Human-Centric Intelligent Systems (2023) 3:161–188
https://doi.org/10.1007/s44230-023-00038-y

REVIEW ARTICLE

Survey on Explainable AI: From Approaches, Limitations and Applications Aspects

Wenli Yang[1] · Yuchen Wei[1] · Hanyu Wei[1] · Yanyu Chen[1] · Guan Huang[1] · Xiang Li[1] · Renjie Li[1] · Naimeng Yao[1] · Xinyi Wang[1] · Xiaotong Gu[1] · Muhammad Bilal Amin[1] · Byeong Kang[1]

Received: 16 March 2023 / Accepted: 24 July 2023 / Published online: 10 August 2023
© The Author(s) 2023

Abstract

In recent years, artificial intelligence (AI) technology has been used in most if not all domains and has greatly benefited our lives. While AI can accurately extract critical features and valuable information from large amounts of data to help people complete tasks faster, there are growing concerns about the non-transparency of AI in the decision-making process. The emergence of explainable AI (XAI) has allowed humans to better understand and control AI systems, which is motivated to provide transparent explanations for the decisions made by AI. This article aims to present a comprehensive overview of recent research on XAI approaches from three well-defined taxonomies. We offer an in-depth analysis and summary of the status and prospects of XAI applications in several key areas where reliable explanations are urgently needed to avoid mistakes in decision-making. We conclude by discussing XAI's limitations and future research directions.

Keywords Explainable AI · Machine learning · Human-centered · Survey

Abbreviations

AI	Artificial intelligence
XAI	Explainable artificial intelligence
GDPR	General data protection regulation
SO	Source-oriented
RO	Representation-oriented
CNN	Convolutional neural network
NNKX	Neural network knowledge extraction
REFNE	Rule extraction from neural network ensemble
ERE	Electric rule extraction
KG	Knowledge graph
LO	Logic-oriented
CT	Computed tomography
MRI	Magnetic resonance imaging
US	Ultrasound
Grad-CAM	Gradient-weighted class activation mapping
LIME	Locally-interpretable model-agnostic explanations
SHAP	Shapley additive explanations
CMGE	Counterfactual multi-granularity graph supporting fact extraction

✉ Yuchen Wei
yuchen.wei@utas.edu.au

Wenli Yang
yang.wenli@utas.edu.au

Hanyu Wei
hanyu.wei@utas.edu.au

Yanyu Chen
yanyu.chen@utas.edu.au

Guan Huang
guan.huang@utas.edu.au

Xiang Li
xiang.li@utas.edu.au

Renjie Li
renjie.li@utas.edu.au

Naimeng Yao
naimeng.yao@utas.edu.au

Xinyi Wang
xinyi.wang@utas.edu.au

Xiaotong Gu
xiaotong.gu@utas.edu.au

Muhammad Bilal Amin
bilal.amin@utas.edu.au

Byeong Kang
byeong.kang@utas.edu.au

[1] School of ICT, University of Tasmania, Hobart, TAS 7005, Australia

EHR Electronic health records
LEMNA Local explanation method using nonlinear approximation

1 Introduction

Deep learning has been contributing to artificial intelligence (AI) systems to speed up and improve numerous tasks, including decision-making, predictions, identifying anomalies and patterns, and even recommendations and so on. Although the accuracy of deep learning models has dramatically improved during the last decade, this improved accuracy has often been achieved through increased model complexity, which may induce common sense mistakes in practice without providing any reasons for the mistakes, making it impossible to fully trust its decisions. It's also challenging to achieve targeted model improvement and optimisation [1]. Without reliable explanations that accurately represent the current AI system processes, humans still consider AI untrustworthy due to a variety of dynamics and uncertainties [2] when deploying AI applications in real-world environments. This motivates the inherent need and expectation from human users that AI systems should be explainable to help confirm decisions.

Explainable AI (explainable artificial intelligence (XAI)) is often considered a set of processes and methods that are used to describe deep learning models, by characterizing model accuracy, transparency, and outcomes in AI systems [3]. XAI methods aim to provide human-readable explanations to help users comprehend and trust the outputs created by deep learning algorithms. Additionally, some regulations such as European General Data Protection Regulation (general data protection regulation (GDPR))[4] have been introduced to drive further XAI research, demanding the important ethics [5], justifications [6], trust [7] and bias [8] to explore reliable XAI solutions.

The need for XAI is multi-factorial and depends on the concerned people (Table 1), whether they are end users, AI developers or product engineers. End-users need to trust the decisions and be reassured based on the explainable process and feedback. On the other hand, AI developers need to understand the limitations of current models to validate and improve future versions. Besides, regarding product engineers in different domains, they need to access and optimise explanations of the decision process for the deployment of AI systems, especially in real-world environments.

In a more detailed manner, XAI should consider different cultural and contextual factors. For example, in different contexts and cultural backgrounds, XAI may need to provide different interpretations for the same objects and phenomena. To address this, scholars have proposed the Contextual Utility Theory [9], which explains the final decision outcome by assessing the importance and influence of different factors. Additionally, tailoring explanations based on user expertise is another crucial aspect of designing effective explainable artificial intelligence (XAI) systems. By considering the varying levels of technical knowledge and expertise among users, XAI systems can provide explanations that are better suited to their individual needs. For example, in healthcare, patients often have varying levels of medical knowledge and technical understanding. When presenting AI-driven diagnoses or treatment recommendations to patients, explanations should be tailored to their level of health literacy. For patients with limited medical expertise, explanations should use plain language and visual aids to help them comprehend the reasoning behind the AI-generated recommendations. On the other hand, for healthcare professionals who have a deeper understanding of medical concepts, explanations can delve into more technical details and provide insights into the model's decision-making process [10].

From a long-term perspective, more focus should be done on usability and maintainability, which requires improved personalisation, evolution with time, and data management [11]. These aspects are essential for ensuring the continued effectiveness and relevance of XAI approaches. One area that requires attention is improved personalization, where XAI systems can be tailored to individual users or specific application domains [12]. Also, as AI models and data evolve over time, XAI systems need to adapt and evolve as well. Data management is another critical aspect for the long-term usability and maintainability of XAI systems. As data volumes increase and data distributions change, XAI methods should be able to handle shifting data characteristics and adapt accordingly [13].

At present, XAI has gained a great deal of attention across different application domains. Accordingly, an increasing number of XAI tools and approaches are being introduced

Table 1 Explainable AI: who need it? Why? For what?

Who?	Why?	For what?
End user	Understand decisions	Persuasive explanation
AI developer	Understand limitations; improve future visions; debug algorithms	Intrinsic explanation; training and validation
Product engineer	System design, integration, and deployment	Complete explanation

both in industry and academia. These advancements aim to address the ongoing trade-off between interpretability and predictive power.There is a growing recognition that highly interpretable models might encounter limitations in capturing complex relationships, which can lead to reduced accuracy. On the other hand, complex models often achieve superior accuracy but at the expense of interpretability. Balancing these considerations becomes crucial and is contingent upon the specific requirements of the application domain. In certain contexts, such as healthcare or finance, interpretability and transparency play pivotal roles in ensuring regulatory compliance and addressing ethical considerations [14, 15]. In other domains, such as medical image or signal recognition, where accuracy is paramount, the focus may be more on predictive power than interpretability [16].

The current XAI methods exhibit various dimensions and descriptions to understand deep learning models and some survey papers [3, 17, 18] have summarized the methods and basic differences among different XAI approaches. However, the state-of-the-art analysis with respect to existing approaches and limitations for different XAI-enabled application domains still lacks investigation.

The field of explainable artificial intelligence (XAI) has witnessed the emergence of numerous methods and techniques aimed at comprehending the intricate workings of deep learning models. Currently, some survey papers have made efforts to summarize these methods and offer a fundamental understanding of the distinctions among various XAI approaches [3, 17, 18]. However, while certain survey papers have focused on specific domains like healthcare [19] or medical applications [20], there still exists a substantial gap in the state-of-the-art analysis pertaining to the existing approaches and their limitations across all XAI-enabled application domains. This gap necessitates a comprehensive investigation encompassing various aspects such as different requirements, suitable XAI approaches, and domain-specific limitations. Conducting such an analysis is crucial as it allows us to gain a deeper understanding of the performance of XAI techniques in real-world scenarios. Additionally, it helps us identify the challenges and opportunities that arise when applying these approaches to different application domains. By bridging this gap, we can make significant strides towards developing more effective and reliable XAI systems tailored to specific domains and their unique characteristics.

In this survey, our primary objective is to provide a comprehensive overview of explainable artificial intelligence (XAI) approaches across various application domains by exploring and analysing the different methods and techniques employed in XAI and their application-specific considerations. We achieve this by utilizing three well-defined taxonomies, as depicted in Fig. 1. Unlike many existing surveys that solely focus on reviewing and comparing methods, we go beyond that by providing domain mapping. This mapping provides insights into how XAI methods are interconnected and utilized across various application domains, and even in cases where domains intersect. Additionally, we delve into a detailed discussion on the limitations of the existing methods, acknowledging the areas where further improvements are necessary. Lastly, we summarize the future directions in XAI research, highlighting potential avenues for advancements and breakthroughs. Our contributions in this survey can be summarized as follows:

- Develop a new taxonomy for the description of XAI approaches based on three well-defined orientations with a wider range of explanation options;
- Investigate and examine various XAI-enabled applications to identify the available XAI techniques and domain insights through case studies;

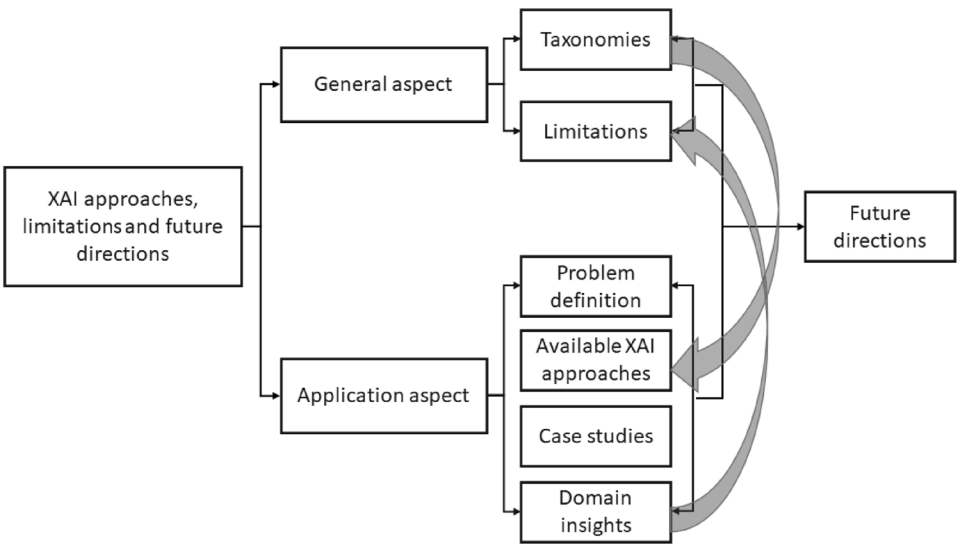

Fig. 1 The proposed organization to discuss the approaches, limitations and future directions in XAI

- Discuss the limitations and gaps in the design of XAI methods for the future directions of research and development.

In order to comprehensively analyze XAI approaches, limitations, and future directions from application perspectives, our survey is structured around two main themes, as depicted in Fig. 1. The first theme focuses on general approaches and limitations in XAI, while the second theme aims to analyze the available XAI approaches and domain-specific insights.

Under each domain, we explore four main sub-themes: problem definition, available XAI approaches, case studies, and domain insights. Before delving into each application domain, it is important to review the general taxonomies of XAI approaches. This provides a foundation for understanding and categorizing the various XAI techniques. In each domain, we discuss the available and suitable XAI approaches that align with the proposed general taxonomies of XAI approaches. Additionally, we examine the domain-specific limitations and considerations, taking into account the unique challenges and requirements of each application area. We also explore cross-disciplinary techniques that contribute to XAI innovations. The findings from these discussions are summarized as limitations and future directions, providing valuable insights into current research trends and guiding future studies in the field of XAI.

2 Taxonomies of XAI Approaches

2.1 Review Scope and Execution

This work is mainly based on a scope of review refers to the specific boundaries and focus of the research being conducted. In the context of an XAI survey, the scope typically includes the following aspects:

- XAI approaches: The review will focus on examining and analyzing different XAI approaches and methods that have been proposed in the literature. This include visualization techniques, symbolic explanations, ante-hoc explanations, post-hoc explanations, local explanations, global explanations and any other relevant techniques.
- Application domains: The review may consider various application domains where XAI techniques have been applied, including medical and biomedical, healthcare, finance, law, cyber security, education and training, civil engineering. The scope involve exploring the usage of XAI techniques in these domains and analyzing their effectiveness and limitations across multiple domains.
- Research papers: The review will involve studying and synthesizing research papers that are relevant to the chosen scope. These papers may include original research articles, survey papers and scholarly publications that contribute to the understanding of XAI approaches and their application in the selected domains through case studies.
- Limitations and challenges: The scope also encompass examining the limitations and challenges of existing XAI methods and approaches. This could involve identifying common issues, gaps in the literature, and areas that require further research or improvement.

Having the scope of review established, the selected databases and a search engine include Scopus, Web of Science and Google Scholar (Search engine) and arXiv between 2013 and 2023. The search terms based on the scopes are:

- XAI keywords: explainable, XAI, interpretable.
- Review keywords: survey, review, overview, literature, bibliometric, challenge, prospect, trend, insight, opportunity, future direction.
- Domain keywords: medical, biomedical, healthcare, wellness, civil, urban, transportation, cyber security, information security, education, training, learning and teaching, coaching, finance, economics, law, legal system.

With the selected search terms, the two-round search strings were designed to effectively retrieve relevant information and narrow down the search results.

The first round, focusing on general research papers, consisted of the following search string: (explainable OR XAI OR interpretable) AND (survey OR review OR overview OR literature OR bibliometric OR challenge OR prospect OR trend OR opportunity OR "future direction").

The second round, aimed at selecting specific application domains, utilized the following search string: (explainable OR XAI OR interpretable) AND (medical, biomedical OR healthcare OR wellness OR civil OR urban OR transportation OR “cyber security” OR “information security” OR education OR training OR “learning and teaching” OR coaching OR finance OR economics OR law OR “legal system”).

Publications that did not clearly align with the scopes based on their title or abstract were excluded from this review. While not all literature explicitly stated this information, the extracted data was organized and served as the foundation for our analysis.

2.2 XAI Approaches

The taxonomies in the existing survey papers generally categorised XAI approaches based on scope (local or global) [21], stage (ante-hoc or post-hoc) [17] and output format (numerical, visual, textual or mixed) [22]. The main

difference between the existing study and our survey is that this paper focuses on the human perspective involving source, representation, and logic reasoning. We summarise the taxonomies categorised in this survey in Fig. 2:

Source-oriented (source-oriented (SO)) the sources that support building explanations can be either subjective (S) or objective (O) cognition, depending on whether the explanations are provided based on the fact or human experience. For example, in the medical field, if the explanation of a diagnosis is provided based on the patient's clinical symptoms and explains the cause and pathology in detail during the AI learning process, this is from objective cognitive concern. In contrast, explanations with subjective cognitive consider patients' current physical conditions and doctors' medical knowledge.

Representation-oriented (representation-oriented (RO)) core representation among the XAI approaches can generally be classified into visualisation-based (V), symbolic-based (S) or even hybrid (H) methods. Visual-based methods are the most common representation ways including input visualisation and model visualisation. Input visualisation methods provide an accessible way to view and understand how input data affect model outputs, while model visualisation methods provide analysis based on the aspect of layers or features inside the model.

Besides visualization-based methods, other formats of explanations, including numerical, graphical, rules, and textual explanations, are covered in symbolic-based methods. Symbolic-based methods tend to describe the process of deep learning models by extracting insightful information, such as meaning and context, and representing them in different formats. The coding symbolic-based explanation is provided directly from the factual features, including numerical, graphical and textual explanations. For instance, a numerical method [23] monitors the features at every layer of a neural network model and measures their suitability for classification, which is beneficial to provide a better understanding of the roles and dynamics of the intermediate layers. Zhang et al. [24] used an explanatory graph to reveal the knowledge hierarchy hidden inside a pre-trained convolutional neural network (CNN) model. In the graph, each node represents a part pattern, and each edge encodes co-activation relationships and spatial relationships between patterns. A phase-critic model is developed to generate a candidate textual explanation for the input image [25].

By contrast, qualifying symbolic explanations, such as rules and graphical explanations, are provided under human knowledge. Rules explanations are usually in the form of "If-Then" rules to obtain the inferential process of neural networks. Rule extraction techniques include neural network knowledge extraction (neural network knowledge extraction (NNKX)) [26], rule extraction from neural network ensemble (rule extraction from neural network ensemble (REFNE)) [27], and Electric Rule Extraction (electric rule extraction (ERE)) [28]. Moreover, knowledge graphs can also help AI models be more explainable and interpretable and are widely used in explainable methods. Andriy and Mathieu [29] applied rule mining using knowledge graph (KG)s to reveal semantic bias in neural network models. Visualisation-based methods and coding symbolic explanations generally belong to objective cognitive, while qualifying symbolic explanations are always subjective cognitive by adding human knowledge and opinions.

Hybrid methods can generate mixed explanations, consisting of visualization explanations and symbolic information, which can be either subjective or objective cognitive. In [30], visual and textual explanations are employed in the visual question answering task. Both subjective and objective cognitive explanations are provided in this work. We

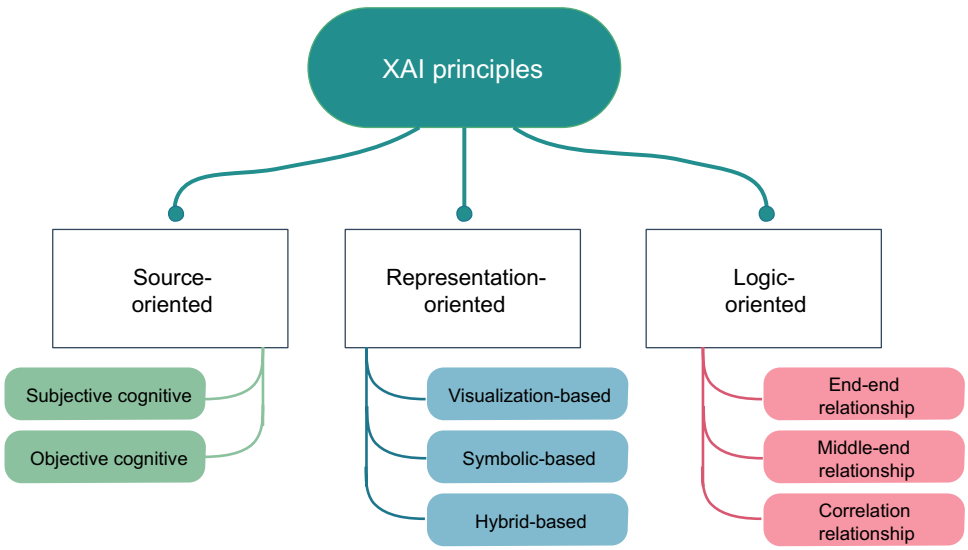

Fig. 2 Taxonomies of XAI approaches in this survey

summarise the above-mentioned XAI representations with surveyed related work in Table 2.

Logic-oriented (logic-oriented (LO)) a well-developed XAI method with a specific representation can be integrated logic reasoning into deep learning models, including end-end (E-E) relationship, middle-end (M-E) relationship, and correlation (Corr) relationship shown in Table 3. The end-end explanations focus on providing how the AI system processes from the first input stage to the final output result. The middle-end performs explanations by considering the internal AI system structure. The correlation is used to represent the correlations among sequence inputs or consecutive outputs of deep learning models. Most XAI approaches target to clarify the relationship between input features and output results, and very few research relates to middle-end and correlation relationships.

3 Applications Using XAI Approaches

Nowadays, applications using XAI approaches have been covered in various domains. This section provides details for different XAI techniques used for each application. Some of the main applications are summarised in Table 4.

3.1 Medical and Biomedical

3.1.1 Problem Definition

The use of AI systems in medical and biomedical research has increasing influences on medical advice and therapeutic decision-making processes, especially in one of the most common areas that apply deep learning techniques, the medical and biomedical image analysis, such as image registration and localisation, detection of anatomical and cellular structures, tissue segmentation, and computer-aided disease diagnosis.

Medical and biomedical image analysis refers to the extraction of meaningful information from digital images, which utilised a variety of medical imaging techniques, including computed tomography (computed tomography (CT)), magnetic resonance imaging (Magnetic resonance imaging (MRI)), ultrasound (ultrasound (US)), and X-rays, covering important body parts such as the brain, heart, breast, lung and kidney [153]. Due to the advantages of deep learning in the field of medical image analysis, in recent years, more and more researchers have adopted deep learning to solve problems of medical image analysis and achieved good performances. Although medical image analysis based on deep learning has made great progress, it still faces some urgent problems in medical practice.

Deep learning methods now can automatically extract abstract features by end-to-end prediction processing, which can obtain direct results, but this is insufficient to provide diagnostic evidence and pathology, and the features cannot be completely trusted or accepted. For example, for glaucoma screening, doctors can use intraocular pressure testing, visual field testing, and manual inspection of the optic disc to diagnose the disease and give the cause and pathology based on the patient's symptoms and pathological reports [154]. However, the deep learning model is difficult to explain the correlation or causality between its input and output in different contexts, lacking an explanation of the process, which is difficult to support reasoning in medical diagnosis and research.

Additionally, due to the data-driven nature of deep learning, models can easily learn the deviations in the training [155]. This phenomenon is common in medical image processing. For example, a deep learning model identifies certain diseases from images during training while the actual diagnosis could be another disease, and this should not occur at all. If users intend to to improve a model, the explanation of the model is a prerequisite, because before being able to solve a problem, its existence and causality need to be identified.

3.1.2 XAI Based Proposals

The research on explainable deep learning can enhance the capabilities for AI-assisted diagnosis by integrating with large-scale medical systems, providing an effective and interactive way to promote medical intelligence. Different from common explainable deep learning methods, the deep learning explanation research of medical image analysis is not only affected by the data, but also related to medical experts' knowledge.

In terms of source-oriented, the objective cogitative explanation is provided based on visible, measurable findings obtained by medical examinations, tests, or images, while the subjective cogitative explanation needs to consider the medical experts' knowledge and patient situations. The existing XAI proposals cover both objective and subjective cognitive aspects. For example, the gradient-weighted class activation mapping [gradient-weighted class activation mapping (Grad-CAM)] method proposed in [36] performs explanation by highlighting the important regions in the image, which refers to objective cogitative. Some researchers also consider using subjective sources, such as in [85], authors presented the explanation by combining time series, histopathological images, knowledge databases as well as patient histories.

In terms of representation-oriented, visualisation methods emphasise the visualisation of training data rules and the visualisation inside the model, which is the most popular XAI

Table 2 Summarising the XAI approaches from representation-oriented

Type	Sub-type	Typical examples	Pros	Cons	References
Visualization-based	Model-agnostic visualization	Attribute-based; attention-based, perturbation-based; CAM-based; counterfactual interpretation, etc.	The outputs are simple to understand because there is no requirement to comprehend the internal model structure	It only offers a user-friendly interface for seeing and comprehending input data and does not provide information on the context of the model learning process	[31–35]
	Model-specific visualization	Concept attribution; dimensionality reduction; graph neural network explanation, etc.	The explanation can uncover the information that is hidden inside the model	The features used to determine interpretability might not be repeatable among different models	[36, 37]
Symbolic-based	Coding symbolic	Knowledge graph	The explanation is more objective and understandable for its close connection with factual features	The explanation may be incomplete with some hidden information	[23–25, 38, 39]
	Qualifying symbolic	Rule-extraction techniques; semantic web; knowledge graph	The explanation can incorporate prior knowledge	The explanation aims at describing the inferential process of a model trained on categorical inputs. It is highly dependent on reliable rules or knowledge graphs	[29, 40–45]
Hybrid-based		Rivelo; explainer; TensorBoard	Users can interactively explore a set of visual and textual instance-level explanations	If the mixed explanation is not well designed and structured, the explanation retrieved by humans might be confusing	[46–51]

Table 3 Summarising the XAI approaches from logic-oriented

Type	Typical examples	Pros	Cons	References
E-E relationship	Regression-based partitioned method; Importance estimation network, etc.	Better applicability with no limitations for various models	Still has challenging to understand the internal logic that how to generate outputs and how to forecast the model behaviour in a variety of situations	[52–54]
M-E relationship	Feature-wise relevance; Layer-wise relevance; etc.	The procedure is quite straightforward, and it is easy to identify internal structure or any bias in the middle	The explanation may be partial just covering particularly part of interesting areas	[55–57]
Corr relationship	Input-level correlation (feature, label, image, etc.)	It has high benefits to understand why the model produces a given result during the entire learning	The explanation may be complex by considering different dimensions of relationships	[58, 59]

approaches used in medical image analysis. Some typical examples include attributed-based and perturbation-based methods for model-agnostic explanations as well as CAM-based and concept attribution for model-specific explanations. Locally-interpretable model-agnostic explanations (locally-interpretable model-agnostic explanations (LIME)) [86] is utilised to generate explanations for the classification of medical image patches. Zhu et al. [87] used rule-based segmentation and perturbation-based analysis to generate the explanation for visualising the importance of each feature in the image. The concept attribution [37] is introduced by quantifying the contribution of features of interest to the CNN network's decision-making. Symbolic methods focus on the symbolic information representations that simulate the doctor's decision-making process with natural language, along with the generated decision results, such as primary diagnosis reports, etc. For example, Kim et al.[66] introduced concept activation vectors (CAVs), which provided textual interpretation of neural network internal state with user-friendly concepts. Lee et al. [73] provided explainable computer-aided diagnoses by combining a visual pointing map and diagnostic sentences based on the predefined knowledge base.

In terms of logical-oriented, explanations focused on end-end logic reasoning, such as the above-mentioned LIME, perturbation-based methods are utilised to explain the relationship between input medical images and predicted results. For example, a linear regression model is embedded into LIME [86] to identify relevant regions by plotting heat maps with varying color scales. Zhang et al. [56] provided a diagnostic reasoning process and translate gigapixels directly to a series of interpretable predictions. Shen et al. [58] used a hierarchical architecture to present a concept learning-based correlation model. The prediction's interpretability is aided by the model's intermediate outputs, which anticipate diagnostic elements connected to the final classification. A correlation-XAI approach proposed by [59] is used for feature selection merging generalised feature importance obtained with shapley additive explanations (shapley additive explanations (SHAP)) and correlation analysis to achieve the optimal feature vector for classification.

3.1.3 Cases Studies

Explainable AI applications in the medical field assist healthcare providers in making accurate diagnoses, treatment decisions, risk assessments, and recommendations. The transparency and interpretability of these AI models ensure that clinicians can trust and validate the outputs, leading to improved patient care and outcomes [92].

Lesion classification: The visualisation of lesion areas mainly refers to heat maps [88], attention mechanisms [89, 90], and associated with other diagnostic means such as structural remodeling [91] and language models to represent report text [156] to find out lesion areas. These methods provide visual evidence to explore the basis for medical decision-making. For example, Biffi et al. [91] used a visualisation method on the original images to measure the specificity of pathology, using interpretable characteristics of specific tasks to distinguish clinical conditions and make the decision-making process transparent. Garcia-Peraza-Herrera et al. [92] used embedded activation charts to detect early squamous cell tumors, showing the focus on the interpretability of the results and use it as a constraint to provide a more detailed attention map. Paschali et al. [88] used a model to activate the fine-grained Logit heat map to explain the medical imaging decision-making process. Lee et al. [90] used head CT scan images to detect acute intracranial hemorrhage, and proposed an interpretable deep learning framework. Liao et al. [89] provided a visual explanation basis for the automatic detection of glaucoma based on the attention mechanism, and in the process of automatic glaucoma detection, the system provides three types of output: prediction result, attention map, and prediction basis, which enhances the result interpretability.

Table 4 Applications of XAI

Application field	References	SO		RO			LO			Case studies
		S	O	I	M	H	E-E	M-E	Corr	
Medical and biomedical	[36, 37, 40, 60–88, 88, 89, 89, 90, 90–97]	✓	✓	✓	✓	✓	✓	✓	✓	Lesion classification; disease diagnosis and treatment
Healthcare	[98–101] [102–112]	✓	✓	✓	✓	✓	✓	✓	✓	Pain detection; quality of life analysis; surgery complication
Cybersecurity	[113, 113–126]	✓	✓	✓	✓	✓	✓	✓	x	Intrusion detection system; malware detection
Finance and law	[17, 127–134]	✓	✓	✓	✓	✓	✓	✓	x	Financial-related figure forecasting; credit risk management; legal text classification
Education and training	[135–143]	✓	✓	✓	✓	✓	✓	x	✓	Feedback providing; intelligent tutoring systems
Civil engineering	[36, 63, 144–152]	✓	✓	✓	✓	✓	✓	✓	✓	Decision Vehicle actions; power system management

Disease diagnosis and treatment: Research on XAI in disease diagnosis and treatment has recently gained much attention. Amoroso et al.[93] applied clustering and dimension reduction to outline the most important clinical feature for patients and designed oncological therapies in the proposed XAI framework. In the context of high-risk diagnoses, explainable AI techniques have been applied to provide visual cues and explanations to clinicians. For example, Sarp et al. [94] utilized LIME (local interpretable model-agnostic explanations) to generate visual explanations for a CNN-based chronic wound classification model. These visual cues help clinicians understand the model's decision-making process and provide transparency in the diagnosis. Moreover, Wu et al. [95] proposed a counterfactual multi-granularity graph supporting fact extraction (counterfactual multi-granularity graph supporting fact extraction (CMGE)) for lymphedema diagnosis. CMGE is a graph-based neural network that extracts facts from electronic medical records, providing explanations and causal relationships among features. These explanations assist clinicians in comprehending the reasoning behind the diagnosis and identifying relevant factors contributing to the condition.

In the domain of relatively low-risk screenings, explainable AI research has explored the integration of medical records and natural language processing methods to provide interpretable diagnostic evidence. For instance, Wang et al. [96] and Lucieri et al. [97] have integrated medical records into map and image processing, creating diagnostic reports directly from medical images using multi-modal medical information. This integration enables the generation of interpretable evidence and explanations for clinicians during the screening process.

Additionally, hybrid XAI approaches have been explored, such as Mimir proposed by Hicks et al. [84]. Mimir learns intermediate analysis steps in deep learning models and incorporates these explanations to produce structured and semantically correct reports that include both textual and visual elements. These explanations aid in understanding the screening results and provide insights into the features contributing to the risk assessment, assisting clinicians in making informed decisions and recommendations for further screenings or preventive measures.

3.1.4 Domain-Specific Insights

In terms of biomedical field, the end-users of XAI are mostly pharmaceutical companies and biomedical researchers. With the use of AI and XAI, they can understand the reasoning behind predictions made for disease diagnosis and diagnostic evidence. This insight into the decision-making process can enhance the transparency and trustworthiness of AI-based predictions, leading to more accurate, reliable and efficient disease diagnosis. Nonetheless, XAI techniques

have their limitations. For example, due to the inherent complexity of disease diagnosis and treatment, XAI may struggle to provide full and concise explanations of all factors influencing a prediction. Further, if a situation is novel or significantly different from past scenarios used in training the AI, the explanations offered by XAI may be insufficient or not entirely accurate. Moreover, given the dynamic and multifaceted nature of public health and disease spread, XAI might struggle to provide real-time explanations or take into account every single factor influencing disease spread. This includes factors like socioeconomic conditions, behavior changes, and environmental changes. Additionally, if the situation changes rapidly, as in the case of a new disease outbreak, the explanations provided by XAI might be outdated or not entirely accurate. Therefore, XAI provides significant advantages in the biomedical field by enhancing transparency and trust in AI predictions, it also faces challenges related to the complexity and dynamic nature of biomedical data and scenarios. More research and advancements are needed to improve the capability of XAI in handling these challenges and providing clear, concise, and real-time explanations.

3.2 Healthcare

3.2.1 Problem Definition

AI-assisted exploration has broad applications in healthcare including drug discovery [104, 105] and disease diagnosis [103, 106]. Deep learning techniques achieved high accuracy on classification problems—e.g., using MRI images to identify different stages of dementia [157]. The Healthcare industry can now examine data at extraordinary rates without sacrificing accuracy due to the development of deep learning. Healthcare offers unique challenges, with typically much higher requirements for interpretability, model fidelity, and performance than most other fields. More interpretable solutions are in demand apart from the binary classification of positive-negative testing, and they can benefit clinicians, allowing them to have an understanding of the results. In healthcare, critical applications like predicting a patient's end-of-life may have more stringent conditions on the fidelity of interpretation than just predicting the cost of a procedure[158]. Researchers have analysed the interpretability of deep-learning algorithm for the detection of acute intracranial haemorrhage from small datasets [90]. Moreover, some studies focus on explaining how AI improves health-associated records of individual patients' electronic health records (electronic health records (EHR))[159]. This is because AI models for treating and diagnosing patients might be complex as these tools are not always sufficient on their own to support the medical community and medical staff may not be exposed to these technologies before[159].

3.2.2 XAI Based Proposals

Adding to the numerous research of SARS-CoV-2 mutations, Garvin et al. [160] used a method called iRF-LOOP [161], an XAI algorithm, in combination with Random Intersection Trees (RIT) [162] for the matrix of variable site mutations analysis. The network as the output of the iRF-LOOP model includes a score of the ability to predict the absence or presence of another variable site mutation for each variable site mutation. Ikemura et al. [163] used an unsupervised XAI method BLSOM (batch-learning self-organizing map) for oligonucleotide compositions of SARS-Cov-2 genomes that reveals new noval characteristics of genome sequences and drivers of species-specific clustering. This BLSOM method presented the contribution levels of the variables at each node by visualising the explanation with a heat map.

In terms of source-oriented, XAI in healthcare is mainly visible, measurable findings, medical histories, numerical records and reports. This information conducts from a series of medical examinations, which is not subjective to the explanation methods or requirements. For example, SHAP is an explainer that helps to visualise the output of the machine learning model and compute the contribution of each feature to the prediction [164].

In terms of representation-oriented, in the healthcare field, explanations of AI models are more realistically applicable to overall AI processes, but individual decisions need to be carefully considered. XAI in the healthcare field mainly includes causality, which is the capacity to identify haphazard connections among the system's many components, and transferability, which is the capacity to apply the information the XAI provides to different issue domains[107].

In terms of logical-oriented, explanations in healthcare mainly focus on correlation analysis. For example, SHAP has been widely used in the healthcare industry to provide explanations for hospital admission [108], quality of life [109], surgery complication [110], Oncology [111] and risk factor analysis of in-hospital mortality [112].

3.2.3 Cases Studies

Pain detection based on facial expressions: Understanding the choices and restrictions of various pain recognition models is essential for the technology's acceptability in high-risk industries like healthcare. Researchers have provided a method for examining the variances in learned representations of models trained on experimental pain (BioVid heat pain dataset) and clinical pain (UNBC shoulder pain dataset). To do this, they first train two convolutional neural networks, one for each dataset, to recognise pain and the absence of pain automatically. The performance of the heat pain model is then assessed using pictures from the shoulder pain dataset, and vice versa. This is known as a cross-dataset

evaluation. Then, to determine which areas of the photographs in the database are most relevant, they employed a Layer-wise Relevance Propagation [165]. In this study, they showed that the experimental pain model is paying more attention to facial expression.

Prediction of caregiver quality of life: The caregiver's quality of life may be impacted by the patient's depression and employment status before the onset of symptoms. Some researchers analysed the quality of life of caregivers, using SHAP to visualize the overall impact of all features and then selecting a subset of the seven most predictive features to establish a simpler model (M7) for a global explanation. Simpler models may be easier to comprehend, less likely to be influenced by unimportant or noisy aspects, more accurate, and more useful. In this study, SHAP was used to provide post-hoc explanations. Studies explored the most predictive features that impact the quality of caregivers' life, including the weekly caregiving duties, age and health of the caregiver, as well as the patient's physical functioning and age of onset [109].

3.2.4 Domain-Specific Insights

In healthcare, the end-users of XAI systems range from clinicians and healthcare professionals to patients and their families. Given the high-stakes nature of many medical decisions, explainability is often crucial to ensuring these stakeholders understand and trust AI-assisted decisions or diagnoses. One of the primary benefits of XAI in the healthcare domain is the potential to make complex medical decisions more transparent and interpretable, leading to improved patient care. By providing clear explanations for AI-driven predictions, such as identifying risk factors for a particular disease, XAI can help clinicians make more informed decisions about a patient's treatment plan. Patients, too, can benefit from clearer explanations about their health status and prognosis, which can lead to better communication with their healthcare providers and a greater sense of agency in their care. For instance, in the context of disease diagnosis, AI models equipped with XAI can interpret complex medical imaging data, such as MRI scans, to accurately diagnose a disease like dementia or cancer. Not only can these models highlight which features are most important in reaching a diagnosis, but they can also provide a visual explanation that assists clinicians in understanding the model's reasoning. Similarly, in drug discovery, XAI can assist in identifying novel therapeutic targets and predicting the efficacy of potential drugs, improving the speed and accuracy of drug development. The transparency provided by XAI in this process can improve trust and confidence in the AI model's suggestions, and potentially speed up the regulatory approval process. However, the implementation of XAI in the healthcare domain is not without challenges. Privacy and data security are significant concerns when dealing with sensitive health data. Additionally, the ability of XAI to provide clear, comprehensible explanations in cases where the underlying AI model is extremely complex remains a challenge. Moreover, the healthcare field is inherently dynamic and complex, with countless interacting variables, so the explainability provided by current XAI methods might not be complete or fully accurate. Finally, there are also significant regulatory and ethical considerations that come with the application of AI and XAI in healthcare. Regulators will need to establish clear guidelines for the use of these technologies to ensure that they are used responsibly and ethically.

3.3 Cybersecurity

3.3.1 Problem Definition

Cybersecurity is the use of procedures, protections, and technologies to defend against potential online threats to data, applications, networks, and systems [166]. Maintaining cybersecurity is becoming more and more challenging because of the complexity and huge amount of cyber threats, including viruses, intrusions, and spam [167].

In recent years, intelligent network security services and management have benefited from the use of AI technology, such as ML and DL algorithms.

There has been a variety of AI methods and tools developed to defend against threats to network systems and applications that may be present inside an organisation or outside of it. However, it is challenging for humans to comprehend how these outcomes are produced since the developed network security-related decision-making model based on artificial intelligence lacks reasons and rational explanations [168]. This is due to the black-box nature of AI models. As a result of the network vulnerability, the network defence mechanism in this situation transforms into a black box system that is susceptible to information leakage and the effects of AI [169]. In order to combat cyber security that takes advantage of AI's flaws, XAI is a solution to the growing issue of the black box in AI. Due to XAI's logical interpretation and key data proof interoperability, experts and general users can comprehend AI-based models [170].

Zhang et al. [171] divided these applications into three categories: defensive network threats applications using XAI, network security XAI applications in different industries, and defence methods against network threats in XAI applications. This section mainly analyses the defensive applications of XAI against network attacks.

3.3.2 XAI Based Proposals

In terms of source-oriented, the objective explanation is based on the detected system or network data, while the

subjective explanation depends on the analysts' expertise and backgrounds. The existing XAI proposals cover both objective and subjective aspects. For example, Hong et al. [113] proposed a framework to provide explanations from the model and also combine with the analysts' knowledge to eliminate the false positive errors of the decision made by the AI Models. Amarasinghe and Manic [114] proposed the method to give the most relevant input features align with the domain experts' knowledge for each concept learned from the system.

In terms of representation-oriented, embedding with human understandable texts to interpret the results of the decision-making process is a common way. For example, Amarasinghe et al.[115] used text summary to interpret the reason to the end user of an explainable DNNs-based DoS anomaly detection in process monitoring. Besides, to present the explanation logic, DENAS [116] is a rule-generation approach that extracts knowledge from software-based DNNs. It approximates the nonlinear decision boundary of DNNs, iteratively superimposing a linearized optimization function. Moreover, an image visualisation method was also used to explain, as Gulmezoglu [117] generated LIME and used saliency maps to examine the most dominant features of the website fingerprinting attack after the trained DL model. Feichtner et al. [118] used LIME to calculate a score for each word, showing the importance of output, and used the heat map visualisation method to interpret the text description and application request permissions based on samples of the correlation between groups. Another interpretation proposal [119] used the image to represent a mobile application and localise the useful salient parts of the model by the Grad-CAM algorithm. In this way, the analyst can gain knowledge by using the image symptom region for a specific prediction.

In terms of logic-oriented, explanations focus on end-end logic reasoning similar to LIME used in other areas, while local explanation method using nonlinear approximation (LEMNA) (Local explanation method using nonlinear approximation) is optimised for security applications based on deep learning, such as PDF malware recognition and binary reverse [120]. In contrast to the strong assumption that the model made by LIME is locally linear, the LEMNA scheme can deal with local nonlinearities and takes into account the dependencies between features. LEMNA can be used for the interpretation of a function starting position detection model in a scene in binary reverse. Yan et al. [121] proposed a technique for extracting a rule tree merged from generated rule tree of the hidden layer and the output layer of the DNN, which shows the more important input feature in the prediction task. Also, the middle-to-end explanation is used in this area. Amarasinghe et al. [115] used the layer-wiser relevance propagation (LRP) method [122] to find the relevance features between the input layer and the last layer to explain what input feature contributes to making that decision.

3.3.3 Cases Studies

Intrusion detection systems XAI studies in intrusion detection systems are normally used to provide explanations regarding different user perspectives. Most approaches use already-developed methods to make the results interpretable, with SHAP being the most adopted. LIME, on the other hand, has been adopted in only a few cases. Shraddha et al. [123] proposed a framework with a DNN at its base and apply an XAI method to add transparency at every stage of the deep learning pipeline in intrusion detection system (IDS). Explanations give users measurable factors as to what features influence the prediction of a cyber-attack and to what degree. Researchers create XAI mechanisms depending on who benefits from them. For data scientists, SHAP and BRCG [124] are proposed, while for analysts Protodash is used. For end-users where an explanation of a single instance is required, researchers suggest SHAP, LIME, and CEM. Hong et al. [113] proposed a network intrusion detection framework called FAIXID making use of XAI and data cleaning techniques to enhance the explainability and understandability of intrusion detection alerts. The proposed XAI algorithms included exploratory data analysis (EDA), Boolean Rule column generation (BRCG), and contrastive explanations method (CEM) that deployed in different explainability modules respectively to provide cybersecurity analysts with comprehensive and high-quality explanations about the detection decisions made by the framework. On the other hand, collecting analysts' feedback through the evaluation module to enhance the explanation models by data cleaning also proved effective in this work as well.

Malware detection The effectiveness of malware detection increases when AI models are applied to signature-based and anomaly-based detection systems. Heuristic-based [172] methods were proposed to understand the behavior of an executable file using data mining and deep learning approaches. The development of interpretable techniques for malware detection in mobile environments-particularly on Android platforms-has received a lot of attention. The adversarial attack method is also used to improve interpretability. Bose et al. [125] provided a framework for interpolating between various classes of samples at various levels to look at many levels of weights and gradients as well as the raw binary bytes to understand how the MalConv architecture [126] learns in an effort to understand the mechanisms at work.

3.3.4 Domain-Specific Insights

AI-based cybersecurity systems carry two different kinds of secondary risks. The possibility of generating misleading negative results that result in erroneous conclusions is present in the first category. The second is the chance of receiving inaccurate notifications or erroneous warnings due to false positive results [173]. In such circumstances, it is imperative to take the required mitigating action to ensure that violations or unique events are handled more accurately, keeping the decision-making process's ability to be understood and supported [174]. Additionally, the use of AI by hackers makes it possible for them to circumvent security measures so that data tampering goes unnoticed. This makes it challenging for businesses to correct the data supplied into AI-based security systems. As a result, compared to conventional model-based algorithms, the current difficulty with AI-based systems is that they make decisions that lack reason [167]. Hence, XAI is required to enhance trust and confidence in AI-based security systems.

In cybersecurity, XAI requires a more structured approach, utilizing various integrated techniques from diverse fields, guided by a dedicated research community focusing on increasing formalism. In particular, for areas like malware detection, there is a need for unified and clearly explained methods, ensuring a coherent understanding for all stakeholders, including users, analysts, and developers, to enhance the analysis and prevention of cyber-attacks. Moreover, currently, there's no recognized system for gauging whether one XAI system is more user-intelligent than another. A well-defined evaluation system to select the most effective explainability technique is needed. XAI also needs to grapple with substantial security and privacy issues. Current XAI models are still vulnerable to adversarial attacks, leading to public concerns about XAI security.

3.4 Finance

3.4.1 Problem Definition

In the finance domain, XAI is mainly applied in financial forecasting [175] and credit risk management [176]. Financial forecasting is the problem of predicting different financial-related metrics, such as profit, financial healthy ratios, etc. Credit risk management is the problem of how to manage credit risks from different subjects. For instance, from a bank perspective, when a model predicts a client to be at high risk of default, it should give a reason for different stakeholders to understand.

3.4.2 XAI Based Proposals

XAI in finance is mainly in the form of explaining which features are more important and which features are less important in the AI model. Under this form, different statistical-based techniques can be applied to find the features' importance. In the legal domain, information is represented as natural language texts, so the objective of XAI models is to identify the important words that contribute to the final prediction.

In terms of source-oriented, most XAI models in the finance domain are objective and cognitive-based, as they are based on fact, not on the human experience. In the legal domain, those XAI models are objective cognitive since the highlighted words are from the input data. As CNNs have been proven useful in text input data, general XAI methods (e.g. Grad-CAM, LIME, SHAP) for CNNs have been used to explain the AI model trained for legal text [134]. To demonstrate the contribution of each word in the given sentence on the final prediction, XAI models, like LIME, can indicate the contribution of the word "awesome" on the positive sentiment prediction result.

In terms of representation-oriented, XAI models in the finance domain vary among visualization-based, symbolic-based and hybrid. For visualisation-based, some tree XAI models [177] can give tree-shape visualisations. For symbolic-based, SHAP and LIME models can give numerical forms. In this case, XAI models try to give the explanation or contribution of different features in making the prediction. Symbolic-based XAI in the financial domain emphasises giving the explanation in a quantitative manner. This is more applied in credit risk management. For hybrid, it is symbolic-based and basically a combination of visualisation. Visualisation in the representation-oriented category has been adopted to the legal domain to visualize the explainability [132].

3.4.3 Cases Studies

Financial data forecasting AI models are being used to predict financial-related figures, such as stocking price, profit of the company, etc. However, sophisticated AI models can only give predictions rather than how these predictions are being made. In a typical application of XAI in financial forecasting cases, Local interpretable model-agnostic explanations (LIME) are used to explain the model predictions locally around each example. LIME is an agnostic-based XAI model which is to simplify the complex model by using a linear regression model locally. For instance, Agarwal et al. [178] applied LIME on a stock price AI prediction model (the AdaBoost model) to explain the prediction. The AI (AdaBoost model) model used 20 previous days' stock prices as the feature to predict the next day's stock price. The

output of LIME is a 2-dimensional bar chart, with the Y-axis showing features and the X-axis showing the importance of features. In this way, the bar chart clearly shows how the AI prediction model makes the prediction (i.e., which features play dominant roles in the prediction model). Similarly, the SHAP XAI model can also be used to explain the financial forecasting model. The principle of the SHAP model is to compute the contribution level of each feature to the prediction based on game theory. In the same work of Agarwal et al. [178], SHAP is used to explain the AdaBoost and random forest-based stock price prediction model. The output of the SHAP model is a 2-dimensional chart, with the X-axis showing the contribution of features to the model and the Y-axis showing each feature used in the prediction model.

Credit risk management In addition to the financial foresting case, XAI has also been applied in the case of credit risk management. Credit risk may happen to a financial institution due to an expected financial loss. In [177], a SHAP model and a minimal spanning tree model were used to explain an extreme gradient boosting (XGBoost) based credit default risk default probability estimation model. Features used in the estimation model include financial information from the balance sheet. The minimal spanning tree model presents a clustering tree based on financial data, and the SHAP model presents the contribution level of each feature. The principle of the minimal spanning tree model is to present the credit risk prediction in a clustered-based visualisation, while the principle of the SHAP model is to present the explainability at a contribution level.

3.4.4 Domain-Specific Insights

In terms of credit risk management, the end-users of XAI are mostly financial institutions like banks and insurance companies. XAI provides transparency and explanations for AI-driven decisions, allowing these institutions to understand and validate the factors influencing risk assessments and fraud detection. This empowers financial institutions to make more informed and accountable decisions. XAI techniques may struggle to provide clear and concise explanations for every aspect of the decision, potentially leading to incomplete or partial explanations. In addition, credit risk is a dynamic and evolving field, influenced by various economic, regulatory, and market factors, so the XAI may not be able to provide real-time explanations of the risk management decisions. XAI techniques may struggle to provide clear and concise explanations for every aspect of the decision, potentially leading to incomplete or partial explanations. In addition, credit risk is a dynamic and evolving field, influenced by various economic, regulatory, and market factors, so the XAI may not be able to provide real-time explanations of the risk management decisions.

In terms of financial data forecasting, the end-users of XAI are mostly financial institutions like fund management companies and asset management companies. With explanations provided by AI models, financial professionals gain insights into the reasoning behind investment recommendations, risk assessments, and other financial decisions. This helps them validate the suggestions and communicate the rationale more effectively to their clients. One limitation of XAI techniques is that they cannot explain events that never happened in the past. When faced with new and unprecedented circumstances, the explanations provided by XAI may not adequately account for these events, leading to potentially inaccurate forecasts. There is limited current research analyzing the specific computational cost in XAI models in finance. The computational cost of XAI used in the finance domain depends on the complexity of the underlying AI model, feature dimensions, and hardware level. If the underlying AI model is a regression or tree-based model and does not include many factors, the computational cost will be relatively low. However, if the underlying AI model is based on complex neural networks and includes tons of factors in the model, the computational cost will be high.

3.5 Law

3.5.1 Problem Definition

In the legal domain, one of the major concerns of XAI is whether the legal decisions are made fairly towards certain individuals or groups [133], since it is the high-stake domain, and decisions have real significant impacts on real human beings. The explainability of AI models provides transparency on how the decisions are made. It is the desired feature for decisions, commendations, and predictions made in the legal domain by AI algorithms. However, there are few works that have been done in the legal domain apart from general XAI methods [17].

3.5.2 XAI Based Proposals

In the legal domain, information is represented as natural language texts, so the objective of XAI models is to identify the important words that contribute to the final prediction.

In the legal domain, those XAI models are objective cognitive since the highlighted words are from the input data. As CNNs have been proven useful in text input data, general XAI methods (e.g. Grad-CAM, LIME, SHAP) for CNNs have been used to explain the AI model trained for legal text [134]. To demonstrate the contribution of each word in the given sentence on the final prediction, XAI models, like LIME, can indicate the contribution of the word “awesome" on the positive sentiment prediction result.

In terms of logic-oriented, there are few XAI models in the finance domain falling into this end. As in the finance domain, the cases are about the prediction of a probability or a numeric value, and the demand for the explanation of AI models tends to be about which factors contribute more to the final prediction. For the legal domain, the XAI models are end-end relationships, since the explanation has been expressed as the relationship between input words and output predictions.

3.5.3 Cases Studies

Legal text classification Authors in [132] provided a case study from a lawyer's perspective to utilise Grad-CAM, LIME, and SHAP to explain the legal text classification outcomes. The AI model consists of DistiBERT for embedding and CNN for classification. The method used to represent the explainability is the heatmap visualisation, i.e., highlighted words with different intensities corresponding to different contributions to the final prediction. Apart from two evaluation metrics, the responses of lawyers on the given explanation have also been collected. The scores on visualisations for the selected six correctly classified sentences are from 4.2 to 6.66 with 0 for worst and 10 for best. The key point made by lawyers is that the explanations made by XAI should be understandable for users who have no professional knowledge of the legal domain.

3.5.4 Domain-Specific Insights

Explainability not only is necessary for AI applications in Law, but also is required by law (e.g., GDPR) for AI applications. In this subsection, we address the necessity of explainability in AI applications in law. The decisions made in law require explainability by nature [179] as it forms the important part of outputs. All judgements need reasons. Lawyers need to explain to the clients, judges need reference to relevant articles or cases to support the decision[132]. For the AI-empowered legal consultation or recommender systems, the more important information is why this is relevant instead of just listing the relevant articles or similar cases. For judge results prediction, it is only helpful to professionals like lawyers when explainability is provided.

Although the necessity of explainability in AI applications in Law, its adoption is faced with challenges and difficulties. The explainability can be the relevant articles or similar case, but more importantly, the analysis to link them to the target case. The heatmap, mentioned above as an example, may provide certain extend of explainability by highlighting the key words used to make decisions. However, the explainability in law applications requires more descriptions in natural language as the most inputs of AI systems in law are texts written in natural language. This explainability requires certain level of reasoning capabilities to have the explain make sense to the users.

Another challenge is the linking of the evidences. Many legal decisions made by AI systems involve multiple parts of the input text or documents. Explains using only one piece of the information are incomplete. The advent of the large language models (LLMs), such as GPT series models, may facilitate the reasoning and explaining of decisions made in AI applications in law. The LLMs can be instructed to give reference for the outputs generated. This provides opportunities for explainable use of AI in law, but the models involve extra resources.

3.6 Education and Training

3.6.1 Problem Definition

As one of the essential methods to improve and optimise learning, AI is now widely applied in the field of educational research [180, 181]. The applications of AI in education (AIED) have shown great benefits in several ways, including support instructions, personalised learning systems, and automated assessment systems [182]. At the same time, there are some risks associated with the use of AI, given the specific nature of education. For example, bias is a prominent issue in discussions on AI ethics in education. When using AI techniques for student performance prediction or risk identification, they are likely to produce more biased results for a particular group of students based only on the different demographic variables (such as gender) of the students [183]. Consequently, concerns about AI in relation to fairness, accountability, transparency, and ethics (FATE) have sparked a growing debate [135].

XAI contributes to making the decision-making process of an AI model more transparent and fair by explaining the internal processes and providing a logical chain for generating outcomes. This is essential for human users to build trust and confidence in AIED. Recently, there has been a growing body of research showing the opportunities and demands of applying advanced XAI for education [135].

3.6.2 XAI Based Proposals

In terms of source-oriented, there are different objective factors that can be used to analyse and evaluate the performance of students in the education domain. Also, since each student comes from a different family environment and has a very distinct personality, subjective perceptions and feelings in learning have significant impacts on educational outcomes due to the different circumstances of each student. Alonso and Casalino [136] proposed a method that uses an XAI tool, named ExpliClas, to analyse objective attributes of students' learning processes. Their method provides both

global and local explanations. There are also many studies that use objective characteristics, such as age, gender, and study hours, to predict and explain students' performance in teaching activities [137].

In terms of representation-oriented, the XAI approaches in education mainly include visualization-based and symbolic-based explanations. In [138], a deep learning-based method was proposed to achieve the automatic classification of online discussion texts in the educational context. Particularly, the authors used gradient-based sensitivity analysis (SA) to visualise the significance of words in conversation texts for recognising the phases of cognitive presence, thus providing the explanation for the deep learning model. Recently, some researchers have also applied the symbolic approach in education, expecting to adopt symbolic knowledge extraction to provide logical reasoning for the AI interpretation. Hooshyar and Yang [139] provided a framework that integrates neural-symbolic computing to address the interpretability of AI models. The framework considered using prior knowledge, such as logic rules and knowledge graphs.

In terms of logical-oriented, XAI in education is primarily required to provide explanations for machine learning black-box and rule-based models. SHAP was employed in [140] to explain the black-box student dropout classification model in relation to Shapley values. Explanation from the rule-based algorithms specialises in showing clear logic from the input data to the output results. For example, in [141], global explanations were provided to train nursing students by analysing the temporal logic between actions.

3.6.3 Cases Studies

Feedback providing For students, getting timely and formative feedback from educators about performance and assignments is an important way of improving the efficiency and quality of learning. Feedback should include not only the student's marks and evaluation but also, and more importantly, an explanation of the problems with the assignment and learning. XAI has been applied in this area, the relevant techniques include sequence mining, natural language processing, logic analysis and machine learning approaches. Take writing analytics as an example, which aims to provide feedback for students to improve their writing. Knight et al. [142] introduced an open-source tool, AcaWriter, which provides informative feedback to fit different learning contexts. AcaWriter employs the Stanford CoreNLP to process each input sentence, and then it uses a rule-based model to extract the matched patterns. Their research demonstrates the great application of XAI in education by explaining to students about their writing.

Intelligent tutoring systems In addition to providing feedback to students, XAI can also help give personalised tutoring instructions based on the student's learning activity performance. Conati et al. [143] attempted to integrate multiple machine learning algorithms into an interactive simulation environment so as to provide hints to students regarding their learning behaviours and needs. The proposed XAI mechanism consists of behaviour discovery, user classification, and hint selection. In the behaviour discovery phase, the authors first apply an unsupervised clustering algorithm to group students and then use association rule mining to analyse student behaviour further. In the user classification phase, they build a supervised classifier to predict students' learning. In the hint selection phase, the previous classification result and association rules will be used to trigger the corresponding hints.

3.6.4 Domain-Specific Insights

The end-users of XAI in education and training mainly include students and educators. XAI can help students understand and interpret the outcomes of AI-driven systems, such as automated grading or recommendation algorithms, providing them with transparency and insights into the feedback. Additionally, XAI provides educators with a great opportunity to gain a deeper understanding of the AI-powered educational tools they employ in their classrooms. By using XAI, educators can acquire insights into the underlying reasons behind specific recommendations or suggestions generated by these systems. Consequently, they can adapt and customize their teaching strategies based on this understanding.

While XAI holds great potential and prospects for application in the field of education, there are currently challenges and limitations that need to be addressed. Many AI algorithms, such as deep learning neural networks, can be complex and difficult to interpret. XAI techniques still have limitations to provide clear and comprehensive explanations for the decisions made by these complex models, which can hinder their adoption in educational settings. Also, there is a Lack of standardization for XAI. We need standardized metrics and a framework to evaluate and assess the explanations provided by XAI, particularly when comparing different XAI techniques and approaches. The absence of standardized practices and guidelines can lead to inconsistency and confusion in implementing XAI solutions. Addressing trade-offs is an essential step in developing machine learning models, and XAI is no exception. Finding the right balance between explainability and performance is crucial, especially in educational contexts where accurate feedback and predictions are necessary.

3.7 Civil Engineering

3.7.1 Problem Definition

AI systems used in civil engineering research have a significant impact on the decision-making processes in road transport and power systems. In particular, autonomous driving techniques in road transport and power system analysis and power systems are the common areas used deep learning techniques, such as navigation and path planning, scene recognition, lane and obstacle detection, as well as planning, monitoring, and controlling the power system [150, 184].

In the field of autonomous driving, deep learning techniques are normally utilised to recognize scenes for digital images [184, 185]. While in the field of power system analysis, deep learning techniques are used to extract features from the underlying data for power system management, such as power grid synthesis, state estimation, and photovoltaic (PV) power prediction [150, 186]. Deep learning explainable techniques are used to automatically extract abstract features of images or depth non-linear features of underlying data through end-to-end predictive processing to obtain results, which is not sufficient to provide the evidence to trust and accept the result of autonomous driving and power system management. For example, one can use traffic lights and signal recognition for driving planning, in which the traffic lights at crosswalks and intersections are an essential function in following traffic rules and preventing traffic accidents. Deep learning methods have achieved prominence in traffic sign and light recognition, but they are hard to explain the correlation between inputs and outputs and lack an explanation to support reasoning in driving planning studies [187]. In power system management, deep learning methods may mislead the output explanations of power stability to provide unreliable recommendations, so explanations can increase user trust [150].

3.7.2 XAI Based Proposals

XAI can improve the management of autonomous driving and power system, providing an effective interaction to promote smart civil engineering. Deep learning interpretation research in autonomous driving and power systems is a common interpretable deep learning method because it is not only influenced by data, but also relates to expert knowledge and ethical principles.

In terms of source-oriented, objective interpretability obtains visible or measurable results from 2D and 3D images or underlying datasets, while subjective interpretability requires consideration of the knowledge from automotive or electrical experts and the ethical standards of their fields. Currently, XAI proposals include objective and subjective cognitive aspects. For example, CAM, as an objective cognition method, is used to explain the highlight of important regions in 2D or 3D images. Time series, 2D images, 3D images, Lidar images, knowledge databases and ethical criteria are utilised as subject sources to explain the model [147, 185, 187].

In terms of representation-oriented, visual interpretation is the highest level semantics to understand which parts of the image impact the model, emphasing on visual structure of data and model, which is the primary XAI method used in autonomous driving. These XAI methods can be divided into gradient-based and back propagation-based. Gradient-based interpretation methods include CAM, and its enhanced variants such as Guided Grad-CAM, Grad-CAM, Grad-CAM++ and Smooth Grad CAM++. CAM can highlight the discriminative regions of a scene image used for scene detection [147]. Backpropagation-based methods contain guided backpropagation, layered relevance propagation, visual backprop and deep lift. Visual Backprop shows which input pixels set contributes to steering self-driving cars [144]. Symbolic interpretation uses understandable language to provide evidence for result recommendations in autonomous driving and power system management. In autonomous driving, proposed AI methods make decisions according to traffic rules. For example, "the traffic light ahead turned red," thus "the car stopped" [185]. In power system management, it uses the data gathered from occupant actions for resources such as room lighting to forecast patterns of energy resource usage [188]. Hybrid interpretation combines visual interpretation and symbolic interpretation to provide steering determination in autonomous driving. For example, Berkeley Deep Drive-X (BDD-X) is introduced in autonomous driving which includes the description of driving pictures and annotations for textual interpretation [49].

In terms of logical-oriented, the end-end explanations are used to explain the relationship between input images including obstacle and scene images and the prediction. For example, LIME is utilised to explain the relationship between input radar image and prediction results [189]. Middle-end explanations reveal reasons behind the autoencoder-based assessment model and how they can help drivers reach a better understanding and trust in the model and its results. For example, a rule-based local surrogate interpretable method is proposed, namely MuRLoS, which focuses on the interaction between features [149]. Correlation expatriation is used in the risk management of self-driving and power systems. For example, SHAP is used to assess and explain collision risk using real-world driving data for self-driving [190].

3.7.3 Cases Studies

Decisive vehicle actions Decisive vehicle actions in autonomous driving are based on multiple tasks, such as scene recognition, obstacle detection, lane recognition, and path

planning. It can use attention mechanisms, heat maps, diagnostic models and texture descriptions to recognise obstacles, scenes and lanes and steer the car operation [147, 185, 187]. As mentioned before, CAM is used to highlight the main area for recognition [63]. Visual Backprop, unlike CAM-based, emphases highlighting pixel-level to filter features of scene images [144]. Grad-CAM is combined with existing fine-grained visualisations to provide a high-resolution class-discriminative visualisation [36]. Visual attention heat maps are used to explain the vehicle controller behaviour through segmenting and filtering simpler and more accurate maps while not degrading control accuracy [145]. A neural motion planner uses 3D detection instances with descriptive information for safe driving [146]. An interpretable tree-based representation as hybrid presentations combines rules, actions, and observation to generate multiple explanations for self-driving [147]. An architecture is used for joint scene prediction to explain object-induced actions [149]. An auto-discern system utilises surroundings observations and common-sense reasoning with answers for driving decisions [148].

Power system management Power system management normally consists of stability assessment, emergency control, power quality disturbance, and energy forecasting. CNN classifier, combined with non-intrusive load monitoring (NILM), is utilised to estimate the activation state and provide feedback for the consumer-user [150]. The shape method is firstly used in emergency control for reinforcement learning for grid control (RLGC) under three different outputs analysis [151]. Deep-SHAP is proposed for the under-voltage load shedding of power systems, and it adds feature classification of the inputs and probabilistic analysis of the outputs to increase clarity [152].

3.7.4 Domain-Specific Insights

In terms of transportation systems, operators, such as drivers and passengers, are the primary end-users in scenarios involving decisive vehicle actions, because they may want to comprehend the reasoning behind the decisions made by the autonomous system. It is very important in high-stake domains which human lives are risk. XAI can provide explanations for AI decisions to enhance the system more transparent and fostering trust. Real-time explanations pose a significant challenge for XAI in decisive vehicle actions, because decisions need to be made with fractions of a second. Rapidly changing environments such as weather conditions, pedestrian movement and other vehicles actions promote XAI should ideally make quick and accurate decisions. Moreover, every driving situation can be unique. XAI needs suitable for diversity situation and adapt its explanations which based on context-aware interoperability. As previously mentioned, XAI demands more computational resources because of real-time explanations based on timely response. Moreover, deceive vehicle actions require high dimensional sensor data, such as the inputs from LiDAR and stereo cameras, which lead the methods, like LIME and SHAP, which adopts approximate local decision boundaries, are expensive for computation and especially for high-dimensional inputs. The requirements in XAI that can generate real-time, informative explanations without overburdening the computational resources of the system.

In terms of infrastructure system management, such as power or water system management, general public, including governments and residents, are the key end-users in power system management. Government bodies want to oversee the safe and fair use of AI in power system management. Meanwhile, residents may be curious about the mechanics of AI used to manage power systems in the city. XAI can be used to evaluate AI systems for safety, fairness, transparency, and adherence to regulatory requirements. Interpretation complexity is a primary challenge for XAI in infrastructure system management due to the multidimensional nature of the data, which includes factors from power generators, transmission lines, and power consumers. Moreover, unlike the case of autonomous driving, power system operations demand more technical expertise and need to adhere to various regulatory requirements. Consequently, XAI is not only to provide coherent and insightful interpretations of the system's operations but also to demonstrate that these operations comply with all relevant regulations. The entire process in infrastructure system management is starting from generation and distribution to monitor consumer usage patterns. The complexity is future amplified by the demands for load balancing and power outages, which influences the public life and the city operation. Moreover, it also need to fix the various regulations and standers. To evidence such compliance, XAI may need to generate more complex or detailed explanations, thus increasing the computational cost.

3.8 Cross-Disciplinary Techniques for XAI Innovations

XAI innovations for cross-disciplinary refers to the advancements and developments in explainable AI (XAI) that span multiple domains and disciplines. It involves the integration and adaptation of XAI techniques and methodologies to address complex problems and challenges that arise in diverse fields.

One aspect of XAI Innovations for cross-disciplinary is the exploration and utilization of common XAI techniques across different domains. These techniques, such as attention-based models, model-agnostic methods, and rule-based methods, can be applied to various fields to provide

transparent and interpretable explanations for AI models. Below are some examples of common XAI techniques:

1. Regression-based partitioned methods:can be applied to any black-box model. For example, LIME approximates the decision boundaries of the model locally and generates explanations by highlighting the features that contribute most to the prediction for a specific instance. LIME can be used in domains such as healthcare, cyber security, finance, or education to provide instance-level interpretability and explainability. SHAP is another common technique based on cooperative game theory, which can be applied to different domains to explain the importance of features in the decision-making process. For example, in medical diagnostics, SHAP can help understand which medical parameters or biomarkers have the most impact on a particular diagnosis.
2. Feature importance: Feature importance techniques assess the relevance and contribution of each feature in the model's predictions. Methods like permutation importance, Gini importance, or gain-based importance are commonly used. Feature importance can be useful in various domains to identify the factors that drive specific outcomes or decisions. For instance, in finance, feature importance can help understand which financial indicators or market factors play a crucial role in investment decisions.
3. Partial dependence plots: Partial dependence plots visualize the relationship between a feature and the model's output while holding other features constant. These plots show how changing the value of a specific feature affects the model's predictions. Partial dependence plots can be employed in domains such as healthcare, where they can provide insights into the impact of certain medical treatments or interventions on patient outcomes.
4. Rule-based models: Rule-based models provide transparent and interpretable decision-making processes by expressing decision rules in the form of "if-then" statements. These models can be used in various domains to generate explanations that are easily understandable by humans. In legal applications, rule-based models can help explain legal reasoning by mapping legal principles and regulations to decision rules.

These are just a few examples of common XAI techniques that can be applied across different domains. The choice of technique depends on the specific requirements and characteristics of each domain. We summarise some typeical suitable XAI approaches for each domain shown in Table 5. By leveraging these techniques, domain experts and practitioners can gain insights into the inner workings of AI models and make informed decisions based on understandable and interpretable explanations.

Another aspect of XAI innovations for cross-disciplinary involves the development of domain-specific XAI approaches. In Table 5, we summarize some typical suitable XAI approaches for different domains. These approaches can be tailored to the unique characteristics and requirements of specific domains, taking into account the specific challenges and complexities of each field. Domain-specific XAI approaches consider various factors, including domain knowledge, regulations, and ethical considerations, to create an XAI framework that is specifically designed for a particular domain. By incorporating domain expertise and contextual information, these approaches provide explanations that are not only interpretable but also relevant and meaningful within their respective domains.

By tailoring XAI approaches to specific domains, practitioners can gain deeper insights into the behavior of AI models within the context of their field. This not only enhances transparency and trust in AI systems but also enables domain-specific considerations to be incorporated into the decision-making process, ensuring the explanations are relevant and aligned with the requirements and constraints of each domain.

Furthermore, XAI innovations for cross-disciplinary emphasize the importance of collaboration and the integration of expertise from different fields. This approach recognizes that the challenges and complexities of XAI extend beyond individual domains and require a multidisciplinary perspective. Collaboration and integration of expertise enable a holistic approach to XAI, where insights from different disciplines can inform the development of innovative and effective solutions. For example, in the field of healthcare, collaboration between medical practitioners, data scientists, and AI researchers can lead to the development of XAI techniques that not only provide interpretable explanations but also align with medical guidelines and regulations. This integration of expertise ensures that the explanations generated by XAI systems are not only technically sound but also relevant and meaningful in the specific healthcare context.

Similarly, in the domain of cybersecurity, collaboration between cybersecurity experts, AI specialists, and legal professionals can lead to the development of XAI techniques that address the unique challenges of cybersecurity threats. By combining knowledge from these different fields, XAI systems can provide interpretable explanations that enhance the understanding of AI-based security measures, assist in identifying vulnerabilities, and facilitate decision-making processes for cybersecurity professionals.

The collaboration and integration of expertise from different fields also foster a cross-pollination of ideas and perspectives, driving innovation and the development of novel XAI techniques. By leveraging the diverse knowledge and experiences of experts from various domains, XAI can evolve and

Table 5 XAI suitability analysis for application domains

Domain	Typical approaches	Explanation importance
Medical and biomedical	Attention-based explanations	It is suitable for medical image analysis tasks, such as identifying regions of interest or explaining the predictions of deep learning models by highlighting the important areas in an image that contribute to the model's decision
	Model-Agnostic explanations	It is suitable for medical applications where feature importance and individual instance explanations are required. It quantifies the contribution of each feature to a model's prediction and provides explanations at the individual patient level
	Rule extraction	It is suitable for medical domains where interpretable decision rules are desired, which is used to generate human-readable rules that align with medical guidelines, making it suitable for decision support systems and improving trust in AI-driven medical applications
Healthcare	Counterfactual explanations	It is suitable for healthcare applications where personalized treatment recommendations or interventions are required which can enable personalized care plans and enhancing patient engagement
	Attention-based explanations	It is suitable to analyse patient records, clinical notes, or time-series data. They enable the model to dynamically attend to important features, leading to improve interpretability and the identification of critical factors influencing healthcare outcomes
	Case-based reasoning	It is suitable to evolving knowledge, integrates expert knowledge, supports decision-making, and facilitates learning from past experiences. By leveraging historical data and real-world examples, it supports healthcare professionals in decision-making, diagnosis, treatment planning, and improving patient outcomes
Cyber security	Model-Agnostic explanations	It is suitable to identify the key factors or variables that contribute to the likelihood of a security breach or attack. By understanding the importance of different features, security analysts can prioritize their efforts and focus on mitigating the most critical vulnerabilities
	Graph-based explanations	It is suitable to aid in detecting complex cyber threats and visualizing attack relationships. It enables the identification of patterns, anomalies, and influential factors, providing explanations that enhance situational awareness, threat detection, and decision-making in cyber security operations
Finance	Model-Agnostic explanations	It is suitable to interpret complex deep learning models used for risk assessment, fraud detection, or portfolio management. It provides individual feature importance scores that help understand the factors contributing to predictions, enabling better decision-making and risk management
	Rule extraction	It is suitable for finance to help identify specific conditions or criteria that drive financial outcomes or decisions, such as loan approvals or investment recommendations
Law	Case-based reasoning	It is suitable in the legal domain as it allows for the retrieval and adaptation of past legal cases to support current decision-making, which can enhance legal decision-making by providing context-specific explanations based on past legal experiences
	Rule extraction	It is suitable in law to uncover the underlying rules and criteria used by legal systems. These techniques can help in understanding the decision-making process of legal systems and provide explanations for legal outcomes
Education and training	Model-Agnostic explanations	It is suitable to identify the key features or factors that contribute to student performance, engagement, or learning outcomes by providing localized explanations that highlight the importance and influence of specific features, such as demographic information, learning activities, or socio-economic factors, on educational outcomes
	Rule extraction	It is suitable to explain the reasoning behind educational decisions or outcomes by explicitly stating the conditions and criteria that influence educational decisions, such as grading rubrics or admission criteria
	Interactive explanations	It is suitable to present complex educational data in an understandable manner, facilitating a deeper understanding of student performance, engagement, and progress

Table 5 (continued)

Domain	Typical approaches	Explanation importance
Civil engineering	Attention-based explanations	It is suitable for image-based tasks such as defect detection, structural damage assessment, or material characterization. This can aid in the interpretation of model outputs, improve trust in the predictions, and guide subsequent inspection or repair actions
	Model-Agnostic explanations	It is suitable to explain the factors influencing the structural integrity or performance of a building or infrastructure. This information can guide engineers in making informed decisions regarding maintenance, repairs, or design modifications
	Rule extraction	It is suitable for models used for tasks such as traffic signal control, behaviour analysis, or demand forecasting. By extracting interpretable rules, such as IF-THEN statements transportation planners and policymakers can gain insights into the decision-making process of the models

adapt to meet the evolving needs and challenges of different industries and societal contexts.

4 Discussion

As the concerns on explainability and the attentions for XAI, regulations such as GDPR set out the transparency rules about the data processing. As most modern AI systems are data-driven AI, these requirements are actually applicable to all application domains. Not only the explainability is necessary, but also the way of explaining is required.

In this section, we will summarize the limitations of existing XAI approaches based on the above review in each application domain, and identify future research directions.

4.1 Limitations

Adaptive integration and explanation: many existing approaches provide explanations in a generic manner, without considering the diverse backgrounds (culture, context, etc.) and knowledge levels of users. This one-size-fits-all approach can lead to challenges in effective comprehension for both novice and expert users. Novice users may struggle to understand complex technical explanations, while expert users may find oversimplified explanations lacking in depth. These limitations hinder the ability of XAI techniques to cater to users with different levels of expertise and may impact the overall trust and usability of the system. Furthermore, the evaluation and assessment of XAI techniques often prioritize objective metrics, such as fidelity or faithfulness, which measure how well the explanations align with the model's internal workings. While these metrics are important for evaluating the accuracy of the explanations, they may not capture the subjective aspects of user understanding and interpretation. The perceived quality of explanations can vary among users with different expertise levels, as well as under different situations or conditions.

Interactive explanation: in the current landscape of XAI research, there is recognition that a single explanation may not be sufficient to address all user concerns and questions in decision-making scenarios. As a result, the focus has shifted towards developing interactive explanations that allow for a dynamic and iterative process. However, there are challenges that need to be addressed in order to effectively implement interactive explanation systems. One of the key challenges is the ability to handle a wide range of user queries and adapt the explanations accordingly. Users may have diverse information needs and may require explanations that go beyond superficial or generic responses. In particular, addressing queries that involve deep domain knowledge or intricate reasoning processes can be complex and requires sophisticated techniques. Another challenge is striking a balance between providing timely responses to user queries and maintaining computational efficiency. Interactive explanation systems need to respond quickly to user interactions to facilitate a smooth and engaging user experience. However, generating accurate and informative explanations within a short response time can be demanding, and trade-offs may need to be made depending on the specific domain and computational resources available. Moreover, the design and implementation of interactive explanation systems should also consider the context and domain-specific requirements. Different domains may have unique challenges and constraints that need to be taken into account when developing interactive explanations. It is important to ensure that the interactive explanation systems are tailored to the specific domain and can effectively address the needs of users in that context.

Connection and consistency in hybrid explanation: in the context of hybrid explanations in XAI, it is crucial to ensure connection and consistency among different sources of explanations. Hybrid approaches aim to leverage multiple techniques to provide users in various domains with different application purposes, achieving robustness and interpretability. However, it is necessary to address potential conflicts and ensure coordinated integration of different components

within these hybrid systems. Currently, many works focus on combining various explanation techniques to complement each other and enhance overall system performance. While this integration is valuable, it is important to acknowledge that different techniques may have inherent differences in their assumptions, methodologies, and outputs. These differences can result in conflicts or inconsistencies when combined within a hybrid explanation system. Therefore, careful attention should be given to the design of complex hybrid explanation systems. The structure and architecture need to be thoughtfully planned to ensure seamless connections between components. This involves identifying potential conflicts early on and developing strategies to resolve them. Additionally, efforts should be made to establish a unified framework that allows for effective coordination and integration of the different techniques used in the hybrid system. Furthermore, the evaluation and validation of hybrid explanation systems should include assessing the consistency of explanations provided by different sources. This evaluation process helps identify any discrepancies or inconsistencies and guides the refinement of the system to ensure a coherent and unified user experience.

Balancing model interpretability with predictive accuracy: currently, researchers are developing hybrid approaches that aim to strike a better balance between interpretability and accuracy, such as using post-hoc interpretability techniques with complex models or designing new model architectures that inherently provide both interpretability and high accuracy.However, they also come with their own limitations. Post-hoc interpretability techniques generate explanations after the model has made its predictions, which means they do not directly influence the model's decision-making process. As a result, the explanations may not capture the full complexity and nuances of the model's internal workings. Furthermore, post-hoc techniques can be computationally expensive and may not scale well to large datasets or complex models with high-dimensional inputs. Designing new model architectures such as rule-based models or attention mechanisms in neural networks may struggle to capture complex interactions and may require a significant amount of manual rule engineering. It is crucial to recognize that there is no universal solution to the interpretability-accuracy trade-off. The choice of approach depends on the specific requirements of the application, available resources, and acceptable trade-offs in the given context. Researchers and practitioners must carefully consider the limitations and benefits of different techniques to strike an appropriate balance based on their specific use cases.

Long-term usability and maintainability: the current XAI methods face several limitations when deployed in real-world scenarios. One significant limitation is the need for continuous explanation updates. XAI systems generate explanations based on training data, and as the underlying AI models or data evolve, the explanations may become outdated or less accurate. To ensure relevance and usefulness, XAI systems should be designed to incorporate mechanisms for updating explanations to reflect the latest model updates or data changes. Another limitation is the assumption of stationary data distributions. XAI methods are typically trained on historical data, assuming that the future data will follow a similar distribution. However, if the data distribution changes over time, the performance of the XAI system may deteriorate. Adapting XAI methods to handle shifting data distributions is essential for maintaining their effectiveness and ensuring reliable explanations in dynamic environments. Scalability is another crucial consideration, particularly for large-scale AI systems. XAI techniques that work well on small-scale or controlled datasets may face challenges when applied to large-scale AI systems with complex models and massive amounts of data. Efficient algorithms and sufficient computational resources are necessary to handle the increased computational demands of explaining large-scale AI systems without sacrificing performance or usability.

4.2 Future Directions

To address the first limitation, building the context-awareness XAI is important, we need to explore how to generate explanations by considering mission contexts (surrounding environment, situations, time-series datasets.), mapping user roles (end-user, domain expert, business manager, AI developer, etc.) and targeted goals (refine the model, debugging system errors, detecting bias, understand AI learning process, etc.) regardless of the type of AI system. So far, most of these studies were still conceptual with limited consideration, the more general context-driven systems and practical implementations will be an important direction for future research.

Secondly, interactive explanations (e.g., conversation system Interfaces, games, using audio, visuals, video, etc.) should be explored further. This is a promising approach to building truly human-centred explanations by identifying users' requirements and providing better human-AI collaboration. These incorporating theories and frameworks allow an iterative process from humans, which is a crucial aspect of building successful XAI systems.

Finally, the hybrid explanation should be applied by concerning fusing heterogeneous knowledge from different sources, managing time-sensitive data, inconsistency, uncertainty, etc. Among these conditions, hybrid explanation has been an interesting and increasing topic in recent years. This will also involve a wide range of criteria and strategies that target a clear structure and consensus on what constitutes success and trustworthy explanations.

5 Conclusion

This paper addresses a wide range of explainable AI topics. XAI is a rapidly growing field of research, as it fills a gap in current AI approaches, allowing people to better understand AI models and therefore trust their outputs. By summarising the current literature, we have proposed a new taxonomy for XAI from the human perspective. The taxonomy considers source-oriented, representation-oriented aspect, and logic-oriented perspectives.

It is very important that we have elaborated on the applications of XAI in multiple areas in the paper, including medical, healthcare, cybersecurity, finance and law, education and training, and civil engineering. We provide a comprehensive review of different XAI approaches and identify the key techniques for case studies. Finally, we discuss the limitations of existing XAI methods and present several corresponding areas for further research: (1) context-awareness XAI, (2) interactive explanations, and (3) hybrid explanations.

Overall, this paper provides a clear survey of the current XAI research and application status from the human perspective. We hope this article will provide a valuable reference for XAI-related researchers and practitioners. We believe XAI will build a bridge of trust between humans and AI.

Acknowledgements This work was supported by the Institute of Information and Communications Technology Planning and Evaluation (IITP) grant funded by the Korean Government (MSIT) (2022-0-00078, Explainable Logical Reasoning for Medical Knowledge Generation).

Author Contributions The authors confirm their contribution to the paper as follows: study conception and design: WY; draft manuscript preparation: YW, HW, YC, GH, XL, RL, NY, XW, and XG. Supervision: MBA and BK. All authors reviewed the results and approved the final version of the manuscript.

Data availability Not applicable.

Declarations

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

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Zhang Y, Tiňo P, Leonardis A, Tang K. A survey on neural network interpretability. IEEE Trans Emerg Top Comput Intell. 2021;20:20.
2. Tomsett R, Preece A, Braines D, Cerutti F, Chakraborty S, Srivastava M, Pearson G, Kaplan L. Rapid trust calibration through interpretable and uncertainty-aware AI. Patterns. 2020;1(4):100049.
3. 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. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fusion. 2020;58:82–115.
4. Regulation GDP. General data protection regulation (GDPR). Intersoft Consult. 2018;24:1.
5. Bostrom N, Yudkowsky E. The ethics of artificial intelligence. In: Artificial intelligence safety and security. New York: Chapman and Hall; 2018. p. 57–69.
6. Weld DS, Bansal G. The challenge of crafting intelligible intelligence. Commun ACM. 2019;62(6):70–9.
7. Das A, Rad P. Opportunities and challenges in explainable artificial intelligence (XAI): a survey. arXiv:2006.11371 (arXiv preprint) (2020).
8. Challen R, Denny J, Pitt M, Gompels L, Edwards T, Tsaneva-Atanasova K. Artificial intelligence, bias and clinical safety. BMJ Qual Saf. 2019;28(3):231–7.
9. Patil MS, Främling K. Context, utility and influence of an explanation. arXiv:2303.13552 (arXiv preprint); 2023.
10. Ooge J, Verbert K. Explaining artificial intelligence with tailored interactive visualisations. In: 27th international conference on intelligent user interfaces; 2022. p. 120–3.
11. Saeed W, Omlin C. Explainable AI (XAI): a systematic meta-survey of current challenges and future opportunities. Knowl Based Syst. 2023;11:0273.
12. Förster M, Klier M, Kluge K, Sigler I. Fostering human agency: a process for the design of user-centric XAI systems; 2020.
13. Kotriwala A, Klöpper B, Dix M, Gopalakrishnan G, Ziobro D, Potschka A. Xai for operations in the process industry-applications, theses, and research directions. In: AAAI spring symposium: combining machine learning with knowledge engineering; 2021.
14. Albahri A, Duhaim AM, Fadhel MA, Alnoor A, Baqer NS, Alzubaidi L, Albahri O, Alamoodi A, Bai J, Salhi A, et al. A systematic review of trustworthy and explainable artificial intelligence in healthcare: assessment of quality, bias risk, and data fusion. Inf Fusion. 2023;20:20.
15. Kurshan E, Chen J, Storchan V, Shen H. On the current and emerging challenges of developing fair and ethical AI solutions in financial services. In: Proceedings of the second ACM international conference on AI in finance; 2021. p. 1–8.
16. Komorowski P, Baniecki H, Biecek P. Towards evaluating explanations of vision transformers for medical imaging. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2023. p. 3725–3731.
17. Adadi A, Berrada M. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access. 2018;6:52138–60.
18. Minh D, Wang HX, Li YF, Nguyen TN. Explainable artificial intelligence: a comprehensive review. Artif Intell Rev. 2021;20:1–66.
19. Chaddad A, Peng J, Xu J, Bouridane A. Survey of explainable AI techniques in healthcare. Sensors. 2023;23(2):634.

20. Tjoa E, Guan C. A survey on explainable artificial intelligence (XAI): toward medical XAI. IEEE Trans Neural Netw Learn Syst. 2020;32(11):4793–813.
21. Angelov PP, Soares EA, Jiang R, Arnold NI, Atkinson PM. Explainable artificial intelligence: an analytical review. Wiley Interdiscip Rev Data Min Knowl Discov. 2021;11(5):1424.
22. Vilone G, Longo L. Classification of explainable artificial intelligence methods through their output formats. Mach Learn Knowl Extract. 2021;3(3):615–61.
23. Alain G, Bengio Y. Understanding intermediate layers using linear classifier probes. arXiv:1610.01644 (arXiv preprint); 2016.
24. Zhang Q, Cao R, Shi F, Wu YN, Zhu S-C. Interpreting CNN knowledge via an explanatory graph. In: Proceedings of the AAAI conference on artificial intelligence, vol. 32; 2018.
25. Hendricks LA, Hu R, Darrell T, Akata Z. Grounding visual explanations. In: Proceedings of the European conference on computer vision (ECCV); 2018. p. 264–79.
26. Bondarenko A, Aleksejeva L, Jumutc V, Borisov A. Classification tree extraction from trained artificial neural networks. Proced Comput Sci. 2017;104:556–63.
27. Zhou Z-H, Jiang Y, Chen S-F. Extracting symbolic rules from trained neural network ensembles. AI Commun. 2003;16(1):3–15.
28. Barakat N, Diederich J. Eclectic rule-extraction from support vector machines. Int J Comput Intell. 2005;2(1):59–62.
29. Nikolov A, d'Aquin M. Uncovering semantic bias in neural network models using a knowledge graph. In: Proceedings of the 29th ACM international conference on information and knowledge management; 2020, p. 1175–84.
30. Riquelme F, De Goyeneche A, Zhang Y, Niebles JC, Soto A. Explaining VQA predictions using visual grounding and a knowledge base. Image Vis Comput. 2020;101:103968.
31. Erion G, Janizek JD, Sturmfels P, Lundberg SM, Lee S-I. Learning explainable models using attribution priors; 2019.
32. Robnik-Šikonja M, Bohanec M. Perturbation-based explanations of prediction models. In: Human and machine learning. Berlin: Springer; 2018. p. 159–75.
33. Laugel T, Lesot M-J, Marsala C, Renard X, Detyniecki M. The dangers of post-hoc interpretability: Unjustified counterfactual explanations. arXiv:1907.09294 (arXiv preprint); 2019.
34. Chefer H, Gur S, Wolf L. Transformer interpretability beyond attention visualization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2021. p. 782–91.
35. Jalaboi R, Faye F, Orbes-Arteaga M, Jørgensen D, Winther O, Galimzianova A. Dermx: an end-to-end framework for explainable automated dermatological diagnosis. Med Image Anal. 2023;83:102647.
36. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision; 2017. p. 618–26.
37. Graziani M, Andrearczyk V, Marchand-Maillet S, Müller H. Concept attribution: explaining CNN decisions to physicians. Comput Biol Med. 2020;123:103865.
38. Zhang Q, Wu YN, Zhu S-C. Interpretable convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2018. p. 8827–836.
39. Liang X, Hu Z, Zhang H, Lin L, Xing EP. Symbolic graph reasoning meets convolutions. Adv Neural Inf Process Syst. 2018;31:25.
40. Li CY, Liang X, Hu Z, Xing EP. Knowledge-driven encode, retrieve, paraphrase for medical image report generation. In: Proceedings of the AAAI conference on artificial intelligence; 2019. vol. 33, p. 6666–73.
41. Ribeiro MT, Singh S, Guestrin C. Anchors: High-precision model-agnostic explanations. In: Proceedings of the AAAI conference on artificial intelligence; 2018. vol. 32.
42. Teng F, Yang W, Chen L, Huang L, Xu Q. Explainable prediction of medical codes with knowledge graphs. Front Bioeng Biotechnol. 2020;8:867.
43. Sun P, Gu L. Fuzzy knowledge graph system for artificial intelligence-based smart education. J Intell Fuzzy Syst. 2021;40(2):2929–40.
44. Panchenko A, Ruppert E, Faralli S, Ponzetto SP, Biemann C. Unsupervised does not mean uninterpretable: the case for word sense induction and disambiguatio; 2017. Association for Computational Linguistics.
45. Bennetot A, Laurent J-L, Chatila R, Díaz-Rodríguez N. Towards explainable neural-symbolic visual reasoning. arXiv:1909.09065 (arXiv preprint); 2019.
46. Tamagnini P, Krause J, Dasgupta A, Bertini E. Interpreting black-box classifiers using instance-level visual explanations. In: Proceedings of the 2nd workshop on human-in-the-loop data analytics; 2017. p. 1–6.
47. Spinner T, Schlegel U, Schäfer H, El-Assady M. Explainer: a visual analytics framework for interactive and explainable machine learning. IEEE Trans Visual Comput Graph. 2019;26(1):1064–74.
48. Hendricks LA, Akata Z, Rohrbach M, Donahue J, Schiele B, Darrell T. Generating visual explanations. In: European conference on computer vision. Springer; 2016. p. 3–19.
49. Kim J, Rohrbach A, Darrell T, Canny J, Akata Z. Textual explanations for self-driving vehicles. In: Proceedings of the European conference on computer vision (ECCV); 2018. p. 563–78.
50. Park DH, Hendricks LA, Akata Z, Rohrbach A, Schiele B, Darrell T, Rohrbach M. Multimodal explanations: justifying decisions and pointing to the evidence. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2018. p. 8779–8788.
51. Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhudinov R, Zemel R, Bengio Y. Show, attend and tell: neural image caption generation with visual attention. In: International conference on machine learning; 2015. p. 2048–2057.
52. Gu D, Li Y, Jiang F, Wen Z, Liu S, Shi W, Lu G, Zhou C. Vinet: a visually interpretable image diagnosis network. IEEE Trans Multimed. 2020;22(7):1720–9.
53. Slack D, Hilgard S, Jia E, Singh S, Lakkaraju H. Fooling lime and shap: adversarial attacks on post hoc explanation methods. In: Proceedings of the AAAI/ACM conference on AI, ethics, and society; 2020. p. 180–86.
54. Zhang Z, Rudra K, Anand A. Explain and predict, and then predict again. In: Proceedings of the 14th ACM international conference on web search and data mining; 2021. p. 418–26.
55. Montavon G, Binder A, Lapuschkin S, Samek W, Müller K-R. Layer-wise relevance propagation: an overview. Explainable AI: interpreting, explaining and visualizing deep learning; 2019. 193–209.
56. Zhang Z, Chen P, McGough M, Xing F, Wang C, Bui M, Xie Y, Sapkota M, Cui L, Dhillon J, et al. Pathologist-level interpretable whole-slide cancer diagnosis with deep learning. Nat Mach Intell. 2019;1(5):236–45.
57. Sarlin P-E, DeTone D, Malisiewicz T, Rabinovich A. Superglue: learning feature matching with graph neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2020. p. 4938–47.
58. Shen S, Han SX, Aberle DR, Bui AA, Hsu W. Explainable hierarchical semantic convolutional neural network for lung cancer diagnosis. In: CVPR workshops; 2019. p. 63–6.

59. Gozzi N, Malandri L, Mercorio F, Pedrocchi A. Xai for myo-controlled prosthesis: explaining EMG data for hand gesture classification. Knowl-Based Syst. 2022;240:108053.
60. Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: visualising image classification models and saliency maps. arXiv:1312.6034 (arXiv preprint); 2013.
61. Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. In: European conference on computer vision. Springer; 2014. p. 818–33.
62. Springenberg JT, Dosovitskiy A, Brox T, Riedmiller M. Striving for simplicity: the all convolutional net. arXiv:1412.6806 (arXiv preprint); 2014.
63. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. p. 2921–9.
64. Olah C, Mordvintsev A, Schubert L. Feature visualization. Distill. 2017;2(11):7.
65. Zhang Z, Xie Y, Xing F, McGough M, Yang L. Mdnet: a semantically and visually interpretable medical image diagnosis network. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2017. p. 6428–36.
66. Kim B, Wattenberg M, Gilmer J, Cai C, Wexler J, Viegas F, et al. Interpretability beyond feature attribution: quantitative testing with concept activation vectors (TCAV). In: International conference on machine learning; 2018. p. 2668–77.
67. Wu B, Zhou Z, Wang J, Wang Y. Joint learning for pulmonary nodule segmentation, attributes and malignancy prediction. In: 2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018); 2018. p. 1109–13. IEEE.
68. de Vos BD, Wolterink JM, Leiner T, de Jong PA, Lessmann N, Išgum I. Direct automatic coronary calcium scoring in cardiac and chest CT. IEEE Trans Med Imaging. 2019;38(9):2127–38.
69. Eitel F, Ritter K, ADNI. Testing the robustness of attribution methods for convolutional neural networks in MRI-based Alzheimer's disease classification. In: Interpretability of machine intelligence in medical image computing and multimodal learning for clinical decision support. Berlin: Springer; 2019. p. 3–11.
70. Clough JR, Oksuz I, Puyol-Antón E, Ruijsink B, King AP, Schnabel J.A. Global and local interpretability for cardiac MRI classification. In: International conference on medical image computing and computer-assisted intervention; Springer. 2019. p. 656–4.
71. Gasimova A. Automated enriched medical concept generation for chest X-ray images. In: Interpretability of machine intelligence in medical image computing and multimodal learning for clinical decision support. Springer; 2019. p. 83–92.
72. Kim ST, Lee J-H, Ro YM. Visual evidence for interpreting diagnostic decision of deep neural network in computer-aided diagnosis. In: Medical imaging 2019: computer-aided diagnosis. 2019; vol. 10950, p. 139–47. SPIE.
73. Lee H, Kim ST, Ro YM. Generation of multimodal justification using visual word constraint model for explainable computer-aided diagnosis. In: Interpretability of machine intelligence in medical image computing and multimodal learning for clinical decision support. Springer; 2019. p. 21–9.
74. Shen S, Han SX, Aberle DR, Bui AA, Hsu W. An interpretable deep hierarchical semantic convolutional neural network for lung nodule malignancy classification. Expert Syst Appl. 2019;128:84–95.
75. Arun N, Gaw N, Singh P, Chang K, Aggarwal M, Chen B, et al. Assessing the (un) trustworthiness of saliency maps for localizing abnormalities in medical imaging (arXiv preprint); 2020.
76. Zeng X, Wen L, Xu Y, Ji C. Generating diagnostic report for medical image by high-middle-level visual information incorporation on double deep learning models. Comput Methods Programs Biomed. 2020;197:105700.
77. Yang S, Niu J, Wu J, Liu X. Automatic medical image report generation with multi-view and multi-modal attention mechanism. In: International conference on algorithms and architectures for parallel processing. Springer; 2020. p. 687–99.
78. Barnett AJ, Schwartz FR, Tao C, Chen C, Ren Y, Lo JY, Rudin C. A case-based interpretable deep learning model for classification of mass lesions in digital mammography. Nat Mach Intell. 2021;3(12):1061–70.
79. Saleem H, Shahid AR, Raza B. Visual interpretability in 3d brain tumor segmentation network. Comput Biol Med. 2021;133:104410.
80. Wang S, Yin Y, Wang D, Wang Y, Jin Y. Interpretability-based multimodal convolutional neural networks for skin lesion diagnosis. IEEE Trans Cybern. 2021;20:20.
81. Ahmed U, Jhaveri RH, Srivastava G, Lin JC-W. Explainable deep attention active learning for sentimental analytics of mental disorder. Trans Asian Low-Resour Lang Inf Proces. 2022;20:22.
82. Lu Y, Perer A. An interactive interpretability system for breast cancer screening with deep learning. arXiv:2210.08979 (arXiv preprint); 2022.
83. Figueroa KC, Song B, Sunny S, Li S, Gurushanth K, Mendonca P, Mukhia N, Patrick S, Gurudath S, Raghavan S, et al. Interpretable deep learning approach for oral cancer classification using guided attention inference network. J Biomed Opt. 2022;27(1):015001.
84. Hicks SA, Eskeland S, Lux M, de Lange T, Randel KR, Jeppsson M, Pogorelov K, Halvorsen P, Riegler M. Mimir: an automatic reporting and reasoning system for deep learning based analysis in the medical domain. In: Proceedings of the 9th ACM multimedia systems conference; 2018. p. 369–74.
85. Holzinger A, Malle B, Saranti A, Pfeifer B. Towards multi-modal causability with graph neural networks enabling information fusion for explainable AI. Inf Fusion. 2021;71:28–37.
86. Palatnik de Sousa I, Maria Bernardes Rebuzzi Vellasco M, Costa da Silva E. Local interpretable model-agnostic explanations for classification of lymph node metastases. Sensors. 2019;19(13):2969.
87. Zhu P, Ogino M. Guideline-based additive explanation for computer-aided diagnosis of lung nodules. In: Interpretability of machine intelligence in medical image computing and multimodal learning for clinical decision support. Springer; 2019; p. 39–47.
88. Paschali M, Ferjadnaeem M, Simson W, et al. Improving the interpretability of medical imaging neural networks. In: Computer vision and pattern recognition; 2019.
89. Liao W, Zou B, Zhao R, Chen Y, He Z, Zhou M. Clinical interpretable deep learning model for glaucoma diagnosis. IEEE J Biomed Health Inform. 2019;24(5):1405–12.
90. Lee H, Yune S, Mansouri M, Kim M, Tajmir SH, Guerrier CE, Ebert SA, Pomerantz SR, Romero JM, Kamalian S, et al. An explainable deep-learning algorithm for the detection of acute intracranial haemorrhage from small datasets. Nat Biomed Eng. 2019;3(3):173–82.
91. Biffi C, Oktay O, Tarroni G, Bai W, De Marvao A, Doumou G, Rajchl M, Bedair R, Prasad S, Cook S, et al. Learning interpretable anatomical features through deep generative models: application to cardiac remodeling. In: International conference on medical image computing and computer-assisted intervention. Springer; 2018. p. 464–71.
92. Garcia-Peraza-Herrera LC, Everson M, Li W, Luengo I, Berger L, Ahmad O, Lovat L, Wang H-P, Wang W-L, Haidry R, et al. Interpretable fully convolutional classification of intrapapillary capillary loops for real-time detection of early squamous neoplasia. arXiv:1805.00632 (arXiv preprint); 2018.
93. Amoroso N, Pomarico D, Fanizzi A, Didonna V, Giotta F, La Forgia D, Latorre A, Monaco A, Pantaleo E, Petruzzellis N, et al.

A roadmap towards breast cancer therapies supported by explainable artificial intelligence. Appl Sci. 2021;11(11):4881.
94. Sarp S, Kuzlu M, Wilson E, Cali U, Guler O. The enlightening role of explainable artificial intelligence in chronic wound classification. Electronics. 2021;10(12):1406.
95. Wu H, Chen W, Xu S, Xu B. Counterfactual supporting facts extraction for explainable medical record based diagnosis with graph network. In: Proceedings of the 2021 conference of the north American chapter of the association for computational linguistics: human language technologies; 2021. p. 1942–55.
96. Wang X, Peng Y, Lu L, Lu Z, Summers RM. Tienet: Text-image embedding network for common thorax disease classification and reporting in chest X-rays. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2018. p. 9049–58.
97. Lucieri A, Bajwa MN, Braun SA, Malik MI, Dengel A, Ahmed S. On interpretability of deep learning based skin lesion classifiers using concept activation vectors. In: 2020 international joint conference on neural networks (IJCNN); 2020. p. 1–10. IEEE.
98. Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, Qin C, Žídek A, Nelson AW, Bridgland A, et al. Improved protein structure prediction using potentials from deep learning. Nature. 2020;577(7792):706–10.
99. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, et al. Highly accurate protein structure prediction with alphafold. Nature. 2021;596(7873):583–9.
100. Merk D, Friedrich L, Grisoni F, Schneider G. De novo design of bioactive small molecules by artificial intelligence. Mol Inf. 2018;37(1–2):1700153.
101. Zhavoronkov A, Ivanenkov YA, Aliper A, Veselov MS, Aladinskiy VA, Aladinskaya AV, Terentiev VA, Polykovskiy DA, Kuznetsov MD, Asadulaev A, et al. Deep learning enables rapid identification of potent ddr1 kinase inhibitors. Nat Biotechnol. 2019;37(9):1038–40.
102. Karimi M, Wu D, Wang Z, Shen Y. Explainable deep relational networks for predicting compound-protein affinities and contacts. J Chem Inf Model. 2020;61(1):46–66.
103. Ezzat D, Hassanien AE, Ella HA. An optimized deep learning architecture for the diagnosis of covid-19 disease based on gravitational search optimization. Appl Soft Comput. 2021;98:106742.
104. Segler MH, Kogej T, Tyrchan C, Waller MP. Generating focused molecule libraries for drug discovery with recurrent neural networks. ACS Cent Sci. 2018;4(1):120–31.
105. Preuer K, Renz P, Unterthiner T, Hochreiter S, Klambauer G. Fréchet chemnet distance: a metric for generative models for molecules in drug discovery. J Chem Inf Model. 2018;58(9):1736–41.
106. Wan Y, Zhou H, Zhang X. An interpretation architecture for deep learning models with the application of covid-19 diagnosis. Entropy. 2021;23(2):204.
107. Loh HW, Ooi CP, Seoni S, Barua PD, Molinari F, Acharya UR. Application of explainable artificial intelligence for healthcare: a systematic review of the last decade (2011–2022). Comput Methods Programs Biomed. 2022;20: 107161.
108. Duckworth C, Chmiel FP, Burns DK, Zlatev ZD, White NM, Daniels TW, Kiuber M, Boniface MJ. Using explainable machine learning to characterise data drift and detect emergent health risks for emergency department admissions during covid-19. Sci Rep. 2021;11(1):1–10.
109. Antoniadi AM, Galvin M, Heverin M, Hardiman O, Mooney C. Prediction of caregiver quality of life in amyotrophic lateral sclerosis using explainable machine learning. Sci Rep. 2021;11(1):1–13.
110. Zeng X, Hu Y, Shu L, Li J, Duan H, Shu Q, Li H. Explainable machine-learning predictions for complications after pediatric congenital heart surgery. Sci Rep. 2021;11(1):1–11.
111. Farhadloo M, Molnar C, Luo G, Li Y, Shekhar S, Maus RL, Markovic S, Leontovich A, Moore R. Samcnet: towards a spatially explainable AI approach for classifying MXIF oncology data. In: Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining; 2022. p. 2860–70.
112. Jiang Z, Bo L, Xu Z, Song Y, Wang J, Wen P, Wan X, Yang T, Deng X, Bian J. An explainable machine learning algorithm for risk factor analysis of in-hospital mortality in sepsis survivors with ICU readmission. Comput Methods Programs Biomed. 2021;204:106040.
113. Liu H, Zhong C, Alnusair A, Islam SR. Faixid: a framework for enhancing AI explainability of intrusion detection results using data cleaning techniques. J Netw Syst Manage. 2021;29(4):1–30.
114. Amarasinghe K, Manic M. Improving user trust on deep neural networks based intrusion detection systems. In: IECON 2018-44th annual conference of the IEEE Industrial electronics society; 2018. p. 3262–68. IEEE.
115. Amarasinghe K, Kenney K, Manic M. Toward explainable deep neural network based anomaly detection. In: 2018 11th international conference on human system interaction (HSI); 2018. IEEE. p. 311–7.
116. Chen S, Bateni S, Grandhi S, Li X, Liu C, Yang W. Denas: automated rule generation by knowledge extraction from neural networks. In: Proceedings of the 28th ACM joint meeting on European software engineering conference and symposium on the foundations of software engineering, 2020; p. 813–25.
117. Gulmezoglu B. Xai-based microarchitectural side-channel analysis for website fingerprinting attacks and defenses. IEEE Trans Depend Sec Comput. 2021;20:10.
118. Feichtner J, Gruber S. Understanding privacy awareness in android app descriptions using deep learning. In: Proceedings of the tenth ACM conference on data and application security and privacy; 2020. p. 203–14.
119. Iadarola G, Martinelli F, Mercaldo F, Santone A. Towards an interpretable deep learning model for mobile malware detection and family identification. Comput Secur. 2021;105:102198.
120. Guo W, Mu D, Xu J, Su P, Wang G, Xing X. Lemna: explaining deep learning based security applications. In: Proceedings of the 2018 ACM SIGSAC conference on computer and communications security; 2018. p. 364–79.
121. Yan A, Chen Z, Zhang H, Peng L, Yan Q, Hassan MU, Zhao C, Yang B. Effective detection of mobile malware behavior based on explainable deep neural network. Neurocomputing. 2021;453:482–92.
122. Bach S, Binder A, Montavon G, Klauschen F, Müller K-R, Samek W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS One. 2015;10(7):0130140.
123. Mane S, Rao D. Explaining network intrusion detection system using explainable AI framework. arXiv:2103.07110 (arXiv preprint); 2021.
124. Dash S, Gunluk O, Wei D. Boolean decision rules via column generation. Adv Neural Inf Process Syst. 2018;31:25.
125. Bose S, Barao T, Liu X. Explaining AI for malware detection: analysis of mechanisms of malconv. In: 2020 international joint conference on neural networks (IJCNN); 2020. IEEE. p. 1–8.
126. Al-Fawa'reh M, Saif A, Jafar MT, Elhassan A. Malware detection by eating a whole APK. In: 2020 15th international conference for internet technology and secured transactions (ICITST); 2020. IEEE. p. 1–7.
127. Ohana JJ, Ohana S, Benhamou E, Saltiel D, Guez B. Explainable AI (XAI) models applied to the multi-agent environment of financial markets. In: International workshop on explainable, transparent autonomous agents and multi-agent systems. Springer; 2021. p. 189–207.

128. Gramegna A, Giudici P. Shap and lime: an evaluation of discriminative power in credit risk. Front Artif Intell. 2021;140:25.
129. Wijnands M. Explaining black box decision-making: adopting explainable artificial intelligence in credit risk prediction for p2p lending. Master's thesis, University of Twente; 2021.
130. El Qadi A, Trocan M, Diaz-Rodriguez N, Frossard T. Feature contribution alignment with expert knowledge for artificial intelligence credit scoring. Signal, Image and Video Processing; 2022. 1–8.
131. de Lange PE, Melsom B, Vennerød CB, Westgaard S. Explainable AI for credit assessment in banks. J Risk Financ Manage. 2022;15(12):556.
132. Górski Ł, Ramakrishna S. Explainable artificial intelligence, lawyer's perspective. In: Proceedings of the eighteenth international conference on artificial intelligence and law; 2021. p. 60–8.
133. Berk RA, Bleich J. Statistical procedures for forecasting criminal behavior: a comparative assessment. Criminol Pub Pol'y. 2013;12:513.
134. Mardaoui D, Garreau D. An analysis of lime for text data. In: International conference on artificial intelligence and statistics; 2021. p. 3493–501. PMLR.
135. Khosravi H, Shum SB, Chen G, Conati C, Tsai Y-S, Kay J, Knight S, Martinez-Maldonado R, Sadiq S, Gašević D. Explainable artificial intelligence in education. Comput Educ Artif Intell. 2022;3:100074.
136. Alonso JM, Casalino G. Explainable artificial intelligence for human-centric data analysis in virtual learning environments. In: International workshop on higher education learning methodologies and technologies online. Springer; 2019. p. 125–38.
137. Ghai B, Liao QV, Zhang Y, Bellamy R, Mueller K. Explainable active learning (XAL) toward AI explanations as interfaces for machine teachers. Proc ACM Human Comput Interact. 2021;4(CSCW3):1–28.
138. Hu Y, Mello RF, Gašević D. Automatic analysis of cognitive presence in online discussions: an approach using deep learning and explainable artificial intelligence. Comput Educ Artif Intell. 2021;2:100037.
139. Hooshyar D, Yang Y. Neural-symbolic computing: a step toward interpretable AI in education. Bull Tech Committee Learn Technol (ISSN: 2306-0212) 2021;21(4), 2–6.
140. Melo E, Silva I, Costa DG, Viegas CM, Barros TM. On the use of explainable artificial intelligence to evaluate school dropout. Educ Sci. 2022;12(12):845.
141. Fernandez-Nieto GM, Echeverria V, Shum SB, Mangaroska K, Kitto K, Palominos E, Axisa C, Martinez-Maldonado R. Storytelling with learner data: guiding student reflection on multimodal team data. IEEE Trans Learn Technol. 2021;14(5):695–708.
142. Knight S, Shibani A, Abel S, Gibson A, Ryan P. Acawriter: a learning analytics tool for formative feedback on academic writing. J Writing Res. 2020;20:20.
143. Conati C, Barral O, Putnam V, Rieger L. Toward personalized XAI: a case study in intelligent tutoring systems. Artif Intell. 2021;298:103503.
144. Bojarski M, Choromanska A, Choromanski K, Firner B, Jackel L, Muller U, Zieba K. Visualbackprop: visualizing cnns for autonomous driving. arXiv:1611.05418 (arXiv preprint); 2016.
145. Kim J, Canny J. Interpretable learning for self-driving cars by visualizing causal attention. In: Proceedings of the IEEE international conference on computer vision; 2017. p. 2942–50.
146. Zeng W, Luo W, Suo S, Sadat A, Yang B, Casas S, Urtasun R. End-to-end interpretable neural motion planner. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2019. p. 8660–9.
147. Omeiza D, Webb H, Jirotka M, Kunze L. Explanations in autonomous driving: a survey. IEEE Trans Intell Transport Syst. 2021;20:20.
148. Kothawade S, Khandelwal V, Basu K, Wang H, Gupta G. Autodiscern: autonomous driving using common sense reasoning. arXiv:2110.13606 (arXiv preprint); 2021.
149. Gao Y, Zhang S, Sun J, Yu S, Yamamoto T, Li Z, Li X. A joint framework based on accountable AI for driving behavior assessment and backtracking. In: 2022 IEEE 25th international conference on intelligent transportation systems (ITSC); 2022. IEEE. p. 268–74.
150. Machlev R, Heistrene L, Perl M, Levy K, Belikov J, Mannor S, Levron Y. Explainable artificial intelligence (XAI) techniques for energy and power systems: review, challenges and opportunities. Energy AI. 2022; 20:100169.
151. Zhang K, Xu P, Zhang J. Explainable AI in deep reinforcement learning models: a shap method applied in power system emergency control. In: 2020 IEEE 4th conference on energy internet and energy system integration (EI2); 2020. IEEE. p. 711–6.
152. Zhang K, Zhang J, Xu P-D, Gao T, Gao DW. Explainable AI in deep reinforcement learning models for power system emergency control. IEEE Trans Comput Soc Syst. 2021;9(2):419–27.
153. Shen D, Wu G, Suk H-I. Deep learning in medical image analysis. Annu Rev Biomed Eng. 2017;19:221–48.
154. Thompson AC, Jammal AA, Medeiros FA. A review of deep learning for screening, diagnosis, and detection of glaucoma progression. Transl Vis Sci Technol. 2020;9(2):42–42.
155. Moolayil J. An introduction to deep learning and Keras. In: Learn Keras for deep neural networks. Berlin: Springer; 2019. p. 1–16.
156. Zhang Z, Chen P, Sapkota M, Yang L. Tandemnet: Distilling knowledge from medical images using diagnostic reports as optional semantic references. In: International conference on medical image computing and computer-assisted intervention. Springer; 2017. p. 320–8.
157. Altinkaya E, Polat K, Barakli B. Detection of Alzheimer's disease and dementia states based on deep learning from MRI images: a comprehensive review. J Inst Electron Comput. 2020;1(1):39–53.
158. Mathews SM. Explainable artificial intelligence applications in nlp, biomedical, and malware classification: a literature review. In: Intelligent computing-proceedings of the computing conference; Springer. 2019. p. 1269–92.
159. Madanu R, Abbod MF, Hsiao F-J, Chen W-T, Shieh J-S. Explainable AI (XAI) applied in machine learning for pain modeling: a review. Technologies. 2022;10(3):74.
160. Garvin MR, Prates ET, Pavicic M, Jones P, Amos BK, Geiger A, Shah MB, Streich J, Gazolla JGFM, Kainer D, et al. Potentially adaptive SARS-COV-2 mutations discovered with novel spatiotemporal and explainable AI models. Genome Biol. 2020;21(1):1–26.
161. Cliff A, Romero J, Kainer D, Walker A, Furches A, Jacobson D. A high-performance computing implementation of iterative random forest for the creation of predictive expression networks. Genes. 2019;10(12):996.
162. Shah RD, Meinshausen N. Random intersection trees. J Mach Learn Res. 2014;15(1):629–54.
163. Ikemura T, Wada K, Wada Y, Iwasaki Y, Abe T. Unsupervised explainable AI for simultaneous molecular evolutionary study of forty thousand sars-cov-2 genomes. bioRxiv; 2020.
164. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst. 2017;30:25.
165. Prajod P, Huber T, André E. Using explainable AI to identify differences between clinical and experimental pain detection models based on facial expressions. In: International conference on multimedia modeling. Springer; 2022. p. 311–22.

166. Dasgupta D, Akhtar Z, Sen S. Machine learning in cybersecurity: a comprehensive survey. J Defense Model Simul. 2022;19(1):57–106.
167. Ucci D, Aniello L, Baldoni R. Survey of machine learning techniques for malware analysis. Comput Secur. 2019;81:123–47.
168. Perarasi T, Vidhya S, Ramya P, et al. Malicious vehicles identifying and trust management algorithm for enhance the security in 5g-vanet. In: 2020 second international conference on inventive research in computing applications (ICIRCA); 2020. p. 269–75. IEEE.
169. Jaswal G, Kanhangad V, Ramachandra R. AI and deep learning in biometric security: trends, potential, and challenges. Boca Raton: CRC Press; 2021.
170. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell. 2019;1(5):206–15.
171. Zhang Z, Hamadi HA, Damiani E, Yeun CY, Taher F. Explainable artificial intelligence applications in cyber security: state-of-the-art in research. arXiv:2208.14937 (arXiv preprint); 2022.
172. Capuano N, Fenza G, Loia V, Stanzione C. Explainable artificial intelligence in cybersecurity: a survey. IEEE Access. 2022;10:93575–600.
173. Buczak AL, Guven E. A survey of data mining and machine learning methods for cyber security intrusion detection. IEEE Commun Surv Tutor. 2015;18(2):1153–76.
174. Chalapathy R, Chawla S. Deep learning for anomaly detection: a survey. arXiv:1901.03407 (arXiv preprint); 2019.
175. Carta S, Podda AS, Reforgiato Recupero D, Stanciu MM. Explainable AI for financial forecasting. In: International conference on machine learning, optimization, and data science; Springer. 2021. p. 51–69.
176. Chromik M, Eiband M, Buchner F., Krüger A, Butz A. I think i get your point, AI! the illusion of explanatory depth in explainable AI. In: 26th international conference on intelligent user interfaces; 2021. p. 307–17.
177. Bussmann N, Giudici P, Marinelli D, Papenbrock J. Explainable machine learning in credit risk management. Comput Econ. 2021;57(1):203–16.
178. Agarwal A, Bhatia A, Malhi A, Kaler P, Pannu HS, et al. Machine learning based explainable financial forecasting. In: 2022 4th international conference on computer communication and the internet (ICCCI); 2022. p. 34–8. IEEE.
179. Eliot DLB. The need for explainable AI (XAI) is especially crucial in the law. Available at SSRN 3975778; 2021.
180. Williamson B. Digital policy sociology: software and science in data-intensive precision education. Crit Stud Educ. 2019;20:1–17.
181. Luan H, Tsai C-C. A review of using machine learning approaches for precision education. Educ Technol Soc. 2021;24(1):250–66.
182. Akgun S, Greenhow C. Artificial intelligence in education: addressing ethical challenges in k-12 settings. AI Ethics. 2021;20:1–10.
183. Gardner J, Brooks C, Baker R. Evaluating the fairness of predictive student models through slicing analysis. In: Proceedings of the 9th international conference on learning analytics and knowledge. p. 225–234; 2019.
184. Atakishiyev S, Salameh M, Yao H, Goebel R. Explainable artificial intelligence for autonomous driving: a comprehensive overview and field guide for future research directions. arXiv:2112.11561 (arXiv preprint); 2021.
185. Ni J, Chen Y, Chen Y, Zhu J, Ali D, Cao W. A survey on theories and applications for self-driving cars based on deep learning methods. Appl Sci. 2020;10(8):2749.
186. Yousuf H, Zainal AY, Alshurideh M, Salloum SA. Artificial intelligence models in power system analysis. In: Artificial intelligence for sustainable development: theory, practice and future applications; Springer. 2021. p. 231–42.
187. Lorente MPS, Lopez EM, Florez LA, Espino AL, Martínez JAI, de Miguel AS. Explaining deep learning-based driver models. Appl Sci. 2021;11(8):3321.
188. Konstantakopoulos IC, Das HP, Barkan AR, He S, Veeravalli T, Liu H, Manasawala AB, Lin Y-W, Spanos CJ. Design, benchmarking and explainability analysis of a game-theoretic framework towards energy efficiency in smart infrastructure. arXiv:1910.07899 (arXiv preprint); 2019.
189. Pannu HS, Malhi A, et al. Deep learning-based explainable target classification for synthetic aperture radar images. In: 2020 13th international conference on human system interaction (HSI); 2020. p. 34–9. IEEE.
190. Nahata R, Omeiza D, Howard R, Kunze L. Assessing and explaining collision risk in dynamic environments for autonomous driving safety. In: 2021 IEEE international intelligent transportation systems conference (ITSC); 2021. p. 223–30. IEEE.