Explainable Artificial Intelligence: Importance, Use Domains, Stages, Output Shapes, and Challenges

NAEEM ULLAH, Department of Electrical Engineering and Information Technology, University of Naples Federico II, Napoli, Italy
JAVED ALI KHAN, Department of Computer Science, University of Hertfordshire, Hatfield, United Kingdom of Great Britain and Northern Ireland
IVANOE DE FALCO, Institute for High Performance Computing and Networking (ICAR), National Research Council (CNR), Napoli, Italy
GIOVANNA SANNINO, Institute for High Performance Computing and Networking (ICAR), National Research Council (CNR), Napoli, Italy

There is an urgent need in many application areas for eXplainable ArtificiaI Intelligence (XAI) approaches to boost people's confidence and trust in Artificial Intelligence methods. Current works concentrate on specific aspects of XAI and avoid a comprehensive perspective. This study undertakes a systematic survey of importance, approaches, methods, and application domains to address this gap and provide a comprehensive understanding of the XAI domain. Applying the Systematic Literature Review approach has resulted in finding and discussing 155 papers, allowing a wide discussion on the strengths, limitations, and challenges of XAI methods and future research directions.

CCS Concepts: • **Computing methodologies** → **Artificial intelligence**; *Machine learning*; • **General and reference** → **Surveys and overviews;**

Additional Key Words and Phrases: Explainable Artificial Intelligence, literature review

ACM Reference Format:
Naeem Ullah, Javed Ali Khan, Ivanoe De Falco, and Giovanna Sannino. 2024. Explainable Artificial Intelligence: Importance, Use Domains, Stages, Output Shapes, and Challenges. *ACM Comput. Surv.* 57, 4, Article 94 (December 2024), 36 pages. https://doi.org/10.1145/3705724

1 Introduction

Recently, **Artificial intelligence (AI)** has been successfully integrated into our lives by addressing various problems in different sectors. People use various AI frameworks in different domains

The authors acknowledge the support of the project FAIR - Future Artificial Intelligence (AI) Research (PE00000013), spoke 3 - Resilient AI (CUP B53C22003630006), and under the NRRP MUR program funded by the NextGenerationEU. PNRR-PNC-D34Health (B53C22006120001).

Authors' Contact Information: Naeem Ullah, Department of Electrical Engineering and Information Technology, University of Naples Federico II, Napoli, Campania, Italy; e-mail: naeem.ullah@unina.it; Javed Ali Khan, Department of Computer Science, University of Hertfordshire, Hatfield, United Kingdom of Great Britain and Northern Ireland; e-mail: j.a.khan@herts.ac.uk; Ivanoe De Falco, Institute for High Performance Computing and Networking (ICAR), National Research Council (CNR), Napoli, Italy; e-mail: ivanoe.defalco@icar.cnr.it; Giovanna Sannino, Institute for High Performance Computing and Networking (ICAR), National Research Council (CNR), Napoli, Italy; e-mail: giovanna.sannino@icar.cnr.it.

ACM 0360-0300/2024/12-ART94
https://doi.org/10.1145/3705724

and applications such as mobile phones [105], vehicles [132], healthcare [131, 142, 144], military [4], law enforcement [61], and insurance companies [66] for different functions such as to prevent accidents and improve safety, help doctors diagnose and detect diseases, retrieve evidence and streamline law enforcement procedures, risk assessment, and so on. Because of the best performance of the AI models, many vendors are currently looking to use AI in their operational processes as a competitor of abilities in a variety of fields [35]. However, such AI-based applications have always been questioned for their proposed result transparency. These applications are usually referred to as black-box systems.

Therefore, **Explainable Artificial Intelligence (XAI)** has emerged to address the need for transparency in AI frameworks and to lower barriers to the widespread application of AI in important sectors. XAI is an approach to develop open techniques that let consumers comprehend and trust the evolving AI systems while being able to govern them successfully [47]. The early attempts to comprehend the operation of expert systems and Bayesian networks may be found when one digs into the history of artificial intelligence [12]. However, the advent of **deep learning (DL)** has made XAI a vibrant and ever-evolving field of study, emphasizing the growing importance of understanding the complex mechanisms at the heart of AI systems. With AI becoming more and more ingrained in society, the importance of XAI cannot be overstated; it is essential to bridging the knowledge gap, guaranteeing accountability, and advancing moral and ethical AI usage.

Recently, many review and research articles addressing different XAI components in general or in particular have been published. This makes it possible to find different problems and possible solutions for future studies. Furthermore, these studies addressed and explored different issues in the domain of XAI. While the body of material already in existence has made a significant contribution to the comprehension of XAI, the proposed study aims to provide a comprehensive viewpoint that goes beyond the confines of certain fields. Our method stands out in this extensive study by providing a more in-depth analysis of XAI than just a list of ideas. To provide our readers with an understanding of the most current breakthroughs, we not only explore the significance, application fields, stages, output formats, and problems of XAI, but we also offer a distinctive synthesis of the most recent achievements. What distinguishes our survey is its focus on the interaction among explainability, application areas, and emerging issues. We hope that by offering a comprehensive overview of XAI, we will provide scholars, professionals, and enthusiasts with a unique viewpoint that goes beyond current surveys and adds to our understanding of the dynamic field of explainability in artificial intelligence. This survey seeks to serve as a resource for scholars looking to work on issues and possible avenues for XAI research. We seek to highlight commonalities, contrasts, and new developments in XAI research by including a range of application fields. With this synthesis, we seek to clarify the meaning of XAI and offer insightful guidance to academics, practitioners, and policymakers navigating the rapidly changing field of artificial intelligence.

Therefore, the key contributions of the current review are:

- *Comprehensive and Holistic Exploration*: Our review offers a thorough and all-encompassing investigation of XAI, addressing its significance, various application fields, phases of development, output formats, and related difficulties. Unlike other existing reviews that consider just one area or, at most, a few, this review considers a wide set of application areas. At the same time, the different output formats provided by the XAI methodologies are shown, differently from many existing reviews.
- *Cutting-edge Synthesis*: To give readers an understanding of the most recent approaches and strategies influencing the field, we provide a cutting-edge synthesis of the most recent developments in XAI.

- *Interdisciplinary Insight*: Going beyond traditional assessments, this study offers interdisciplinary insights into the various aspects of XAI by exploring the intersectionality of explainability, application areas, and emergent difficulties.

The article is organized as follows: Section 2 provides detailed information about the fundamental concepts and background of XAI, examining its necessity from various perspectives. Section 3 details the **systematic literature review (SLR)** methodology employed in this study, outlining the comprehensive approach used to identify and analyze relevant scholarly publications. Section 4 provides information about different stages of explainability, providing insights into approaches that contribute to improving the transparency of AI models. The exploration of various forms of explanations or output formats is presented in Section 5, highlighting the diversity in presentation methods and their implications for interoperability. Section 6 provides information about the diverse domains in which XAI can be used. Section 7 provides a critical discussion. Finally, Section 8 concludes the key findings of the proposed SLR.

2 Fundamental Concepts and Background

This section focuses on explainability and its types. Below, we discuss in detail:

2.1 Artificial Intelligence

AI makes computer systems capable of doing tasks, referred to as human intelligence. These skills include memorization, problem-solving, perceiving and thinking, understanding ordinary language, self-learning, and even recognizing emotions. Moreover, AI applications are designed to comprehend data by processing it, analyzing it, identifying hidden patterns, and then utilizing that knowledge to predict the future. It tries to replicate human cognitive abilities so robots can solve problems and complete challenging tasks. AI technologies are applied in various fields, including natural language processing, robotics, decision-making systems, image recognition, and so on, to create intelligent gadgets that can improve human productivity and decision-making processes.

2.2 Explainability

After the success of AI in various application domains, as a future development, we should see the emergence of self-sufficient systems capable of perceiving, learning, making decisions, and acting independently. The inability of the machines in these systems to justify their choices and behaviors to human users, however, limits their efficacy. In the literature, there are different definitions for explainability. Arrieta et al. [5] defined explainability as "an active property of a model, denoting any action or procedure taken by a model with the intent of clarifying or detailing its internal functions." Also, Gunning et al. [48] describe the explainability as "XAI will create a suite of machine learning techniques that enable human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners." In line with the previous definitions, we defined explainability as a characteristic of AI systems that gives humans understandable explanations for decisions made, outputs produced, and actions taken by the model in a transparent way. These XAI definitions enlist explainability as an active feature that the AI model utilizes to help users better understand the logic behind the decision-making and elaborate on the system's inner workings. This proactive aspect is essential in fields such as banking and healthcare, where awareness of AI decision-making procedures may greatly impact results and moral behavior. The eXplainable Artificial Intelligence (XAI) initiative intends to provide tools that let people interact with the burgeoning class of AI partners with confidence, understanding, and efficiency. New AI systems can justify their actions, identify their advantages and disadvantages, and express predictions about how they will behave. The approach to achieve that objective

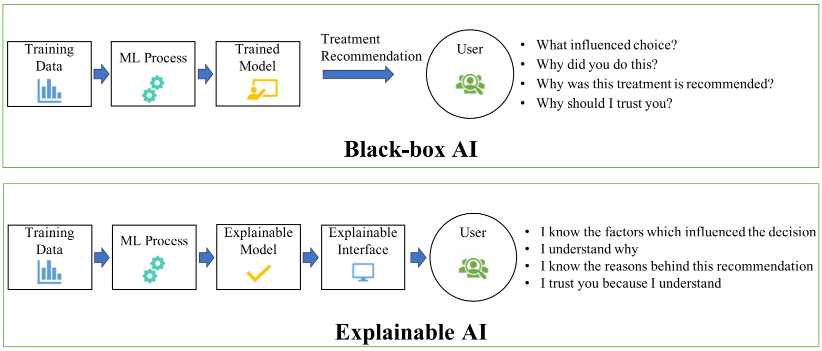

Fig. 1. General concept of XAI.

is to create new or improved AI methods that will result in models that are easier to understand. Modern human–computer interface approaches will be used in combination with these models to provide explanation dialogues for the user that are clear and helpful. Figure 1 compares ordinary and explainable AI systems. Think of a case study from the healthcare sector. XAI is essential to ensure that medical practitioners (users) can understand and trust the judgments made by AI systems in the healthcare context. For healthcare professionals to make educated judgments, guarantee the security of patients, and uphold faith in the AI-driven diagnostic process, transparency is crucial. Incorporating XAI enables efficient cooperation between AI tools and healthcare experts, improving patient outcomes. Medical practitioners may thoroughly know how the AI system arrived at a particular diagnosis.

Experts in vital domains, not only healthcare but also, for example, finance and the military, look for effective problem-solving techniques in addition to relevant outcomes they can comprehend and rely on. Engineers benefit from this need for applicability, since it encourages more in-depth analysis of system behavior, and domain experts may review and validate the AI's findings. XAI techniques enable assessing and enhancing current knowledge, expanding knowledge boundaries, and creating new ideas and hypotheses. Researchers also employ XAI to accomplish a range of objectives, including enhanced control, progress, justification, and discovery. Investigating these otherwise transparent AI systems has many advantages, including helping people deal with any unfavorable effects of automated decision-making, helping them make better decisions, spotting and fixing security flaws, bringing algorithms into line with moral standards, and raising industry standards for AI-powered manufacturing, all of which will boost trust with clients and companies. The benefits of offering information on these "black-box" technologies are listed below (Figure 2):

- *Transparency*: To make the AI decision-making processes clear and transparent.
- *Trust*: To increase confidence and trust in AI outputs.
- *Understanding*: To ensure people can easily comprehend AI decisions.
- *Accountability*: Making AI accountable for its deeds.
- *Ethics*: To ensure that AI models abide by moral principles.
- *Innovation*: To promote inventive, new, and fresh approaches.

Research on the evolving theories, methodology, and tools of XAI has been quite busy over the past few years, and XAI's standing as a research field has grown with time. In the next section, an overview of several reviews regarding XAI is reported, which demonstrates the importance and interest of XAI in the international research context.

Transparency
Trust
Understanding
Accountability
Ethics
Innovation
XAI Goals

Fig. 2. Six key objectives or goals of XAI.

2.3 Related Work

Recently, researchers have published several review studies (as shown in Table 1) that compile and discuss the most current advancements in the XAI domain.

The XAI methodologies and assessment criteria in the papers reported in Table 1 were extensively clustered, making them more reliable than previous literature reviews. Researchers from certain fields also examined the opportunities and difficulties from their viewpoints. Most of the publications in the literature are in the medical and healthcare fields. However, the literature contains a few review articles that contain information about different fields where XAI can be used. In the research, the authors studied and analyzed the theories and practices of XAI, the problems they may address, and the potential course of action. This was done in relation to or without consideration for the application domains and tasks. To our knowledge, no research has used XAI methodologies while considering various application areas, relevance, approaches, methods, and challenges. Additionally, a survey that would examine the many application areas, significance, methodologies, methods, and challenges across the entire research is still lacking. To gather and study the ways to add explainability to AI/ML models, an established guideline for SLR [71] was followed. Additionally, based on the chosen publications, this survey study generated a broad idea about the use of XAI in several application fields.

3 SLR Methodology

SLR approach [71] was used to analyze the literature for this review study. SLR is a well-known method used to quickly find, assess, and analyze the most important literature in a particular field of scientific study. The SLR's goal in this study is to identify the best XAI approaches, challenges, and the domains in which we can apply these approaches. The recommendations include procedures for locating and studying possible research works with the intention of considering potential directions for future research as well as instructions for properly reporting the SLR. Figure 3 briefly depicts the SLR phases. Three phases comprise the methodology of the SLR: planning the review, conducting the review, and reporting the SLR results. Before developing a data extraction strategy, the study's goals and objectives must first be established. Next, research questions must be developed followed by the development of a search strategy to target only literature that is relevant to the research questions. While the remaining phases are discussed here, the study's goal was already indicated.

3.1 Planning the SLR

Developing a thorough research strategy for the SLR is the first step. At this step, the requirement for conducting the SLR is determined, the research questions are outlined, and a thorough procedure for the research tasks is decided.

3.1.1 Identifying the Need for Performing SRL. Keeping with the discussion in Sections 1 (introduction) and 2 (related work), the background information gets more and more disorganized as there are more research works on XAI. However, very little secondary research has been carried

Table 1. Survey Papers Available through Google Scholar Published in 2022 and 2023

Work	Year	Focus Area	Review Type	Applications	# of Cited Papers
[30]	2022	XAI	**Systematic Literature Review (SLR)**	Cybersecurity, smart healthcare and transportation, economy and justice systems, and networking and communications	330
[82]	2022	Explainability in GNN	Experimental Survey	Real-world citation networks	62
[27]	2022	Explainability in GNN	Comprehensive survey	Healthcare and recommendation systems	283
[161]	2022	Explainability in GNN	Taxonomic Survey	Chemical molecules, financial data, and social networks	283
[156]	2022	Trustworthy graph learning	Comprehensive review	Finance and e-commerce	171
[134]	2022	Human-machine collaboration	Comprehensive review	Medical domain	171
[166]	2022	Deep learning	SLR	Medical domain	56
[97]	2022	Deep learning	SLR	Medical domain	167
[146]	2022	Deep learning	SLR	Medical domain	290
[24]	2022	Human-centered design	SLR	Medical image analysis	128
[89]	2022	SHAP-based Explanation Methods	SLR	Medical imaging AI	128
[49]	2022	Satellite imagery and deep machine learning	SLR	Human poverty domain	75
[20]	2022	Human-centered design	SLR	Medical informatics	77
[46]	2022	Deep learning for radiology	SLR	CT scan, radiography, ultrasound, sMRI, PET, and mammography	77
[125]	2022	Deep neural networks and global interpretation	Comprehensive review	Safety-critical applications and healthcare	112
[158]	2022	Multi-modal and multi-center data fusion	Mini-review	Medical and healthcare domains	112
[10]	2022	Distributed Ledger Technology and AI	SLR	Self-driving cars, chatbots, Bitcoin blockchain, SMART CONTRACTS, Non-fungible token, and Decentralized Finance and Autonomous Organization	140
[18]	2022	CyberSecurity	SLR	Malware and intrusion detection systems, spam and phishing detection, fraud, and BotNets detection, and digital forensics	244
[18]	2022	CyberSecurity	SLR	Malware and intrusion detection systems, spam and phishing detection, fraud, and BotNets detection, and digital forensics	244

(Continued)

Table 1. Continued

Work	Year	Focus Area	Review Type	Applications	# of Cited Papers
[58]	2022	**Convolutional Neural Network (CNN)**-based Brain-computer interface	SLR	Diagnosis of brain disorders, prevention of motion sickness, rehabilitation after stroke and robotic arm control	54
[122]	2023	XAI techniques	Comprehensive Survey	Medical systems, autonomous vehicles, and social networks.	158
[2]	2023	XAI	SLR	Healthcare, military, and banking	516
[33]	2023	XAI core ideas and techniques	Comprehensive Survey	Healthcare, livestock mart, and user privacy	63
[103]	2023	Evaluation of XAI	SLR	Not mentioned	307
[22]	2023	Trustworthy and Explainable AI components	SLR	Healthcare, banking, IoT, and autonomous system	103
[120]	2023	XAI methods	Comprehensive Review	Cybersecurity	108
[93]	2023	Historical perspective of XAI	SLR	Biomedical and COVID-19	64
[56]	2023	XAI concepts and challenges	SLR	Healthcare	84
[7]	2023	XAI	SLR	Fake news detection	49
[8]	2023	XAI models	SLR	Federated learning, 5G/6G networks	60
[72]	2023	XAI for IOT	SLR	Energy Management, Autonomous Systems and Robotics, Environmental Monitoring, Industrial Domain, Financial System, Healthcare, Security, and Privacy	118
[101]	2023	XAI opportunities and solutions	SLR	Cyber defenses, Internet of things	215
[14]	2023	Explainable AI in medical imaging	Comprehensive Review	Healthcare	76
[21]	2023	Explainable AI Techniques	Comprehensive Survey	Healthcare	119
[104]	2023	XAI techniques for biomedical imaging, DNN	Comprehensive Survey	Healthcare	291
[109]	2023	Trustworthy AI	Mini Survey	Medical Applications	61

The data mentioned in the table was updated on 10 October 2023 from Google Scholar bibliographic database.

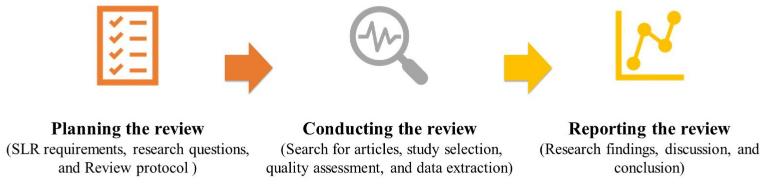
Planning the review
(SLR requirements, research questions, and Review protocol)
Conducting the review
(Search for articles, study selection, quality assessment, and data extraction)
Reporting the review
(Research findings, discussion, and conclusion)

Fig. 3. SLR methodology stages following the guidelines from Kitchenham and Charters [71].

out specifically to organize the vast knowledge of the XAI techniques. Additionally, no proof of an SLR was identified in the literature that broadly discusses the XAI and its implications in various fields. As a result, it is necessary to undertake an SLR to assemble, analyze, and offer a thorough and objective assessment of the significant publications on the XAI measurements and techniques.

3.1.2 ***Research Questions (RQs)***. The goal of this study is to examine the best alternatives that are now accessible in the field and assess the options that have been utilized in the literature to explain the explainability of AI and ML. To perform an SLR of the methods used to provide explainability for AI and **Machine Learning (ML)** systems and their evaluations in various application domains and tasks, a number of RQs were established. The main goal of the questions is to investigate the most popular techniques for developing understandable AI and ML models as well as their advantages and disadvantages. Included here are the significant application domains that employed XAI technology. We developed the following research questions with this in mind:

- RQ1. Is XAI essential to use with ML and DL approaches?
- RQ2. When will we generate the explanation for the model (i.e., stage of explainability)?
- RQ3. What are the different forms (output formats) of providing explanations?
- RQ4. In which application domains or tasks can the XAI methods be used?

3.1.3 Motivation of the Questions. To fully appreciate the importance of research in XAI, it is essential to know the rationales for the questions answered in this survey. Each inquiry explores crucial elements that not only influence the course of XAI research but also have significant ramifications for its real-world applications and broader social effect.

RQ1: The rising dependence on AI systems in several crucial fields, where the requirement for transparency, accountability, and ethical decision-making is vital, highlights the significance of XAI research. The adoption of AI systems in critical industries such as healthcare, banking, and security may be improved with a comprehensive knowledge of how these systems make their judgments.

RQ2: When to construct model explanations, or what stage of explainability to use, is a topic that motivates a more detailed study of the temporal dynamics of AI decision-making (RQ3). Understanding the explanation production time is crucial for optimizing the user experience and facilitating informed decision-making. Researchers may ensure that users get pertinent and actionable explanations by determining the most effective time to provide explanations either before or after the model's decision-making process. This information depends on gaining trust, influencing users' perceptions of AI systems, and facilitating the easy integration of AI into complex decision-making processes.

RQ3: Addressing question RQ3 requires understanding the variety of explanation forms used in AI systems. Suppose researchers possess a comprehensive understanding of the range of viable explanation formats. In that case, they may adapt how information is presented to suit different study groups' specific needs and preferences. This knowledge is crucial to ensuring that explanations are engaging and instructive in addition to being clear and accessible to a variety of stakeholders. Recognizing the variety of explanation strategies leads to developing a consistent and adaptable explanation strategy, which enhances the overall interpretability and user-friendliness of AI systems.

RQ4: For AI to reach its full potential, it is essential to understand the many areas and activities where XAI technologies might be used. Identifying these application areas not only shows how versatile XAI is, but it also allows the customization of thorough explanations to fit the unique requirements and challenges of many businesses, promoting moral and practical use of AI.

3.1.4 SLR Protocol. To address the above RQs and help achieve the review's goal, the SLR technique was created. Most of the specifics of how to conduct the SLR were included in the protocol. First, the section thoroughly describes the selection of research papers, the development of inclusion/exclusion criteria and quality rating questions, and the identification of potential bibliographic databases. The second stage involved a thorough scan of every object, during which relevant data were gathered and assembled into a feature matrix. Finally, a detailed report was created after all the authors contributed to the study of the obtained features to illustrate this SLR outcome.

Search Strategy. The published literature in scholarly conferences and journals was studied for this purpose. These digital libraries were searched throughout the search:

- Google Scholar
- Science Direct
- IEEE Explore
- Springer
- Elsevier.

Search Terms. To formulate the most effective search string, we combine keywords produced from research questions and goals with related and synonym terms for each topic. To link important terms, we utilized Boolean AND, and to add alternatives and synonyms, we used Boolean OR. The terms used are "explainable AI" or "XAI" or "model explanation" or "model explanation" and "explainable machine learning." Conference papers and journal articles were searched using the previously specified search terms in electronic databases. Since the search engines of various databases use various search string syntaxes, the search terms were altered to fit many databases.

Inclusion Criteria.

- A study will be included if it satisfies the search criteria listed in the aforementioned section.
- Only the journal version is included for studies with a conference edition and a journal variant.
- Only the most recent and comprehensive study published more than once will be considered.
- English-language studies will be included.
- The manuscripts published from 2017 to 2023 will be considered for the proposed SLR.
- In this review study, only survey and review papers from 2022 and 2023 will be used.
- Articles focusing on XAI's significance, strategies, techniques, application areas, and problems.

Exclusion Criteria.

- Studies that do not specifically address the explainability of AI or ML will not be considered.
- Duplicate studies will not be considered.
- Studies that are not in English will not be considered.
- Literature that is unrelated to XAI directly.

Quality Assessment Criteria. After evaluating the results using the quality evaluation criterion, the studies that met the final list were selected. The substance of the completed papers is the main focus of quality evaluation. The quality checklist includes the following questions:

- Are discussions of the explainable context clear?
- Is the research's purpose clearly stated?
- Are the research findings relevant to the goal of the study?
- Have all queries about the study been answered?
- How are the results compared against prior papers?
- Do the researchers specify how the data are presented (such as images, statistics, etc.)?

3.1.5 Conducting the Review. SLR demands a careful analysis of all relevant sources. Finding pertinent research papers included many procedures, including identification, screening, eligibility checks, and sorting of the selected articles. Figure 3 shows a detailed flow chart of this identification process. With pertinent keywords, a systematic search was carried out in five well-known academic databases. The snowballing method was used in addition to database searches to find pertinent articles. This strategy examined the important publications found through the first database search for references and citations. Using the snowballing approach, we aimed to gather any relevant research that could not have been found using standard database queries. The search was limited to research publications between 2017 and 2023 (May) to ensure the inclusion of recent and relevant literature in the field of XAI. The review's chosen timeline, which extends from 2017 to 2023, corresponds with the amount of research material found in the examined databases. The chosen era was justified, since, following a preliminary search for previous publications, no relevant research that fits the given search parameters was discovered. After initial screening, the inclusion criteria are applied to evaluate the relevance of the research articles identified using the search criteria. The paper's titles and abstracts were skimmed to align with the screening process to ensure they were aligned with the research objectives. After that, the whole texts of a few selected publications were examined for further assessment. A thorough quality evaluation was conducted on a few pieces of research to determine their reliability and methodological soundness. The robustness of the research technique, the accuracy of the conclusions, and the dependability of the arguments made in the chosen literature were among the quality assessment criteria.

The initiation of the process took place in March 2023. A preliminary search was conducted using Google Scholar, using keyword phrases to ascertain the origins of the research papers. Google Scholar has been used to extract 106 publications. However, most of the papers were obtained from Springer Link and IEEE Xplore, all accessed in June 2023, according to the search results. There were other comparable sources, but we did not consider them, as they primarily contained data from the sources already mentioned. To narrow the search specifically to the AI sector, databases were selected as the main sources of research articles for this evaluation. By searching for the terms explainable AI, XAI, and explainable machine learning, we retrieved the first 280 articles from IEEE Explore and Science Direct. After searching the three electronic databases separately, we gathered 386 papers. Table 2 shows their distribution over time. These studies were reviewed, and 23 duplicate papers were removed. Then, the studies were reviewed based on "Title and Abstract," which returned 213 relevant studies. In the next phase, the selected studies were further reviewed based on the introduction and other details, including the Conclusion (i.e., text review), which shrank the candidate papers to 148. These studies are finally selected after applying quality assessment criteria. Then, we added the 7 most recent studies from Google Scholar published in 2023. Hence, the total number of final papers reached 155. These results are graphically shown in Figure 4.

The proposed SLR is conducted methodically and structurally; the selection process is designed to incorporate the most noteworthy and pertinent studies within the scope of XAI. Following

Table 2. The Distribution over Time of the Published XAI Literature Articles

Year	2017	2018	2019	2020	2021	2022	2023
No Papers	14	31	48	56	69	77	91

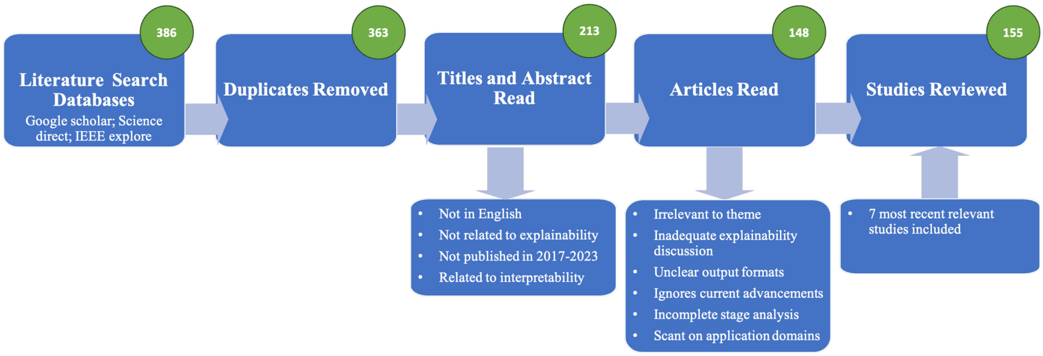
386
Literature Search Databases
Google scholar; Science direct; IEEE explore
363
Duplicates Removed
213
Titles and Abstract Read
148
Articles Read
155
Studies Reviewed
Not in English
Not related to explainability
Not published in 2017-2023
Related to interpretability
Irrelevant to theme
Inadequate explainability discussion
Unclear output formats
Ignores current advancements
Incomplete stage analysis
Scant on application domains
7 most recent relevant studies included

Fig. 4. Summary of review results.

the inclusion and exclusion criteria of the letter initially resulted in a quantity of 386 papers after removing 23 duplicates and further refining based on relevancy and a thorough quality assessment. Later, 155 key papers that met the inclusion requirements and covered the years 2017 to 2023 were selected. Notably, according to our knowledge, no study that met the inclusion criteria and made a significant effect was intentionally omitted from the SLR. This rigorous selection process ensures that the proposed assessment reflects the current state-of-the-art, addressing the most significant and impactful findings in the field of XAI and providing a strong foundation for further advancements in the field. The purpose of the exclusion of publications before 2017 was to link the research aim of reviewing recent accomplishments with the notion and growing research interest in XAI, which predominantly evolved around that time. Studies that did not specify the inclusion criteria were excluded from the SLR analysis to better represent and answer the research questions with more recent and relevant research articles. Furthermore, we have reviewed the most current articles till submission to ensure no noteworthy recent research was missed meeting the inclusion criteria.

4 Stage of Explainability

Models are created as part of the ML and AI processes to assess data and make predictions or categorical determinations. In this sense, explainability refers to a model's capacity to offer a concise and intelligible justification for its judgments or decisions. This justification can be produced during and after the model has been trained and used on fresh data. According to Vilone and Longo [149, 150], these two phases are called ante hoc and post hoc. Both "ante hoc" and "post hoc," taken as a whole, can be described as strategies for gaining explainability in AI. These phrases refer to the moment or timing of the efforts undertaken to achieve explainability.

4.1 Ante Hoc

In this stage, the decision-justification procedure from the first model training is integrated. "Ante hoc" refers to actions taken before the AI model's implementation that are intended to make certain the model is transparent and understandable from the start. Ante hoc methods are commonly used

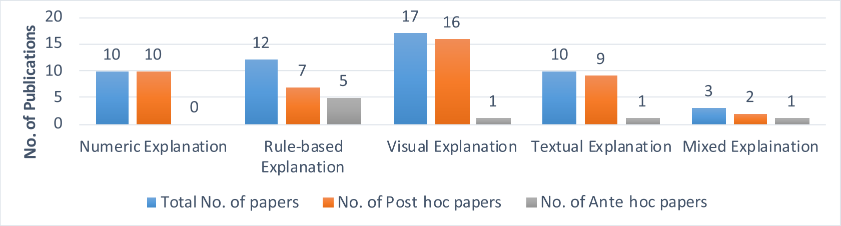

Fig. 5. Details of latest papers published within 2017 and 2023 about various output formats.

for transparent models such as fuzzy and decision trees. The aim is to make the model intrinsically interpretable, meaning that explanations will be given in addition to predictions or classifications.

4.2 Post Hoc

A surrogate or external model is combined with the main model in post hoc approaches. It elaborates on how a trained AI approach interprets decisions made after it has already made a prediction. To make the decision process understandable for the possible users, it sheds light on the complex internal structure of the model by reproducing the model's decision-making behavior. Such approaches are applied to machine learning approaches such as neural networks and support vector machines whose internal decision-making structures for classification problems are complex and difficult to grasp. Moreover, post hoc is divided into model-agnostic and model-specific.

4.2.1 Model-specific. These techniques have been specific to certain ML models to add explainability for decision-making by providing insights into the model's internal working mechanism. This type of XAI approach sheds light on the model decision-making logic and feature importance, as they are connected to the model's structure and individual features. The ML classifiers using Model-specific XAI techniques include gradient boosting and decision tree for feature importance and decision tree analysis, respectively.

4.2.2 Model-agnostic. The Model-agnostic XAI techniques are not specific to any ML classifier. Regardless of their internal complexity and structure, these approaches can be intertwined with any ML model. The main driver of model-agnostic techniques is to shed light on the model's inputs or features that greatly impact its decision-making ability. LIME, **Shapley Values (SHAP)** [76] and permutation feature significance [85] are a few popular XAI approaches independent of the model.

5 Different Forms of Explanation or Output Formats

This section elaborates on the different explanations included in various AI/ML models identified through the proposed SLR. The chosen articles revealed the choices that the models produced. The various explanation types identified are numerical, logical, textual, and visual. The authors of several articles combined these forms to make the explanation more understandable and approachable (mixed approach). Figure 5 provides the number of publications within 2017 and 2023 for each output format covered in this review. The descriptions of each explanation type are followed by related work identified using the proposed SLR to highlight its importance and implications.

5.1 Numeric Explanations

In XAI, numerical values and statistical metrics explain how AI models make decisions. These explanations seek clarity and transparency by mathematically portraying the elements impacting the

Table 3. Recent Works in XAI that Use the Numeric Form of Explanation

Ref.	**Method for explainability**	**Stage of explainablity**	**Year**
Ten et al. [163]	**Failure Diagnosis Explainability (FDE)**	Post hoc	2018
Lundberg et al. [87]	VisExp	Post hoc	2022
Carletti et al. [19]	Depth-based Isolation Forest Feature Importance (DIFFI)	Post hoc	2019
Ponn et al. [112]	SHAP	Post hoc	2020
Serradilla et al. [130]	LIME	Post hoc	2020
Amit et al. [3]	LIME	Post hoc	2022
Salih et al. [126]	LIME, SHAP	Post hoc	2023
Nikith et al. [106]	LIME	Post hoc	2022
Ullah et al. [143]	LIME	Post hoc	2023
Khedkar et al. [68]	LIME	Post hoc	2019

model's output or choice. Numerous metrics, including the confidence measures of features [100], saliency, causal importance [39], feature importance [116], and mutual importance [59], serve to quantify the contribution. Using numerical data, XAI improves the interpretability of complicated AI models, allowing stakeholders to learn more about the important factors or parameters influencing the model's behavior. Sarathy et al. [127] computed and compared the cases' quadratic means to offer the decision with justifications. The **depth-based isolation forest feature importance (DIFFI)** was used by Carletti et al. to support the decisions made from depth-based **isolation forests (IFs)** in anomaly detection for industrial applications [19], and the FDE measure was developed to add exact explainability for inability diagnosis in automated industries [163]. Furthermore, certain model-neutral tools, such as SHAPley Values (SHAP) [112], LORE [152], and **Locally Interpretable Model-agnostic Explanations (LIME)** [130], generate numerical explanations. Table 3 shows numerical techniques according to stage and explanation scope, providing further instances of numerical explanations. However, the numerical explanations need a high level of expertise in the pertinent domains, because they are related to the features. Table 3 lists studies that employ numerical explainability measures.

5.2 Rule-based Explanations

Rule-based explanations entail the application of predetermined rules or logical criteria to describe how AI models make decisions. These explanations depend on clearly stated guidelines that clarify the standards models utilized to make a particular prediction or judgment. Rule-based explanations may be expressed as logical statements, trees, lists, or if-then clauses, and they offer a simple and understandable foundation for comprehending how the model handles incoming data and produces outputs. Most models that create rule-based explanations generate justifications that apply to the entire model or global explanations. De et al. [28] proposed using the existing TREPAN decision tree as a surrogate model with a **feed-forward neural network (FFNN)** to generate rules that describe the information flow within the neural network. Similar to this, Confalonieri et al. [26] changed TREPAN, an algorithm that explains **Artificial Neural Networks (ANN)** using decision trees, to become TREPAN Reloaded by using ontologies that define domain knowledge in the process of providing explanations. Rutkowski et al. [121] presented a novel approach to developing explainable recommender systems. It was based on the Wang-Mendel approach for constructing fuzzy rules. A method for learning and decreasing fuzzy recommenders is recommended in addition to feature encoding. An explainable intelligence model was proposed by Keneni et al. [65] to

Table 4. Recent Works in XAI that Use the Rule-based Form of Explanation

Ref.	Method for explainability	Stage of explainablity	Year
De et al. [28]	TREPAN decision tree	Post hoc	2020
Li et al. [83]	SIRUS	Post hoc	2023
Nimmy et al. [107]	Belief-Rule-Based	Ante hoc	2023
Nimmy et al. [107]	Belief-Rule-Based	Ante hoc	2023
Macha et al. [90]	RuleXAI	Post hoc	2022
Zhang et al. [165]	CF-MABLAR	Post hoc	2022
Confalonieri et al. [26]	TREPAN decision tree	Post hoc	2019
Rutkowski et al. [121]	Wang-Mendal	Post hoc	2019
Keneni et al. [65]	Mamdani fuzzy model	Post hoc	2019
Soares et al. [135]	ALMMo-0*	Ante hoc	2020
Hatwell et al. [51]	Ada-WHIPS	Ante hoc	2020
Ferreyra et al. [41]	BB-BC IT2FLS	Ante hoc	2019

explain the decisions an **unmanned aerial vehicle (UAV)** makes when it is on a preset mission and decides to deviate from it. The explainable model of the Sugeno-type fuzzy inference model is provided as if-then rules on a visual platform. In addition to providing rules of engagement when the UAV encounters hostile positions and adverse weather, the Mamdani fuzzy model is utilized to instruct the UAV along its chosen mission. The novel neuro-fuzzy system ALMMo-0* was proposed by Soares et al. [135]. Two further model-specific approaches have been proposed to produce rule-based explanations: eUD3.5, an explainable form of UD3.5, and Ada-WHIPS, which supports the AdaBoost ensemble approach [51]. Post hoc explanation techniques, however, describe the results after they have been produced rather than in a step-by-step glass-box fashion to explain how output is attained. By creating a **Belief-Rule-Based (BRB)** framework that explains in a glass-box fashion why a given decision has been made, the suggested solution in Reference [107] tackles these shortcomings. Table 4 lists studies that employ rule-based explainability measures.

5.3 Visual Explanations

In XAI, the term "visual explanations" refers to using graphical or visual representations to explain the decision-making process of complex AI models. To make it simpler for users to understand the elements impacting the model's outputs, these explanations try to explain complicated data and model behavior clearly and interpretably. Heatmaps, line graphs, bar charts, decision trees, scatter plots, and other graphical representations are frequently used as visual explanations to highlight the roles played by various elements or variables in the model's predictions. The most common kind of justification was found to be visualizations. Most of the time, post hoc methods were employed to offer visual explanations in both global and local scopes throughout the stage of adding explanations, and the study investigations were carried out as studies from the healthcare domain that were not domain-specific. **Class activation maps (CAM)** [6, 138] and attention maps are popular visualization methods. The Uncertain-CAM method was proposed in Reference [1] to enhance the COVID-19 classification system's accuracy and deep learning explainability. Grad-CAM was proposed by Selvaraju et al. [129] when CAM was further expanded to include gradient weights. Based on X-rays, Brunese et al. employed Grad-CAM to identify COVID-19 infection [16]. To explain how to identify banknote fraud, Han and Kim [50] chose the pGrad-CAM form.

Table 5. Recent Works in XAI that Use the Visual Form of Explanation

Ref.	Method for explainability	Stage of explainablity	Year
Stodt et al. [137]	Grad-CAM and Eigen-CAM	Post hoc	2023
Yuan et al. [162]	LIME	Post hoc	2023
Kawakura et al. [64]	Grad-CAM++, ScoreCAM, and Grad-CAM	Post hoc	2023
Muddamsetty et al. [102]	Similarity Difference and Uniqueness (SIDU)	Post hoc	2022
Muddamsetty et al. [102]	Similarity Difference and Uniqueness (SIDU)	Post hoc	2022
Wang et al. [153]	MetaMatrix	Post hoc	2022
Vielhaben et al. [148]	DFT-LRP	Post hoc	2023
Sun et al. [138]	CAM	Post hoc	2020
Assaf et al. [6]	CAM	Post hoc	2019
Riquelme et al. [117]	Attention maps	Post hoc	2020
Aldhahi et al. [1]	Uncertain-CAM	Post hoc	2023
Brunese et al. [16]	Grad-CAM	Post hoc	2020
Han et al. [50]	pGrad-CAM	Post hoc	2019
Li et al. [84]	MLG-CAM	Post hoc	2023
Graziani et al. [45]	Concept attribution	Post hoc	2020
Shalaeva et al. [133]	MTDTs	Ante hoc	2018
Jung et al. [60]	SLRP	Post hoc	2021

A plausible path to understandable AI is provided by the **Gradient-weighted Class Activation Map (Grad-CAM)** approach, which uses gradient information to explain DNN. However, when network layers deepen for vibration signals utilized in machine fault diagnosis, Grad-CAM's feature resolution declines, lowering network explainability. Li et al. [84] propose a novel **Multilayer Grad-CAM (MLG-CAM)** as a valuable tool to show what networks have been discovered to handle this problem. Heatmaps of significant pixels were used by Graziani et al. [45] in addition to the concept-based explanation. They created a framework for idea attribution for **Deep Learning (DL)** to measure the contribution of intriguing characteristics to the deep network's decision-making. Furthermore, several explanation techniques have been proposed using attribution-based visualizations, including **Multi-Operator Temporal Decision Trees (MTDTs)** [133] and Selective **layer-wise relevance propagation (LRP)** (SLRP) [60]. The authors of Reference [148] created the virtual inspection layer by converting the time series into a comprehensible representation and enabled the transfer of relevance attributions to this representation using local XAI techniques like LRP. They broadened the number of domains in which a family of XAI techniques may be deployed by doing this, including speech, where the input may only be comprehended after being modified. They focused on the Fourier transformation, often used in interpreting time series, and referred to the method as DFT-LRP. Table 5 provides additional methods for supplementing the procedures for including visual justifications in different AI/ML model types.

5.4 Textual Explanations

In XAI, textual explanations explain how AI models make decisions using natural language or textual descriptions. With the help of these explanations, users will be able to appreciate the variables impacting the model's outputs without needing specific technical expertise. These explanations try to communicate complicated model behaviors and predictions in a human-understandable way.

Table 6. Recent Works in XAI that Use the Textual Form of Explanation

Ref.	Method for explainability	Stage of explainablity	Year
Stodt et al. [137]	Grad-CAM and Eigen-CAM	Post hoc	2023
Fernández et al. [40]	Counterfactual sets	Post hoc	2020
Zhong et al. [167]	Template-based natural language generation	Ante hoc	2019
Law et al. [78]	Counterfactual explanations	Post hoc	2023
Xia et al. [157]	Co-attention-based fine-grained counterfactual explanation	Post hoc	2023
Zhan et al. [164]	ExBERT	Post hoc	2023
Weber et al. [155]	TCBR	Post hoc	2018
Van et al. [147]	ICM	Post hoc	2020
Le et al. [79]	GRACE	Post hoc	2020
Benlabiod et al. [15]	CNN-CBR	Post hoc	2023

Textual justifications frequently offer thorough insights into the thinking behind a given choice, highlighting the essential characteristics or factors that influence the model's predictions. Textual explanations are the least often utilized, since they need more advanced computation and **Natural Language Processing (NLP)**. The majority of textual reasons are developed locally or for a specific judgment. Significant experiments have generated textual arguments using counterfactual sets [40], template-based natural language generation [167], and so on. The authors of [78] presented a novel sort of plausible counterfactual explanation to explain the behavior of computer vision systems utilized in urban analytics that make predictions based on characteristics throughout the full picture, as opposed to particular areas. To demonstrate the advantages of their approach, they talked about computer vision algorithms that are increasingly being employed in Geo AI and urban analytics to analyze street photography. Urban analytics need such arguments, since practitioners and academics increasingly rely on them to make decisions. In Reference [157], the authors proposed an aspect representation learning and co-attention-based fine-grained counterfactual explanation model that directly incorporates user preferences towards various objects for recommendation and explanation. The authors of Reference [164] proposed ExBERT, a fresh Explainable recommender system with a BERT-guided explanation generator, to deliver reliable explanations with more specificity. More specifically, adopting a multi-head self-attention-based encoder allows for inserting fake user and object profiles into a semantic representation. They further provided a special matched explanation prediction job with discriminative abilities to enable modification of the output sentence. Weber et al. [155] proposed textual CBR (TCBR) employing patterns of input-to-output links to recommend citations for academics using linguistic justifications. In contrast to TCBR, Waa et al. combined CBR with interpretable confidence metrics (ICM) to provide textual explanations [147]. GRACE, which may provide logical textual reasons for judgments, was developed by Le et al. [79]. The authors of Reference [15] used DL and CBR to boost the accuracy of breast cancer detection using mammography pictures. While DL provides exact mammogram segmentation, CBR provides a clear and accurate classification. Both simulated and real experiments showed that humans more easily understood textual explanations created using GRACE. Additionally, it is observed that textual justifications are created locally and tied to academic research, judicial systems, and so on. Additional proposed methods for creating textual reasons are shown in Table 6.

5.5 Mixed Explanations

In XAI, the term "mixed explanations" refers to the combining of several explanation approaches, such as numerical, visual, and textual methods, to offer a thorough and all-encompassing

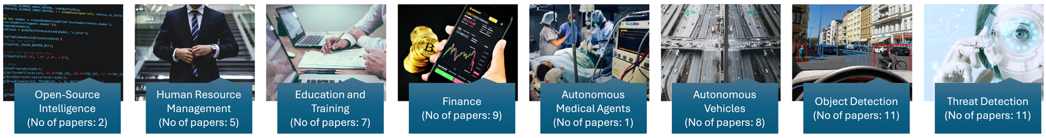
Open-Source Intelligence (No of papers: 2)
Human Resource Management (No of papers: 5)
Education and Training (No of papers: 7)
Finance (No of papers: 9)
Autonomous Medical Agents (No of papers: 1)
Autonomous Vehicles (No of papers: 8)
Object Detection (No of papers: 11)
Threat Detection (No of papers: 11)

Fig. 6. Eight key XAI application domains with the number of published articles. Pictures taken from https://unsplash.com/ with free licences.

comprehension of AI model decisions. Mixed explanations try to use the advantages of each technique by including a variety of explanation styles, providing a more complex and comprehensive view of the variables impacting the outputs of AI models. When stakeholders with various degrees of technical skill must work together to make choices based on the outputs of AI models, using mixed explanations is very beneficial. Several explainability methods produce graphical and numerical explanations that are simpler for average people to grasp information. Two user interfaces that display a combination of visual, textual, and numerical explanations are ExplAIner [136] and Rivelo [140]. Model understanding, diagnosis, and refinement are the three steps of the iterative workflow that lead to the ML models displayed in ExplAIner. ExplAIner creates an interactive graph representation of the model to be described by starting with TensorBoard, a visualization tool for machine learning created by Google. The edges of the graph reflect the connections between the components. In contrast, the graph nodes represent the model's constituent parts, such as inputs, parameters, and outputs, along with textual definitions.

6 Application Domains

Explainable AI has developed several methods that have been applied in various domains with a range of expected outcomes [2]. Although ML systems are presently fully functional in many fields (Figure 6), implementing them still presents considerable difficulties. These can occasionally result from an inability to explain how these methods work. In other words, their effectiveness is acknowledged, but the precise results are yet unknown. Because many of these applications are either from sectors where safety is critical or where sensitive personal information is present, a lot of emphasis is on illuminating how trained models generate conclusions, which are frequently predictions or classifications [86].

6.1 Explainable Threat Detection

The detection of hazards and efficient triage have been fundamental challenges in IT security for at least three decades. This topic subsequently proceeded to automated code analysis, recompilation, and assistance for evaluating network monitoring data from research on code analysis and signature-based antivirus software. Despite several diverse methods, fully automated threat detection and triage are presently not feasible in real-world systems due to the complexity of the task and the problem with false positives. These also include methods that reduce the amount of data that have to be manually evaluated by filtering out all known valid network traffic rather than focusing on identifying real threats [110]. However, these methods' opaqueness presents a significant challenge, since it makes it challenging to fully understand how an inference was generated and how they operate inside. Explainability has the potential to greatly enhance detection abilities, especially because dynamic consequences, such as changing user behavior, may be modeled and incorporated into algorithms earlier without causing a considerable number of false positives. XAI approaches are used in the process known as explainable threat detection to locate, analyze, and present clear justifications for potential dangers or hazards in complex systems. By providing

Table 7. Recent Work on Thread Detection Using Different XAI Approaches

Ref.	Thread Type	Method	Dataset	Year
Rahman et al. [114]	Intrusion detection	Tree-based classifiers and XAI using LIME and SHAP	ToN-IoTDataset	2023
Caforio et al. [17]	Network intrusion detection	CNN and Grad-CAM	KDDCUP99, NSL-KDD, UNSWNB15	2021
Malik et al. [92]	Cyber-threat detection	CNN and SHAP	MalDroid20	2022
Kao et al. [62]	Adversarial attacks	Grad-CAM	Mnist, Cifar10, Mini-Imagenet	2022
Koutsoubis et al. [74]	Drone and Aerial Threat Detection	ML and Grad-CAM	DyViR	2022
Khan et al. [67]	Threat detection Internet of Medical Things networks	Bidirectional SRU-driven DL model and LIME	ToN_IoT	2022

insights into the decision-making procedures of threat detection models, this technique makes it possible to comprehend the fundamental elements that contribute to identifying threats. By incorporating XAI into threat detection systems, it is possible to improve the interpretability and accountability of these systems, enabling users to understand the justification behind the identification of different threats, such as cybersecurity breaches, fraudulent activities, or anomalous behaviors.

6.1.1 Protection against Adversarial ML. By introducing carefully crafted data during the learning process, adversarial ML attackers try to distort the results of learning algorithms [55], leading a model to learn the incorrect things. Finding such a modification is challenging, especially when dealing with large amounts of data and having no prior model to base your search on. There are several approaches to solving this issue [38], but some of them, like Reference [98], employ neural sub-networks to discern between bad and excellent input data. In this particular scenario, explainability would considerably increase confidence in ML results, as it would speed up the process of identifying such manipulation without actually locating the changed data [53].

6.1.2 Importance of XAI in Thread Detection. Explainable Threat Detection promotes a safer and more resilient operating environment by facilitating not only a better knowledge of possible risks but also the creation of more dependable and efficient ways for mitigating and managing these threats. XAI can assist in spotting and protecting against adversarial attacks, in which harmful inputs are purposefully created to trick or manipulate AI systems, producing inaccurate outputs or judgments. Security professionals can comprehend and trust the outputs produced by threat detection systems thanks to XAI, which offers clear and understandable insights into the decision-making process. Additionally, XAI promotes public trust in security systems by providing clear insights into identifying possible risks. As a result, the general public is more certain of the accuracy and impartiality of threat detection processes. Some recent works on thread detection using different XAI approaches are given in Table 7.

6.2 Explainable Object Detection

Many other ANN designs, like YOLO [13], that have been trained on many labeled data are frequently used for object detection. It could be difficult, if not impossible, to grasp object detections in some circumstances due to the extreme complexity of the hyperparameters (number of layers, filters, regularizers, optimizers, and loss functions). Only qualities that are available in the data and have been represented as saliency maps [23] or, at their most, examples [81] or prototypes [70] can be utilized to explain an object identification job. The data frames that feed the ANNs have a limit on what can be explained, although they are cutting-edge techniques. Applications for industrial

Table 8. Recent Work on Object Detection Using Different XAI Approaches

Ref.	Ojective	XAI approach	Dataset	Year
Mankodiya et al. [94]	DL model and XAI approach is used to detect and segment the road that the Autonomous Vehicles will follow	Grad-CAM	KITTI road	2023
Sahatova et al. [123]	A comparison of XAI approaches for object detection in Computer Tomography	Grad-CAM, LIME, occlusion, and EigenCAM	KITTI road	2022
Honig et al. [54]	An object detection method was proposed for detecting handwheels of gate-valves	Layerwise relevance propagation and spectral relevance analysis	Novel handwheel	2023
Farabi et al. [36]	Used ResNet50V2, MobileNetV2, and VGG19 for breast cancer detection using histopathology images	Grad-CAM	BreakHis	2023
Noori et al. [108]	A deep learning and XAI-based approach for myopia detection	LIME	Real myopia	2022
Dworak et al. [34]	Visualization of CNN findings using an XAI camera image processing technique applied to LiDAR object detection	Grad-CAM	KITTI	2022

object detection, such as identifying railway obstacles, require human-like reasoning to ensure the system can be validated and even authorized [111]. Table 8 lists the recent developments in object detection domain using various XAI techniques.

6.2.1 Importance of XAI in Object Detection. Using XAI techniques for object detection in computer vision applications is known as Explainable Object Detection. In it, we explain the detection process of the complex and advanced techniques used to detect different objects in image or video streams. The integration of XAI with object detection system is more important from the user's perspective, because it will allow users to get a clear idea of the variables affecting the decision-making process of object detection algorithms, hence, resulting in a simpler, accurate, accountable, and reliable object detection system with fewer errors and biases. Furthermore, this integration leads to a more transparent and readable computer vision system, promoting the development of efficient applications in various domains, including security, image analysis, and autonomous driving.

6.3 Trustworthy or Autonomous Medical Agents (AMA)

AMA are powerful and independent systems that do many medical tasks such as patient monitoring, diagnosis, and treatment planning. These automatic systems make decisions (perform medical operations) using data and advanced algorithms without human involvement. By providing prompt and precise medical help, they can improve efficiency, expedite healthcare procedures, and improve patient care. In the past, to combine ML and medical decision-making, many unique architectural solutions have been recommended. These are predicated on the "doctor-in-the-loop," wherein healthcare providers offer ML algorithms feedback. These may result in a course of treatment, which the physicians themselves could subsequently assess. After that, they may continuously comment on improving the modeling [69]. The mechanism can also incorporate external knowledge to support decision-making that attempts to account for the latest developments in the underlying medical profession. The inclusion of XAI to AMA makes the decision-making process transparent, enabling users to know about the logic and reasoning behind different clinical decisions made by AMA. Explainable AMA will contribute to more effective and accurate healthcare delivery systems.

6.3.1 Importance of AMA. The transparent decision-making process of the AMA assists the users (medical professionals and patients) in understanding the underlying behavior of

Table 9. Recent Work on Autonomous Vehicles Using Different XAI Approaches

Ref.	Objective	Method	Dataset	Year
Mankodia et al. [95]	Use the XAI system in VANETs to give the system intelligence and stop any harmful information from spreading farther over the network, reducing the likelihood of chaos	Random forest and a heat map for XAI (no specific XAI method is given)	Vehicular reference misbehavior	2021
Rjoub et al. [119]	Using of edge computing, ML, and AI to aid newly developed autonomous cars in making trajectory and motion planning decisions	LIME with One-shot Federated Learning	Driving	2023
Saravanarajan et al. [128]	Improving semantic segmentation under hazy weather for autonomous vehicles using XAI and adaptive dehazing approach	Adaptive dehazing, DeepLabV3+ ResNet-101, Grad-CAM	Cambridge-driving Labeled Video	2023
Madhav et al. [91]	Introduce an explicable navigational intelligence that tries to coordinate the autonomous vehicle decision-making procedures	InceptionV3, LIME, and Grad-CAM	NSL-KDD	2022
Dong et al. [31]	Improve trustworthiness in autonomous driving systems through the development of XAI methods	Multi-modal DL architecture and Attention visualization	BDD Object Induced Actions	2023
Kolekar et al. [73]	An approach for semantic segmentation of images taken from unpredicted and unstructured traffic roads	Inception-based U-Net model with Grad-CAM	Indian Driving Lite	2022
Rjoub et al. [118]	An intelligent method for autonomous vehicles to help in trajectory and motion-planning decisions	Federated reinforcement learning with SHAP	Real life (level-5.global)	2022

pharmaceutical and medical devices. XAI helps to find mistakes in AMAs' decision-making process, guarantees that AMAs follow ethical and legal standards, fosters responsibility, and builds trust in their abilities by providing lucid justifications for medical assessments. Furthermore, to support patient-centric care, XAI provides straightforward justifications for health choices, enabling patients to engage in their healthcare actively and supporting them in making knowledgeable treatment decisions.

6.4 Autonomous Vehicles

Autonomous vehicles (self-driving cars) include cutting-edge technology, allowing them to travel without the driver. These cars utilize cameras, sensors, computers, and GPS to perceive their surroundings, identify obstacles, and make decisions in real-time efficient and secure transportation. Explainability is advantageous for the autonomous vehicle industry, particularly regarding technological advancements. Explanations can be used to determine the reasons behind the actions of an autonomous vehicle in car-accident situations. This might result in quicker court rulings and safer automobiles, substantially increasing public trust in these innovative ML-based technologies, particularly the associated artifacts [44].

6.4.1 Importance of XAI in Autonomous Vehicles. XAI is crucial to advancing autonomous vehicles, because it provides openness in their decision-making processes. Integrating XAI, techniques lead autonomous automobiles to provide explicit explanations for their actions. This transparency fosters trust among authorities, passengers, and the general public by making it possible for them to comprehend the logic underlying route planning, navigation, and vehicle reaction to changing driving conditions. XAI also makes it easy to eliminate potential flaws inside autonomous vehicles, ensuring the safety, reliability, and moral adherence of autonomous automobiles on the road. XAI promotes accountability, which helps with the adoption of autonomous vehicles to create a safer transportation system. Table 9 lists a few recent studies on autonomous cars using XAI techniques.

6.5 Open Source Intelligence (OSINT)

The process of collecting and analyzing data from publicly available sources to produce actionable intelligence is known as OSINT. Newspapers, public websites, social media platforms, commercial data, and other things that are available to the general public are some examples of these sources. To acquire information, evaluate risks, and make wise judgments, a variety of organizations, including intelligence agencies, corporations, and researchers, frequently use OSINT. Information retrieval in **open-source intelligence (OSI)**, in contrast to **signals intelligence (SIGINT)**, is tightly constrained to publicly available content. However, OSINT has several important obstacles, notably in context, languages, and the amount of information delivered. A related question is how much information can be trusted and how much impact news from sources should have on the results of their aggregates. When considering adversarial attacks against OSINT techniques and systems, it is important to keep this in mind [29]. Explainability may help identify these attacks, which would help lessen their impact. In Reference [139], the authors addressed how **cyber threat intelligence (CTI)** can be widened with a comprehensible CTI framework. They proposed an approach for identifying **domain generation algorithm (DGA)**-based traffic utilizing statistical features with datasets of 55 DGA families. They also used LIME and SHAP to enhance the explainability of the results of the model by combining XAI and OSINT for trust problems.

6.5.1 Importance of XAI in OSINT. Since knowing that an attack on an intelligent system was launched is also a crucial input from an intelligence viewpoint, explainability may provide additional pertinent information. However, even when reporting on current events, some false information may be erroneous, confusing, or unknowable at the time of reporting rather than being purposefully manufactured. OSINT is more difficult when there are ongoing events, since the information continually changes due to new intelligence or the event itself. Explainability would enable, for example, the reporting of error margins on reported values, making it easier to ascertain the influence of erroneous information particles on the overall outcomes of machine learning. By offering clear explanations for data patterns and trends, XAI promotes more effective data use, allowing analysts to concentrate on the most important information and base their judgments on accurate and timely knowledge. Additionally, by assisting in the thorough evaluation of risks and threats discovered by OSINT, XAI enables analysts to prioritize and deal with possible problems while having a full grasp of the underlying variables driving the outcomes of intelligence gathering.

6.6 Human Resource Management (HRM)

HRM is the strategic approach to managing an organization's most valuable assets, its employees, in a way that maximizes their performance and contribution to the company's overall objectives. The goal of HRM is to maximize human capital to support organizational success. It includes a variety of tasks, such as recruiting, training, performance evaluation, employee engagement, and strategic workforce planning. Integrating transparent and understandable AI methods into managing human resources is necessary to connect XAI with HRM. The authors of Reference [151] emphasized that, given the increasing usage of AI in many different industries, HRM decisions must be transparent. They stress the need for XAI methodologies and the relevance of unbiased, moral, and equitable procedures inside AI-using businesses. The SLR is focused on the design, accuracy, accountability, and data-processing efforts of XAI approaches. The proposed integrated framework aims to bridge the knowledge gap between open HRM practices and AI by offering business and academia relevant data on potential XAI applications inside HRM processes.

6.6.1 Importance of XAI in HRM. Organizations may create a more effective and equitable workplace by integrating XAI into their HRM procedures, assuring fair decision-making, encouraging employee growth, and maximizing workforce tactics. Improved employee satisfaction and

Table 10. Recent Work on HRM Using Different XAI Approaches

Ref.	Objective	Method	Dataset	Year
Votto et al. [151]	Reveal the existence of AI within HRM	-	-	2023
Baun et al. [9]	Enhance AI Adoption in HRM	-	-	2023
Hofeditz et al. [52]	Examine how AI-driven candidate suggestions influence hiring choices and the potential moderating role of an XAI approach	Online experimental survey	-	2023
Fischer et al.	Focused on unwanted job turnover and HRM support	Utilized LIME, SHAP, DiCE, and PDP to showcase their explanatory abilities for data-driven decision-making	IBM HR Analytics Employee Attrition & Performance	2023
Bhattacharya et al. [11]	Integrate blockchain and explainable AI (xAI) to enhance transparency and efficiency in staffing and recruitment	Used AaJeeViKa, a fusion scheme, which combines XAI and blockchain	Business school placement dataset	2022

organizational success result from HR professionals making more informed and objective decisions thanks to XAI, which promotes a deeper knowledge of the elements driving HR decisions. Additionally, XAI supports the development of a culture of trust and responsibility among the workforce by ensuring that HR procedures comply with ethical standards and legal obligations. By utilizing XAI's capabilities, HRM can become a more strategic and employee-focused section, fostering long-term growth and giving the company a competitive edge. Some of the recent works in the domain of HRM using different XAI approaches are given in Table 10.

6.7 XAI for Education and Training

Education and training are key elements of personal and professional development, providing individuals with the skills and knowledge necessary to navigate various aspects of life. Instilling critical thinking, problem-solving skills, and practical knowledge necessary for success in a world that is always changing serves as the cornerstone of human capital development. Even though they are essential, education and training may be improved even further by integrating XAI. To create a more open and understandable learning environment, educators and students may use XAI to obtain deeper insights into the decision-making processes of AI-driven educational systems (for example, Educational Data Mining and Learning Management Systems). This transparency not only encourages confidence in AI systems but also allows teachers to customize learning opportunities, understand student requirements by evaluating performance, and help students comprehend and navigate the educational environment powered by AI.

6.7.1 Importance of XAI in Education and Training. XAI can assist instructors in better understanding how automated grading software evaluates student work so they can give more useful and individualized comments. To build confidence in AI-driven recommendations, XAI may provide transparency when proposing educational materials. This will allow students and teachers to understand why particular items are recommended. Learners may identify their strengths and areas for progress by using XAI to understand better how AI models evaluate their skill development. Additionally, XAI can guarantee that AI systems deployed in educational settings follow moral standards and support impartiality, openness, and accountability in decision-making procedures. Table 11 gives some of the recent works in education and training that are based on the use of different approaches to XAI.

Table 11. Recent Work on XAI in Education and Training

Ref.	Objective	Methodology	Dataset	Year
Farrow et al. [37]	Understand the role of XAI in education and identify practical or ethical limitations concerning transparency in learning and teaching	-	-	2023
Lee et al. [80]	Assess the impact of an XAI-driven Self-Regulated Learning Training System on student learning outcomes and self-regulated learning behavior	Online experimental survey	-	2023
Reeder et al. [115]	Develop a dual-phase strategy for improving both the efficacy and comprehensibility of AI systems in educational contexts	Three types of explanations: technical text, visual word clouds, simple text	IN-VALSI	2023
Rachha et al. [113]	Analyze the unique requirements of XAI in education and propose tailored adaptation strategies while identifying research gaps and future directions	-	-	2023
Kartik et al. [63]	Improve student performance prediction accuracy using ML and XAI	Random forest and LSTM; employed CNN and SHAP for model explainability	Jordan	2023
Fiok et al. [42]	Review the capabilities, limitations, and desiderata of XAI tools, focusing on their application in AI in education and presenting varied user groups and expectations	-	-	2022
Melo et al. [96]	Apply and evaluate XAI methods to predict school dropout among students in a Brazilian school	LIME	IFRN	2022

6.8 XAI in Finance

Referring to finance within AI, we mean the use of complex computer algorithms and data-driven models for financial data analysis, market trend prediction, and investment strategy optimization.

Algorithmic trading, fraud detection, risk management, personalized financial advising, and credit scoring are just a few of the many uses of AI in finance. Technology and digitalization have an impact on many different fields, offering benefits but also posing hazards. Therefore, regulators mandate high degrees of transparency, ensuring the traceability of choices for third parties, when decision-makers in highly regulated industries like finance apply these technology advances—especially AI. In this situation, XAI is crucial: Integrating transparent and understandable AI methods into financial decision-making processes is necessary for integrating XAI with finance.

6.8.1 Importance of XAI in Finance. Organizations may ensure improved transparency and comprehension of intricate financial models, investment plans, and risk assessments by implementing XAI in finance. While simultaneously assuring adherence to ethical norms and legal requirements, XAI provides comprehensive explanations for the variables impacting financial decisions, empowering stakeholders to make more responsible and informed judgments. A more dependable and trustworthy financial system is promoted by XAI's assistance in recognizing potential biases and faults inside financial algorithms. The financial sector may create a more transparent and accountable financial ecosystem, better risk management, and improve decision-making by utilizing XAI's capabilities. Table 12 includes some works on finance using different XAI approaches.

7 Discussion

Recently, researchers have increased interest in proposing various approaches utilizing XAI for better understandability of the results produced by the AI-based approaches. In the proposed SLR study, various methodologies, challenges, and applications of contemporary XAI are considered

Table 12. Recent Work on XAI in Finance

Ref.	Objective	Method	Dataset	Year
Torky et al. [141]	Create a novel XAI model for automatic recognition of financial crisis roots	Pigeon optimizer for feature selection; Gradient Boosting classifier for classification	JSTdatasetR3	2023
Weber et al. [154]	Provide a comprehensive overview of XAI applications in Finance, categorizing relevant articles based on utilized methods and objectives	-	-	2023
Ghosh et al. [43]	Utilize hybrid predictive frameworks to anticipate global financial stress patterns, leveraging EEMD, Facebook's Prophet algorithms, and LSTM alongside XAI techniques for improved model interpretation	-	Financial stress data regulated by the OFR	2023
Chen et al. [25]	Conduct a bibliometric study of the XAI literature in finance	-	-	2023
Zhou et al. [168]	Enhance the explainability of DL models in the financial industry and categorize them based on characteristics	Explainable framework based on SHAP	-	2023
Yeo et al. [160]	Create a method for detecting financial fraud that is both accurate and transparent	-	Observations of Chinese non-financial companies	2023
Kovvuri et al. [75]	Apply XAI to a global equity fund dataset to uncover insights into portfolio diversification's impact on fund performance across G10 countries.	Used XGBoost and Shapley values	Morningstar Direct database	2023
Sai et al. [124]	Develop a novel model to classify the transaction as fraudulent or legitimate	Five ML algorithms and two DNN algorithms, i.e., Artificial Neural Networks and CNN are used	Morningstar Direct database	2023
Mill et al. [99]	Promote increased research in XAI for credit card fraud detection and emphasize the regulatory changes and the operational environment	-	-	2023

for potential readers to enhance their understanding and knowledge. However, there are various challenges related to technological, legal, and practical aspects of explainability approaches. Below, we highlighted some potential challenges to be considered as a future research direction.

Challenges in the Development of Explainability Approaches for Complex AI Models. There are significant challenges to providing clear and understandable justifications when dealing with complex model architectures, high-dimensional data, and non-linear decision limits [5]. Complex strategies are needed for an effective resolution due to the opaque nature of black-box models, feature significance and relevance assessment, and the interpretability of time-series data [32]. Additional levels of complexity are added by ensuring the robustness and uncertainty assessment of AI models [77], constructing interpretable representation learning methods, and creating universally applicable model-agnostic explanations. Furthermore, creating effective and responsive XAI strategies is required to deliver real-time explanations in time-sensitive settings [88]. To create transparent and understandable AI systems, XAI, research must constantly evolve, incorporating a variety of techniques and multidisciplinary approaches.

Possible Solutions. To address the challenges associated with explaining complex AI frameworks, it is recommended to use ensemble models for black-box transparency, interpretable representation learning for time-series data, LIME for local feature interpretation, Bayesian methods and ensemble models for robustness and uncertainty assessment, and SHAP values for broadly applicable model-agnostic explanations. Using strategies such as caching and parallelization with real-time visualization approaches like dynamic dashboards may maximize computational efficiency for real-time explanations in time-sensitive circumstances. Fostering transparent and intelligible AI systems requires ongoing progress in XAI research that integrates these strategies.

Tradeoff Accuracy vs. Explanations Provided. It is difficult to maintain a balance between the demand for transparent decision-making processes and the necessity for reliable predictions when trying to overcome the interpretability-accuracy tradeoff in DL, hybrid, and ensemble models, although it is identified using the proposed SLR accuracy and generalizability. Therefore, it is recommended that to overcome these obstacles, research must be done on ways to improve model explainability without sacrificing accuracy, such as using model-agnostic explanation approaches or adding simpler models. Finding the optimal balance is a continuous and dynamic process in DL.

Solutions. To tackle the tradeoff between interpretability and accuracy in complicated models, explanation methodologies must be employed that are not dependent on a particular model. XAI techniques, such as LIME or SHAP values, offer clear insights into model choices without sacrificing accuracy. Alternatively, complex hybrid modules might be replaced with basic ones. This method achieves a better understanding without sacrificing total prediction accuracy. Decision-making procedures must be transparent to achieve the greatest possible balance between accuracy and explanations in DL, which can only be achieved by ongoing dynamic methodological adaptation.

Best Practices for Real-world Applications. Techniques that preserve both qualities without sacrificing either are necessary to manage the interpretability vs. model accuracy tradeoff in real-world XAI applications. Numerous less-complex models are successfully combined using ensemble approaches, such as Random Forests, to improve performance without sacrificing interpretability. Hybrid models integrate interpretable decision-making approaches, such as decision trees with DL for feature extraction, to provide results that are easy to grasp. Furthermore, regularization techniques such as L2 and L1, which optimize algorithms during training, may increase interpretability. Tools for post hoc explanation, such as LIME or SHAP, provide thorough evaluations of feature contribution and help comprehend difficult model choices. Validating user input ensures that explanations satisfy relevant requirements while balancing interpretability and correctness fairly.

Which Forms of XAI Are Commonly Used? Researchers have developed different explanation formats to meet the demands of different end-users in different disciplines. Nonetheless, a major gap persists in understanding what defines an explanation and its necessary prominent features. To operate the software system according to their demands, different user types, such as domain experts, decision-makers, AI practitioners, consumers, and researchers, look for explanations.

Solutions. The field of XAI can be effectively analyzed using a detailed study to determine the best explanation format for every situation. Understanding the demands of users can lead to agreement on explanation forms. Making sure that XAI methods work well in different situations and helping everyone agree on what a good explanation looks like in AI are two advantages of creating explanations that fit the needs of all the people involved.

Which Is Preferable, XAI or ***Interpretable Machine Learning (IML)****?* The key aim of both XAI and IML is transparency. However, selecting between IML and XAI might still be difficult because of the different nature of both approaches. IML includes models that are inherently interpretable and simple by nature (decision tree and linear regression models), whereas XAI provides models to explain the prediction of both inherently interpretable models and black boxes. The selection between highly accurate XAI models and inherently interpretable IML models depends on the needs of the particular application domain. Maintaining a balance between interpretability and accuracy is the most crucial issue. So, the disadvantages and advantages of both approaches must be kept in mind during the final selection process.

Solutions. Understanding the needs of the application domain can lead to agreement on using XAI or IML approaches. Use IML approaches to explain small and naturally interpretable models and XAI approaches to explain large and complex models. Also, IML is beneficial in situations where accuracy is secondary, whereas XAI is beneficial in situations where accuracy is crucial.

Emerging Challenges in XAI. Including cutting-edge technologies in AI, such as blockchain, IoT, and quantum computing, results in new challenges to the field of XAI. Ensuring the explanation works well for various applications, and user types in dynamic contexts such as intelligent and real-time data processing systems is challenging. It becomes difficult to justify the decision-making process in the case of the integration of quantum computing into AI. Privacy concerns must also be considered while managing sensitive data, which requires careful balancing between data protection and transparency. Moreover, there is still debate over a sufficient explanation, making it challenging to standardize AI thinking across various industries and application cases.

Solutions. XAI researchers can propose and develop new algorithms that aim to generate solutions that are context-specific, scalable to various disciplines, and that can adapt to explainability by considering the situation, context, and user groups. When AI is integrated with advanced technologies such as IoT, blockchain, and quantum computing, creating hybrid XAI models to handle the complexity of AI systems can be the best choice. To maintain transparency and address privacy concerns, differentiable privacy techniques can be integrated into XAI to ensure the security of sensitive data. Finally, close cooperation with regulatory bodies and business leaders is also crucial for developing improved standards for XAI across various applications.

7.1 Ethical Considerations and Frameworks for Responsible AI Development

It is important to intertwine ethical framework with XAI-based approaches by ensuring that human values are not violated with the recent advancements [57]. The XAI-based system developed must ensure fairness, accountability, and transparency. For example, the explanation generated by the XAI applications must be easy for various related stakeholders to understand and comprehend. It should be understandable to those who lack technical education and expertise. Also, the primary purpose of XAI systems is to help decision-makers identify and mitigate model biases. The XAI-based should be able to provide explanations and traces for the decision-making that hints about bias. For example, the model should be able to highlight the Skewness of the model to one particular group when performing a classification problem. For unsupervised problems, the model should be able to ensure the training process is based on equal distributions from the different groups, which would result in fair result generation that is acceptable to each group. Moreover, researchers and practitioners need to propose XAI-based approaches that could support accountability for decision-making. It is crucial to identify and fix responsibilities, including those of practitioners, vendors, and operators of the system, for certain decisions made by the XAI system. Another important aspect that researchers and practitioners need to work with is ensuring users' data protection by adhering to **General Data Protection Regulation (GDPR)** standards. For example, the explanation provided by the XAI-based system should not portray personal and sensitive information that might be exploited. Additionally, with the increasing popularity of AI-based systems in various domains that foster autonomy, there is a debate about human roles and responsibilities connected to these systems. Therefore, it is evident that vendors and developers should adopt human-centric software development approaches when developing XAI-based applications aiming to enhance human capabilities in decision-making and ensure human autonomy.

Researchers and developers must closely liaise with cultural values and social norms when developing XAI-based applications to improve productivity and trust. Researchers, vendors, and developers need to consider the above-mentioned discussion to enhance the performance of XAI-based applications further by proposing research methodologies and practices.

7.2 Fostering Interdisciplinary Collaboration in XAI

Experts from various disciplines collaborate to solve the social, ethical, and technological issues AI raises in creating XAI. Effective collaboration is essential in ethics, cognitive science, computer science, economics, and healthcare. This interdisciplinary approach guarantees the technological strength and social responsibility of XAI technologies. Academic and research institutions can facilitate these links by planning cross-departmental initiatives and interdisciplinary seminars. Funding agencies can also help these initiatives by giving preference to initiatives promoting cooperation between specialists in other sectors. In this manner, we may create XAI systems that are comprehensive and mindful of the wider effects on society. Industry relationships ensure that AI systems are built with a thorough awareness of their wider social consequences, further enhancing this integration by applying theoretical findings to real-world circumstances. International XAI conferences and symposia offer essential venues for knowledge exchange and network building, promoting a deeper, more inclusive conversation about the direction of AI technology.

7.3 Emerging Domains for Future Research in XAI

Although our review goes into great detail about several significant areas where XAI is becoming more and more common, such as education, training, finance, HRM, autonomous cars, AMA, explainable object detection, and threat detection, there are still other developing areas that need more research. These include uses of XAI in smart city applications, where XAI may increase urban planning and governance, and environmental monitoring, where explainability can improve the transparency of AI-driven choices connected to conservation and climate change. Additional investigation may focus on the application of AI in cybersecurity, where XAI helps explain threat assessments and defense strategies, and in agriculture, particularly in precision farming techniques, where it may optimize and clarify agricultural decisions. Furthermore, using XAI in software engineering might facilitate troubleshooting, optimizing, and elucidating software behavior and performance. Last, XAI can greatly assist sophisticated decision-making systems in various sectors, guaranteeing that AI suggestions are clear and intelligible. We suggest that, to increase the applications of XAI and boost its social advantages across a wider variety of businesses, these domains should be the focus of future studies.

7.4 Cross-domain Applicability of XAI Methodologies

Recently, XAI has become increasingly useful and important across many application areas due to its explainability advantage. With this, researchers are proposing various XAI-based approaches. Therefore, scholars and professionals must understand which XAI techniques best apply in certain situations. In particular, techniques such as LIME may elucidate intricate model choices in medical image analysis, essential to assisting physicians in comprehending and having confidence in diagnostic findings. Software developers may debug and optimize algorithms with the help of SHAP values, which provide insights into the many factors impacting code performance. In legal NLP, LIME is also helpful for deciphering intricate language patterns, while SHAP values improve transparency and transparency in financial fraud detection. Similarly, SHAP can be used to understand the complex decision-making process of various ML classifiers in classifying end-user reviews into distant rationale elements [145]. Moreover, counterfactual explanations improve safety in self-driving car applications, whereas TreeSHAP clarifies the reasons that affect forecasts.

Integrated gradients are useful for environmental monitoring, and anchors enhance e-commerce recommendation systems. Transparency and confidence in AI applications are enhanced by these unique XAI techniques, which guarantee precise forecasts and offer a thorough comprehension of model decisions.

7.5 Threads to Validity: Challenges Encountered in the SLR Process

In this section, we elaborated on the challenges and threats that threaten the review process that must be resolved to guarantee the robustness and coherence of the proposed SLR. One possible threat is the search procedure, which is a major limitation of this study; even with an extensive search procedure in place that includes many databases, there is still a possibility that certain important articles on XAI might be missed. This might be happening due to several factors. First, how research databases indexed the published articles in their system might impact how well relevant articles can be retrieved for analysis. Also, XAI is an emerging research field where researchers frequently propose and publish new approaches that pose another challenge to the search criteria, as authors might have used emerging research tags that do not match the search keywords and criteria. However, as evident from the proposed SLR, XAI is a multidisciplinary research area where researchers use it in various emerging domains; it becomes challenging to search related articles from several perspectives, making the search even more difficult. Moreover, we put extensive efforts into identifying the most related articles on XAI to be included in the proposed SLR that satisfies the inclusion and exclusion criteria; still, the inherent complexity of information structure and retrieval systems can lead to unintentional omissions from the manuscript dataset. Another threat could be publication bias, as we might have missed important and critical XAI information sources that do not fit under the inclusion and exclusion criteria, such as non-English reports and other sources that could be identified through gray literature. Currently, researchers are using gray literature as an alternative method to extract the latest information and industry trends about the topic of concentration. The proposed approach can be extended to overcome this threat by including gray literature to identify processes, methodologies, best practices from white papers, and so on [159]. Additionally, the time constraints of the review process can affect how thorough the literature search is. However, we included the latest research articles in the proposed SLR to present the latest information on XAI. Still, there is a possibility that some research papers will be published after the submission of the proposed SLR. Variations in methodological rigor between studies may affect the final evaluation, even with efforts to standardize the procedure. By transparently discussing these challenges, we aim to enhance the interpretability of the proposed SLR findings and provide readers with a nuanced understanding of the potential limitations inherent in the SLR methodology. Due to the wide range of study designs and the subjective nature of quality assessment, which is an essential part of SLRs, challenges arise. Variations in methodological rigor between studies may affect the final evaluation, even with efforts to standardize the procedure. By transparently discussing these challenges, we aim to enhance the interpretability of our review findings and provide readers with a nuanced understanding of the potential limitations of the SLR methodology.

8 Conclusion and Future Directions

Research on methods to elucidate the inner logic of a learning algorithm, the induced model, or a knowledge-based approach for inference is recognized as a core area of XAI. While terms like "interpretable machine learning" exist, XAI emphasizes a broader application scope. Though numerous studies and events contribute to this knowledge, it remains fragmented. This systematic review organizes this extensive knowledge into four clusters: (I) reviews, (II) theories, (III) methods for explaining inferential processes, and (IV) applications of XAI (in different domains). Scholars have advanced various explanation formats, from textual to visual aids, rules, and dialogues, tailored for different end-users and applications, incorporating insights from diverse fields. However, defining explanations and establishing their properties, particularly for non-experts, presents a challenge. The formalization and operationalization of explainability are ongoing tasks, necessitating a comprehensive XAI framework adaptable to diverse contexts and end-users.

In this quest for transparent and comprehensible AI systems, a few prospects need analysis.

First, focusing on developing hybrid models incorporating symbolic reasoning with the benefits of DL has much potential. More accurate and efficient explanatory models may be possible with developments in defensible reasoning and argumentation and an emphasis on incorporating human-in-the-loop techniques for creating large knowledge bases.

Balancing interpretability and model correctness requires combining modern learning strategies with reasoning methods. Mixing the connectionist and symbolic paradigms and continuously investigating their convergence to produce precise explanations is advisable.

The use of XAI in a range of industries, including banking, healthcare, and autonomous systems, necessitates the development of customized methods and guidelines compatible with each industry's unique challenges and constraints.

Examining the ethical ramifications and establishing guidelines for responsible AI research and use is essential to ensuring the ethical use of AI technologies and public acceptance of them. Future ethical AI system development can benefit from interdisciplinary collaboration and thoughtful assessment of XAI's social impacts.

References

[1] Waleed Aldhahi and Sanghoon Sull. 2023. Uncertain-CAM: Uncertainty-based ensemble machine voting for improved COVID-19 CXR classification and explainability. *Diagnostics* 13, 3 (2023), 441.

[2] Sajid Ali, Tamer Abuhmed, Shaker El-Sappagh, Khan Muhammad, Jose M. Alonso-Moral, Roberto Confalonieri, Riccardo Guidotti, Javier Del Ser, Natalia Díaz-Rodríguez, and Francisco Herrera. 2023. Explainable artificial intelligence (XAI): What we know and what is left to attain trustworthy artificial intelligence. *Inf. Fusion* 99 (2023), 101805.

[3] Guy Amit, Fabiana Fournier, Shlomit Gur, and Lior Limonad. 2022. Model-informed LIME extension for business process explainability. In *International Workshop on Process Management in the AI Era (PMAI'23) co-located with 31st International Joint Conference on Artificial Intelligence (IJCAI'23)*.

[4] Shama Ams. 2023. Blurred lines: The convergence of military and civilian uses of AI & data use and its impact on liberal democracy. *Int. Polit.* 60, 4 (2023), 879–896.

[5] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins et al. 2020. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* 58 (2020), 82–115.

[6] Roy Assaf and Anika Schumann. 2019. Explainable deep neural networks for multivariate time series predictions. In *28th International Joint Conference on Artificial Intelligence (IJCAI'19)*. International Joint Conferences on Artificial Intelligence Organization, 6488–6490. DOI: https://doi.org/10.24963/ijcai.2019/932

[7] A. B. Athira, S. D. Madhu Kumar, and Anu Mary Chacko. 2023. A systematic survey on explainable AI applied to fake news detection. *Eng. Applic. Artif. Intell.* 122 (2023), 106087.

[8] José Luis Corcuera Bárcena, Pietro Ducange, Francesco Marcelloni, Giovanni Nardini, Alessandro Noferi, Alessandro Renda, Fabrizio Ruffini, Alessio Schiavo, Giovanni Stea, and Antonio Virdis. 2023. Enabling federated learning of explainable AI models within beyond-5G/6G networks. *Comput. Commun.* 210 (2023), 356–375.

[9] Lorenz Baum, Patrick Weber, and Laura-Marie Kolb. 2023. The explanation matters: Enhancing AI adoption in human resource management. In *PACIS 2023 Proceedings*, Vol. 17. https://aisel.aisnet.org/pacis2023/17

[10] Jagger S. Bellagarda and Adnan M. Abu-Mahfouz. 2022. An updated survey on the convergence of distributed ledger technology and artificial intelligence: Current state, major challenges and future direction. *IEEE Access* 10 (2022), 50774–50793.

[11] Pronaya Bhattacharya, Mohd Zuhair, Debanjana Roy, Vivek Kumar Prasad, and Darshan Savaliya. 2022. AaJeeViKa: Trusted explainable AI based recruitment scheme in smart organizations. In *5th International Conference on Contemporary Computing and Informatics (IC3I'22)*. IEEE, 1002–1008.

[12] Or Biran and Courtenay Cotton. 2017. Explanation and justification in machine learning: A survey. In *IJCAI-17 Workshop on Explainable AI (XAI'17)*. 8–13.

[13] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. 2020. YOLOv4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934* (2020).

[14] Katarzyna Borys, Yasmin Alyssa Schmitt, Meike Nauta, Christin Seifert, Nicole Krämer, Christoph M. Friedrich, and Felix Nensa. 2023. Explainable AI in medical imaging: An overview for clinical practitioners–Beyond saliency-based XAI approaches. *European Journal of Radiology* 162 (2023), 110786. https://doi.org/10.1016/j.ejrad.2023.110786

[15] Lydia Bouzar-Benlabiod, Khaled Harrar, Lahcen Yamoun, Mustapha Yacine Khodja, and Moulay A. Akhloufi. 2023. A novel breast cancer detection architecture based on a CNN-CBR system for mammogram classification. *Comput. Biol. Med.* 163 (2023), 107133.

[16] Luca Brunese, Francesco Mercaldo, Alfonso Reginelli, and Antonella Santone. 2020. Explainable deep learning for pulmonary disease and coronavirus COVID-19 detection from X-rays. *Comput. Meth. Prog. Biomed.* 196 (2020), 105608.

[17] Francesco Paolo Caforio, Giuseppina Andresini, Gennaro Vessio, Annalisa Appice, and Donato Malerba. 2021. Leveraging Grad-CAM to improve the accuracy of network intrusion detection systems. In *International Conference on Discovery Science*. Springer, 385–400.

[18] Nicola Capuano, Giuseppe Fenza, Vincenzo Loia, and Claudio Stanzione. 2022. Explainable artificial intelligence in cybersecurity: A survey. *IEEE Access* 10 (2022), 93575–93600.

[19] Mattia Carletti, Chiara Masiero, Alessandro Beghi, and Gian Antonio Susto. 2019. Explainable machine learning in Industry 4.0: Evaluating feature importance in anomaly detection to enable root cause analysis. In *IEEE International Conference on Systems, Man and Cybernetics (SMC'19)*. IEEE, 21–26.

[20] Luciano Caroprese, Eugenio Vocaturo, and Ester Zumpano. 2022. Argumentation approaches for explanaible AI in medical informatics. *Intell. Syst. Applic.* 16 (2022), 200109.

[21] Ahmad Chaddad, Jihao Peng, Jian Xu, and Ahmed Bouridane. 2023. Survey of explainable AI techniques in healthcare. *Sensors* 23, 2 (2023), 634.

[22] Vinay Chamola, Vikas Hassija, A. Razia Sulthana, Debshishu Ghosh, Divyansh Dhingra, and Biplab Sikdar. 2023. A review of trustworthy and explainable artificial intelligence (XAI). *IEEE Access* 11 (2023), 78994–79015. https://doi.org/10.1109/ACCESS.2023.3294569

[23] Chun-Hao Chang, Elliot Creager, Anna Goldenberg, and David Duvenaud. 2017. Interpreting neural network classifications with variational dropout saliency maps. In *International Conference on Neural Information Processing Systems (NIPS'17)*. 1–9.

[24] Haomin Chen, Catalina Gomez, Chien-Ming Huang, and Mathias Unberath. 2022. Explainable medical imaging AI needs human-centered design: Guidelines and evidence from a systematic review. *NPJ Digit. Med.* 5, 1 (2022), 156.

[25] Xun-Qi Chen, Chao-Qun Ma, Yi-Shuai Ren, Yu-Tian Lei, Ngoc Quang Anh Huynh, and Seema Narayan. 2023. Explainable artificial intelligence in finance: A bibliometric review. *Finance Research Letters* 56 (2023), 104145. https://doi.org/10.1016/j.frl.2023.104145

[26] Roberto Confalonieri, Tillman Weyde, Tarek R. Besold, and Fermín Moscoso del Prado Martín. 2019. Trepan reloaded: A knowledge-driven approach to explaining artificial neural networks. arXiv preprint arXiv:1906.08362 (2019).

[27] Enyan Dai, Tianxiang Zhao, Huaisheng Zhu, Junjie Xu, Zhimeng Guo, Hui Liu, Jiliang Tang, and Suhang Wang. 2022. A comprehensive survey on trustworthy graph neural networks: Privacy, robustness, fairness, and explainability. *arXiv preprint arXiv:2204.08570* (2022).

[28] Tanusree De, Prasenjit Giri, Ahmeduvesh Mevawala, Ramyasri Nemani, and Arati Deo. 2020. Explainable AI: A hybrid approach to generate human-interpretable explanation for deep learning prediction. *Proced. Comput. Sci.* 168 (2020), 40–48.

[29] Sean M. Devine and Nathaniel D. Bastian. 2019. Intelligent systems design for malware classification under adversarial conditions. *arXiv preprint arXiv:1907.03149* (2019).

[30] Weiping Ding, Mohamed Abdel-Basset, Hossam Hawash, and Ahmed M. Ali. 2022. Explainability of artificial intelligence methods, applications and challenges: A comprehensive survey. *Information Sciences* 615 (2022), 238–292. https://doi.org/10.1016/j.ins.2022.10.013

[31] Jiqian Dong, Sikai Chen, Mohammad Miralinaghi, Tiantian Chen, Pei Li, and Samuel Labi. 2023. Why did the AI make that decision? Towards an explainable artificial intelligence (XAI) for autonomous driving systems. *Transport. Res. Part C: Emerg. Technol.* 156 (2023), 104358.

[32] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).

[33] Rudresh Dwivedi, Devam Dave, Het Naik, Smiti Singhal, Rana Omer, Pankesh Patel, Bin Qian, Zhenyu Wen, Tejal Shah, Graham Morgan et al. 2023. Explainable AI (XAI): Core ideas, techniques, and solutions. *Comput. Surv.* 55, 9 (2023), 1–33.

[34] Daniel Dworak and Jerzy Baranowski. 2022. Adaptation of Grad-CAM method to neural network architecture for LiDAR pointcloud object detection. *Energies* 15, 13 (2022), 4681.

[35] Ida Merete Enholm, Emmanouil Papagiannidis, Patrick Mikalef, and John Krogstie. 2022. Artificial intelligence and business value: A literature review. *Inf. Syst. Front.* 24, 5 (2022), 1709–1734.

[36] Farhan Farabi, Farhan Hossen, Farhan Monsur, Mir Araf Hossain, and Md Mohibul Hasan. 2023. *Explainable Breast Cancer Detection from Histopathology Images Using Transfer Learning and XAI.* Ph. D. Dissertation. Brac University.

[37] Robert Farrow. 2023. The possibilities and limits of XAI in education: A socio-technical perspective. *Learning, Media and Technology* 48, 2 (2023), 266–279. https://oro.open.ac.uk/87862/

[38] Reuben Feinman, Ryan R. Curtin, Saurabh Shintre, and Andrew B. Gardner. 2017. Detecting adversarial samples from artifacts. *arXiv preprint arXiv:1703.00410* (2017).

[39] Raphael Féraud and Fabrice Clérot. 2002. A methodology to explain neural network classification. *Neural Netw.* 15, 2 (2002), 237–246.

[40] Rubén R. Fernández, Isaac Martín De Diego, Víctor Aceña, Alberto Fernández-Isabel, and Javier M. Moguerza. 2020. Random forest explainability using counterfactual sets. *Inf. Fusion* 63 (2020), 196–207.

[41] Emmanuel Ferreyra, Hani Hagras, Mathias Kern, and Gilbert Owusu. 2019. Depicting decision-making: A type-2 fuzzy logic based explainable artificial intelligence system for goal-driven simulation in the workforce allocation domain. In *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE'19).* IEEE, 1–6.

[42] Krzysztof Fiok, Farzad V. Farahani, Waldemar Karwowski, and Tareq Ahram. 2022. Explainable artificial intelligence for education and training. *J. Defense Model. Simul.* 19, 2 (2022), 133–144.

[43] Indranil Ghosh and Pamucar Dragan. 2023. Can financial stress be anticipated and explained? Uncovering the hidden pattern using EEMD-LSTM, EEMD-prophet, and XAI methodologies. *Complex Intell. Syst.* 9, 4 (2023), 4169–4193.

[44] Jon Arne Glomsrud, André Ødegårdstuen, Asun Lera St. Clair, and Øyvind Smogeli. 2019. Trustworthy versus explainable AI in autonomous vessels. In *International Seminar on Safety and Security of Autonomous Vessels (ISSAV'19) and European STAMP Workshop and Conference (ESWC'19).*

[45] Mara Graziani, Vincent Andrearczyk, Stéphane Marchand-Maillet, and Henning Müller. 2020. Concept attribution: Explaining CNN decisions to physicians. *Comput. Biol. Med.* 123 (2020), 103865.

[46] Arjan M. Groen, Rik Kraan, Shahira F. Amirkhan, Joost G. Daams, and Mario Maas. 2022. A systematic review on the use of explainability in deep learning systems for computer aided diagnosis in radiology: Limited use of explainable AI? *European Journal of Radiology* 157 (2022), 110592. https://doi.org/10.1016/j.ejrad.2022.110592

[47] David Gunning. 2016. *Broad Agency Announcement Explainable Artificial Intelligence (XAI).* Technical Report. Defense Advanced Research Projects Agency.

[48] David Gunning, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf, and Guang-Zhong Yang. 2019. XAI—Explainable artificial intelligence. *Sci. Robot.* 4, 37 (2019), eaay7120.

[49] Ola Hall, Mattias Ohlsson, and Thorsteinn Rögnvaldsson. 2022. A review of explainable AI in the satellite data, deep machine learning, and human poverty domain. *Patterns* 3, 10 (2022), 100600.

[50] Miseon Han and Jeongtae Kim. 2019. Joint banknote recognition and counterfeit detection using explainable artificial intelligence. *Sensors* 19, 16 (2019), 3607.

[51] Julian Hatwell, Mohamed Medhat Gaber, and R. Muhammad Atif Azad. 2020. Ada-WHIPS: Explaining AdaBoost classification with applications in the health sciences. *BMC Med. Inform. Decis. Mak.* 20, 1 (2020), 1–25.

[52] Lennart Hofeditz, Sünje Clausen, Alexander Rieß, Milad Mirbabaie, and Stefan Stieglitz. 2022. Applying XAI to an AI-based system for candidate management to mitigate bias and discrimination in hiring. *Electron. Mark.* 32, 4 (2022), 2207–2233.

[53] Katharina Holzinger, Klaus Mak, Peter Kieseberg, and Andreas Holzinger. 2018. Can we trust machine learning results? Artificial intelligence in safety-critical decision support. *ERCIM News* 112 (2018), 42–43.

[54] Peter Hönig and Wilfried Wöber. 2023. Explainable object detection in the field of search and rescue robotics. In *International Conference on Robotics in Alpe-Adria Danube Region.* Springer, 37–44.

[55] Ling Huang, Anthony D. Joseph, Blaine Nelson, Benjamin I. P. Rubinstein, and J. Doug Tygar. 2011. Adversarial machine learning. In *4th ACM Workshop on Security and Artificial Intelligence.* 43–58.

[56] Tim Hulsen. 2023. Explainable artificial intelligence (XAI): Concepts and challenges in healthcare. *AI* 4, 3 (2023), 652–666.

[57] Aisha Zahid Huriye. 2023. The ethics of artificial intelligence: Examining the ethical considerations surrounding the development and use of AI. *Am. J. Technol.* 2, 1 (2023), 37–44.

[58] Maksims Ivanovs, Beate Banga, Valters Abolins, and Krisjanis Nesenbergs. 2022. Methods for explaining CNN-based BCI: A survey of recent applications. In *IEEE 16th International Scientific Conference on Informatics (Informatics'22).* IEEE, 137–141.

[59] Alexander Jung and Pedro H. J. Nardelli. 2020. An information-theoretic approach to personalized explainable machine learning. *IEEE Signal Process. Lett.* 27 (2020), 825–829.

[60] Yeon-Jee Jung, Seung-Ho Han, and Ho-Jin Choi. 2021. Explaining CNN and RNN using selective layer-wise relevance propagation. *IEEE Access* 9 (2021), 18670–18681.

[61] Md Shahin Kabir and Mohammad Nazmul Alam. 2023. IoT, big data and AI applications in the law enforcement and legal system: A review. *International Research Journal of Engineering and Technology (IRJET)* 10, 5 (2023), 1777–1789.

[62] Ching-Yu Kao, Junhao Chen, Karla Markert, and Konstantin Böttinger. 2022. Rectifying adversarial inputs using XAI techniques. In *30th European Signal Processing Conference (EUSIPCO'22)*. IEEE, 573–577.

[63] N. Kartik, R. Mahalakshmi, and K. A. Venkatesh. 2023. XAI-based student performance prediction: Peeling back the layers of LSTM and random forest's black boxes. *SN Comput. Sci.* 4, 5 (2023), 699.

[64] Shinji Kawakura, Masayuki Hirafuji, Seishi Ninomiya, and Ryosuke Shibasaki. 2023. Visual analysis of agricultural workers using explainable artificial intelligence (XAI) on class activation map (CAM) with characteristic point data output from OpenCV-based analysis. *Eur. J. Artif. Intell. Mach. Learn.* 2, 1 (2023), 1–8.

[65] Blen M. Keneni, Devinder Kaur, Ali Al Bataineh, Vijaya K. Devabhaktuni, Ahmad Y. Javaid, Jack D. Zaientz, and Robert P. Marinier. 2019. Evolving rule-based explainable artificial intelligence for unmanned aerial vehicles. *IEEE Access* 7 (2019), 17001–17016.

[66] Ausaf Ahmad Khan and Noor Alam Khan. 2023. Role of artificial intelligence in life insurance organization: In respect of human resource functions. *Journal of Management and Entrepreneurship* 8, 2 (2023), 1–14.

[67] Izhar Ahmed Khan, Nour Moustafa, Imran Razzak, Muhammad Tanveer, Dechang Pi, Yue Pan, and Bakht Sher Ali. 2022. XSRU-IoMT: Explainable simple recurrent units for threat detection in internet of medical things networks. *Fut. Gen. Comput. Syst.* 127 (2022), 181–193.

[68] Sujata Khedkar, Vignesh Subramanian, Gayatri Shinde, and Priyanka Gandhi. 2019. Explainable AI in healthcare. In *2nd International Conference on Advances in Science & Technology: Healthcare (ICAST'19)*.

[69] Peter Kieseberg, Bernd Malle, Peter Frühwirt, Edgar Weippl, and Andreas Holzinger. 2016. A tamper-proof audit and control system for the doctor in the loop. *Brain Inform.* 3 (2016), 269–279.

[70] Been Kim, Rajiv Khanna, and Oluwasanmi O. Koyejo. 2016. Examples are not enough, learn to criticize! Criticism for interpretability. In *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (Eds.). Vol. 29. Curran Associates, Inc., n.a.

[71] B. A. Kitchenham. 2004. *Procedures for Undertaking Systematic Reviews*. Joint Technical Report TR/SE0401 and 0400011T.1. Computer Science Department, Keele University and National ICT Australia Ltd.

[72] İbrahim Kök, Feyza Yıldırım Okay, Özgecan Muyanlı, and Suat Özdemir. 2023. Explainable artificial intelligence (XAI) for internet of things: A survey. *IEEE Internet of Things Journal* 10, 16 (2023), 14764–14779. https://doi.org/10.1109/JIOT.2023.3287678

[73] Suresh Kolekar, Shilpa Gite, Biswajeet Pradhan, and Abdullah Alamri. 2022. Explainable AI in scene understanding for autonomous vehicles in unstructured traffic environments on Indian roads using the inception U-Net Model with Grad-CAM visualization. *Sensors* 22, 24 (2022), 9677.

[74] Nikolas Koutsoubis. 2023. *Machine Learning-based Drone and Aerial Threat Detection for Increased Turret Gunner Survivability*. Ph. D. Dissertation. Rowan University.

[75] Veera Raghava Reddy Kovvuri, Hsuan Fu, Xiuyi Fan, and Monika Seisenberger. 2023. Fund performance evaluation with explainable artificial intelligence. *Fin. Res. Lett.* 58 (2023), 104419.

[76] I. Elizabeth Kumar, Suresh Venkatasubramanian, Carlos Scheidegger, and Sorelle Friedler. 2020. Problems with Shapley-value-based explanations as feature importance measures. In *International Conference on Machine Learning (ICML'20)*. PMLR, 5491–5500.

[77] Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Jure Leskovec. 2017. Interpretable & explorable approximations of black box models. *arXiv preprint arXiv:1707.01154* (2017).

[78] Stephen Law, Rikuo Hasegawa, Brooks Paige, Chris Russell, and Andrew Elliott. 2023. Explaining holistic image regressors and classifiers in urban analytics with plausible counterfactuals. *International Journal of Geographical Information Science* 37, 12 (2023), 1–22.

[79] Thai Le, Suhang Wang, and Dongwon Lee. 2020. GRACE: Generating concise and informative contrastive sample to explain neural network model's prediction. In *26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 238–248.

[80] HaeJin Lee and Nigel Bosch. 2023. EXcel My SRL: Explainable AI-driven Self-Regulated Learning Training System. OSF Registries. , n.a. pages. https://doi.org/10.17605/OSF.IO/5Z9JK

[81] Oscar Li, Hao Liu, Chaofan Chen, and Cynthia Rudin. 2018. Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions. In *AAAI Conference on Artificial Intelligence*.

[82] Peibo Li, Yixing Yang, Maurice Pagnucco, and Yang Song. 2022. Explainability in graph neural networks: An experimental survey. *arXiv preprint arXiv:2203.09258* (2022).

[83] Rui Li and Olga Gadyatskaya. 2023. Evaluating rule-based global XAI malware detection methods. In *International Conference on Network and System Security*. Springer, 3–22.

[84] Sinan Li, Tianfu Li, Chuang Sun, Ruqiang Yan, and Xuefeng Chen. 2023. Multilayer Grad-CAM: An effective tool towards explainable deep neural networks for intelligent fault diagnosis. *J. Manuf. Syst.* 69 (2023), 20–30.

[85] Wei-Yin Loh and Yu-Shan Shih. 1997. Split selection methods for classification trees. *Statistica Sinica* 7, 4 (1997), 815–840.

[86] Luca Longo, Randy Goebel, Freddy Lecue, Peter Kieseberg, and Andreas Holzinger. 2020. Explainable artificial intelligence: Concepts, applications, research challenges and visions. In *International Cross-domain Conference for Machine Learning and Knowledge Extraction*. Springer, 1–16.

[87] Hampus Lundberg, Nishat I. Mowla, Sarder Fakhrul Abedin, Kyi Thar, Aamir Mahmood, Mikael Gidlund, and Shahid Raza. 2022. Experimental analysis of trustworthy in-vehicle intrusion detection system using explainable artificial intelligence (XAI). *IEEE Access* 10 (2022), 102831–102841.

[88] Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Vol. 30. Curran Associates, Inc.

[89] Qing Lyu, Marianna Apidianaki, and Chris Callison-Burch. 2022. Towards faithful model explanation in NLP: A survey. *arXiv preprint arXiv:2209.11326* (2022).

[90] Dawid Macha, Michał Kozielski, Łukasz Wróbel, and Marek Sikora. 2022. RuleXAI—A package for rule-based explanations of machine learning model. *SoftwareX* 20 (2022), 101209.

[91] A. V. Shreyas Madhav and Amit Kumar Tyagi. 2022. Explainable artificial intelligence (XAI): Connecting artificial decision-making and human trust in autonomous vehicles. In *3rd International Conference on Computing, Communications, and Cyber-Security (IC4S'21)*. Springer, 123–136.

[92] AL-Essa Malik, Giuseppina Andresini, Annalisa Appice, and Donato Malerba. 2022. An XAI-based adversarial training approach for cyber-threat detection. In *IEEE International Conference on Dependable, Autonomic and Secure Computing, International Conference on Pervasive Intelligence and Computing, International Conference on Cloud and Big Data Computing, International Conference on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech'22)*. IEEE, 1–8.

[93] Luca Malinverno, Vesna Barros, Francesco Ghisoni, Giovanni Visonà, Roman Kern, Philip J. Nickel, Barbara Elvira Ventura, Ilija Šimić, Sarah Stryeck, Francesca Manni, et al. 2023. A historical perspective of biomedical explainable AI research. *Patterns* 4, 9 (2023), 1–9.

[94] Harsh Mankodiya, Dhairya Jadav, Rajesh Gupta, Sudeep Tanwar, Wei-Chiang Hong, and Ravi Sharma. 2022. OD-XAI: Explainable AI-based semantic object detection for autonomous vehicles. *Appl. Sci.* 12, 11 (2022), 5310.

[95] Harsh Mankodiya, Mohammad S. Obaidat, Rajesh Gupta, and Sudeep Tanwar. 2021. XAI-AV: Explainable artificial intelligence for trust management in autonomous vehicles. In *International Conference on Communications, Computing, Cybersecurity, and Informatics (CCCI'21)*. IEEE, 1–5.

[96] Elvis Melo, Ivanovitch Silva, Daniel G. Costa, Carlos M. D. Viegas, and Thiago M. Barros. 2022. On the use of explainable artificial intelligence to evaluate school dropout. *Educ. Sci.* 12, 12 (2022), 845.

[97] Pablo Messina, Pablo Pino, Denis Parra, Alvaro Soto, Cecilia Besa, Sergio Uribe, Marcelo Andia, Cristian Tejos, Claudia Prieto, and Daniel Capurro. 2022. A survey on deep learning and explainability for automatic report generation from medical images. *ACM Comput. Surv.* 54, 10s (2022), 1–40.

[98] Jan Hendrik Metzen, Tim Genewein, Volker Fischer, and Bastian Bischoff. 2017. On detecting adversarial perturbations. *arXiv preprint arXiv:1702.04267* (2017).

[99] Eleanor Ruth Mill, Wolfgang Garn, Nicholas F. Ryman-Tubb, and Christopher Turner. 2023. Opportunities in real time fraud detection: An explainable artificial intelligence (XAI) research agenda. *Int. J. Advan. Comput. Sci. Applic.* 14, 5 (2023), 1172–1186.

[100] Milad Moradi and Matthias Samwald. 2021. Post-hoc explanation of black-box classifiers using confident itemsets. *Expert Syst. Applic.* 165 (2021), 113941.

[101] Nour Moustafa, Nickolaos Koroniotis, Marwa Keshk, Albert Y. Zomaya, and Zahir Tari. 2023. Explainable intrusion detection for cyber defences in the internet of things: Opportunities and solutions. *IEEE Communications Surveys Tutorials* 25, 3 (2023), 1775–1807. https://doi.org/10.1109/COMST.2023.3280465

[102] Satya M. Muddamsetty, Mohammad N. S. Jahromi, Andreea E. Ciontos, Laura M. Fenoy, and Thomas B. Moeslund. 2022. Visual explanation of black-box model: Similarity difference and uniqueness (SIDU) method. *Pattern Recog.* 127 (2022), 108604.

[103] Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice van Keulen, and Christin Seifert. 2023. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI. *Comput. Surv.* 55, 13s (2023), 1–42.

[104] Sajid Nazir, Diane M. Dickson, and Muhammad Usman Akram. 2023. Survey of explainable artificial intelligence techniques for biomedical imaging with deep neural networks. *Computers in Biology and Medicine* 156 (2023), 106668. https://doi.org/10.1016/j.compbiomed.2023.106668

[105] Christina Neumayer and Miguel Sicart. 2023. Probably not a game: Playing with the AI in the ritual of taking pictures on the mobile phone. *New Media Societ.* 25, 4 (2023), 685–701.

[106] Boda Venkata Nikith, Masabattula Teja Nikhil, Mutyala Sai Sri Siddhartha, and K. Murali. 2022. LIME explainability on flower classification. In *6th International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS'22)*. IEEE, 1–4.

[107] Sonia Farhana Nimmy, Omar K. Hussain, Ripon K. Chakrabortty, Farookh Khadeer Hussain, and Morteza Saberi. 2023. An optimized belief-rule-based (BRB) approach to ensure the trustworthiness of interpreted time-series decisions. *Knowledge-based Syst.* 271 (2023), 110552.

[108] Worood Esam Noori and A. S. Albahri. 2023. Towards trustworthy myopia detection: Integration methodology of deep learning approach, XAI visualization, and user interface system. *Applied Data Science and Analysis* 2023 (Feb. 2023), 1–15. https://doi.org/10.58496/ADSA/2023/001

[109] Mohsen Abbaspour Onari, Isel Grau, Marco S. Nobile, and Yingqian Zhang. 2023. Trustworthy artificial intelligence in medical applications: A mini survey. In *IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB'23)*. IEEE, 1–8.

[110] Martin Pirker, Patrick Kochberger, and Stefan Schwandter. 2018. Behavioural comparison of systems for anomaly detection. In *13th International Conference on Availability, Reliability and Security*. 1–10.

[111] Tanguy Pommellet and Freddy Lecue. 2019. Feeding machine learning with knowledge graphs for explainable object detection. In *Proceedings of the ISWC 2019 Satellite Tracks (Posters Demonstrations, Industry, and Outrageous Ideas) Co-Located with 18th International Semantic Web Conference (ISWC 2019)*, Vol. 2456. CEUR-WS, 277–280.

[112] Thomas Ponn, Thomas Kröger, and Frank Diermeyer. 2020. Performance Analysis of Camera-based Object Detection for Automated Vehicles. *Sensors (Basel, Switzerland)* 20, 13 (2020), E3699–E3699.

[113] Ashwin Rachha and Mohammed Seyam. 2023. Explainable AI in education: Current trends, challenges, and opportunities. In *IEEE SoutheastCon'23*. 232–239.

[114] Masroor Rahman, Reshad Karim Navid, Md Muballigh Hossain Bhuyain, Farnaz Fawad Hasan, and Naima Ahmed Nup. 2023. *Exploring the Intersection of Machine Learning and Explainable Artificial Intelligence: An Analysis and Validation of ML Models through XAI for Intrusion Detection*. Ph. D. Dissertation. Brac University.

[115] Samuel Reeder, Joshua Jensen, and Robert Ball. 2023. Evaluating explainable AI (XAI) in terms of user gender and educational background. In *International Conference on Human–computer Interaction*. Springer, 286–304.

[116] Jana-Rebecca Rehse, Nijat Mehdiyev, and Peter Fettke. 2019. Towards explainable process predictions for Industry 4.0 in the DFKI-Smart-Lego-Factory. *KI-Künstliche Intelligenz* 33 (2019), 181–187.

[117] Felipe Riquelme, Alfredo De Goyeneche, Yundong Zhang, Juan Carlos Niebles, and Alvaro Soto. 2020. Explaining VQA predictions using visual grounding and a knowledge base. *Image Vis. Comput.* 101 (2020), 103968.

[118] Gaith Rjoub, Jamal Bentahar, and Omar Abdel Wahab. 2022. Explainable AI-based federated deep reinforcement learning for trusted autonomous driving. In *International Wireless Communications and Mobile Computing (IWCMC'22)*. IEEE, 318–323.

[119] Gaith Rjoub, Jamal Bentahar, and Omar Abdel Wahab. 2023. Explainable trust-aware selection of autonomous vehicles using LIME for one-shot federated learning. In *International Wireless Communications and Mobile Computing (IWCMC'23)*. IEEE, 524–529.

[120] Gaith Rjoub, Jamal Bentahar, Omar Abdel Wahab, Rabeb Mizouni, Alyssa Song, Robin Cohen, Hadi Otrok, and Azzam Mourad. 2023. A survey on explainable artificial intelligence for cybersecurity. *IEEE Transactions on Network and Service Management* 20, 4 (2023), 5115–5140. https://doi.org/10.1109/TNSM.2023.3282740

[121] Tomasz Rutkowski, Krystian Łapa, and Radosław Nielek. 2019. On explainable fuzzy recommenders and their performance evaluation. *Int. J. Appl. Math. Comput. Sci.* 29, 3 (2019), 595–610.

[122] Waddah Saeed and Christian Omlin. 2023. Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities. *Knowl.-based Syst.* 263 (2023), 110273.

[123] Kseniya Sahatova and Ksenia Balabaeva. 2022. An overview and comparison of XAI methods for object detection in computer tomography. *Proced. Comput. Sci.* 212 (2022), 209–219.

[124] Chaithanya Vamshi Sai, Debashish Das, Nouh Elmitwally, Ogerta Elezaj, and Md Baharul Islam. 2023. Explainable Ai-Driven Financial Transaction Fraud Detection Using Machine Learning and Deep Neural Networks. Available at SSRN (2023), pages n.a. https://doi.org/10.2139/ssrn.4439980

[125] Rabia Saleem, Bo Yuan, Fatih Kurugollu, Ashiq Anjum, and Lu Liu. 2022. Explaining deep neural networks: A survey on the global interpretation methods. *Neurocomputing* 513, C (2022), 165–180. https://doi.org/10.1016/j.neucom.2022.09.129

[126] Ahmed Salih, Zahra Raisi-Estabragh, Ilaria Boscolo Galazzo, Petia Radeva, Steffen E. Petersen, Gloria Menegaz, and Karim Lekadir. 2023. Commentary on explainable artificial intelligence methods: SHAP and LIME. *arXiv preprint arXiv:2305.02012* (2023).

[127] Nikhil Sarathy, Mohammad Alsawwaf, and Zenon Chaczko. 2020. Investigation of an innovative approach for identifying human face-profile using explainable artificial intelligence. In *IEEE 18th International Symposium on Intelligent Systems and Informatics (SISY'20)*. IEEE, 155–160.

[128] Vani Suthamathi Saravanarajan, Rung-Ching Chen, Cheng-Hsiung Hsieh, and Long-Sheng Chen. 2023. Improving semantic segmentation under hazy weather for autonomous vehicles using explainable artificial intelligence and adaptive dehazing approach. *IEEE Access* 11 (2023), 38194–38207.

[129] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision (ICCV'17)*. 618–626.

[130] Oscar Serradilla, Ekhi Zugasti, Carlos Cernuda, Andoitz Aranburu, Julian Ramirez de Okariz, and Urko Zurutuza. 2020. Interpreting remaining useful life estimations combining explainable artificial intelligence and domain knowledge in industrial machinery. In *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE'20)*. IEEE, 1–8.

[131] A. K. Seth. 2023. Artificial intelligence in health care. *Journal of Integrated Health Sciences* 11, 1 (2023), 1–2. https://doi.org/10.4103/2347-6486.386493

[132] Jash Minesh Shah, N. A. Natraj, Giri G. Hallur, and Avinash Aslekar. 2023. Artificial intelligence (AI) in the automotive industry and the use of exoskeletons in the manufacturing sector of the automotive industry. In *International Conference on Sustainable Computing and Data Communication Systems (ICSCDS'23)*. IEEE, 428–432.

[133] Vera Shalaeva, Sami Alkhoury, Julien Marinescu, Cécile Amblard, and Gilles Bisson. 2018. Multi-operator decision trees for explainable time-series classification. In *17th International Conference on Information Processing and Management of Uncertainty in Knowledge-based Systems. Theory and Foundations (IPMU'18)*. Springer, 86–99.

[134] Ruey-Kai Sheu and Mayuresh Sunil Pardeshi. 2022. A survey on medical explainable AI (XAI): Recent progress, explainability approach, human interaction and scoring system. *Sensors* 22, 20 (2022), 8068.

[135] Eduardo Soares, Plamen Angelov, and Xiaowei Gu. 2020. Autonomous learning multiple-model zero-order classifier for heart sound classification. *Appl. Soft Comput.* 94 (2020), 106449.

[136] Thilo Spinner, Udo Schlegel, Hanna Schäfer, and Mennatallah El-Assady. 2019. explAIner: A visual analytics framework for interactive and explainable machine learning. *IEEE Trans. Visualiz. Comput. Graph.* 26, 1 (2019), 1064–1074.

[137] Dominik Stoffels, Susanne Grabl, Thomas Fischer, and Marina Fiedler. 2023. How explainable AI methods support data-driven decision making. In *Wirtschaftsinformatik 2023 Proceedings*, Vol. 31. https://aisel.aisnet.org/wi2023/31

[138] Kyung Ho Sun, Hyunsuk Huh, Bayu Adhi Tama, Soo Young Lee, Joon Ha Jung, and Seungchul Lee. 2020. Vision-based fault diagnostics using explainable deep learning with class activation maps. *IEEE Access* 8 (2020), 129169–129179.

[139] Hatma Suryotrisongko, Yasuo Musashi, Akio Tsuneda, and Kenichi Sugitani. 2022. Robust botnet DGA detection: Blending XAI and OSINT for cyber threat intelligence sharing. *IEEE Access* 10 (2022), 34613–34624.

[140] Paolo Tamagnini, Josua Krause, Aritra Dasgupta, and Enrico Bertini. 2017. Interpreting black-box classifiers using instance-level visual explanations. In *2nd Workshop on Human-in-the-loop Data Analytics*. 1–6.

[141] Mohamed Torky, Ibrahim Gad, and Aboul Ella Hassanien. 2023. Explainable AI model for recognizing financial crisis roots based on pigeon optimization and gradient boosting model. *Int. J. Comput. Intell. Syst.* 16, 1 (2023), 50.

[142] Naeem Ullah, Ivanoe De Falco, and Giovanna Sannino. 2023. A novel deep learning approach for colon and lung cancer classification using histopathological images. In *IEEE 19th International Conference on e-Science (e-Science'23)*. 1–10. DOI: https://doi.org/10.1109/e-Science58273.2023.10254909

[143] Naeem Ullah, Muhammad Hassan, Javed Ali Khan, Muhammad Shahid Anwar, and Khursheed Aurangzeb. 2023. Enhancing explainability in brain tumor detection: A novel DeepEBTDNet model with LIME on MRI images. *International Journal of Imaging Systems and Technology* 34, 1 (2023), e23012. https://doi.org/10.1002/ima.23012

[144] Naeem Ullah, Ali Javed, Ali Alhazmi, Syed M. Hasnain, Ali Tahir, and Rehan Ashraf. 2023. TumorDetNet: A unified deep learning model for brain tumor detection and classification. *PLoS One* 18, 9 (2023), e0291200.

[145] Tahir Ullah, Javed Ali Khan, Nek Dil Khan, Affan Yasin, and Hasna Arshad. 2023. Exploring and mining rationale information for low-rating software applications. *Soft Computing* n.a. (2023), 1–26.

[146] Bas H. M. Van der Velden, Hugo J. Kuijf, Kenneth G. A. Gilhuijs, and Max A. Viergever. 2022. Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. *Med. Image Anal.* 79 (2022), 102470.

[147] Jasper van der Waa, Tjeerd Schoonderwoerd, Jurriaan van Diggelen, and Mark Neerincx. 2020. Interpretable confidence measures for decision support systems. *Int. J. Hum.-comput. Stud.* 144 (2020), 102493.

[148] Johanna Vielhaben, Sebastian Lapuschkin, Grégoire Montavon, and Wojciech Samek. 2023. Explainable AI for time series via virtual inspection layers. *arXiv preprint arXiv:2303.06365* (2023).

[149] Giulia Vilone and Luca Longo. 2020. Explainable artificial intelligence: A systematic review. *arXiv preprint arXiv:2006.00093* (2020).

[150] Giulia Vilone and Luca Longo. 2021. Classification of explainable artificial intelligence methods through their output formats. *Mach. Learn. Knowl. Extract.* 3, 3 (2021), 615–661.

[151] Alexis Votto and Charles Zhechao Liu. 2023. Transparent artificial intelligence and human resource management: A systematic literature review. In *Hawaii International Conference on System Sciences 2023 (HICSS-56)*. https://aisel.aisnet.org/hicss-56/da/xai/2

[152] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y Lim. 2019. Designing theory-driven user-centric explainable AI. In *CHI Conference on Human Factors in Computing Systems*. 1–15.

[153] Qianwen Wang, Kexin Huang, Payal Chandak, Marinka Zitnik, and Nils Gehlenborg. 2022. Extending the nested model for user-centric XAI: A design study on GNN-based drug repurposing. *IEEE Trans. Visualiz. Comput. Graph.* 29, 1 (2022), 1266–1276.

[154] Patrick Weber, K. Valerie Carl, and Oliver Hinz. 2024. Applications of explainable artificial intelligence in finance—a systematic review of finance, information systems, and computer science literature. *Management Review Quarterly* 74, 2 (2024), 867–907.

[155] Rosina O. Weber, Adam J. Johs, Jianfei Li, and Kent Huang. 2018. Investigating textual case-based XAI. In *26th International Conference on Case-based Reasoning Research and Development (ICCBR'18)*. Springer, 431–447.

[156] Bingzhe Wu, Jintang Li, Junchi Yu, Yatao Bian, Hengtong Zhang, Chaochao Chen, Chengbin Hou, Guoji Fu, Liang Chen, Tingyang Xu et al. 2022. A survey of trustworthy graph learning: Reliability, explainability, and privacy protection. *arXiv preprint arXiv:2205.10014* (2022).

[157] Haiyang Xia, Qian Li, Zhichao Wang, and Gang Li. 2023. Toward explainable recommendation via counterfactual reasoning. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 3–15.

[158] Guang Yang, Qinghao Ye, and Jun Xia. 2022. Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond. *Inf. Fusion* 77 (2022), 29–52.

[159] Affan Yasin, Rubia Fatima, Lin Liu, Javed Ali Khan, Raian Ali, and Jianmin Wang. 2022. On the utilization of non-quality assessed literature in software engineering research. *Journal of Software: Evolution and Process* 34, 7 (2022), e2464. https://doi.org/10.1002/smr.2464

[160] Wei Jie Yeo, Wihan van der Heever, Rui Mao, Erik Cambria, Ranjan Satapathy, and Gianmarco Mengaldo. 2023. A comprehensive review on financial explainable AI. *arXiv preprint arXiv:2309.11960* (2023).

[161] Hao Yuan, Haiyang Yu, Shurui Gui, and Shuiwang Ji. 2022. Explainability in graph neural networks: A taxonomic survey. *IEEE Trans. Pattern Anal. Mach. Intell.* 45, 5 (2022), 5782–5799.

[162] Jun Yuan, Kaustav Bhattacharjee, Akm Zahirul Islam, and Aritra Dasgupta. 2024. TRIVEA: Transparent ranking interpretation using visual explanation of black-box algorithmic rankers. *The Visual Computer* 40, 5 (2024), 3615–3631.

[163] S. G. Zeldam. 2018. *Automated Failure Diagnosis in Aviation Maintenance Using Explainable Artificial Intelligence (XAI).* Master's thesis. University of Twente.

[164] Huijing Zhan, Ling Li, Shaohua Li, Weide Liu, Manas Gupta, and Alex C. Kot. 2023. Towards explainable recommendation via BERT-guided explanation generator. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'23)*. IEEE, 1–5.

[165] Te Zhang, Christian Wagner, and Jonathan M. Garibaldi. 2022. Counterfactual rule generation for fuzzy rule-based classification systems. In *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE'22)*. IEEE, 1–8.

[166] Yiming Zhang, Ying Weng, and Jonathan Lund. 2022. Applications of explainable artificial intelligence in diagnosis and surgery. *Diagnostics* 12, 2 (2022), 237.

[167] Qiaoting Zhong, Xiuyi Fan, Xudong Luo, and Francesca Toni. 2019. An explainable multi-attribute decision model based on argumentation. *Expert Syst. Applic.* 117 (2019), 42–61.

[168] Ying Zhou, Haoran Li, Zhi Xiao, and Jing Qiu. 2023. A user-centered explainable artificial intelligence approach for financial fraud detection. *Fin. Res. Lett.* 58 (2023), 104309.

Received 24 January 2024; revised 7 October 2024; accepted 18 November 2024