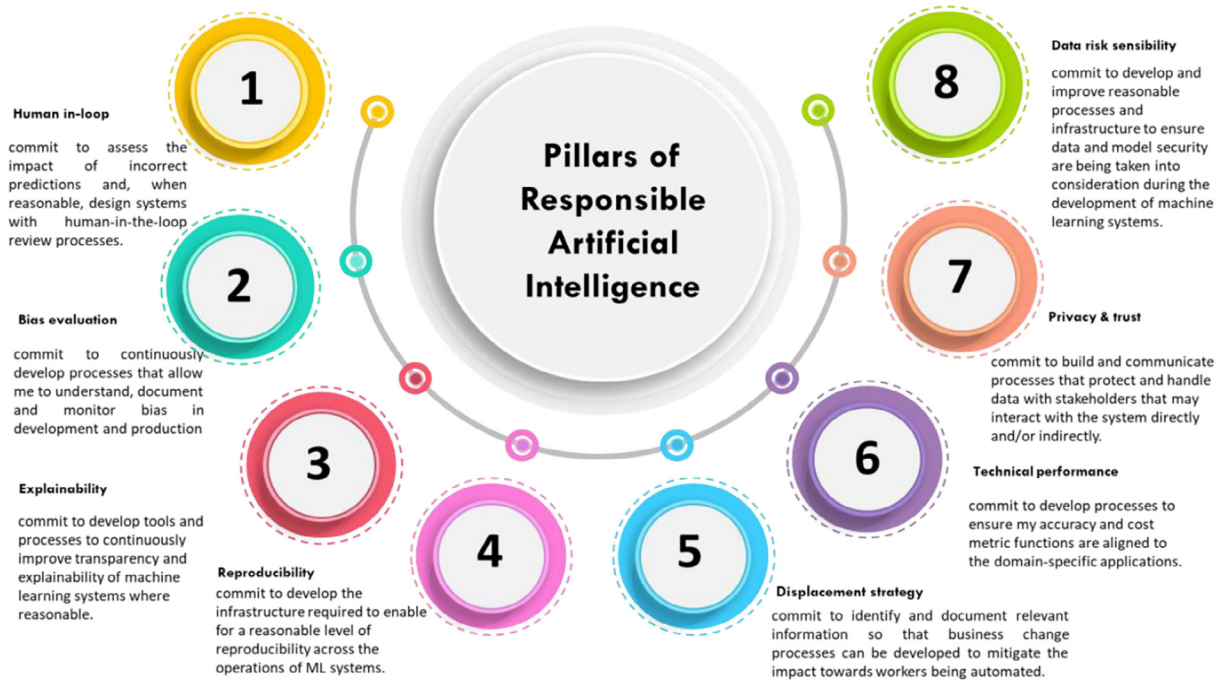

Fig. 20. Main pillars for developing responsible Artificial intelligence solutions.

Table 8
An overview of the scientific studies proposed to evaluate different explainability measures, categorized based on evaluation aspect, measures, and the construction method.

Evaluation aspect	Evaluation Measures	Evaluation Methods	Studies
Mental Models	model understandability	Interview and Self-explanation	[62,197,74]
		Likert-scale Questionnaire	[252,207]
	output predictability	Predicting Model Output	[237], [238,140]
	failure predictability	Model Failure analysis	[20,142]
Effectiveness & Satisfaction	Explanation Effectiveness	Task Duration and Cognitive Load	[92,162]
		Commitment with Explanations	[63]
	User Satisfaction	Interview and Self-explanation	[40,173]
		Likert-scale Questionnaire	[252,207]
		Expert Case Study	[135,157,175,176]
Confidence & Dependence	Subjective Measures	Interview and Self-explanation	[41,43]
		Likert-scale Questionnaire	[24,41,43]
	Objective Measures	User Perceived System Competence	[206,227,311]
		User Compliance with System	[80]
		User Perceived Understandability	[205,311]
Human-AI performance	Human Performance	Task Performance	[174,135,160]
		Task Throughput	[174,160,164]
		Model Failure Prediction	[20,142]
	Model Performance	Model Accuracy	[176,159,221]
		Model Tuning and Selection	[292]
Functional evaluation	Explainer Fidelity	Simulated Experiments	[237,229]
		Sanity Check	[243,314,319]
		Comparative Evaluation	[244,260]
	Explainer Sparsity	Sparsity computation	[213,301,89]
	Explainer Sensitivity	Sensitivity analysis	[268]
	Model Stability	Debugging Model and Training	[129,187,102,158]
		Human-Grounded Evaluation	[192,213]

human individuals to establish a mental model for characterizing the way the AI algorithms act [197]. In this regard, we propose to further taxonomize the mental model measures into three subcategories namely model understandability, output predictability, and failure predictability (see Fig. 19). ML-driven explanation is a means for assisting individuals to build a more precise mental model. Researching individuals' mental models driven by XAI explanations can support validating the efficacy of explanation in expressing the decisions of AI algorithms [302]. The literature of studies in human-AI collab-

orations has also investigated construction, categories, and tasks of explainability to discover important elements of perfect explanation for superior human perception and additional precise intellectual models [114,246]. For example, the authors of [179] examined how various kinds of explanations could assist in building theoretical representation. To discover how an intellectual system ought to describe its conduct for non-specialists, studies on ML explanations have analyzed how individuals can interpret AI algorithms and agents to discover what humans demand from machinery explanations. In another way, the mental model can be measured by the means of failure analysis. for example, the authors of [20] created a game, where members obtain regulatory incentives according to the ultimate performance score. Though experimentations were performed on a straightforward 3D task, the findings demonstrate a decline in the capability of human individuals in forecasting the failure of the model because of the growing complexity of the model and data. in addition, Perlmutter et al. [218] show how representations affect the correctness of people's mental models of what robots can perceive and the effectiveness and efficiency for communicating mission orders. A practical method of analyzing individuals' understanding of AI solutions is to promptly question them concerning the rational decision-making operation. Scrutinizing individuals' interviews and self-explanations offers important knowledge about the thinking procedures and mental models of human individuals [160].

5.2. Effectiveness and satisfaction

The satisfaction and effectiveness of explanations for human individuals are of great prominence when assessing explanations of algorithms [29]. Scholars exploited a variety of personal and factual measures for comprehensibility, effectiveness, and abundance of details to evaluate instructive value for individuals [192]. Though the existence of inherent techniques to quantify the satisfaction of humans [111], a substantial portion of the literature comes out of a qualitative assessment of fulfillment in explanations i.e., interviews and questionnaires. In this scenario, the authors of [92] assessed ten distinct explanation categories via the individual grades of satisfaction and clarity of explanation. The findings demonstrated a powerful correlation between human satisfaction and apparent simplicity. In a similar way, the study of [174] investigated the explanation efficacy and productivity of an interpretable context-aware system by introducing various categories of explanations based on "why", "why not" and "what if" queries by assessing the reaction time of individuals.

5.3. Confidence and dependence

The human trustworthiness in an AI solution can be considered an affecting and intellectual aspect that impacts negative or positive experiences of the underlying system [113]. Primary human confidence and the improvement of confidence over time have been analyzed and stated with diverse names like default trust [190], suspicious confidence, and swift confidence [191]. Earlier information and attitudes are vital in defining the preliminary state of confidence; nevertheless, confidence and dependence can adjust in response to investigating and confronting the system with edge situations [112]. Thus, individuals can have various thoughts of confidence and distrust throughout various periods of knowledge with any certain system. Scholars characterize and evaluate confidence with various methods. User experience, practical competency, familiarity, trust, attitudes, trust, reactions, and individual additions are general conditions employed to evaluate and inspect trust [183]. For these consequences, confidence and dependence can be quantified by clearly questioning individuals' opinions throughout and after dealing with the system, which could be carried out via questionnaires and interviews. The authors of [311] examined the significance of the accuracy of AI algorithm on human confidence, and the outcomes revealed that human confidence was impacted by both the asserted accuracy of the AI system and the accuracy perceived by the user over time. Furthermore, confidence evaluation dimensions might be certain to the application domain and XAI design objectives. For example, many dimensions would evaluate the human feedback about the dependability, consistency, and security of the system individually. In this regard, a thorough confidence measurement framework is presented in [43], which evaluates human confidence with compound confidence levels (i.e., video recording, and interviews) to evaluate three forms of system presentation It is worth noting that though the research literature did not contain an explicit of measurement of confidence to be generally highlighted in analysis tools for data scientists and AI experts, the human dependence on tools and the propensity to maintain make use of tools are regularly deemed as a portion of the assessment procedure during case studies. Ribera et al [239] divided system targets justifications to distinct groups of human individuals into developers, AI researchers, domain experts, and novice users. Putnam et al [228] conduct a study to identify student attitudes toward incorporating explanations of Intelligent Tutoring Systems by soliciting recommendations from participants on the kinds of explanations, and the results show an overall positive sentiment toward wanting an explanation. In addition, Weidele et al. [294] provide visualized AutoML named AutoAIViz that wants to free data scientists from tedious manual labor with model explanations and results so that consumers can trust the outputs, and it has been assessed by ten data scientists (domain experts) of an experimental system. Moreover, Paudyal et al. [214] evaluate the usability and utility of the SignGuru system by a user study using 14 Aided Sign Language (ASL) signs with 26 users.

5.4. AI vs human performance

A crucial objective of XAI is to assist human individuals in being effective in their responsibilities including AI systems [122]. Thus, the performance of human-AI tasks is a gauge related to all levels of audiences. The authors of [174] evaluated individuals' performance according to success ratio and task end-time to assess the influence of various categories of expla-

nations. They utilized a standard interface that used for different kinds of sensor-centered context-aware systems, i.e., climate forecast, where explanations could be help individuals in adapting the AI system to cope with their requirements. On the other hand, visual analytics techniques also support domain professionals to well achieve their tasks by delivering model understandings. Envisaging model construction, specifics, and ambiguity in AI outcomes could permit domain professionals to analyze models and modify hyper-parameters to their particular data for improved analysis. This research direction has investigated the necessity for model explanation in multimedia [38,59], and text [127,175] analysis, by revealing the significance of incorporating human feedback to enhance AI decisions and outcomes. In another direction, instead of domain professionals, AI specialists can also use visual analytics to discover training weaknesses and imperfections in the model building in DL algorithms to enhance the performance of the underlying task [176]. For example, the author of [135] developed an approach to envisage case-level and subclass-level neuron activation in a lasting exploration and improvement with AI engineers. Their case studies considered questioned three data scientists and AI specialists who utilized the tool and stated their crucial opinions. Besides, Kuleszka et al. [159] evaluated approach efficiency to demonstrate that explanatory debugging improved client and machine effectiveness and efficiency and allow users to explain adjustments to the learning system. In addition, Das et al [68] created automatically from an AI's internal task model that explains what the system is doing to teach the user, therefore improving task performance. moreover, the authors of [109] developed task descriptions that can be easily comprehended by humans to promote interpretability for both humans and robots. Furthermore, Billiet et al. [30] deliver findings that provide a fair balance of system performance and accuracy, as well as an assessment module through a color-coded graphical interface that includes a variety of evaluation criteria.

5.5. Computational measures

As a popular way to evaluate the explainability methods, computational measures are presented to assess the acceptability and comprehensiveness of generated explanations [143]. The literature studies [79] show that dependence on humans in assessing explanations could result in convincing explanations instead of translucent systems because of the human preference for shortened explanations. This in turn leads to the contention that the fidelity of explanations of XAI algorithms must be evaluated by computational techniques rather than human-centered analyses [265,181]. The Fidelity of a particular XAI method describes the exactness of the explanation technique in producing the real explanations for the AI decisions. Consequently, a sequence of computational techniques was developed to assess the acceptability of obtained explanations, stability of explanations, and fidelity of explanations. Ross et al. [245] developed empirical assessments and compare the computational cost and consistency of their explanations with the LIME technique [237] which evaluates explanation fidelity by comparison with intrinsically interpretable models that quantify the feature importance. Besides, Bucinca et al. [39] XAI systems are evaluated using proxy tasks such as how effectively people anticipate the choice from the supplied justifications and subjective measures of trustworthiness and preference as real predicting performance.

5.5.1. Fidelity measure

The generated explanations must be devoted to the underlying AI solution by recognizing the key features necessary to the AI algorithm, not the targeted audience. For instance, the Fidelity+ [223] performance measure was recently designed to compare this. Subjectively, if significant input features recognized by explanation methods are discriminatory to the AI model, removing these features should cause the forecasts to substantially differ. Fidelity + is thus described as the variation in accuracy between the last and fresh forecasts after filtering significant input features [223,123]. Let G_i denoting the i-th input and $f(.)$ representing the AI algorithm that need to be interpreted in formal terms, and the outcome can be represented as$\bar{y}_i = \arg\max f(G_i)$. The explanations can then be thought of as a hard importance map m_i, with each element indicating whether the corresponding feature is important or not. It is worth noting that the generated explanations for methods [95,15] are discrete masks that can be used directly as the importance map. Furthermore, if the importance scores for methods [128,312] are constant ideals, the significance of the metric m_i could be acquired using thresholding and normalisation. Therefore, the *Fidelity*+ prediction accuracy score can be calculated as.

$$Fidelity+^{acc} = \frac{1}{N}\sum_{i=1}^{N}\left(1\left(\bar{y}_i = y_i\right) - 1\left(\bar{y}_i^{1-m_i} = y_i\right)\right) \tag{23}$$

where as y_i denotes the initial forecasting of input i and N denotes the number of inputs. $1 - m_i$ denotes the opposite mask which means eliminating the significant input features, and $\bar{y}_i^{1-m_i}$ denotes the outcome obtained by passing new input to trained AI algorithm $f(\cdot)$. If $\bar{y}_i$ and y_i are equal, the indicator function $1\left(\bar{y}_i = y_i\right)$ returns 1; otherwise, it returns 0. It should be noted that the $Fidelity+^{acc}$ metric investigates the change in prediction accuracy. The *Fidelity*+ of probability can be defined by focusing on the predicted probability.

$$Fidelity+^{prob} = \frac{1}{N}\sum_{i=1}^{N}(f(G_i)_{y_i} - f\left(G_i^{1-m_i}\right)_{y_i}), \tag{24}$$

where $G_i^{1-m_i}$ denotes the new input formed by retaining features of G_i relying on the complimentary mask $1 - m_i$.

It is significant to mention that *Fidelity*$+^{prob}$ tracks the transformation in forecasted probability, which is more delicate than *Fidelity*$+^{acc}$. Increasing scores for both performance measures allow improved explanation outcomes and the identification of more key features [98,223]. The *Fidelity*+ measurement investigates forecasting transformation by removing important features. conversely, the measurement *Fidelity*− experiments forecast to change by retaining significant input features while discarding insignificant ones. Intuitively, the notable features should encompass contextual power, resulting in forecasts that are equivalent to the previous forecasts even when insignificant features are eliminated. Formally, the measurement *Fidelity*− can be calculated as follows:

$$Fidelity+^{acc} = \frac{1}{N}\sum_{i=1}^{N}\left(1\left(\bar{y}_i = y_i\right) - 1\left(\bar{y}_i^{m_i} = y_i\right)\right) \tag{25}$$

$$Fidelity+^{prob} = \frac{1}{N}\sum_{i=1}^{N}(f(G_i)_{y_i} - f(G_i^{m_i})_{y_i}), \tag{26}$$

where G_i^m is the new graph created by retaining significant features of G_i relying on explanation m_i and $\bar{y}_i^{m_i}$ is the fresh forecasting It is worth noting that for both *Fidelity*$+^{acc}$ and *Fidelity*$+^{prob}$, smaller values indicate that less important information is deleted in order to improve the explanations outcomes.

5.5.1.1. Sparsity measure. Better explanations should be sparse, capturing only the major significant input features while ignoring the irrelevant features. Sparsity is an indicator that shows such an estate. It particularly represents the ratio of features chosen as significant by explanation techniques [223]. Formally, given the input G_i and its tough significance map m_i, the Sparsity measurement can be estimated as:

$$Sparsity = \frac{1}{N}\sum_{i=1}^{N}\left(1 - \frac{|m_i|}{|M_i|}\right), \tag{27}$$

wherein $|m_i|$ symbolizes the total of significant input features recognized in m_i and $|M_i|$ embodies the total number of features in G_i. A higher score indicates that the explanations are sparser and seem to encapsulate only the major significant input details.

5.5.2. Stability measure

Furthermore, better explanations should be consistent. Intuitively, when tiny differences to the input are made without influencing the forecasts, the explanations should stay constant. The previously approved Stability measurement assesses the stability of an explanation technique [249]. Based on a given input G_i, the explanations m_i are considered the ground truth. The input G_i is then perturbed by slight differences, to yield a new input$\dddot{G}_i$. It is significant to mention that G_i and $\dddot{G}_i$ must have the same forecasts. The explanations of $\dddot{G}_i$ are then acquired, signified asm_i. We can calculate the Stability score by looking at the difference between m_i andm_i'. Lower value shows that the explanation method is more durable and resistant to noisy information. Furthermore, because some input formats are sensitive, hence, determining the appropriate number of perturbations may be difficult.

5.5.2.1. Performance measures. Furthermore, for synthesis datasets, the Accuracy metric is proposed [249,312]. Even though it is unknown whether the XAI algorithms make predictions in our expected manner in synthesis datasets, the rules for building these datasets can be used as reasonable approximations of the ground truths. Then we can compare the explanations for any input sample to such ground truths. When studying important edges, for example, we can compare the matching rate for important edges in explanations to those in-ground truths. Common accuracy, precision, recall, and area under the curve are common metrics for such comparisons. Higher values indicate that the explanations are closer to the truth and can be considered better results. Unfortunately, due to a lack of ground truths, the Accuracy metric cannot be applied to real-world datasets.

Different from computer vision, qualitative evaluations could have a restricted potential with the time series data. In fact, the non-instinctive characteristics of time series make it challenging even for specialists to qualitatively evaluate the quality of the obtained explanations. These findings imply that the research community should concentrate on quantitative assessments for the XAI algorithms that are based on time-series data. In the nutshell, the quantitative or qualitative evaluations could be less or more suitable depending on what we intend to undertake with the explanations. Quantitative evaluations can be well fitted to explanations that struggle to discern new analytical knowledge in the data [33], while Qualitative evaluations can be better fitted to explanations that focus on decision-makers and end-users [27].

6. Explainable AI: Applications

Remarkably, XAI can bring noteworthy advantages to a broad range of life sectors and application domains that may involve AI decisions. In this regard, this section provides an analytical overview of XAI literature in various applications and figures out the potential application domains that might require research work on explainable algorithms.

6.1. Healthcare

AI applications in the healthcare domain witnessed an increasing trend since 2017. AI algorithms have demonstrated their ability to synthesize and extract valuable knowledge from enormous amounts of medical data that might be difficult or even impossible for humans to comprehend [302]. In fact, the most important thing about healthcare AI is the requirement to be *informed consent* of the patient, which means sharing the decision-making between doctors and patients in such a way that the patients have the final opinion. Healthcare AI can thus only stand applied if and only if the patients are informed about its crucial functionalities in advance—definitely in a comprehensible manner [274]. This way, the European essential rights fundamentally necessitate the integration of XAI in healthcare. To achieve that, recent research efforts have been dedicated to developing XAI systems to generate interpretations and explanations for healthcare practitioners that can improve the reasoning and decision-making chain in healthcare [274].

To this end, the XAI can be examined and studied in different subdomains of healthcare including medical image analysis, health record analysis, and drug discovery.

Medical image analysis: AI approaches have been demonstrating great success in detecting, localizing, and identifying different diseases from medical images such as computed tomography (CT), X-rays, Magnetic Resonance Images (MRI), etc. Accordingly, they become an essential component for building computer-aided diagnosis (CAD) systems. To increase the trust in such systems, the explainability of medical AI algorithms turned out to be a mandatory requirement to enable verifying the generated diagnostic decisions thereby guaranteeing the dependence on correct image features. In this regard, more research attention has been devoted to XAI-based medical image analysis. for example, an explainable CNN, called CA-Net, was presented by [97] for automatic segmentation from medical images, whereas channel attention is presented for adjustive recalibration of channel-wise feature reactions and emphasized the most important feature channels, and a scale attention module was developed to accentuate learning on the salient feature maps among manifold scales to be adapted to the objects' size. Besides, the authors of [126] proposed an interpretable multimodal fusion approach for the detection of biological mechanisms of brains, called Grad-CAM guided convolutional collaborative learning (gCAM-CCL), that integrates gradient-based parameters with intermediary feature maps to produce interpretable activation maps for computing pixel-level donations of the input features. Moreover, the authors of [289] presented an interpretable convolutional dictionary network, termed DICDNet, for metal artifact reduction (MAR) in CT images. it presents a proximal gradient-based optimization technique to iteratively unfold into matching building blocks according to particular physical implications.

Recently, XAI has attracted more attention for the diagnosis of COVID-19 pandemics from medical images. for example, the authors of [299] presented an explainable DL framework that combines classification and segmentation networks to offer a real-time diagnosis of COVID- 19 from chest CT scans. Similarly, the DeepSC-COVID was introduced in [291] for 3D classification and segmentation of COVID-19 from CT scans, where three sub-networks id developed for extracting features necessary for classification and segmentation of infection tasks. It also applies task-aware loss to capture the interaction between segmentation and classification sub-networks, thereby offering an interpretable diagnosis. In addition, Malhotra et al [184] presented a multi-task explainable deep network, called COMiT-Net, for COVID-19 screening from X-ray images by just predicting if COVID-19 features exist or not, then, conducting semantic segmentation for the regions of interest to provide an explanation of the outcomes. In another way, Shi et al [263] Presented a knowledge distillation based explainable attention-transfer classification network for automatic discriminating between community-acquired pneumonia and COVID-19 based on radiographic imaging by transferring deformable attention maps (as a knowledge) from a teacher network (responsible for global feature extraction from infection regions) to the student network to emphasis learning erratically shaped lesion areas.

Electronic Health Record (EHR) analysis: The authors of [166] developed an XAI approach for early warning scores that contains an efficient and precise temporal convolutional network (TCN) for forecasting acute serious disease from electronic health records of patients. They also introduced to offer straightforward visual explanation method to explain the TCN-generated predictions for three emergency cases namely acute kidney injury (AKJ), sepsis, and acute lung injury (ALJ). In addition, the authors of [320] presented a computing approach, named Patient2Vec, that provides an interpretable learning longitudinal pattern from personalized EHR data to predict future hospitalizations. The Patient2Vec generates a vector space with profound architecture to estimate the feature importance that generates clinical intuitions that could be visualized and understood by human individuals. Besides, the authors of [49] presented an interpretable DNN for predicting an essential clinical 1-year mortality from multi-modal data comprising echocardiographic videos and EHRs. Moreover, the authors of [264] proposed an interpretable DL framework, named DeepSOFA, to exploit the temporal representation necessary to evaluate sickness severity (i.e., Sequential Organ Failure Assessment (SOFA)) at any point in the course of an ICU stay. The DeepSOFA is considerably more precise forecasting performance for in-hospital mortality. Furthermore, the authors of [35] have formulated the challenge of forecasting chronic disease hospitalization as a binary classification problem that can be

addressed by sparse logistic regression, kernelized and sparse SVMs, and random forests. Whereas the tradeoff between accuracy and interpretability was tackled by distinguishing unseen patient groups and adjusting classifiers to each group using two techniques namely Joint Clustering and Classification (JCC) and likelihood ratio test-based technique.

The research on the role of XAI in healthcare has witnessed a great increase in the last years as a result of the increase in the amount of clinical information from different modalities. In view of the presented taxonomies (see previous sections), this section categorizes the XAI studies in healthcare according to five categories of techniques namely Attention techniques, dimension reduction techniques, feature interaction techniques, knowledge distillation techniques, and surrogate representations techniques (see Table 9).

6.2. Robotics and automation

As robots become more sophisticated and independent, there is a growing need for humans to comprehend what they do and think. As robots grow more common in human-inhabited situations, it is becoming increasingly vital for people to retain a reasonable level of confidence in robots. The capacity to explain the knowledge and rationale behind decisions, like in human relationships, substantially enhances trust by developing an understanding of why a decision was taken and consequently offers insight into future decisions. Several research efforts within Human-Robot Interaction (HRI) focus on how to make robots understand human actions and thoughts. The robot projects assembly information and planned robot movements onto the worktable in order to improve collaboration with a human operator [163]. Our tutorial considers the papers studying robotics and how they achieve transparency in their work, and that mostly use DNN techniques and fuzzy systems in their work. Robots almost include post-hoc Explanations to be transparent, and understandable to humans and show their collaboration with humans using a dialogue system that includes conversational, chatbots, and textual explanations with humans to increase human trust in robots that provide understanding into robotic behaviors by allowing a robot to respond to inquiries about its activities posed by humans (e.g., Q: "Why did you turn left there?" A: "I noticed someone at the end of the corridor.") like in [261]. Visual question answering and graph-based explanations play the main role in robot explanations. Atman et al. [14] investigate two similar architectures for a semi-autonomous robotic swarm that differs only in the visual feedback given to a passivity-short assumption human operator and graphical explanations [298]. Most of these papers include human-grounded evaluation in which humans evaluate the robot explanations and User Study for the effective measures of evaluations of the results papers which use end-users to participate and evaluate its explanations as with [195].

6.3. Justice system

The major goal of establishing explainable judgments is to allow the information subject to understand the findings of fact and to avoid discrimination or other illegal acts [36]. XAI can help experts (police, forensics, psychologists, and politicians) to understand the criminal behavior in the city by using a feature representation based on the weather. Deeks et al. [71] believe that justices should need and seek interpretations for algorithms decisions and that developing systems that explain how algorithms derive their findings, suggestions, or forecasts is one method to solve opaque model concerns, and to avoid models that affect people's liberty, safety, or privacy, and that criminal justice algorithms predict recidivism less accurately than humans. As a result, XAI has a reason to better estimate recidivism risks and reduce the expenses associated with both crime and jail [2].

Organizations in many domains, such as health care, financial services, logistics, insurance, advertising, and telecommunication are using forecasting models to gain strategic competitive advantage, and software-supported forecasting of legal decisions to predict legal outcomes have been launched in recent years world [285]. DNN and rule-based classification models are mostly used in this field, for example, Chhatwal et al. [57] present explainable predictive coding that provides lawyers with greater confidence in the results and locates responsive snippets within responsive documents. So, it uses interpretable models and rule extraction for explanations and dialogues in conversational systems; as an example, Hepenstal et al. [110] offer a unique architecture for creating a clear chatbot AI-based information retrieval system to help crime justifications.

6.4. Cybersecurity

Massive attacks made by malware such as viruses, backdoors, spyware, trojans, and worms have presented a high-security threat to users' devices, so virus analysts need transparent and interpretable classifiers to detect malware from the large and imbalanced gray list. AfzaliSeresht et al. [3] present an explainable based on Association Rules to assist the analysts' judgment to know what is happening from the security event logs through interpretable and visualization techniques. Soviany [270] provides a better comprehension of the deployment of a rule-based approach to online payment to recognize malicious events and explain the logic behind the decisions. The cybersecurity field use DL models in most of its work. for instance, Usama et al [85] examine and explain the robustness of Self Organizing Networks (SON) against adversarial attacks example through DNN. Yoon et al. [313] Analyze the propagation of adversarial attacks from the standpoint of XAI, which investigates the progress of adversarial perturbations using CNN designs. Marino et al. [185] develop justifications for data-driven intrusion detection systems' erroneous categorization through visualizing the most relevant features that explain the reason for the misclassification [34]. This field mostly uses rule extraction and interpretable and intrinsic

Table 9
Overview of XAI studies proposed for healthcare different tasks, classified by the type of applied explainability technique.

Ref	Explainability technique	Problem addressed	L/G	S/A	I/P
XAI based on attention techniques					
[161]	Attention	Cardiac failure forecasting	G, L	S	I
[320]	Attention	potential hospitalization forecast from EHR	L	S	I
[58]	Attention	Cardiac failure forecasting	L	S	I
[136]	Attention	Forecasting medical happenings in ICU	G, L	S	I
[264]	Attention	Evaluation of successive organ failure in-hospital death	G, L	S	I
[321]	Attention	Forecasting the site of HIV genome blending	L	S	I
[131]	MLCAM	Localization of brain tumors	L	S	I
[325]	Respond-CAM	Macro-molecular complexes	L	S	I
[64]	Super-pixel maps	Classification of histological tumor subclasses	L	S	I
[168]	CAM	Detecting acute intracranial hemorrhage	L	S	I
[147]	CAM	Analyzing breast neoplasm ultrasonography	L	S	I
[234]	Grad-CAM	Appendicitis classification	L	S	I
[224]	Grad-CAM	Detecting hypoglycemia from ECG	L	S	I
[125]	Multiscale CAM	Diagnosis of COVID-19	L	S	I
XAI based on dimension reduction methods					
[108]	Sparse deep learning	Forecasting longstanding survival for glioblastoma multiforme	G	A	I
[322]	Optimal feature selection	Estimating the side effects of drugs	G	S	I
[305]	Laplacian Eigenmaps	Classifying brain tumors from MRS	G	S	I
[150]	Ideal feature selection	Forecasting cell-type given enhancers	G	A	I
[327]	Cluster analysis and LASSO	Delamination of lung cancer cases	G	A	I
[25]	Sparse-balanced SVM	Classification of type 2 diabetes	G	A	I
XAI based on feature importance methods					
[79]	Feature marginalisation	Diagnosing of inflammatory bowel and skin microbiota diseases	G, L	A	P
[91]	Feature weighting	Forecasting ICU deaths	G	A	P
[330]	DeepLIFT	Detecting splice site	G	A	P
[275]	Shapley value	Diagnosis of prostate cancer	G, L	A	P
[267]	DeepLIFT and others	Diagnosis of ophthalmic	G, L	A	P
XAI based on knowledge distillation					
[48]	Rule-based system	Forecasting the pneumonia threat	G	S	I
[300]	Complex relationships distilling	Forecasting the readmission-caused heart failure in hospital	G	S	I
[54]	Mimic learning	Forecasting outcomes of ICU for ALJ	G	S	I
[193]	Visualisation of rules	Classification of diabetes as well as breast cancer	G	S	I
[69]	Fuzzy rules	Forecasting death in hospitals	G	S	I
[167]	Visual/textual justification	Classifying breast mass	G	S	I
[169]	Bayesian rule lists	Stroke forecasting	G	S	I
[226]	Decision rules		G	S	I
XAI based on surrogate models					
[212]	LIME	Prediction of main bright puberty	L	A	P
[156]	LIME	Building survival models	L	A	P
[213]	LIME like with rule based XAI	Forecasting patient diagnosis, medications, and readmission	L	A	P
[166]	LRP	Forecasting acute serious disease from EHR	L	A	P
[93]	LIME	Classifying autism spectrum disorders	L	A	P
[188]	LIME	Segmenting lung lesion	L	A	P

approaches like Decision Trees, rule classifiers, and association rules. Also, visualization explanation techniques have a good part in Cybersecurity [85].

6.5. Finance

Transparent ML models can be automatically extracting earmarks from congressional bills and reports that allow us to extract earmarks cheaply, reliably, and consistently from historical congressional documents. Chou [60] presents a practical grafting approach that proposed a hybrid model of local with global approaches through explainable AI of the decision tree model for predicting corporate financial distress. Han et al. [106] through CNN architectures, present the XAI technique to display areas to banknote and counterfeit detection. In financial services, DNN like in [106], Association Rules, fuzzy rule-based, and Rule-based classifications are mostly used like in [247]. So interpretable and intrinsic models and rule extraction methods are used widely in this field, and also human-grounded and function-grounded evaluations are mostly used.

6.6. Smart transportation

One of the key features of driver assistance systems, autonomous vehicles, and self-driving automobiles is re- liable perception and detection of objects, thus they must be explainable, offering rationales for their actions [208]. The rapid growth of unmanned traffic is anticipated to introduce new challenges and have a significant effect on traffic management and clear implications for both social beings and the architecture necessary to allow highly automated and reliable unmanned oper-

ations [151]. Many state and municipal agencies are presently confronted with problems in collecting and estimating traffic volumes, particularly in collecting annual average daily traffic (AADT) on low-volume routes. To address these obstacles, a novel interpretable ML framework is required. Predicting daily transportation mode usage (auto, public transportation, or active travel) and interpreting commonly utilized travel modes are critical tasks in transportation research. DL techniques have proven to be effective in the identification and categorization of objects from pictures and transportation Systems, frequently outperforming humans, So the DL models are widely used and have the most articles in this field [171]. Visualization techniques are mostly used to achieve explainability as with [326].

A demonstrative subset of earlier studies where explanation methods were mainly debated or applied in the field of smart transportation is displayed in Table 10. It is worth noting that attention maps and heatmaps are frequently considered as explanations in ML studies; however, they may be ineffective in some cases. Thus, this section incorporates related works on heatmaps and attention maps for smart transportation as reviewed in Table 10. It could be seen that the XAI can be investigated for transportation applications according to six criteria defined in our previously debated taxonomies. It is also observable that some categories of explanations such as global explanations contrastive, demographics, sensitivity, model-agnostic, and counterfactual, are uncommon in the literature on smart transportation. This could be attributed to the burgeoning nature of the explainable intelligent transportation systems.

6.7. Education

Intelligent Tutoring System (ITS) is primarily used to teach students about its underlying user modeling approaches, and to explore student attitudes toward incorporating explanations to ITSs [228]. The tutor assists students in reviewing their exercise, eliciting where and how they might have done better, and employing XAI to allow students to ask questions regarding the virtual human's actions. The behavior assessment system for children is a norm-referenced diagnostic instrument for assessing the behavior and self-perceptions of children and young adults aged 3 to 18 years [134,297]. Because there is a growing request for sign language learners, an explainable ML system can give effective and relevant input on position, movement, and hand shape to sign language users [214]. Robots can learn from human instructors, and the openness of the machine's internal state might enhance the learning process by informing teachers of what is established and what is unknown [256]. Rule-based explanations and interpretable models are widely used in the education domain such as Naive Bayes and rule-based classification and DL models also play a major role in this domain [214].

7. Explainable AI: Challenges & future directions

This section uncovers the up-to-date challenges and future trends for improving the quality and transparency of AI solutions in real-world applications.

7.1. Standardization and formalization

Although XAI has been studied from different perspectives, the research community still lacks a standardized definition of explainability. It is vital to take on a general base that researchers and technicians must follow to develop and design XAI systems and solutions. The existing research can offer much preferential information about the explainability, it is difficult to evaluate them in a methodical way. This makes it challenging to accommodate explainability methods into one standardized framework. The design of such a framework will necessitate characterizing explainability from different aspects including data, model, learning strategy, outcomes, and gradient information.

7.2. Explainability-efficiency tradeoffs

XAI generally has the very little predictive capability or is stiff and computationally difficult, despite having predictions that are easy to understand and explain. Which of the subsequent should we use, then? White-Box models that are less accurate but simpler to grasp, or Black-Box models that are extremely accurate? The performance of XAI algorithms can unavoidably deteriorate under the explainability method, according to a thorough review of the XAI literature. In other words, performance suffers as explanation methods become more intricate. More study will be required to achieve an appropriate balance between performance and explainability as a result of this. The resource-constrained nature of IIoT devices places strict constraints on the XAI solution's complexity. Therefore, the research efforts should take into account easy-to-use explainability techniques.

7.3. Visual explanation

The debate above highlights some critical issues with XAI visualizations and interpretability methods. First, is the unwillingness of human attention to infer XAI explanation maps for use in judgment. Second is the lack of a quantifiable way to assess the explanatory map's precision and completeness. This shows that going forward, mission-critical applications' continued utilization of visualization approaches needs to be reevaluated. It needs to be thought about how to express and

Table 10
Overview of XAI studies that fully or partially argue the explanation categories for different smart transportation tasks: conversational (C), non-conversational (N), local (L), global (G), model-specific (S), model-agnostic (A), goal-driven (GD), data-driven (DD).

Ref	Topic covered	Causal			Style				C/N	L/G	S/A	GD/DD
		Factual	Contrastive	Counterfactual	Input Influence	Sensitivity	Case-based	Demographic				
[149]	Textual Explanations of driving	✓	×	×	×	×	✓	×	N	L	A	DD
[51]	beyond explanation as soliloquy	✓	×	×	✓	×	×	×	N	G	A	GD
[235]	Explaining Impracticable Robot	✓	×	×	✓	×	×	×	N	G,	A	GD
[303]	Explaining autonomous action	✓	×	×	×	×	✓	×	N	G, L	A	DD
[148]	Interpretable delf-driving	✓	×	✓	×	×	✓	×	N	L	A	DD
[66]	Visual attention of self-driving	✓	×	×	✓	×	×	×	N	L	A	DD
[251]	autonomous driving	✓	×	×	✓	×	×	×	N	L	A	DD
[230]	Context aware estimation	×	×	×	✓	×	×	×	N	L	A	DD
[262]	necessity of explanations	✓	×	×	×	×	✓	×	N	L	A	DD
[22]	multi-level explanation	✓	×	×	×	×	✓	×	N	L	A	DD
[202]	explaining collision risk	✓	×	✓	✓	×	×	×	C	L	S	DD
[105]	autonomous vehicles	✓	×	×	✓	×	×	×	N	L, G	A	GD
[155]	Explanation of semi-autonomous driving	✓	×	×	✓	×	×	×	N	L	A	GD
[31]	Visualization of CNN based driving	✓	×	×	×	×	✓	×	N	G	A	DD
[200]	Attention for explaining self-driving	✓	×	×	✓	×	×	×	N	G	A	DD
[177]	Hazard detection failure	✓	×	×	✓	×	×	×	N	G	A	GD
[209]	Explaining driver perceptions	✓	✓	✓	✓	×	×	×	N	L, G	A	GD
[242]	Explaining traffic control signals	✓	×	×	✓	×	×	×	N	G	A	DD
[178]	Attention to interpreting driving behavior	✓	×	×	✓	×	×	×	N	L	A	DD
[211]	Explaining autonomous driving	✓	✓	✓	✓	×	×	×	N	L, G	A	GD

explain in the best manner. Weerts et al.[293], for instance, investigated the effect of SHAP explanations on enhancing the human ability to alert computational requirements. The authors conducted a human-centered study to see whether giving decision justification could make particular decision-making scenarios better. The findings revealed that adding more SHAP explanations to the class output likelihood had no positive impact on people's capacity to make decisions. Authors noticed a greater interest in the final grade after deciding that might result in disastrous consequences in mission-critical situations. In a comparable way, the authors of [196] introduced a human-grounded evaluation baseline in [111] and assessed the efficacy of the LIME technique by contrasting the explanation map produced by LIME with an explanatory map that was evaluated according to Ten human annotations. According to the findings, LIME generates certain extrapolations that are unrelated to humans explanations, which leads to a lower explanation quality than the balanced explanation map produced by human annotators. This clarifies the significance of comprehending the different types of explanations, such as application-grounded, human-grounded, and functionally grounded interpretations, in order to enhance explanatory maps by using relevant data produced by people, putting more restrictions on interpretations, or trying to introduce a detailed definition of interpretations to the optimization process.

7.4. Explainability metrics

An important aspect that directly deeply relates to explainability is the presence of evaluation metrics. One or more evaluation metrics must permit a profound assessment of the ability of the XAI algorithm to fits the meaning of explainability. Without these metrics, any assumption in this regard weakens the literature, not offering a consistent base on which to depend. These metrics should communicate how well the AI algorithm behaves in a particular facet of explainability.

Unluckily, findings obtained from the literature demonstrate that more measurable, typical XAI metrics are certainly required to provide for the present measurement XAI methods and tools. It is advocated to dedicate more efforts towards designing new metrics to assess and comparatively analyze XAI algorithms in various settings, application contexts, data, and models.

7.5. Explainable security

With the ever-increasing improvements in the digital world, systems, and software are usually prone to a wide spectrum of security threats and cyber-attacks. AI algorithms have been shown as a better solution for detecting these cyberattacks. However, almost all AI-based security solution gives Black-Box decisions that may be uninterpretable even for security specialist. Hence, advancing toward explaining security solutions is anticipated as a promising direction for improving the security of public and private systems. Nevertheless, the information uncovered by XAI algorithms may be exploited to either create adversarial attacks to confuse the AI algorithm or develop methods to better defend against confidential content disclosure. Moreover, generative networks can be exploited to engender counterfactuals, for enabling the individuals to interpret the capabilities and limits of the AI algorithm by taking into account his/her enhanced responsibility and knowledgeable critique. Considering this latest trend, this work absolutely thinks that there is a path forward for generative networks to participate in situations requiring comprehensible machine decisions.

7.6. Privacy preservation

The digital world is known to contain a large amount of multi-site data that entails privacy concerns during storage or transmission. This comes to be specifically crucial when AI algorithms deal with private data as people's privacy rights must be addressed. Since XAI algorithms are designed to be interpretable for humans (White-Box), which poses serious concerns about the privacy-explainability trade-offs. This in turn advocates more investigation on this point, particularly, in confirming that XAI algorithms do not threaten the privacy of the data during the training or inference. Also, it is advocated to design a new method for explaining the loss of privacy by providing a subjective metric of the seriousness of the privacy breach based on the training behavior.

7.7. Structural complexity

Another ML explainability challenge is the Systematic instability that complicates automated generated explanations in the multiplicity of good models, because of the complicated structure of ML models, Complex ML algorithms can generate numerous correct approaches for the same number of input and forecast objectives by following extremely similar but not identical internal network pathways, thus specifics of explanations might differ among multiple accurate models [186].

7.8. Deep learning

Recent advances in post-hoc explainability methods have highlighted the exchange between forecasting and description associated with deep networks—these methodologies approximate deep opaque models with simpler interpretable approaches that can be inspected to clarify the Opaque approaches. Because they turn Opaque models into transparent models, these techniques are referred to as XAI. They are becoming increasingly popular because they enable AI applications to achieve both prediction accuracy and interpretability requirements [231].

7.9. Uncertainty and confidence estimation

Some AI applications are known to have critical outputs that can threaten the safety of humans i.e., automatic surgery, self-driving vehicles, data-driven medical diagnosis, insurance risk assessment, etc. In all these circumstances, the incorrect output of algorithms can result in a damaging outcome, which has generated thorough supervisory efforts intended at guaranteeing that decisions are not taken exclusively based on data handling. This motivates the research to reduce the uncertainty of damages resulting from AI decisions. This way the epistemic uncertainty can be estimated for training data and also the confidence of the model's output can be estimated to avoid erroneous decisions. Fuzzy theory can be integrated with XAI to offer a solution for this dilemma.

7.10. Tabular data

Though of the presence of different categories of XAI methods, they are still inadequate for *tabular data*, where all samples share the same feature space. Essentially, XAI designed for visual or textual data cannot easily be used for tabular data. This is attributed to the fact that tabular information has distinctive attributes such as possible interdependencies and relationships between the features, the existence of both constant and definite characteristics, and the chronological view of the data. Hence, it is imperative to evidently differentiate the methods that are consistent with tabular data.

7.11. Causality

Intelligent systems that are context-aware use absolute inputs and carry out judgments according to sophisticated regulations and AI solutions that are difficult to understand by humans. The absence of a smart system to generate causative explanation could result in losing the clients' satisfaction and confidence which can make the XAI systems unqualifiable critical applications. Nevertheless, automatic generation of causative explanations for the AI decision can assist alleviate this issue. An important research direction challenge is a call to address the issue of fidelity (or causality) more critically and to ask hard questions about whether a claimed explanation is faithfully explaining the model's prediction.

7.12. User interface

The human experts still face obstacles with a gap between the explanations obtained from machines and those obtained from humans. They shed the light on an interdisciplinary research area that not merely explains outcomes of AI algorithms in a transparent way but derives insight from the practice humans explain beliefs to one another. This advocates the research community toward a question-guided XAI interface that helps satisfy the needs of domain specialists. Another direction for improving the quality of the explanation interface is to consider the human-in-loop strategy.

8. Responsible AI

In recent years, the multiplicity of private and public institutions, companies, and organizations have been devoting many efforts to establish and settle the principles and regulations to imply in what way AI has to be developed, implemented, and applied in our world. These regulations are popularly identified as AI *principles,* which are proposed to solve the issues associated with possible AI risks on persons, organizations, and the overall public community. All these efforts seek to make the development of AI more responsible in practice. For example, Benjamins et al [23] described five principles for AI development including Human-centric AI, Fairness of AI, Privacy and Secure design, Transparent and Explainable AI, and extensibility to any third party. In another way, the European Commission has just announced ethical regulations for Trustworthy AI [44] by evaluating predefined checklists based on different AI systems stakeholders. This evaluation is established on six principles namely 1) diversity, transparency, non-bias, and fairness; 2) human agency and supervision; 3) technological toughness and safety; 4) accountability; 5) privacy and data governance; and 6) social and environmental welfare. Some organizations like IBM, Google, Microsoft have released their principles for responsible AI. Motivated by these efforts, this section study, analyze, and associate different principles from academic and technical viewpoint to carefully define the new pillars for developing responsible AI solutions. Based on the findings obtained from the reviewed studies' previous sections, it can be stated with confidence that shows that XAI is enough to achieve responsible AI solutions in practical and other complementary aspects should be addressed to achieve practical applicability of XAI. In this regard, this work defines eight pillars of responsible AI which include 1) explainability; 2) human-in loop; 3) bias evaluation; 4) reproducibility; 5) data risk sensibility; 6) privacy & trust; 7) technical performance; and 8) displacement strategy (see Fig. 20. Each of these pillars represents a great challenge for the research community to achieve responsible AI and exhibits a considerable tradeoff with other pillars, which pave the way for further investigation in such a multitude and interrelated requirements.

9. Conclusions

Through a systematic analysis of the literature of XAI, this survey studied widely a framework of explainability elements and structured them from multiple viewpoints, as well as mapped the research space from varied fields linked to explainable systems. This survey presented a fine-grained and multilevel taxonomy of XAI methods that takes into account all aspects and criteria of classifying the existing XAI methods. The taxonomy also extends to categorizing the different metrics for evaluating the explanations generated by different methods. To promote the applicability of XAI, this work charts out the state-of-the-art application domains that show a growing necessity to interpret AI decisions. Finally, we presented a road map of responsible AI to promote further research in this promising emergent subfield of AI. Overall, this work is meant to be the most comprehensive in terms of the XAI methods, metrics, applications, and challenges it covers and the most detailed in terms of the multi-level taxonomy it offers.

CRediT authorship contribution statement

Weiping Ding: Conceptualization, Methodology, Supervision, Writing - review & editing. **Mohamed Abdel-Basset:** Conceptualization, Methodology, Writing - original draft, Validation. **Hossam Hawash:** Methodology, Writing - original draft, Writing - review & editing. **Ahmed M. Ali:** Writing - original draft, Writing - review & editing.

Data availability

No data was used for the research described in the article.

Declaration of Competing Interest

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

Acknowledgment

The authors would like to express our sincere appreciation to the editor and anonymous reviewers for their insightful comments, which greatly improved the quality of this paper.

References

[1] A. Abujabal, R.S. Roy, M. Yahya, G. Weikum, Quint: Interpretable question answering over knowledge bases, in EMNLP 2017 - Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2017.
[2] A. Adadi, M. Berrada, Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI), IEEE Access 6 (2018) (2018) 52138–52160, https://doi.org/10.1109/ACCESS.2018.2870052.
[3] Neda AfzaliSeresht, Qing Liu, and Yuan Miao. 2019. An Explainable Intelligence Model for Security Event Analysis. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 315–327. 10.1007/978-3-030-35288-2_26.
[4] Rishabh Agarwal, Levi Melnick, Nicholas Frosst, Xuezhou Zhang, Ben Lengerich, Rich Caruana, and Geoffrey Hinton. 2020. Neural Additive Models: Interpretable Machine Learning with Neural Nets. arXiv Prepr. arXiv2004.13912 (2020). Retrieved from http://arxiv.org/abs/2004.13912.
[5] Fatemeh Aghaeipoor, Mohammad Masoud Javidi, and Alberto Fernandez. 2021. IFC-BD: An Interpretable Fuzzy Classifier for Boosting Explainable Artificial Intelligence in Big Data. IEEE Trans. Fuzzy Syst. (2021). 10.1109/TFUZZ.2021.3049911.
[6] X. Alameda-Pineda, M. Redi, E. Celis, N. Sebe, S.F. Chang, FAT/MM'19: 1st international workshop on fairness, accountability, and transparency in multimedia, in: In MM 2019 - ProceedIngs of the 27th ACM International Conference on Multimedia, 2019, pp. 2728–2729, https://doi.org/10.1145/3343031.3350555.
[7] Emanuele Albini, Antonio Rago, Pietro Baroni, and Francesca Toni. 2020. Relation-based counterfactual explanations for Bayesian network classifiers. In IJCAI International Joint Conference on Artificial Intelligence, 451–457. 10.24963/ijcai.2020/63.
[8] Syed Imran Ali and Sungyoung Lee. 2020. Ensemble based Cost-Sensitive Feature Selection for Consolidated Knowledge Base Creation. In Proceedings of the 2020 14th International Conference on Ubiquitous Information Management and Communication, IMCOM 2020. 10.1109/IMCOM48794.2020.9001751.
[9] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. Concrete Problems in AI Safety. Lect. Notes Comput. Sci. (2016). Retrieved from http://arxiv.org/abs/1606.06565.
[10] Sule Anjomshoae, Kary Främling, and Amro Najjar. 2019. Explanations of Black-Box model predictions by contextual importance and utility. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 95–109. 10.1007/978-3-030-30391-4_6.
[11] S. Anjomshoae, T. Kampik, K. Främling, Py-CIU: A Python Library for Explaining Machine Learning Predictions Using Contextual Importance and Utility, In IJCAI-PRICAI 2020 Workshop on ExplaInable Artificial Intelligence (XAI), 2020.
[12] Vijay Arya, Rachel K. E. Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Q. Vera Liao, Ronny Luss, Aleksandra Mojsilovic, Sami Mourad, Pablo Pedemonte, Ramya Raghavendra, John Richards, Prasanna Sattigeri, Karthikeyan Shanmugam, Moninder Singh, Kush R. Varshney, Dennis Wei, and Yunfeng Zhang. 2021. AI Explainability 360: Impact and Design. (September 2021). Retrieved from http://arxiv.org/abs/2109.12151.
[13] Vijay Arya, Rachel K. E. Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Q. Vera Liao, Ronny Luss, Aleksandra Mojsilović, Sami Mourad, Pablo Pedemonte, Ramya Raghavendra, John Richards, Prasanna Sattigeri, Karthikeyan Shanmugam, Moninder Singh, Kush R. Varshney, Dennis Wei, and Yunfeng Zhang. 2019. One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques. (September 2019). Retrieved from http://arxiv.org/abs/1909.03012.
[14] Made Widhi Surya Atman, Julian Hay, Junya Yamauchi, Takeshi Hatanaka, and Masayuki Fujita. 2018. Two variations of passivity-short-based semi-autonomous robotic swarms. In SICE ISCS 2018 - 2018 SICE International Symposium on Control Systems, 12–19. 10.23919/SICEISCS.2018.8330150.
[15] Anonymous Authors. 2021. Hard Masking for Explaining Graph Neural Networks. Iclr 2021 (2021), 1–12.
[16] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus Robert Müller, and Wojciech Samek. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS One 10, 7 (2015), (e0130140) 1–46. 10.1371/journal.pone.0130140.
[17] X. Bai, X. Wang, X. Liu, Q. Liu, J. Song, N. Sebe, B. Kim, Explainable deep learning for efficient and robust pattern recognition: A survey of recent developments, Pattern Recognition 120 (2021), https://doi.org/10.1016/j.patcog.2021.108102.
[18] Federico Baldassarre and Hossein Azizpour. 2019. Explainability Techniques for Graph Convolutional Networks. arXiv Prepr. arXiv1905.13686 (2019). Retrieved from http://arxiv.org/abs/1905.13686.
[19] H. Baniecki, W. Kretowicz, P. Piatyszek, J. Wisniewski, P. Biecek, dalex: Responsible machine learning with interactive explainability and fairness in python, J. Mach. Learn. Res. 22 (2021) (2021) 1–7.
[20] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S Lasecki, Daniel S Weld, and Eric Horvitz. 2019. Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance. Proc. AAAI Conf. Hum. Comput. Crowdsourcing 7, 1 (2019), 19. Retrieved from www.aaai.org.
[21] A.B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, F. Herrera, Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, Inf. Fusion 58 (2020) (2020) 82–115, https://doi.org/10.1016/j.inffus.2019.12.012.
[22] H. Ben-Younes, É. Zablocki, P. Pérez, M. Cord, Driving behavior explanation with multi-level fusion, Pattern Recognit. 123 (2022) (2022), https://doi.org/10.1016/j.patcog.2021.108421 108421.
[23] Richard Benjamins, Alberto Barbado, and Daniel Sierra. 2019. Responsible AI by Design in Practice. (September 2019). Retrieved from http://arxiv.org/abs/1909.12838.
[24] S. Berkovsky, R. Taib, D. Conway, How to recommend? User trust factors in movie recommender systems, In International Conference on Intelligent User Interfaces, Proceedings IUI 287–300 (2017), https://doi.org/10.1145/3025171.3025209.
[25] Michele Bernardini, Luca Romeo, Paolo Misericordia, and Emanuele Frontoni. 2020. Discovering the Type 2 Diabetes in Electronic Health Records Using the Sparse Balanced Support Vector Machine. IEEE J. Biomed. Heal. Informatics 24, 1 (2020), 235–246. 10.1109/JBHI.2019.2899218.
[26] Przemyslaw Biecek. 2018. Dalex: Explainers for complex predictive models in R. J. Mach. Learn. Res. 19, (2018). 10.5281/zenodo.3670940.
[27] Przemyslaw Biecek and Tomasz Burzykowski. 2021. Explanatory Model Analysis. 10.1201/9780429027192.
[28] Jacob Bien and Robert Tibshirani. 2011. Prototype selection for interpretable classification. Ann. Appl. Stat. 5, 4 (2011), 2403–2424. 10.1214/11-AOAS495.
[29] Mustafa Bilgic and Raymond J Mooney. 2005. Explaining Recommendations: Satisfaction vs. Promotion. Proc. Beyond Pers. 2005 A Work. Next Stage Recomm. Syst. Res. 2005 Int. Conf. Intell. User Interfaces (2005).

[30] Lieven Billiet, Sabine Van Huffel, and Vanya Van Belle. 2018. Interval Coded Scoring: A toolbox for interpretable scoring systems. PeerJ Comput. Sci. 2018, 4 (2018), (e150) 1–28. 10.7717/peerj-cs.150.
[31] M. Bojarski, A. Choromanska, K. Choromanski, B. Firner, L.J. Ackel, U. Muller, P. Yeres, K. Zieba, VisualBackProp: Efficient visualization of CNNs for autonomous driving, In Proceedings - IEEE International Conference on Robotics and Automation 4701–4708 (2018), https://doi.org/10.1109/ICRA.2018.8461053.
[32] Tiago Botari, Rafael Izbicki, and Andre C.P.L.F. de Carvalho. 2020. Local interpretation methods to machine learning using the domain of the feature space. Commun. Comput. Inf. Sci. 1167 CCIS, (2020), 241–252. 10.1007/978-3-030-43823-4_21.
[33] O. Boz, Extracting decision trees from trained neural networks, in: In ProceedIngs of the ACM SIGKDD International Conference on Knowledge Discovery and Data MInIng, 2002, pp. 456–461, https://doi.org/10.1145/775107.775113.
[34] W. Briguglio, S. Saad, Interpreting Machine Learning Malware Detectors Which Leverage N-gram Analysis, In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 82–97 (2020), https://doi.org/10.1007/978-3-030-45371-8_6.
[35] Theodora S. Brisimi, Tingting Xu, Taiyao Wang, Wuyang Dai, William G. Adams, and Ioannis Ch Paschalidis. 2018. Predicting Chronic Disease Hospitalizations from Electronic Health Records: An Interpretable Classification Approach. Proc. IEEE 106, 4 (2018), 690–707. 10.1109/JPROC.2017.2789319.
[36] Maja Brkan and Grégory Bonnet. 2020. Legal and Technical Feasibility of the GDPR's Quest for Explanation of Algorithmic Decisions: Of Black Boxes, White Boxes and Fata Morganas. Eur. J. Risk Regul. 11, 1 (2020), 18–50. 10.1017/err.2020.10.
[37] C.B. Browne, E. Powley, D. Whitehouse, S.M. Lucas, P.I. Cowling, P. Rohlfshagen, S. Tavener, D. Perez, S. Samothrakis, S. Colton, A survey of Monte Carlo tree search methods, IEEE Transactions on Computational Intelligence and AI in Games 4 (2012) 1–43, https://doi.org/10.1109/TCIAIG.2012.2186810.
[38] N.J. Bryan, G.J. Mysore, An efficient posterior regularized latent variable model for interactive sound source separation, In 30th International Conference on MachIne LearnIng, 2013.
[39] Z. Buçinca, P. Lin, K.Z. Gajos, E.L. Glassman, Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems, In International Conference on Intelligent User Interfaces, Proceedings IUI 454–464 (2020), https://doi.org/10.1145/3377325.3377498.
[40] A. Bunt, M. Lount, C. Lauzon, Are explanations always important? A study of deployed, low-cost intelligent interactive systems, In International Conference on Intelligent User Interfaces, Proceedings IUI 169–178 (2012), https://doi.org/10.1145/2166966.2166996.
[41] A. Bussone, S. Stumpf, D. O'Sullivan, The role of explanations on trust and reliance in clinical decision support systems, in: In ProceedIngs - 2015 IEEE International Conference on Healthcare Informatics, 2015, https://doi.org/10.1109/ICHI.2015.26.
[42] L.A. Bygrave, Article 22 Automated individual decision-making, including profiling, In The EU General Data Protection Regulation (GDPR) (2020), https://doi.org/10.1093/oso/9780198826491.003.0055.
[43] Béatrice Cahour and Jean François Forzy. 2009. Does projection into use improve trust and exploration? An example with a cruise control system. Saf. Sci. 47, 9 (2009), 1260–1270. 10.1016/j.ssci.2009.03.015.
[44] M. Cannarsa, Ethics Guidelines for Trustworthy AI, In The Cambridge Handbook of Lawyering in the Digital Age. 283–297 (2021), https://doi.org/10.1017/9781108936040.022.
[45] A. Cano, B. Krawczyk, Evolving rule-based classifiers with genetic programming on GPUs for drifting data streams, Pattern Recognit. 87 (2019) (2019) 248–268, https://doi.org/10.1016/j.patcog.2018.10.024.
[46] Bin Cao, Jianwei Zhao, Xin Liu, Jaroslaw Arabas, Mohammad Tanveer, Amit Kumar Singh, and Zhihan Lv. 2022. Multiobjective Evolution of the Explainable Fuzzy Rough Neural Network with Gene Expression Programming. IEEE Trans. Fuzzy Syst. (2022), 1–1. 10.1109/tfuzz.2022.3141761.
[47] André M. Carrington, Paul W. Fieguth, Hammad Qazi, Andreas Holzinger, Helen H. Chen, Franz Mayr and Douglas G. Manuel. 2020. A new concordant partial AUC and partial c statistic for imbalanced data in the evaluation of machine learning algorithms. BMC Medical Informatics and Decision Making (2020) 20: 4. 10.1186/s12911-019-1014-6.
[48] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, N. Elhadad, Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission, In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 1721–1730 (2015), https://doi.org/10.1145/2783258.2788613.
[49] Diogo V. Carvalho, Eduardo M. Pereira, and Jaime S. Cardoso. 2019. Machine learning interpretability: A survey on methods and metrics. Electron. 8, 8 (2019), 832. 10.3390/electronics8080832.
[50] Alvaro E. Ulloa Cerna, Marios Pattichis, David P. vanMaanen, Linyuan Jing, Aalpen A. Patel, Joshua V. Stough, Christopher M. Haggerty, and Brandon K. Fornwalt. 2019. Interpretable Neural Networks for Predicting Mortality Risk using Multi-modal Electronic Health Records. (January 2019). Retrieved from http://arxiv.org/abs/1901.08125.
[51] Tathagata Chakraborti, Sarath Sreedharan, Yu Zhang, and Subbarao Kambhampati. 2017. Plan explanations as model reconciliation: Moving beyond explanation as soliloquy. In IJCAI International Joint Conference on Artificial Intelligence, 156–163. 10.24963/ijcai.2017/23.
[52] R. Chandra, R.P. Rocha, Knowledge-Based Framework for Human-Robots Collaborative Context Awareness in USAR Missions, in: In ProceedIngs - 2016 International Conference on Autonomous Robot Systems and Competitions, 2016, https://doi.org/10.1109/ICARSC.2016.50.
[53] A. Chattopadhay, A. Sarkar, P. Howlader, V.N. Balasubramanian, Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks, in: In ProceedIngs - 2018 IEEE WInter Conference on Applications of Computer Vision, 2018, https://doi.org/10.1109/WACV.2018.00097.
[54] Z. Che, S. Purushotham, R. Khemani, Y. Liu, Interpretable Deep Models for ICU Outcome Prediction, AMIA Annu. Symp. proceedings. AMIA Symp. 2016 (2016) (2016) 371–380.
[55] Chaofan Chen, Oscar Li, Chaofan Tao, Alina Jade Barnett, Jonathan Su, and Cynthia Rudin. 2019. This looks like that: Deep learning for interpretable image recognition. In Advances in Neural Information Processing Systems.
[56] Jianbo Chen, Le Song, Martin J. Wainwright, and Michael I. Jordan. 2018. Learning to explain: An information-theoretic perspective on model interpretation. In 35th International Conference on Machine Learning, ICML 2018, 1386–1418.
[57] Rishi Chhatwal, Peter Gronvall, Nathaniel Huber-Fliflet, Robert Keeling, Jianping Zhang, and Haozhen Zhao. 2019. Explainable Text Classification in Legal Document Review A Case Study of Explainable Predictive Coding. In Proceedings - 2018 IEEE International Conference on Big Data, Big Data 2018, 1905–1911. 10.1109/BigData.2018.8622073.
[58] E. Choi, M.T. Bahadori, J.A. Kulas, A. Schuetz, W.F. Stewart, J. Sun, RETAIN: An interpretable predictive model for healthcare using reverse time attention mechanism, in: in Advances in Neural Information Processing Systems, 2016, pp. 3512–3520.
[59] Jaegul Choo, Hanseung Lee, Jaeyeon Kihm, and Haesun Park. 2010. iVisClassifier: An interactive visual analytics system for classification based on supervised dimension reduction. In VAST 10 - IEEE Conference on Visual Analytics Science and Technology 2010, Proceedings, 27–34. 10.1109/VAST.2010.5652443.
[60] Tsung Nan Chou. 2019. A Practical Grafting Model Based Explainable AI for Predicting Corporate Financial Distress. In Lecture Notes in Business Information Processing, 5–15. 10.1007/978-3-030-36691-9_1.
[61] Yu Liang Chou, Catarina Moreira, Peter Bruza, Chun Ouyang, and Joaquim Jorge. 2022. Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms, and applications. Inf. Fusion 81, (May 2022), 59–83. 10.1016/j.inffus.2021.11.003.
[62] William J. Clancey and Robert R. Hoffman. 2021. Methods and standards for research on explainable artificial intelligence: Lessons from intelligent tutoring systems. Appl. AI Lett. 2, 4 (2021). 10.1002/ail2.53.
[63] S. Coppers, J. Van Den Bergh, K. Luyten, K. Coninx, I. Van Der Lek-Ciudin, T. Vanallemeersch, V. Vandeghinste, Intellingo: An intelligible translation environment, In Conference on Human Factors in Computing Systems - Proceedings (2018), https://doi.org/10.1145/3173574.3174098.
[64] H.D. Couture, J.S. Marron, C.M. Perou, M.A. Troester, M. Niethammer, Multiple instance learning for heterogeneous images: Training a CNN for histopathology, In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 254–262 (2018), https://doi.org/10.1007/978-3-030-00934-2_29.

[65] Xiaocong Cui, Jung min Lee, and J. Po-An Hsieh. 2019. An integrative 3C evaluation framework for explainable artificial intelligence. 25th Am. Conf. Inf. Syst. AMCIS 2019 (2019).
[66] L. Cultrera, L. Seidenari, F. Becattini, P. Pala, A. Del Bimbo, Explaining autonomous driving by learning end-to-end visual attention, In IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops 1389–1398 (2020), https://doi.org/10.1109/CVPRW50498.2020.00178.
[67] Arun Das and Paul Rad. 2020. Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey. (June 2020). Retrieved from http://arxiv.org/abs/2006.11371.
[68] D. Das, S. Chernova, Leveraging rationales to improve human task performance, In International Conference on Intelligent User Interfaces, Proceedings IUI 510–518 (2020), https://doi.org/10.1145/3377325.3377512.
[69] Raheleh Davoodi, Mohammad Hassan Moradi. 2018. Mortality prediction in intensive care units (ICUs) using a deep rule-based fuzzy classifier. J. Biomed. Inform. 79, (2018), 48–59. 10.1016/j.jbi.2018.02.008.
[70] R. Dazeley, P. Vamplew, F. Cruz, 2021, A Conceptual Framework and Survey, Explainable Reinforcement Learning for Broad-XAI, August 2021, Retrieved from http://arxiv.org/abs/2108.09003.
[71] Ashley Deeks. 2019. The judicial demand for explainable artificial intelligence. Columbia Law Rev. 119, 7 (2019), 1829–1850.
[72] K. Dembczyński, W. Kotłowski, R. Słowiński, Maximum likelihood rule ensembles, in: In ProceedIngs of the 25th International Conference on MachIne LearnIng, 2008, pp. 224–231, https://doi.org/10.1145/1390156.1390185.
[73] Amit Dhurandhar, Pin Yu Chen, Ronny Luss, Chun Chen Tu, Paishun Ting, Karthikeyan Shanmugam, and Payel Das. 2018. Explanations based on the Missing: Towards contrastive explanations with pertinent negatives. In Advances in Neural Information Processing Systems, 592–603.
[74] J. Dieber, S. Kirrane, A novel model usability evaluation framework (MUsE) for explainable artificial intelligence, Inf. Fusion 81 (2022) (2022) 143–153, https://doi.org/10.1016/j.inffus.2021.11.017.
[75] Finale Doshi-Velez and Been Kim. 2017. Towards A Rigorous Science of Interpretable Machine Learning. arXiv Prepr. arXiv1702.08608 (2017). Retrieved from http://arxiv.org/abs/1702.08608.
[76] Mengnan Du, Ninghao Liu, and Xia Hu. 2020. Techniques for interpretable machine learning. Commun. ACM 63, 1 (2020), 68–77. 10.1145/3359786.
[77] D. Duvenaud, D. Maclaurin, J. Aguilera-Iparraguirre, R. Gómez-Bombarelli, T. Hirzel, A. Aspuru-Guzik, R.P. Adams, Convolutional networks on graphs for learning molecular fingerprints, In Advances in Neural Information Processing Systems (2015) 2224–2232.
[78] Martin Ebers, Veronica R. S. Hoch, Frank Rosenkranz, Hannah Ruschemeier, and Björn Steinrötter. 2021. The European Commission's Proposal for an Artificial Intelligence Act—A Critical Assessment by Members of the Robotics and AI Law Society (RAILS). J (2021). 10.3390/j4040043.
[79] A. Eck, L. M. Zintgraf, E. F.J. de Groot, T. G.J. de Meij, T. S. Cohen, P. H.M. Savelkoul, M. Welling, and A. E. Budding. 2017. Interpretation of microbiota-based diagnostics by explaining individual classifier decisions. BMC Bioinformatics 18, 1 (2017). 10.1186/s12859-017-1843-1.
[80] M. Eiband, D. Buschek, A. Kremer, H. Hussmann, The impact of placebic explanations on trust in intelligent systems, In Conference on Human Factors in Computing Systems - Proceedings (2019), https://doi.org/10.1145/3290607.3312787.
[81] R. ElShawi, Y. Sherif, M. Al-Mallah, S. Sakr, ILIME: Local and Global Interpretable Model-Agnostic Explainer of Black-Box Decision, In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 53–68 (2019), https://doi.org/10.1007/978-3-030-28730-6_4.
[82] European Commission. 2021. Artificial Intelligence Act (2021) 206 final.
[83] F.u. Cheng Fan, C.Y. Xiao, C. Liu, Z. Li, J. Wang, A novel methodology to explain and evaluate data-driven building energy performance models based on interpretable machine learning, Appl. Energy 235 (2019) (2019) 1551–1560, https://doi.org/10.1016/j.apenergy.2018.11.081.
[84] Feng Lei Fan, Jinjun Xiong, Mengzhou Li, and Ge Wang. 2021. On Interpretability of Artificial Neural Networks: A Survey. IEEE Trans. Radiat. Plasma Med. Sci. 5, 6 (2021), 741–760. 10.1109/TRPMS.2021.3066428.
[85] Salah Ud Din Farooq, Muhammad Usama, Junaid Qadir, and Muhammad Ali Imran. 2019. Adversarial ML Attack on Self Organizing Cellular Networks. In 2019 UK/China Emerging Technologies, UCET 2019. 10.1109/UCET.2019.8881842.
[86] N. Fouladgar, M. Alirezaie, K. Främling, Exploring Contextual Importance and Utility in Explaining Affect Detection, In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 3–18 (2021), https://doi.org/10.1007/978-3-030-77091-4_1.
[87] Jerome H. Friedman and Bogdan E. Popescu. 2008. Predictive learning via rule ensembles. Ann. Appl. Stat. 2, 3 (2008), 916–954. 10.1214/07-AOAS148.
[88] F.u. Ruigang, H.u. Qingyong, X. Dong, Y. Guo, Y. Gao, B. Li, 2020, Towards Accurate Visualization and Explanation of CNNs, Axiom-based Grad-CAM, August 2020, Retrieved from http://arxiv.org/abs/2008.02312.
[89] Thorben Funke, Megha Khosla, and Avishek Anand. 2021. Zorro: Valid, Sparse, and Stable Explanations in Graph Neural Networks. (May 2021). Retrieved from http://arxiv.org/abs/2105.08621.
[90] Giuseppe Futia and Antonio Vetrò. 2020. On the integration of knowledge graphs into deep learning models for a more comprehensible AI-Three challenges for future research. Inf. 11, 2 (2020), (122) 1–10. 10.3390/info11020122.
[91] W. Ge, J.W. Huh, Y.R. Park, J.H. Lee, Y.H. Kim, A. Turchin, An Interpretable ICU Mortality Prediction Model Based on Logistic Regression and Recurrent Neural Networks with LSTM units, AMIA Annu. Symp. proceedings. AMIA Symp. 2018 (2018) (2018) 460–469.
[92] Fatih Gedikli, Dietmar Jannach, and Mouzhi Ge. 2014. How should i explain? A comparison of different explanation types for recommender systems. Int. J. Hum. Comput. Stud. 72, 4 (2014), 367–382. 10.1016/j.ijhcs.2013.12.007.
[93] Soudeh Ghafouri-Fard, Mohammad Taheri, Mir Davood Omrani, Amir Daaee, Hossein Mohammad-Rahimi, and Hosein Kazazi. 2019. Application of Single-Nucleotide Polymorphisms in the Diagnosis of Autism Spectrum Disorders: A Preliminary Study with Artificial Neural Networks. J. Mol. Neurosci. 68, 4 (2019), 515–521. 10.1007/s12031-019-01311-1.
[94] A. Ghorbani, J. Wexler, J. Zou, B. Kim, Towards automatic concept-based explanations, in Advances in Neural Information Processing Systems, 2019.
[95] M.L. Gonzalgo, A.J. Stephenson, I.M. Thompson, Causal Screening To Interpret Graph Neural networks, Education 2 (2011) (2011) 1–13.
[96] Yash Goyal, Amir Feder, Uri Shalit, and Been Kim. 2019. Explaining Classifiers with Causal Concept Effect (CaCE). arXiv Prepr. arXiv1907.07165 (2019). Retrieved from http://arxiv.org/abs/1907.07165.
[97] Ran Gu, Guotai Wang, Tao Song, Rui Huang, Michael Aertsen, Jan Deprest, Sebastien Ourselin, Tom Vercauteren, and Shaoting Zhang. 2021. CA-Net: Comprehensive Attention Convolutional Neural Networks for Explainable Medical Image Segmentation. IEEE Trans. Med. Imaging 40, 2 (2021), 699–711. 10.1109/TMI.2020.3035253.
[98] R. Guidotti, A. Monreale, S. Matwin, D. Pedreschi, Explaining image classifiers generating exemplars and counter-exemplars from latent representations, in: In AAAI 2020–34th AAAI Conference on Artificial Intelligence, 2020, pp. 13665–13668, https://doi.org/10.1609/aaai.v34i09.7116.
[99] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Dino Pedreschi, Franco Turini, and Fosca Giannotti. 2018. Local Rule-Based Explanations of Black Box Decision Systems. arXiv Prepr. arXiv1805.10820 (2018). Retrieved from http://arxiv.org/abs/1805.10820.
[100] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. ACM Comput. Surv. 51, 5 (2018), 1–42. 10.1145/3236009.
[101] Riccardo Guidotti, Anna Monreale, Francesco Spinnato, Dino Pedreschi, and Fosca Giannotti. 2020. Explaining any time series classifier. In Proceedings - 2020 IEEE 2nd International Conference on Cognitive Machine Intelligence, CogMI 2020, 167–176. 10.1109/CogMI50398.2020.00029.
[102] Han Guo, Nazneen Rajani, Peter Hase, Mohit Bansal, and Caiming Xiong. 2021. FastIF: Scalable Influence Functions for Efficient Model Interpretation and Debugging. (December 2021), 10333–10350. 10.18653/v1/2021.emnlp-main.808.
[103] Zhiwei Guo, Keping Yu, Alireza Jolfaei, Ali Kashif Bashir, Alaa Omran Almagrabi, and Neeraj Kumar. 2021. Fuzzy Detection System for Rumors through Explainable Adaptive Learning. IEEE Trans. Fuzzy Syst. 29, 12 (December 2021), 3650–3664. 10.1109/TFUZZ.2021.3052109.
[104] K.S. Gurumoorthy, A. Dhurandhar, G. Cecchi, C. Aggarwal, Efficient data representation by selecting prototypes with importance weights, In Proceedings - IEEE International Conference on Data Mining, ICDM 260–269 (2019), https://doi.org/10.1109/ICDM.2019.00036.

[105] T. Ha, S. Kim, D. Seo, S. Lee, Effects of explanation types and perceived risk on trust in autonomous vehicles, Transp. Res. Part F Traffic Psychol. Behav. 73 (2020) (2020) 271–280, https://doi.org/10.1016/j.trf.2020.06.021.
[106] Miseon Han and Jeongtae Kim. 2019. Joint banknote recognition and counterfeit detection using explainable artificial intelligence. Sensors (Switzerland) 19, 16 (2019), (3607) 1–18. 10.3390/s19163607.
[107] Gaofeng Hao, Zhuang Fu, Xin Feng, Zening Gong, Peng Chen, Dan Wang, Weibin Wang, and Yang Si. 2021. A Deep Deterministic Policy Gradient Approach for Vehicle Speed Tracking Control With a Robotic Driver. IEEE Trans. Autom. Sci. Eng. (2021). 10.1109/TASE.2021.3088004.
[108] Jie Hao, Youngsoon Kim, Tae Kyung Kim, and Mingon Kang. 2018. PASNet: Pathway-associated sparse deep neural network for prognosis prediction from high-throughput data. BMC Bioinformatics 19, 1 (2018). 10.1186/s12859-018-2500-z.
[109] A. Heimerl, T. Baur, F. Lingenfelser, J. Wagner, E. Andre, NOVA-A tool for eXplainable Cooperative Machine Learning, in: In 2019 8th International Conference on Affective ComputIng and Intelligent Interaction, 2019, https://doi.org/10.1109/ACII.2019.8925519.
[110] S. Hepenstal, L. Zhang, N. Kodagoda, B.L. William Wong, What are you thinking? Explaining conversational agent responses for criminal investigations, In CEUR Workshop ProceedIngs, 2020.
[111] R.R. Hoffman, Theory → concepts → measures but policies → metrics, In Macrocognition Metrics and Scenarios: Design and Evaluation for Real-World Teams. 3–10 (2012), https://doi.org/10.1201/9781315593173-2.
[112] Robert R. Hoffman, John K. Hawley, and Jeffrey M. Bradshaw. 2014. Myths of automation, part 2: Some very human consequences. IEEE Intell. Syst. 29, 2 (2014), 82–85. 10.1109/MIS.2014.25.
[113] R.R. Hoffman, M. Johnson, J.M. Bradshaw, A.I. Underbrink, Trust in automation, IEEE Intell. Syst. 28 (1) (2013) 84–88, https://doi.org/10.1109/MIS.2013.24.
[114] R.R. Hoffman, G. Klein, S.T. Mueller, Explaining explanation for ‘‘explainable AI, in: In ProceedIngs of the Human Factors and Ergonomics Society, 2018, pp. 197–201, https://doi.org/10.1177/1541931218621047.
[115] F. Hohman, A. Srinivasan, S.M. Drucker, TeleGam: Combining Visualization and Verbalization for Interpretable Machine Learning. In 2019 IEEE Visualization Conference, VIS 2019 (2019) 151–155, https://doi.org/10.1109/VISUAL.2019.8933695.
[116] A. Holzinger, M. Dehmer, F. Emmert-Streib, R. Cucchiara, I. Augenstein, J. Del Ser, W. Samek, I. Jurisica, N. Díaz-Rodríguez, Information fusion as an integrative cross-cutting enabler to achieve robust, explainable, and trustworthy medical artificial intelligence, Inf. Fusion 79 (2022) (2022) 263–278, https://doi.org/10.1016/j.inffus.2021.10.007.
[117] Andreas Holzinger, Georg Langs, Helmut Denk, Kurt Zatloukal, and Heimo Müller. 2019. Causability and explainability of artificial intelligence in medicine. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 9, 4 (2019), (e1312) 1–13. 10.1002/widm.1312.
[118] Andreas Holzinger, Bernd Malle, Anna Saranti, and Bastian Pfeifer. 2021. Towards multi-modal causability with Graph Neural Networks enabling information fusion for explainable AI. Inf. Fusion (2021). 10.1016/j.inffus.2021.01.008.
[119] Andreas Holzinger, André Carrington, and Heimo Müller. 2020. Measuring the quality of explanations: The system causability scale (SCS). Comparing human and machine explanations. KI-Künstliche Intelligenz (German J. Artif. Intell.), 34, 2 (2020),193-198. 10.1007/s13218-020-00636-z.
[120] Andreas Holzinger, Heimo Mueller (2021). Toward Human-AI Interfaces to Support Explainability and Causability in Medical AI. IEEE Computer, 54,10 (2021), 78-86. 10.1109/MC.2021.3092610.
[121] Chris Jay Hoofnagle, Bart van der Sloot, and Frederik Zuiderveen Borgesius. 2019. The European Union general data protection regulation: What it is and what it means. Inf. Commun. Technol. Law 28, 1 (2019), 65–98. 10.1080/13600834.2019.1573501.
[122] K. Höök. 2000. Steps to take before intelligent user interfaces become real. Interact. Comput. 12, 4 (2000), 409–426. 10.1016/S0953-5438(99)00006-5.
[123] S. Hooker, D. Erhan, P.J. Kindermans, B. Kim, A benchmark for interpretability methods in deep neural networks, in Advances in Neural Information Processing Systems, 2019.
[124] Benjamin Hoover, Hendrik Strobelt, and Sebastian Gehrmann. 2020. exBERT: A Visual Analysis Tool to Explore Learned Representations in Transformer Models. 187–196. 10.18653/v1/2020.acl-demos.22.
[125] H.u. Shaoping, Y. Gao, Z. Niu, Y. Jiang, L. Li, X. Xiao, M. Wang, E.F. Fang, W. Menpes-Smith, J. Xia, H. Ye, G. Yang, Weakly Supervised Deep Learning for COVID-19 Infection Detection and Classification from CT Images, IEEE Access 8 (2020) (2020) 118869–118883, https://doi.org/10.1109/ACCESS.2020.3005510.
[126] Wenxing Hu, Xianghe Meng, Yuntong Bai, Aiying Zhang, Gang Qu, Biao Cai, Gemeng Zhang, Tony W. Wilson, Julia M. Stephen, Vince D. Calhoun, and Yu Ping Wang. 2021. Interpretable Multimodal Fusion Networks Reveal Mechanisms of Brain Cognition. IEEE Trans. Med. Imaging 40, 5 (2021), 1474–1483. 10.1109/TMI.2021.3057635.
[127] H.u. Yuening, J. Boyd-Graber, B. Satinoff, A. Smith, Interactive topic modeling, Mach. Learn. 95 (3) (2014) 423–469, https://doi.org/10.1007/s10994-013-5413-0.
[128] Qiang Huang, Makoto Yamada, Yuan Tian, Dinesh Singh, Dawei Yin, and Yi Chang. 2020. GraphLIME: Local Interpretable Model Explanations for Graph Neural Networks. arXiv Prepr. arXiv2001.06216 (2020). Retrieved from http://arxiv.org/abs/2001.06216.
[129] Maximilian Idahl, Lijun Lyu, Ujwal Gadiraju, and Avishek Anand. 2021. Towards Benchmarking the Utility of Explanations for Model Debugging. 68–73. 10.18653/v1/2021.trustnlp-1.8.
[130] Muhammad Aminul Islam, Derek T. Anderson, Anthony J. Pinar, Timothy C. Havens, Grant Scott, and James M. Keller. 2020. Enabling Explainable Fusion in Deep Learning with Fuzzy Integral Neural Networks. IEEE Trans. Fuzzy Syst. 28, 7 (2020), 1291–1300. 10.1109/TFUZZ.2019.2917124.
[131] M. Izadyyazdanabadi, E. Belykh, C. Cavallo, X. Zhao, S. Gandhi, L.B. Moreira, J. Eschbacher, P. Nakaji, M.C. Preul, Y. Yang, Weakly-supervised learning-based feature localization for confocal laser endomicroscopy glioma images, In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 300–308 (2018), https://doi.org/10.1007/978-3-030-00934-2_34.
[132] P.T. Jiang, C.B. Zhang, Q. Hou, M.M. Cheng, Y. Wei, LayerCAM: Exploring hierarchical class activation maps for localization, IEEE Trans. Image Process. 30 (2021) 5875–5888, https://doi.org/10.1109/TIP.2021.3089943.
[133] J. Jiménez-Luna, F. Grisoni, G. Schneider, Drug discovery with explainable artificial intelligence, Nature Machine Intelligence 2 (2020) 573–584, https://doi.org/10.1038/s42256-020-00236-4.
[134] Fernando Jiménez, Rosalia Jódar, Maria del Pilar Martín, Gracia Sánchez, and Guido Sciavicco. 2017. Unsupervised feature selection for interpretable classification in behavioral assessment of children. Expert Syst. 34, 4 (2017), e12173. 10.1111/exsy.12173.
[135] Minsuk Kahng, Pierre Y. Andrews, Aditya Kalro, and Duen Horng Polo Chau. 2018. ActiVis: Visual Exploration of Industry-Scale Deep Neural Network Models. IEEE Trans. Vis. Comput. Graph. 24, 1 (2018), 88–97. 10.1109/TVCG.2017.2744718.
[136] Deepak A. Kaji, John R. Zech, Jun S. Kim, Samuel K. Cho, Neha S. Dangayach, Anthony B. Costa, and Eric K. Oermann. 2019. An attention based deep learning model of clinical events in the intensive care unit. PLoS One 14, 2 (2019). 10.1371/journal.pone.0211057.
[137] Olli Kanerva. 2019. Evaluating explainable AI models for convolutional neural networks with proxy tasks. (2019), 68. Retrieved from https://pdfs.semanticscholar.org/d910/62a3e13ee034af6807e1819a9ca3051daf13.pdf.
[138] A. Kapishnikov, T. Bolukbasi, F. Viegas, M. Terry, XRAI: Better attributions through regions, in: In ProceedIngs of the IEEE International Conference on Computer Vision, 2019, pp. 4947–4956, https://doi.org/10.1109/ICCV.2019.00505.
[139] Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. 2020. A survey of algorithmic recourse: definitions, formulations, solutions, and prospects. (October 2020). Retrieved from http://arxiv.org/abs/2010.04050.
[140] M. Kay, T. Kola, J.R. Hullman, S.A. Munson, When (ish) is my bus? User-centered visualizations of uncertainty in everyday, mobile predictive systems, In Conference on Human Factors in Computing Systems - Proceedings 5092–5103 (2016), https://doi.org/10.1145/2858036.2858558.
[141] M.T. Keane, B. Smyth, Good Counterfactuals and Where to Find Them: A Case-Based Technique for Generating Counterfactuals for Explainable AI (XAI), In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (2020), https://doi.org/10.1007/978-3-030-58342-2_11.

[142] Eoin M. Kenny, Courtney Ford, Molly Quinn, and Mark T. Keane. 2021. Explaining Black-Box classifiers using post-hoc explanations-by-example: The effect of explanations and error-rates in XAI user studies. Artif. Intell. 294, (2021). 10.1016/j.artint.2021.103459.
[143] Eoin M. Kenny and Mark T. Keane. 2019. Twin-systems to explain artificial neural networks using case-based reasoning: Comparative tests of feature-weighting methods in ANN-CBR twins for XAI. In IJCAI International Joint Conference on Artificial Intelligence, 2708–2715. 10.24963/ijcai.2019/376.
[144] Rame Khasawneh and Ruth Kornreich. 2014. Explaining Data-Driven Document Classifications. MIS Q. Manag. Inf. Syst. (2014).
[145] B. Kim, R. Khanna, O. Koyejo, Examples are not enough, learn to criticize! Criticism for interpretability, In Advances in Neural Information Processing Systems (2016) 2288–2296.
[146] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. 2018. Interpretability beyond feature attribution: Quantitative Testing with Concept Activation Vectors (TCAV). In 35th International Conference on Machine Learning, ICML 2018, 4186–4195.
[147] J. Kim, H.J. Kim, C. Kim, W.H. Kim, Artificial intelligence in breast ultrasonography, Ultrasonography 40 (2021) 183–190, https://doi.org/10.14366/usg.20117.
[148] J. Kim, J. Canny, Interpretable Learning for Self-Driving Cars by Visualizing Causal Attention, in: In ProceedIngs of the IEEE International Conference on Computer Vision, 2017, pp. 2961–2969, https://doi.org/10.1109/ICCV.2017.320.
[149] J. Kim, A. Rohrbach, T. Darrell, J. Canny, Z. Akata, Textual Explanations for Self-Driving Vehicles, In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 577–593 (2018), https://doi.org/10.1007/978-3-030-01216-8_35.
[150] Seong Gon Kim, Nawanol Theera-Ampornpunt, Chih Hao Fang, Mrudul Harwani, Ananth Grama, and Somali Chaterji. 2016. Opening up the blackbox: An interpretable deep neural network-based classifier for cell-type specific enhancer predictions. BMC Syst. Biol. 10, (2016). 10.1186/s12918-016-0302-3.
[151] Trevor Kistan, Alessandro Gardi, and Roberto Sabatini. 2018. Machine learning and cognitive ergonomics in air traffic management: Recent developments and considerations for certification. Aerospace 5, 4 (2018), (103) 1–18. 10.3390/aerospace5040103.
[152] Janis Klaise, Arnaud Van Looveren, Giovanni Vacanti, and Alexandru Coca. 2021. Alibi explain: Algorithms for explaining machine learning models. J. Mach. Learn. Res. 22, (2021).
[153] Pang Wei Koh and Percy Liang. 2017. Understanding Black-Box predictions via influence functions. In 34th International Conference on Machine Learning, ICML 2017, 2976–2987.
[154] N. Kokhlikyan, V. Miglani, M. Martin, E. Wang, B. Alsallakh, J. Reynolds, A. Melnikov, N. Kliushkina, C. Araya, S. Yan, O. Reblitz-Richardson, 2020, A unified and generic model interpretability library for PyTorch, Captum, September 2020, Retrieved from http://arxiv.org/abs/2009.07896.
[155] Jeamin Koo, Jungsuk Kwac, Wendy Ju, Martin Steinert, Larry Leifer, and Clifford Nass. 2015. Why did my car just do that? Explaining semi-autonomous driving actions to improve driver understanding, trust, and performance. Int. J. Interact. Des. Manuf. 9, 4 (2015), 269–275. 10.1007/s12008-014-0227-2.
[156] Maxim S. Kovalev, Lev V. Utkin, and Ernest M. Kasimov. 2020. SurvLIME: A method for explaining machine learning survival models. Knowledge-Based Syst. 203, (2020). 10.1016/j.knosys.2020.106164.
[157] Josua Krause, Adam Perer, and Enrico Bertini. 2014. INFUSE: Interactive feature selection for predictive modeling of high dimensional data. IEEE Trans. Vis. Comput. Graph. 20, 12 (2014), 1614–1623. 10.1109/TVCG.2014.2346482.
[158] Sanjay Krishnan and Eugene Wu. 2017. PALM: Machine learning explanations for iterative debugging. In Proceedings of the 2nd Workshop on Human-In-the-Loop Data Analytics, HILDA 2017. 10.1145/3077257.3077271.
[159] T. Kulesza, M. Burnett, W.K. Wong, S. Stumpf, Principles of Explanatory Debugging to personalize interactive machine learning, In International Conference on Intelligent User Interfaces, Proceedings IUI 126–137 (2015), https://doi.org/10.1145/2678025.2701399.
[160] T. Kulesza, S. Stumpf, M. Burnett, W.K. Wong, Y. Riche, T. Moore, I. Oberst, A. Shinsel, K. McIntosh, Explanatory debugging: Supporting end-user debugging of machine-learned programs. In Proceedings - 2010 IEEE Symposium on Visual Languages and Human-Centric Computing, VL/HCC 2010 (2010) 41–48, https://doi.org/10.1109/VLHCC.2010.15.
[161] Bum Chul Kwon, Min Je Choi, Joanne Taery Kim, Edward Choi, Young Bin Kim, Soonwook Kwon, Jimeng Sun, and Jaegul Choo. 2019. RetainVis: Visual Analytics with Interpretable and Interactive Recurrent Neural Networks on Electronic Medical Records. IEEE Trans. Vis. Comput. Graph. 25, 1 (2019), 299–309. 10.1109/TVCG.2018.2865027.
[162] I. Lage, E. Chen, J. He, M. Narayanan, B. Kim, S.J. Gershman, F. Doshi-Velez, Human evaluation of models built for interpretability, Proc. AAAI Conf. Hum. Comput. Crowdsourcing 2019 (2019) 59–67.
[163] S.G. Lakhmani, J.L. Wright, M. Schwartz, D. Barber, Exploring the effect of communication patterns and transparency on the attitudes towards robots, In Advances in Intelligent Systems and Computing 27–36 (2020), https://doi.org/10.1007/978-3-030-20148-7_3.
[164] H. Lakkaraju, S.H. Bach, J. Leskovec, Interpretable decision sets: A joint framework for description and prediction, In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 1675–1684 (2016), https://doi.org/10.1145/2939672.2939874.
[165] O. Lampridis, R. Guidotti, S. Ruggieri, Explaining Sentiment Classification with Synthetic Exemplars and Counter-Exemplars, In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 357–373 (2020), https://doi.org/10.1007/978-3-030-61527-7_24.
[166] Simon Meyer Lauritsen, Mads Kristensen, Mathias Vassard Olsen, Morten Skaarup Larsen, Katrine Meyer Lauritsen, Marianne Johansson Jørgensen, Jeppe Lange, and Bo Thiesson. 2020. Explainable artificial intelligence model to predict acute critical illness from electronic health records. Nat. Commun. 11, 1 (2020). 10.1038/s41467-020-17431-x.
[167] H. Lee, S.T. Kim, Y.M. Ro, Generation of multimodal justification using visual word constraint model for explainable computer-aided diagnosis, In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 21–29 (2019), https://doi.org/10.1007/978-3-030-33850-3_3.
[168] Hyunkwang Lee, Sehyo Yune, Mohammad Mansouri, Myeongchan Kim, Shahein H. Tajmir, Claude E. Guerrier, Sarah A. Ebert, Stuart R. Pomerantz, Javier M. Romero, Shahmir Kamalian, Ramon G. Gonzalez, Michael H. Lev, and Synho Do. 2019. An explainable deep-learning algorithm for the detection of acute intracranial haemorrhage from small datasets. Nat. Biomed. Eng. 3, 3 (2019), 173–182. 10.1038/s41551-018-0324-9.
[169] Benjamin Letham, Cynthia Rudin, Tyler H. McCormick, and David Madigan. 2015. Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. Ann. Appl. Stat. 9, 3 (2015), 1350–1371. 10.1214/15-AOAS848.
[170] Jianqiang Li, Cheng Wang, Jie Chen, Heng Zhang, Yuyan Dai, Lingwei Wang, Li Wang, and Asoke K Nandi. 2022. Explainable CNN with Fuzzy Tree Regularization for Respiratory Sound Analysis. IEEE Trans. Fuzzy Syst. (2022), 1–1. 10.1109/tfuzz.2022.3144448.
[171] L. Li, L. Ma, L. Jiao, F. Liu, Q. Sun, J. Zhao, Complex Contourlet-CNN for polarimetric SAR image classification, Pattern Recognit. 100 (2020) (2020), https://doi.org/10.1016/j.patcog.2019.107110 107110.
[172] Y.u. Liang, S. Li, C. Yan, M. Li, C. Jiang, Explaining the Black-Box model: A survey of local interpretation methods for deep neural networks, Neurocomputing 419 (2021) (2021) 168–182, https://doi.org/10.1016/j.neucom.2020.08.011.
[173] B.Y. Lim, A.K. Dey, Assessing demand for intelligibility in context-aware applications, In ACM International Conference Proceeding Series 195–204 (2009), https://doi.org/10.1145/1620545.1620576.
[174] B.Y. Lim, A.K. Dey, D. Avrahami, Why and why not explanations improve the intelligibility of context-aware intelligent systems, In Conference on Human Factors in Computing Systems - Proceedings 2119–2128 (2009), https://doi.org/10.1145/1518701.1519023.
[175] Mengchen Liu, Shixia Liu, Xizhou Zhu, Qinying Liao, Furu Wei, and Shimei Pan. 2016. An Uncertainty-Aware Approach for Exploratory Microblog Retrieval. IEEE Trans. Vis. Comput. Graph. 22, 1 (2016), 250–259. 10.1109/TVCG.2015.2467554.
[176] Mengchen Liu, Jiaxin Shi, Zhen Li, Chongxuan Li, Jun Zhu, and Shixia Liu. 2017. Towards Better Analysis of Deep Convolutional Neural Networks. IEEE Trans. Vis. Comput. Graph. 23, 1 (2017), 91–100. 10.1109/TVCG.2016.2598831.

[177] T. Liu, H. Zhou, M. Itoh, S. Kitazaki, The Impact of Explanation on Possibility of Hazard Detection Failure on Driver Intervention under Partial Driving Automation, In IEEE Intelligent Vehicles Symposium, Proceedings 150–155 (2018), https://doi.org/10.1109/IVS.2018.8500521.
[178] Y.C. Liu, Y.A. Hsieh, M.H. Chen, C.H. Huck Yang, J. Tegner, Y.C. James Tsai, Interpretable Self-Attention Temporal Reasoning for Driving Behavior Understanding, In ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings 2338–2342 (2020), https://doi.org/10.1109/ICASSP40776.2020.9053783.
[179] Tania Lombrozo. 2009. Explanation and categorization: How "why?" informs "what?" Cognition 110, 2 (2009), 248–253. 10.1016/j.cognition.2008.10.007.
[180] A. Van Looveren, J. Klaise, Interpretable Counterfactual Explanations Guided by Prototypes, In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 650–665 (2021), https://doi.org/10.1007/978-3-030-86520-7_40.
[181] S.M. Lundberg, S.I. Lee, A unified approach to interpreting model predictions, in: in Advances in Neural Information Processing Systems, 2017, pp. 4766–4775.
[182] Dongsheng Luo, Wei Cheng, Dongkuan Xu, Wenchao Yu, Bo Zong, Haifeng Chen, and Xiang Zhang. 2020. Parameterized explainer for graph neural network. In Advances in Neural Information Processing Systems.
[183] M. Madsen, S. Gregor, 2000, Measuring Human-Computer Trust. Proc. Elev. Australas. Conf. Inf, Syst, 2000.
[184] Aakarsh Malhotra, Surbhi Mittal, Puspita Majumdar, Saheb Chhabra, Kartik Thakral, Mayank Vatsa, Richa Singh, Santanu Chaudhury, Ashwin Pudrod, and Anjali Agrawal. 2022. Multi-task driven explainable diagnosis of COVID-19 using chest X-ray images. Pattern Recognit. 122, (2022). 10.1016/j.patcog.2021.108243.
[185] D.L. Marino, C.S. Wickramasinghe, M. Manic, An adversarial approach for explainable AI in intrusion detection systems, in: In ProceedIngs: IECON 2018–44th Annual Conference of the IEEE Industrial Electronics Society, 2018, pp. 3237–3243, https://doi.org/10.1109/IECON.2018.8591457.
[186] Charles T. Marx, Flavio Du Pin Calmon, and Berk Ustun. 2020. Predictive multiplicity in classification. In 37th International Conference on Machine Learning, ICML 2020, 6721–6730.
[187] W. Mayer, M. Stumptner, Evaluating models for model-based debugging, in: In ASE 2008–23rd IEEE/ACM International Conference on Automated Software EngIneerIng, 2008, pp. 128–137, https://doi.org/10.1109/ASE.2008.23.
[188] Anna Meldo, Lev Utkin, Maxim Kovalev, and Ernest Kasimov. 2020. The natural language explanation algorithms for the lung cancer computer-aided diagnosis system. Artif. Intell. Med. 108, (2020). 10.1016/j.artmed.2020.101952.
[189] Jerry M. Mendel and Piero P. Bonissone. 2021. Critical Thinking about Explainable AI (XAI) for Rule-Based Fuzzy Systems. IEEE Trans. Fuzzy Syst. 29, 12 (2021), 3579–3593. 10.1109/TFUZZ.2021.3079503.
[190] Stephanie M. Merritt, Heather Heimbaugh, Jennifer Lachapell, and Deborah Lee. 2013. I trust it, but i don't know why: Effects of implicit attitudes toward automation on trust in an automated system. Hum. Factors 55, 3 (2013), 520–534. 10.1177/0018720812465081.
[191] D. Meyerson, K.E. Weick, R.M. Kramer, Swift Trust and Temporary Groups, In Trust in Organizations: Frontiers of Theory and Research. 166–195 (2012), https://doi.org/10.4135/9781452243610.n9.
[192] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, Artificial Intelligence 267 (2019) 1–38, https://doi.org/10.1016/j.artint.2018.07.007.
[193] Yao Ming, Huamin Qu, and Enrico Bertini. 2019. RuleMatrix: Visualizing and Understanding Classifiers with Rules. IEEE Trans. Vis. Comput. Graph. 25, 1 (2019), 342–352. 10.1109/TVCG.2018.2864812.
[194] B. Mittelstadt, C. Russell, S. Wachter, in: ExplAining Explanations in AI, and Transparency, 2019, pp. 279–288, https://doi.org/10.1145/3287560.3287574.
[195] Y. Mizuchi, T. Inamura, Estimation of Subjective Evaluation of HRI Performance Based on Objective Behaviors of Human and Robots, In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 201–212 (2019), https://doi.org/10.1007/978-3-030-35699-6_16.
[196] Sina Mohseni, Jeremy E. Block, and Eric D. Ragan. 2018. A Human-Grounded Evaluation Benchmark for Local Explanations of Machine Learning. (January 2018). Retrieved from http://arxiv.org/abs/1801.05075.
[197] Sina Mohseni, Niloofar Zarei, and Eric D. Ragan. 2021. A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. ACM Trans. Interact. Intell. Syst. 11, 3–4 (2021), 1–45. 10.1145/3387166.
[198] Ioannis Mollas, Nikolaos Bassiliades, and Grigorios Tsoumakas. 2020. LioNets: Local interpretation of neural networks through penultimate layer decoding. Commun. Comput. Inf. Sci. 1167 CCIS, (2020), 265–276. 10.1007/978-3-030-43823-4_23.
[199] I. Mollas, N. Bassiliades, G. Tsoumakas, LioNets: Local interpretation of neural networks through penultimate layer decoding, In Communications in Computer and Information Science 265–276 (2020), https://doi.org/10.1007/978-3-030-43823-4_23.
[200] K. Mori, H. Fukui, T. Murase, T. Hirakawa, T. Yamashita, H. Fujiyoshi, Visual explanation by attention branch network for end-to-end learning-based self-driving, In IEEE Intelligent Vehicles Symposium, Proceedings 1577–1582 (2019), https://doi.org/10.1109/IVS.2019.8813900.
[201] R.K. Mothilal, A. Sharma, C. Tan, in: Explaining Machine Learning Classifiers Through Diverse Counterfactual Explanations, and Transparency, 2020, pp. 607–617, https://doi.org/10.1145/3351095.3372850.
[202] R. Nahata, D. Omeiza, R. Howard, L. Kunze, Assessing and Explaining Collision Risk in Dynamic Environments for Autonomous Driving Safety, In IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC 223–230 (2021), https://doi.org/10.1109/ITSC48978.2021.9564966.
[203] Rakshit Naidu, Ankita Ghosh, Yash Maurya, Shamanth R Nayak K, and Soumya Snigdha Kundu. 2020. IS-CAM: Integrated Score-CAM for axiomatic-based explanations. (October 2020). Retrieved from http://arxiv.org/abs/2010.03023.
[204] H. Nori, S. Jenkins, P. Koch, R. Caruana, 2019, A Unified Framework for Machine Learning Interpretability, InterpretML, September 2019, Retrieved from http://arxiv.org/abs/1909.09223.
[205] F. Nothdurft, F. Richter, W. Minker, Probabilistic human-computer trust handling, in: In SIGDIAL 2014–15th Annual MeetIng of the Special Interest Group on Discourse and Dialogue, ProceedIngs of the Conference, 2014, pp. 51–59, https://doi.org/10.3115/v1/w14-4307.
[206] M. Nourani, S. Kabir, S. Mohseni, and E. D. Ragan. 2019. View of The Effects of Meaningful and Meaningless Explanations on Trust and Perceived System Accuracy in Intelligent Systems. Proc. AAAI Conf. Hum. Comput. Crowdsourcing 7, 1 (2019), 97–105. Retrieved from https://ojs.aaai.org/index.php/HCOMP/article/view/5284/5136.
[207] M. Nourani, C. Roy, J.E. Block, D.R. Honeycutt, T. Rahman, E. Ragan, V. Gogate, Anchoring Bias Affects Mental Model Formation and User Reliance in Explainable AI Systems, In International Conference on Intelligent User Interfaces, Proceedings IUI 340–350 (2021), https://doi.org/10.1145/3397481.3450639.
[208] T. Nowak, M.R. Nowicki, K. Cwian, P. Skrzypczynski, How to improve object detection in a driver assistance system applying explainable deep learning, In IEEE Intelligent Vehicles Symposium, Proceedings 226–231 (2019), https://doi.org/10.1109/IVS.2019.8814134.
[209] D. Omeiza, K. Kollnig, H. Web, M. Jirotka, L. Kunze, Why Not Explain? Effects of Explanations on Human Perceptions of Autonomous Driving, In Proceedings of IEEE Workshop on Advanced Robotics and its Social Impacts, ARSO 194–199 (2021), https://doi.org/10.1109/ARSO51874.2021.9542835.
[210] Daniel Omeiza, Skyler Speakman, Celia Cintas, and Komminist Weldermariam. 2019. Smooth Grad-CAM++: An Enhanced Inference Level Visualization Technique for Deep Convolutional Neural Network Models. (August 2019). Retrieved from http://arxiv.org/abs/1908.01224.
[211] D. Omeiza, H. Web, M. Jirotka, L. Kunze, Towards accountability: Providing intelligible explanations in autonomous driving, In IEEE Intelligent Vehicles Symposium, Proceedings 231–237 (2021), https://doi.org/10.1109/IV48863.2021.9575917.
[212] Liyan Pan, Guangjian Liu, Xiaojian Mao, Huixian Li, Jiexin Zhang, Huiying Liang, and Xiuzhen Li. 2019. Development of prediction models using machine learning algorithms for girls with suspected central precocious puberty: Retrospective study. JMIR Med. Informatics 7, 1 (2019). 10.2196/11728.

[213] C. Panigutti, A. Perotti, D. Pedreschi, in: Doctor XAI An Ontology-based Approach to Black-Box Sequential Data Classification ExplAnations, and Transparency, 2020, pp. 629–639, https://doi.org/10.1145/3351095.3372855.
[214] P. Paudyal, J. Lee, A. Kamzin, M. Soudki, A. Banerjee, S.K.S. Gupta, Learn2Sign: Explainable AI for sign language learning, In CEUR Workshop ProceedIngs, 2019.
[215] J. Pearl, Causality: Models, Reasoning, and Inference, 2nd Edition., Cambridge University Press, Cambridge, 2009.
[216] Tomi Peltola. 2018. Local Interpretable Model-agnostic Explanations of Bayesian Predictive Models via Kullback-Leibler Projections. (October 2018). Retrieved from http://arxiv.org/abs/1810.02678.
[217] Luís Moniz Pereira, António Barata Lopes. 2020. Cognitive prerequisites: The special case of counterfactual reasoning. In Studies in Applied Philosophy, Epistemology and Rational Ethics. 10.1007/978-3-030-39630-5_14.
[218] L. Perlmutter, E. Kernfeld, M. Cakmak, Situated language understanding with human-like and visualization-based transparency, In Robotics: Science and Systems. (2016), 10.15607/rss.2016.xii.040.
[219] Vitali Petsiuk, Abir Das, and Kate Saenko. 2019. RisE: Randomized input sampling for explanation of Black-Box models. In British Machine Vision Conference 2018, BMVC 2018.
[220] Pouya Pezeshkpour, Yifan Tian, and Sameer Singh. 2019. Investigating robustness and interpretability of link prediction via adversarial modifications. In NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, 3336–3347. 10.18653/v1/n19-1337.
[221] Nicola Pezzotti, Thomas Höllt, Jan Van Gemert, Boudewijn P.F. Lelieveldt, Elmar Eisemann, and Anna Vilanova. 2018. DeepEyes: Progressive Visual Analytics for Designing Deep Neural Networks. IEEE Trans. Vis. Comput. Graph. 24, 1 (2018), 98–108. 10.1109/TVCG.2017.2744358.
[222] G. Plumb, D. Molitor, A. Talwalkar, Model agnostic supervised local explanations, In Advances in Neural Information Processing Systems (2018) 2515–2524.
[223] P.E. Pope, S. Kolouri, M. Rostami, C.E. Martin, H. Hoffmann, Explainability methods for graph convolutional neural networks, in: In ProceedIngs of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2019, pp. 10764–10773, https://doi.org/10.1109/CVPR.2019.01103.
[224] Mihaela Porumb, Saverio Stranges, Antonio Pescapè, and Leandro Pecchia. 2020. Precision Medicine and Artificial Intelligence: A Pilot Study on Deep Learning for Hypoglycemic Events Detection based on ECG. Sci. Rep. 10, 1 (2020). 10.1038/s41598-019-56927-5.
[225] R. Poyiadzi, K. Sokol, R. Santos-Rodriguez, T. De Bie, P. Flach, in: FACE: Feasible and Actionable Counterfactual Explanations, and Society, 2020, pp. 344–350, https://doi.org/10.1145/3375627.3375850.
[226] N. Prentzas, A. Nicolaides, E. Kyriacou, A. Kakas, C. Pattichis, Integrating machine learning with symbolic reasoning to build an explainable ai model for stroke prediction, in: In ProceedIngs - 2019 IEEE 19th International Conference on BioInformatics and BioengIneerIng, 2019, https://doi.org/10.1109/BIBE.2019.00152.
[227] P.u. Pearl, L.i. Chen, Trust building with explanation interfaces, In International Conference on Intelligent User Interfaces, Proceedings IUI 93–100 (2006), https://doi.org/10.1145/1111449.1111475.
[228] V. Putnam, C. Conati, Exploring the need for explainable artificial intelligence (XAI) in intelligent tutoring systems (ITS), In CEUR Workshop ProceedIngs, 2019.
[229] E. Rader, R. Gray, Understanding user beliefs about algorithmic curation in the facebook news feed, In Conference on Human Factors in Computing Systems - Proceedings 173–182 (2015), https://doi.org/10.1145/2702123.2702174.
[230] A. Rahimpour, S. Martin, A. Tawari, H. Qi, Context aware road-user importance estimation (iCARE), In IEEE Intelligent Vehicles Symposium, Proceedings 2337–2343 (2019), https://doi.org/10.1109/IVS.2019.8814210.
[231] Arun Rai. 2020. Explainable AI: from black box to glass box. J. Acad. Mark. Sci. 48, 1 (2020), 137–141. 10.1007/s11747-019-00710-5.
[232] N.F. Rajani, B. McCann, C. Xiong, R. Socher, Explain Yourself! Leveraging language models for commonsense reasoning, in: In ACL 2019–57th Annual MeetIng of the Association for Computational LInguistics, ProceedIngs of the Conference, 2020, pp. 4932–4942, https://doi.org/10.18653/v1/p19-1487.
[233] D. Rajapaksha, C. Bergmeir, W. Buntine, LoRMIkA: Local rule-based model interpretability with k-optimal associations, Inf. Sci. (Ny) 540 (2020) (2020) 221–241, https://doi.org/10.1016/j.ins.2020.05.126.
[234] Pranav Rajpurkar, Allison Park, Jeremy Irvin, Chris Chute, Michael Bereket, Domenico Mastrodicasa, Curtis P. Langlotz, Matthew P. Lungren, Andrew Y. Ng, and Bhavik N. Patel. 2020. AppendiXNet: Deep Learning for Diagnosis of Appendicitis from A Small Dataset of CT Exams Using Video Pretraining. Sci. Rep. 10, 1 (2020). 10.1038/s41598-020-61055-6.
[235] Vasumathi Raman, Constantine Lignos, Cameron Finucane, Kenton C. T. Lee, Mitch Marcus, and Hadas Kress-Gazit. 2016. Sorry Dave, I'm Afraid I Can't Do That: Explaining Unachievable Robot Tasks Using Natural Language. 10.15607/rss.2013.ix.023.
[236] Gabrielle Ras, Ning Xie, Marcel Van Gerven, and Derek Doran. 2022. Explainable Deep Learning: A Field Guide for the Uninitiated. J. Artif. Intell. Res. 73, (January 2022), 329–397. 10.1613/jair.1.13200.
[237] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?" Explaining the Predictions of Any Classifier. In NAACL-HLT 2016 - 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Demonstrations Session, 97–101. 10.18653/v1/n16-3020.
[238] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-precision model-agnostic explanations. In Proceedings of the AAAI conference on artificial intelligence, Washington, 1527–1535.
[239] M. Ribera, A. Lapedriza, Can we do better explanations? A proposal of user-centered explainable AI, In CEUR Workshop Proceedings 38 (2019).
[240] Michael Ridley. 2022. Explainable Artificial Intelligence (XAI). Inf. Technol. Libr. 41, 2 (June 2022). 10.6017/ital.v41i2.14683.
[241] Heather Riley and Mohan Sridharan. 2019. Integrating Non-monotonic Logical Reasoning and Inductive Learning With Deep Learning for Explainable Visual Question Answering. Front. Robot. AI 6, (2019), (125) 1–20. 10.3389/frobt.2019.00125.
[242] S.G. Rizzo, G. Vantini, S. Chawla, Reinforcement Learning with Explainability for Traffic Signal Control. In 2019 IEEE Intelligent Transportation Systems Conference, ITSC 2019 (2019) 3567–3572, https://doi.org/10.1109/ITSC.2019.8917519.
[243] Andrew Slavin Ros and Finale Doshi-Velez. 2018. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. In 32nd AAAI Conference on Artificial Intelligence, AAAI 2018, 1660–1669.
[244] Andrew Slavin Ross, Michael C. Hughes, and Finale Doshi-Velez. 2017. Right for the right reasons: Training differentiable models by constraining their explanations. In IJCAI International Joint Conference on Artificial Intelligence, 2662–2670. 10.24963/ijcai.2017/371.
[245] Andrew Slavin Ross, Michael C. Hughes, and Finale Doshi-Velez. 2017. Right for the right reasons: Training differentiable models by constraining their explanations. IJCAI Int. Jt. Conf. Artif. Intell. 0, (2017), 2662–2670. 10.24963/ijcai.2017/371.
[246] M. Rueben, S. Nikolaidis, M. De Graaf, E. Phillips, L. Robert, D. Sirkin, M. Kwon, S. Thellman, Half day workshop on mental models of robots, In ACM/IEEE International Conference on Human-Robot Interaction 658–659 (2020), https://doi.org/10.1145/3371382.3374856.
[247] S. Sachan, J.B. Yang, X.u. Dong Ling, D.E. Benavides, Y. Li, An explainable AI decision-support-system to automate loan underwriting, Expert Syst. Appl. 144 (2020) (2020), https://doi.org/10.1016/j.eswa.2019.113100 113100.
[248] M. Sahakyan, Z. Aung, T. Rahwan, Explainable Artificial Intelligence for Tabular Data: A Survey, IEEE Access 9 (2021) (2021) 135392–135422, https://doi.org/10.1109/ACCESS.2021.3116481.
[249] Benjamin Sanchez-Lengeling, Jennifer Wei, Brian Lee, Emily Reif, Peter Y. Wang, Wesley Wei Qian, Kevin McCloskey, Lucy Colwell, and Alexander Wiltschko. 2020. Evaluating attribution for graph neural networks. In Advances in Neural Information Processing Systems.
[250] Michael Sejr Schlichtkrull, Nicola De Cao, and Ivan Titov. 2020. Interpreting Graph Neural Networks for NLP With Differentiable Edge Masking. arXiv Prepr. arXiv2010.00577 (2020). Retrieved from http://arxiv.org/abs/2010.00577.
[251] T. Schneider, J. Hois, A. Rosenstein, Explain yourself! transparency for positive ux in autonomous driving, In Conference on Human Factors in Computing Systems - Proceedings (2021), https://doi.org/10.1145/3411764.3446647.

[252] Jan Maarten Schraagen, Pia Elsasser, Hanna Fricke, Marleen Hof, and Fabyen Ragalmuto. 2020. Trusting the X in XAI: Effects of different types of explanations by a self-driving car on trust, explanation satisfaction and mental models. Proc. Hum. Factors Ergon. Soc. Annu. Meet. 64, 1 (2020), 339–343. 10.1177/1071181320641077.
[253] Robert Schwarzenberg, Marc, die;bner Hu, David Harbecke, Christoph Alt, and Leonhard Hennig. 2019. Layerwise relevance visualization in convolutional text graph classifiers. In EMNLP-IJCNLP 2019 - Graph-Based Methods for Natural Language Processing - Proceedings of the 13th Workshop, 58–62. 10.18653/v1/d19-5308.
[254] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2020. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. Int. J. Comput. Vis. 128, 2 (2020), 336–359. 10.1007/s11263-019-01228-7.
[255] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization, In International Journal of Computer Vision 336–359 (2020), https://doi.org/10.1007/s11263-019-01228-7.
[256] A. Sena, Y. Zhao, M.J. Howard, Teaching human teachers to teach robot learners, In Proceedings - IEEE International Conference on Robotics and Automation 5675–5681 (2018), https://doi.org/10.1109/ICRA.2018.8461194.
[257] M. Setzu, R. Guidotti, A. Monreale, F. Turini, Global explanations with local scoring, In Communications in Computer and Information Science 159–171 (2020), https://doi.org/10.1007/978-3-030-43823-4_14.
[258] Mattia Setzu, Riccardo Guidotti, Anna Monreale, Franco Turini, Dino Pedreschi, and Fosca Giannotti. 2021. GLocalX - From Local to Global Explanations of Black Box AI Models. Artif. Intell. 294, (2021). 10.1016/j.artint.2021.103457.
[259] S.M. Shankaranarayana, D. Runje, ALIME: Autoencoder based approach for local interpretability, In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 454–463 (2019), https://doi.org/10.1007/978-3-030-33607-3_49.
[260] Xiaoting Shao, Arseny Skryagin, Wolfgang Stammer, Patrick Schramowski, and Kristian Kersting. 2021. Right for Better Reasons: Training Differentiable Models by Constraining their Influence Functions. Proc. AAAI Conf. Artif. Intell. 35, 11 (2021), 9533–9540.
[261] R.K. Sheh, "Why did you do that?" Explainable intelligent robots, In AAAI Workshop - Technical Report (2017) 628–634.
[262] Yuan Shen, Shanduojiao Jiang, Yanlin Chen, Eileen Yang, Xilun Jin, Yuliang Fan, and Katie Driggs Campbell. 2020. To Explain or Not to Explain: A Study on the Necessity of Explanations for Autonomous Vehicles. arXiv Prepr. arXiv2006.11684 (2020). Retrieved from http://arxiv.org/abs/2006.11684.
[263] Wenqi Shi, Li Tong, Yuanda Zhu, and May D. Wang. 2021. COVID-19 Automatic Diagnosis with Radiographic Imaging: Explainable Attention Transfer Deep Neural Networks. IEEE J. Biomed. Heal. Informatics 25, 7 (2021), 2376–2387. 10.1109/JBHI.2021.3074893.
[264] Benjamin Shickel, Tyler J. Loftus, Lasith Adhikari, Tezcan Ozrazgat-Baslanti, Azra Bihorac, and Parisa Rashidi. 2019. DeepSOFA: A Continuous Acuity Score for Critically Ill Patients using Clinically Interpretable Deep Learning. Sci. Rep. 9, 1 (2019). 10.1038/s41598-019-38491-0.
[265] A. Shrikumar, P. Greenside, A. Kundaje, Learning important features through propagating activation differences, In 34th International Conference on MachIne LearnIng, 2017.
[266] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep inside convolutional networks: Visualising image classification models and saliency maps. 2nd Int. Conf. Learn. Represent. ICLR 2014 - Work. Track Proc. (2014).
[267] A. Singh, A.R. Mohammed, J. Zelek, V. Lakshminarayanan, Interpretation of deep learning using attributions: application to ophthalmic diagnosis. 9 (2020), https://doi.org/10.1117/12.2568631.
[268] D. Slack, S. Hilgard, E. Jia, S. Singh, H. Lakkaraju, in: Fooling LIME and SHAP: Adversarial Attacks on Post Hoc Explanation Methods, and Society, 2020, pp. 180–186, https://doi.org/10.1145/3375627.3375830.
[269] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. 2017. SmoothGrad: removing noise by adding noise. arXiv Prepr. arXiv1706.03825 (2017). Retrieved from http://arxiv.org/abs/1706.03825.
[270] Cristina Soviany. 2018. The benefits of using artificial intelligence in payment fraud detection: A case study. J. Payments Strateg. Syst. 12, 2 (2018), 102–110.
[271] A. G. Stepanian. 2021. The European Union Artificial Intelligence Act: the first look at the project. Cour. Kutafin Moscow State Law Univ. (2021). 10.17803/2311-5998.2021.83.7.093-098.
[272] I. Stepin, J.M. Alonso, A. Catala, M. Pereira-Farina, A Survey of Contrastive and Counterfactual Explanation Generation Methods for Explainable Artificial Intelligence, IEEE Access 9 (2021) 11974–12001, https://doi.org/10.1109/ACCESS.2021.3051315.
[273] I. Stepin, J.M. Alonso, A. Catala, M. Pereira-Farina, Generation and evaluation of factual and gounterfaetual explanations for decision trees and fuzzy rule-based classifiers, In IEEE International Conference on Fuzzy Systems (2020), https://doi.org/10.1109/FUZZ48607.2020.9177629.
[274] Karl Stöger, David Schneeberger, and Andreas Holzinger. 2021. Medical artificial intelligence. Commun. ACM 64, 11 (November 2021), 34–36. 10.1145/3458652.
[275] Jungyo Suh, Sangjun Yoo, Juhyun Park, Sung Yong Cho, Min Chul Cho, Hwancheol Son, and Hyeon Jeong. 2020. Development and validation of an explainable artificial intelligence-based decision-supporting tool for prostate biopsy. BJU Int. 126, 6 (2020), 694–703. 10.1111/bju.15122.
[276] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, In 34th International Conference on MachIne LearnIng, 2017.
[277] S. Tan, M. Soloviev, G. Hooker, M.T. Wells, Tree Space Prototypes: Another Look at Making Tree Ensembles Interpretable, in: In FODS 2020 - ProceedIngs of the 2020 ACM-IMS Foundations of Data Science Conference, 2020, pp. 23–34, https://doi.org/10.1145/3412815.3416893.
[278] Erico Tjoa and Cuntai Guan. 2021. A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI. IEEE Trans. Neural Networks Learn. Syst. 32, 11 (2021), 4793–4813. 10.1109/TNNLS.2020.3027314.
[279] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, In Advances in Neural Information Processing Systems (2017) 5999–6009.
[280] S. Verma, J. Dickerson, K. Hines, 2020, A Review, Counterfactual Explanations for Machine Learning, October 2020, Retrieved from http://arxiv.org/abs/2010.10596.
[281] G. Vilone, L. Longo, Notions of explainability and evaluation approaches for explainable artificial intelligence, Inf. Fusion 76 (2021) (2021) 89–106, https://doi.org/10.1016/j.inffus.2021.05.009.
[282] M.N. Vu, M.T. Thai, PGM-explainer: Probabilistic graphical model explanations for graph neural networks, in Advances in Neural Information Processing Systems, 2020.
[283] Sandra Wachter, Brent Mittelstadt, and Luciano Floridi. 2017. Transparent, explainable, and accountable AI for robotics. Sci. Robot. 2, 6 (2017). 10.1126/scirobotics.aan6080.
[284] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2017. Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. SSRN Electron. J. (2017). 10.2139/ssrn.3063289.
[285] Bernhard Waltl and Roland Vogl. 2018. Explainable artificial intelligence - The new frontier in legal informatics. Jusletter IT 4, February (2018), 1–10.
[286] Douglas Walton. 2007. Dialogical models of explanation. AAAI Work. - Tech. Rep. WS-07-06, 1 (2007), 1–9.
[287] Haofan Wang, Rakshit Naidu, Joy Michael, and Soumya Snigdha Kundu. 2020. SS-CAM: Smoothed Score-CAM for Sharper Visual Feature Localization. (June 2020). Retrieved from http://arxiv.org/abs/2006.14255.
[288] H. Wang, Z. Wang, D.u. Mengnan, F. Yang, Z. Zhang, S. Ding, P. Mardziel, H.u. Xia, Score-CAM: Score-weighted visual explanations for convolutional neural networks, In IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops 111–119 (2020), https://doi.org/10.1109/CVPRW50498.2020.00020.
[289] Hong Wang, Yuexiang Li, Nanjun He, Kai Ma, Deyu Meng, and Yefeng Zheng. 2021. DICDNet: Deep Interpretable Convolutional Dictionary Network for Metal Artifact Reduction in CT Images. IEEE Trans. Med. Imaging (2021), 1–1. 10.1109/TMI.2021.3127074.
[290] M. Wang, K. Zheng, Y. Yang, X. Wang, An Explainable Machine Learning Framework for Intrusion Detection Systems, IEEE Access 8 (2020) (2020) 73127–73141, https://doi.org/10.1109/ACCESS.2020.2988359.

[291] Xiaofei Wang, Lai Jiang, Liu Li, Mai Xu, Xin Deng, Lisong Dai, Xiangyang Xu, Tianyi Li, Yichen Guo, Zulin Wang, and Pier Luigi Dragotti. 2021. Joint Learning of 3D Lesion Segmentation and Classification for Explainable COVID-19 Diagnosis. IEEE Trans. Med. Imaging 40, 9 (2021), 2463–2476. 10.1109/TMI.2021.3079709.
[292] Xiting Wang, Shixia Liu, Junlin Liu, Jianfei Chen, Jun Zhu, and Baining Guo. 2016. TopicPanorama: A Full Picture of Relevant Topics. IEEE Trans. Vis. Comput. Graph. 22, 12 (2016), 2508–2521. 10.1109/TVCG.2016.2515592.
[293] Hilde J. P. Weerts, Werner van Ipenburg, and Mykola Pechenizkiy. 2019. A Human-Grounded Evaluation of SHAP for Alert Processing. (July 2019). Retrieved from http://arxiv.org/abs/1907.03324.
[294] K.I. Daniel, J.D. Weidele, E.O. Weisz, M. Muller, J. Andres, A. Gray, D. Wang, AutoAIViz: Opening the blackbox of automated artificial intelligence with conditional parallel coordinates, In International Conference on Intelligent User Interfaces, Proceedings IUI 308–312 (2020), https://doi.org/10.1145/3377325.3377538.
[295] S.M. Weiss, N. Indurkhya, Lightweight Rule Induction. Proc. 17th Int Retrieved from Conf. Mach. Learn. 2000 (2000) 1135–1142. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.34.4619&rep=rep1&type=pdf.
[296] B.M. Wilamowski, Understanding neural networks. Intell. Syst. (2016), https://doi.org/10.2307/1270439.
[297] J.J. Williams, W.S. Lasecki, A.N. Rafferty, A. Ang, D. Tingley, J. Kim, Connecting instructors and learning scientists via collaborative dynamic experimentation, In Conference on Human Factors in Computing Systems - Proceedings 3012–3018 (2017), https://doi.org/10.1145/3027063.3053247.
[298] R.H. Wortham, A. Theodorou, J.J. Bryson, Robot transparency: Improving understanding of intelligent behaviour for designers and users, In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 274–289 (2017), https://doi.org/10.1007/978-3-319-64107-2_22.
[299] W.u. Yu Huan, S.H. Gao, J. Mei, X.u. Jun, D.P. Fan, R.G. Zhang, M.M. Cheng, JCS: An Explainable COVID-19 Diagnosis System by Joint Classification and Segmentation, IEEE Trans. Image Process. 30 (2021) (2021) 3113–3126, https://doi.org/10.1109/TIP.2021.3058783.
[300] Cao Xiao, Tengfei Ma, Adji B. Dieng, David M. Blei, and Fei Wang. 2018. Readmission prediction via deep contextual embedding of clinical concepts. PLoS One 13, 4 (2018). 10.1371/journal.pone.0195024.
[301] Qianqian Xie, Prayag Tiwari, Deepak Gupta, Jimin Huang, and Min Peng. 2021. Neural variational sparse topic model for sparse explainable text representation. Inf. Process. Manag. 58, 5 (2021). 10.1016/j.ipm.2021.102614.
[302] Yao Xie, Xiang Anthony Chen, and Ge Gao. 2019. Outlining the design space of explainable intelligent systems for medical diagnosis. CEUR Workshop Proc. 2327, (2019).
[303] Yiran Xu, Xiaoyin Yang, Lihang Gong, Hsuan Chu Lin, Tz Ying Wu, Yunsheng Li, and Nuno Vasconcelos. 2020. Explainable object-induced action decision for autonomous vehicles. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 9520–9529. 10.1109/CVPR42600.2020.00954.
[304] C. Yang, A. Rangarajan, S. Ranka, Global Model Interpretation Via Recursive Partitioning, in: In ProceedIngs - 20th International Conference on High Performance ComputIng and Communications, 16th International Conference on Smart City and 4th International Conference on Data Science and Systems, 2019, https://doi.org/10.1109/HPCC/SmartCity/DSS.2018.00256.
[305] Guang Yang, Felix Raschke, Thomas R. Barrick, and Franklyn A. Howe. 2015. Manifold Learning in MR spectroscopy using nonlinear dimensionality reduction and unsupervised clustering. Magn. Reson. Med. 74, 3 (2015), 868–878. 10.1002/mrm.25447.
[306] G. Yang, Q. Ye, J. Xia, Unbox the Black-Box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond, Inf. Fusion 77 (2022) (2022) 29–52, https://doi.org/10.1016/j.inffus.2021.07.016.
[307] H. Yang, C. Rudin, M. Seltzer, Scalable Bayesian rule lists, In 34th International Conference on MachIne LearnIng, 2017.
[308] Zebin Yang, Aijun Zhang, and Agus Sudjianto. 2021. GAMI-Net: An explainable neural network based on generalized additive models with structured interactions. Pattern Recognit. 120, (2021). 10.1016/j.patcog.2021.108192.
[309] Mojtaba Yeganejou, Scott Dick, and James Miller. 2020. Interpretable Deep Convolutional Fuzzy Classifier. IEEE Trans. Fuzzy Syst. (2020). 10.1109/TFUZZ.2019.2946520.
[310] Chih Kuan Yeh, Been Kim, Sercan Arik, Chun Liang Li, Tomas Pfister, and Pradeep Ravikumar. 2020. On completeness-aware concept-based explanations in deep neural networks. Adv. Neural Inf. Process. Syst. 2020-Decem, (2020).
[311] M. Yin, J.W. Vaughan, H. Wallach, Understanding the effect of accuracy on trust in machine learning models, In Conference on Human Factors in Computing Systems - Proceedings (2019), https://doi.org/10.1145/3290605.3300509.
[312] R. Ying, D. Bourgeois, J. You, M. Zitnik, J. Leskovec, GNNExplainer: Generating explanations for graph neural networks, in Advances in Neural Information Processing Systems, 2019.
[313] J. Yoon, K. Kim, J. Jang, Propagated perturbation of adversarial attack for well-known CNNs: Empirical study and its explanation, in: In ProceedIngs - 2019 International Conference on Computer Vision Workshop, 2019, https://doi.org/10.1109/ICCVW.2019.00520.
[314] Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. 2015. Understanding Neural Networks Through Deep Visualization. 2005 Proc. - 22nd Int. VLSI Multilevel Interconnect. Conf. VMIC 2005 (2015). Retrieved from http://arxiv.org/abs/1506.06579.
[315] H. Yuan, J. Tang, H.u. Xia, S. Ji, XGNN: Towards Model-Level Explanations of Graph Neural Networks, in: In ProceedIngs of the ACM SIGKDD International Conference on Knowledge Discovery and Data MInIng, 2020, pp. 430–438, https://doi.org/10.1145/3394486.3403085.
[316] Hao Yuan, Haiyang Yu, Shurui Gui, and Shuiwang Ji. 2020. Explainability in Graph Neural Networks: A Taxonomic Survey. (December 2020). Retrieved from http://arxiv.org/abs/2012.15445.
[317] Hao Yuan, Haiyang Yu, Jie Wang, Kang Li, and Shuiwang Ji. 2021. On Explainability of Graph Neural Networks via Subgraph Explorations. arXiv Prepr. arXiv2102.05152 (2021). Retrieved from http://arxiv.org/abs/2102.05152.
[318] Muhammad Rehman Zafar and Naimul Mefraz Khan. 2019. DLIME: A Deterministic Local Interpretable Model-Agnostic Explanations Approach for Computer-Aided Diagnosis Systems. (June 2019). Retrieved from http://arxiv.org/abs/1906.10263.
[319] T. Zahavy, N.B. Zrihem, S. Mannor, Graying the black box: Understanding DQNs, In 33rd International Conference on MachIne LearnIng, 2016.
[320] J. Zhang, K. Kowsari, J.H. Harrison, J.M. Lobo, L.E. Barnes, Patient2Vec: A Personalized Interpretable Deep Representation of the Longitudinal Electronic Health Record, IEEE Access 6 (2018) (2018) 65333–65346, https://doi.org/10.1109/ACCESS.2018.2875677.
[321] Lei Zhang, Hailin Hu, An Xiao, Sai Zhang, Yangyang Li, Xuanling Shi, Tao Jiang, Linqi Zhang, and Jianyang Zeng. 2019. DeepHINT: Understanding HIV-1 integration via deep learning with attention. Bioinformatics 35, 10 (2019), 1660–1667. 10.1093/bioinformatics/bty842.
[322] Wen Zhang, Feng Liu, Longqiang Luo, and Jingxia Zhang. 2015. Predicting drug side effects by multi-label learning and ensemble learning. BMC Bioinformatics 16, 1 (2015). 10.1186/s12859-015-0774-y.
[323] Y.u. Zhang, P. Tino, A. Leonardis, K.e. Tang, A Survey on Neural Network Interpretability, IEEE Transactions on Emerging Topics in Computational Intelligence 5 (2021) 726–742, https://doi.org/10.1109/TETCI.2021.3100641.
[324] Yue Zhang, David Defazio, and Arti Ramesh. 2021. RelEx: A Model-Agnostic Relational Model Explainer. In AIES 2021 - Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, 1042–1049. 10.1145/3461702.3462562.
[325] G. Zhao, B.o. Zhou, K. Wang, R. Jiang, X.u. Min, Respond-CAM: Analyzing deep models for 3D imaging data by visualizations, In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 485–492 (2018), https://doi.org/10.1007/978-3-030-00928-1_55.
[326] Juanping Zhao, Mihai Datcu, Zenghui Zhang, Huilin Xiong, and Wenxian Yu. 2019. Contrastive-regulated CNN in the complex domain: A method to learn physical scattering signatures from flexible polsar images. IEEE Trans. Geosci. Remote Sens. 57, 12 (2019), 10116–10135. 10.1109/TGRS.2019.2931620.
[327] Lue Ping Zhao and Hamid Bolouri. 2016. Object-oriented regression for building predictive models with high dimensional omics data from translational studies. J. Biomed. Inform. 60, (2016), 431–445. 10.1016/j.jbi.2016.03.001.

[328] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning Deep Features for Discriminative Localization, in: In ProceedIngs of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016, pp. 2921–2929, https://doi.org/10.1109/CVPR.2016.319.
[329] Yichen Zhou and Giles Hooker. 2016. Interpreting Models via Single Tree Approximation. arXiv Prepr. arXiv1610.09036 (2016). Retrieved from http://arxiv.org/abs/1610.09036.
[330] Jasper Zuallaert, Fréderic Godin, Mijung Kim, Arne Soete, Yvan Saeys, and Wesley De Neve. 2018. Splicerover: Interpretable convolutional neural networks for improved splice site prediction. Bioinformatics 34, 24 (2018), 4180–4188. 10.1093/bioinformatics/bty497.