Cognitive Computation (2024) 16:45–74
https://doi.org/10.1007/s12559-023-10179-8

Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence

Vikas Hassija[1] · Vinay Chamola[2] · Atmesh Mahapatra[2] · Abhinandan Singal[3] · Divyansh Goel[3] · Kaizhu Huang[4] · Simone Scardapane[5] · Indro Spinelli[6] · Mufti Mahmud[7,8,9] · Amir Hussain[10]

Received: 16 March 2023 / Accepted: 10 July 2023 / Published online: 24 August 2023

Abstract

Recent years have seen a tremendous growth in Artificial Intelligence (AI)-based methodological development in a broad range of domains. In this rapidly evolving field, large number of methods are being reported using machine learning (ML) and Deep Learning (DL) models. Majority of these models are inherently complex and lacks explanations of the decision making process causing these models to be termed as 'Black-Box'. One of the major bottlenecks to adopt such models in mission-critical application domains, such as banking, e-commerce, healthcare, and public services and safety, is the difficulty in interpreting them. Due to the rapid proleferation of these AI models, explaining their learning and decision making process are getting harder which require transparency and easy predictability. Aiming to collate the current state-of-the-art in interpreting the black-box models, this study provides a comprehensive analysis of the explainable AI (XAI) models. To reduce false negative and false positive outcomes of these back-box models, finding flaws in them is still difficult and inefficient. In this paper, the development of XAI is reviewed meticulously through careful selection and analysis of the current state-of-the-art of XAI research. It also provides a comprehensive and in-depth evaluation of the XAI frameworks and their efficacy to serve as a starting point of XAI for applied and theoretical researchers. Towards the end, it highlights emerging and critical issues pertaining to XAI research to showcase major, model-specific trends for better explanation, enhanced transparency, and improved prediction accuracy.

Keywords Machine learning · XAI · Black-box models · Interpretability · Transparency · Responsible AI

✉ Vinay Chamola
vinay.chamola@pilani.bits-pilani.ac.in;
f20190560@pilani.bits-pilani.ac.in

✉ Mufti Mahmud
mufti.mahmud@ntu.ac.uk; muftimahmud@gmail.com

Vikas Hassija
vikas.hassijafcs@kiit.ac.in

Abhinandan Singal
abhi.singal7@gmail.com

Divyansh Goel
divyanshgoel10@gmail.com

Kaizhu Huang
kaizhu.huang@dukekunshan.edu.cn

Simone Scardapane
Italysimone.scardapane@uniroma1.it

Indro Spinelli
indro.spinelli@roma1.infn.it

Amir Hussain
A.Hussain@napier.ac.uk

[1] School of Computer Engineering, Kalinga Institute of industrial Technology, Bhubaneswar, India

[2] Department of Electrical and Electronics, and APPCAIR, BITS-Pilani, Pilani Campus 333031, India

[3] Department of Computer Science and IT, Jaypee Institute of Information Technology, Noida 201304, India

[4] Duke Kunshan University, Jiangsu 215316, China

[5] Sapienza University of Rome, Rome, Italy

[6] INFN Sezione di Roma, Rome, Italy

[7] Department of Computer Science, Nottingham Trent University, Clifton Lane, Nottingham NG11 8NS, UK

[8] Computing and Informatics Research Centre, Nottingham Trent University, Clifton Lane, Nottingham NG11 8NS, UK

[9] Medical Technologies Innovation Facility, Nottingham Trent University, Clifton Lane, Nottingham NG11 8NS, UK

[10] School of Computing, Edinburgh Napier University, Scotland, UK

Introduction

A Brief Historical Perspective

Artificial Intelligence (AI) technology is finding its way into transforming various application domains [1, 2]. These machine learning (ML) and deep learning (DL)-empowered methods are proving their dominance with their utilisation going from automated chess-playing computers to self-driving cars. The implementation of DL-based methods in the field of computer vision (CV) has been very successful and outperformed traditional methods. Human beings have been defeated for the first time in related open challenges [3] (e.g., ImageNet image-classification [4], COCO object-detection [5]) since the introduction of AlexNet [6]. Following their success, DL methods gradually got employed in numerous fields, such as natural language tasks which involved automated translation [7] and visually-guided question answering [8]. One of the significant milestones was witnessed in 2016 when an AI player AlphaGo [9, 10], was able to defeat the human world champion Sedol Lee in a game of Go. Since then, deep reinforcement learning (RL)-based programs and applications have been developed to determine the degree to which they can compete with human champion players in various games like Texas hold'em poker [11] and Dota 2 [12]. Taking into consideration the revolutionary success of AI, researchers predict that it will snowball in the next few years, reaching $190.61 billion market value in 2025 [13–15]. Consequently, the proliferation of AI has made people think, "How comfortable are we trusting blindly on these AI-generated predictions and results? Who will be held accountable when things go wrong?" It is essential to note that the highly efficient predictions of AI models are derived from Deep Neural Networks (DNNs) which originate from extremely complex non-linear statistical models and innumerable parameters, thus compromising the aforementioned algorithms' transparency [16, 17]. Due to this, AI algorithms suffer from opacity i.e., the situation in which a system is unable to offer any reason or suitable explanation involved behind its decisions, commonly referred to as "the black-box problem." A black-box nature is poorly understandable by humans. Entrusting crucial decisions to a black-box model creates a necessary need for AI algorithms to be explainable for their decision-making process [18].

Over the last few years, ML-based predictions have faced a lot of scepticism, especially when it comes to life-changing decisions such as the early detection of a terminal disease in the healthcare field or AI-engineered military drones. The topic of how to explain ML predictions has generated a lot of discussions [19]. Methods and techniques have advanced at such a rapid rate that a new field has been created around them: explainable artificial intelligence (XAI). The field seeks to develop AI systems that not only provide accurate predictions but also provide explicit and interpretable explanations for their decisions and actions, thereby making them more trustworthy for human users.

XAI aims to equip engineers with extensive resources to understand the elusive black-box nature of AI, emphasising transparency and the interpretability of AI models employed to reach conclusions. [20].

Objectives and Structure

Existing surveys [21–24] focus on answering "What, Why, and How" to encompass all aspects of XAI. The "What" tries to explain the existing definitions of explainable AI and the importance of explaining a user's role. The "Why" provides an overview of key factors driving XAI research, including building trust, meeting regulatory requirements, identifying bias, ensuring generalisation in AI models, and debugging. The "How" section examines the methods for attaining explainability before the modeling stage, including techniques for thoroughly understanding and documenting the datasets utilised in modeling. Meanwhile, only some of them have attempted to adopt a more formalistic approach towards addressing the taxonomies and evaluation metrics of XAI. Unlike the existing literature and studies around XAI, we reflect upon a multidisciplinary approach towards research and development in this emerging field that opens up a new path towards exploring the importance of human–computer interaction (HCI) skills for making transparent models [25, 26]. The existing methodology surveys the need to keep humans in the loop as a key prospect to consider while making the ML model human-understandable. Following that, we come to the idea of responsible artificial intelligence, which is an AI that takes into account societal values as well as moral and ethical issues to improve the usability of AI models in real-world applications [27, 28]. Figure 1 shows the overall structure of this survey.

We organise this survey study with the following contributions to open up new directions for future research:

- This paper provides a completely revised hierarchical taxonomy that lays down the building block for future researchers to learn about the key aspects of XAI.
- Using popular works in this field, we stress on the latest findings about XAI and its implications.

This article is organised as the following: the "Introduction" section introduces fundamental concepts and background. After a deep understanding of taxonomy, we proceed to the "The Need for XAI" section, which addresses the need for XAI and how it can promote explainability and trustfulness in AI. The "XAI Evaluation Framework" section

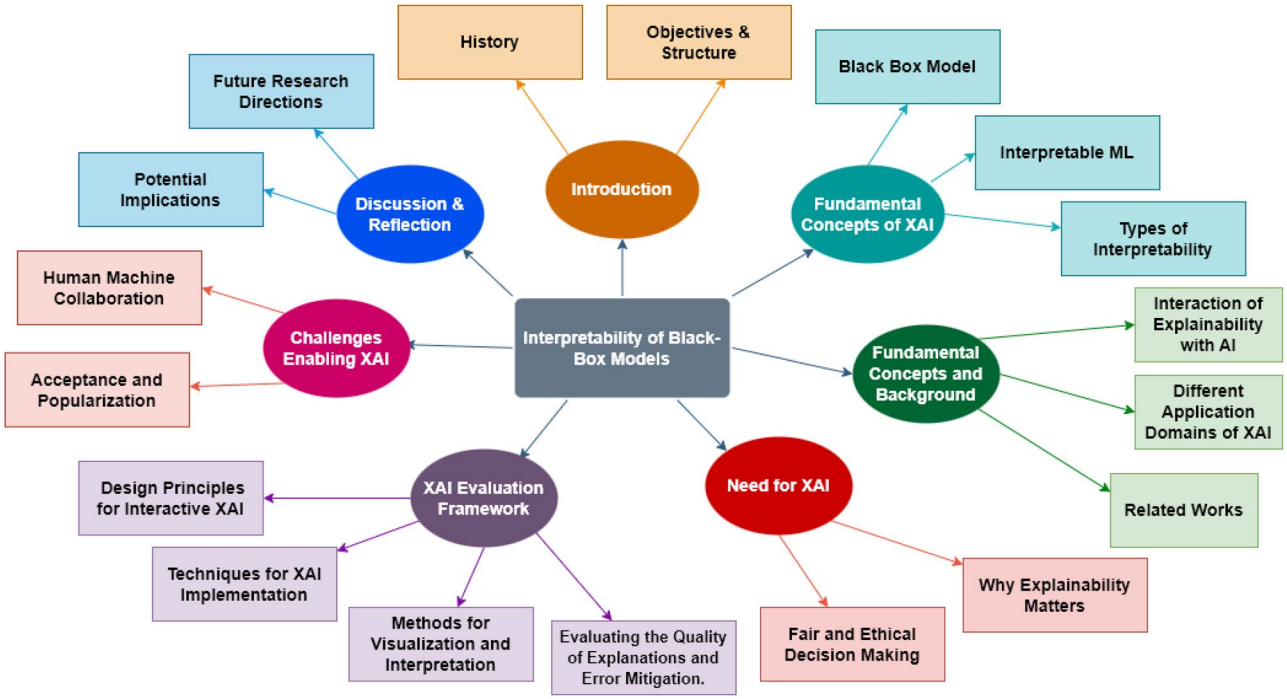


Fig. 1 Overview of the survey

provides a comprehensive and in-depth evaluation of the XAI framework by reviewing the techniques for XAI as well as inspecting their efficacy. The "Challenges for Enabling XAI" section points out the main challenges of enabling XAI. The "Discussion and Reflection" section provides a holistic discussion and reflection on XAI's prospects and implications. The survey paper is finally concluded in the "Conclusion" section.

Fundamental Concepts of XAI

Since the early 1980s, research scholars have been seeking to expand the scope of AI by incorporating the explainability factor and enhancing trust among end-users. With XAI now being recognised as a necessity rather than just a choice, explanation methods have come a long way, from textual formats to visual aids [29, 30]. However, to ensure that the concept of XAI is fully grasped, we put forward in the following section the underlying concepts that will serve as a solid baseline for comprehensive and up-to-date navigation of this rapidly growing area of research. Distinctively, we first address the black-box problem which led to the beginning of XAI. It is followed by an approach aimed at making ML algorithms more interpretable. Finally, we put forward the different types of interpretability offered to achieve explainability.

Black-Box Model

A black-box model in XAI refers to a machine learning model that operates as an opaque system where the internal workings of the model are not easily accessible or interpretable. These models make predictions based on input data, but the decision-making process and reasoning behind the predictions are not transparent to the user [31]. This lack of transparency makes it strenuous for users to understand the model's behavior, detect potential biases or errors, or hold the model accountable for its decisions.

In XAI, the term "black box" is often used to contrast with "white box" or "transparent" models, where the internal workings and reasoning behind the predictions are easily accessible and interpretable. Overall, it helps users deeply understand and trust the decisions made by these systems. In general, highly successful prediction models, such as DNNs, have some inherited drawbacks in terms of transparency that need to be addressed to justify the use of these models in many scenarios.

Interpretable Machine Learning

The first thing that springs to mind whenever black-box models are brought up in a conversation is always a basic interpretation of these models. When ML models are utilised in a product, interpretable systems are frequently a decisive

element. In machine learning, interpretability is a crucial component. Nevertheless, it is still unclear how to quantify it. Because of this ambiguity, academics frequently conflate the terms "interpretability" and "explainability." Only when machine learning models are explicable can they be audited and debugged. Even in a trustworthy field, like movie reviews, it is difficult to interpret whether a review is positive or negative because the movie rating and the emotion do not match [32, 33]. When a product is put into use, things can go wrong. An incorrect prediction's interpretation aids in determining its root cause. It provides guidance on how to repair the system. An excellent (artificial) example of ambiguity is the task of classifying wolf vs. Siberian husky from [34], where a DNN is shown to incorrectly label some canines as wolves. The experiment predicts a "Wolf" if there is snow and a "Husky" otherwise, regardless of animal color, position, pose, etc. The experiment begins as follows: First, a wolf without a snowy background is presented (which is classified as a husky) and then one husky with a snowy background is presented (which is classified as a wolf) [34].

Another example of an incorrect prediction by ML that could be fixed by interpretability is the case of a deep learning model that was developed to predict which patients would benefit from an antidepressant medication called escitalopram [35]. A large set of clinical data, including patient demographics, symptom severity, and genetic information, was used to train the model. However, when the model was evaluated on a new set of patients, in some instances, it made inaccurate predictions. In particular, the model predicted that some patients who benefited from the medication would not, and vice versa. This could have severe consequences for patients, as prescribing the incorrect medication could result in ineffective treatment and potentially dangerous adverse effects. The researchers utilised the SHapley Additive exExplanations (SHAP) technique to construct an interpretable version of the deep learning model for predicting treatment outcomes in depression. SHAP is a procedure that can be applied to any machine learning model in order to provide explanations for specific predictions. Using SHAP, the researchers were able to identify the most influential factors influencing the model's predictions for each patient. These characteristics included demographic variables such as age and gender, in addition to genetic markers associated with treatment response. By providing these explanations to clinicians, the researchers hoped to increase the accuracy and reliability of the model's predictions and to identify potential errors or biases in the underlying data. With approximately 70% accuracy, the interpretable version of the model was able to identify patients who were more likely to benefit from escitalopram.

Modern methods are being created every day to make AI more understandable. Trying to keep up with everything that is published would be absurd and impossible.

Types of Interpretability

The degree to which a person can comprehend and foresee the results of an ML model is known as interpretability. To date, numerous frameworks have been proposed for achieving interpretability which justifies their work through one criterion or another [36–38]. Tjoa et al. [39] proposed two major classes of interpretability, i.e., perceptive interpretability and interpretability by mathematical structures.

- Perceptive interpretability unifies all the interpretabilites that are well perceived by humans as they generally provide visual evidence. However, this obvious nature of the class lacks in fulfilling the true motive behind XAI because the black-box algorithm is yet to be unboxed. One of the integral methods to achieve perceptive interpretability is saliency which formulates its explanation based on the relative importance of all the input features. The resultant values could be in the form of probabilities (the LIME model [34]), superpixels (ACE algorithm [40]), and heatmaps (CAM and LRP [41–45]).
- Interpretability by mathematical structures unifies all the interpretabilities that reveal the mechanisms behind deeper layers (which store all the complex information) of NN algorithms [46]. An example of this approach is testing with concept activation vectors (TCAV) [47]. Several other methods, such as t-distributed stochastic neighbor embedding (t-SNE) and correlation-based singular vector canonical correlation analysis (SVCCA) [48], play a significant role in directing towards the subspace of input for error-free predictions.

The effectiveness and efficiency of XAI models and strategies can depend on many factors. A few of the components of a black-box model that play a role in Interpretability are Model Architecture, Feature Selection, and even Explainability Techniques itself.

The model's architecture can have a substantial effect on its interpretability. Some architectures, including decision trees and rule-based models, are intrinsically more interpretable than others, including deep neural networks. The selection of model features can also affect its interpretability. Using readily understandable and explainable features can make the model more interpretable, whereas using complex or abstract features can make it more challenging to comprehend. The specific techniques used to generate explanations for the decision-making process of the model can also play a significant role in interpretability. Techniques such as saliency maps, feature importance scores, and counterfactual explanations can assist users in comprehending the model's decision-making process. Even the design and usability of the model's user interface can influence its interpretability

Table 1 List of abbreviations

Abbreviation	Description
XAI	Explainable artificial intelligence
FAT	Fairness-accountability-transparency
ML	Machine learning
DNN	Deep neural networks
AGI	Artificial general intelligence
MLP	Multilayer perceptrons
IAI	Interpretable artificial intelligence
DL	Deep learning
CNN	Convolutional neural network
RNN	Recurrent neural networks
HCI	Human–computer interaction
NLU	Natural language understanding
ANI	Artificial narrow intelligence
NLP	Natural language processing
NN	Neural network

(Table 1). Providing plain and intuitive visualisations of the model's decision-making process can assist users in comprehending its behavior and gaining confidence in its outputs.

Fundamental Concepts and Background

Interaction of Explainability With AI

Explainability is becoming increasingly important as AI systems are being used to make decisions that have significant impacts on people's lives, such as in healthcare, finance, and criminal justice [49]. Here are some ways that explainability is influencing AI:

- Model interpretability: There is a growing focus on developing AI models that are interpretable, meaning that their decision-making process can be understood and explained to users.
- Regulatory requirements: In some industries and regions, regulations are being introduced to require AI systems to be explainable in order to ensure accountability and transparency.
- Trust and adoption: Explainability can play a role in building trust in AI systems, which is crucial for their widespread adoption and use.
- Model validation: Explainability can help validate the decisions made by AI models, ensuring that they are free from biases and errors.
- Debugging and improvement: Explainability can provide insights into how AI models are making decisions, making it easier to identify and address sources of error or bias.

Overall, the interaction of explainability with AI highlights the importance of developing AI systems that are transparent, interpretable, and trustworthy in order to ensure their responsible and effective deployment in various domains [50–52].

Different Application Domains of XAI

It is natural to state that the importance of a thing is derived from its degree of need in society. Considering the significant increase in the emphasis on explainability in AI algorithms, we put forward some of the critical domains where XAI can prove to be a revolutionary change (Fig. 2).

Automated Transport

Automated transportation has gone from being a topic of science fiction to being a reality thanks to the development of AI [53, 54]. Artificial intelligence is undoubtedly the most significant and complex part of self-driving cars, with a huge potential to eliminate unsafe driving behaviors.

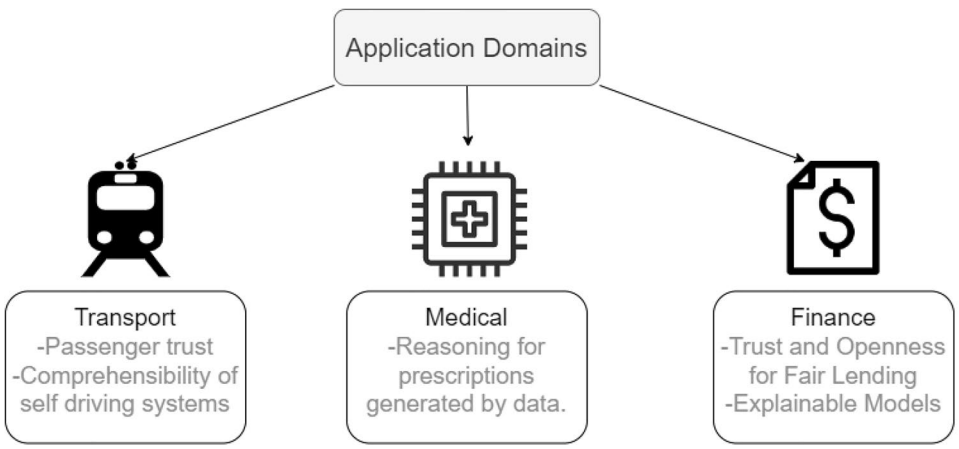


Fig. 2 XAI can have deep impacts in mission critical fields, such as finance, healthcare and public service

However, a technology that may take the role of a human's cognitive and motor talents is currently lacking in dependable and secure autonomous cars. A 2016 incident in which a Tesla on autopilot attempted to plow through an 18-wheeler truck trying to cross the highway lends credence to the aforementioned claims. According to the National Highway Traffic Safety Administration, the vehicle was unable to discern between the white side of the truck and a brightly lit sky, preventing the need to brake that may have caused the passengers to suffer significant injuries [55]. These tragic events demonstrate the importance of self-driving systems being comprehensible.

Recent times have observed some of the notable works attempting to theoretically build humans' trust in self-driving vehicles. Mittu et al. [55]. study revealed that when an example supported the decision of a self-driving car abruptly changing lanes, passengers were more comfortable. Petersen et al. [56] specifically focused on trying to increase passengers' trust through augmentation of their situational awareness and consideration of actual applications of semi-autonomous vehicles. Contrary to popular belief, Haspiel et al. [57] presented another point of view by investigating the influence of explanations on the trust-building process that were provided at different time frames. The study successfully proved that explanations generated before vehicle-taking actions yielded the best trust among humans.

Medical

A significant body of work on automated diagnosis and prognosis using machine learning has lately acquired popularity in the medical sector [58, 59]. The issue of interpretability holds significantly more magnitude in the medical area than just intellectual curiosity. A range of elements that other fields typically neglect must be taken into account more carefully when making critical decisions where human lives are at stake.

Extensively data-driven and context-sensitive AI predictions are being used as decision support for delirium, a highly relative syndrome in elderly patients. One example of this is the use of machine learning algorithms to analyse electronic health records (EHRs) to identify patients at risk of delirium. These algorithms can take into account a wide range of patient-specific data, including demographic information, laboratory test results, and medication use, to make predictions about a patient's risk of delirium [60].

Although these models are highly optimised on the training data set using [61, 62], user acceptability remains under question. Therefore, an explanatory component capable of providing appropriate reasoning for its predictions was introduced. Supported by the patient's medical history, it played a pivotal role in establishing confidence and increasing patient safety [63]. Thus, it can be inferred that the healthcare industry is one of the few industries where the accuracy and explainability of AI algorithms need to go hand in hand to gain user acceptance.

Financial

As AI developed over the years, the modern era saw financial institutions adopting these algorithms to reduce risks and optimise efficiency to supplement banking practices such as Anti-Money Laundering (AML), Counter-Terrorism Financing (CTF), risk management, and market abuse. Parallel to this developments, financial giants firmly believe that interpretability and explainability are essential prerequisites for the use of AI models in thin, highly regulated, opaque, and uncontrollable sectors [64]. Thus, relying on black-box models for crucial decisions such as "fair lending" could pose some severe problems. Explainability of the results and functionality of AI systems have appeared as an obligation since the financial sector is subject to higher societal requirements for trust and openness. Sincere attempts have been undertaken to increase the transparency of these financial algorithms' decision-making capabilities. Lecue and Wu [65] successfully predicted abnormal expenses through combined knowledge of Semantic Web and ML technologies. The study of generated data through visualisation was an integral part of the Artificial Intelligence Finance System (AIFS). Akur8 [66] used Generalised Linear Models (GLM) to devise interpretable as well as automated systems to determine insurance prices, with the main objective being to achieve a competitive level of performance.

Despite the aforementioned application fields, XAI is also finding use in a number of additional sectors as explainability's significance and requirement grow progressively each day. The Future of Privacy Forum [67] cites a number of areas of life where automated decisions could be risky and where justification could make them wise and reliable choices.

Related Works

A thorough and systematic survey is still lacking despite the large number of surveys that have been conducted to understand the explainability and interpretability of AI. According to researchers, there are currently only a few research papers in this area that give us a brief overview of the best XAI practices.

An early survey on the matter by Angelov et al. [68] provides an analytical review of explainable artificial intelligence (XAI) methods with a focus on developing AI models that are transparent and interpretable to humans. The authors provide an overview of the current state of XAI, including a review of the various techniques and approaches proposed for achieving explainability in AI models. In addition, they

discuss the difficulties and limitations of current approaches, such as the trade-off between accuracy and interpretability and the difficulty of determining the efficacy of XAI methods. The review highlights the potential advantages of XAI for facilitating greater human-AI collaboration and enhancing AI systems' credibility and accountability.

Guidotti et al. [69] have conducted a comprehensive assessment of approaches for illuminating black-box models that combine machine learning and data mining. They provided a thorough taxonomy that classified the different challenges encountered. Even though the survey was extremely clear about the concepts underlying the idea of explainability, the absence of evaluation as an essential explainability component suggested its inadequacy.

Dosilovic et al. [70]. provided a general review of XAI in their paper's closing remarks. They discussed improvements in machine learning models' explainability, paying particular attention to DNNs. This essay demonstrates how human intelligence and artificial general intelligence are interdependent and offers the reader fresh ideas.

Table 2 summarises the contributions of various works done in the field, along with their strengths and weaknesses.

A general overview of XAI is provided, along with a detailed breakdown of its contributions, as seen from several angles. While investigating various explicable techniques, we adhere to cognitivism and clarity while exploring different explainable approaches.

The Need For XAI

Why Explainability Matters

With the primary goal of building trust and confidence when putting AI models into production, XAI was developed. The newly developed AI discipline may be able to effectively explain AI results irrespective of any potential biases and prejudices. However, As seen in Fig. 3, the results of these AI Models (Black Box) often give Unexplainable, Unjustifiable, and Unaccountable results. In this section, we shall dig deeper into the root causes that lead to the need for XAI.

Need For Reasoning

The reasoning for a decision made by AI algorithms mainly involves providing explanations and justifications for that particular outcome. Humans generally look for reasoning rather than an incomprehensive description of the inner workings of the algorithms and logic behind the decision-making process.

XAI holds the promise to provide the equivalent predictions as black-box models without their drawbacks i.e., lack of explainability. It also reaffirms that besides being accurate, these predictions are trustworthy and accountable for their decisions. With the aim of providing an explanation for legal judgment predictions generated by Legal Judgement Prediction (LJP), Zhong et al. [71] designed the QAjudge. The model used the concept of reinforcement learning, which followed a process of asking a series of human-readable questions and then generating explanations based on the responses by humans.

Further, much more than societal needs, the legislation demands AI to be explainable under the "right to explanation" act. This act was included in the General Data Protection Regulation (GDPR) that comes into effect across the EU on 25 May 2018 [72].

Need For Innovation

Explanation fulfils the constant desire to innovate for more effective algorithms and sophisticated neural networks. Recent years have also witnessed humans discovering that these algorithms may learn novel concepts and ideas. This newly gained knowledge can be applied to open a new course of action and hidden laws in various fields like neuroscience, astronomy, etc. Medical research has also confirmed the predictions and early conclusions that were presented by the LIME model and various DNNs, which were not possible by humans at that time [78]. Doctors are now able to use medical image technology to diagnose diseases and understand the patient's condition [79, 80]. XAI helps us to get knowledge about the hidden layers of some black-box models which are neural networks that fosters a closer link between humans and machines.

Need For Regulation

Despite the excellent performance of DNNs in predicting outcomes, it is shown to be fragile and vulnerable to adversarial perturbation [81–85]. The Local Interpretable Model-Agnostic explanation (LIME) model, when subjected to such perturbations, developed self-learned patterns that produced false results [34]. It creates synthetic data based on the input data, trains a basic ML model (with the synthetic data) that performs similarly to the complex black-box model, and utilises the weights of this model to determine the significance of features.

Need For Advancement

At present, advancements in technology such as AI are aimed towards gaining acceptance and becoming flawless. This highlights that the key to the success of AI is

Table 2 Related work on explainable approaches

References	Contributions	Strengths	Weaknesses
van der Velden et al. [73]	It contributes to improved trust and transparency in medical decision-making, allowing medical experts to understand why a specific diagnosis was made, leading to more accurate diagnoses and better patient outcomes.	Trends and future perspectives for XAI in medical image analysis are identified.	Struggles to cover all work in the field.
Rudin et al. [74]	Identifies the top 10 challenges in interpretable ML along with their comprehensive background and solution approach.	Focused in assisting readers and new researchers to steer clear through common but problematic techniques related to interpretability in AI.	No discussion on socio-technical challenges, human–computer-interaction challenges and how robustness, as well as fairness, interact with interpretability.
Abdul et al. [75]	The paper reveals fading and burgeoning trends in explainable systems and identifies closely related domains or mostly isolated.	Investigates how HCI researchers can help to develop accountable systems.	Lack of a range of philosophical theories about XAI and deep dive to extract information was lacking.
Machlev et al. [76]	Highlights the potential of using XAI for power system applications.	This paper highlights the potential of using XAI for energy and power systems applications and covers the challenges and limitations of adopting and implementing XAI techniques in the field of energy and power systems.	The paper might favor specific industries or sectors and not take into account the needs and challenges faced by other industries using XAI in energy and power systems.
Li et al. [3]	The progress in methodology, evaluation, and application of XAI is covered.	A new hierarchical taxonomy is explained which introduces the use of previous knowledge of XAI.	Overview of external knowledge is lacking and there are multiple open and unanswered questions.
Tjoa et al. [39]	Aim to provide a comprehensive overview of the current state of XAI in the medical domain.	Aimed to give clinicians a perspective on the use of interpretable algorithms.	Less suitable for technically non-oriented readers due to some mathematical details.
Adadi et al. [77]	Proposes the main concepts of enabling explainability in intelligent systems.	Includes a detailed discussion on the challenges and limitations of XAI methods, which allows for a more realistic understanding of the field and the potential areas of improvement.	It mainly focuses on the technical side of XAI and does not give enough attention to the ethical and societal implications of XAI.

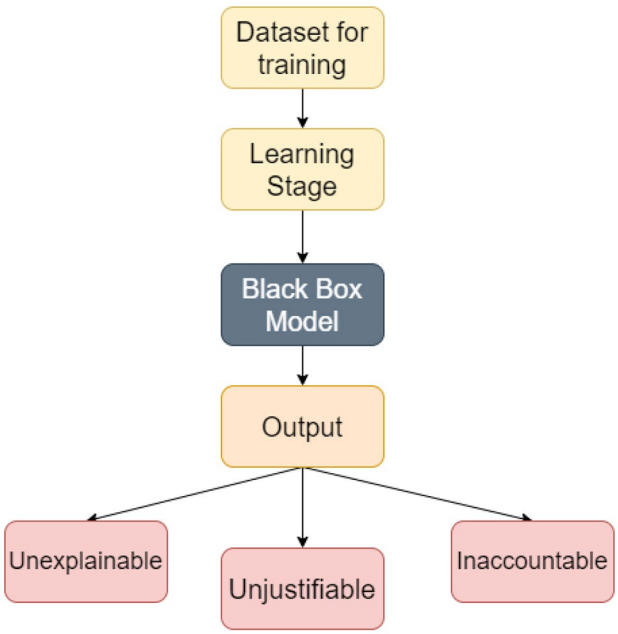


Fig. 3 Most ML models behave as black-box models

the continuous progress made in this field. With XAI, new insights into the system and understanding will guide engineers to improve specific parameters. One such example is the introduction of the CrystalCandle tool in a Linkedin software sales team [86], whose explanations increased its subscription revenue by 8%. Thus, XAI can act as a stepping stone for AI to achieve accuracy and recognition.

Fair and Ethical Decision Making

AI ethics are a set of principles that guide what is right or wrong, good, or bad in a way. The extant research on XAI has focused on technical issues, but it is more desirable to have a branch that deals with trade-off problems and ethical issues. Human judgment is imperfect and does not contain qualitative and quantitative simplification. On the basis of three components of the Independent High-Level Expert Group on AI (LEG), an algorithm can be said to be ethical [87].

- It should comply with all laws and regulations and be lawful.
- It should have good intentions and should be robust, both from a social and technical perspective.
- It should demonstrate respect and should adhere to all principle and values.
- For XAI to achieve this goal, it is important to carefully consider and address the key components of fair and ethical decision-making, including data quality, algorithmic bias, explainability, ethical principles, and human oversight.

Here are some methods to achieve algorithm fairness:

- Counterfactuals: This is a well-liked strategy for illuminating AI and making sense of algorithmic fairness. An excellent application of counterfactuals is risk management. A classic example of this is a bank. A bank may use counterfactual analysis to evaluate the potential outcomes of a loan application under different scenarios. For example, if the borrower loses their job or if interest rates increase unexpectedly. By simulating these potential scenarios, the bank can better understand the risks associated with the loan and make more informed decisions about whether or not to approve the application.

Proxy fairness methods: These methods use proxies, such as demographic information, to correct for bias in AI systems. For example, a model trained on data that is not representative of a particular demographic group may be corrected by adjusting the model's output based on demographic information.

- Fairness through awareness: This approach involves training AI models on data that is specifically designed to capture relevant factors contributing to fairness and ethical considerations. This can help to ensure that the models produce fair and ethical outcomes.
- Fairness constraints: Fairness constraints are mathematical formulations of fairness criteria that can be used to optimise AI models. For example, a fairness constraint may be used to ensure that an AI model does not discriminate against certain demographic groups.
- Human-in-the-loop approaches: Human-in-the-loop approaches involve incorporating human oversight into the decision-making process of AI systems. For example, a human may be involved in validating or adjusting the decisions made by an AI model to ensure that they are fair and ethical.

XAI Evaluation Frameworks

Design Principles For Interactive XAI

Since the emergence of XAI, researchers have tried to address the multidisciplinary nature of the process involved in interpreting a black-box model. To gain continual advancement in this field, a multifaceted approach involving collaborative efforts from independent research horizons is advocated. In addition, a determinant factor is to adhere

to the demands and interests of multiple stakeholders: the designers, decision-makers as well as end consumers. Given the disciplinary efforts being carried out to keep humans in the loop, seeking a more formal and holistic standard of procedure for the explanation of opaque AI algorithms is the need of the hour. As a result, it is advised that the integration and application of interpretation in AI design and deployment workflows adhere to the principal principles listed below:

Adeptness to the User Behavior

XAI's effectiveness depends on the degree of motivation it brings about in the user to interact with the AI. Zhu et al. [88] noticed that instead of focusing on the practicality and efficacy of current XAI practices, most works are coming up with new methods of explanations. Thus, acknowledging the importance of user behavior will help the XAI models to better utilise their functions and parameters to interpret the black-box model.

Align Perception of XAI With Human Understandability

Human-centered design principles are considered a valuable resource in order to develop explanations that carry a contextual value with them. Developers should aim to involve the user in the early stages of the development process, which will ensure their active participation. Empowering users to self-explain the logic involved in an AI algorithm aids in making the XAI model more sensible as well as practical.

Collaborative Techniques Should Be Implemented More

An ideal explanation is one which incorporates expertise from multiple domains of knowledge. HCI skills such as philosophy, psychology, and cognitive science have proven their worth in the field of XAI by developing explanations that were able to stimulate the human explanation process [89]. It is imperative to shed some light on how some model-agnostic methods (later discussed in this section) combine the global and local scope of interpretability to reinforce the objective of explainable AI models.

Explanations Have More Dimensions Than Just Performance

Appropriately evaluating the capability of the explanation method goes beyond determining whether they just work or not. Since each explanation generated carries its own impact, they should be assessed along dimensions such as qualitative performance (satisfaction, trust, and understanding), achievement of the end task, mitigation, and error analysis. Designers should aim at developing an evaluation benchmark that could measure the qualitative and quantitative ability of the explanations. Goal-centric explanations should be encouraged, which would enable designers to decide beforehand what consequences the explanation will have on the AI algorithm.

Contradictions Provide Alternative Approaches

The utilisation of contradictions and counterfactuals is termed as one of the best practices that XAI designers need to involve in their development process. One example of this statement is the development of AI-based medical diagnostic systems [90]. Utilising contradictions and counterfactuals in the design process can help ensure that the diagnostic system is able to identify and handle rare or unusual cases, which can be critical in medical diagnosis. For example, by providing the system with counterfactual examples of patients who have similar symptoms but different diagnoses, the system can learn to identify the unique characteristics that distinguish one diagnosis from another. Additionally, by exposing the system to contradictory examples where the same symptoms can indicate different diseases, the system can learn to evaluate multiple possibilities and make a more accurate diagnosis [91]. This can increase trust in the system's diagnosis and improve patient outcomes. Besides giving valuable knowledge about the system's limiting conditions, they also help in discovering its vulnerabilities.

Are Explanations Always Important?

In the spirit of holism, Bunt et al. [92] originally raised this question in their work mainly due to the efficiency of explanation techniques deployed in systems that offer users low-cost decisions. They found that although these systems were opaque and provided no reasoning behind their predictions, they were still positively perceived by the users. Therefore, an important question was raised: "Does the cost of getting an explanation outweigh its benefit."

Amendments Go a Long Way

To date, XAI is considered a new concept in the technological field. As more and more research is done, the gained experience will help in changing the beliefs of people. Therefore, it is crucial to understand that the process of generating an explanation is never a "one-off." Especially from the viewpoint of dynamic environments, constantly flourishing users with modified explanations will ultimately advance the state of the art of explainability [93]. Addressing questions such as "How do the modifications affect the algorithm" and "Why should I consider this explanation instead of the previous one" will ensure positive feedback and will benefit XAI

designers given their long-term interaction with the users. So, along with a strong foundation, necessary amendments will always help in adding more to the structure of XAI, which will lead to a greater impact [94, 95].

To sum up, we distilled a range of design principles that have been suggested for the design and development of explainable systems. However, simply making out some guidelines does not necessarily guarantee success. While implementation of some principles is ongoing, some have resulted in no progress also. XAI designers have stated that in order to get strong conclusions in explainability, rigorous research and testing will be required, accompanied by reasonable pieces of evidence and relevant work context.

Techniques For XAI Implementation

While designing an explanation method for a typical ML model, it is expected from the method to answer some foundational questions: Why the model produced this prediction, and what are the logic and reasoning involved behind the model's decisions? However, as progress towards advancement continues, researchers came across other questions that could not be answered by the current design of the explanation method. Hence, different types of explanations were designed to serve the purpose of distinct behaviors, problems, and types of users. Since a typical end-user is focused on case-specific (local) interests and a domain expert needs to have a full understanding (global) of the AI prediction, the method of explainability needs to be exhaustive in its explanation.

It is interesting to note that most of the preliminary work in XAI is done from the perspective of an explainer, for example, a domain expert who was able to understand the working of an AI prediction and the logic used behind it. Conversely, methods to provide a satisfactory explanation to an explainee, for example, end-users, are rarely seen. This hindrance to exhaustive explanation stems from the fact that domain experts currently working in the field of developing AI models are the ones that also develop their XAI counterparts. Thus, their works in XAI seldom involve a focus on the explainee.

Therefore, as our framework suggests, it is important to keep end-users in mind while designing and developing XAI models.

From our conducted survey of the literature, we propose a categorisation of all the types of explanations which aim to interpret the logic of black-box algorithms. Figure 4 shows the general structure of our approach.

Based on Scope

On the basis of scope, interpretability mainly involves deriving explanations in either of the two directions: by considering a holistic view of all explanations of the model, i.e., global explanations, or by considering individual instances of explanations provided by the model, i.e., local explanations.

Global Interpretability

Global interpretable approaches are intended to make it easier to comprehend a model's overarching logic as well as the whole justification used to produce specific predictions. These strategies frequently aid large-scale decisions like drug usage or climatic changes because they concentrate on compiling explanations for all potential scenarios. We categorise the various approaches to achieving explainability globally into the following subclasses:

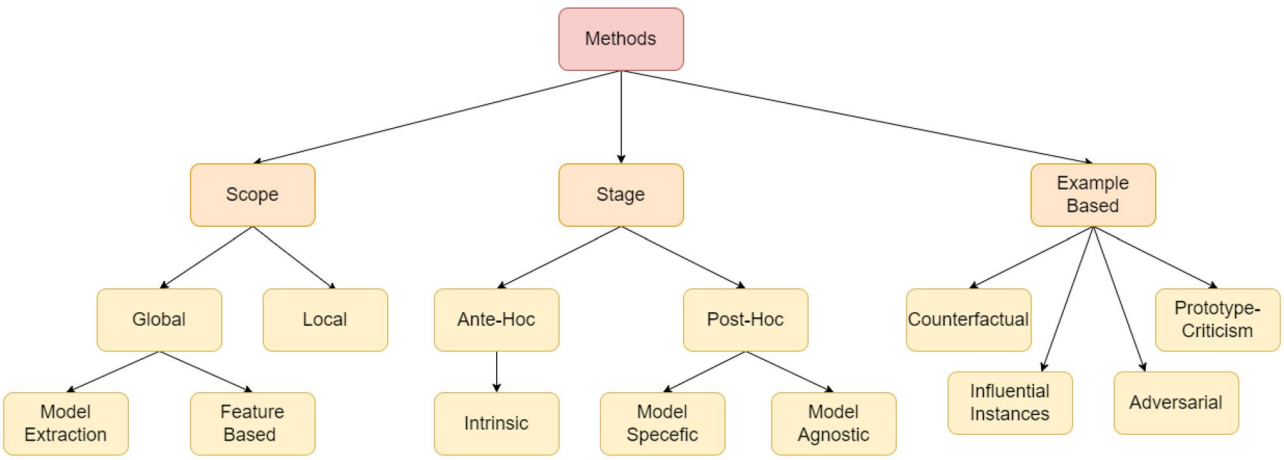


Fig. 4 Types of interpretabilities

Model Extraction In order to properly mimic the black-box model's judgments, model extraction entails training an interpretable model (such as a linear model or a decision tree) on the predictions of the black-box model. The global surrogate model is another name for the model produced by this procedure. The following are the steps needed to create a global surrogate model [3]:

- A desirable dataset X is chosen, which may either be the same black-box model's training dataset or a brand-new dataset with comparable distributions.
- The trained black-box model generates predictions from the chosen dataset.
- To fit the black-box model's predictions, an interpretable model is chosen and trained on dataset X.

The aforementioned procedure has been reproduced in other studies to draw information from the black-box models. Rule extraction and model distillation are the two strategies that are primarily used in the creation of model extraction algorithms.

Rule Extraction One of the most popular methods for extracting models from highly complex black-box models is rule extraction. An approach was provided in the study by Craven et al. [96] in 1994 that entailed the iterative creation and updating of a set of rules applied across the input and output classes of the ANN until all of the target classes have been processed. Ras et al. [97] proposed three modes to extract rules: (i) pedagogical rule extraction, (ii) decompositional rule extraction, and (iii) eclectic rule extraction. They did this by using the categorisation of different rule extraction methodologies provided.

The G-REX approach was used by Johanson et al. [98] to extract rules from genetic programming. To handle regression and classification issues using regression trees and fuzzy rules, respectively, the authors further created G-REX [99]. The Rule Extraction from Neural Network Ensemble (REFNE), developed by Zhou et al. [100], was able to prevent pointless discretisations by using adaptive intervals. To address categorisation issues, Biswak et al. [101] developed Rule Extraction by Reverse Engineering the Neural Networks (RxREN).

Model Distillation Only explainer-specific models with a narrow range of black-box techniques can be created using the rule extraction method. Hinton et al. [102] proposed transparent model distillation, a unified technique for model extraction, which was used to tackle the above-stated issue. Specifically, their paper reflected on how dark knowledge i.e., hidden knowledge from a complex and sophisticated model (teacher model) can be transferred to a simpler (student) model that will be able equally competent to the deep models in terms of predictions while at the same time being more interpretable.

Distillation turns out to be a more effective method for obtaining interpretable models. Tan et al. [103]. expanded on the notion of distilling intricate black-box models into transparent ones known as iGAMs through model distillation.

Che et al. proposed Interpretable Mimic Learning, a model distillation-based method that has the capacity to learn interpretable phenotypic traits for providing potent predictions [104]. Xu et al. [105] developed DarkSight, a visualisation technique for producing interpretations of a black-box model on a specific dataset, which was motivated by the idea of dark knowledge. This technique also supplemented the extraction of informative patterns and purposeful features from deep models.

Feature-Based Methods Although model extraction was successful in providing global explanations of black-box models, researchers found compromised accuracy due to the oversimplification of the model complexity. Hence, a method that could provide explanations as well as maintain the desired level of accuracy was needed. Methods looking at the impact or significance of input features in an algorithm were investigated to meet these needs. The proposed approach was further extended into two alternative paths, which were used to measure the input feature's relevance: (i) feature importance and (ii) feature interaction.

Feature Interaction The impact of one feature depends on the value of the other if an AI algorithm bases a forecast on two features. Due to how the individual feature effects interact with one another, the prediction that is created cannot be described as the total of these effects. Freidman established the H-statistic and discovered a mechanism to gauge the strength of feature interaction [106].

The following equation [107] is a mathematical representation of the H-statistic developed by Friedman and Popescu for interaction between any two features j and k. It collates the variation between the observed partial dependence (PD) function and the decomposed one without any interactions between the features.

A random forest is trained to predict cervical cancer and is examined for the interaction between its properties. The chart (Fig. 5) shows that IUD has the least proportional interaction effect with other variables compared to years of using hormonal contraception [107].

Feature Importance By calculating each feature's contribution to the predictions, a model's feature importance is calculated. One of the noteworthy methods is the permutation feature importance (PFI) by Breiman [108] for random forests. The degree to which the prediction error increases when the values of the feature are shuffled determines the importance. Fisher et al. [109] expanded on this notion by proposing Model Class Reliance, a model-independent

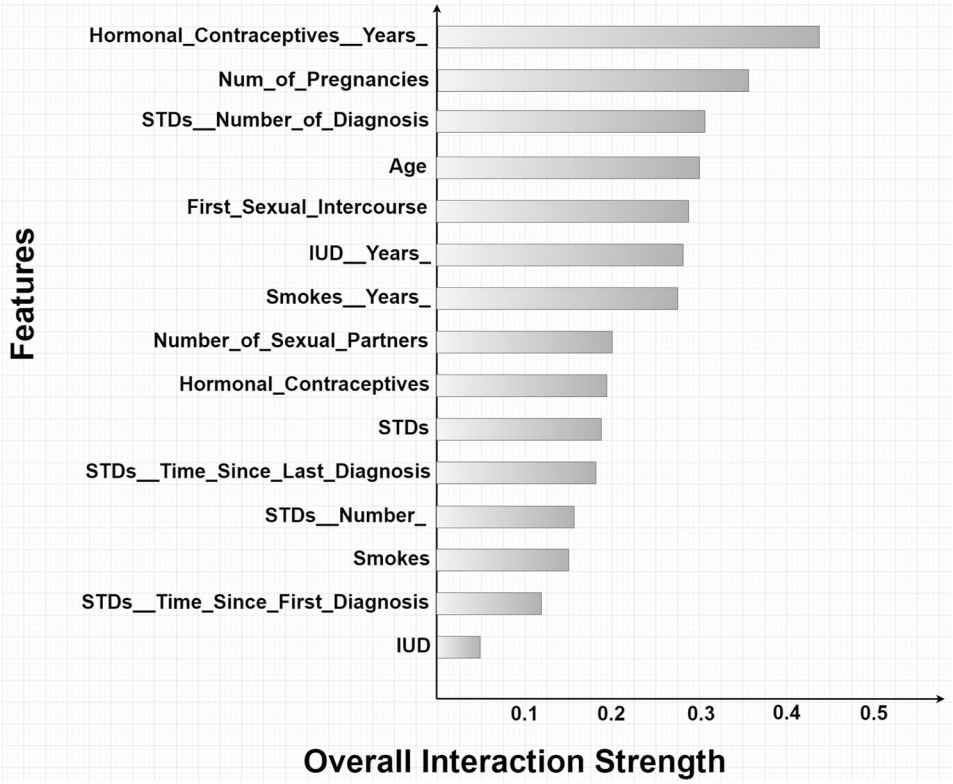


Fig. 5 The degree of interaction (H-statistic) between each feature and each other feature in a random forest that predicts the likelihood of cervical cancer

variant of feature significance theory (MCR). This paper presents an approach to model interpretability that offers several advantages. By studying an entire class of models simultaneously, the approach is able to provide a more comprehensive understanding of the relationships between variables and predictions compared to approaches that study individual models. This approach can also help to identify variables that are consistently important across different models, providing more robust and trustworthy explanations. However, one potential disadvantage of this approach is that it may be computationally expensive, as it requires training and evaluating multiple models. Another disadvantage is that the approach may struggle with handling complex models, especially those with many variables or non-linear relationships. The success of this work for explainability depends on the specific context and application. For some datasets and problems, this approach may provide more accurate and trustworthy explanations than other methods [110, 111].

At the same time, for other datasets and problems, other methods may be more suitable or perform better. It is important to note that interpretability and explainability are challenging and active areas of research, and there is no one-size-fits-all solution. The effectiveness and success of any interpretability method will depend on the specific context and

Local Interpretability

Local interpretability focuses on providing explanations separately for each choice and prediction rather than providing a detailed description of the intricate mechanism underlying the entire black-box model. In these models, each input feature is associated with a weight, and the final prediction is made by taking the dot product of the input features and their corresponding weights plus a bias term. This means that the importance of each input feature can be easily determined by looking at the corresponding weight. Additionally, it is possible to understand the impact of a specific feature by holding all other features constant and varying the feature of interest.

For example, in a linear regression model, if one wants to understand the effect of a specific feature on the predicted output, one can calculate the partial derivative of the output with respect to that feature and interpret it as the average change in the output for a one-unit change in that feature, holding all other features constant. Compared to global interpretability methods, developing explanations for black-box model local behavior is easier. However, this might not be always true. It is dependent on the black-box model. Furthermore, straightforward explanations are more useful than complex ones. Numerous research publications have put forth strategies to investigate local explanation

techniques. The next section provides a thorough overview of some of the key justification techniques found in the publications we analysed.

To develop local explanations for black-box models, Ribeiro et al. introduced the Local Interpretable Model-Agnostic explanation (LIME) [34]. It operates by training nearby substitute models to approximate certain model predictions. Below is a succinct explanation of LIME's working theory:

- From the predictions made by the black box, an instance of interest is chosen for which explanation is desired.
- A fresh dataset is formed consisting of perturbed samples, and their respective predictions are extracted from the black-box model.
- The new samples are weighted depending on their proximity to the instance of interest accordingly.
- Now, the black-box model can be explained via an interpretable model trained on the perturbed dataset with help of the following equation [107]:

$$\text{Explanation}\,(x) = \arg\min_{g \epsilon G} L\left(f,\, g,\, \pi_x\right) + \Omega(g) \quad (1)$$

It can be noted LIME generated surrogate models that have high local fidelity, meaning that they accurately represent the behavior of the black-box model in the local region around a specific instance. The goal of LIME is to create a simple, interpretable model that mimics the behavior of the black-box model in the vicinity of the instance being explained. By doing this, LIME provides an accurate and trustworthy explanation of the predictions made by the black-box model without requiring a deep understanding of its inner workings. The local fidelity of LIME-generated surrogate models is a key factor in their ability to provide useful and reliable explanations of black-box model predictions.

An extension of LIME can be found in anchors [112] proposed by the same creators. Anchor is a method that uses LIME to explain the predictions of machine learning models in a more efficient and scalable manner. The main idea behind Anchors is to precompute explanations for a subset of the instances in the dataset, called anchors, and then use these explanations to generate explanations for other instances. This approach reduces the computational cost of generating explanations and makes it possible to explain the predictions of large, complex models. The extension of LIME found in Anchors is designed to provide more accurate and efficient explanations of machine learning models while still preserving the local fidelity of the original LIME method.

Another significant work to be noted is by Ying et al. [113], who tried to devote more attention to the feasible local approximations for Graph Neural Networks (GNN). This work exhausted the limited information available to explain the predictions of GNN with the help of a sub-graph correlated with nodes and edges. Such work is renowned for becoming the first successful investigation in the field of local approximation of GNN [114, 115].

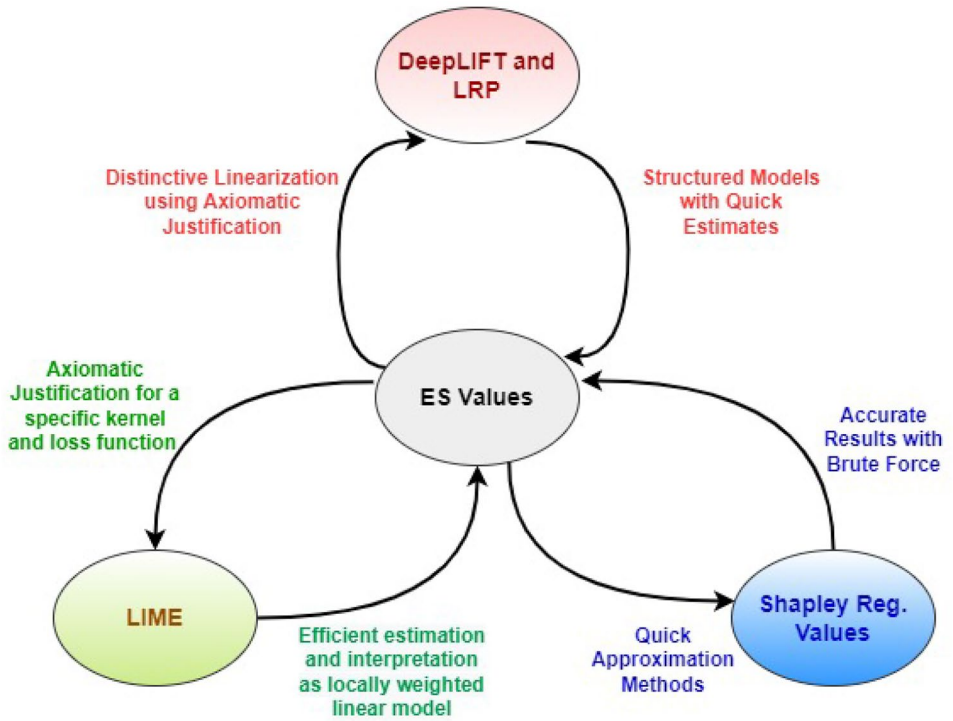


Fig. 6 Expectation Shapley (ES) values help connect different interpretation methods and provide a clear view of their correlation with each other

Further development was marked by the introduction of Shapley Values by Lundberg and Lee [116]. They used the preliminary information available from the work proposed by Štrumbelj and Kononenko [117] to design a new set of values called the expectation Shapley (ES) values which were able to unify and justify a broad spectrum of approaches (e.g., LIME, DeepLift, and Layer-Wise Relevance Propagation) for black-box model interpretations.

In Fig. 6, the arrows signify how different prediction methods gain the advantage over the ES values and vice versa [116]. The combined bubble of LRP and DeepLIFT symbolises their equivalency as proved by Shrikumar et al. [118].

An interesting feature to note about local explainability is that besides being the most applied forms of explanation methods for DNNs, they hold the potential to generate explanations for various classes of neural networks. Local explainability methods are indeed widely applied to DNNs and other types of machine learning models, as they provide insight into how the model makes specific decisions. Additionally, some local explainability methods are considered to be model-agnostic, meaning that they can be applied to a variety of different types of models. However, not all local explainability methods are model-agnostic, and other areas of research, such as model-specific explainability, also exist.

The aforementioned sections uncovered some crucial methods and strategies for achieving interpretability on a worldwide scale. Despite this, it appears that this mode of interpretability is difficult to use in practice, particularly for models with a lot of different parameters. Similar to how humans naturally seek out the logic and reasoning behind a particular model component in order to comprehend the entire model, local interpretability is more widely applicable and acceptable.

Based on Stage

The next category of model explanatory methods is based on the stage of interpretability. This includes ante hoc interpretations, which comprise classical approaches for analysing black-box models before their training, and post hoc interpretations, which comprise approaches for analysing black-box models after their training.

Ante-Hoc Interpretability

Ante Hoc interpretability techniques mostly consist of traditional AI practices. Designed with the idea of keeping an uncomplicated structure so that complexity is limited to a certain extent, such techniques are termed glass box techniques. Due to this, ante hoc interpretability has a close relation to intrinsic interpretability. Ante hoc interpretability mostly involves dealing with the data itself since data evaluation gives the much-needed insight and understanding of the model to be explained. Following are some commonly used explainable techniques:

Linear regression leverages a linear relationship among features of the input model as a means of providing an explanation for its predictions.

$$y = \beta_0 + \beta_{1x1} + \cdots + \beta_{pxp} + \epsilon \tag{2}$$

The weighted sum of the input features(β_i) [119] equals the projected outcome (y). The linearity of a model holds a significant weightage in its interpretation. However, this strategy is only shown to be workable for a small set of features. L1 and L2 regularisation provide a novel solution in handling the overfitting and feature selection for such cases [120, 121].

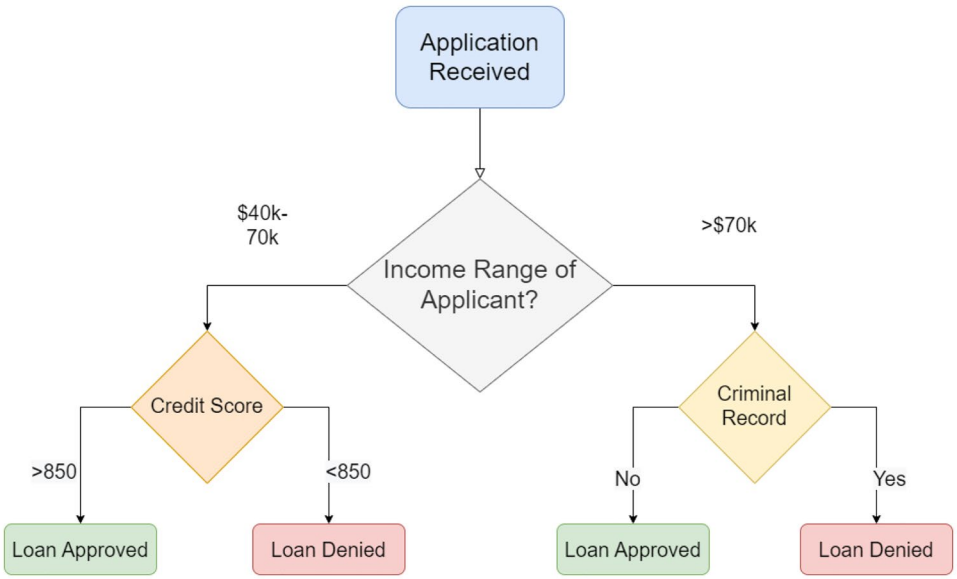


Fig. 7 An algorithm designed to predict loan approval applicants visualised through decision tree

Logistic regression models are a modification of linear regression to serve as a solution for classification problems with two possible outcomes.

Leaf node, explanations are given. Figure 7 shows one such decision tree [123].

Rule-based learners: This class of model makes rules to characterise data from its input data. These rules can either be if-else rules or a set of more complex combinations. Fuzzy-rule-based systems (FRBS) are models which are rule-based learners and are based on fuzzy sets. These models tackle real-world problems involving uncertainty and imprecision. An explanation of studies using FRBS can be found in [124].

K-NNs: This method is non-parametric and classifies instances simply and methodically. In XAI, KNN can be considered an interpretable algorithm as it provides a clear explanation for its predictions. The explanation is based on the principle that the algorithm classifies a new data point based on the class labels of its nearest neighbos in the training dataset [125]. This makes it relatively easy for a human to understand and verify the reasoning behind the predictions. However, KNN may not always provide the most accurate predictions compared to other, more complex algorithms.

Letham et al. [126] proposed a model called Bayesian Rule Lists (BRL), which is based on decision trees as mentioned in [127]. It was able to pick out certain data patterns that may be used as criteria to produce decision lists. Initial models produced encouraging results and had room for refinement to build trust at the domain level.

$$\text{logistic}(n) = \frac{1}{1 + \exp(-n)} \tag{3}$$

Trees are not found to be efficient in the case of linear relationships among input features. They are also vulnerable. Using the above logistic function [107], the regression model fits the output of the linear model between 0 and 1, thus making the linear influence of weights negligible. One of the major disadvantages that the logistic regression model poses is its comparatively more challenging interpretation because of the multiplicative nature of the weights considered.

Decision trees are used to overcome the problem of non-linearity and correlation in features for which linear regression models fail to generate interpretations. Decision trees function by continuously segmenting the input data into nodes of a tree, each of which represents a subset that belongs to a particular instance from the dataset [122]. By moving through a specific section of the tree from the root node to even the slightest of changes in the training dataset, thus providing unstable results.

Due to their ease in providing explanations, ante hoc interpretable methods have always been superior to other "black-box" methods. However, accuracy suffers as a result of this intrinsic interpretability [128]. This unfeasible tradeoff between accuracy and interpretability suggested post hoc procedures to be taken into account in a deliberate manner [129].

Post-Hoc Interpretability

Contradictory to ante hoc methods, post hoc interpretability refers to the class of techniques which involve the research and development of black-box models post their training. One interesting feature to note about post hoc methods is their diversified applications in the field of XAI, which also extends to applications in intrinsically interpretable models. The permutation feature, a post hoc interpretation method, is utilised for the computation of decision trees.

Model-Specific Methods Though helpful, model-specific methods of explainability offer a very finite range of interpretations for predictions provided by opaque AI algorithms. Thus, the availability of limited choices hinders their acceptance into the mainstream research of XAI methods. Regardless, a silver lining can be found in their specificity, which is leveraged in the case of a dominant model representation and prediction. To counter this incapability, researchers came up with model-agnostic methods of interpretability, which are model-independent and provide competitive results. We shall discuss these methods in detail in the following section.

Model-Agnostic Methods Model-agnostic methods of interpretability are applicable to different types of ANN and black-box models. Their universal nature is achieved by simultaneous analysis of the feature's input and output. But their structural definition restricts them from gaining model insights such as weights and crucial parameters. Collaborative work from researchers around the globe has witnessed a surge in the development of model-agnostic methods to cover a broader aspect of XAI. Upon review of the literature, we propose a classification of this class into the following methods (summary of method and its advantages and disadvantage in Table 3):

Visualisation Visualisation of a black-box model helps us dive deep into the hidden patterns and internal reasonings of the algorithm, which naturally enhances the understanding related to its predictions. The flexible representation of the technique helps it to stand out among other methods of explanation. One of the most popular domains where visualisation finds its application is supervised learning algorithms. Some of the popularly implemented visualisation techniques are:

Table 3 Comparing model agnostic methods

Method	Advantages	Disadvantages
PDP	1. Intuitive 2. Easy to implement 3. Interpretation is clear and causal	1. Valid for maximum three features 2. Assumes absence of correlation between features 3. Heterogeneous effects may be hidden
ICE	1. Intuitive 2. Specific 3. Reveal heterogeneous relationships	1. Overcrowding leads to unreadability 2. Need PDP to see the average
ALE	1. Unbiased towards correlated features 2. Faster Computation	1. Unsteady with high number of intervals 2. Comparatively much more complex 3. Not accompanied by ICE plots
LIME	1. Inherently interpretable 2. Widely acceptable especially in DNN's 3. Human-Freindly explanations	1. Unsatisfactory global approximation 2. Easily manipulated to hide biases

Partial dependence plot (PDP) uses graph-based explanations for visualising the relationship between one or more features (at maximum three features) and the prediction generated by the black-box model [130]. Being global in nature, it not only provides a comprehensive grasp of the model interpretations but also differentiates target-feature relations into linearity, monotonicity, or complexity accordingly. To understand the associative effects of predictors on the conditional average treatment outcome derived from a voter mobilisation experiment, Green and Kern [131] Elith et al. [132] made major contributions outside of the research area by replacing stochastic gradient boosting with PDP to comprehend how different environmental factors affect the distribution of freshwater. Averaging the marginal effects hides their interactions with the data, according to some academics, who argue that this is bad for the black-box model. Therefore, a new interpretation technique had to be developed to address this issue.

Individual conditional expectation (ICE) plots are an extended version of partial dependence plots which reveal the heterogeneous effects hidden by PDPs. They can be considered the local equivalent model of PDP since they work by visualising a feature's influence over the specified instance of the prediction under scrutiny. In other words, a PDP is an aggregation of all the resultant lines generated by an ICE plot. Due to the above-mentioned stress on individuality, ICE curves are naturally more comprehensive than PDP plots. Goldstein et al. supported the same [133]. ICE curves also undergo some severe problems. The presence of numerous instances at once creates overcrowding leaving the plot unclear. Casalicchio et al. [134]. were able to break the dilemma between PDP and ICE by proposing an approach that could use both of the visualisation tools as an aid to make black-box models transparent. An example of such tools is the "lossless visualisation methods," which are techniques used to represent high-dimensional data generated by ML models in a way that preserves all of the information in the data. These methods aim to provide clear, easy-to-interpret visualisations that avoid the issue of "quasi-explanations," which appear to be meaningful but are based on misleading or irrelevant features. Examples of lossless visualisation methods include dimensionality reduction techniques [135, 136] (such as PCA or t-SNE), scatter plots, and heat maps. These methods play an important role in ensuring that the information in the data is not lost or distorted and that the resulting explanations are meaningful and accurate.

Accumulated Local Effects (ALE) Introduced by Apley and Zhu [137], ALE plots are always the forerunner when it comes to visualising features that are correlated. While partial dependence plots result in a significantly biased computed feature effect, ALEs provide a faster and unbiased substitute for visualisation [138, 139]. Now, both PDP and ALE share the common characteristic of reducing the complex prediction function to a simpler one that deals with one or two features at a time. However, they mainly differ in the fact that whether they are utilising averages of predictions or differences in predictions. ALE goes with the latter one and aggregates them over the grid. Despite their faster computation, ALE plots are found to be complex in nature.

None of the above-discussed visualisation techniques was successfully able to interpret models with strong feature correlation. But, with work being developed at a much faster pace, ALEs are as good as it gets.

Example-Based Explanations Example-based interpretability represents a class of methods that elaborate the black-box model's predictions using specific instances from the model's training dataset. In layman's terms, their approach can be simplified into "Since A is equivalent to B and B generated C, A will also generate C." Example-based methods act on a specific instance of the ML model while model-agnostic methods interpret the model by engaging with its features or performing alterations to it.

From a research point of view, some of the widely accepted example-based interpretability techniques are the following:

Prototypes and Criticisms Prototypes are a group of chosen examples that accurately depict all the data [140–142]. The selection of data instances that are poorly represented by the corresponding prototypes that must be made in order for the model's output forecast to alter. These justifications are predicated on the idea that "If a certain thing X would have been done differently, its effect Y would likewise have been different." Prototypes are representative examples taken directly from the data, whereas counterfactuals can be created by combining new input instances. But despite how human-friendly they may seem, counterfactuals nevertheless have certain drawbacks. The "Rashomon Effect" is a phenomenon in which users overwhelmed with too many options tend to choose inadequate counterfactual explanations resulting in subpar performance. Thus, it inhibits users from fully utilising the potential of counterfactuals. Hasan et al. [146] realised the need to mitigate this effect and proposed a game-theory perspective to make counterfactuals more intriguing. They were successful in narrowing down the counterfactual possibilities by optimising it to a more informed process.

Since counterfactual explanations do not need access to the model or dataset itself, their implementation for extracting explanations seems to be in the best interests of the owner. Thus, they have recently been receiving growing attention from companies offering explanations without interfering with their model and data.

The selection of data instances that are poorly represented by the corresponding prototypes is criticised. They can be used to characterise the data (independently) or interpret a black-box model depending on their relationship (dependent). According to reviewed literature, there are many methods for locating prototypes in data (such as the k-medoids algorithm), but very few for locating criticisms. We use the example of Google's image classifier mistaking black individuals for gorillas to demonstrate how important the existence of criticisms is. It was discovered that including black people's photographs as complaints would have increased the dataset's diversity (Fig. 8) [143]. Kim et al. [144] provided the MMD-critic, a novel framework that provides the optimum number of prototypes and criticisms needed to describe the dataset as a whole.

Counterfactuals Counterfactual explanations were suggested by Wachter et al. [145]. as a way to explain predictions and choices made by enigmatic AI algorithms. Counterfactual explanations outline the minimal adjustments to the input feature values.

Adversarial Adversarial examples are basically instances accompanied by a minor perturbation with the intention to deceive an ML model [85, 147]. For instance, attackers were able to successfully fool a facial recognition AI system by introducing various patterns on glasses or hats. With the help of adversarial examples, one can gain insights into the internal structure of algorithms, find their vulnerabilities, and improve interpretability. Upon investigation, we find some of the techniques for designing adversarial examples.

In their study, Szegedy et al. [147]. suggested a gradient-based method that may be utilised to provide adversarial instances for DNNs. The rapid gradient sign method for implementing adversarial pictures was created by Goodfellow et al., according to [81]. Su et al. [148] demonstrated how

Fig. 8 Users found viewing prototypes and criticisms led to results being more accurate and efficient in comparison to viewing a random subset of data

image classifiers may be tricked by altering just one input picture pixel. Athalye et al. [149] successfully fooled a DNN by creating a 3D-printed turtle that the DNN mistakenly thought was a weapon.

But how do these deceiving tools help in studying the interpretability of AI models? The answer lies in a recent effort to produce "robust" models which resist such adversarial perturbations and also offer higher-quality explanations [83, 150]. Ilyas et al. [151] classified features based on their robustness i.e., the ability to counter adversarial examples, and proved that robust models held a more plausible explanation than non-robust models.

Influential Instances ML-based methods generate their predictions based on the learning of their training data. Even a small modification in a training instance could alter the resultant model significantly. An "influential" training instance is one which considerably influences the parameter determination and decisions of the model. These influential instances are crucial in debugging and examining the behavior of the model. One of the approaches to finding influential instances is called deletion diagnostics [152]. In this, we simply delete the concerned instance and analyse the difference in the predictions of the model with and without that instance.

Methods For Visualising and Interpretation Data visualisation can be defined as complex algorithms which use data to create images so that humans can understand and respond more effectively. AI development is a hunt for algorithms to be better and more responsive than humans. AI learning technique is based upon writing a model, but instead of a human using your model, the system takes some data as

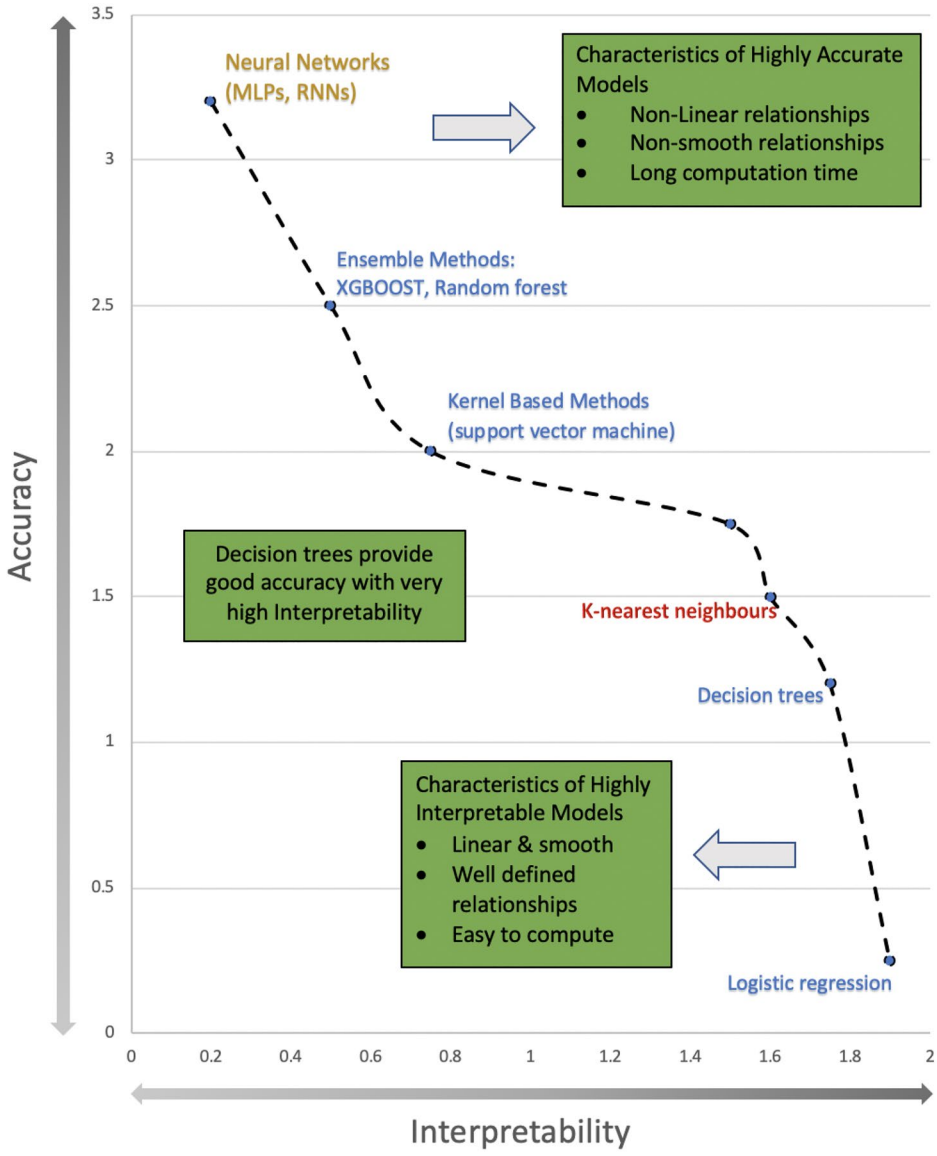


Fig. 9 Graphical representation of accuracy vs interpretability for widely used AI algorithms

input and creates a new model (Fig. 9). Exploratory data analysis with the help of visualisation tools such as Tensorflow library or Microsoft Azure ML studio is done to make things sensible [107]. Using advanced data processing software, developers can integrate multiple coordinated views to seamlessly explore large-scale deep learning models and predict results, as well as discover patterns [127, 153, 154]. Here are some new methods and approaches that AI software can use for data visualisation:

- Interactive visualisations: Using dynamic, interactive visualisations to allow users to explore data and model behavior in real-time.
- Augmented Reality (AR) and Virtual Reality (VR) visualisations: Using AR and VR technology to create immersive visualisations that can help users better understand and interact with data.
- 3D and 4D visualisations: Using 3D and 4D visualisations to represent data in new and more informative ways, allowing users to better understand complex relationships and patterns.
- Automated visualisation generation: Using machine learning algorithms to automatically generate visualisations based on the data and desired output.
- Real-time streaming visualisations: Creating visualisations that can be updated in real-time as new data is received, allowing users to monitor the data and model performance in real-time.
- Multi-view visualisations: Using multiple views or perspectives to represent data, allowing users to explore and understand the data from different angles.

With increasing efforts to feed quality data to AI models, developers are getting better at visualising data rather than just what is hidden in numbers; this concept holds immense significance in the AI world. To understand trends and spot anomalies, presented data can help AI visualise and understand the context of the problem. A few examples include DGMTracker and GANViz, which focus on helping a developer understand training dynamics and train these complex models. Research in this area includes the creation of various tools and frameworks for democratisation and interoperability, but new work is immediately open-sourced without being published at a reputed conference. There may be many languages to perform data visualisation, but the much widely used are *python* and *R*. Which one would be best and has better scope? The answer is simple; it is purely the user's choice. In this paper, our main goal is to understand the concepts, such as which conditions are optimal for visualising concepts, and deep dive into coding/implementation to help sample plots [155].

Data interpretation is one of the most arising innovations that joins techniques and fields from various fields of study.

Fig. 10 Input–output model of recommendation

Enormous data are present, and a huge effort is required to organise all the data systematically. Humans simply cannot read all data and organise them. The prominence of AI is real and cannot be ignored. In today's time, such good PCs are available that even our grandmasters get defeated by them. Data analytics alongside AI are less labor-intensive and highly efficient. An AI automated bot is useful software to interact with millions of users daily by gathering knowledge through ML. It stimulates human interaction and reduces workload. It interprets data and breaks down each sentence into individual words, and each word is used as data for ANN [156] (Fig. 10).

Another emerging topic that can help to interpret data is big data. It has been evolving in the field of AI in recent years and has been tackling almost all human challenges. Big data refers to a large number of instances (and maybe features) that traditional data processing software is unable to capture and process. With big data analysis, real-time problems such as fraud detection, financial risk analysis and price optimisation can be done in fractions of seconds [157]. Have you ever noticed that fries ordered at McDonald's or Carl's Jr. are always on time or sometimes a bit early too? Well, the answer is simple, it is big data that monitors the number of customers, and if the waiting line is too long, it will reflect only those items which can be quickly prepared. At Universal Studios, they give us bands that are installed with RFID tags. Thousands of sensors are installed across the park, which gathers information about activities. Thus, big data helps us to enhance customer experience. According to a survey, big data has some demerits due to the lack of proper maintenance. At all times, there is a high risk of data piracy and leak, and businesses are at potential risk of cyber-attacks [158]. Also, a major issue rising these days is that there is a lack of awareness about big data [159]. People who possess skills and can work on big data are few people who want to work on it.

Evaluating the Quality of Explanations and Error Mitigation In addition to providing the explanation for an AI decision, incorporating an evaluative approach under the XAI umbrella should be considered while the field is still in its early stages. By doing this, XAI will be able to cater to a wider spectrum of stakeholders. However, the subjective nature of explainability poses a strenuous challenge for researchers.

A rudimentary principle for benchmarking the quality of explanations would be if the evaluation is demonstrating the claimed contribution. Subsequently, other factors such as

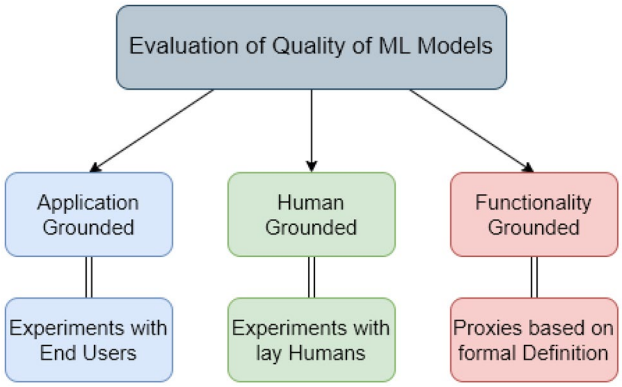


Fig. 11 Taxonomy of explanation quality

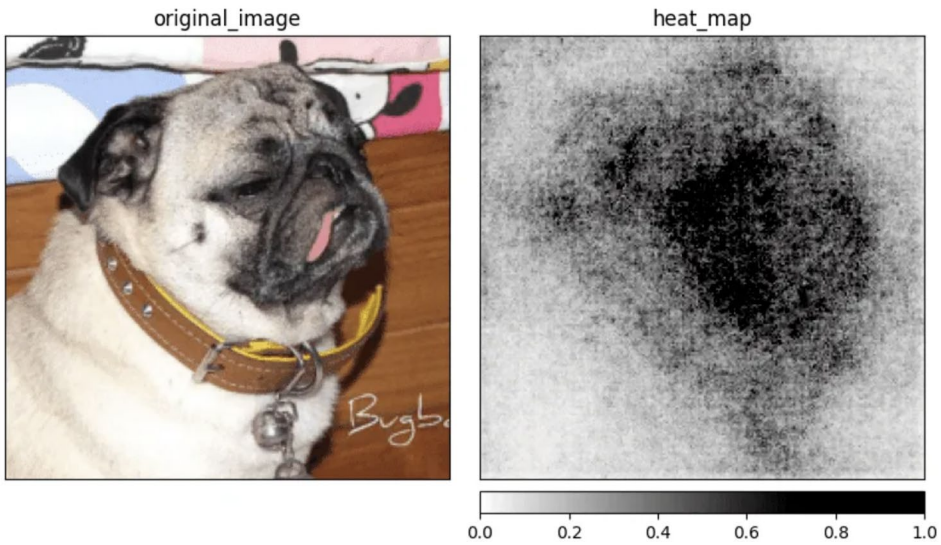
optimal use of resources, minimal time consumption, and degree of explanation may embrace the benchmark process. Doshi-Velez and Kim [36] provided a paradigm which categorised evaluation techniques into (also shown in Fig. 11):

- Application grounded—This methodology deals with domain experts experimenting in real-life situations to validate the delivery of end-task. More specifically, it acknowledges any discovery of new facts, debugging of new errors and elimination of biases. For example, to diagnose a disease, the most suitable way is for a doctor to perform diagnosis [160].
- Human grounded—This methodology deals with lay humans performing general experiments to address a wider pool of evaluation concepts. This proves to be a cost-effective method of maintaining the crux of the target application. For example, humans are given a choice between two theories, and they need to choose the one with higher accuracy [161].
- Functionally Grounded—This methodology deals with a proxy measure for evaluating interpretability, and therefore additional research is needed. No requirement for human interaction and minimal cost are some appealing factors for its wide usage. For example, decision trees are considered interpretable in many situations [162].

Holzinger et al. [163] devised a notion of causability, which they later combined with a widely-accepted usability scale to form the System Causability Scale (SCS). SCS measures the extent to which a certain explanation behind the AI model decision attains a causal understanding with the user. As a demonstration, SCS was utilised with the Framingham Risk Tool (FRT) by a medical doctor from the Ottawa Hospital [164].

Error mitigation via DL is referred to as the simple technique used to reduce specific errors in quantum computing algorithms. Many XAI methods produce saliency maps. These maps highlight and increase the pixel intensity of a particular image that has similar salient properties. Despite its success, its black-box nature is a serious hurdle in pivotal fields like medicine and autonomous driving. These maps fail if they are subjected to data poisoning or if the model hasn't been trained sufficiently. Saliency maps assume that all features in the model are interpretable, but in some cases, the model may be making decisions based on features that are incoherent to humans. Figure 12 shows how when a dog's image is subjected to a Grad-CAM heat map produces a lot of noise. To improve the accuracy of saliency maps, Ismail et al. [165] take a different approach by proposing a new training procedure called saliency-guided training. This

Fig. 12 Saliency map of a dog clearly showing noise



procedure presents outputs that are less noisy and clear and do not degrade the model's performance.

In machine learning pipelines, the black box is not an option anymore, and there is a tradeoff between performance and interpretability. It is argued that many complex models are full of errors. To prove this, Aman et al. [166] gave a classic example by comparing laboratory testing and CDSS. With the availability of more datasets, there is an increased benefit that allows more complex functions to be approximated for future developments in XAI [21, 167].

In order for AI models to provide quality data in return, candidates need to train their models hard enough to procure quality results. Error analysis helps to investigate and diagnose error patterns. In conclusion, every model has its own unique set of errors and problems. If we follow a formal approach every time, we can avoid reinventing the wheel every time.

Depending on the specific application and context, the efficacy of XAI's current models and strategies can vary. Some models and strategies have been demonstrated to be highly effective and efficient at providing interpretable explanations for AI systems, whereas others may still be in the early stages of development or encounter scaling or generalisability challenges.

Challenges For Enabling XAI

Human-Machine Collaboration

Machine-Human collaboration is a system where humans collude with AI and other machines. Both have a symbiotic relationship with each other, where the human uses the machine's intelligence and its superpower to compute millions of threads while the machine uses the human's ability to interact with employees and customers to accumulate data. This type of collaboration allows humans to understand the decisions made by AI systems and to provide feedback and guidance to improve the performance of these systems over time. Building a successful human-AI collaboration will allow big IT companies to identify new strategies to overcome challenges humans face to foster a positive relationship between machines and humans in the workforce. According to Xiong et al. [168], machines will be helpful in gathering data and conveying key messages, while humans can obligate emotional influences and make an unbiased judgment. Damacharla et al. [169]. found that a combination of two non-expert chess players and three personal computers was more successful than either a group of supercomputers or a group of grandmasters on their own.

A recent survey by Deloitte tells us that this technology is being extensively used in their organisation. The robotics section uses 22% AI automation, while the company's patented cognitive technology uses around 43% [170, 171]. Such intelligent automation hands off the workload and phases the era from manual execution to the tactical automation age era. This enables organisations to identify algorithmic analyses of the data to make predictions. Accenture's Paul Daugherty, Chief Technology and Innovation Officer, and H. James Wilson, Managing Director of Information Technology and Business Research, claim that such collaborations could increase revenue by 38% by 2022. More than 68% of businessmen agreed that this intersection would help them achieve strategic priorities faster [172]. Figure 14 shows the benefit of AI in the workplace, according to the survey. They conclude uman collaboration based on four terms:

- Improved decision quality: By leveraging the strengths of both humans and machines, human–machine collaboration can result in improved decision quality compared to relying on either alone. This idea emerges because of the compulsive nature to improve the reliability of the system [173].
- Increased transparency: Human–machine collaboration can increase transparency and accountability in AI systems, as humans can understand the reasoning behind the decisions made by these systems. This will allow shared awareness and intent for optimal teamwork [174].
- Mitigation of bias: By incorporating human feedback and oversight, human–machine collaboration can help to mitigate bias and ensure that AI systems are fair and ethical. Using an end-to-end machine learning pipeline which includes pre-processing, in-processing, and post-processing, data scientists can detect and eliminate any form of bias in their models.
- Better performance over time**:** Human–machine collaboration allows AI systems to continuously improve over time, as human experts can provide feedback and guidance to improve the performance of these systems.

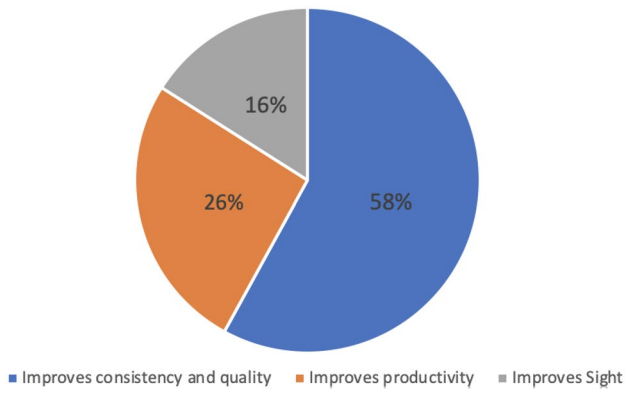


Fig. 13 Benefit of AI in the workplace

Of course, with time, there will be some human roles that will be eliminated due to technological evolution. For now, new talent who are proficient at developing AI models which are explainable needs to be recruited [175]. In the past 22 years, the people of the United States have lost five million jobs due to this revolution, but they have been re-recruited at job openings in the manufacturing sector. In the continuous circle of learning between humans and machines, there is a mutual transfer of knowledge [176]. The outcome of the decision is susceptible to unpredictability. This refers to some risky decisions that are made using AI [177] (Fig. 13). Such decisions are made without knowing the exact consequence. These kinds of problems are omnipresent in technology, daily life, and several choices we make each day. With machines, emotional influences can be restrained, and empirical analysis can demonstrate the benefits of human-machine collaboration. These AI models are created with limited data and within a pre-defined scope. Although these machines can surpass humans in terms of data processing [178], they lack accountability and interpretability and hence are referred to as "black boxes."

Acceptance and Popularisation

Studies have shown that when people see AI technology as being simple to use, they are more likely to adopt and trust it. Therefore, it is necessary to make AI clear and understandable in order to improve its adoption in people's lives. Consumer technologies have undergone extensive research and are constructed using a number of predetermined principles, including justice, accountability, and transparency [179]. Future research will look into whether there is a significant empirical requirement to understand what supports AI if the role of central trust is assumed.

Collaborating with industry, government, and academic stakeholders can help to promote XAI and educate the public on its potential benefits and limitations. For example, IBM [180] is working to develop technology that will enable the machine also to be able to explain to you what it is advising, which is likely to happen within the next five years, according to Rachel Bellamy, IBM research manager for human-agent collaboration. A technology's level of human acceptance can be predicted using the well-known Technology Acceptance Model (TAM), which was created by Mr. Fred David [181] in the late 1980s. Perceived usability and perceived ease of use are the two criteria that affect this model.

The rising usage of AI, on the other hand, has raised ethical concerns and opposing arguments that directly challenge traditional approaches in XAI. According to a woman in her 60 s, she demonstrates how the excessive use of AI is displacing workers. She worries that soon humans will not be able to make decisions. "Too Much, Too Little, or Just Right?," a work proposed by Kulesza et al. [182] presented their findings to reveal how explanations are impacting end users' mental models. Another encouraging area of research in literature is the economic perspective of XAI, which examines the cost requirements of integrating XAI into the mainstream.

Akyol et al. [183]. initially attempted a quantitative analysis of the cost of transparency (PoT) in ML algorithms. The work of Igami [184] about the connections between machine learning and econometrics laid the foundation for "Structural Econometrics for XAI."

It is crucial to realise that theory says very little about technology but a lot about our beliefs and how we interpret it. Similar to how people adapt to new revolutionising technologies, people will gradually come to embrace XAI in their lives too.

Discussion and Reflection

Future Research Directions

The wide range of reviewed methodologies in this survey demonstrates how quickly XAI has advanced in the creation and application of open and accountable AI systems. Due to the infancy of this field, there are a number of areas where sophisticated ML and AI algorithms do not appear to benefit from conventional approaches. Therefore, it is advised to search for non-conventional sources of explanations. We suggest some potential future directions that interested researchers could investigate in addition to the current initiatives, keeping in mind the long-term objective of XAI.

Responsible AI

The limitations of AI gave origin to the field of XAI. However, some researchers stressed the fact that if somehow the discussed limitations of AI could be eradicated, it could save precious resources as well as produce more efficient AI models. We move forward with our discussion from the XAI realm towards Responsible AI, a paradigm that is based on the principle of protecting societal values and interests of the stakeholders. With considerable weightage given to the ethical implications of the decisions made by such AI algorithms, not only does it guarantee transparency but also inherited responsibility makes the system truly intelligent.

To direct the development of Responsible AI, the following factors should be kept in mind:

- Acceptance of responsibility will determine public attitude towards the acceptance of responsible AI in society. Governments and citizens need to act together to resolve issues of reliability concerned with AI.

- Self-justification will enable AI models to develop reasoning and a code of conduct based on human values and ethics. Current research shows that an adequate link between decisions and ethical context is missing.
- Participation involves the real-life application of AI in everyday life to develop the guidelines for responsible AI. Here, education plays a significant role in creating awareness among people that their input is crucial in shaping the societal character of responsible AI.

To ensure that the design and development of responsible AI reflect the ethical concerns of humans, we propose to incorporate the above-discussed factors with the principles proposed by Dignum [185], i.e., accountability, responsibility, and transparency (ART) as depicted in Fig. 14.

Presently, responsible AI does not hold any concrete approach to its design, due to which its theory seems to cease to exist. Also, responsible AI suffers from moral deliberation, which brings unnecessarily increased computation cost and complexity.

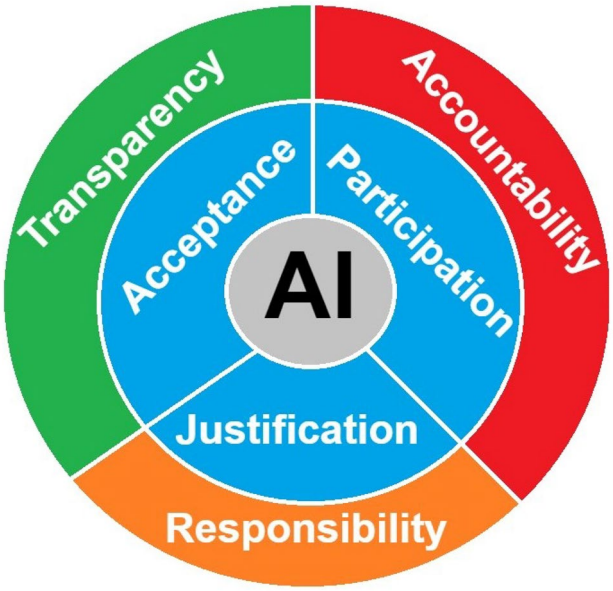


Fig. 14 Benefit of AI in the workplace

Universal Framework and Formalisation

We could see the gap between interdisciplinary research on how people explain and formalise patterns in algorithmic forms and XAI techniques and frameworks, which are constantly evolving. A novel and unified framework should be aimed at the formalism of XAI, given the rise in research proposals in the field. This framework will be supported by the two pillars of XAI: "explanation" and "interpretation," acting as an agent for addressing the growing heterogeneity and consolidating developed methods.

Potential Implications

Due to the wide-ranging approaches of XAI, it will be nearly impossible and highly tiring to survey all research papers about XAI. The papers surveyed are selectively chosen based on their content and impact on XAI. More preference is given to fresh work to give an overview to interested researchers about recent trends. Until now, four main areas of focus have been identified: A way to explain complex black-box models, performance and analysis of neural networks, improving the popularity of white box models among developers, and methods to eliminate discrimination and improve fairness [186]. Exploring new concepts increases the meta information of a person and helps to evaluate individual class predictions by local algorithms [187]. By decomposing and breaking each explanation, the importance of each concept increases. With the ups and downs of life, there will be both positive and negative impacts of XAI on a person's daily life. There are multiple impacts of this technology. Every organisation needs to apply it to their website, working algorithms, games, etc. The future revolves around ethics, new technologies, and governance. Top-tier organisations are continuously developing AI-rooted.technology in technological rigor. The XAI market is growing ceaselessly at a high rate. Until now, there are very few companies that produce only XAI models. While there is a crunch of startups in the market, monitoring this growing field helps to stay on top of new mechanisations. Prioritizing today's leaders will impact tomorrow's future. XAI will enable federal leaders to make calculated investments to automate workflow.

Other Research Directions (Future GPT Models)

XAI, or explainable artificial intelligence, is gaining importance for GPTs (Generative Pretrained Transformers) as these models become more sophisticated and capable. GPTs are notorious for their lack of interpretability and transparency, despite achieving remarkable results in several applications. This makes it difficult to comprehend how they arrive at their predictions, making it challenging to identify and rectify errors, biases, and other problems. By providing clear and understandable explanations for GPTs, XAI can assist in overcoming these obstacles. This can assist users in comprehending the model's decision-making process, identifying potential biases or errors, and building confidence and accountability in the system.

Several XAI techniques, such as saliency maps, feature centrality scores, and counterfactual explanations, can be

applied to GPTs. These techniques can assist users in comprehending which aspects of the input data the GPT is focusing on, which features are most crucial for its predictions, and how modifying the input data would affect the output. As the complexity and strength of GPTs continue to increase, XAI is becoming increasingly vital to their success. By providing transparent and interpretable explanations for these models, XAI can aid in ensuring that they are used ethically and responsibly and that their outputs are accurate and reliable.

There are initiatives to develop tools and techniques that can help explain why a model makes particular predictions or generates particular outputs. For instance, techniques such as attention visualisation can help determine which portions of the input text are most crucial to the output of the model.

Transformer models that have been pre-trained are frequently fine-tuned for specific tasks such as query answering and sentiment analysis. During fine-tuning, the model is trained on task-specific data, which helps to identify any biases or errors in the output of the previously trained model.

Moreover, GPTs can be employed to explain and interpret black-box AI models. Training GPT on a large corpus of explanations for various types of models and problems is one method to use to explain black-box AI models. This can include explanations of the models' fundamental algorithms and techniques, as well as the models' inputs and outputs and how they are used in decision-making.

Once trained, the GPT model can generate explanations in natural language for specific black-box models based on their inputs and outputs [22]. These explanations can cast light on the inner workings of the black-box model and provide insight into how it arrived at its conclusions or predictions.

Notably, the quality and accuracy of the explanations produced by GPT may depend on the quality and accuracy of the training data, as well as the complexity and character of the black-box model being explained [188, 189]. Hence, it would be advisable to combine GPT-generated explanations with other techniques, such as model-agnostic methods and model-specific interpretability techniques, when attempting to explain black-box AI models.

Conclusion

Ethics issues and the requirement for control of the infamous AI black box have drawn a lot of attention in the last 10 years. XAI was introduced as a multidisciplinary field to make the "black box" transparent. By elaborating on a conceptual understanding of XAI taxonomy, diverse XAI applications, methodologies for explainability, as well as the limitations and challenges faced by XAI, in this survey, we aimed to create a unified framework that can navigate through this literature.

As far as we know, XAI's constantly expanding field still requires a more formal approach to address the diverse range of user questions that can be influenced by various factors such as motivation, context, and individuality. Our research highlights the shortcomings of current XAI algorithms and proposes opportunities for the human–computer interaction community. Despite the significant progress that has been made in XAI, it still faces challenges such as balancing transparency and privacy, addressing the diversity of user needs, and creating effective explanations for complex models. Nonetheless, XAI has the potential to bring AI closer to the human domain, making it more accessible and trustworthy.

Funding This work was partially supported by the CHIST-ERA grant CHIST-ERA-19-XAI-009. M. Mahmud was supported by supported by the AI-TOP (2020-1-UK01-KA201-079167) and DIVERSASIA (618615-EPP-1-2020-1-UKEPPKA2-CBHEJP) projects funded by the European Commission under the Erasmus+ programme.

Data Availability Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

Declarations

Ethical Approval This article does not contain any studies with human participants or animals performed by any of the authors.

Conflict of Interest The authors declare no competing interests.

References

1. Anbar M, Abdullah R, Al-Tamimi BN, Hussain A. A machine learning approach to detect router advertisement flooding attacks in next-generation ipv6 networks. Cogn Comput. 2018;10:201–14.
2. Osaba E, Del Ser J, Martinez AD, Hussain A. Evolutionary multitask optimization: A methodological overview, challenges, and future research directions. Cogn Comput. 2022;14(3):927–54.
3. Li XH, Cao CC, Shi Y, Bai W, Gao H, Qiu L, Wang C, Gao Y, Zhang S, Xue X, Chen L. A survey of data-driven and knowledge-aware explainable ai. IEEE Trans Knowl Data Eng. 2022;34(1):29–49.
4. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg AC, Fei-Fei L. Imagenet large scale visual recognition challenge. Int J Comput Vision. 2015;115(3):211–52.

5. Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL. Microsoft coco: Common objects in context. 2014. p. 740–55.
6. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: Pereira F, Burges CJC, Bottou L, Weinberger KQ, editor. Advances in Neural Information Processing Systems 25. Curran Associates, Inc.; 2012. p. 1097–105. http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf.
7. Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. arXiv. 2014.
8. Antol S, Agrawal A, Lu J, Mitchell M, Batra D, Zitnick CL, Parikh D. VQA: Visual question answering. CoRR. 2015;abs/1505.00468. http://arxiv.org/abs/1505.00468.
9. Silver D, Huang A, Maddison CJ, Guez A, Sifre L, van den Driessche G, Schrittwieser J, Antonoglou I, Panneershelvam V, Lanctot M, Dieleman S, Grewe D, Nham J, Kalchbrenner N, Sutskever I, Lillicrap T, Leach M, Kavukcuoglu K, Graepel T, Hassabis D. Mastering the game of Go with deep neural networks and tree search. Nature. 2016;529(7587):484–9.
10. Sharma P, Jain S, Gupta S, Chamola V. Role of machine learning and deep learning in securing 5g-driven industrial iot applications. Ad Hoc Netw. 2021;123:102685.
11. Brown N, Sandholm T. Superhuman ai for multiplayer poker. Science. 2019;365:eaay2400.
12. Berner C, Brockman G, Chan B, Cheung V, Debiak P, Dennison C, Farhi D, Fischer Q, Hashme S, Hesse C, Józefowicz R, Gray S, Olsson C, Pachocki J, Petrov M, de Oliveira Pinto HP, Raiman J, Salimans T, Schlatter J, Schneider J, Sidor S, Sutskever I, Tang J, Wolski F, Zhang S. Dota 2 with large scale deep reinforcement learning. CoRR. 2019;abs/1912.06680. http://arxiv.org/abs/1912.06680.
13. Todorov G. 65 artificial intelligence statistics for 2021 and beyond. 2021. https://www.semrush.com/blog/artificial-intelligence-stats/.
14. Roy A, Banerjee B, Hussain A, Poria S. Discriminative dictionary design for action classification in still images and videos. Cogn Comput. 2021;13:698–708.
15. Bansal G, Chamola V, Narang P, Kumar S, Raman S. Deep3dscan: Deep residual network and morphological descriptor based framework forlung cancer classification and 3d segmentation. IET Image Proc. 2020;14(7):1240–7.
16. Li B, Xu Z, Hong N, Hussain A. A bibliometric study and science mapping research of intelligent decision. Cogn Comput. 2022;14(3):989–1008.
17. Mahmud M, Kaiser MS, McGinnity TM, Hussain A. Deep learning in mining biological data. Cogn Comput. 2021;13:1–33.
18. Hassija V, Chamola V, Bajpai BC, Zeadally S, et al. Security issues in implantable medical devices: Fact or fiction? Sustain Cities Soc. 2021;66: 102552.
19. Rohmetra H, Raghunath N, Narang P, Chamola V, Guizani M, Lakkaniga NR. Ai-enabled remote monitoring of vital signs for covid-19: Methods, prospects and challenges. Computing. 2021;1–27.
20. Alladi T, Kohli V, Chamola V, Yu FR, Guizani M. Artificial intelligence (ai)-empowered intrusion detection architecture for the internet of vehicles. IEEE Wirel Commun. 2021;28(3):144–9.
21. Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, Garcia S, Gil-Lopez S, Molina D, Benjamins R, Chatila R, Herrera F. Explainable artificial intelligence (xai): concepts, taxonomies, opportunities and challenges toward responsible ai. Inf Fusion. 2020;58:82–115.
22. Das A, Rad P. Opportunities and challenges in explainable artificial intelligence (XAI): A survey. CoRR. 2020;abs/2006.11371. https://arxiv.org/abs/2006.11371.
23. Khaleghi B. An Explanation of What, Why, and How of eXplainable AI (XAI). 2020. https://towardsdatascience.com/an-explanation-of-what-why-and-how-of-explainable-ai-xai-117d9c441265.
24. Anand T, Sinha S, Mandal M, Chamola V, Yu FR. Agrisegnet: Deep aerial semantic segmentation framework for IoT-assisted precision agriculture. IEEE Sens J. 2021;21(16):17581–90.
25. Chhikara P, Tekchandani R, Kumar N, Chamola V, Guizani M. Dcnn-ga: A deep neural net architecture for navigation of uav in indoor environment. IEEE Internet Things J. 2020;8(6):4448–60.
26. Chamola V, Goyal A, Sharma P, Hassija V, Binh HTT, Saxena V. Artificial intelligence-assisted blockchain-based framework for smart and secure EMR management. Neural Comput Appl. 2022;1–11.
27. Shen Y, Ding N, Zheng HT, Li Y, Yang M. Modeling relation paths for knowledge graph completion. IEEE Trans Knowl Data Eng. 2021;33(11):3607–17.
28. Lu S, Liu M, Yin L, Yin Z, Liu X, Zheng W, Kong X. The multimodal fusion in visual question answering: A review of attention mechanisms. PeerJ Comput Sci. 2023;9.
29. Wazid M, Das AK, Chamola V, Park Y. Uniting cyber security and machine learning: Advantages, challenges and future research. ICT Express. 2022;8(3):313–21.
30. Hassija V, Batra S, Chamola V, Anand T, Goyal P, Goyal N, Guizani M. A blockchain and deep neural networks-based secure framework for enhanced crop protection. Ad Hoc Netw. 2021;119: 102537.
31. Garg P, Chakravarthy AS, Mandal M, Narang P, Chamola V, Guizani M. Isdnet: Ai-enabled instance segmentation of aerial scenes for smart cities. ACM Trans Internet Technol (TOIT). 2021;21(3):1–18.
32. Ahmed F, Sultana S, Reza MT, Joy SKS, Golam M. Interpretable movie review analysis using machine learning and transformer models leveraging xai. 2023.
33. Singh S, Sulthana R, Shewale T, Chamola V, Benslimane A, Sikdar B. Machine-learning-assisted security and privacy provisioning for edge computing: A survey. IEEE Internet Things J. 2021;9(1):236–60.
34. Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?": Explaining the predictions of any classifier. CoRR. 2016;abs/1602.04938. http://arxiv.org/abs/1602.04938.
35. Chekroud AM, Zotti RJ, Shehzad Z, Gueorguieva R, Johnson MK, Trivedi MH, Cannon TD. Cross-trial prediction of treatment outcome in depression: A machine learning approach. Lancet Psychiat. 2018;5(5):417–25.
36. Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. 2017. https://arxiv.org/abs/1702.08608.
37. Wang D, Yang Q, Abdul A, Lim B. Designing theory-driven user-centric explainable ai. 2019.
38. Lapuschkin S, Wäldchen S, Binder A, Montavon G, Samek W, Müller K. Unmasking clever hans predictors and assessing what machines really learn. CoRR. 2019;abs/1902.10178. Available: http://arxiv.org/abs/1902.10178.
39. Tjoa E, Guan C. A survey on explainable artificial intelligence (xai): Toward medical xai. IEEE Trans Neural Netw Learn Sys. 2021;32(11):4793–813.
40. Ghorbani A, Wexler J, Zou J, Kim B. Towards automatic concept-based explanations. 2019. https://arxiv.org/abs/1902.03129.
41. Selvaraju RR, Das A, Vedantam R, Cogswell M, Parikh D, Batra D. Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization. CoRR. 2016;abs/1610.02391. Available: http://arxiv.org/abs/1610.02391.
42. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. CoRR. 2015;abs/1512.04150. http://arxiv.org/abs/1512.04150.
43. Samek W, Binder A, Montavon G, Bach S, Müller K. Evaluating the visualization of what a deep neural network has learned. CoRR. 2015;abs/1509.06321. http://arxiv.org/abs/1509.06321.
44. Becker S, Ackermann M, Lapuschkin S, Müller K, Samek W. Interpreting and explaining deep neural networks for classification

of audio signals. CoRR. 2018;abs/1807.03418. http://arxiv.org/abs/1807.03418.
45. Arras L, Horn F, Montavon G, Müller KR, Samek W. "What is relevant in a text document?": An interpretable machine learning approach. PLoS ONE. 2017;12:E0181142.
46. Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. CoRR. 2013;abs/1311.2901. http://arxiv.org/abs/1311.2901.
47. Kim B, Wattenberg M, Gilmer J, Cai C, Wexler J, Viegas F, Sayres R. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). 2017. Available: https://arxiv.org/abs/1711.11279.
48. Raghu M, Gilmer J, Yosinski J, Sohl-Dickstein J. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. 2017. https://arxiv.org/abs/1706.05806.
49. Silva A, Schrum M, Hedlund-Botti E, Gopalan N, Gombolay M. Explainable artificial intelligence: Evaluating the objective and subjective impacts of xai on human-agent interaction. Int J Hum-Comput Interact. 2022;1–15.
50. Mohseni S, Zarei N, Ragan ED. A multidisciplinary survey and framework for design and evaluation of explainable ai systems. ACM Trans Interact Intell Syst. 2021;11(3–4). https://doi.org/10.1145/3387166.
51. Liu D, Cao Z, Jiang H, Zhou S, Xiao Z, Zeng F. Concurrent low-power listening: A new design paradigm for duty-cycling communication. ACM Trans Sen Netw. 2022;19(1).
52. Shen X, Jiang H, Liu D, Yang K, Deng F, Lui JCS, Luo J. Pupilrec: leveraging pupil morphology for recommending on smartphones. IEEE Internet Things J. 2022;9(17):15538–53.
53. Ren Y, Jiang H, Ji N, Yu H. Tbsm: A traffic burst-sensitive model for short-term prediction under special events. Knowl-Based Syst. 2022;240: 108120.
54. Ren Y, Jiang H, Feng X, Zhao Y, Liu R, Yu H. Acp-based modeling of the parallel vehicular crowd sensing system: Framework, components and an application example. IEEE Trans Intell Veh. 2022;8(2):1536–48.
55. Mittu R, Sofge D, Wagner A, Lawless W. Robust intelligence and trust in autonomous systems. 2016.
56. Petersen L, Tilbury DM, Yang XY, Robert LP. Effects of augmented situational awareness on driver trust in semi-autonomous vehicle operation. 2017.
57. Haspiel J, Du N, Meyerson J, Robert LP, Tilbury D, Yang XJ, Pradhan AK. Explanations and expectations: Trust building in automated vehicles. In: Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction, ser. HRI '18. New York, NY, USA: Association for Computing Machinery; 2018. p. 119–20. https://doi.org/10.1145/3173386.3177057.
58. Xie X, Huang L, Marson SM, Wei G. Emergency response process for sudden rainstorm and flooding: Scenario deduction and Bayesian network analysis using evidence theory and knowledge meta-theory. Nat Hazards. 2023;117(3):3307–29.
59. Chen P, Liu H, Xin R, Carval T, Zhao J, Xia Y, Zhao Z. Effectively detecting operational anomalies in large-scale IoT data infrastructures by using a gan-based predictive model. Comput J. 2022;65(11):2909–25.
60. Cresswell K, Callaghan M, Khan S, Sheikh Z, Mozaffar H, Sheikh A. Investigating the use of data-driven artificial intelligence in computerised decision support systems for health and social care: A systematic review. Health Inform J. 2020;26(3):2138–47.
61. Li B, Tan Y, Wu A, Duan G. A distributionally robust optimization based method for stochastic model predictive control. IEEE Trans Autom Control. 2021;67(11):5762–76.
62. Qu Z, Liu X, Zheng M. Temporal-spatial quantum graph convolutional neural network based on schrödinger approach for traffic congestion prediction. IEEE Trans Intell Transp Syst. 2022.
63. Leodolter W. Ai-based prediction in clinical settings: Can we trust it? 2019. https://healthmanagement.org/c/hospital/issuearticle/ai-based-prediction-in-clinical-settings-can-we-trust-it.
64. Zhao K, Jia Z, Jia F, Shao H. Multi-scale integrated deep self-attention network for predicting remaining useful life of aero-engine. Eng Appl Artif Intell. 2023;120: 105860.
65. Lecue F, Wu J. Explaining and predicting abnormal expenses at large scale using knowledge graph based reasoning. J Web Semant. 2017;44:89–103. https://www.sciencedirect.com/science/article/pii/S1570826817300252.
66. Akur8. 2021. https://akur8-tech.com/. Accessed 31 July 2023.
67. F. of Privacy Forum. Unfairness by algorithm: Distilling the harms of automated decision-making. 2017. https://fpf.org/wp-content/uploads/2017/12/FPF-AutomatedDecision-Making-Harms-and-Mitigation-Charts.pdf.
68. Angelov P, Soares E, Jiang R, Arnold N, Atkinson P. Explainable artificial intelligence: An analytical review. Wiley Interdiscip Rev: Data Min Knowl Discov. 2021;11.
69. Guidotti R, Monreale A, Turini F, Pedreschi D, Giannotti F. A survey of methods for explaining black box models. CoRR. 2018;abs/1802.01933. http://arxiv.org/abs/1802.01933.
70. Dosilovic FK, Bri M, Hlupic N. Explainable artificial intelligence: A survey. 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO). 2018. p. 210–5.
71. Zhong H, Wang Y, Tu C, Zhang T, Liu Z, Sun M. Iteratively questioning and answering for interpretable legal judgment prediction. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34. 01 ed. 2020. p. 1250–7. https://ojs.aaai.org/index.php/AAAI/article/view/5479.
72. European union general data protection regulation (gdpr). 2016. https://gdpr.eu/. Accessed 31 July 2023.
73. van der Velden BH, Kuijf HJ, Gilhuijs KG, Viergever MA. Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. Med Image Anal. 2022;79: 102470.
74. Rudin C, Chen C, Chen Z, Huang H, Semenova L, Zhong C. Interpretable machine learning: Fundamental principles and 10 grand challenges. CoRR. 2021;abs/2103.11251. https://arxiv.org/abs/2103.11251.
75. Abdul A, Vermeulen J, Wang D, Lim BY, Kankanhalli M. Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. 2018.
76. Machlev R, Heistrene L, Perl M, Levy K, Belikov J, Mannor S, Levron Y. Explainable artificial intelligence (XAI) techniques for energy and power systems: Review, challenges and opportunities. Energy AI. 2022;9.
77. Adadi A, Berrada M. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). IEEE Access. 2018;6:52138–60.
78. Gabbay F, Bar-lev S, Montano O, Hadad N. A lime-based explainable machine learning model for predicting the severity level of covid-19 diagnosed patients. Appl Sci. 2021;11:10417.
79. Ahmed AM, Kun Y, Chunqing G, Yuehui G. An optimized lime scheme for medical low light level image enhancement. Comput Intell Neurosci. 2022;2022:9613936.
80. Zhu H, Xue M, Wang Y, Yuan G, Li X. Fast visual tracking with siamese oriented region proposal network. IEEE Signal Process Lett. 2022;29:1437.
81. Goodfellow IJ, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. 2014. https://arxiv.org/abs/1412.6572.
82. Lyu C, Huang K, Liang HN. A unified gradient regularization family for adversarial examples. IEEE Int Conf Data Min. 2015;301–9.
83. Zhang S, Qian Z, Huang K, Wang Q, Zhang R, Yi X. Towards better robust generalization with shift consistency regularization. Intl Conf Mach Learn. 2021;12524–34.

84. Yuan X, He P, Zhu Q, Li X. Adversarial examples: Attacks and defenses for deep learning. IEEE Trans Neural Netw Learn Syst. 2019;30(9):2805–24.
85. Qian Z, Huang K, Wang QF, Zhang XY. A survey of robust adversarial training in pattern recognition: Fundamental, theory, and methodologies. Pattern Recogn. 2023;132.
86. Dave P. Ai is explaining itself to humans. And it's paying off. 2022. https://www.reuters.com/technology/ai-is-explaining-itself-humans-its-paying-off-2022-04-06/.
87. Jobin A, Ienca M, Vayena E. The global landscape of ai ethics guidelines. Nat Mach Intell. 2019;1:389–99.
88. Zhu J, Liapis A, Risi S, Bidarra R, Youngblood GM. Explainable AI for designers: A human-centered perspective on mixed-initiative co-creation. In: IEEE Conference on Computational Intelligence and Games (CIG). 2018. p. 1–8.
89. Miller T. Explanation in artificial intelligence: Insights from the social sciences. Artif Intell. 2019;267:1–38. https://www.sciencedirect.com/science/article/pii/S0004370218305988.
90. Kaur S, Singla J, Nkenyereye L, Jha S, Prashar D, Joshi GP, El-Sappagh S, Islam MS, Islam SMR. Medical diagnostic systems using artificial intelligence (AI) algorithms: principles and perspectives. IEEE Access. 2020;8:228049–69.
91. Chou YL, Moreira C, Bruza P, Ouyang C, Jorge J. Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms, and applications. Inf Fusion. 2022;81:59–83.
92. Bunt A, Lount M, Lauzon C. Are explanations always important? A study of deployed, low-cost intelligent interactive systems. In: Proceedings of the 2012 ACM International Conference on Intelligent User Interfaces, ser. IUI '12. New York, NY, USA: Association for Computing Machinery; 2012. p. 169–78. https://doi.org/10.1145/2166966.2166996.
93. Palacio S, Lucieri A, Munir M, Hees J, Ahmed S, Dengel A. XAI handbook: Towards a unified framework for explainable AI. CoRR. 2021;abs/2105.06677. https://arxiv.org/abs/2105.06677.
94. Jiang H, Wang M, Zhao P, Xiao Z, Dustdar S. A utility-aware general framework with quantifiable privacy preservation for destination prediction in lbss. IEEE/ACM Trans Netw. 2021;29(5):2228–41.
95. Han S, Ding H, Zhao S, Ren S, Wang Z, Lin J, Zhou S. Practical and robust federated learning with highly scalable regression training. IEEE Trans Neural Netw Learn Syst. 2023.
96. Craven MW, Shavlik JW. Using sampling and queries to extract rules from trained neural networks. In: Cohen WW, Hirsch H, editors. Machine Learning Proceedings 1994. San Francisco (CA): Morgan Kaufmann; 1994. p. 37–45. https://www.sciencedirect.com/science/article/pii/B9781558603356500131.
97. Ras G, van Gerven M, Haselager P. Explanation methods in deep learning: Users, values, concerns and challenges. CoRR. 2018;abs/1803.07517. http://arxiv.org/abs/1803.07517.
98. Johansson U, König R, Niklasson L. Rule extraction from trained neural networks using genetic programming. In: 13th International Conference on Artificial Neural Networks. 2003. p. 13–6.
99. Johansson U, König R, Niklasson L. The truth is in there - rule extraction from opaque models using genetic programming. 2004.
100. Zhou ZH, Jiang Y, Chen SF. Extracting symbolic rules from trained neural network ensembles. AI Commun. 2003;16(1):3–15.
101. Biswas SK, Chakraborty M, Purkayastha B, Roy P, Thounaojam DM. Rule extraction from training data using neural network. Int J Artif Intell Tools. 2017;26(3):1750006. https://doi.org/10.1142/S0218213017500063.
102. Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. 2015. https://arxiv.org/abs/1503.02531.
103. Tan S, Caruana R, Hooker G, Lou Y. Distill-and-compare. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. 2018. https://doi.org/10.1145/3278721.3278725.
104. Che Z, Purushotham S, Khemani R, Liu Y. Distilling knowledge from deep networks with applications to healthcare domain. 2015. https://arxiv.org/abs/1512.03542.
105. Xu K, Park DH, Yi C, Sutton C. Interpreting deep classifier by visual distillation of dark knowledge. 2018. https://arxiv.org/abs/1803.04042.
106. Friedman JH, Popescu BE. Predictive learning via rule ensembles. Ann Appl Stat. 2008;2(3). https://doi.org/10.1214/07-AOAS148.
107. Molnar C. Interpretable Machine Learning. 2nd ed. 2022. https://christophm.github.io/interpretable-ml-book.
108. Breiman L. Random forests. Mach Learn. 2001;45:5–32.
109. Fisher A, Rudin C, Dominici F. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. 2018. https://arxiv.org/abs/1801.01489.
110. Adhikari A, Tax DMJ, Satta R, Faeth M. Leafage: Example-based and feature importance-based explanations for black-box ml models. In: 2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE). 2019. p. 1–7.
111. Saarela M, Jauhiainen S. Comparison of feature importance measures as explanations for classification models. SN Appl Sci. 2021;3:02.
112. Ribeiro MT, Singh S, Guestrin C. Anchors: High-precision model-agnostic explanations. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32. 1st ed. 2018. https://ojs.aaai.org/index.php/AAAI/article/view/11491.
113. Ying R, Bourgeois D, You J, Zitnik M, Leskovec J. Gnnexplainer: Generating explanations for graph neural networks. 2019. https://arxiv.org/abs/1903.03894.
114. Sato R, Yamada M, Kashima H. Random features strengthen graph neural networks. 2020.
115. Kadir M, Mosavi A, Sonntag D. Assessing xai: Unveiling evaluation metrics for local explanation, taxonomies, key concepts, and practical applications. 2023.
116. Lundberg SM, Lee S. A unified approach to interpreting model predictions. CoRR. 2017;abs/1705.07874. http://arxiv.org/abs/1705.07874.
117. Štrumbelj E, Kononenko I. Explaining prediction models and individual predictions with feature contributions. In: Knowledge and information systems. 2014;41(3):647–665. https://doi.org/10.1007/s10115-013-0679-x.
118. Shrikumar A, Greenside P, Shcherbina A, Kundaje A. Not just a black box: Learning important features through propagating activation differences. CoRR. 2016;abs/1605.01713. http://arxiv.org/abs/1605.01713.
119. Islam SR, Eberle W, Ghafoor SK, Ahmed M. Explainable artificial intelligence approaches: A survey. 2021. https://arxiv.org/abs/2101.09429.
120. Nagpal A. L1 and l2 regularization methods, explained. 2022. https://builtin.com/data-science/l2-regularization.
121. Demir-Kavuk O, Kamada M, Akutsu T, Knapp EW. Prediction using step-wise l1, l2 regularization and feature selection for small data sets with large number of features. BMC Bioinform. 2011;12:412.
122. Huynh-Cam TT, Chen LS, Le H. Using decision trees and random forest algorithms to predict and determine factors contributing to first-year university students' learning performance. Algorithms. 2021;14(11). https://www.mdpi.com/1999-4893/14/11/318.
123. Sanjeevi M. Chapter 4: Decision trees algorithms. 2017. https://medium.com/deep-math-machine-learning-ai/chapter-4-decision-trees-algorithms-b93975f7a1f1.
124. Fayek A. Fuzzy logic and fuzzy hybrid techniques for construction engineering and management. J Constr Eng Manag. 2020;146:04020064.
125. Guo G, Wang H, Bell D, Bi Y. Knn model-based approach in classification. 2004.

126. Letham B, Rudin C, McCormick TH, Madigan D. Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. Ann Appl Stat. 2015;9(3). https://doi.org/10.1214/15-AOAS848.
127. Cheng L, Yin F, Theodoridis S, Chatzis S, Chang T. Rethinking bayesian learning for data analysis: The art of prior and inference in sparsity-aware modeling. IEEE Signal Process Mag. 2022;39(6).
128. Breiman L. Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). Stat Sci. 2001;16(3):199–231. https://doi.org/10.1214/ss/1009213726.
129. Sarkar S, Weyde T, d'Avila Garcez AS, Slabaugh GG, Dragicevic S, Percy C. Accuracy and interpretability trade-offs in machine learning applied to safer gambling. In: CoCo@NIPS. 2016.
130. Friedman JH. Greedy function approximation: A gradient boosting machine. Ann Stat. 2001;29(5):1189–232. https://doi.org/10.1214/aos/1013203451.
131. Green D, Kern H. Modeling heterogeneous treatment effects in large-scale experiments using Bayesian additive regression trees. Iowa City: The Annual Summer Meeting of the Society of Political Methodology; 2010.
132. Elith J, Leathwick JR, Hastie T. A working guide to boosted regression trees. J Anim Ecol. 2008;77(4):802–13. https://besjournals.onlinelibrary.wiley.com.
133. Goldstein A, Kapelner A, Bleich J, Pitkin E. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. J Comput Graph Stat. 2013;24.
134. Casalicchio G, Molnar C, Bischl B. Visualizing the feature importance for black box models. In: Machine Learning and Knowledge Discovery in Databases. 2019. p. 655–70. https://doi.org/10.1007/978-3-030-10925-7_40.
135. Han H, Li W, Wang J, Qin G, Qin X. Enhance explainability of manifold learning. Neurocomputing. 2022;500:877–95. https://www.sciencedirect.com/science/article/pii/S0925231222007044.
136. Liu S, Wang X, Liu M, Zhu J. Towards better analysis of machine learning models: A visual analytics perspective. Vis Inform. 2017;1(1):48–56. https://www.sciencedirect.com/science/article/pii/S2468502X17300086.
137. Apley DW, Zhu J. Visualizing the effects of predictor variables in black box supervised learning models. 2019.
138. Moustafa N, Koroniotis N, Keshk M, Zomaya AY, Tari Z. Explainable intrusion detection for cyber defences in the internet of things: Opportunities and solutions. IEEE Commun Surv Tutor. 2023;1–1.
139. Clement T, Kemmerzell N, Abdelaal M, Amberg M. Xair: A systematic metareview of explainable ai (xai) aligned to the software development process. Mach Learn Knowl Extr. 2023;5(1):78–108. https://www.mdpi.com/2504-4990/5/1/6.
140. Gurumoorthy KS, Dhurandhar A, Cecchi G, Aggarwal C. Efficient data representation by selecting prototypes with importance weights. 2017. https://arxiv.org/abs/1707.01212.
141. Kim B, Rudin C, Shah J. The Bayesian case model: A generative approach for case-based reasoning and prototype classification. 2015. https://arxiv.org/abs/1503.01161.
142. Bien J, Tibshirani R. Prototype selection for interpretable classification. Ann Appl Stat. 2011;5(4). https://doi.org/10.1214/11-AOAS495.
143. Olsson C. How to make your data and models interpretable by learning from cognitive science. 2017. https://medium.com/south-park-commons/how-to-make-your-data-and-models-interpretable-by-learning-from-cognitive-science-a6a29867790.
144. Kim B, Khanna R, Koyejo OO. Examples are not enough, learn to criticize! criticism for interpretability. In: Lee D, Sugiyama M, Luxburg U, Guyon I, Garnett R, editors. Advances in Neural Information Processing Systems. 29th ed. Curran Associates, Inc.; 2016. https://proceedings.neurips.cc/paper/2016/file/5680522b8e2bb01943234bce7bf84534-Paper.pdf.
145. Wachter S, Mittelstadt B, Russell C. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. Harv J Law Technol. 2018;31:841–87.
146. Mehedi Hasan MGM, Talbert D. Mitigating the rashomon effect in counterfactual explanation: A game-theoretic approach. In: The International FLAIRS Conference Proceedings, vol. 35. 2022.
147. Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, Fergus R. Intriguing properties of neural networks. 2013. https://arxiv.org/abs/1312.6199.
148. Su J, Vargas D, Sakurai K. One pixel attack for fooling deep neural networks. IEEE Trans Evolut Comput. 2017.
149. Athalye A, Engstrom L, Ilyas A, Kwok K. Synthesizing robust adversarial examples. 2017. https://arxiv.org/abs/1707.07397.
150. Leino K. Ai explainabilityrequires robustness. 2021. https://towardsdatascience.com/ai-explainability-requires-robustness.
151. Ilyas A, Santurkar S, Tsipras D, Engstrom L, Tran B, Madry A. Adversarial examples are not bugs, they are features. 2019. https://arxiv.org/abs/1905.02175.
152. Sadiku M, Shadare A, Musa S, Akujuobi C, Perry R. Data visualization. Int J Eng Res Adv Technol (IJERAT). 2016;12:2454–6135.
153. Lu Z, Cheng R, Jin Y, Tan KC, Deb K. Neural architecture search as multiobjective optimization benchmarks: Problem formulation and performance assessment. IEEE Trans Evol Comput. 2022.
154. Yang S, Li Q, Li W, Li X, Liu A. Dual-level representation enhancement on characteristic and context for image-text retrieval. IEEE Trans Circuits Syst Video Technol. 2022;32(11):8037–50.
155. Wu A, Wang Y, Shu X, Moritz D, Cui W, Zhang H, Zhang D, Qu H. Survey on artificial intelligence approaches for visualization data. 2021.
156. Khanna A, Pandey B, Vashishta K, Kalia K, Bhale P, Das T. A study of today's AI through chatbots and rediscovery of machine intelligence. Int J of u- and e-Serv, Sci Technol. 2015;8:277–84.
157. Yelekeri Jagadeesha RG. Artificial intelligence for data analysis and management. 2020.
158. Zhang J, Peng S, Gao Y, Zhang Z, Hong Q. Apmsa: Adversarial perturbation against model stealing attacks. IEEE Trans Inf Forensics Secur. 2023.
159. Acharjya DP, Ahmed KA. A survey on big data analytics: Challenges, open research issues and tools. Int J Adv Comput Sci Appl. 2016;7(2). https://doi.org/10.14569/IJACSA.2016.070267.
160. Jesus S, Belém C, Balayan V, Bento JA, Saleiro P, Bizarro P, Gama JA. How can i choose an explainer? An application-grounded evaluation of post-hoc explanations. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, ser. FAccT '21. New York, NY, USA: Association for Computing Machinery; 2021. p. 805–15. https://doi.org/10.1145/3442188.3445941.
161. Chang J, Boyd-Graber J, Gerrish S, Wang C, Blei D. Reading tea leaves: how humans interpret topic models. vol 32. 2009. p. 288–96.
162. Freitas A. Comprehensible classification models: A position paper. ACM SIGKDD Explorations Newsl. 2014;15:1–10.
163. Holzinger A, Carrington AM, Müller H. Measuring the quality of explanations: The system causability scale (SCS). Comparing human and machine explanations. CoRR. 2019;abs/1912.09024. http://arxiv.org/abs/1912.09024.
164. Grundy SM, Pasternak R, Greenland P, Smith S, Fuster V. Assessment of cardiovascular risk by use of multiple-risk-factor assessment equations. Circulation. 1999;100(13):1481–92. https://www.ahajournals.org.
165. Ismail AA, Bravo HC, Feizi S. Improving deep learning interpretability by saliency guided training. CoRR. 2021;abs/2111.14338. https://arxiv.org/abs/2111.14338.
166. Amann J, Blasimme A, Vayena E, Frey D, Madai VI, Precise4Q consortium. Explainability for artificial intelligence in

healthcare: A multidisciplinary perspective. BMC Med Inform Decis Mak. 2020;20(1):310.
167. Lipton ZC. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue. 2018;16(3):31–57. https://doi.org/10.1145/3236386.3241340.
168. Xiong W, Fan H, Ma L, Wang CM. Challenges of human—machine collaboration in risky decision-making. Front Eng Manag. 2022;9.
169. Damacharla P, Javaid AY, Gallimore JJ, Devabhaktuni VK. Common metrics to benchmark human-machine teams (HMT): A review. CoRR. 2020;abs/2008.04855. https://arxiv.org/abs/2008.04855.
170. Perelman BS, Mueller ST, Schaefer KE. Evaluating path planning in human-robot teams: Quantifying path agreement and mental model congruency. In: IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA). 2017. p. 1–7
171. Martin L, González-Romo M, Sahnoun M, Bettayeb B, He N, Gao J. Effect of human-robot interaction on the fleet size of AIV transporters in FMS. In: 2021 1st International Conference On Cyber Management And Engineering (CyMaEn). 2021. p. 1–5.
172. Ballav A, Ghosh M. Human factors of human machine interaction: Analyzing future trends through the past and the present. Int J Res. 2017;4:138–44.
173. Han K, Cook K, Shih P. Exploring effective decision making through human-centered and computational intelligence methods. 2016.
174. Lyons JB, Havig PR. Transparency in a human-machine context: Approaches for fostering shared awareness/intent. In: Shumaker R, Lackey S, editors. Virtual, Augmented and Mixed Reality. Designing and Developing Virtual and Augmented Environments. Cham: Springer International Publishing; 2014. p. 181–90.
175. Raheem F, Iqbal N. Artificial Intelligence and Machine Learning for the Industrial Internet of Things (IIoT). 2022. p. 1–20.
176. Qian M, Qian D. Defining a human-machine teaming model for ai-powered human-centered machine translation agent by learning from human-human group discussion: dialog categories and dialog moves. In: Degen H, Reinerman-Jones L, editors. Artificial Intelligence in HCI. Cham: Springer International Publishing; 2020. p. 70–81.
177. Xiong W, Fan H, Ma L, Wang C. Challenges of human—machine collaboration in risky decision-making. Front Eng Manag. 2022;9(1):89–103.
178. Jarrahi MH. Artificial intelligence and the future of work: human-AI symbiosis in organizational decision making. Bus Horiz. 2018;61(4):577–86.
179. Shin D. User perceptions of algorithmic decisions in the personalized ai system: perceptual evaluation of fairness, accountability, transparency, and explainability. J Broadcast Electron Media. 2020;64(4):541–65.
180. IBM. Building trust in ai. 2018. https://www.ibm.com/watson/advantage-reports/future-of-artificial-intelligence/building-trust-in-ai.html.
181. Ma Q, Liu L. The Technology Acceptance Model. 2005.
182. Kulesza T, Stumpf S, Burnett M, Yang S, Kwan I, Wong WK. Too much, too little, or just right? ways explanations impact end users' mental models. 2013.
183. Akyol E, Langbort C, Basar T. Price of transparency in strategic machine learning. 2016. https://arxiv.org/abs/1610.08210.
184. Igami M. Artificial intelligence as structural estimation: deep blue, bonanza, and alphago. J Econom. 2020;23.
185. Dignum V. Responsible artificial intelligence: Designing ai for human values. 2017.
186. Antoniadi AM, Du Y, Guendouz Y, Wei L, Mazo C, Becker BA, Mooney C. Current challenges and future opportunities for XAI in machine learning-based clinical decision support systems: A systematic review. Appl Sci. 2021;11(11).
187. Nie W, Bao Y, Zhao Y, Liu A. Long dialogue emotion detection based on commonsense knowledge graph guidance. IEEE Trans Multimed. 2023.
188. Liu X, He J, Liu M, Yin Z, Yin L, Zheng W. A scenario-generic neural machine translation data augmentation method. Electronics. 2023;12(10):2320.
189. Liu X, Shi T, Zhou G, Liu M, Yin Z, Yin L, Zheng W. Emotion classification for short texts: An improved multi-label method. Humanit Soc Sci Commun. 2023;10(1):306.