Artificial Intelligence Review (2022) 55:3503–3568
https://doi.org/10.1007/s10462-021-10088-y

Explainable artificial intelligence: a comprehensive review

Dang Minh[1] · H. Xiang Wang[2] · Y. Fen Li[2] · Tan N. Nguyen[3]

Published online: 18 November 2021

Abstract
Thanks to the exponential growth in computing power and vast amounts of data, artificial intelligence (AI) has witnessed remarkable developments in recent years, enabling it to be ubiquitously adopted in our daily lives. Even though AI-powered systems have brought competitive advantages, the black-box nature makes them lack transparency and prevents them from explaining their decisions. This issue has motivated the introduction of explainable artificial intelligence (XAI), which promotes AI algorithms that can show their internal process and explain how they made decisions. The number of XAI research has increased significantly in recent years, but there lacks a unified and comprehensive review of the latest XAI progress. This review aims to bridge the gap by discovering the critical perspectives of the rapidly growing body of research associated with XAI. After offering the readers a solid XAI background, we analyze and review various XAI methods, which are grouped into (i) pre-modeling explainability, (ii) interpretable model, and (iii) post-modeling explainability. We also pay attention to the current methods that dedicate to interpret and analyze deep learning methods. In addition, we systematically discuss various XAI challenges, such as the trade-off between the performance and the explainability, evaluation methods, security, and policy. Finally, we show the standard approaches that are leveraged to deal with the mentioned challenges.

Keywords Explainable artificial intelligence · Interpretability · Black-box models · Deep learning · Machine learning

1 Introduction

Artificial intelligence (AI) has been considered the most prevalent technology over the last couple of decades. According to a report by the International Data Corporation (IDC), the AI global expenditures are forecasted to reach nearly $100 billion in 2023, which is more than double the spending of $37.5 billion in 2019 (IDC 2020). In

Dang Minh and Tan N. Nguyen are co-first authorship and have been contributed equally to the work.

✉ Dang Minh
minhdl3@fe.edu.vn

✉ Tan N. Nguyen
tnnguyen@sejong.ac.kr

Extended author information available on the last page of the article

the meantime, Statista, which is a well-known online portal for statistics, predicts that yield from the global AI software market is forecast to expand significantly from $9.51 billion in 2018 to $118.6 billion by 2025 (Statista 2020). Gartner identifies AI-driven development as a leading trend in the *Gartner Top 10 Strategic Technology Trends 2020* (Gartner 2020). The statistics demonstrate that AI has already been widely adopted worldwide, and the rapid expansion of AI has had a huge impact on society. Consequently, humans have increasingly relied on decisions made by AI, which can be simple decisions, such as product recommendations, movie recommendations, and friend suggestions, to complicated decisions, such as autonomous vehicles in transportation.

Machine learning (ML) is a subclass of AI that depends on mathematical models to enhance machine intelligence. An ML model is constructed and trained on a specific dataset in order to automatically produce the predictions for any test sample without being explicitly programmed to do the job. Deep learning is currently the most popular subset of ML, which mimics how the human brain processes data and patterns to make a decision. For example, Rajkomar et al. (2018) and other researchers from Google achieved a remarkable accuracy of 95% in predicting the probability of a patients? death using the electronic health record (EHR) data of over 200,0000 US patients. Some typical applications that deep neural learning has been increasingly deployed are computer vision (CV) (Chan et al. 2015), natural language processing (NLP) (Minh et al. 2018), and the Internet of Things (IoT) (Dang et al. 2019). Even though deep learning has outperformed the traditional ML algorithms and achieved state-of-the-art performance across the industries, it is often referred to as the backbox that lacks opacity and transparency, because it cannot explain how a specific decision was made (Adadi and Berrada 2018; Guidotti et al. 2019).

A considerable number of ML algorithms are black-box models that do not reveal how the predictions were made so that humans can understand, because there is a trade-off between the model?s performance and its explainability (Deeks 2019), and the previous studies only focused solely on improving the system?s performance and ignored its transparency. However, it is challenging to convince the users to entrust applications that are based on the conventional algorithms in order to make crucial decisions, because they lack transparency, flexibility, and trustworthiness (Wang et al. 2019b). As a result, there has been a growing trend to develop a new generation of interpretable models that achieve comparable performance to the current state-of-the-art model. The possibility of an entirely interpretable model can help the researchers correct the model?s flaws and build the user confidence and trust. During the implementation of an AI system, the additional explainability can enhance its practicability for three reasons, which include guaranteeing fairness during the learning process, such as identifying and removing the bias in a dataset, improving the system?s robustness by indicating the possible noise that could affect the performance, and ensuring that the model uses only the essential features to infer the output. As a result, explainable artificial intelligence (XAI) was proposed to enhance the model transparency by proposing various methods that enable better model interpretability while maintaining the model performance (Escalante et al. 2018).

In this survey, we (1) provide a theoretical foundation of XAI, (2) categorize the latest XAI studies into three primary groups, which include pre-modeling explainability, interpretable mode, and post-modeling explainability, (3) discuss and compare the advantages and drawbacks of each approach from multiple perspectives, (4) focus on analyzing the research that equips explainability to the deep learning models, and (5) discuss various challenges and show the future research ideas.

1.1 XAI landscape

The increasing interest in XAI is due to the growing number of recent scientific events. XAI has progressively become an essential topic of committee discussions/tutorials at particular sessions at major conferences, such as ICCV (2019), ICML (2021), and BMVC (2020). Moreover, it has also become the key topic for the special issue of the top-ranking journals. Table 1 shows various XAI topics, which have been discussed in several scientific events.

The potential benefits of XAI lead to the introduction of important organizations and influencers that back it. Indeed, up until now, two of the leading players of the XAI topic include (1) a group of researchers and practitioners that operate under the *ACM Conference fon fairness, accountability, and transparency* or the ACM FAT* (ACM 2020) and (2) a group of experts backed by the *Defense Advanced Research Projects Agency* (DARPA) (Darpa 2020). FAT* is an annual conference that promotes and enables the explainability and fairness in AI systems and analyzes the social and economic impact. Since 2017, DARPA funded an XAI project with the ambition to develop a set of new methods that can explain AI systems. The program contains a total of 11 subprojects, and it is expected to run until 2021. The research groups backed by DARPA, which involves people from multiple educational institutions and various corporations, mainly focus on enhancing the explainability of complicated AI models for crucial security applications.

1.2 Relevant surveys

Table 2 provides detailed contributions of the recent comprehensive review papers, which investigated various aspects of the XAI. Overall, there is a growing interest in the XAI topic, because the number of XAI review papers increased significantly between 2017 and 2021.

In 2021, Ivanovs et al. (2021) released a survey that emphasized the pressing need for XAI and showed the current progress of the perturbation-based XAI approaches. On the other hand, Langer et al. (2021) paid attention to analyzing the XAI stakeholders and their requirements. Moreover, a unified framework was proposed to predict the required concepts and relations needed to develop a specific XAI model. Four XAI reviews were published in 2020. Among them, two research papers were dedicated to reviewing the main XAI approaches in specific fields. Guo (2020) concentrated on summarizing XAI for the 6G field, whereas Tjoa and Guan (2020) discussed the recent XAI approaches for the medical. The two remaining research focused on comprehensive XAI review. While Meske et al. (2020) showed the primary motivations for the XAI research and described essential stakeholders and requirements for the XAI studies, Arrieta et al. (2020) introduced a more comprehensive explanation of XAI that was based on the latest XAI studies. In 2019, three significant reviews about XAI were published. The survey conducted by Miller (2019) established a new definition of XAI by investigating over 250 papers. Moreover, they listed the significant challenges of XAI and showed the future directions. Carvalho et al. (2019a) surveyed the main achievements of the interpretable ML field. In addition, the author focused extensively on the societal impact of interpretable ML research. Finally, Guidotti et al. (2019) classified several XAI components, which included the algorithms, data, and problems, and investigated previous XAI research using these components. In 2018, the study conducted by Adadi and Berrada (2018) analyzed the key aspects of the

Table 1 XAI landscape summary

Type	Host	Year	Title	References
C	Neural Information Processing Systems (NIPS)	2017	Interpreting, explaining and visualizing deep learning	NIPS (2017)
	International Conference on Computer Vision (ICCV)	2019	Interpreting and explaining visual AI models	ICCV (2019)
	International Conference on Intelligent User Interfaces (IUI)	2019	Explainable smart systems	IUI (2019)
	International Joint Conference on Artificial Intelligence (IJCAI)	2019	Explainable AI	IJCAI (2019)
	International Conference on Automated Planning and Scheduling (ICAPS)	2020	Explainable planning	ICAPS (2020)
	British Machine Vision Conference (BMVC)	2020	Interpretable and explainable machine vision	BMVC (2020)
	International Conference on Machine Learning (ICML)	2021	Theoretic foundation, criticism, and application trend of XAI	ICML (2021)
J	Artificial Intelligence	2019	Ethics for autonomous systems	AI (2019)
	Pattern Recognition	2019	Explainable deep learning for efficient and robust pattern recognition	PR (2019)
	MDPI Electronics	2019	Interpretable deep learning in electronics, computer science and medical imaging	Electronics (2019)
	Signal Processing: Image Communication	2019	Explainable AI on emerging multimedia technologies	SP (2019)
	Artificial Intelligence	2020	Explainable artificial intelligence	AI (2020)
	Journal of the Academy of Marketing Science	2020	Explainable AI: from black box to glass box	Rai (2020)
	Future Generation Computer Systems	2021	Explainable AI for healthcare	FGCS (2021)
	Data Mining and Knowledge Discovery	2021	Explainable and interpretable machine learning and data mining	DMKD (2021)
	IEEE Computational Intelligence Magazine	2021	Explainable and trustworthy AI	CIM (2021)

C Conference & Workshop, *J* Journal & Special issue

Table 2 Summary of the previous XAI reviews, which include references, research field, and main contributions

References	Field	Contributions
Chakraborty et al. (2017)	General	• Presents the primary interpretability groups and categorizes the previous research based on these groups • Analyzes the current XAI challenges • Conducts a gap analysis of the future direction to advance the model interpretability
Adadi and Berrada (2018)	General	• Provides a comprehensive survey on the key aspects of the XAI topic • Presents the trending XAI methods and the primary research trajectories • Discusses the main concepts, motivations, and implications of implementing XAI
Zhang and Zhu (2018)	General	• Reviews the explainable deep learning models through various visualization methods • Examines the latest approaches to perform the pre-trained model interpretability • Discusses the current XAI challenges and future trends
Miller (2019)	Social science	• Analyzes over 250 XAI publications from the social science aspect • Presents the existing XAI challenges and discusses the future research direction • Addresses the related XAI concepts and shows the experimental results to support the concepts
Guidotti et al. (2019)	General	• Discusses several XAI components, which include the problem, the algorithm type, and the data type • Presents open challenges, which are related to the black-box models and explanations
Carvalho et al. (2019a)	General	• Reviews the current state of the interpretable ML • Concentrates on the societal impact, evaluation methods, and benchmark metrics of the interpretable ML • Discusses the future directions for the interpretable ML research to motivate more research on this field
Guo (2020)	Networking	• Outlines the primary XAI methods for wireless network configurations • Summarizes the fundamental XAI research in the 6G area • Deploys various XAI case studies for the optimization of both wireless PHY and MAC layer
Tjoa and Guan (2020)	Medical	• Reviews explainability and interpretability of the ML models • Categorizes previous interpretation approaches into three distinct groups • Standardizes interpretability mathematically and provides a medical case study
Meske et al. (2020)	General	• Shows the main drawbacks of the black-box AI models and motivations for XAI research • Generalizes the main goals, stakeholders, and requirements for implementing XAI techniques • Discusses challenges and directions for the future work
Arrieta et al. (2020)	General	• Introduces a general XAI concept that is based on the previous studies • Discusses and analyze the significant contributions from the previous XAI research • Shows the new XAI research trends to solve the existing drawbacks

Table 2 (continued)

References	Field	Contributions
Ivanovs et al. (2021)	General	• Explains the black-box problem in deep learning that leads to XAI • Reviews the previous perturbation-based attribution approaches for various types of input data • Outlines the future work for the perturbation-related studies
Langer et al. (2021)	General	• Presents five primary classes of stakeholders who demand the XAI and shows their requirements • Offers a unified framework to produce the primary concepts and relations required during the development of the XAI

emerging XAI topic and presented the trending XAI techniques, while Zhang and Zhu (2018) reviewed the visualization techniques for XAI and deep learning visualization. Finally, a survey was conducted by Chakraborty et al. (2017), which introduced the main interpretability approaches and categorized the previous work that was based on these approaches.

1.3 Contributions

The existing XAI reviews that were described in Sect. 1.2 proved that there is a growing activity in XAI research across sectors and disciplines (Arrieta et al. 2020; Guidotti et al. 2019). In addition, the establishment of several deep learning-based systems has recently added additional challenges for the implementation of XAI models (Chakraborty et al. 2017; Chan et al. 2015). Therefore, this manuscript summarizes and analyzes the fundamental topics that help the interested readers to gain comprehensive and latest knowledge regarding the XAI topic. Moreover, with a different approach from the other surveys, we reviewed and examined over 225 XAI publications by three levels of explainability, which include (1) pre-modeling explainability, which is gaining an insight into the dataset used to train the models, (2) interpretable model that contains the ML models that is explainable by nature, and (3) post-modeling explainability, which refers to a set of techniques implemented to enable the ML model explainability. Next, exciting ideas about enabling XAI in deep learning were conducted. Finally, we identified a list of challenges of XAI that need to be studied. All things considered, the primary contents of this review include the four items that are listed below.

1. An up-to-date comprehensive review of the explainable and interpretable ML models.
2. Categorizes the previous XAI methods based on three levels of explainability.
3. Focuses on systematically describing the latest collection of XAI techniques for deep learning.
4. Discusses several challenges and reveals future trends for XAI research.

1.4 Investigation methods

A difficult barrier during the paper composing period was to cover the most recent XAI literature, so we did various stages of search in order to cover as much literature as possible. Initially, the XAI papers were collected by querying the keywords related to the topic, such as *explainable AI*, *interpretable machine learning*, *XAI*, *model interpretation*, *interpretable AI*, *model visualization*, and *deep learning interpretation*. After that, more related papers from the literature review section of the previously collected articles were added. Moreover, we also collect notable papers from the previous XAI surveys. The layout of the manuscript was then outlined by thoroughly investigating all the collected papers. The fundamental techniques from each group of methods are presented in order to enable the readers to have a comprehensive understanding of the XAI topic.

The rest of the review is divided into six main sections. Section 2 discusses the background and various aspects of the XAI. The pre-modeling explainability and the main pre-modeling approaches are then described thoroughly in Sect. 3. After that, a detailed description of the characteristics of numerous interpretable models is represented in Sect. 4. Section 5 explains the trending post-modeling explainability methods, which have been widely adopted recently. The challenges and future research trends for the XAI topic are discussed in Sect. 6. Finally, the conclusion, which summarizes the contents of this review and provides some concluding remarks, is shown in Sect. 7.

2 Background

It is crucial to establish a common understanding of why it is necessary to promote explainability in the AI context, particularly ML algorithms, before proceeding with the main contents of this survey. Therefore, the main goal of this section is to analyze the previously introduced definitions that are related to XAI. After that, it demonstrates the importance of the explainability topic in AI. Finally, a general categorization of the XAI methods is established to guide the following sections.

2.1 General terminologies

Figure 1 describes the most commonly mentioned terminologies in the XAI domain, the relationships between them, and the essential characteristics of each term.

Among the listed terms, *understandability* appears as the most fundamental XAI notion that is linked to the other concepts (Hagras 2018). *Comprehensibility* and *understandability* are both dependent on the users? ability to perceive the knowledge that is learned by a model (Páez 2019). For example, *interpretability*, *explainability*, and *succinctness* are highly related to *understandability*. While *succinctness* indicates how concise and compact is the generated explanations to be understandable is for humans themselves (Abdollahi and Nasraoui 2018), the *interpretability* and *explainability* estimate the level that the observers can comprehend the outputs of the AI models. *interpretability* and *explainability* are among the two terms that seem to be related and usually misused, which can cause confusion and prevent the establishment of a standard term (Carvalho et al. 2019a). However, they are notably different in the XAI domain. *Explainability* refers to the active nature of an AI model that expresses any

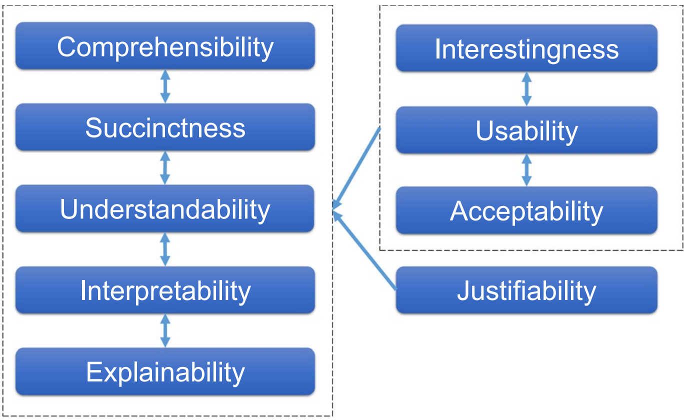


Fig. 1 Outline of the relationships between the common XAI terminologies. X → Y means that the assessment of Y requires the assessment of X, whereas X ↔ Y indicates that assessing X is equal to assessing Y. The boxes emphasize the equivalent classes of problems

ability or any procedure that the model takes in order to clarify or reveal its internal functions (Adadi and Berrada 2018). On the other hand, *interpretability* indicates the degree that an AI model becomes clear to humans in a passive way. Based on the mentioned terminologies, some user-related terminologies can be comprehensively assessed. *Justifiability* offers a simple way for non-technical users to perceive the inner learning processes of a learning model and allows them to justify the model. When an AI model becomes explainable, it attracts the users (*Interestingness*). As a result, in general, XAI improves the *usability* and *acceptability* of the existing AI models as it allows the users to get involved in the process of debugging and building the models.

2.2 What is XAI?

The idea of interpreting the AI system was first introduced in the mid-90s when Swartout and Moore (1993) proposed an XAI prototype that could describe what hard-coded rules contributed to a decision. The XAI phrase was initially suggested by Van Lent et al. (2004) in order to define the system capability to describe the internal process of the objects that were controlled by a game simulation. It contradicts the current *black box* nature of the AI systems, where the researchers and developers find it hard to explain the decisions made by AI. The XAI development was delayed for a long period when AI entered a point of inflection, where AI algorithms showed remarkable results in several research areas. The main objective of AI research since then has been turned into improving the algorithms? predictive power. Consequently, the ability to interpret and explain the decisions that are predicted by AI algorithms has been ignored. In recent years, the XAI term has gained increasing attention from academia and developers as an immediate outcome of the massive integration of AI/ML in everyday life (Páez 2019). As a result, the pressure from society, ethics (Muller et al. 2021), and legislation (Schneeberger et al. 2020) demands a new generation of AI that can explain its inner functions and allow the users to interpret the logic chain that brings about its decisions. Figure 2 demonstrates the remarkable revival of XAI research, which is based on observing Google trends and the rising number of XAI publications during the last decades. Fig. 2(1) shows that there is a gradually increasing interest in the XAI, because the volume of Google searches for the explainable AI keyword rose significantly during the period between 2014 and 2020.

The research community has acknowledged the necessity of a comprehensive and standard definition that covers all the crucial characteristics of XAI for years. As a result, many theories have been proposed to try to explain the XAI concept thoroughly. However, there is no unified definition of the XAI term because it is usually associated with the changes, efforts, and initiatives to establish transparent AI and solve the trust concerns instead of being a standard concept. For the moment, the most accepted definition can be viewed from different explanations that are accepted by two well-known organizations. As presented by DARPA (Darpa 2020), the XAI program tries to build a series of ML methods that ?produce more explainable models, while maintaining a high level of learning performance (prediction accuracy); and enable human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners.? On the other hand, according to FAT* (ACM 2020), ?XAI is the study of explainability and transparency for socio-technical systems, including AI.?

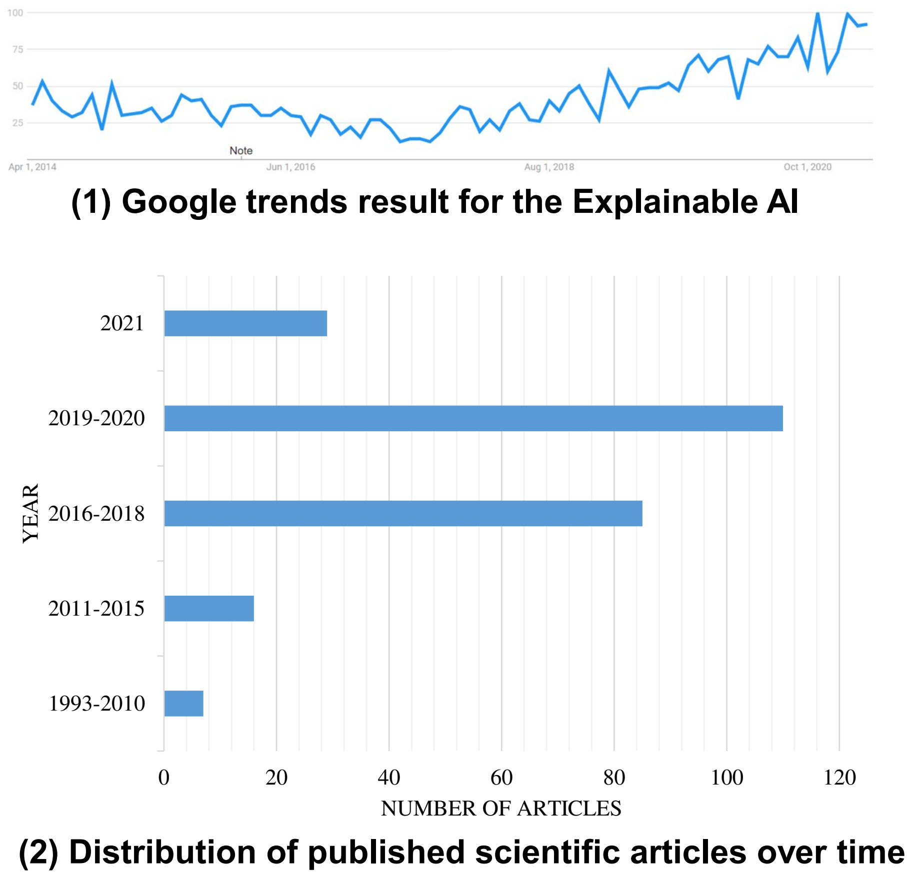


Fig. 2 Proofs of the increasing interest in XAI research from the community, which include (1) Google trends for the XAI topic and (2) distribution of published XAI-related articles by the years that were investigated in this survey

2.3 Why is XAI important?

It is intuitively obvious that we are the sole person responsible for our own decisions and actions. However, the liability for a decision made by an AI algorithm is vague, because the AI systems cannot explain their internal process, which is mentioned in the introduction section. The AI system relies on a mathematical model to learn the fundamental features from a dataset in order to make a prediction or a suggestion. In addition, it is laborious to shed some light on the complicated internal procedures of an AI model with the current technology (Adadi and Berrada 2018). Therefore, there is an increasing demand for a new generation of XAI technology to completely understand how AI models make predictions. The explanation of the decision-making process in AI systems is particularly critical for various industries, such as financial, healthcare, and security. For example, a self-driving Uber hit a woman and caused her death in Arizona in 2019 (UberAccident 2020). It is troublesome to decide who is responsible for this profoundly significant and moral situation. Therefore, an insight into the decision rationale of AI is required to guarantee the trustworthiness and the responsibility.

The lack of the ability to explain the logical reason why some ML algorithms have reached human-level performance is rooted in two primary problems. The first problem is the huge differences between the research community and the business sectors that prevent a complete integration or replacement of the newest ML systems into the rigorously controlled industries, because the new technologies can put existing systems at risk if the users do not fully understand them (Asadi et al. 2017). The second issue is the uncertainty regarding the AI?s performance. ML models have been implemented in numerous applications, and some of them are starting to reach levels of human (Deeks 2019). Due to the adoption of the new generation AI and ML techniques, these applications can process an enormous amount of data with high accuracy. Even though each research paper showed that an AI system achieved high performance in particular disciplines, the performance is just part of the users? concern, and XAI is the main element that allows a better understanding of the model and improves its applicability (Chakraborty et al. 2017).

2.4 XAI notable features

Even though the notable characteristics of XAI systems have been revealed slowly through numerous XAI studies (Ding 2018; Lawless et al. 2019), to the best of our knowledge, there still exists no XAI survey that mentions and discusses every significant characteristic of an XAI system. These characteristics are essential in order to discriminate against the primary objective of an XAI model. As a result, this section attempts to explain and discuss them in detail in Table 3. In total, XAI has eight fundamental characteristics, which include reliability, causability, transferability, informativeness, confidence, fairness, accessibility, and privacy.

2.5 How is XAI implemented?

A well-known XAI classification introduced by Doran et al. (2017) divided the XAI techniques into three levels of explainability, which include (1) the opaque models, which are models where the users are unable to comprehend how those models produce an output for a specific input, (2) the interpretable models, which are models where the users can mathematically analyze the connection between the model input and output, and (3) the comprehensible models, which is where the models can provide both the output and a set of rules to support the users in order to gain an insight into how the model works. The mentioned categorization criterion is included in the new XAI classification that is introduced in this article, which gives a more precise categorization based on the latest progress of XAI.

Figure 3 describes a systematic categorization for all the existing XAI approaches which include the pre-modeling explainability, the interpretable model, and the the post-modeling explainability. Important methods for the pre-modeling explainability group are the *data analysis*, *data summarization*, and *data transformation*. There are several approaches for the interpretable model group, which include the *inherently interpretable model* and the *hybrid interpretable model*. Finally, for the post-modeling explainability, some common methods are *textual justification*, *visualization*, *simplification*, and *feature relevance*.

Table 3 Description of eight notable XAI characteristics

Characteristic	Explanation	References
Reliability	The certainty of whether a model perform as designed when it is assigned a task	García-Magariño et al. (2019)
Causability	The quality of explanations by delivering a specified level of causal understanding to the human experts	Holzinger et al. (2019)
Transferability	A standard explainable model can be transferred to be used in other topics and obtain robust results	Long et al. (2018)
Informativeness	The explainable model gives more information about the problem being tackled	Kapelner et al. (2018)
Confidence	The explainable frameworks are robust, stable, and trustful	Arrieta et al. (2020)
Fairness Justifiability	An output of an explainable model can be judged and subjected to examination	Ahn and Lin (2019)
Accessibility Interactivity	The ability to directly interact with the decision-making process of an explainable model	Baniecki and Biecek (2019)
Privacy	The ability to describe the internal operations of a model by a third party	Baron and Musolesi (2020)

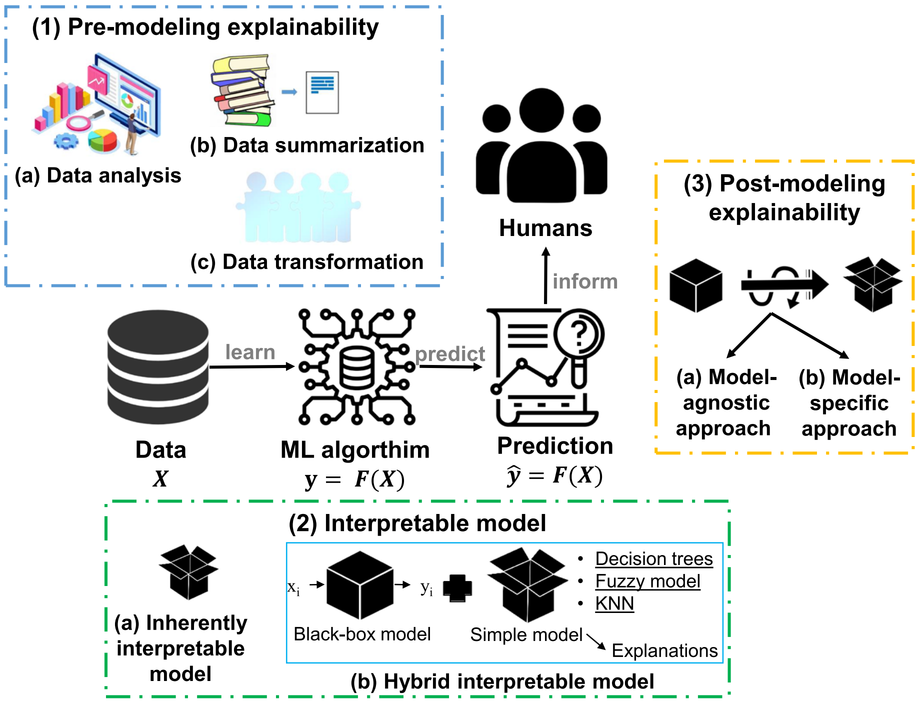


Fig. 3 Conceptual diagram that represents the three degrees of explainability, which include (1) the pre-modeling explainability, (2) the interpretable mode, and (3) the post-modeling explainability

2.5.1 Pre-modeling explainability

Pre-modeling explainability refers to a series of data processing approaches, which is implemented to gain an insight into the datasets collected to train the ML models.

- The *data analysis* implements a set of techniques in order to obtain an overview of the various statistical information of a dataset, which includes the dimensionality, mean, standard deviation, range, and the missing samples (Zhuang et al. 2017). Consider a road defect classification application as an example, which was based on a huge road defect dataset (defect and normal images) (Dang et al. 2018). By implementing the *data analysis*, the frequency of the images between the defect and the normal classes exposes the imbalanced data problem, where the number of defect images is far less than the normal ones. As a result, many solutions can then be used in order to mitigate the problem and increase the classifier?s performance (Hu et al. 2018b).
- In the digital transformation era, where the large-scale deployment of AI technologies has grown at a rapid rate (Holzinger et al. 2021b). There is a pressing need to collect huge datasets to support those AI applications. Although the number and the quality of the datasets have improved significantly due to a more accessible and straightforward data collection process, the datasets were often published without sufficient documentation, so it was challenging for other researchers to apply these datasets in their studies (Yu et al. 2017). *Transformations* can guarantee decent interactions between

the creators and the users of the datasets and can further reduce the common problems, such as data bias and the misuse of the data (Anysz et al. 2016). In recent years, novel data transformation approaches, which include data statements (Bender and Friedman 2018), datasheets, and nutrition labels, have been introduced. Each method essentially proposes various solutions for important metadata for a dataset in order to describe the data creation, data preparation, data collection, and legal/ethical consideration.

- *Data summarization* techniques attempt to find a minimal subset from the original dataset (Ahmed 2019). The model?s performance that is trained on the subset is comparable to the original dataset, because the subset contains representative samples, which can represent the entire dataset. The conventional data summarization techniques include K-medoid clustering (Mohit et al. 2019) and K-means clustering. The research about data summarization has increased remarkably in recent years, mainly due to the increasing number of publicly available big datasets. Yang and Shafto (2017) proposed a Bayesian-based teaching model, which could pick a representative small number of data samples that would yield the same results as when the learner is trained using the whole dataset. Wu et al. (2017) showed that various *data summarization* techniques were introduced for document summarization, video summarization, and classification tasks. While the document summarization and the classification tasks are formulated as an optimization problem, the video summarization is performed by extracting the keyframes that best describe the video or by applying the video skimming method. In addition to data summarization, data squashing, which contains a set of techniques that create a subset from the original dataset to generate a similar analysis result as the original dataset, was also investigated (DuMouchel 2002). The samples in the subset of the data squashing method usually contain weights, which is different from the data summarization. The recent work on so-called Bayesian coresets is a typical example of the data squashing that is expressed in the Bayesian learning environment (Campbell and Broderick 2019).

Hu et al. (2018a) suggested DIVE, which is a mixed-initiative data exploration system that combines several methods, such as *data summarization*, *visualization*, *statistical analysis*, and *storytelling* to gain knowledge about the data. The authors proved that using the DIVE helped the data scientists to perform the data visualization and data analysis faster and more efficiently. However, this technique merely considered the statistical information when analyzed the datasets. It is challenging to conceive this type of data, because most of the datasets are complicated, suffer from the high-dimensional space problem (Jagadish et al. 2014), and humans only perceive up to three-dimensional data. Therefore, the following approaches were proposed in order to enhance the model explainability.

2.5.2 Interpretable model

The ML models with complicated processes and architectures, such as deep learning, have successfully been applied to solve various ML and AI challenging problems over the past decade (Nguyen et al. 2019, 2020b). However, the current generation of models faces the black-box problem because they are trained directly from data by an algorithm, indicating that researchers who create them cannot explain how the model uses the variables to produce predictions. Even though the input variables are available, the black-box models are considered complex functions of the variables that we cannot understand how the final output is created from these variables. The black-box problem can ideally be prevented in the early stages by constructing interpretable models. A model is considered an interpretable model if it can be interpreted by humans all by itself by looking at the model summary or the model parameters. Some examples of the interpretable models include linear/logistic regression and rule-based learners. Those models will be discussed in detail in Sect. 4.

- *Inherently interpretable model* approach is a conventional way to achieve interpretability, which contains a group of models and algorithms that are considered understandable by design. The standard algorithms include rule sets, linear models, decision trees, case-based reasoning, and generalized additive algorithms. Lipton (2018) classified this group of models into three degrees of transparency, which include *simulatability*, which is the whole system level, *decomposability*, which is the smaller parts level, and *algorithmic transparency*, which is the algorithm level. As illustrated in Fig. 4, the previous group includes its successors. A model that belongs to the *simulatability* group has the *algorithmic transparency* and the *decomposability* properties. However, merely adopting a simple and plain model does not automatically ensure the model explainability in practical. For example, feeding high-dimensional input data into a linear regression model can affect the simulatable characteristic of the model (Deleforge et al. 2015). There have been many solutions proposed to solve the issue. For example, Luo et al. (2016) investigated the implementation of the L1 norm regularization method before training the model in order to minimize the number of crucial input features. However, the coefficients that were computed for a linear regression model could likely be unsteady when feature collinearity happens, which caused some variables to be correlated to each other due to an observed or an unobserved confounder. In this situation, the L2 norm could be used to mitigate the problem (Fang et al. 2017).
- The *hybrid interpretable model* approach includes a set of methods that attempts to combine a complex black-box model with an *inherently interpretable model* in order to build an interpretable model that achieves comparable performance to the black-box model. For instance, Gallego et al. (2018) demonstrated a clustering-based k-nearest neighbor classification that combined an approximated similarity search (ASS) method with the clustering model to lower the complexity of the k-nearest neighbors (k-NN) algorithm. Deep learning was also implemented to acquire a proper representation of the classification task. The experimental results on eight distinct datasets confirmed

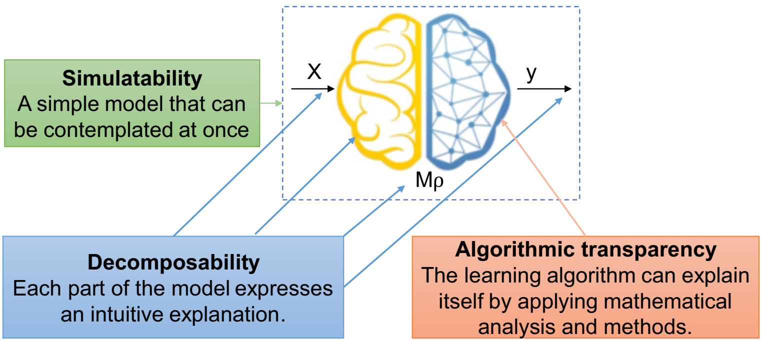


Fig. 4 Visualization of the inherently interpretable model approach, which show three main characteristics that include simulatability, decomposability, and algorithmic transparency

that the combination of various techniques enabled a notable performance improvement with a substantial decline in the number of parameters required to categorize an item.

Even though there were many techniques proposed to limit the weaknesses of the *interpretable model* approaches, these models, in general, are simple and inefficient to cope with the real-world applications (Salmeron et al. 2019). This hypothesis has led to the tradeoff between the model?s performance and explainability, because the better the algorithm performs, the lower the interpretability becomes (Ren et al. 2019). The main obstacle is to develop a simple model that can be comprehended easily but complicated enough to fit the data correctly. Recent research has shown that the previously mentioned tradeoff does not always correct, because many new XAI methodologies have been proposed in this category. Those methods were organized using ideas, key underlying factors, and probably overlapping to help the models be more straightforward to comprehend.

2.5.3 Post-modeling explainability

The post-modeling explainability approach enhances the explainability of the existing black-box ML models by applying a set of techniques, such as *visualization*, *textual justification*, *simplification*, and *feature relevance* techniques. These methods were inspired by how humans interpret a system and its processes. Even though each group also specifies the type of data it needs, some techniques from different groups may perform well on different data types. The latest post-modeling explainability research is classified into the proposed categories in order to reduce the time and effort that the readers have to spend looking for specific research that suits their demands.

- The *textual justification* facilitates the interpretability ability for the ML models by producing detailed explanations in the form of text for every decision made by the ML models. This group also comprises the techniques that create symbols to describe a model?s function or algorithm through semantic mapping.
- While *textual justification* explains a model by generating explanations in the form of text, *visualization* interprets a model?s behavior by visual representations. *Visualization* techniques are considered the best way to explain the complicated inner interactions of the variables of the model, and they can be combined with other methods in order to increase their interpretability ability (Spinner et al. 2019).
- The *simplification* aims to make a simplified version of the original model that has an optimized function, significantly reduces the complexity, has a simpler implementation process, and performance is comparable to the original version. *Local explanations* and *Examples generation* are considered belong to the *simplification* approach. *Local explanations* help the researchers interpret the ML models by splitting a black-box model into several simple subprocesses. Each subprocess can be explained using a distinctive technique that only solves part of the model?s functioning. *Examples generation* technique extracts the data samples that are associated with the outputs of a particular model to allow the users to gain a better comprehension of the model, which is similar to human behavior when solving a specific task. This approach concentrates on obtaining representative samples that describe the internal relations correlations of the model under consideration.
- The *feature relevance* approach describes the inner function or process of a model by calculating the relevance score for the available variables. The importance (sensi-

tivity) of a feature to the model?s prediction can be analyzed based on the computed score. The score comparison between the variables can then be computed to show the model?s attention to a list of variables during the testing process. In addition, the relevance score between the variables can also be calculated in order to explain their relationship.

- The *joint prediction and explanation* approach assumes that an ML model can predict and explain the output at the same time. Alternatively, a complicated model can be forced to explain its prediction during the training process. For example, Hind et al. (2019) proposed the Teaching Explanations for Decisions (TED) framework with the primary objective to increase the training data, which comprises a list of essential features, a prediction, and the corresponding explanation. The provided prediction and corresponding explanation were then combined into a single label during the training. During the testing process, the output prediction was decoded in order to create the prediction and the corresponding explanation. The TED model was proved to show accurate explanations with no loss in prediction.

2.6 Real-world applications of XAI

The last three years have witnessed a sharp rise in XAI research activity, which focused on various aspects of the XAI topic. Table 4 discusses the potential applications of XAI in some representative domains, which include finance, healthcare, transportation, military, legal, and human-computer interactions.

In addition, Table 5 describes many open-source XAI platforms that were funded by giant tech companies. Most of the open-source platforms support the pre-modeling and post-modeling explainability for the black-box models. The big corporations, such as Google, Microsoft, and Oracle, started to focus more on integrating XAI into their ecosystems, proving XAI?s important role in recent years.

3 Pre-modeling explainability

Pre-modeling explainability refers to various data pre-processing and feature exploration methods in order to obtain an overview of any dataset and pre-process it before the training process. Table 6 summarizes the previous research that worked on pre-modeling explainability. There are three major pre-modeling explainability categories, which include the *data analysis*, *data summarization*, and *data transformation*.

3.1 Data analysis

The data analysis is described as a process of extracting, transforming, loading, and modeling data in order to identify the crucial features required for the ML models to make decisions. The four stages of data analysis that are usually encountered in data science include descriptive analysis, predictive analysis, diagnostic analysis, and prescriptive analysis.

Table 4 The review of XAI research in various application domains

Domain	References	Contents
Finance	Zheng et al. (2019)	• Explains financial intelligence and its role in the fintech field • Analyzes the state-of-the-art financial intelligence systems in numerous sectors, such as financial consulting, financial security, and risk management • Introduces FinBrain, which solves four open problems of the XAI
	Liberati et al. (2017)	• Introduces a linear kernel reconstruction that enables explainability • the reconstruction method stabilized the loss and brought good interpretability in the practical credit scoring experiment
Healthcare	Vellido (2019)	• Reviews recent studies about the interpretability and explainability of ML algorithms in healthcare • Concentrates on the data and model visualization
	Wang et al. (2019b)	• Presents a view of the AI *black box* of medicine • Introduces and analyzes the current research on AI *black box* of medicine • Shows challenges that must be solved to develop a more explainable and interpretable healthcare model
	Holzinger (2016)	• Shows the importance of the human-in-the-loop for the health informatics, which brings human experience and conceptual knowledge to the AI processes • Proves that the human model is also crucial to the development of human-AI interfaces
Self-driving	Lee et al. (2019)	• Implements the interpretable gradient boosting method to enhance the model?s interpretability • The proposed model contains numerous interpretable features, which enable it to achieve higher predictive performance
	Kim and Canny (2018)	• Applies a visual attention approach, which enables an explainable convolutional network system that is in charge of the steering angle • Implements a filtering approach to explain what input regions affect the prediction results and remove unimportant features • The trained model describes the explainable features that affect the automated steering system while driving
Robotics	Felzmann et al. (2019)	• Discusses why transparency to stakeholders is critical for autonomous systems, such as robotics • Introduces a list of requirements for designers to achieve transparency for the autonomous systems
	O?sullivan et al. (2019)	• Shows the main challenges of implementing explainable robotic surgery • Recommends necessary agents for creating and developing appropriate explainable frameworks or standards • Focuses on analyzing accountability, liability, and culpability when develop a new system

Table 4 (continued)

Domain	References	Contents
Military	Keneni et al. (2019)	• Proposes an XAI method that shows and explains how the systems make a decision • Can be integrated into the existing autonomous systems to make them more transparent, understandable, and trustworthy
	Wasilow and Thorpe (2019)	• Introduces an ethics evaluation benchmark for emerging AI and robotics systems • Validates the proposed assessment framework in a contextual environment • Shows how the benchmark helps the developers and other stakeholders discover potential ethical issues
Legal	Deeks (2019)	• Proposes an explainable framework that reveals how the algorithms make predictions to support the judge in making a decision • Presents the advantages of the explainable framework that is built from the bottom-up and based on a case-by-case consideration to make decisions
	Raaijmakers (2019)	• Shows the existing *black-box* problem of AI frameworks for law enforcement that need to be solved to make them trustworthy • Proves that explainable and auditable AI is crucial, especially in the legal field • Analyzes fundamental factors of the XAI frameworks for law enforcement

Table 5 Open-source XAI platforms

Name	Backed by	Explainability		Characteristics
		Pre-modeling	Post-modeling	
AI fairness 360	IBM (2019)	✓	✓	Mitigates bias for datasets and models Available in both Python and R
What-If tool	Google (2021)	✓	✓	Analyzes data features and model behavior Extension in Jupyter and Google Cloud
Model interpretability	Microsoft (2021)	✓	✓	Presents feature importance for the model Extracts data patterns using interactive GUI
Skater	Oracle (2021)		✓	Supports black-box models demystification Open-source python library
H2O platform	H2oai (2017)		✓	Supports post-hoc explainability toolkit White-box modeling with AutoML

After that, machine learning (ML) and AI are used to predict the outcomes and suggest options to respond to those predictions.

3.1.1 Descriptive analysis

The descriptive analysis or descriptive statistics describes, shows, or summarizes raw data points to provide insightful information about the data. This process is often used to represent data in the past to enable the data scientists to study earlier behaviors and figure out how they can impact future outcomes. Typically, the fundamental data is described by applying a series of statistics in order to perform simple to complex operations, such as aggregate amounts or the counting of a filtered column. Examples of descriptive analysis are documents that summarize the company?s finances, production, inventory, sales, operations, and customers.

3.1.2 Predictive analysis

The predictive analysis is rooted in the capability of a model to predict future outcomes based on probabilities. Predictive analysis equips the researchers with actionable insights on estimating the likelihood of a prediction using raw data. The most typical application that uses predictive analysis is the creation of a credit score. The score describes an individual?s creditworthiness, which is managed by financial services in order to check whether a customer can pay off loans on time. Other applications include estimating how sales end at year-end, discovering items that customers frequently bought together, and managing inventory using historical data.

Table 6 Summary of the previous pre-modeling research, which are grouped by category

Category	Year	Dataset	Approach	Results	References
Ds	2019	Self-collected	Long short-term memory (LSTM) & single-layer CNN	ROUGE-1 of 34.9%; ROUGE-2 of 17.8%.	Song et al. (2019)
	2017	VSUMM and VYT video databases	A novel clustering method	Recall and F-score of 0.63; Precision of 0.68 Running time of 0.014s	Wu et al. (2017)
	2017	SKE and BC3 email corpora	Ensemble Noisy Auto-Encoder	The ROUGE-2 recall improved on average 11.2%	Yousefi-Azar and Hamey (2017)
	2019	Object information and Places365 datasets	A tree-based method with a two-step optimization approach	Accuracy of 73.2%	Pan et al. (2019)
	2016	Three video clips	SpiralTape approach	The output is aesthetically pleasing. Intuitively and naturally personalizing of video browsing	Liu et al. (2016)
	2016	CAVIAR, ViSOR, and CUHK benchmark surveillance datasets	Cumulative moving average (CMA) and the preceding segment average (PSA)	Obtains higher performance than previous research	Dogra et al. (2016)
Da	2016	2004 KDD Contest (10498 rows and 77 columns)	Interaction and visualization techniques for analyzing high-dimensional data	The model has a robust exploratory analysis ability on high-dimensional data	Turkay et al. (2016)
	2018	Publicly available animal dataset (49 animals with 72 attributes)	Methods to visualize and interact with high-dimensional data	Clarifies the distinction of observation-level interaction for interacting with dimension reduction models	Self et al. (2018)
	2018	Five high-dimensional genomic datasets	Computationally fast heuristic variable importance	The proposed method requires considerably less computation time compared to other methods	Janitza et al. (2018)
	2019	Self-made	Scented widgets	Expands the number of questions requested about data Expands the analysis ability without sacrificing depth	Sarvghad et al. (2016)

Table 6 (continued)

Category	Year	Dataset	Approach	Results	References
Dt	2020	Benchmark datasets (iRoads, Caltech-256, Caltech-101)	Construction of an ontology for data standardization (PCLiON)	The proposed PCLiON has standardized 320 attribute annotations and 11 object attributes	Chen et al. (2020b)
	2018	Clinical dataset	Proposes necessary steps to prepare data for a research study	Increases and improves data transparency. Provides guidelines for accurate data management	Lapchak and Zhang (2018)
	2020	Yahoo finance dataset	SVM	Significantly improves prediction performance	Kumari and Swarnkar (2020)

Da, Ds, and Dt stands for data analysis, data summarization, and data transformation, respectively

3.1.3 Diagnostic analysis

The diagnostic analysis involves many methods, such as data discovery, drill-down, correlations, and data mining, to describe why some events occurred. More importantly, diagnostic analytics lets researchers comprehend the data, identify anomalies promptly, and figure out the potential hidden relationships between multiple anomalies.

3.1.4 Prescriptive analysis

The prescriptive analysis enables the users to decide various feasible actions based on outputs from the algorithms. Moreover, the prescriptive analysis forecasts the output of a decision and why that output will happen. Although it involves both the descriptive analysis and the predictive analysis, it exceeds them by offering a wider range of methods. The prescriptive analysis implements a series of methods, such as algorithms, business rules, ML, and computational modeling procedures on various types of datasets, including real-time data feeds, historical/transactional data, and big data.

3.2 Data summarization

The summarization approach refers to the process of creating an informative and compact summary of the initial, which includes unstructured data and structured data. The unstructured data refers to a huge collection of plain text, dates, numbers, and punctuations. Therefore, text summarization is a fundamental preprocessing process before performing the training. On the other hand, the structured data indicates any data in an established field, such as rows and columns within a file or matrix, which involves spreadsheets and relational databases. The summarization approach has been widely implemented in various application domains, which include text mining, traffic network monitors, the financial sector, the healthcare sector, and several others (Kolyshkina and Simoff 2021). The meaning of summarization depends on the intention of implementing it. For example, the purpose of a network traffic summary and a text summary are different. A text summary keeps the essential contents and reduces a large amount of unnecessary text. On the other hand, a network traffic summary helps the administrator get an overview of the network in real-time.

3.3 Data transformation

The data transformation is a basic rule-based data processing method that is applied to map the structure of the source datasets into a target structural format. However, the data consistency is ensured, which means the data has a similar format and content. The client?s name is a good data transformation example that can be expressed in various forms. A proper data transformer can parse distinct parts of the client?s name, which include titles, last names, first names, and middle names, and it then organizes them into a rule-based representation in order to assist the data manipulation from other services.

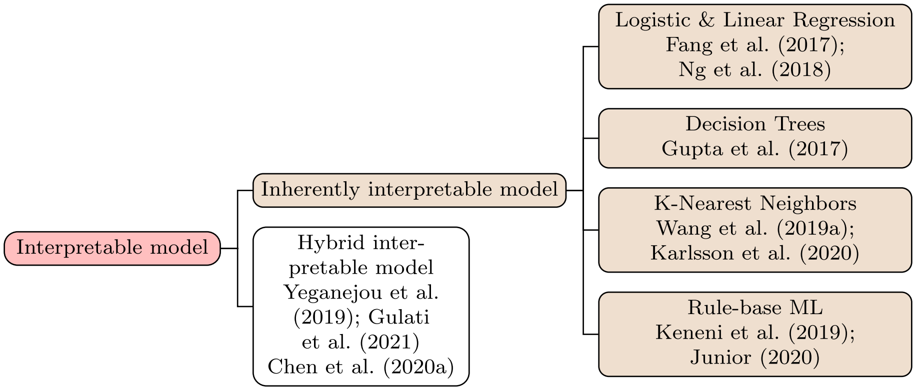


Fig. 5 Categorization of the interpretable model topic based on the previous notable research, which include the inherently interpretable model and the hybrid interpretable model

4 Interpretable model

The interpretable group refers to a collection of ML algorithms that is interpretable due to its simple structure. Previous research are divided into two main groups, which include the inherently interpretable model and hybrid interpretable model, as described in Fig. 5.

4.1 Inherently interpretable model

Most conventional ML algorithms have already been deployed to support the users in order to make high-stakes decisions. However, they still caused several issues across domains, such as public health, criminal justice, and others. The research community has expected that introducing new approaches in order to solve these black-box algorithms is going to solve some of these dilemmas. However, investing efforts in order to explain the existing black-box algorithms instead of constructing algorithms that are inherently interpretable is prone to cause bad practices, which can lead to catastrophic harm to society. Therefore, designing inherently interpretable models is a way forward. As suggested by Lipton (2018), the inherently interpretable approaches can be divided into three levels of explainability, which include simulatability, decomposability, and algorithmic transparency, as described in Table 7.

4.1.1 Linear/logistic regression

The linear regression algorithm is applied in order to predict the continuous dependent variables using a certain set of independent variables. On the other hand, the logistic method is implemented to model the possibility of a predefined class or event existing within a provided set of independent variables. Even though linear/logistic regression is considered transparent because it assumes that predictors and the predicted variables are linearly dependent, some post-modeling explainability approaches can be implemented to provide more explanations in the form of visualization for non-expert audiences. For example, Latouche et al. (2018) combined logistic regression and a residual network to build

Table 7 Descriptions of three levels of inherently interpretable ML models, which include simulatability, decomposability, and algorithmic transparency for the standard ML algorithms

Algorithm	Interpretable model			Post-modeling process
	Simulatability	Decomposability	Algorithmic transparency	
Linear/Logistic regression	Humans can interpret the predictor variables, and the interaction among the variables is simple	Humans can interpret the variables. However, the variables and their interactions were broadened to facilitate decomposition	The number of variables and interactions increase significantly to the level that it is exceedingly complicated to be interpreted without a special method	Not required
Decision trees	The results of a decision tree algorithm can be comprehended and reproduced by humans without any mathematical background	The algorithm includes a set of basic rules that are understandable to humans without any data revision	Interpretable rules that describe the knowledge obtained by interpreting the data and provides comprehensible explanations about the prediction phase	Not required
K-nearest neighbors	Humans can simulate the model complexity (the number of variables, interpretability, and similarity function)	Although the model includes a huge number of variables or uses complicated similarity functions that are hard to be simulated totally, they still can be independently interpreted and decomposed	A huge number of variables and the complex similarity function that cannot be decomposed without appropriate statistical and mathematical tools	Not required
Rule-based learners	The number of variables included in the rules is small and simple, and the size of the ruleset is manageable by humans without additional tools	The ruleset size is huge to be comprehended without dissecting it into smaller rule chunks	The ruleset gets too complicated that additional tools are required to interpret the model	Not required
Naïve Bayes	The statistical relationships that are modeled between the variables is understandable by the target audience directly	Statistical variables comprise of many variables so they should be decomposed to facilitate the model interpretation	Statistical relationships cannot be interpreted even they were already decomposed, and individual variables are too complicated that the model can only be interpreted using mathematical tools	Not required

Table 7 (continued)

Algorithm	Interpretable model			Post-modeling process
	Simulatability	Decomposability	Algorithmic transparency	
Generalized linear model	The variables, interaction among variables, and smooth functions associated with the model can be comprehended by humans	The interactions get too complicated to be imitated, so decomposition methods are implemented to explain the model	The complexity of variables and their interactions increase significantly to the level that they cannot be interpreted without the statistical tools	Not required
Ensembles of decision trees	×	×	×	Required: the *feature relevance* or *simplification* approach
Support vector machines	×	×	×	Required: the *local explanation* or *simplification* approach
Artificial neuron network	×	×	×	Required: the *feature relevance* or *visualization*

a generic model. The network checked whether the independent variables of the model were adequate to describe the entire network topology. Moreover, a variational Bayes network was applied to calculate the residual graph function by averaging the block-wise constant sequences. After comparing with the other eight networks from ecology and social sciences, the proposed network showed that it could be applied to various applications, because the control variables were usually provided when the binary networks were investigated. Ahn and Lin (2019) introduced a visual analytics software for the interactive analysis of logistic regression models, which supported the researchers in efficiently creating, analyzing, and comparing various models using the initial model development workflow. In addition, the tool sufficiently revealed general patterns from the candidate models? parameters. With a similar idea, Dingen et al. (2018) proposed a visual analytic software to allow the users to explore logistic regression models interactively. The tool facilitated a quick generation, evaluation, and comparison of several models that are based on the model development workflow as a starting point. Global patterns in the parameter values of the models under consideration could be investigated adequately in order to make new theories or improve the model.

4.1.2 Decision trees

Decision tree learning, which is a commonly used hierarchical structure-based ML algorithm for the regression and the classification topics (Eiras-Franco et al. 2019; Gupta et al. 2017; Wang et al. 2019d), can be interpreted easily and fulfill all the constraints of the inherently transparent model. It has consistently stayed between various groups of transparent models. In the simplest form, the decision tree is a simulatable model. However, its properties can be revised to become a decomposable model and an algorithmic transparent model. The algorithm complexity and understandability are assumed to be a critical factor, because the decision tree models were closely related to the decision-making process. The evidence for the mentioned information can be witnessed through the increasing number of research on the generation and explanation of the decision tree (Sagi and Rokach 2020; Främling 2020; Zhang et al. 2019). Even though the decision tree algorithm fits into the three degrees of the transparent model, the decision tree?s attributes can drive it toward the algorithmic transparent class. For example, a single decision tree is the decision tree that can be simulated and comprehended by humans comfortably, because it is small and contains a limited number of features. An increase in the model size converts it into a decomposable model, because its size interferes with the humans? full simulation. Finally, a significant increase in the size and the number of complicated features turns the model into an algorithmic transparent model that loses the early explainable features. For example, Eiras-Franco et al. (2019) proposed a novel decision tree-based model that contained a binary decision tree and a clustering algorithm. The proposed model was scalable and interpretable, which was designed mainly to acquire the global explanation of important information from the dyadic dataset. The preliminary outcomes indicated that the suggested model obtained good performance and achieved the explainability ability. In another research, Nguyen et al. (2020a) demonstrated two new multivariate decision tree (MDT) methods, which included an exact-convertible decision tree (EC-DT) and an Extended C-Net algorithm to extract important rules from an artificial neural network (ANN) model efficiently. The experimental results suggested that the extracted rules contained multiple attributes that supported a precise interpretation of the decision-making processes.

The decision tree is a conventional ML algorithm that was implemented extensively in the decision support contexts. Several of its applications were different from the computation and AI fields, which indicated that the researchers from other areas felt that the algorithm was straightforward to interpret and comprehend (Eiras-Franco et al. 2019). However, the decision tree is less likely to be applied to applications where the model?s performance is the foremost requirement, because the simple architecture leads to bad generalization ability compared to other complicated models. As a result, the ensemble learning approach was recently proposed to overcome to partly solve the poor performance by gathering the outputs predicted by several trees in order to learn from various training subsets of the original data (Park et al. 2018; Sagi and Rokach 2020). The combination of several trees significantly increased the complexity and made the algorithm lost the transparent characteristics. Therefore, a set of post-modeling explainability methods is applied, which will be discussed in the following section of the review.

4.1.3 K-Nearest neighbors (k-NN)

k-NN is a typical approach that belongs to the group of transparent models. Without the training process, it solves the classification problem by directly predicting a label for an input sample using its k-NN? votes, which measures the distance. The voting mechanism can be replaced with an accumulation of the nearest neighbors? target values to deal with the regression problem. The k-NN model can be customized for the specific problem that is being researched.

Regarding the model interpretability, it is crucial to state that the k-NN model?s output mainly depends on computing the similarity and the distance between samples. The mechanism of the k-NN resembles the experience-based decision-making of humans, which determines the output based on past cases. The user?s interaction with the model is straightforward due to the model?s interpretable nature, which allows the researchers to discover why a new input was put into a specific group and how the output is updated when K was changed. As a result, the k-NN has been extensively considered for applications that required the model interpretability (Zheng and Ding 2020). As noted earlier, the k-NN interpretability relies heavily on three main factors, which include the number of features, the total number of instances, and the distance metric applied to calculate the correlation among data samples. For example, a high K value hinders the complete simulation of the model. Likewise, a large number of features or a complicated distance function impedes the model?s decomposability and restricts its interpretability to algorithmic transparency. Wang et al. (2019a) demonstrated a novel k-NN rough set model that incorporated the strong points of both δ-neighborhood and k-NN, which dealt with heterogeneous data more effectively compared to the existing models. An iterative process was used to depict a decision through rough approximations and describe its monotonic. Moreover, an attribute reduction method was introduced to offer higher performance than the previous approaches, especially for the δ-neighborhood rough set and k-NN rough set approaches. Zheng and Ding (2020) suggested a novel classifier motivated by the original k-NN algorithm to increase the classification accuracy and the model explainability. The classifier was robust, because a sparse group lasso was applied to the group level in order to choose K most related groups and eliminate all the irrelevant groups rather than the sample level. Moreover, K-SVD, which was a dictionary learning algorithm, was implemented to precisely extract wanted sparsity (nonzero entries) to overcome the hyperparameter optimization challenge. Eventually, the possibility of a sample that belonged to a particular group was expressed clearly by the summary of the regression weights of each class in compliance with the XAI for the proposed model. The experimental results showed that the proposed algorithm outperformed eight other algorithms in terms of classification accuracy.

4.1.4 Rule-based machine learning

Rule-based ML describes a group of models that generate rules to represent the data being learned. The rules can be basic and straightforward conditional, such as if-then rules, or a combination of several rules to represent the knowledge. The well-known fuzzy rule-based approaches also belong to the rule-based ML models, which were proposed for a wider range of objects and enabled the concept of verbally formed rules across vague domains. The rule-based ML algorithms belong to the interpretable model because they use fuzzy logic and fuzzy sets to express various forms of knowledge and model the existing relationships and interactions between the variables. The rule-based ML algorithms were proven to perform better than the traditional rule-based systems when some levels of uncertainty exist. The fuzzy rule-based algorithm is considered transparent models, because it solved the explainability problem by creating rules to justify their outputs (Adriana da Costa et al. 2013; Yeganejou et al. 2019). For instance, Fernandez et al. (2019) suggested the ?4 W? questions with the main goal to show how the evolutionary fuzzy systems were crucial from an explainable point of view.

Rule-based learning methods have been utilized widely to perform knowledge extraction in many application domains (Bologna 2019; Keneni et al. 2019; Singh et al. 2019). The primary design objective of a rule-based algorithm is interpretable and straightforward. However, the model interpretability is affected by the length and the number of generated rules. A high number of rules improves the model?s performance but with a risk of neglecting its interpretability. Likewise, the length of the rules also works against interpretability, because it becomes hard to interpret when it gets longer. Using the same deductive reasoning, these two features collaborate with the transparent model?s classes, which is shown in Sect. 4.1. Lengthy rules or a huge number of rules turn a model into an algorithmically transparent model. One solution is to convert the basic rules into fuzzy rules to ease the rule size limitations, because a larger rule size can be implemented with smaller tension on interpretability.

The resemblance of the rule-based learners to natural human behavior makes them an excellent choice in order to understand and interpret other models. When a specific value for the number of rules is decided, a rule wrapper can be applied to hold sufficient data of a model in order to describe its operations to the regular users without yielding to the likelihood of utilizing the created rules as a standalone prediction model. Hatzilygeroudis and Prentzas (2015) demonstrated neurules, which are a type of neuro-symbolic rules that combine the symbolic rule with neurocomputing. A neurule-base model contains several autonomous neural units with a symbolically oriented syntax. Two reasoning phases of the neurules are connectionism, which focuses on the neurocomputing approach, and symbolism, which concentrates on the symbolic backward chaining-like method. The experimental results showed that the symbolism approach was more productive than the connectionism approach regarding computation complexity and speed, even though both demanded an equal number of input variables. In addition, the neurule-based explainable system significantly improved the efficiency and comprehensibility of the explanations compared to the existing rule-based expert systems. In another work, Keneni et al. (2019) deployed an explainable steering control framework for unmanned aerial vehicles (UAV) by implementing a rule-based Sugeno fuzzy inference model. The data collection process was performed by directing the UAV to fly along with the assigned task and recorded a series of actions when it ran into predefined weather and enemy patterns. The collected data was then applied to train a Sugeno-type fuzzy inference model using the subtractive clustering algorithm on the data. The subtractive clustering parameters were optimized by accessing the model?s performance and the number of rules. The model was fine-tuned with an adaptive neuro-fuzzy logic model (ANFIS) to provide explainable features of the decisions that were made by the UAV. The experimental results showed that ANFIS was effective in enabling the XAI feature for the framework. The output model involves six rules with a root mean square deviation less than 0.05.

4.2 Hybrid interpretable model

The hybrid interpretable model approach assumes that it is feasible to combine an inherently interpretable model with a black-box model in order to form a hybrid model that achieves both high performance and model interpretability instead of implementing a single black-box model, which is challenging to interpret. Wang and Yeung (2016) proposed a novel Bayesian deep learning (BDL) framework to obtain combined intelligence that supported the inference of the output. The BDL estimates the uncertainty of the black-box deep learning models by either putting distributions over the model weights or by seeking straight mapping to the probabilistic outputs. The importance of each feature was evaluated efficiently through the weight distributions of the outputs and classes using the proposed model. Recently, Yeganejou et al. (2019) introduced a hybrid convolutional fuzzy classifier to perform model interpretability. The idea was based on the fact that even though CNN models have obtained state-of-the-art performances on various application domains, it was impossible to persuade the users to trust them because they were black-box models. In contrast, the fuzzy system was much easier to interpret due to its simple architecture. As a result, the authors used CNN as a feature extractor, and then implemented a fuzzy clustering to cluster the extracted features. An explanation mechanism for the proposed model was then created by identifying each cluster?s medoid and assessing the significance of each pixel in the input data. The preliminary results on the three benchmark datasets revealed that the CNN feature extractor considerably improved the fuzzy classifier?s performance and interpretation ability. With the same intention, Gulati et al. (2021) demonstrated a hybrid interpretable solution to identify the 17 types of gestures in a user-specific approach. A GradCAM method was also developed to optimize and the generalized model architecture to explain its predictions. Through analyzing the GradCAM results, the last CNN layer was removed, because it demonstrated minimum contributions towards the prediction. Therefore, the approach preserved the accuracy, sensitivity, specificity with a shallower structure, which the trainable parameters were reduced by 20%. Recently, a hybrid model that combines two prior algorithms, which include the TREPAN decision tree and the clustering of a hidden layer representation, was proposed to deconstruct a deep learning network (De et al. 2020). The proposed model aimed to visualize the information flow of an underlying model to make it comprehensible to humans. The experimental results revealed that the hybrid model provided brief human-interpretable evidence for the framework predictions. Another example is the contextual explanation networks (CEN) framework (Al-Shedivat et al. 2020), which generated parameters for the intermediate interpretable models to make predictions and produce explanations. The preliminary experiments on the image classification and NLP topics demonstrated that the CENs performance was comparable to the well-known models and gave additional explanations behind each prediction. Chen et al. (2020a) introduced adaptive explainable neural networks (AxNN)?a novel ML framework that supported two primary goals of model accuracy and explainability. The model included ensembles of additive index models and generalized additive model networks. After that, the outputs of AxNN were separated into high-order interactions to perform interpretation.

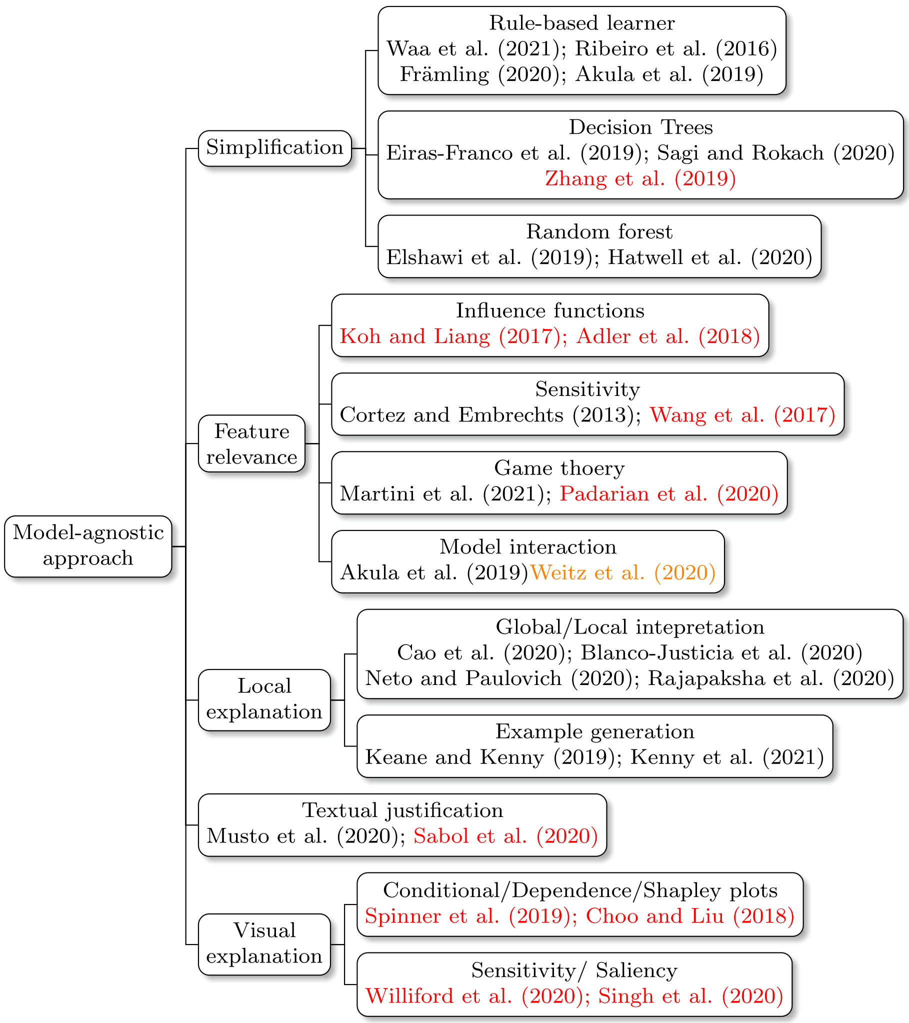


Fig. 6 Categorization of the XAI trends that are based on the previous notable research associated with various ML algorithms. The XAI research trends are classified by analyzing the previous related studies in depth in order to determine if a post-modeling explainability can be effortlessly implemented for a specific ML algorithm. The boxes in black, red, and orange refer to the XAI approaches on the text, image, or audio data. (Color figure online)

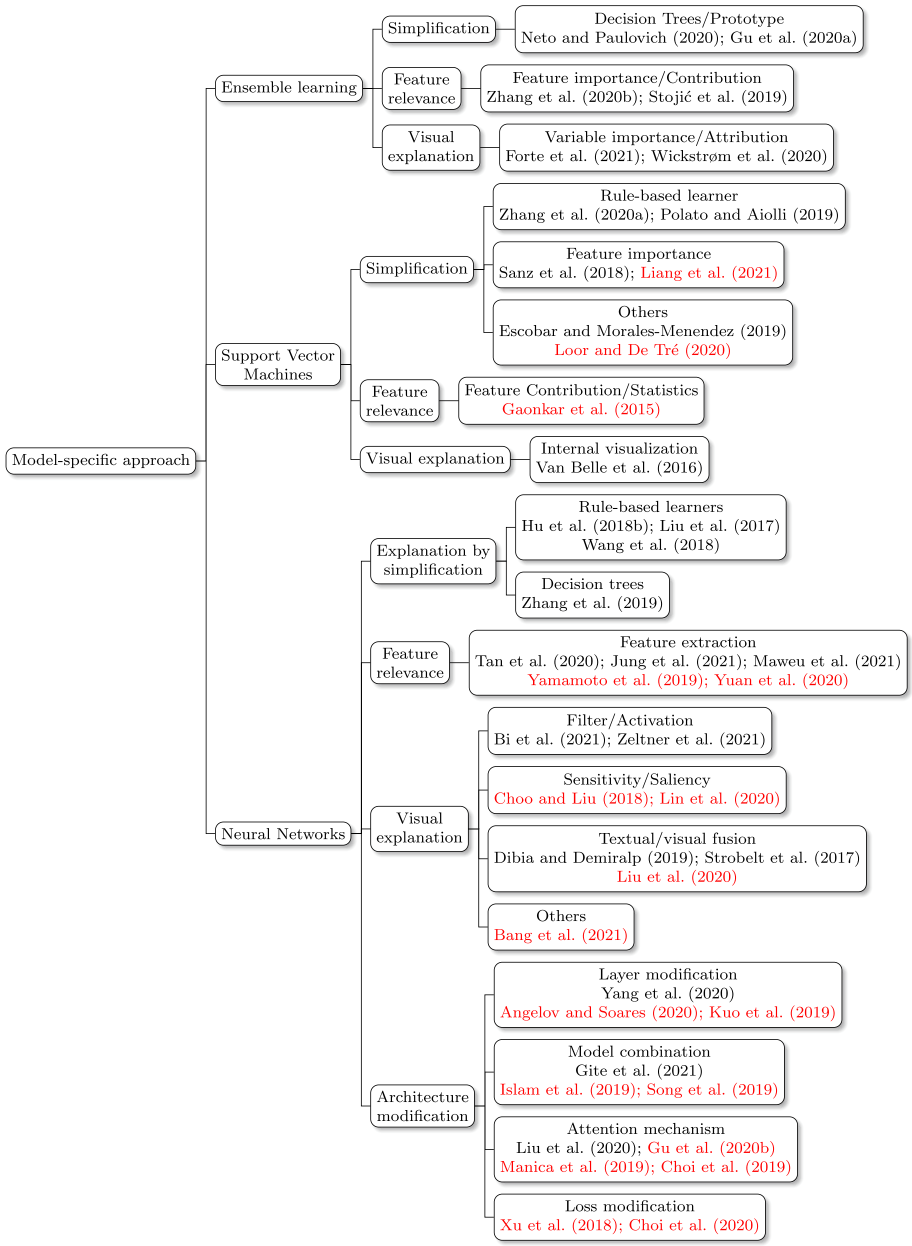


Fig. 6 (continued)

The main advantage of the hybrid interpretable approach is that it offers robustness and interpretability to the black-box models (Yeganejou et al. 2019; Gulati et al. 2021). Other approaches have shown that the hybrid interpretable models simultaneously learn and provide explanations with both symbolic descriptions, sub-symbolic descriptions, and inferences (Al-Shedivat et al. 2020).

5 Post-modeling explainability

Suppose that the ML algorithms did not satisfy any standards to consider them an interpretable model. A group of approaches referred to as post-modeling explainability can be proposed to enable their explainability. The methods that belong to the post-modeling explainability aim to explain how an algorithm performs during the training process or how it generates predictions for any provided input. In this section, various algorithmic techniques for post-modeling explainability are first classified and evaluated, which include (1) the model-agnostic approaches that were devised to be implemented on any ML algorithms, and (2) the model-specific algorithms that were proposed for a particular ML algorithm which are difficult to be implemented on other learners or achieve a low performance. Next, the post-modeling explainability research trends for distinct ML branches are evaluated using a hierarchical graph, as demonstrated in Fig. 6. For each approach, a thorough survey of the most recent post-modeling explainability methods introduced by the research community and the identifying trends that arose with these types of contributions. A summary of Fig. 6 is described below.

- The model-agnostic approach contains all research that concentrates on implementing the post-modeling explainability for lightweight ML algorithms except for the family of deep learning models, which is described in Sect. 5.1.
- The model-specific approach aims at addressing the explainability and the interpretability for deep learning, such as CNN, RNN, and hybrid models that combine the interpretable models and the deep neural networks, which is discussed in Sect. 5.2.

5.1 Model-agnostic approach for the post-modeling explainability

This section discusses any method that applies the ML model to obtain some crucial knowledge from the model?s training and prediction scheme or visualized by a set of specific techniques to facilitate model explainability. In addition, proxy models that mimic the original models can be created to allow explainability and reduce complexity. The model-agnostic approach involves standard techniques, such as *textual justification*, *simplification*, *feature relevance*, and *visualization* methods, which is displayed in Fig. 6.

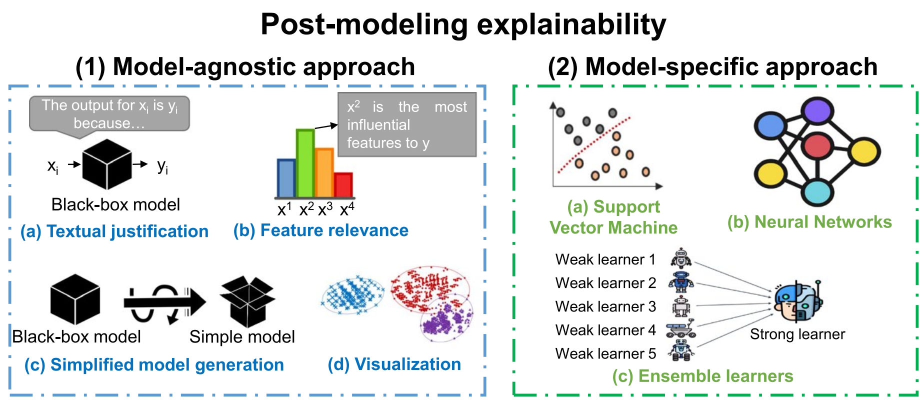


Fig. 7 Conceptual diagram that depicts two categories of the post-modeling explainability for the ML models, which include (1) the model-agnostic and (2) the model-specific groups

5.1.1 Textual justification

Textual justification offers explainability for a model by generating a text explanation in the form of phrases or sentences using the natural language generation (NLG) methods in order to explain the model predictions directly to the general users and the experts (Shi et al. 2018). Moreover, it also refers to the methods that create symbols to portray the algorithm?s function logic through semantic mapping. An example of the *textual justification* is shown in Fig. 7(1).

One primary approach to *textual justification* is to train a model using visual features extracted from another classifier in order to generate text explanations. For example, Sabol et al. (2020) introduced a textual justification model that helped categorize eight types of tissue from a histopathological test. The model was competitive with the latest RNN models and provided human-friendly explanations about the credibility of a prediction made by the classifier. Moreover, the experts can use it as a diagnostic tool in the medical domain with high confidence due to the textual explanation ability. Another major approach is to train an XAI model to justify the outputs of a black-box model. Musto et al. (2020) proposed a post-modeling textual justifications framework, which interpreted the suggestions created by a decision support system. The framework accepted a suggestion and a collection of reviews as the input and generated textual justification, independent of the primary model. The experimental results showed that the review-based justifications could rely on simple features-based explanations. Moreover, the text summarization method led to more pleasing justifications. The proposed framework showed that it made the recommendation process clearer, more appealing, and brought more trust to the users, which proved the approach?s potential.

5.1.2 Visualization

A typical approach to explaining an ML model, especially complex black-box models, is the *visualization*, which analyzes how a model learns the hidden patterns during the training process or how a prediction is made during the testing process. Therefore, many studies have followed this trend by highlighting decision-relevant parts of machine representations in order to explain the black-box models. For example, the parts that contributed to model accuracy during training or to a particular prediction.

The latest research on this topic can be observed in several studies. For example, Ahn and Lin (2019) applied a ranking mechanism in order to achieve model-agnostic post-modeling explainability for various prediction approaches, which included binary and multi-class classification. The proposed framework?s main advantage is that it does not rely on any specific algorithm and offers the users a fair evaluation of each learning phase from the input data to the prediction. After that, a visual analytic tool, which is FairSight, with various visualization functions, was introduced based on the framework to support the real-world deployment of fair decision-making. The experimental results revealed that the suggested framework provided more model explainability over the existing tools. Moreover, it was proven to effectively estimate and minimize the bias on the benchmark datasets. Another means of a model-agnostic post-modeling approach is presented by Spinner et al. (2019). This approach involves the use of explAIner, a visualization tool introduced to support interactive and explainable ML algorithms. The tool supports various processes, such as provenance tracking, model comparison, quality monitoring, interactive investigation of the graph, on-demand collection, visual analytics of the evaluation metrics, TensorBoard debugging environment, and the latest visualization methods, such as layer-wise relevance propagation (LRP) (Lapuschkin et al. 2016), and local interpretable model-agnostic explanation (LIME) (Ribeiro et al. 2016). Zhang et al. (2018a) suggested Manifold, which is a model-agnostic framework for ML model visualization that supports debugging, interpretation, and comparison more interactively and transparently. The framework focused on a different aspect of the model by relying entirely on observing the input, which was the feature importance, and the output, which was the predicted results. The visual modules were comprised of a scatterplot-based visual representation, which analyzed the system?s prediction, and a custom tabular view, which was the feature discrimination.

The *visualization* approach is less likely to be applied to the model-agnostic approach than the model-specific approach, because it has to inspect the model structures and create many kinds of visualizations from the inputs and outputs. The *visualization* method can also be combined with other methods in order to enhance their explainability. Therefore, most *visualization* techniques in this section partly or fully involve the implementation of the *feature relevance* method in order to extract crucial features that are ultimately presented to the end-users.

5.1.3 Simplification

The *simplification* approach is probably the broadest category of the model-agnostic post-modeling that builds a new and simple system from the complex ML model. The created simple model regularly seeks to optimize its similarity to the original function while minimizing the complexity and achieving comparable performance to the original model. There are two methods that can be considered subsets of the *simplification* approach, which include the *local explanation* (Neto and Paulovich 2020; Karlsson et al. 2020) and the *example generation* (Chen et al. 2021). The *local explanation* assumes that simplification can be achieved by segmenting the solution space into smaller segments and performing an analysis on the particular segments of a model. After that, the method explains the less

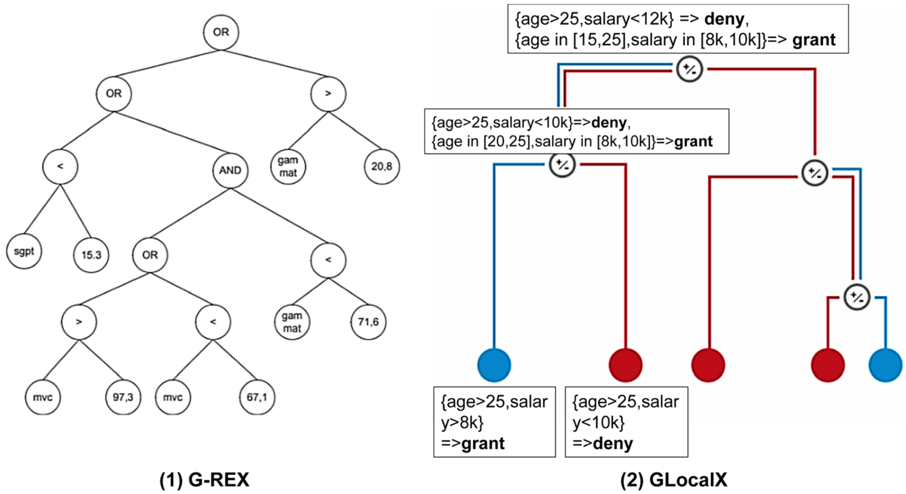


Fig. 8 Examples of the model-agnostic simplification approach based on rule-based extraction (1) G-REX framework (Konig et al. 2008) and (2) GLocalX framework (Setzu et al. 2021)

complicated solution subspaces that are associated with the original model. These explanations can be created by separating the property methods that only describe a portion of the model?s operation. On the other hand, the *example generation* method involves the process of showing the data samples that are related to the predictions of a specific model, which allows the users to obtain comprehensive explanations about the model itself. Those methods primarily concentrate on obtaining the representative samples that cover the model?s internal correlations, which simulate human behavior while explaining a given problem. Most of the *simplification* methods are based on the rule extraction techniques, which are explained in Fig. 8.

The *simplification* approach is a well-known XAI approach, because it has commonly appeared in the latest XAI studies, such as G-REX (Konig et al. 2008) and LIME (Ribeiro et al. 2016). LIME produces locally linear models to analyze and interpret the outputs of a complex algorithm. Therefore, it involves both the *simplification* and the *local explanation* approaches. Setzu et al. (2021) demonstrated a novel GLocalX recently as an alternative to the rule extraction, which applies the logic rules to black-box ML models. GLocalX hierarchically aggregates the local explanations extracted from the local decision rules to create global explanations. The experimental results showed that GLocalX correctly created the simplified versions, which obtained an even higher performance than the original models.

The main benefit of this technique is that the implementation of the simplified model is generally more straightforward due to the decreased complexity compared to the original model. As a result, this approach is forecast to continue playing crucial roles in XAI research.

5.1.4 Feature relevance

The *feature relevance* is considered an indirect approach to perform the post-modeling explainability, which evaluates the algorithm?s internal processes by calculating the relevance score for all the variables that it manages. The computed score quantifies the importance or the sensitivity, which reveals what features are crucial, which is what the model depends on when making its prediction.

One well-known research regarding this topic was introduced by ?trumbelj and Kononenko (2014), which is called SHapley Additive exPlanations (SHAPE). SHAPE proposes a novel way to compute the important scores for each specific output using a collection of valuable attributes, such as the missing items, consistency, and local accuracy that is missing in the original model. Chen et al. (2018) suggested two extensions of the SHAPE framework, which include L-Shapley and the C-Shapley algorithm, to score instance-wise feature importance in order to perform the model interpretation. After that, the degree that the features affect the output is well-explained using a graph-structured factorization. With a different method, Henelius et al. (2014) analyzed the connections and the dependencies between the variables of the model by merging the features to deliver the model explainability. Molnar et al. (2019) suggested the analysis of the ML model complexity based on the relevance of the features like the number of features, feature interactions, and complexity. The researchers can effectively select and compare different types of models using the proposed system. Moreover, the post-modeling explainability was proved to be improved significantly by investigating and minimizing these measures. Most recently, Moradi and Samwald (2021) presented a post-modeling explanation framework called confident itemsets explanation (CIE), which extracted a collection of features that highly affected the model in order to predict a specific class label. CIE provided class-wise and instance-wise explanations that precisely showed how the black-box works. The CIE framework outperformed the former rule-based explanation through several experiments regarding the explanations? interpretability and descriptive accuracy. Krishnamurthy et al. (2021) offered a so-called sentences in feature subsets (SiFS) in order to implement a genetic-based algorithm to choose a subset of distinctive features that affect the prediction process. The obtained features were then converted into compact decision rules and held Boolean logic sentences that humans can easily comprehend.

The *feature relevance* is also considered a well-known model-agnostic post-modeling explainability approach that is similar to the simplification approach because a large number of publications related to this topic were found between 2018 and 2021. The *feature relevance* method has also become a vibrant XAI research topic in recent years.

5.2 Model-specific post-modeling explainability

Given that several complex ML models have achieved notable performance in predictive applications and particularly the dominance of deep learning, this section focuses on adopting the model-specific post-modeling explainability approach in order to interpret these models. It includes a set of techniques proposed to enhance the explainability of a specific black-box ML model. After the analysis of recent literature related to the model-specific post-modeling explainability, various trending topics emerged. Firstly, the rule-based extraction method predominates this topic, which is intuitively anticipated, because it is challenging to enable the explainability for the complicated ML models themselves. Many studies that followed the rule-based learning approach for model explainability will be discussed in the Sect. 5.1.3. *Feature relevance*, which can also be implemented to perform the model-specific post-modeling explainability, has attracted a lot of attention from researchers recently. This approach can be used to create hybrid models that are independent of the model being explained. Finally, *textual justification* and *visualization* methods propose engaging means that affect important features that are detected by the *feature relevance* approach in order to facilitate the post-modeling explainability tasks. The application of the *textual justification* and the *visualization* on other ML model characteristics, such as the structure, firmly rely on the specific model that is being analyzed.

5.2.1 Ensemble-based models

Ensemble learning has been primarily implemented to improve the model?s performance for specific tasks, and it is arguably a reliable way to improve a model?s performance (Deng 2019). It merges the predictions of different models/trees in order to achieve an aggregated output. The most crucial benefit of ensemble learning is the generalization improvement as a single model/tree, which regularly faces the overfitting issue. Even though it is highly effective against overfitting, the aggregation of several models makes it more complex to interpret compared to individual compounding models. Therefore, the post-modeling explainability techniques need to be implemented. The standard methods found in this topic are *simplification* and *feature relevance*.

First of all, many ensemble-based studies that concentrated on the post-modeling *simplification* approach in order to partially explain the ensemble-based model and reserve the model performance have been proposed recently. For example, Hara and Hayashi (2018) solved the simplification of the tree ensemble algorithms by making a simplified model that is similar to the original model while preserving the original model?s performance.

They implemented a Bayes factor algorithm that automatically determined the model complexity. Deng (2019) proposed another method for the post-modeling simplification by offering an interpretable tree framework (inTrees) that performs extraction, measurement, pruning, selection, and summarization of the rules. The proposed framework was used in various ensemble-based algorithms, which include boosted trees, random forests, and regularized random forests. Likewise, Obregon et al. (2019) offered an innovative system in order to merge and simplify the ensemble trees? outputs into a single output ruleset. Each decision tree?s weight was put into a matrix in order to generate a unique set of reduced rules that brought a comparable accuracy to the rules of the initial ensemble model. Hatwell et al. (2020) presented a novel Collection of High Importance Random Path Snippets algorithm (CHIRPS) to interpret the random forest classification for each data instance and bring comparable performance to the state-of-the-art algorithm with a higher coverage rate. In addition to the *simplification* approach, the *feature relevance* method has been increasingly applied to the tree ensembles. For example, the generic framework proposed by Elghazel and Aussem (2015) was the first to evaluate the out-of-bag error and the feature importance from the individual learners in the ensemble for the unsupervised random forest algorithm. Each partition was created using the random subset features and a distinct bootstrap sample.

The *feature relevance* and the *simplification* approaches appear to be the most common strategies for the ensemble models, which are similar to the trend of the model-agnostic approach. However, many studies, which date back to 2018, focused primarily on the bagging technique, so a limited amount of research has been observed lately when the research interest was shifted to other ensemble techniques, such as stacking and boosting. Among the studies, it is noticeable that a group of authors tried to explain why an individual classifier from the ensemble classifier produced a particular output for a given input. The stacking with interpretable rule approach introduced by Wang et al. (2019d) focused on constructing the model and extracting the interpretable rules. It was based on a multi-objective optimization method in order to enhance the model performance while allowing it to be interpretable. Other fascinating studies about interpreting the ensemble models are DeepSHAP (Chen et al. 2021), which stacked the ensembles and various classification models in order to explain the neural networks. Ribeiro et al. (2019) applied a dimensionality reduction method to facilitate the visualization framework to explain the ensemble models and explained how an individual classifier contributes to the final output. Most recently, Konstantinov and Utkin (2021) implemented an ensemble of gradient boosting machines (GBMs), where each GBM was determined by a single feature and produced a shape function of the feature. The model was built parallel using randomized decision trees of depth 1, which provided a simple architecture. Experiments on the synthetic and real datasets demonstrated that the model was efficient for local and global interpretation.

5.2.2 Support vector machine

Another well-known shallow ML algorithm that was dominant before the deep learning appearance is the support-vector machines (SVM). The SVM algorithm has a more complicated structure compared to the ensemble approach. The standard SVM model is based on a supervised discriminative classifier that investigates data for the classification/regression task using the hyperplane separation mechanism, which obtains high performance and has robust generalization ability. The algorithm tries to find an optimal hyperplane to classify the new samples by learning the discriminative features from the labeled training dataset. For example, the hyperplane is a 2d line in a two-dimensional space, separating the data plane into two regions, where each region belongs to one class. Many post-modeling explainability approaches have been previously proposed to describe SVM?s internal processes, including *simplification*, *local explanation*, and *visualization*.

The four trends of the *simplification* group can be derived from the existing simplification approaches that are based on how deep they operate on the algorithm?s internal process. The first trend is the rule-based post-explainability, where a group of researchers constructed them using the trained SVM models. For example, Polato and Aiolli (2019) introduced a method to obtain the explanation rules from the trained SVM model using boolean kernels with the feature spaces makeup of logical statements. Moreover, a searching strategy was implemented to extract the most important features/rules that efficiently explained the trained model. Adriana da Costa et al. (2013) proposed FREx_SVM, which is a specific framework that extracts the fuzzy rules from a trained SVM for multi-class classification. It contains three simple steps, which lead to a fast and efficient rule-extraction process and enable a more linguistically explainable prediction. Haasdonk (2005) started the second simplification approach, which concentrated on simplifying the SVM kernel by providing arbitrary symmetric kernels. The optimal hyperplane was obtained by minimizing the distance between the convex hulls. The interpretation was generated in order to explain the behavior of the SVM. In the third trend to simplify the SVM, a group of authors suggested adding an SVM?s hyperplane to the XGBoost module, which was implemented to create the rules (Singh et al. 2019). The final category refers to a group of research that implements an add-on model in order to enable the interpretation of the SVM algorithm. For example, Zhou et al. (2018b) applied a genetic algorithm (GA) to concurrently optimized the variables of the SVM classifier and the input features. The GA model improved the accuracy of the SVM and allowed the visualization of the important features that affected the model?s output.

Apart from the *simplification* using rule extraction, other novel approaches that contributed to the explanation of the SVMs were also reported. The *visualization* approach is an effective and user-friendly method to interpret the SVM models that are deployed for real-world applications. For instance, Ma et al. (2017) presented an open-box visual analysis tool, which was called EasySVM, in order to support the construction of the SVM models. The tool allowed the users to analyze the training data and the relations between the input data and the model. Moreover, it also supported the visual rule extraction method from the SVM?s predictions. Sanz et al. (2018) proposed to visualize the most relevant features and the relationship between the predictors and the response variables through non-linear kernels. The proposed method achieved higher performance than the standard RFE algorithm. Gaonkar et al. (2015) presented an intuitive and straightforward tool to conduct multivariate analyses and interpret the SVM model. The research focused on analyzing what piece of the data significantly affected the SVM decision by providing a statistical p-value based answer. The experimental results showed that the accompanied *p-value* was an asymptotic distribution. Furthermore, this statistic was proven to show a better performance than the weight-based randomization test, and it was quite sufficient to reveal the multivariate patterns of the input neuroimaging.

A notable remark between the post-modeling model-agnostic techniques and the SVM technique is that the *simplification* approach in a general view dominated both the post-modeling model-agnostic techniques and the SVM algorithm. However, the *visualization* and the *local explanation* methods have been considered more often in recent years, because the simplification approach was, on average, proposed a long time ago compared to the other approaches. In addition, most of the studies related to the post-modeling SVM explainability were published before 2018, which was due to the growing research interest in deep learning in all disciplines, and due to the fact that the suggested models were already interpretable, so it was challenging to enhance further the explainability based on what has already been performed.

5.2.3 Deep learning model

Post-modeling *feature relevance*, *local explanation*, *textual justification*, and *visualization methods* are becoming the standard means to interpret the deep learning models. This section discusses the XAI research that is aimed at enabling the post-modeling explainability for standard deep learning models, which include feed-forward neural networks, convolutional neural networks (CNN), and recurrent neural networks (RNN).

- Feed-forward neural networks

  Since its first appearance (Wang 2003), ANN has been accepted by the research community and has become the foundation of various deep learning models (Ostad-Ali-Askari et al. 2017; Ostad-Ali-Askari and Shayannejad 2021). ANN contains three main components, which include the input layer that receives the training samples, the hidden layer that automatically learns from the data, and the output layer that generates the outputs. ANN automatically learns the complex relationships between the variables to cope with the issues that would be impossible or difficult to solve by statistical or human standards and outperforms conventional algorithms by a wide margin. However, even the researchers who research the ANN-based models or the developers who deploy them in real-life applications find it hard to explain how the models work. Therefore, the family of deep learning models has always been regarded as the black-box models and has led to reluctance in the model deployment. The post-modeling explainability and interpretability of these models have become a hot topic in recent years in order to improve practical values. Some of the standard post-modeling approaches are the *simplification*, *feature relevance*, *local explanation*, *textual justification*, and *visualization*.

  Numerous *simplification* approaches have been introduced to the ANN-based models. For example, Craven and Shavlik (2014) proposed to train the ANN networks using the NofM extraction method to perform weight clustering during the training process in order to obtain concise representative rules. The obtained rules were proved to be logically understandable. Li et al. (2015) applied an improved particle swarm optimization method (iPSO) in order to optimize the threshold values and the structure?s weights of the ANN. Moreover, a principal component analysis (PCA) was also implemented in order to simplify the original model and choose the essential inputs. The experimental results showed that the proposed hybrid ANN model achieved higher accuracy and required a shorter modeling time. Although the post-modeling simplification-related studies have proved that they partially supported the model interpretation, they became ineffective when more layers were added. Therefore, the *feature relevance* approach has been considered more often. For instance, Lapuschkin et al. (2016) demonstrated the LRP algorithm that visualizes the ANN?s output using the given input image. It attached the relevance scores to the crucial parts of the image by adopting the learned model?s topology. In another work, Zhang et al. (2018c) proposed the Garson?s algorithm, which represents the related importance of the role of a predictor in its relationship with the output variables by analyzing the model learned features. Similarly, Shrikumar et al. (2017) introduced DeepLIFT, which was a framework to calculate the feature relevance scores in a feed-forward neural network. It analyzed a neuron?s activation to the reference activation and used the computed difference to distribute the score.

- Convolutional Neural Networks

  In the last decade, the introduction of a huge publicly available dataset, namely the ImageNet dataset that contains millions of carefully labeled data, and the availability of the computing resources (GPU) has motivated the creation of various state-of-the-art CNN models on a wide range of CV tasks, such as image classification, object detection, and object segmentation. Generally, the CNN contains a series of layers, which includes the convolutional layer, the pooling layer, and the fully connected layer. Each layer contains many neurons in order to automatically learn the increasingly high-level features when the network?s depth increases. At the end of the CNN?s structure, one or a series of the fully connected layers is placed in order to reflect the extracted feature maps to output. One important property of the CNN is *nonlinearity*, which is achieved by applying the nonlinear mapping using the activation function (Zeltner et al. 2021). An activation function is a decision-making function that decides the presence of a particular neural feature, which is mapped between 0 and 1, where zero means the absence of the feature, while one means its presence. The selection of activation function has a significant impact on the capability and performance of the neural network, and multiple activation functions can be used in different parts of the CNN model. The mentioned CNN architecture involves overwhelmingly complicated inner processes that are challenging to be interpreted by humans. As a result, XAI is required in order to improve the model?s trustworthiness. Even though CNN has a complex structure, its explainability ability is actually more straightforward than other algorithms, because human cognition promotes the perception of the data visualization. Previous studies that focused on the CNN explainability can be classified into three primary groups, which include (1) the research that interprets the prediction process of the CNN by the simplification process, (2) the studies that examine the model structure and intermediate layers in order to see how it learned from the input data in general, and (3) the textual justification and visualization techniques.

  Some examples of the *simplification* XAI-based explanation are a rule generation method that was suggested by Bologna (2019) in order to interpret a CNN model that was global to all the input data. The rule extraction module was implemented on the fully connected layer with a transparent Discretized Interpretable Multi-Layer Perceptron (DIMLP) subnetwork that could be mapped back into the input layer in order to locate axis-parallel hyperplanes accurately. In another work, Montavon et al. (2017) introduced a novel framework in order to interpret the nonlinear classification decisions using the propagation rules on the input variables that were extracted from deep Taylor decomposition. The proposed method could be applied to various input data, applications, and models. Recently, Zhang et al. (2019) trained a decision tree in order to simplify the CNN explainability and analyzed the outputs at the semantic level. The decision model decomposed the feature representations from the higher layers of the model into the fundamental information of the object regions and hierarchically represented all possible model outputs in a coarse-to-fine style. As explained earlier in Sect. 5.1.3, the *simplification* approach becomes ineffective when more layers are added into the model. As a result, the number of research that follows this path has significantly reduced because the latest CNN models have been getting more complicated.

Numerous research papers have shifted their attention toward customizing or analyzing the model structure in order to enhance the CNN interpretation. In particular, Islam and Lee (2019) introduced a clustering-based interpretation method in order to extract the feedback weights in the CNN models to enable the interpretation of neurons? activities at various layers toward the classification of the testing data. The experimental results revealed that the introduced system efficiently reconstructed the input data from the class probabilities to pick the centroids of the symbolic clusters during the feedback process. In another research, Kuo et al. (2019) suggested an explainable CNN model, which could automatically show the higher-layer descriptions as the object parts instead of the mixed patterns. The model?s loss function was customized to represent the CNN model, which favored the particular parts of the objects from a class and remains still on the images from the other classes. The model required no annotation data for the object parts and encoded more semantic features (knowledge) from the high-layer filters after the training process. In addition, the presented model achieved comparable performance to unexplainable models (binary classification) and even outperformed conventional models for the multi-class classification regarding the prediction performance. Regularization is a standard approach that is usually implemented to increase the performance and the robustness of an AI framework. Moreover, it can also be applied to improve the model explainability. Wu et al. (2018) demonstrated a tree regularization approach that improved the model explainability to allow the users to evaluate the prediction process. The central concept is to promote training a model that performs well while being approximately modeled by a small decision tree, which is interpretable by humans. The idea was implemented when the authors applied a novel regularization method to the model?s loss function. The experimental results on various datasets showed that the trained model was explainable without losing the model performance. Recently, an explainable CNN model developed by Zhang et al. (2018b) that could represent the higher layer activations as the parts of an object, which was contrasted to the normal deep learning models that showed only a collection of abstract patterns. A custom loss function was added to the feature maps of the standard deep learning model in order to push the representation of the particular object parts of the ground truth class and remain still on inputs from other classes. One main advantage of the proposed model is that it did not require any annotation data to represent the object parts. The authors proved that the proposed model encoded semantic patterns about the objects in the high layers compared to standard CNN through several experiments. Moreover, the interpretable model obtained a comparable prediction accuracy to the original complex CNN models for the binary classification and performed slightly better than the standard CNN models for the multi-class classification. Beside the previous approaches, an increasing number of studies have utilized the regularization to explicitly force the explanations of the model outputs to guarantee the model explainability. For instance, Wu et al. (2010) added a tree regularization approach to the loss function to improve the explainability of the CNN models. The primary idea of the approach was to use the decision tree algorithm to find models that have well-approximated decision boundaries. Thus, the model output was simulatable to humans. The experimental results on various tasks show that the model was more explainable without losing the performance. There is an increasing amount of research on customizing regularization to directly constraint explanations of output predictions in order to guarantee they are correct in recent years.

The *textual justification* and the *visualization* are the last category of the CNN explainability. Many studies followed this category, because it is user-oriented, so the end-users can comprehend how a CNN model works. The Network Dissection, which is a novel framework for assessing the interpretability of the hidden features from any CNN model through the alignment between the specific hidden units and a collection of conceptual semantics, was introduced recently by Bau et al. (2017). The experimental results proved that the representations at different layers hold diverse levels of meaning. Jung et al. (2021) proposed a novel relevance-based algorithm, which was called elective layer-wise relevance propagation, to explain the image classification and the text classification model?s output through the visualization. The authors used the activations selectively to calculate each activation?s gradient for the output label, which incorporated the true positive gradients during the computational process. The experimental results revealed that the framework provided the class-discriminative output with less noise and contained the whole objects compared to the existing methods. However, this approach failed when many activations contained negative gradients, even though the test samples were predicted correctly. The DeepClue framework was built in order to give the end-users an insight into the critical features discovered by the stock price prediction model (Shi et al. 2018). The explainability splits the predictable variables from the unpredictable variables using a risk visualization design and a model parameters intercept. The experimental results showed the success of DeepClue in supporting the stock market analysis and investment tasks. However, the model requires more improvement for future work, such as news headlines and contents in the dataset can be used to enhance the model?s performance. Moreover, more types of graphs can be applied to support the interpretability process. A novel approach, which is called deep visual representation, was introduced by Zhou et al. (2018a) to transform the previous qualitative visualization in order to interpret and measure the interpretability of the CNN networks quantitatively. They fed a huge number of images into a deep learning model and analyzed the relation between each hidden unit and visual semantic concepts.

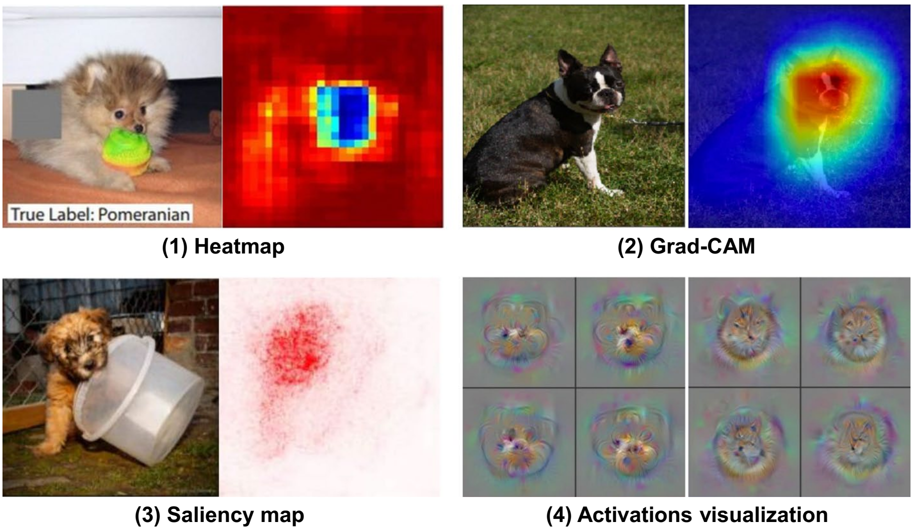


Fig. 9 Examples of various XAI visualization methods on images, which include (1) Heatmap (Payer et al. 2019), (2) Grad-CAM (Selvaraju et al. 2020), (3) Saliency Map (Adebayo et al. 2018), and (4) Activations visualization (Yosinski et al. 2015)

After that, the units were assigned the interpretable labels that were varied from the textures, parts, colors, scenes, materials, and objects. This article also investigated the performance of the traditional methods on the learned model interpretability. The attribution methods, such as the activations visualization (Yosinski et al. 2015), heatmaps (Payer et al. 2019), class activation methods (GradCAM Selvaraju et al. 2020), and saliency maps (Adebayo et al. 2018), were introduced to further enhance the visualization of CNN models, as shown in Fig. 9. They relied on the gradients that flew into any specific layer to generate the coarse localization maps to mark the meaningful areas in the image in order to predict the target label. They can be easily implemented in any applications, because they were supported in the standard deep learning libraries, such as TensorFlow (tf-explain), Torch (Captum).

Equally important to the visualization, some researchers introduced the textual justification to train AI models to simultaneously produce a prediction and corresponding explanation. For instance, the critical illness early warning score framework that used electronic health records (EHR) to explain its prediction was proposed by Lauritsen et al. (2020). The framework brought the user?s trust by outputting a prediction and the corresponding information on the EHR to explain it. The proposed model covered two aspects of the model explanations, which include the individual and population aspects. The model provided clinicians the explanations for the model?s inner processes even though they do not have any technical understanding of the inner processes behind it. The mention framework has some weaknesses. Firstly, its explanations did not precisely reflect how the predictions were decided. Secondly, it relied on the training dataset, such as the stock dataset and EHR dataset, to explain the prediction, which is usually not the case. As a result, some textual justification techniques that did not demand the training set that contained descriptions for the prediction have been suggested lately. For instance, Park et al. (2018) introduced an explainable multi-modal framework that could give the natural language explanations of decisions and show the evidence in an image. The proposed approach needed the textual description of an image and its label as the ground truth during the training process to predict the image label and the visual explanation during the testing process. In another study, Holzinger et al. (2021a) proved that graph neural networks (GNN) is a potential approach for multi-modal embeddings and explainability, because the causal associations between the features can be mapped directly to the graph structures. Even though most of the mentioned techniques required the *local explanations* to perform the explainability, some studies explicitly concentrated on extracting the global descriptors using the locally found descriptors. Through the literature in Assaf et al. (2019) and Neto and Paulovich (2020), the authors empirically confirmed that the global descriptors were dependent on local descriptors. They showed that the extraction of the global descriptors provided a solid reference that prevented the attempt to modify how certain local representations were obtained.

Rather than performing only a single interpretability method, the DeepDream framework was suggested by Mordvintsev et al. (2015) as an attempt to examine and comprehend how the CNN model interpreted the images using the psychedelic images. Initially, the authors started with an image that contained random noise, then they overturned the network and forced it to enhance the image in order to obtain a particular interpretation. Three years later, the same group of authors Olah et al. (2018) proposed a new framework that improved the DeepDream framework, which focused on explaining the role of individual neurons during the prediction process. The combination of the attribution module, which explains the relationship between neurons, and the feature visualization module, which shows what each neuron detects, allows a complete

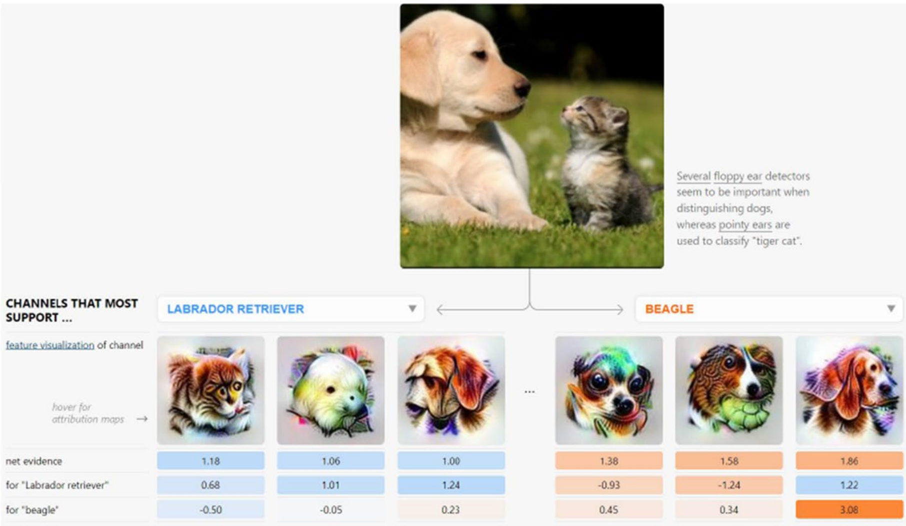


Fig. 10 Samples that were generated by the visualization framework (Olah et al. 2018), which combines the feature visualization (what does each neuron detect) with the attribution (how does each neuron affect the prediction) to show how the model determines the output between Labrador retriever and beagle classes

analysis of the role of each neuron. The visual inspection interface of the libraries contains two main functions are based on users? intention, which include attribution and feature visualization. It consists of a combination of individual components to visualize the content (activations and attribution), the atoms (group, spatial, channel, or neuron), the layers (input, hidden layers, output), and the presentation (feature and information visualization). The example of the framework is illustrated in Fig. 10.

In a more recent study, Carter et al. (2019) introduced an activation atlas with the primary purpose was to support the experts to discover the unanticipated issues in the deep learning models by visualizing the neuron interactions. For example, the library detected false associations or similar features between the two distinct classes during the classification process. Figure 11 explains one case, where the model views the appearance of the noodles as a crucial characteristic of a ?wok? label but not a ?frying fan? label. As a result, it becomes straightforward for the users to explain why the model incorrectly classified a frying pan filled with spaghetti as ?wok? using the activation atlas library.

The well-known model-agnostic libraries, such as LIME (Ribeiro et al. 2016), and SHapley Additive exPlanation (SHAP) (Chen et al. 2021), have gained a lot of attention from the research community, because they use a more straightforward strategy than the previously discussed techniques that can be integrated into any black-box ML models. The depiction of each library is described in Fig. 12. Both methods exploit and utilize the local explainability?s characteristics in order to develop the proxy models to explain the black-box ML models. They slightly change the input, which is close to the original data point, and test the differences in the prediction. While LIME forms the sparse linear models for each output to interpret how the model works at the local level, SHAPE leverages Shapley values to score the feature importance. Shapley values examine all the possible outputs of an instance using all the potential combinations of the inputs.

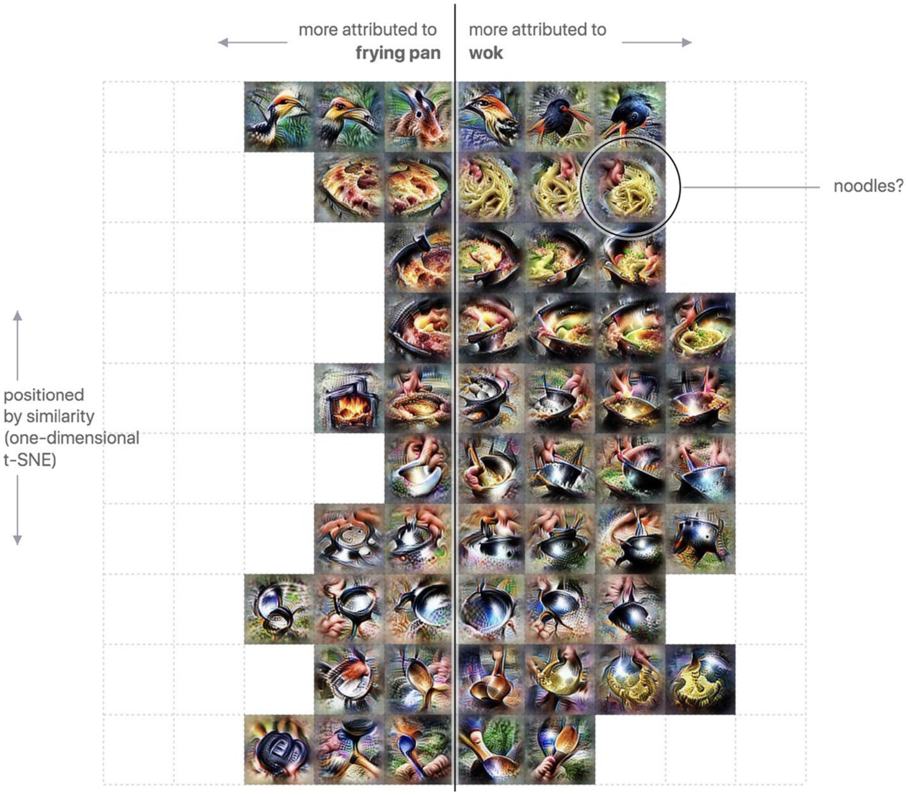


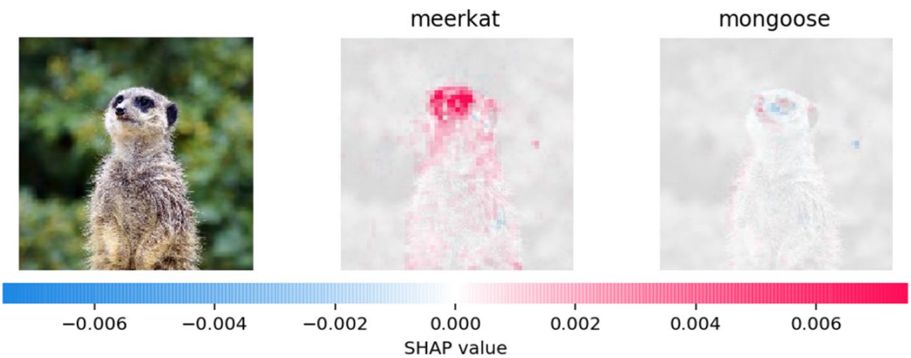
Fig. 11 Example of the explanations from the activation atlas library for a prediction between the ?wok? and the ?frying fan? classes (Carter et al. 2019)

All things considered, the *visualization* strategy is probably the most common approach to enable explainability for the CNN models.

- Recurrent Neural Networks

  Even though the CNN models worked well with 2D data, such as images, they failed to process sequential data, such as natural language or time-series data. The sequential data involves the long-term dependencies that are difficult for the CNN models to capture. As a result, the RNN models are devised to analyze sequential or temporal data to extract this type of time-dependent or sequence-dependent relationship efficiently. The RNNs take in the input data and utilize the activations from the previous nodes or the following nodes in a sequence in order to obtain better predictions.

  There have been a limited number of research methods that involved the RNNs explainability. It is possible to group them into two primary trends, which include (1) the analysis of information the RNNs learned using *feature relevance* in order to perform the explainability, and (2) the modification of RNNs? structures to offer explanations about how the output is determined using local explanations. For the first trend, Arras et al. (2019) adapted the LRP algorithm that was usually used in the feed-forward networks to interpret the RNN predictions. Moreover, they suggested a distinct propagation scheme, which was performed with multiplicative



(1) SHAP explanation

(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar*

(2) LIME explanation

Fig. 12 Examples of explanations that are generated by the SHAP and LIME frameworks

connections using the gated recurrent units (GRU) and long short term memory (LSTM). The novel modifications made the explanations of the proposed model more straightforward and achieved comparable performance to the original model. In another approach, Wang et al. (2018) implemented the RNNs explainability through the rule extraction from the trained RNNs models. The extracted rules outperformed the original RNN model, because they could recognize unseen long sequences of Tomita grammar through several experiments. Strobelt et al. (2017) created LSTMVis, which is an interactive visualization framework for the RNNs that enables the users to comprehend the RNN hidden state dynamics. The tool relies on direct inference when the users pick a range of text to describe a hypothesis, making it easy to adopt for a broad range of visual analyses of various tasks, datasets, and models. Strobelt et al. (2018) proposed Seq2Seq-Vis, which is a novel visual exploration tool for interpreting all stages of the sequence-to-sequence models. The tool facilitated the identification of the learned patterns, the error detection, the explainability, and the real-time interaction with input samples. Apart from the studies that did not alter the RNN structure, many studies have been proposed in order to improve the explainability by altering the RNN architecture. Ceni et al. (2020) presented a novel approach that modeled and explained RNNs using the input data. The authors used a theoretical framework called excitable network attractors that contained the constant attractors and the excitable connections attractors. The experimental results suggested that the regularization parameter directly impacted the number of attracting regions, which were created during the training process. Not long ago, Van Luong et al. (2021) presented a unique interpretable deep RNN network to perform video reconstruction from low-dimensional measurements. It was constructed by unfolding the iterations of proximal algorithms and enhancing the reweighed version, which led to better sparse approximations. Choi et al. (2016) recently devised a novel model, which was called REverse Time AttentIoN (RETAIN), to enable the explainability of black-box RNN models. The proposed model obtained a similar accuracy but with a higher explainability level. The fundamental concept of RETAIN was based on a complex attention generation method of training two RNNs while preserving the representation learning process straightforward for the interpretation. There has also been an exciting trend that builds a hybrid model to combine the CNNs and RNNs to jointly and efficiently predict an output class and provide proof to demonstrate that the output is correct. For example, Hendricks et al. (2016) demonstrated a unique reinforcement learning-based image visual explanation framework that used a novel sentence-level loss to decide the generated sentences. The experimental results confirmed that the created explanations were in line with the input image and more informative than the text generated by the previous approaches. Similarly, Zhang et al. (2020c) used the Hierarchical Refined Attention mechanism to sequentially connect the semantic attributes and the image features to refine the visual information from an input image. The generated captions proved that the proposed model created more comprehensive captions compared to the previous approach.

6 XAI: opportunities, challenges and future research needs

This section evaluates the cited literature to offer the final remarks on the accomplishments, the trending research topics, and the remaining difficulties that require more attention in the field of XAI. Even though the analysis of the XAI research in the previous sections already introduced some of the difficulties, they were not complete. As a result, they are mentioned again in this section, along with the new research possibilities for XAI and the potential techniques that can be researched to address them efficiently in the future.

- The tradeoff between the accuracy and the explainability was briefly discussed during the introduction of the XAI in Sect. 2.5.2. This problem happens during the improvement of the explainability, because when a model becomes more interpretable, its performance is significantly affected. Section 6.1 discusses the potential research paths to enable the explainability and maintain the model?s performance.
- The pressing demand to achieve an agreement on the concept of XAI within the AI domain was emphasized, and the motivations for researching XAI were discussed in Sect. 1.1. Section 6.2 will study these problems in detail. In addition, the standard metrics to evaluate XAI models will also be investigated.

- Given that the deep learning-based models have dominated many AI domains and reached a remarkable performance, XAI has been increasingly applied to solve the black-box AI models and bring about more trust from the users. Section 5.2.3 concentrates on categorizing various strategies for the deep learning models? explainability, exploring the progress reported previously on a particular taxonomy. Section 6.3 delves into this topic and reveals several difficulties regarding the explainability of the deep learning models.

6.1 Considering the tradeoff between the performance and the explainability

Even though the tradeoff between the explainability and the mode?s performance has been discussed from time to time, this issue was still covered in myths and misunderstandings like the other big XAI difficulties. Guo et al. (2019) confirmed that it is not entirely correct that the most complex black-box model consistently delivers the highest accuracy for a prediction problem. For instance, one case that this statement is proved wrong is when the data used for training is well-prepared and contains only the fundamental features. The stated case commonly appears in the industrial applications because the constraints are set for the data and the features being investigated within the controlled experimental environments in order to make the collected features deeply correlated to the problem under consideration. Rudin (2019) showed that simple models, which are linear regression and rule-based learners, achieved comparable performance to the complicated models, such as deep learning-based models, ensemble-based learners, and random forests. Moreover, there exists no noticeable difference in their performance. It can only be concluded that the models with complicated structures are more adaptable than the simple models, which allows them to compress the complicated functions. The mentioned statement is correct when the problem under consideration has a certain complexity, and the training dataset is hugely available to train a complicated black-box model. For instance, as more input data are fed into a deep learning model and contain more hidden layers, the prediction accuracy is significantly improved. However, the explainability of the predictions becomes difficult (Moradi and Samwald 2021; Wang et al. 2017). In this case, the tradeoff between the predictive accuracy and the explainability can be considered.

The thought that the mentioned tradeoff always exists has driven the researchers to sacrifice the explainability to further improve the model?s performance. Therefore, the additional complexity of the model only supports finding a more accurate solution for the problem, whereas the explainability finds itself on a downwards trend that up to now seemed inevitable. This problem has led to the introduction of many recent studies to reverse or at least reduce that trend. For instance, Dziugaite et al. (2020) studied the tradeoff between the model?s performance and the explainability using a simplified model. They concentrated on minimizing the risk for binary classification and considered the explainability a constraint enforced during the training phase. The experimental results demonstrated that the model provided an accurate analysis of the parts that contribute to the model?s performance and how they were influenced. Figure 13 presents the tradeoff between the model?s performance and its explainability and how the future XAI techniques can be proposed to improve it. Overall, the model explainability capability is forecast to increase significantly compared to the model?s performance, because the number of research that proposed novel XAI methods has grown remarkably in recent years.

Another perspective worth discussing is that the performance versus explainability tradeoff should be decided according to the application domain and the target users concerned. If one requires the model to be explainable, they need to use a simple model that is not as powerful

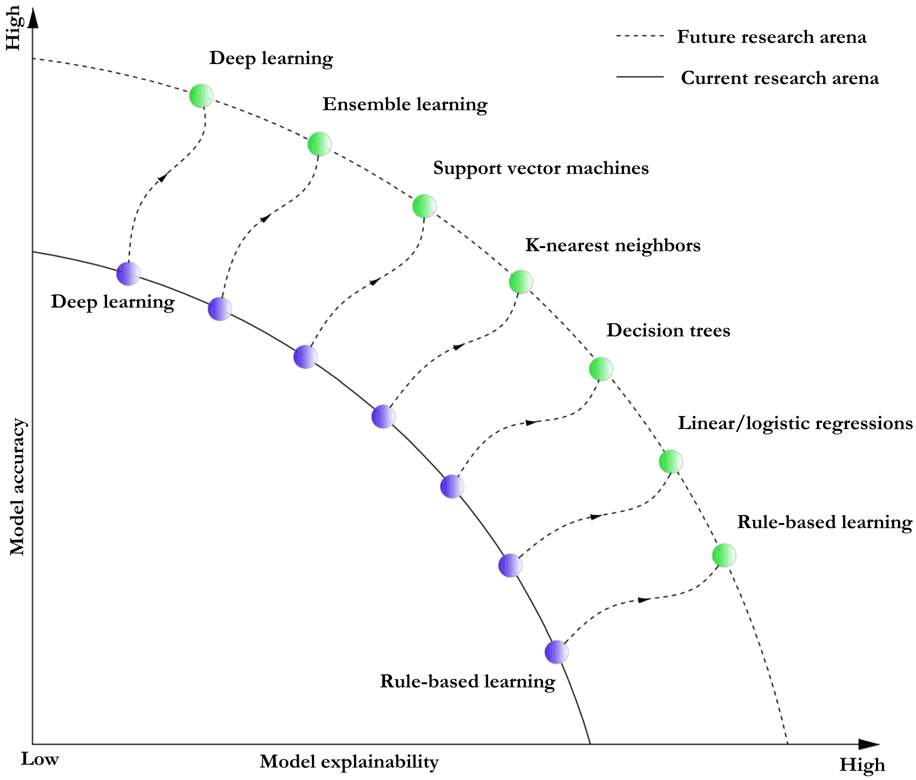


Fig. 13 Relationship between the model accuracy and the explainability of the current XAI research and the future XAI research for seven standard models, ranging from the simple models, such as rule-based learning to complex models, such as deep learning

or precise as black-box models. When performance matters most, even with a complicated model, one should opt for the trust that comes from ensuring that one can check that the model works correctly.

6.2 Considering the XAI concept and evaluation methods

This paper stressed the need for a unified concept of the XAI in Sect. 2.2 to express the demands of XAI in the AI domain. It is crucial to find common ground for the XAI topic in order to put in the cornerstone to develop novel XAI approaches. The literature showed the two concepts suggested by two disciplines DARPA and FAT* that were widely accepted in Sect. 2.2. In addition to these concepts, many review papers also attempted to define the XAI concept. For instance, Arrieta et al. (2020) defined the XAI as the ability of any AI model in order to make its internal processes more transparent to the audiences by implementing the post-modeling methods. Even though the XAI concepts represented in this review can become insufficient as the XAI research is still in the early development stage, they can be relied on as a baseline in order to convey a valuable reference on the topic. It is believed that the XAI research community will eventually reach a unified concept of the XAI by connecting the shattered contributions of an increasing number of XAI studies.

Evaluation metrics are fundamental metrics required to evaluate a particular model. Any assertion that a proposed model is effective without showing the evaluation metrics statistics is considered hard to accept, because it does not provide any solid proof of the model?s effectiveness in order to persuade the readers. In the context of XAI, the evaluation metrics enable a thorough measurement of the quality of how well a model meets the XAI requirements. As with the standard evaluation metrics for classification, such as accuracy, precision, and recall, evaluation metrics for XAI should prove the model performance in a specific viewpoint of XAI. Many efforts have been made lately to discuss the evaluation metrics for XAI (Hoffman et al. 2018; Carvalho et al. 2019b). In general, the suggested metrics enable researchers to evaluate the XAI model?s performance and improve the confidence and trust of the end-users in the model. According to Hoffman et al. (2018), evaluation metrics for XAI measure the explanation satisfaction, explanation goodness, and scale validations, which are discriminant and content validity, whereas Carvalho et al. (2019b) considered the XAI metrics include both the quantitative indicators and the qualitative indicators. The two studies that concentrated on the XAI evaluation metrics appear to be good examples of the importance of evaluating the XAI models. However, the final remarks and future works discussed in these studies agreed with the XAI prospects presented in this survey that more standard and quantifiable XAI evaluation metrics are required in order to measure the increasing number of XAI tools and techniques that are introduced by the community. This review does not address the urgent need to devise widely acceptable evaluation metrics because it is better to be solved in future research before the agreement of the standard theory of explainability, which is one of the primary intentions of this survey. Nonetheless, this study supports spending more effort towards the novel introduction of the XAI evaluation metrics to effectively and efficiently evaluate the performance of the XAI algorithms. Moreover, the comparison methods between different XAI methods that support comparing these methods quantitatively under various application contexts also need to be considered.

6.3 Considering the XAI for deep learning

Even though enormous efforts have been put into explaining the deep learning models in recent years, many difficulties remain to be solved before obtaining the full explainability for the DL models. The compromise on the definition of the main XAI categories for the deep learning models is still missing, because a lot of research is still in progress, which is mentioned in the previous sections. In addition, the research community does not have a standard terminology for XAI yet. For instance, it can be witnessed that the *feature relevance* and the *feature importance* terms can be used interchangeably. Another example is the *post-modeling visualization* category, which is where no agreement supports the definition of the methods, such as saliency maps, heatmaps, neuron activations, and similar definitions.

In addition, the challenge of interpretation is inevitable for the complex models, such as deep learning with various non-linear activation functions, which is different from the linear models that are fully interpretable, and each unit change in the inputs will lead to a constant change in the output. The deep learning models in high-dimensional space are usually more complicated than the simple models in low-dimensional space, which makes it challenging to perform the interpretation. Even though many approaches have been introduced to visualize the inner process of the deep learning models, each approach has its weaknesses. For example, some studies proposed compressing the deep learning models

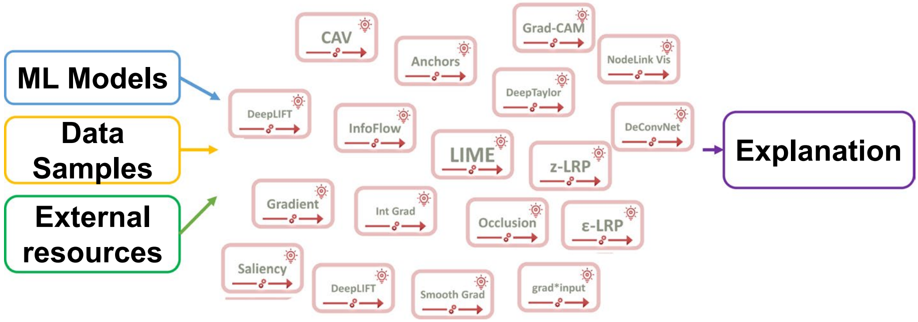


Fig. 14 The diversity of the XAI interpretation methods that have been introduced recently for deep learning (Spinner et al. 2019)

from a high-dimensional representation into a low-dimensional representation (Zhao and Du 2016; Erfani et al. 2016). These methods still demand the readers to have a strong background in AI in order to perceive the process. Most of the current various visualization methods that were proposed to interpret the deep learning models are prone to simple attacks that can completely fool these methods. Ghorbani et al. (2019) and Su et al. (2019) revealed that the interpretation methods, such as saliency maps for black-box models, were weak against adversarial perturbations. Adversarial perturbations refer to the generation of perceptively indistinguishable inputs that hugely affect the interpretation outputs. In another study, Wang et al. (2019c) proved that ignoring bias terms in the deep learning models could provide the wrong input feature attributions. Kindermans et al. (2019) used a standard and simple pre-processing method to prove that a transformation could indirectly manipulate the deep learning networks. The authors proposed to fulfill input invariance in order to guarantee the reliability. Moreover, as more explainable methods are introduced for deep learning, which is illustrated in Fig. 14, it becomes challenging to implement and manage them, because each technique has its dependencies and produces distinct outputs. Therefore, standardization and policies are required in order to unify these techniques further to improve the explainability of the deep learning models.

Although the conventional deep learning models consistently demonstrate impressive learning performance that matches and exceeds people?s cognitive skills in various domains, they are data-specific models that require to be trained on specific data to perform a particular task. Meta-learning is a trending AI topic recently that can solve the mentioned problem by learning many tasks together. For example, the Google DeepMind research group implemented different meta-reinforcement learning models to mimic dopamine?s role during the training phase and then provided a comparison between the activity dynamics of the repetitive network and the accurate data from the existing findings in neuroscience trials. Recently, explainable meta-reinforcement learning (xMRL) was proposed due to the increasing interest in XAI (Da?larli 2020). For example, an xMRL-based agent can develop its strategy by identifying the cause-effect relationships. It can be trained to play chess, Go, checkers, and even learn and adjust the strategy when encounters a new game. In terms of explainability, the agent can explain why a specific action is made upon a move made by the enemy. It is recommended that the interested readers referred to the latest reviews of reinforcement learning (Gronauer and Diepold 2021) and meta-learning (Huisman et al. 2021) to get an in-depth and comprehensive view of the topic.

The researchers from various domains must continuously contribute to this challenging topic in order to eventually reach a set of widely acceptable XAI evaluation metrics until a standard XAI evaluation metric is introduced, which is discussed in Sect. 6.2. The researchers can borrow ideas from other fields, such as social studies, in order to speed up the process by answering a predetermined set of evaluation questions and deciding the evaluation data samples (Carvalho et al. 2019a). One more emerging challenge is enabling the deep learning models to produce comprehensible explanations for users from various sectors, especially those with high demands for explainable AI, such as healthcare, law, and banking. Delivering explainable and interpretable outputs to end-users who have no technical knowledge requires the researchers to concentrate on reducing uncertainties and reaching common goals for explanations (Wang et al. 2019b).

6.4 Considering the XAI for AI security

AI has become ubiquitous nowadays. It has enormous possibilities to build an innovative and intelligent world, even though it is challenged by severe security risks. As a consequence of ignoring the security aspect during the development of AI algorithms, the attackers can implement many techniques to manipulate the inference results. In crucial domains, such as surveillance, healthcare, and transportation, it can lead to devastating consequences when a model is manipulated, such as misinterpretation of the results, property loss, and threaten personal safety. There are various types of AI security attacks that include adversarial examples generation, transferability, poisoning, backdoor, and model extraction, which are described in Table 8.

The attacks mentioned above and the methods to defend against them have been studied extensively in recent years (Dunn et al. 2020; Juuti et al. 2019). The researchers have proved that they are precise and have excellent transferability, which has resulted in errors during the predictions of any AI model. In order to prevent known attacks and any unknown attacks in the future, it is fundamental to improve the confidentiality and the security of the AI models. Figure 15 represents some of the security improvements, which are based on XAI.

Before and after the development of an AI model, it is imperative to implement various security evaluations. For example, a pre-processing agent can be programmed to detect and eliminate the manipulated samples before they are fed into the training model. In order to further reduce the false positives, a post-processing agent can be applied to verify the integrity of the predictions. It is possible to improve the robustness of AI models before and after deployment with these additional agents. The XAI methods can be implemented while constructing and training a model in order to enhance the model?s explainability by automatically examining and explaining the model problems that are difficult for humans to detect, such as logical errors and data blind spots.

6.5 Considering the XAI for community, policymakers, and the law

The development programs and policies that have been established by nations worldwide are a fundamental factor that stimulates the evolution of XAI. This section addresses the ongoing policies and programs that have been released to motivate XAI development from international organizations and countries across the world.

Fairness, transparency, privacy, and explainability are four key aspects of the EU General Data Protection Regulation (GDPR) compliance (Wachter et al. 2017). In particular,

Table 8 Typical AI security attacks

Attack	Contents	References
Adversarial examples	• Adds small perturbations into the original data • Invisible to the human eyes but significantly affect the predictions of the ML models	Su et al. (2019), Ghorbani et al. (2019)
Transferability	• Expects the target model to have the same training data with a model • The information about the model parameters is unnecessary	Huang et al. (2019)
Poisoning	• Introduces carefully manipulated samples to pollute the training data • Includes statistical optimization, global optimum, and optimal gradient attacks	Dunn et al. (2020)
Backdoor	• Injects some particular features into a model • It is usually conducted during the model creation process	Dai et al. (2019)
Model inversion	Uses model outputs to compromise user privacy	Veale et al. (2018)
Model extraction	Analyzes the input, output, and available information of a model in order to obtain the model parameters and eventually hijack the model	Juuti et al. (2019), Wang and Gong (2018)

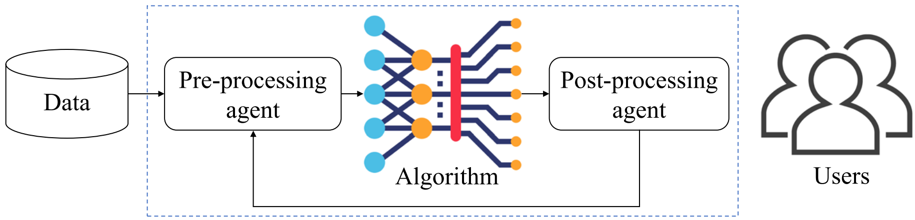


Fig. 15 The enhancement of security for AI models through XAI

Article 22 of GDPR empowers individuals with the right to demand explanations of how an AI system reached a decision that directly affects them. Currently, a position paper piloted by Denmark and endorsed by 13 additional member states established that regardless of the policies enforced toward the AI/ML technologies, they need to abide by the GDPR (Peloquin et al. 2020). In addition, the UK Data Protection Act 2018 mentioned the principles of fairness, transparency, and accountability (Carey 2018). With the foundation of the two mentioned regulations, *Explaining decisions made with AI*, or the *ICO guidance* has been introduced to lay out a general framework for the organizations that apply AI to assist individuals during the decision-making process or to reach decisions itself (Butterworth 2018).

The regulators require AI/ML models to follow the standards of the US Federal Reserve?s SR 11-7 (Kiritz and Sarfati 2018), and Comptroller of the Currency (OCC) 2011-12,1 guidance (Comizio et al. 2011) on the model risk management (MRM). They guided the general model development and validation and established the requirements that users understand the model?s weaknesses and its primary intention to prevent practicing the model in the wrong way that is different from the model intention.

7 Conclusion

The eXplainable artificial intelligence (XAI) is experiencing a rapid transformation due to recent deep learning advances, where many earlier unsolved obstacles have become step by step solvable. The latest progress proved that applications with real-world complexity could gradually be interpretable (Kuo et al. 2019; Rudin 2019; De et al. 2020; Nguyen et al. 2020a). Nevertheless, XAI is a young topic that attracts growing interest, and the number of published articles rises quickly. This survey concentrated on different aspects of eXplainable artificial intelligence (XAI). We first covered the important XAI concepts and the relevant XAI surveys that have been conducted in recent years. After that, the XAI background was then presented by answering several questions, which included what, why, and how that revolved around the XAI topic. These conceptual analyses are good motivation for a comprehensive survey of the latest XAI research. After that, this report introduced a global taxonomy in order to classify over 250 XAI articles under a consistent standard by systematically categorizing the existing XAI research into (1) the pre-modeling explainability, which focuses on analyzing and explaining the data, (2) the interpretable model that indicates the simple models that have some levels of explainability, so it is therefore interpretable to some extent by themselves, and (3) the post-modeling explainability, which is proposed in order to turn the existing black-box models into the interpretable models. Moreover, this survey put special attention on deep learning by creating a subsection to carefully discuss the existing literature that delivers the explainability to the deep learning models. The speculations about the prospect of XAI were mentioned throughout the survey, where we highlighted XAI potential and exposed notable opportunities as well as its limitations. XAI needs to be simultaneously solved with other AI characteristics, such as accessibility, privacy, fairness, and transferability, in order to enable reliable employment and adaptation of the AI algorithms in companies and organizations worldwide.

With those contributions, we hope to provide the interested readers with the necessary means to grasp some basic XAI-related knowledge. In the near future, we anticipate that there will be an abundance of new literature emerge. Therefore, we hope to encourage the AI research community for extra contributions to this exciting and young field of research.

References

Abdollahi B, Nasraoui O (2018) Transparency in fair machine learning: the case of explainable recommender systems. In: Human and machine learning. Springer, Berlin, pp 21?35

ACM (2020) ACM conference on fairness, accountability, and transparency. https://fatconference.org. Accessed 24 Jan 2020

Adadi A, Berrada M (2018) Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access 6:52138?52160

Adebayo J, Gilmer J, Muelly M, Goodfellow I, Hardt M, Kim B (2018) Sanity checks for saliency maps. Adv Neural Inf Process Syst 31:9505?9515

Adler P, Falk C, Friedler SA, Nix T, Rybeck G, Scheidegger C, Smith B, Venkatasubramanian S (2018) Auditing black-box models for indirect influence. Knowl Inf Syst 54(1):95?122

Adriana da Costa FC, Vellasco MMB, Tanscheit R (2013) Fuzzy rules extraction from support vector machines for multi-class classification. Neural Comput Appl 22(7):1571?1580

Ahmed M (2019) Data summarization: a survey. Knowl Inf Syst 58(2):249?273

Ahn Y, Lin YR (2019) Fairsight: visual analytics for fairness in decision making. IEEE Trans Vis Comput Graph 26(1):1086?1095

AI (2019) Ethics for autonomous systems. https://www.journals.elsevier.com/artificial-intelligence/call-for-papers/special-issue-on-ethics-for-autonomous-systems. Accessed 3 Mar 2020

AI (2020) Explainable artificial intelligence. https://www.journals.elsevier.com/artificial-intelligence/call-for-papers/special-issue-on-explainable-artificial-intelligence. Accessed 3 Mar 2020

Akula AR, Todorovic S, Chai JY, Zhu SC (2019) Natural language interaction with explainable AI models. In: CVPR workshops, pp 87?90

Al-Shedivat M, Dubey A, Xing E (2020) Contextual explanation networks. J Mach Learn Res 21(194):1?44

Angelov P, Soares E (2020) Towards explainable deep neural networks (xDNN). Neural Netw 130:185?194

Anysz H, Zbiciak A, Ibadov N (2016) The influence of input data standardization method on prediction accuracy of artificial neural networks. Proc Eng 153:66?70

Arras L, Arjona-Medina J, Widrich M, Montavon G (2019) Explaining and interpreting lstms. In: Explainable AI: interpreting, explaining and visualizing deep learning, vol 11700, p 211

Arrieta AB, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, García S, Gil-López S, Molina D, Benjamins R et al (2020) Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fusion 58:82?115

Asadi S, Nilashi M, Husin ARC, Yadegaridehkordi E (2017) Customers perspectives on adoption of cloud computing in banking sector. Inf Technol Manag 18(4):305?330

Assaf R, Giurgiu I, Bagehorn F, Schumann A (2019) Mtex-cnn: Multivariate time series explanations for predictions with convolutional neural networks. In: 2019 IEEE international conference on data mining (ICDM). IEEE, pp 952?957

Bang JS, Lee MH, Fazli S, Guan C, Lee SW (2021) Spatio-spectral feature representation for motor imagery classification using convolutional neural networks. IEEE Trans Neural Netw Learn Syst

Baniecki H, Biecek P (2019) modelStudio: Interactive studio with explanations for ML predictive models. J Open Source Softw 4(43):1798

Baron B, Musolesi M (2020) Interpretable machine learning for privacy-preserving pervasive systems. IEEE Pervasive Comput
Bau D, Zhou B, Khosla A, Oliva A, Torralba A (2017) Network dissection: Quantifying interpretability of deep visual representations. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 6541?6549
Bender EM, Friedman B (2018) Data statements for natural language processing: toward mitigating system bias and enabling better science. Trans Assoc Comput Linguist 6:587?604
Bi X, Zhang C, He Y, Zhao X, Sun Y, Ma Y (2021) Explainable time?frequency convolutional neural network for microseismic waveform classification. Inf Sci 546:883?896
Blanco-Justicia A, Domingo-Ferrer J, Martínez S, Sánchez D (2020) Machine learning explainability via microaggregation and shallow decision trees. Knowl-Based Syst 194:105532
BMVC (2020) Interpretable & explainable machine vision. https://arxiv.org/html/1909.07245. Accessed 3 Mar 2020
Bologna G (2019) A simple convolutional neural network with rule extraction. Appl Sci 9(12):2411
Butterworth M (2018) The ICO and artificial intelligence: the role of fairness in the GDPR framework. Comput Law Secur Rev 34(2):257?268
Campbell T, Broderick T (2019) Automated scalable Bayesian inference via Hilbert coresets. J Mach Learn Res 20(1):551?588
Cao HE, Sarlin R, Jung A (2020) Learning explainable decision rules via maximum satisfiability. IEEE Access 8:218180?218185
Carey P (2018) Data protection: a practical guide to UK and EU law. Oxford University Press, Inc, Oxford
Carter S, Armstrong Z, Schubert L, Johnson I, Olah C (2019) Activation atlas. Distill 4(3):e15
Carvalho DV, Pereira EM, Cardoso JS (2019a) Machine learning interpretability: a survey on methods and metrics. Electronics 8(8):832
Carvalho DV, Pereira EM, Cardoso JS (2019b) Machine learning interpretability: a survey on methods and metrics. Electronics 8(8):832
Ceni A, Ashwin P, Livi L (2020) Interpreting recurrent neural networks behaviour via excitable network attractors. Cogn Comput 12(2):330?356
Chakraborty S, Tomsett R, Raghavendra R, Harborne D, Alzantot M, Cerutti F, Srivastava M, Preece A, Julier S, Rao RM et al (2017) Interpretability of deep learning models: a survey of results. In: 2017 IEEE SmartWorld, ubiquitous intelligence & computing, advanced & trusted computed, scalable computing & communications, cloud & big data computing, internet of people and smart city innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI). IEEE, pp 1?6
Chan TH, Jia K, Gao S, Lu J, Zeng Z, Ma Y (2015) PCANet: a simple deep learning baseline for image classification? IEEE Trans Image Process 24(12):5017?5032
Chen J, Song L, Wainwright MJ, Jordan MI (2018) L-shapley and c-shapley: efficient model interpretation for structured data. In: International conference on learning representations
Chen J, Vaughan J, Nair V, Sudjianto A (2020a) Adaptive explainable neural networks (AxNNs). Available at SSRN 3569318
Chen Y, Yu C, Liu X, Xi T, Xu G, Sun Y, Zhu F, Shen B (2020b) PCLiON: an ontology for data standardization and sharing of prostate cancer associated lifestyles. Int J Med Inform 145:104332
Chen H, Lundberg S, Lee SI (2021) Explaining models by propagating Shapley values of local components. In: Explainable AI in Healthcare and Medicine. Springer, Berlin, pp 261?270
Choi E, Bahadori MT, Kulas JA, Schuetz A, Stewart WF, Sun J (2016) Retain: an interpretable predictive model for healthcare using reverse time attention mechanism. In: Advances in Neural Information Processing Systems, pp 3512?3520
Choi KS, Choi SH, Jeong B (2019) Prediction of IDH genotype in gliomas with dynamic susceptibility contrast perfusion MR imaging using an explainable recurrent neural network. Neuro Oncol 21(9):1197?1209
Choi H, Som A, Turaga P (2020) AMC-loss: angular margin contrastive loss for improved explainability in image classification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, pp 838?839
Choo J, Liu S (2018) Visual analytics for explainable deep learning. IEEE Comput Graph Appl 38(4):84?92
CIM I (2021) Explainable and trustworthy artificial intelligence. https://sites.google.com/view/special-issue-on-xai-ieee-cim. Accessed 1 Aug 2021
Comizio VG, Petrasic KL, Lee HY (2011) Regulators take steps to eliminate differences in thrift, bank and holding company reporting requirements. Banking LJ 128:426
Cortez P, Embrechts MJ (2013) Using sensitivity analysis and visualization techniques to open black box data mining models. Inf Sci 225:1?17

Craven MW, Shavlik JW (2014) Learning symbolic rules using artificial neural networks. In: Proceedings of the tenth international conference on machine learning, pp 73?80
Daglarli E (2020) Explainable artificial intelligence (XAI) approaches and deep meta-learning models. In: Advances and applications in deep learning, p 79
Dai J, Chen C, Li Y (2019) A backdoor attack against lstm-based text classification systems. IEEE Access 7:138872?138878
Dang LM, Hassan SI, Im S, Mehmood I, Moon H (2018) Utilizing text recognition for the defects extraction in sewers CCTV inspection videos. Comput Ind 99:96?109
Dang LM, Piran M, Han D, Min K, Moon H et al (2019) A survey on internet of things and cloud computing for healthcare. Electronics 8(7):768
Darpa (2020) Explainable artificial intelligence (XAI). https://www.darpa.mil/program/explainable-artificial-intelligence. Accessed 24 Jan 2020
De T, Giri P, Mevawala A, Nemani R, Deo A (2020) Explainable AI: a hybrid approach to generate human-interpretable explanation for deep learning prediction. Procedia Comput Sci 168:40?48
Deeks A (2019) The judicial demand for explainable artificial intelligence. Columbia Law Rev 119(7):1829?1850
Deleforge A, Forbes F, Horaud R (2015) High-dimensional regression with gaussian mixtures and partially-latent response variables. Stat Comput 25(5):893?911
Deng H (2019) Interpreting tree ensembles with intrees. Int J Data Sci Anal 7(4):277?287
Dibia V, Demiralp Ç (2019) Data2vis: automatic generation of data visualizations using sequence-to-sequence recurrent neural networks. IEEE Comput Graph Appl 39(5):33?46
Ding L (2018) Human knowledge in constructing AI systems?neural logic networks approach towards an explainable AI. Procedia Comput Sci 126:1561?1570
Dingen D, van?t Veer M, Houthuizen P, Mestrom EH, Korsten EH, Bouwman AR, Van Wijk J (2018) Regressionexplorer: interactive exploration of logistic regression models with subgroup analysis. IEEE Trans Vis Comput Graph 25(1):246?255
DMKD (2021) Data mining and knowledge discovery. https://www.springer.com/journal/10618/updates/18745970. Aceessed 1 Aug 2021
Dogra DP, Ahmed A, Bhaskar H (2016) Smart video summarization using mealy machine-based trajectory modelling for surveillance applications. Multimed Tools Appl 75(11):6373?6401
Doran D, Schulz S, Besold TR (2017) What does explainable AI really mean? A new conceptualization of perspectives. arXiv preprint arXiv:171000794
DuMouchel W (2002) Data squashing: constructing summary data sets. In: Handbook of massive data sets. Springer, Cham, pp 579?591
Dunn C, Moustafa N, Turnbull B (2020) Robustness evaluations of sustainable machine learning models against data poisoning attacks in the internet of things. Sustainability 12(16):6434
Dziugaite GK, Ben-David S, Roy DM (2020) Enforcing interpretability and its statistical impacts: trade-offs between accuracy and interpretability. arXiv preprint arXiv:201013764
Eiras-Franco C, Guijarro-Berdiñas B, Alonso-Betanzos A, Bahamonde A (2019) A scalable decision-tree-based method to explain interactions in dyadic data. Decis Support Syst 127:113141
Electronics (2019) Interpretable deep learning in electronics, computer science and medical imaging. https://www.mdpi.com/journal/electronics/special_issues/interpretable_deep_learning. Accessed 3 Mar 2020
Elghazel H, Aussem A (2015) Unsupervised feature selection with ensemble learning. Mach Learn 98(1):157?180
Elshawi R, Al-Mallah MH, Sakr S (2019) On the interpretability of machine learning-based model for predicting hypertension. BMC Med Inform Decis Mak 19(1):1?32
Erfani SM, Rajasegarar S, Karunasekera S, Leckie C (2016) High-dimensional and large-scale anomaly detection using a linear one-class SVM with deep learning. Pattern Recogn 58:121?134
Escalante HJ, Escalera S, Guyon I, Baró X, Güçlütürk Y, Güçlü U, van Gerven M, van Lier R (2018) Explainable and interpretable models in computer vision and machine learning. Springer, Cham
Escobar CA, Morales-Menendez R (2019) Process-monitoring-for-quality?a model selection criterion for support vector machine. Procedia Manuf 34:1010?1017
Fang X, Xu Y, Li X, Lai Z, Wong WK, Fang B (2017) Regularized label relaxation linear regression. IEEE Trans Neural Netwo Learn Syst 29(4):1006?1018
Felzmann H, Fosch-Villaronga E, Lutz C, Tamo-Larrieux A (2019) Robots and transparency: the multiple dimensions of transparency in the context of robot technologies. IEEE Robotics Autom Mag 26(2):71?78

Fernandez A, Herrera F, Cordon O, del Jesus MJ, Marcelloni F (2019) Evolutionary fuzzy systems for explainable artificial intelligence: why, when, what for, and where to? IEEE Comput Intell Mag 14(1):69?81

FGCS (2021) Future generation computer systems. https://www.journals.elsevier.com/future-generation-computer-systems/call-for-papers/explainable-artificial-intelligence-for-healthcare. Accessed 1 Aug 2021

Forte JC, Mungroop HE, de Geus F, van der Grinten ML, Bouma HR, Pettilä V, Scheeren TW, Nijsten MW, Mariani MA, van der Horst IC et al (2021) Ensemble machine learning prediction and variable importance analysis of 5-year mortality after cardiac valve and CABG operations. Sci Rep 11(1):1?11

Främling K (2020) Decision theory meets explainable AI. In: International workshop on explainable, transparent autonomous agents and multi-agent systems. Springer, Cham, pp 57?74

Gallego AJ, Calvo-Zaragoza J, Valero-Mas JJ, Rico-Juan JR (2018) Clustering-based k-nearest neighbor classification for large-scale data with neural codes representation. Pattern Recogn 74:531?543

Gaonkar B, Shinohara RT, Davatzikos C, Initiative ADN et al (2015) Interpreting support vector machine models for multivariate group wise analysis in neuroimaging. Med Image Anal 24(1):190?204

García-Magariño I, Muttukrishnan R, Lloret J (2019) Human-centric AI for trustworthy IoT systems with explainable multilayer perceptrons. IEEE Access 7:125562?125574

Gartner (2020) Gartner identifies the top 10 strategic technology trends for 2020. https://www.gartner.com/en/newsroom/press-releases/2019-10-21-gartner-identifies-the-top-10-strategic-technology-trends-for-2020. Accessed 24 Jan 2020

Ghorbani A, Abid A, Zou J (2019) Interpretation of neural networks is fragile. In: Proceedings of the AAAI conference on artificial intelligence, vol 33, pp 3681?3688

Gite S, Khatavkar H, Kotecha K, Srivastava S, Maheshwari P, Pandey N (2021) Explainable stock prices prediction from financial news articles using sentiment analysis. PeerJ Comput Sci 7:e340

Google (2021) Google what-if toolkit. https://pair-code.github.io/what-if-tool/. Accessed 26 Apr 2021

Gronauer S, Diepold K (2021) Multi-agent deep reinforcement learning: a survey. Artif Intell Rev 1?49

Gu D, Su K, Zhao H (2020a) A case-based ensemble learning system for explainable breast cancer recurrence prediction. Artif Intell Med 107:101858

Gu R, Wang G, Song T, Huang R, Aertsen M, Deprest J, Ourselin S, Vercauteren T, Zhang S (2020b) Ca-net: comprehensive attention convolutional neural networks for explainable medical image segmentation. IEEE Trans Med Imaging

Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D (2019) A survey of methods for explaining black box models. ACM Comput Surv (CSUR) 51(5):93

Gulati P, Hu Q, Atashzar SF (2021) Toward deep generalization of peripheral EMG-based human-robot interfacing: a hybrid explainable solution for neurorobotic systems. IEEE Robotics Autom Lett

Guo S, Yu J, Liu X, Wang C, Jiang Q (2019) A predicting model for properties of steel using the industrial big data based on machine learning. Comput Mater Sci 160:95?104

Guo W (2020) Explainable artificial intelligence for 6G: improving trust between human and machine. IEEE Commun Mag 58(6):39?45

Gupta B, Rawat A, Jain A, Arora A, Dhami N (2017) Analysis of various decision tree algorithms for classification in data mining. Int J Comput Appl 163(8):15?19

H2oai (2017) Comparative performance analysis of neural networks architectures on h2o platform for various activation functions. In: 2017 IEEE International young scientists forum on applied physics and engineering (YSF). IEEE, pp 70?73

Haasdonk B (2005) Feature space interpretation of SVMs with indefinite kernels. IEEE Trans Pattern Anal Mach Intell 27(4):482?492

Hagras H (2018) Toward human-understandable, explainable AI. Computer 51(9):28?36

Hara S, Hayashi K (2018) Making tree ensembles interpretable: a Bayesian model selection approach. In: International conference on artificial intelligence and statistics. PMLR, pp 77?85

Hatwell J, Gaber MM, Azad RMA (2020) Chirps: explaining random forest classification. Artif Intell Rev 53:5747?5788

Hatzilygeroudis I, Prentzas J (2015) Symbolic-neural rule based reasoning and explanation. Expert Syst Appl 42(9):4595?4609

Hendricks LA, Akata Z, Rohrbach M, Donahue J, Schiele B, Darrell T (2016) Generating visual explanations. In: European conference on computer vision. Springer, Cham, pp 3?19

Henelius A, Puolamäki K, Boström H, Asker L, Papapetrou P (2014) A peek into the black box: exploring classifiers by randomization. Data Min Knowl Disc 28(5):1503?1529

Hind M, Wei D, Campbell M, Codella NC, Dhurandhar A, Mojsilovi? A, Natesan Ramamurthy K, Varshney KR (2019) TED: teaching AI to explain its decisions. In: Proceedings of the 2019 AAAI/ACM conference on AI, ethics, and society, pp 123?129
Hoffman RR, Mueller ST, Klein G, Litman J (2018) Metrics for explainable AI: challenges and prospects. arXiv preprint arXiv:181204608
Holzinger A (2016) Interactive machine learning for health informatics: when do we need the human-in-the-loop? Brain Inform 3(2):119?131
Holzinger A, Langs G, Denk H, Zatloukal K, Müller H (2019) Causability and explainability of artificial intelligence in medicine. Wiley Interdiscip Rev Data Min Knowl Discov 9(4):e1312
Holzinger A, Malle B, Saranti A, Pfeifer B (2021a) Towards multi-modal causability with graph neural networks enabling information fusion for explainable AI. Inf Fusion 71:28?37
Holzinger A, Weippl E, Tjoa AM, Kieseberg P (2021b) Digital transformation for sustainable development goals (SDGS)?a security, safety and privacy perspective on AI. In: International cross-domain conference for machine learning and knowledge. Springer, Cham, pp 103?107
Hu K, Orghian D, Hidalgo C (2018a) Dive: a mixed-initiative system supporting integrated data exploration workflows. In: Proceedings of the workshop on human-in-the-loop data analytics, pp 1?7
Hu R, Andreas J, Darrell T, Saenko K (2018b) Explainable neural computation via stack neural module networks. In: Proceedings of the European conference on computer vision (ECCV), pp 53?69
Huang Q, Katsman I, He H, Gu Z, Belongie S, Lim SN (2019) Enhancing adversarial example transferability with an intermediate level attack. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 4733?4742
Huisman M, van Rijn JN, Plaat A (2021) A survey of deep meta-learning. Artif Intell Rev 1?59
IBM (2019) AI fairness 360: an extensible toolkit for detecting and mitigating algorithmic bias. IBM J Res Dev 63(4/5):4?1
ICAPS (2020) Explainable planning. https://icaps20.icaps-conference.org/workshops/xaip/. Accessed 3 Mar 2020
ICCV (2019) Interpretating and explaining visual artificial intelligence models. http://xai.unist.ac.kr/workshop/2019/. Accessed 3 Mar 2020
ICML (2021) Theoretic foundation, criticism, and application trend of explainable AI. https://icml2021-xai.github.io/. Accessed 1 Aug 2021
IDC (2020) Worldwide spending on artificial intelligence systems will be nearly 98 billion dollars in 2023. https://www.idc.com/getdoc.jsp?containerId=prUS45481219. Accessed 24 Jan 2020
IJCAI (2019) Explainable artificial intelligence(XAI). https://sites.google.com/view/xai2019/home. Accessed 3 Mar 2020
Islam MA, Anderson DT, Pinar AJ, Havens TC, Scott G, Keller JM (2019) Enabling explainable fusion in deep learning with fuzzy integral neural networks. IEEE Trans Fuzzy Syst 28(7):1291?1300
Islam NU, Lee S (2019) Interpretation of deep CNN based on learning feature reconstruction with feedback weights. IEEE Access 7:25195?25208
IUI (2019) Explainable smart systems. https://explainablesystems.comp.nus.edu.sg/2019/. Accessed 3 Mar 2020
Ivanovs M, Kadikis R, Ozols K (2021) Perturbation-based methods for explaining deep neural networks: a survey. Pattern Recognit Lett
Jagadish H, Gehrke J, Labrinidis A, Papakonstantinou Y, Patel JM, Ramakrishnan R, Shahabi C (2014) Big data and its technical challenges. Commun ACM 57(7):86?94
Janitza S, Celik E, Boulesteix AL (2018) A computationally fast variable importance test for random forests for high-dimensional data. Adv Data Anal Classif 12(4):885?915
Jung YJ, Han SH, Choi HJ (2021) Explaining CNN and RNN using selective layer-wise relevance propagation. IEEE Access 9:18670?18681
Junior JRB (2020) Graph embedded rules for explainable predictions in data streams. Neural Netw 129:174?192
Juuti M, Szyller S, Marchal S, Asokan N (2019) PRADA: protecting against DNN model stealing attacks. In: 2019 IEEE European symposium on security and privacy (EuroS&P). IEEE, pp 512?527
Kapelner A, Soterwood J, Nessaiver S, Adlof S (2018) Predicting contextual informativeness for vocabulary learning. IEEE Trans Learn Technol 11(1):13?26
Karlsson I, Rebane J, Papapetrou P, Gionis A (2020) Locally and globally explainable time series tweaking. Knowl Inf Syst 62(5):1671?1700
Keane MT, Kenny EM (2019) How case-based reasoning explains neural networks: A theoretical analysis of XAI using post-hoc explanation-by-example from a survey of ANN-CBR twin-systems. In: International conference on case-based reasoning. Springer, Cham, pp 155?171

Keneni BM, Kaur D, Al Bataineh A, Devabhaktuni VK, Javaid AY, Zaientz JD, Marinier RP (2019) Evolving rule-based explainable artificial intelligence for unmanned aerial vehicles. IEEE Access 7:17001?17016
Kenny EM, Ford C, Quinn M, Keane MT (2021) Explaining black-box classifiers using post-hoc explanations-by-example: the effect of explanations and error-rates in XAI user studies. Artif Intell 294:103459
Kim J, Canny J (2018) Explainable deep driving by visualizing causal attention. In: Explainable and interpretable models in computer vision and machine learning. Springer, Cham, pp 173?193
Kindermans PJ, Hooker S, Adebayo J, Alber M, Schütt KT, Dähne S, Erhan D, Kim B (2019) The (un) reliability of saliency methods. In: Explainable AI: interpreting, explaining and visualizing deep learning. Springer, Cham, pp 267?280
Kiritz N, Sarfati P (2018) Supervisory guidance on model risk management (SR 11-7) versus enterprise-wide model risk management for deposit-taking institutions (E-23): a detailed comparative analysis. Available at SSRN 3332484
Koh PW, Liang P (2017) Understanding black-box predictions via influence functions. In: International conference on machine learning. PMLR, pp 1885?1894
Kolyshkina I, Simoff S (2021) Interpretability of machine learning solutions in public healthcare: the CRISP-ML approach. Front Big Data 4:18
Konig R, Johansson U, Niklasson L (2008) G-REX: a versatile framework for evolutionary data mining. In: 2008 IEEE international conference on data mining workshops. IEEE, pp 971?974
Konstantinov AV, Utkin LV (2021) Interpretable machine learning with an ensemble of gradient boosting machines. Knowl Based Syst 222:106993
Krishnamurthy P, Sarmadi A, Khorrami F (2021) Explainable classification by learning human-readable sentences in feature subsets. Inf Sci 564:202?219
Kumari B, Swarnkar T (2020) Importance of data standardization methods on stock indices prediction accuracy. In: Advanced computing and intelligent engineering. Springer, Cham, pp 309?318
Kuo CCJ, Zhang M, Li S, Duan J, Chen Y (2019) Interpretable convolutional neural networks via feedforward design. J Vis Commun Image Represent 60:346?359
Langer M, Oster D, Speith T, Hermanns H, Kästner L, Schmidt E, Sesing A, Baum K (2021) What do we want from explainable artificial intelligence (XAI)??A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. Artif Intell 296:103473
Lapchak PA, Zhang JH (2018) Data standardization and quality management. Transl Stroke Res 9(1):4?8
Lapuschkin S, Binder A, Montavon G, Müller KR, Samek W (2016) The LRP toolbox for artificial neural networks. J Mach Learn Res 17(1):3938?3942
Latouche P, Robin S, Ouadah S (2018) Goodness of fit of logistic regression models for random graphs. J Comput Graph Stat 27(1):98?109
Lauritsen SM, Kristensen M, Olsen MV, Larsen MS, Lauritsen KM, Jørgensen MJ, Lange J, Thiesson B (2020) Explainable artificial intelligence model to predict acute critical illness from electronic health records. Nat Commun 11(1):1?11
Lawless WF, Mittu R, Sofge D, Hiatt L (2019) Artificial intelligence, autonomy, and human-machine teams: interdependence, context, and explainable AI. AI Mag 40(3)
Lee D, Mulrow J, Haboucha CJ, Derrible S, Shiftan Y (2019) Attitudes on autonomous vehicle adoption using interpretable gradient boosting machine. Transp Res Rec, p 0361198119857953
Li K, Hu C, Liu G, Xue W (2015) Building?s electricity consumption prediction using optimized artificial neural networks and principal component analysis. Energy Build 108:106?113
Liang S, Sabri AQM, Alnajjar F, Loo CK (2021) Autism spectrum self-stimulatory behaviours classification using explainable temporal coherency deep features and SVM classifier. IEEE Access
Liberati C, Camillo F, Saporta G (2017) Advances in credit scoring: combining performance and interpretation in kernel discriminant analysis. Adv Data Anal Classif 11(1):121?138
Lin YC, Lee YC, Tsai WC, Beh WK, Wu AYA (2020) Explainable deep neural network for identifying cardiac abnormalities using class activation map. In: 2020 Computing in cardiology. IEEE, pp 1?4
Lipton ZC (2018) The mythos of model interpretability. Queue 16(3):31?57
Liu YJ, Ma C, Zhao G, Fu X, Wang H, Dai G, Xie L (2016) An interactive spiraltape video summarization. IEEE Trans Multimed 18(7):1269?1282
Liu Z, Tang B, Wang X, Chen Q (2017) De-identification of clinical notes via recurrent neural network and conditional random field. J Biomed Inform 75:S34?S42
Liu P, Zhang L, Gulla JA (2020) Dynamic attention-based explainable recommendation with textual and visual fusion. Inf Process Manag 57(6):102099
Long M, Cao Y, Cao Z, Wang J, Jordan MI (2018) Transferable representation learning with deep adaptation networks. IEEE Trans Pattern Anal Mach Intell 41(12):3071?3085

Loor M, De Tré G (2020) Contextualizing support vector machine predictions. Int J Comput Intell Syst 13(1):1483?1497
Luo X, Chang X, Ban X (2016) Regression and classification using extreme learning machine based on L1-norm and L2-norm. Neurocomputing 174:179?186
Ma Y, Chen W, Ma X, Xu J, Huang X, Maciejewski R, Tung AK (2017) EasySVM: a visual analysis approach for open-box support vector machines. Comput Vis Media 3(2):161?175
Manica M, Oskooei A, Born J, Subramanian V, Sáez-Rodríguez J, Rodriguez Martinez M (2019) Toward explainable anticancer compound sensitivity prediction via multimodal attention-based convolutional encoders. Mol Pharm 16(12):4797?4806
Martini ML, Neifert SN, Gal JS, Oermann EK, Gilligan JT, Caridi JM (2021) Drivers of prolonged hospitalization following spine surgery: a game-theory-based approach to explaining machine learning models. JBJS 103(1):64?73
Maweu BM, Dakshit S, Shamsuddin R, Prabhakaran B (2021) CEFEs: a CNN explainable framework for ECG signals. Artif Intell Med 102059
Meske C, Bunde E, Schneider J, Gersch M (2020) Explainable artificial intelligence: objectives, stakeholders, and future research opportunities. Inf Syst Manag 1?11
Microsoft (2021) Azure model interpretability. https://docs.microsoft.com/en-us/azure/machine-learning/how-to-machine-learning-interpretability. Accessed 26 Apr 2021
Miller T (2019) Explanation in artificial intelligence: insights from the social sciences. Artif Intell 267:1?38
Minh DL, Sadeghi-Niaraki A, Huy HD, Min K, Moon H (2018) Deep learning approach for short-term stock trends prediction based on two-stream gated recurrent unit network. IEEE Access 6:55392?55404
Mohit, Kumari AC, Sharma M (2019) A novel approach to text clustering using shift k-medoid. Int J Soc Comput Cyber Phys Syst 2(2):106?118
Molnar C, Casalicchio G, Bischl B (2019) Quantifying model complexity via functional decomposition for better post-hoc interpretability. In: Joint European conference on machine learning and knowledge discovery in databases. Springer, Cham, pp 193?204
Montavon G, Lapuschkin S, Binder A, Samek W, Müller KR (2017) Explaining nonlinear classification decisions with deep Taylor decomposition. Pattern Recogn 65:211?222
Moradi M, Samwald M (2021) Post-hoc explanation of black-box classifiers using confident itemsets. Expert Syst Appl 165:113941
Mordvintsev A, Olah C, Tyka M (2015) Inceptionism: going deeper into neural networks, 2015. https://research.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html
Muller H, Mayrhofer MT, Van Veen EB, Holzinger A (2021) The ten commandments of ethical medical AI. Computer 54(07):119?123
Musto C, de Gemmis M, Lops P, Semeraro G (2020) Generating post hoc review-based natural language justifications for recommender systems. User Model User Adapt Interact 1?45
Neto MP, Paulovich FV (2020) Explainable matrix?visualization for global and local interpretability of random forest classification ensembles. IEEE Trans Vis Comput Graph
Ng SF, Chew YM, Chng PE, Ng KS (2018) An insight of linear regression analysis. Sci Res J 15(2):1?16
Nguyen TN, Lee S, Nguyen-Xuan H, Lee J (2019) A novel analysis-prediction approach for geometrically nonlinear problems using group method of data handling. Comput Methods Appl Mech Eng 354:506?526
Nguyen DT, Kasmarik KE, Abbass HA (2020a) Towards interpretable neural networks: an exact transformation to multi-class multivariate decision trees. arXiv preprint arXiv:200304675
Nguyen TN, Nguyen-Xuan H, Lee J (2020b) A novel data-driven nonlinear solver for solid mechanics using time series forecasting. Finite Elem Anal Des 171:103377
NIPS (2017) Interpreting, explaining and visualizing deep learning. http://www.interpretable-ml.org/nips2017workshop/. Accessed 3 Mar 2020
Obregon J, Kim A, Jung JY (2019) RuleCOSI: combination and simplification of production rules from boosted decision trees for imbalanced classification. Expert Syst Appl 126:64?82
Olah C, Satyanarayan A, Johnson I, Carter S, Schubert L, Ye K, Mordvintsev A (2018) The building blocks of interpretability. Distill 3(3):e10
Oracle (2021) Oracle skater. https://oracle.github.io/Skater/overview.html. Accessed 26 Apr 2021
Ostad-Ali-Askari K, Shayannejad M (2021) Computation of subsurface drain spacing in the unsteady conditions using artificial neural networks (ANN). Appl Water Sci 11(2):1?9
Ostad-Ali-Askari K, Shayannejad M, Ghorbanizadeh-Kharazi H (2017) Artificial neural network for modeling nitrate pollution of groundwater in marginal area of Zayandeh-rood river, Isfahan, Iran. KSCE J Civ Eng 21(1):134?140
Osullivan S, Nevejans N, Allen C, Blyth A, Leonard S, Pagallo U, Holzinger K, Holzinger A, Sajid MI, Ashrafian H (2019) Legal, regulatory, and ethical frameworks for development of standards in

artificial intelligence (AI) and autonomous robotic surgery. Int J Med Robotics Comput Assist Surg 15(1):e1968
Padarian J, McBratney AB, Minasny B (2020) Game theory interpretation of digital soil mapping convolutional neural networks. Soil 6(2):389?397
Páez A (2019) The pragmatic turn in explainable artificial intelligence (XAI). Mind Mach 29(3):441?459
Pan X, Tang F, Dong W, Ma C, Meng Y, Huang F, Lee TY, Xu C (2019) Content-based visual summarization for image collections. IEEE Transa Vis Comput Graph
Park DH, Hendricks LA, Akata Z, Rohrbach A, Schiele B, Darrell T, Rohrbach M (2018) Multimodal explanations: justifying decisions and pointing to the evidence. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 8779?8788
Payer C, Stern D, Bischof H, Urschler M (2019) Integrating spatial configuration into heatmap regression based CNNs for landmark localization. Med Image Anal 54:207?219
Peloquin D, DiMaio M, Bierer B, Barnes M (2020) Disruptive and avoidable: GDPR challenges to secondary research uses of data. Eur J Hum Genet 28(6):697?705
Polato M, Aiolli F (2019) Boolean kernels for rule based interpretation of support vector machines. Neurocomputing 342:113?124
PR (2019) Explainable deep learning for efficient and robust pattern recognition. https://www.journals.elsevier.com/pattern-recognition/call-for-papers/call-for-paper-on-special-issue-on-explainable-deep-learning. Accessed 3 Mar 2020
Raaijmakers S (2019) Artificial intelligence for law enforcement: challenges and opportunities. IEEE Secur Priv 17(5):74?77
Rai A (2020) Explainable AI: from black box to glass box. J Acad Mark Sci 48(1):137?141
Rajapaksha D, Bergmeir C, Buntine W (2020) LoRMIkA: local rule-based model interpretability with k-optimal associations. Inf Sci 540:221?241
Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, Liu PJ, Liu X, Marcus J, Sun M et al (2018) Scalable and accurate deep learning with electronic health records. NPJ Digit Med 1(1):1?10
Ren X, Xing Z, Xia X, Lo D, Wang X, Grundy J (2019) Neural network-based detection of self-admitted technical debt: from performance to explainability. ACM Trans Softw Eng Methodol (TOSEM) 28(3):1?45
Ribeiro MT, Singh S, Guestrin C (2016) ?Why should I trust you?? explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp 1135?1144
Ribeiro PC, Schardong GG, Barbosa SD, de Souza CS, Lopes H (2019) Visual exploration of an ensemble of classifiers. Comput Graph 85:23?41
Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 1(5):206?215
Sabol P, Sinčák P, Hartono P, Kočan P, Benetinová Z, Blichárová A, Verbóová Ľ, Štammová E, Sabolová-Fabianová A, Jašková A (2020) Explainable classifier for improving the accountability in decision-making for colorectal cancer diagnosis from histopathological images. J Biomed Inform 109:103523
Sagi O, Rokach L (2020) Explainable decision forest: transforming a decision forest into an interpretable tree. Inf Fusion 61:124?138
Salmeron JL, Correia MB, Palos-Sanchez PR (2019) Complexity in forecasting and predictive models. Complexity 2019
Sanz H, Valim C, Vegas E, Oller JM, Reverter F (2018) SVM-RFE: selection and visualization of the most relevant features through non-linear kernels. BMC Bioinform 19(1):1?18
Sarvghad A, Tory M, Mahyar N (2016) Visualizing dimension coverage to support exploratory analysis. IEEE Trans Visual Comput Graph 23(1):21?30
Schneeberger D, Stöger K, Holzinger A (2020) The European legal framework for medical AI. In: International cross-domain conference for machine learning and knowledge extraction. Springer, Cham, pp 209?226
Self JZ, Dowling M, Wenskovitch J, Crandell I, Wang M, House L, Leman S, North C (2018) Observation-level and parametric interaction for high-dimensional data analysis. ACM Trans Interact Intell Syst (TIIS) 8(2):1?36
Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2020) Grad-cam: visual explanations from deep networks via gradient-based localization. Int J Comput Vis 128(2):336?359
Setzu M, Guidotti R, Monreale A, Turini F, Pedreschi D, Giannotti F (2021) Glocalx-from local to global explanations of black box AI models. Artif Intell 294:103457
Shi L, Teng Z, Wang L, Zhang Y, Binder A (2018) Deepclue: visual interpretation of text-based deep stock prediction. IEEE Trans Knowl Data Eng 31(6):1094?1108

Shrikumar A, Greenside P, Kundaje A (2017) Learning important features through propagating activation differences. In: International conference on machine learning. PMLR, pp 3145?3153
Singh N, Singh P, Bhagat D (2019) A rule extraction approach from support vector machines for diagnosing hypertension among diabetics. Expert Syst Appl 130:188?205
Singh A, Sengupta S, Lakshminarayanan V (2020) Explainable deep learning models in medical image analysis. J Imaging 6(6):52
Song S, Huang H, Ruan T (2019) Abstractive text summarization using LSTM-CNN based deep learning. Multimed Tools Appl 78(1):857?875
SP (2019) Explainable AI on emerging multimedia technologies. https://www.journals.elsevier.com/signal-processing-image-communication/call-for-papers/emerging-multimedia-technologies. Accessed 3 Mar 2020
Spinner T, Schlegel U, Schäfer H, El-Assady M (2019) explAIner: a visual analytics framework for interactive and explainable machine learning. IEEE Trans Vis Comput Graph 26(1):1064?1074
Statista (2020) Revenues from the artificial intelligence software market worldwide from 2018 to 2025. https://www.statista.com/statistics/607716/worldwide-artificial-intelligence-market-revenues/. Accessed 24 Jan 2020
Stojić A, Stanić N, Vuković G, Stanišić S, Perišić M, Šoštarić A, Lazić L (2019) Explainable extreme gradient boosting tree-based prediction of toluene, ethylbenzene and xylene wet deposition. Sci Total Environ 653:140?147
Strobelt H, Gehrmann S, Pfister H, Rush AM (2017) Lstmvis: a tool for visual analysis of hidden state dynamics in recurrent neural networks. IEEE Trans Vis Comput Graph 24(1):667?676
Strobelt H, Gehrmann S, Behrisch M, Perer A, Pfister H, Rush AM (2018) SEQ2SEQ-VIS: a visual debugging tool for sequence-to-sequence models. IEEE Trans Vis Comput Graph 25(1):353?363
Štrumbelj E, Kononenko I (2014) Explaining prediction models and individual predictions with feature contributions. Knowl Inf Syst 41(3):647?665
Su J, Vargas DV, Sakurai K (2019) One pixel attack for fooling deep neural networks. IEEE Trans Evol Comput 23(5):828?841
Swartout WR, Moore JD (1993) Explanation in second generation expert systems. In: Second generation expert systems. Springer, Cham, pp 543?585
Tan Q, Ye M, Ma AJ, Yang B, Yip TCF, Wong GLH, Yuen PC (2020) Explainable uncertainty-aware convolutional recurrent neural network for irregular medical time series. IEEE Trans Neural Netw Learn Syst
Tjoa E, Guan C (2020) A survey on explainable artificial intelligence (XAI): Toward medical XAI. IEEE Trans Neural Netw Learn Syst
Turkay C, Kaya E, Balcisoy S, Hauser H (2016) Designing progressive and interactive analytics processes for high-dimensional data analysis. IEEE Trans Vis Comput Graph 23(1):131?140
UberAccident (2020) What happens when self-driving cars kill people. https://www.forbes.com/sites/cognitiveworld/2019/09/26/what-happens-with-self-driving-cars-kill-people/#4b798bcc405c. Accessed 17 Mar 2020
Van Belle V, Van Calster B, Van Huffel S, Suykens JA, Lisboa P (2016) Explaining support vector machines: a color based nomogram. PLoS ONE 11(10):e0164568
Van Lent M, Fisher W, Mancuso M (2004) An explainable artificial intelligence system for small-unit tactical behavior. In: Proceedings of the national conference on artificial intelligence. AAAI Press; MIT Press, Menlo Park, London, pp 900?907
Van Luong H, Joukovsky B, Deligiannis N (2021) Designing interpretable recurrent neural networks for video reconstruction via deep unfolding. IEEE Trans Image Process 30:4099?4113
Veale M, Binns R, Edwards L (2018) Algorithms that remember: model inversion attacks and data protection law. Philos Trans Royal Soc A Math Phys Eng Sci 376(2133):20180083
Vellido A (2019) The importance of interpretability and visualization in machine learning for applications in medicine and health care. Neural Comput Appl 1?15
Waa J, Nieuwburg E, Cremers A, Neerincx M (2021) Evaluating XAI: a comparison of rule-based and example-based explanations. Artif Intell 291:103404
Wachter S, Mittelstadt B, Floridi L (2017) Why a right to explanation of automated decision-making does not exist in the general data protection regulation. Int Data Privacy Law 7(2):76?99
Wang SC (2003) Artificial neural network. In: Interdisciplinary computing in java programming. Springer, Cham, pp 81?100
Wang B, Gong NZ (2018) Stealing hyperparameters in machine learning. In: 2018 IEEE symposium on security and privacy (SP). IEEE, pp 36?52
Wang H, Yeung DY (2016) Towards Bayesian deep learning: a framework and some existing methods. IEEE Trans Knowl Data Eng 28(12):3395?3408

Wang Y, Aghaei F, Zarafshani A, Qiu Y, Qian W, Zheng B (2017) Computer-aided classification of mammographic masses using visually sensitive image features. J Xray Sci Technol 25(1):171?186

Wang Q, Zhang K, Ororbia AG II, Xing X, Liu X, Giles CL (2018) An empirical evaluation of rule extraction from recurrent neural networks. Neural Comput 30(9):2568?2591

Wang C, Shi Y, Fan X, Shao M (2019a) Attribute reduction based on k-nearest neighborhood rough sets. Int J Approx Reason 106:18?31

Wang F, Kaushal R, Khullar D (2019b) Should health care demand interpretable artificial intelligence or accept ?black box? medicine? Ann Intern Med

Wang S, Zhou T, Bilmes J (2019c) Bias also matters: bias attribution for deep neural network explanation. In: International conference on machine learning. PMLR, pp 6659?6667

Wang Y, Wang D, Geng N, Wang Y, Yin Y, Jin Y (2019d) Stacking-based ensemble learning of decision trees for interpretable prostate cancer detection. Appl Soft Comput 77:188?204

Wasilow S, Thorpe JB (2019) Artificial intelligence, robotics, ethics, and the military: a Canadian perspective. AI Mag 40(1)

Weitz K, Schiller D, Schlagowski R, Huber T, André E (2020) ?Let me explain!?: exploring the potential of virtual agents in explainable AI interaction design. J Multimodal User Interfaces 1?12

Wickstrøm KK, ØyvindMikalsen K, Kampffmeyer M, Revhaug A, Jenssen R (2020) Uncertainty-aware deep ensembles for reliable and explainable predictions of clinical time series. IEEE J Biomed Health Inform

Williford JR, May BB, Byrne J (2020) Explainable face recognition. In: European Conference on computer vision. Springer, Cham, pp 248?263

Wu Q, Burges CJ, Svore KM, Gao J (2010) Adapting boosting for information retrieval measures. Inf Retr 13(3):254?270

Wu J, Zhong Sh, Jiang J, Yang Y (2017) A novel clustering method for static video summarization. Multimed Tools Appl 76(7):9625?9641

Wu M, Hughes M, Parbhoo S, Zazzi M, Roth V, Doshi-Velez F (2018) Beyond sparsity: tree regularization of deep models for interpretability. In: Proceedings of the AAAI conference on artificial intelligence, vol 32

Xu J, Zhang Z, Friedman T, Liang Y, Broeck G (2018) A semantic loss function for deep learning with symbolic knowledge. In: International conference on machine learning. PMLR, pp 5502?5511

Yamamoto Y, Tsuzuki T, Akatsuka J, Ueki M, Morikawa H, Numata Y, Takahara T, Tsuyuki T, Tsutsumi K, Nakazawa R et al (2019) Automated acquisition of explainable knowledge from unannotated histopathology images. Nat Commun 10(1):1?9

Yang SCH, Shafto P (2017) Explainable artificial intelligence via Bayesian teaching. In: NIPS 2017 workshop on teaching machines, robots, and humans, pp 127?137

Yang Z, Zhang A, Sudjianto A (2020) Enhancing explainability of neural networks through architecture constraints. IEEE Trans Neural Netw Learn Syst

Yeganejou M, Dick S, Miller J (2019) Interpretable deep convolutional fuzzy classifier. IEEE Trans Fuzzy Syst 28(7):1407?1419

Yosinski J, Clune J, Nguyen A, Fuchs T, Lipson H (2015) Understanding neural networks through deep visualization. arXiv preprint arXiv:150606579

Yousefi-Azar M, Hamey L (2017) Text summarization using unsupervised deep learning. Expert Syst Appl 68:93?105

Yu H, Yang S, Gu W, Zhang S (2017) Baidu driving dataset and end-to-end reactive control model. In: 2017 IEEE intelligent vehicles symposium (IV). IEEE, pp 341?346

Yuan J, Xiong HC, Xiao Y, Guan W, Wang M, Hong R, Li ZY (2020) Gated CNN: Integrating multi-scale feature layers for object detection. Pattern Recogn 105:107131

Zeltner D, Schmid B, Csiszár G, Csiszár O (2021) Squashing activation functions in benchmark tests: towards a more explainable artificial intelligence using continuous-valued logic. Knowl Based Syst 218:106779

Zhang Qs, Zhu SC (2018) Visual interpretability for deep learning: a survey. Fronti Inf Technol Electron Eng 19(1):27?39

Zhang J, Wang Y, Molino P, Li L, Ebert DS (2018a) Manifold: a model-agnostic framework for interpretation and diagnosis of machine learning models. IEEE Trans Vis Comput Graph 25(1):364?373

Zhang Q, Nian Wu Y, Zhu SC (2018b) Interpretable convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 8827?8836

Zhang Q, Yang Y, Ma H, Wu YN (2019) Interpreting CNNs via decision trees. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 6261?6270

Zhang A, Teng L, Alterovitz G (2020a) An explainable machine learning platform for pyrazinamide resistance prediction and genetic feature identification of mycobacterium tuberculosis. J Am Med Inform Assoc
Zhang M, You H, Kadam P, Liu S, Kuo CCJ (2020b) Pointhop: an explainable machine learning method for point cloud classification. IEEE Trans Multimed 22(7):1744?1755
Zhang W, Tang S, Su J, Xiao J, Zhuang Y (2020c) Tell and guess: cooperative learning for natural image caption generation with hierarchical refined attention. Multimed Tools Appl 1?16
Zhang Z, Beck MW, Winkler DA, Huang B, Sibanda W, Goyal H et al (2018c) Opening the black box of neural networks: methods for interpreting neural network models in clinical applications. Ann Transl Med 6(11)
Zhao W, Du S (2016) Spectral-spatial feature extraction for hyperspectral image classification: a dimension reduction and deep learning approach. IEEE Trans Geosci Remote Sens 54(8):4544?4554
Zheng S, Ding C (2020) A group lasso based sparse KNN classifier. Pattern Recogn Lett 131:227?233
Zheng Xl, Zhu My, Li Qb, Chen Cc, Tan Yc (2019) FinBrain: when finance meets AI 2.0. Front Inf Technol Electron Eng 20(7):914?924
Zhou B, Bau D, Oliva A, Torralba A (2018a) Interpreting deep visual representations via network dissection. IEEE Trans Pattern Anal Mach Intell 41(9):2131?2145
Zhou X, Jiang P, Wang X (2018b) Recognition of control chart patterns using fuzzy SVM with a hybrid kernel function. J Intell Manuf 29(1):51?67
Zhuang Yt, Wu F, Chen C, Pan Yh (2017) Challenges and opportunities: from big data to knowledge in AI 2.0. Front Inf Technol Electron Eng 18(1):3?14

Authors and Affiliations

Dang Minh[1] · H. Xiang Wang[2] · Y. Fen Li[2] · Tan N. Nguyen[3]

H. Xiang Wang
hanxiang@sju.ac.kr

Y. Fen Li
1826535091@sju.ac.kr

[1] Department of Information Technology, FPT University, Ho Chi Minh City, Vietnam

[2] Department of Computer Science and Engineering, Sejong University, 209 Neungdong-ro, Gwangjin-gu, Seoul 05006, Republic of Korea

[3] Department of Architectural Engineering, Sejong University, 209 Neungdong-ro, Gwangjin-gu, Seoul 05006, Republic of Korea