Information Fusion 96 (2023) 156–191

Contents lists available at ScienceDirect

Information Fusion

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

A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion

A.S. Albahri [a,*], Ali M. Duhaim [b], Mohammed A. Fadhel [c], Alhamzah Alnoor [d], Noor S. Baqer [e], Laith Alzubaidi [f,g,*], O.S. Albahri [h,i], A.H. Alamoodi [j], Jinshuai Bai [f,g], Asma Salhi [k], Jose Santamaría [l], Chun Ouyang [m], Ashish Gupta [f,g], Yuantong Gu [f,g], Muhammet Deveci [n,o]

[a] Iraqi Commission for Computers and Informatics (ICCI), Baghdad, Iraq
[b] Ministry of Education, ThiQar, Iraq
[c] College of Computer Science and Information Technology, University of Sumer, Thi Qar , Iraq
[d] Southern Technical University, Basrah, Iraq
[e] Ministry of Education, Baghdad, Iraq
[f] School of Mechanical, Medical, and Process Engineering, Queensland University of Technology, Brisbane, QLD 4000, Australia
[g] ARC Industrial Transformation Training Centre – Joint Biomechanics, Queensland University of Technology, Brisbane, QLD 4000, Australia
[h] Computer Techniques Engineering Department, Mazaya University College, Nasiriyah, Iraq
[i] Department of Computer Science and Information Technology, La Trobe University, Melbourne, VIC, Australia
[j] Faculty of Computing and Meta-Technology (FKMT), Universiti Pendidikan Sultan Idris (UPSI), Perak, Malaysia
[k] Akunah Company for Medical Technology, Brisbane, QLD 4120, Australia
[l] Department of Computer Science, University of Jaén, Jaén 23071, Spain
[m] School of Psychology and Counselling, Queensland University of Technology, Brisbane, QLD 4000, Australia
[n] The Bartlett School of Sustainable Construction, University College London, 1-19 Torrington Place, London WC1E 7HB, UK
[o] Department of Industrial Engineering, Turkish Naval Academy, National Defence University, 34940, Tuzla, Istanbul, Turkey

ARTICLE INFO

Keywords:
Trustworthiness
Explainability
Artificial intelligence
Healthcare
Information fusion

ABSTRACT

In the last few years, the trend in health care of embracing artificial intelligence (AI) has dramatically changed the medical landscape. Medical centres have adopted AI applications to increase the accuracy of disease diagnosis and mitigate health risks. AI applications have changed rules and policies related to healthcare practice and work ethics. However, building trustworthy and explainable AI (XAI) in healthcare systems is still in its early stages. Specifically, the European Union has stated that AI must be human-centred and trustworthy, whereas in the healthcare sector, low methodological quality and high bias risk have become major concerns. This study endeavours to offer a systematic review of the trustworthiness and explainability of AI applications in healthcare, incorporating the assessment of quality, bias risk, and data fusion to supplement previous studies and provide more accurate and definitive findings. Likewise, 64 recent contributions on the trustworthiness of AI in healthcare from multiple databases (i.e., ScienceDirect, Scopus, Web of Science, and IEEE Xplore) were identified using a rigorous literature search method and selection criteria. The considered papers were categorised into a coherent and systematic classification including seven categories: explainable robotics, prediction, decision support, blockchain, transparency, digital health, and review. In this paper, we have presented a systematic and comprehensive analysis of earlier studies and opened the door to potential future studies by discussing in depth the challenges, motivations, and recommendations. In this study a systematic science mapping analysis in order to reorganise and summarise the results of earlier studies to address the issues of trustworthiness and objectivity was also performed. Moreover, this work has provided decisive evidence for the trustworthiness of AI in health care by presenting eight current state-of-the-art critical analyses regarding those more relevant research gaps. In addition, to the best of our knowledge, this study is the first to investigate the feasibility of utilising trustworthy and XAI applications in healthcare, by incorporating data fusion techniques and connecting various important pieces of information from available healthcare datasets and AI algorithms. The analysis of the revised contributions revealed crucial implications for academics and practitioners, and then potential methodological aspects

* Corresponding authors.
E-mail addresses: ahmed.bahri1978@gmail.com, ahmed.bahri1978@iips.icci.edu.iq (A.S. Albahri), l.alzubaidi@qut.edu.au (L. Alzubaidi), Osamahsh89@gmail.com (O.S. Albahri), alamoodi.abdullah91@gmail.com (A.H. Alamoodi), yuantong.gu@qut.edu.au (Y. Gu), muhammetdeveci@gmail.com (M. Deveci).

https://doi.org/10.1016/j.inffus.2023.03.008
Received 18 January 2023; Received in revised form 6 March 2023; Accepted 10 March 2023
Available online 15 March 2023

to enhance the trustworthiness of AI applications in the medical sector were reviewed. Successively, the theoretical concept and current use of 17 XAI methods in health care were addressed. Finally, several objectives and guidelines were provided to policymakers to establish electronic health-care systems focused on achieving relevant features such as legitimacy, morality, and robustness. Several types of information fusion in healthcare were focused on in this study, including data, feature, image, decision, multimodal, hybrid, and temporal.

1. Introduction

One of the most promising fields interested in the adoption of artificial intelligence (AI) applications is the medical sector. Many successful applications based on clinical decision support systems (CDSSs) have been developed in the last few years. Consequently, AI applications addressing the medical sector have gradually changed the landscape of the field [1]. Many doctors and computer scientists are working together in order to improve medical services and deploy automated systems that are capable of increasing the accuracy of disease diagnosis and prescribing the necessary treatments [2]. In particular, AI applications targeting the medical sector have helped support clinicians in developing diagnostic assumptions, providing explanations for clinical reasoning, and selecting appropriate treatments [3]. These kinds of applications have achieved considerable success in identifying and diagnosing health risks in clinical practices in a number of hospitals worldwide.

Furthermore, the U.S. Food and Drug Administration (FDA) has approved the integration of AI technologies into the medical sector to improve health services and mitigate health risks [1]. Recently, health centres and hospitals have announced the adoption of new technologies and breakthroughs at an accelerated rate to build smart entities that are unparalleled in health care [4]. This supports the FDA and European Union's (EU) efforts to develop smart apps for disease diagnosis and treatment that consider ethical issues such as privacy, transparency, safety, and accountability. In this context, medical breakthroughs that represent the adoption of applications of AI in healthcare are useful and vital, and it has expanded to areas such as translational medical research, clinical practice, basic biomedical research, and medical image diagnostics systems [5]. In addition, recent AI research in healthcare has leveraged deep learning (DL) and machine learning (ML) approaches to identify data patterns and explain complex interactions [6]. Moreover, DL and ML methods have shaped the framework of AI in health care [7]. According to the International Data Corporation, spending on AI applications will rise to 97.9 billion dollars (USD) in 2023 [8]. However, systems and applications of AI targeting the healthcare sector are fragile since adopting such applications has led to changes in decision rules and business ethics [1].

The recent renaissance in AI faces the critical challenge of trustworthiness. Several policymakers and academics have been occupied with tackling the issue of trust, transparency, and ethics in AI-based healthcare applications [9]. The trustworthiness of AI applications in biomedicine and healthcare is decisive for adoption of AI. Trust in the applications of AI by practitioners and policymakers in the healthcare sector will increase the provision of superior medical services and care to patients [8]. By contrast, the lack of trustworthiness of AI applications is a critical barrier to deploying modern technology. There are many challenges due to increase in legal and ethical issues for this kind of applications, since clinical and medical decisions affect the well-being of people. Furthermore, the fragile trustworthiness of AI exacerbates decision-making problems for patients and clinicians and weakens accountability for errors [10]. Therefore, the definition of trustworthy AI applications is still unclear. Nevertheless, the literature emphasises three key concepts for achieving trustworthy AI: legitimacy, morality, and robustness. In particular, the legal aspect represents adherence to regulations and laws, the ethical aspect indicates a commitment to values and moral principles, and the robustness aspect refers to safety and security issues [8,11]. Furthermore, translating key concepts of trustworthy AI into practice is an open-ended challenge.

Subsequently, the number of scientific contributions to AI-based healthcare applications over the last 20 years has reached a peak [3]. In particular, the number of academic publications about trustworthy AI-based healthcare has increased in the last few years, and keeping pace with academic work in such a field is becoming a challenging task. Moreover, huge streams of research on trustworthy AI-based healthcare applications being published makes it even more difficult to keep up with the literature. This challenge hinders obtaining conclusive evidence from the earlier studies [12]. As an outcome, several attempts have been made, e.g., mini review, review, and bibliometric and content analysis, each one of them focused on the study of the ethics of AI in healthcare [13–16]. Specifically, both bibliometric and content analysis and mapping analysis of the ethics of AI in healthcare have been carried out in [14]. The main concern of such a method is that the bibliometric analysis uses a single database, e.g., Scopus. Furthermore, bibliometric analysis uses research papers that may not be sufficiently specialised [12]. In line with the previous two, in [17] the authors provided a study that can be considered as a mini review addressing explainable AI (XAI) in medicine and digital healthcare. Two XAI methods have been suggested in such a contribution. To the best of the authors' knowledge, the literature has not provided conclusive evidence about the various case studies of trustworthy AI in healthcare in one taxonomy. A review of the trustworthiness and transparency of AI in the medical sector was conducted in [13,18]. However, the objectivity and reliability of such a review, mini review, and bibliometric analysis approaches is still a challenge [12]. Both academics and scientists adopt different methods that capture knowledge and reorganise the outcomes of previous studies. In the modest opinion of this study's authors, following a more systematic review approach can both summarise results in a more suitable manner and reorganise findings of previous studies with high transparency and reliability [19]. As stated above, further research on the trustworthy AI research pipeline in healthcare is vitally important.

In this study, it is aimed to draw a comprehensive and coherent map of previous works in the medical sector regarding the applications of trustworthy AI in the literature by adopting a systematic review approach. The assessment of quality, bias risk, and data fusion in trustworthy AI-based healthcare systems is crucial to ensuring that healthcare policymakers make sound and reliable decisions. Assessment of quality refers to the evaluation of the performance and accuracy of an AI system. This can include methods such as cross-validation and performance comparison with other models. Bias risk refers to the potential of an AI system to produce unfair or discriminatory results, often due to imbalances in the training data. To mitigate this risk, techniques such as algorithmic fairness and data balancing can be used. Data fusion refers to the process of combining multiple data sources to improve the performance of an AI system. This can involve methods such as data preprocessing, feature selection, and feature engineering. Information fusion is important for achieving robust, explainable, and trustworthy medical AI because it allows for the integration of multiple sources of information and data. This can lead to more accurate and reliable AI models, as well as increased transparency and explainability of the decision-making process [20]. Additionally, the integration of multiple data sources can also help to mitigate potential biases in the data, leading to more fair and trustworthy AI models. Overall, information fusion is a key enabler for achieving more robust and reliable AI systems in the field of medicine [21].

The ongoing research question of this study is in line with what are

the outcomes, motivations, challenges, recommendations, and gaps of trustworthy AI in healthcare in the context of quality assessment, bias risk, and data fusion? Consequently, the authors' proposal for a systematic review of the literature seeks to provide information on trustworthy AI in healthcare and to help scholars to identify current gaps and options. Moreover, this study also makes a cutting-edge contribution by constructing a comprehensive map of trustworthy AI in the healthcare sector to provide a coherent taxonomy system. As a result, reviews of previous studies are enhanced by using multiple databases to identify author highlights in response to trustworthy AI trends in healthcare. This contributes to a comprehensive overview of trustworthy AI in healthcare. However, the issue of bias and reliability was needed to be addressed in a more effective way. Therefore, bibliometric analysis was used to reorganise and summarise the results of the previous studies and the overall knowledge picture by providing a mapping analysis for the research stream of trustworthy AI applications in healthcare [12]. The goal of this study is to aid in highlighting trustworthy AI in healthcare and providing support to the medical sector. An analysis of characteristics and eight research gaps has also been provided to further enhance its contribution to the field.

The structure of this paper is organised as follows. The systematic literature review methodology is presented in Section 2. Section 3 provides five comprehensive science mapping analysis using a bibliometric approach to identify trends and gaps in the existing literature and further deepen the understanding of the topic. Section 4 presents the findings of the review highlighting seven critical categories. Section 5 discusses the enrichment of motivations, challenges, and recommendations in trustworthy AI in healthcare. In Section 6, eight characteristics and research gaps are analysed to identify areas for future research and development in the field of trustworthy AI in healthcare. Finally, Section 7 concludes this contribution.

2. Methodology

Recommended reporting items for a systematic review and meta-analyses approach were followed in this analysis section (Fig. 1) [22, 23]. Several bibliographic citation databases, covering a variety of medical, scientific, and social science journals from multidisciplinary fields, were used in the process. Specifically, four well-known digital databases were considered to search for the target papers, i.e., Science Direct (SD), Scopus, IEEE Xplore (IEEE), and Web of Science (WoS). SD provides reliable technological, scientific, and engineering references.

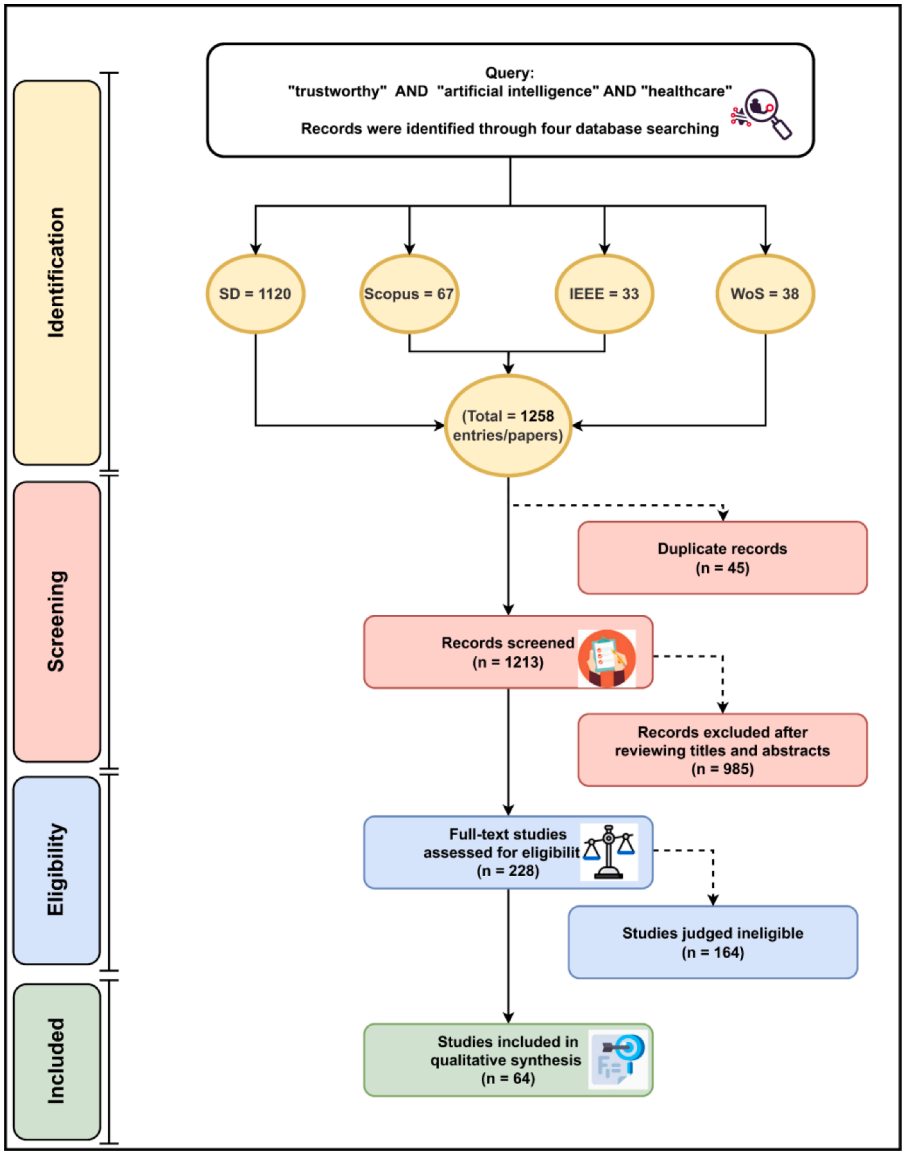

Fig. 1. An outline of the approach to identify, select, and include relevant contributions.

Scopus contains reliable resources in different fields, e.g., medicine, health, technology, science, and engineering. IEEE includes all the technical and scientific literature, providing abstracts and full texts of papers on electrical engineering, electronics, and computer science. The WoS database includes research papers from a wide range of disciplines, including science, technology, art, and social science, making it a cross-disciplinary resource. These databases offer valuable insights to researchers by providing comprehensive coverage of research across scientific and technological fields.

2.1. Search strategy

A comprehensive bibliographic search for English-language academic publications was performed in the four considered databases (SD, Scopus, IEEE, and WoS). This search included all scientific publications from the start of scientific production till February 2023. In particular, this search utilised a boolean query to conclude only one operator (AND) to link the keywords, i.e., trustworthy, AI, healthcare (see Fig. 1). These keywords were selected based on the recommendations of experts from both AI and medical fields. Prospects related to the application of trustworthy components in AI were also determined, e.g., trust, explainability, and auditability.

2.2. Inclusion and exclusion criteria

The criteria considered for the inclusion/selection of papers (see Fig. 1) are the most important part of this conducted systematic revision of the literature. And the following criteria were considered for this study:

- The papers had to be written in English and published in a journal or a conference proceeding.
- The papers had to consider one or more trustworthy components applied to integrate different AI techniques/methods for the healthcare domain.
- Each one of the components had to be significantly related to trustworthy AI.
- The papers that considered XAI and information fusion techniques for healthcare data were also included.

On the other hand, based on the following exclusion criteria, studies outside the scope of this study were excluded:

- Papers written in a language other than English
- Contributions that discussed trustworthy AI in sectors other than healthcare, e.g., industry, finance, and tourism.

2.3. Study selection

As done in earlier contributions [22,23], this work adopted the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) statement to conduct a systematic review of the literature. Such an approach includes several steps that start by excluding duplicate papers. The Mendeley software was used to scan the titles and abstracts of the contributions. All the authors participated in this process, and many unrelated works of literature were excluded. The corresponding author resolved differences and inconsistencies amongst the authors. The third process included reading the full text and excluding articles that did not meet the earlier mentioned inclusion criteria (see Section 2.2). The filtering process was carried out by three experts to judge its efficiency (see Fig. 1).

This study included the articles that met the criteria. A total of 1258 entries were retrieved from the initial search, with 1120 of these being from SD, 67 from Scopus, 33 from IEEE, and 38 from WoS. The search covered all articles published from the beginning of scientific production up to February 2023. After eliminating approximately 45 duplicates from the four databases, the total number of articles was reduced to 1213. Upon reviewing the titles and abstracts, 985 papers were excluded. After a comprehensive and critical evaluation of the remaining 228 contributions, 164 studies were deemed ineligible, and only 64 studies were deemed relevant and included in the final set of articles based on the inclusion criteria. The next subsection discusses how the analysis of collected articles can be traced using several bibliometric methods.

3. Comprehensive science mapping analysis

Increased contributions and applied research have made it more challenging to identify critical evidence from the earlier studies. As there is a vast stream of practical and theoretical contributions, keeping up with the literature has been a significant challenge. Several academics suggested the adoption of the PRISMA approach to reorganise the findings of the previous studies, summarise problems, and identify potential research gaps. By contrast, systematic reviews expand the knowledge base, enhance the research plan, and synthesise the literature results. However, systematic reviews still suffer from the issue of reliability and objectivity, as such approaches rely on the viewpoint of the authors to reorganise the findings of the previous studies. To increase transparency in summarising the results of the previous studies, several works have suggested methods for carrying out a more suitable comprehensive science mapping analysis based on R-tool and VOS-viewer [12]. The bibliometric approach provides conclusive results, explores research gaps, and concludes literature findings with high reliability and transparency. In addition, the tools presented here do not require high skills and are considered to be open-source. Therefore, this study adopted the bibliometric method described in detail in the following subsections.

3.1. Annual scientific production

Trustworthy AI in healthcare has evolved in the last decade. Specifically, the annual scientific production shown in Fig. 2 explains the production of earlier theoretical and practical studies on trustworthy AI.

Fig. 2 shows the annual scientific production for systematic review papers. It can be observed that the number of articles increased significantly in recent years, with only a few articles being published in the earlier years of 2012–2015. There was an increase in the number of articles published in 2017 and 2018, with one article each year. The number of articles continued to increase in 2019 and 2020, with a significant jump to 13 articles in 2020. This trend continued in 2021 and 2022, with 24 and 21 articles being published, respectively. It is still early in 2023, but only 2 articles have been published. The overall trend indicates growth in published trustworthy AI papers.

3.2. Three-field plot

A three-field plot is a visualisation tool used to display data with three parameters. In this particular case, the left field represents sources (SO), the middle field represents cited sources (CR_SO), and the right field represents keywords (DE). The plot is often used to analyse relationships between the three parameters (see Fig. 3).

The analysis, as recognised in the middle field (CR_SO) of Fig. 3, shows that the *Information Fusion, IEEE Access*, and *Nature* journals have been the most frequently cited by the sources (SO) located on the left side. Additionally, the *Information Fusion* journal is the most important journal among the sources (SO) that focus on the topic of trustworthy and explainable AI. Furthermore, as recognised in the right field (DE), across all keywords, the top-used keywords such as ‘deep learning’, ‘artificial intelligence’, ‘explainable AI’, ‘machine learning’, ‘information fusion’, and ‘trust’ are the most commonly matched by journals shown in the middle field (CR_SO).

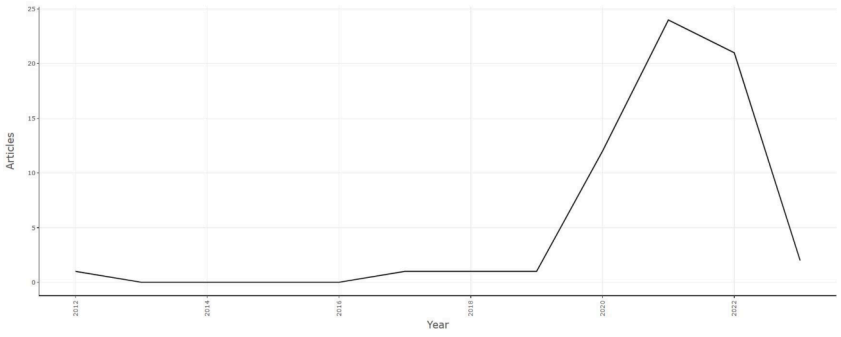

Fig. 2. Annual scientific production.

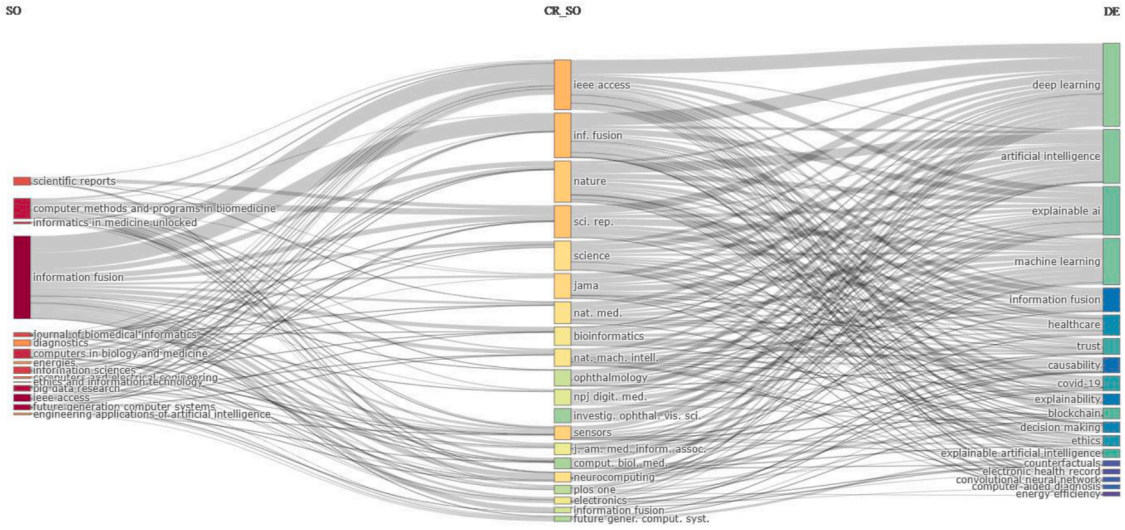

Fig. 3. Three-field plot: left (SO), middle (CR_SO), and right (DE).

3.3. Word cloud

The word cloud has helped us identify the most frequent and essential keywords in earlier studies. In particular, Fig. 4 presents those critical keywords gathered from the findings of earlier studies to summarise a big picture of them and reorganise knowledge.

Fig. 4 displays keywords of different sizes. The large size of the keywords indicates that they appear more frequently in the literature. Conversely, the small size of keywords indicates that their occur less frequently. Based on the frequency of terms listed in Fig. 4, some of the most common topics in the field of trustworthy AI include DL, AI, and ML, with DL having the highest frequency. Additionally, the figure shows that XAI and explainable artificial intelligence (XAI) are important topics in the field. Other related terms, such as ethics, privacy, and security, also have a relatively high frequency, indicating the importance of considering these aspects in developing and deploying AI systems. Fig. 4 also highlights some of the specific applications of AI in various domains, such as healthcare, decision making, and information fusion. Additionally, it mentions several techniques used in AI, such as convolutional neural networks (CNNs), natural language processing, and robotics. Overall, the word cloud of trustworthy AI papers suggests that the field is diverse and covers a wide range of topics, from technical

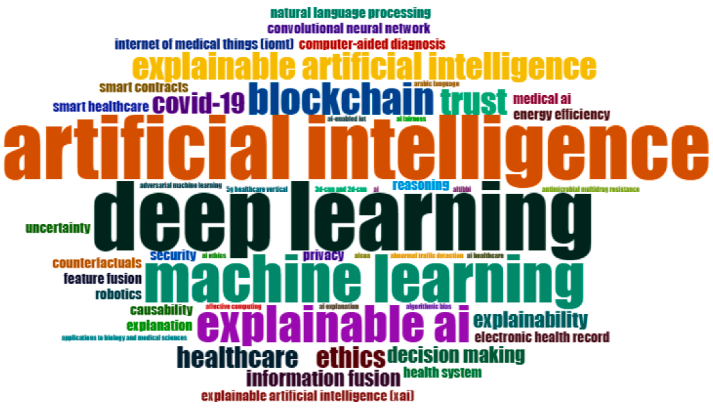

Fig. 4. Word cloud.

aspects of AI to its ethical, legal, and social implications.

3.4. Co-occurrence

Another tool used in bibliometric analysis is a co-occurrence network. Common words established by earlier studies and its analysis are based on a semantic network that provides critical clues to practitioners, policymakers, and academics about the conceptual structure of a field of specialisation. In particular, Fig. 5 provides information about a co-occurrence network based on the titles of trustworthy AI papers.

The network comprises nodes, the individual words in the titles, and the edges between the nodes indicate the frequency with which the words co-occur in the same title. Fig. 5 shows various nodes, along with the cluster they belong to and their closeness, which measures how well connected a node is to the other nodes in the network. It can be seen that the nodes are grouped into eight different clusters, and the words in each cluster are related to a particular theme or concept related to trustworthy AI. For example, cluster 1 includes words such as 'ai', 'medical', 'trustworthy', 'systems', 'blockchain', and 'challenges', which suggest that the cluster is related to the implementation of trustworthy AI systems in the medical field. Cluster 2 includes words such as 'explainable', 'artificial', 'intelligence', 'clinical', 'ethics', and 'causability', which suggest that the cluster is related to the ethical and explainable aspects of AI.

Similarly, other clusters are related to topics such as decision analysis, diagnosis and prediction, ML, and healthcare. The closeness of a node measures its centrality within the network, and the closeness of a node can be seen as an indicator of its importance within the network. The words with higher closeness values are more closely connected to other nodes in the network, indicating that they are more central to the topic of trustworthy AI. Overall, the figure provides a snapshot of the relationships between different concepts and words related to trustworthy AI, as seen through the titles of papers in the field. This information can be useful in understanding the current state of research in this area and in identifying areas where more research is needed.

3.5. Country collaboration map

Last, the country collaboration map shows the scientific cooperation network amongst universities, countries, and authors. Co-authorship increases the skills and experience of countries and researchers in developing a field of expertise. In addition, scientific cooperation is a vital for increasing the development of educational and industrial institutions. Fig. 6 presents a country collaboration map for trustworthy AI applications in healthcare around the world.

Specifically, this tool consists of three colours (see Fig. 6). The dark blue area indicates the countries with highest scientific production. The light blue region refers to countries with less significant scientific production. The grey area indicates countries lacking scientific production. Finally, the red line represents the scientific cooperation between countries. Fig. 6 shows that Australia and Canada seem to have a high level of collaboration with several countries, with a maximum of four collaborations. On the other hand, countries such as Austria, Belgium, Denmark, Finland, France, Italy, and Switzerland seem to have fewer collaborations but still collaborate considerably with other countries. Australia and Canada have a collaboration with Austria (four collaborations), China (four collaborations), and the United Kingdom (two collaborations). Iran has a collaboration with Australia (two collaborations), China (two collaborations), and Singapore (two collaborations). Asia seems to have a significant level of collaboration, with Japan collaborating with multiple countries, including Australia, Belgium, Canada, China, Germany, Italy, and Singapore. South Korea also has a collaboration with Egypt (three collaborations). Middle Eastern countries such as Iran, Saudi Arabia, and Egypt seem to collaborate with several countries, including Australia, Canada, China, India, Korea, and Pakistan. However, it is important to note that the contributions among countries shown in Fig. 6 are solely based on the number of publications found in our search and do not necessarily reflect the countries that are actually applying ethical and highly transparent AI technologies and tools in healthcare. Additionally, it is possible that some countries may not produce papers on this topic, but they may still be actively involved in the application of trustworthy AI in healthcare.

In conclusion, the data suggest that collaboration in trustworthy AI research is widespread across different countries and continents. However, specific countries and regions seem to have a higher level of collaboration, suggesting that these countries may have a strong research community in the field of trustworthy AI.

4. Findings and analysis: a taxonomy

AI in healthcare has been identified through the conducted method, and the final set of papers met both the considered inclusion and exclusion criteria (see Section 2.2). In addition, the 64 articles were divided into seven major categories based on the objective evidences across studies that met these criteria. Each category was then analysed, and attempts were made to find or create subcategories and then sections and subsections within each category based on various trustworthy AI components in healthcare contexts. The first major category contains $n = 64$ papers related to:

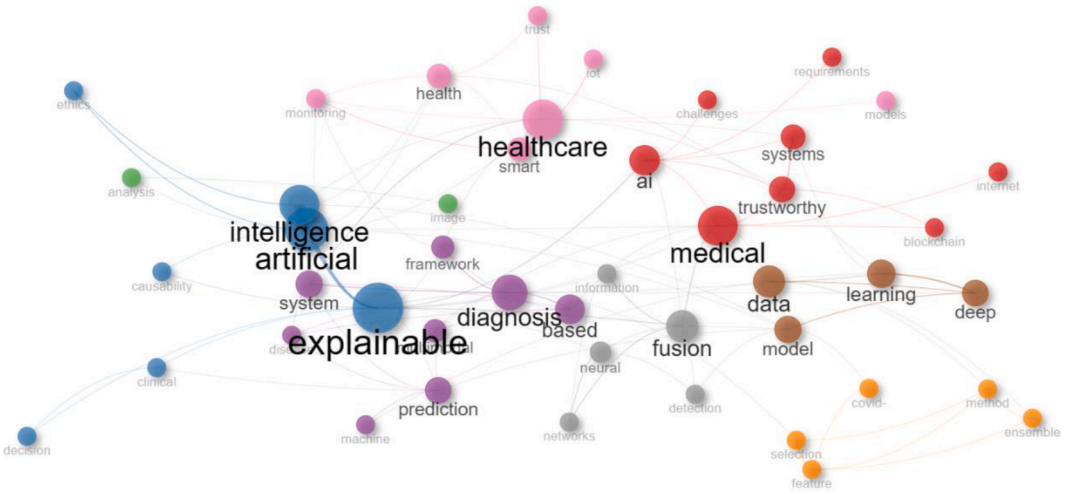

Fig. 5. Co-occurrence network.

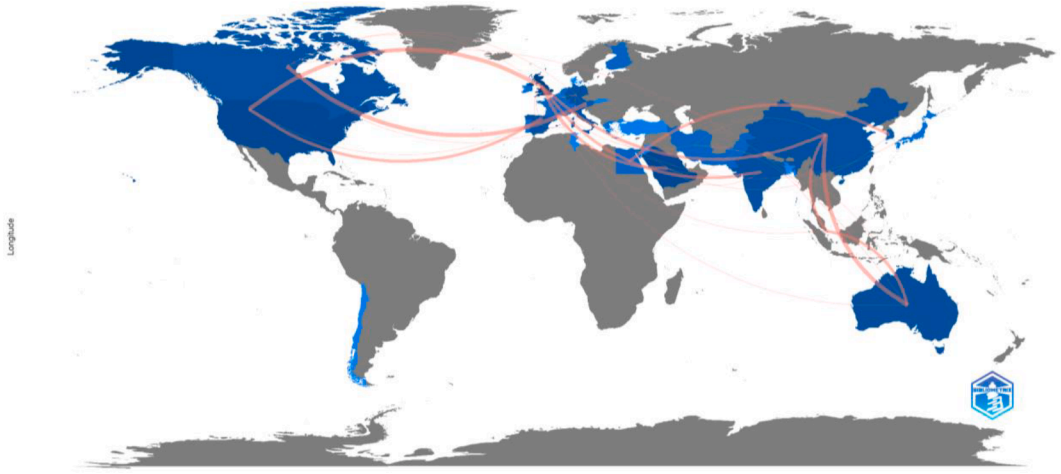

Fig. 6. Country collaboration map.

1) **Explainable Robotics:** including 2 of 64 papers
2) **Prediction:** 17 of 64 papers
3) **Decision Support:** 13 of 64 papers
4) **Blockchain:** 7 of 64 papers
5) **Transparency:** 6 of 64 papers
6) **Trustworthy Digital Health:** 11 of 64 papers
7) **Review:** 8 of 64 papers

Specifically, the following subsections will discuss the categories in a comprehensive manner to provide academics and practitioners with insights regarding trustworthy AI in healthcare (see Fig. 7). It is important to note that some of these categories will include additional subsections.

4.1. Explainable robotics

The explainable robotics category describes the context of human–robot interaction and the explainability of such interaction. Several efforts and attempts in the context of the explainable robotics category have been reported to increase trustworthy robotics in a healthcare context. Therefore, this category includes 2 of 64 articles. The literature discusses in detail human–robot interactions in the context of healthcare [24]. The objective is to create new computational models, methodologies, and algorithms that generate explanations, which in turn enable robots to function with different degrees of independence and communicate with people in a trustworthy and user-friendly way. In [25], the authors discussed the development of several AI-based models aimed at identifying cases of COVID-19. While AI shows promise, only a limited number of models have successfully integrated human-centred and machine-centred approaches to disease diagnosis. To address this issue, the study in [25] proposes a new method for XAI that employs graph analysis to visualise features and optimise blood test sample diagnosis for COVID-19, under the framework of human–computer interaction design.

4.2. Prediction

This category describes the issue of prediction in the medical sector

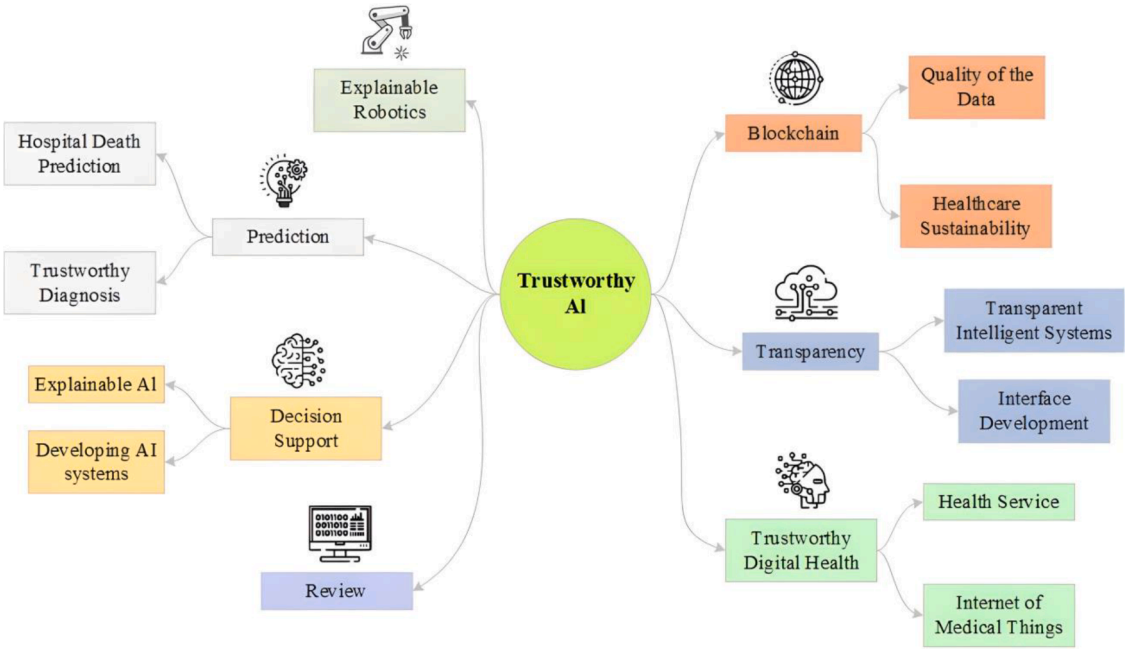

Fig. 7. Taxonomy of trustworthy AI in healthcare.

using AI mechanisms. Prediction is important in diagnosing disease and mortality in many patients. In this context, this category includes 17 of 64 articles.

4.2.1. Hospital death prediction

This subcategory includes 2 of 17 contributions. Many hospitals use invalid and unreliable predictive methods. In [26], the authors highlight that real and simulated data were used to predict in-hospital mortality. Moreover, a framework was developed in that work based on the local fit principle and the principle of density. In particular, the local fit principle emphasises that the trained model must have a good performance. By contrast, the intensity principle based on the evaluated example is similar to the training set. The estimation of deaths in health centres and hospitals is a vital issue. Several of the studies in the literature neglect the issue of cause-of-death privacy and is self-explainable for clinicians. Earlier studies have proposed a deep interpretation model that uses repeat units, multisource embedding, and focal loss techniques to predict mortality. Such a model increases the trustworthiness of AI in predicting mortality [27].

4.2.2. Trustworthy diagnosis

In this subcategory, 15 of 17 papers are included. This category discusses the diagnosis of a disease based on trustworthy AI. Computer-aided diagnostic systems lead to transparent and reliable decision-making in the context of disease prediction and detection [28]. The authors in [29] introduced measures to assess the trustworthiness and credibility of workers in diagnosing autism. Controls for neural patterns and behavioural trait scores were used for the videos of children with autism. In addition, diagnosing malignant neoplasms of skin lesions is a critical issue and requires a reliable method. Accordingly, XAI for dermatology was introduced in [30] in order to analyse medical images to provide visual maps and easy-to-understand interpretations to justify the predictions. Several attempts have been made in earlier studies to provide a reliable and trusted diagnostic method, especially for hospitals [31]. In [32], a system approach based on decision tree algorithms was introduced to diagnose diseases and provide medical advice to patients using the collaborative filtering method. Support clinical decisions, an interpretable learning approach, was adopted in [33] to predict multidrug-resistant antimicrobials that pose a threat to the health system worldwide. This suggests that AI has the potential to improve the accuracy of diagnosis for treatments and diseases. However, AI must have the ability to solve medical problems ethically. Trustworthy AI in healthcare requires the proposal and implementation of international and national regulations and rules that enhance the ethics of AI in the medical sector. To do so, in [34] the authors proposed several ethical rules and regulations that increase trustworthy AI for prediction in healthcare. In addition, the use of conceptual knowledge as a model of reality was proposed in the study of [20] to enhance the robustness, explainability, reduced bias, and improved learning from fewer data of ML models. Achieving these goals is believed to be attainable through the integration of three frontier research areas: complex networks and their inference, graph causal models and counterfactuals, and verification and explainability techniques. The full integration of these three fields is expected to bridge the gap between research and actual applications, thereby advancing trustworthy medical AI. The evaluation of binary classifiers and diagnostic tests was suggested to have been carried out using ‘deep ROC analysis’ in the study of [35]. A compromise between AUC and accuracy, sensitivity, specificity, positive predictive value, and F1 score was presented by this approach. The measures that were either too broad or too narrow were criticised. The definition of AUC as a balanced average accuracy relevant to individuals was provided by the study. The efficacy of the method was demonstrated through three case studies using a Python toolkit.

The study of [37] proposed a generic data-driven method based on integrating models and fusing data in deep ensemble learning for predicting COVID-19 outbreaks as a decision-support tool to diagnose and enhance surveillance, control, and management of epidemics. The validity of the proposed method in predicting the daily number of COVID-19-positive cases was confirmed through two case studies in Tunisia and China. Data were merged, and learners were performed from the China case study, and then the data were reused for the Tunisia case study to aid stakeholders in making and reviewing their epidemic response plans. An adversary-aware multimodal convolutional autoencoder (MCAE) model for the prediction of cancer susceptibility from multi-omics data, including copy number variations, miRNA expression, and gene expression, was presented in [38]. Multimodal feature representations from the multi-omics data are learned by the MCAE model through the use of different representational learning techniques, and patient cohorts are then classified into different cancer types on the multimodal embedding space, exhibiting similar characteristics in the end-to-end setting. To enhance robustness and provide consistent diagnoses, the stability of predictions with regard to small variations in the input is formulated as a property in the MCAE model.

A novel feature fusion model for the accurate detection of COVID-19 cases using CT scan and X-ray data was proposed in [39]. The uncertainty in the proposed feature fusion model was quantified using the effective Ensemble MC Dropout technique. The robustness to data contamination, also known as data noise, was demonstrated to be strong by the proposed feature fusion model. The results obtained from the model in terms of unknown data detection were very encouraging. A novel feature selection method based on rough set theory was proposed in [40] to reduce the dimensionality of large medical datasets. The aim was to support the decision-making process of medical experts in diagnosing diseases with less computational complexity, as evaluated on benchmark medical datasets. The use of feature selection was thought to contribute to the trustworthy process in ML by reducing the dimensionality of medical datasets and increasing classification accuracy, thus improving the reliability of medical diagnoses. In [41], COVID-19 artifacts in real-world chest X-rays were classified. A novel CNN model for chest X-ray recognition, based on Bayesian optimisation, was proposed. The proposed model was demonstrated to be more accurate and trustworthy in practice. A cloud-based framework for diagnosing, detecting, and monitoring Parkinson’s Disease (PD) for remote healthcare applications was developed in [42]. The system allows for easy detection of PD patients through the use of ML and the provision of voice samples through their phones for diagnosis.

4.3. Decision support

This category explains the role of AI in healthcare decision support and includes 13 of 64 papers explaining two subcategories: XAI and the development of AI systems.

4.3.1. Explainable AI

Trustworthy AI enables healthcare policymakers to make sound and reliable decisions. Therefore, there is a need for XAI methods to support decisions. This subcategory includes 9 of 13 articles. Trustworthy AI must enable biomedical experts to bear responsibility for their decision-making, necessitating the use of explicable methodologies. XAI techniques, such as layer-wise relevance propagation, can aid in highlighting and displaying the relevant parts of inputs and representations in a neural network that led to a specific outcome [43]. In [44], the authors presented an investigation about the use of XAI methods for in vitro diagnostic (IVD) devices to support the decisions of healthcare practitioners. Similar to how usability measures usage quality, causability measures the quality of explanations generated by XAI systems. XAI was used for diagnosing glaucoma through image analysis to assist trustworthy healthcare decisions [45]. This study endeavours to implement a hybrid image processing solution and XAI-supported DL to guarantee the dependability of diagnostic decisions. Furthermore, it contributes to the XAI literature by presenting a different XAI perspective on a specific condition. This study showcases that it is still feasible to develop

dependable AI systems. However, there are few studies on human-assisted XAI evaluations. As the application of XAI in the medical environment requires particular evaluations by physicians, this study adopts a rigorous evaluation phase that includes physician participation. This study contributes to the XAI literature by bridging human-side comments on the recent application of XAI in DL. Due to its success in medical applications, CNN is a renowned DL model. However, it still requires XAI components since it contains numerous parameters that have been optimised to demonstrate a suitable level of success for target issues. As the target data is an image, Class Activation Map (CAM), a human-readable interface, was used to develop a successful XAI solution. Specifically, this evaluation was conducted by combining CNN and image processing techniques [46]. The results of this study are deemed valuable not only for researchers but also for physicians and medical personnel. It is a wonderful motivation to recall that stronger DL models can still be supported by XAI to enhance their human–user interaction capabilities.

XAI methods have been employed to ensure reliable decision-making in the presence of uncertain information and to support clinical decision-making. To facilitate interpretability, non-ambiguous categorical traits are converted to numerical and interpretable traits [47]. This study emphasises the importance of trusting modified data to make trustworthy and explainable conclusions. The effectiveness of three AI models (a DL model based on artificial neural networks, a tree-based model using decision trees, and a rule-based model using Belief Rule-Based model) was compared with the performance of four commonly used data imputation techniques using two real-world datasets on early asthma symptoms and mortgage loans. The suggested method, along with I-MAKER and CMAKER, was applied to the data and transformed, and the results were analysed. An explainable ML-based CDSS that identified gestational diabetes mellitus (GDM) women who needed tailored prenatal treatments was developed in [48]. Five-fold cross-validation was used to optimise five ML techniques using PEARS maternal characteristics and blood biomarkers. Shapley additive explanations increased system trustworthiness and acceptability. Theoretical, antenatal GDM screening, and remote GDM risk assessment models had different performance measures. The explainable CDSS academic web server can help doctors identify at-risk patients for early pregnancy GDM prophylaxis. The impact of AI explanations and fairness on human-AI trust and perceived fairness in AI-based decision-making scenarios was examined in [21].

The examination was conducted through an online survey simulating AI-assisted health insurance and decision-making in medical treatment. It was found that the improvement of trust was brought about by the explanation of AI-assisted decision-making, while the effect of fairness on trust was observed only at low levels. An increase in fairness was observed with the addition of explanations while a decline was observed with low levels. A new data fusion scheme, based on type-2 fuzzy logic (T2FL) and Dempster–Shafer theory (DST), was proposed in [49] for accurate inference of Internet of Things (IoT)-based healthcare systems. Type-2 fuzzy logic effectively determines the patient data membership values, and the DST in the decision-making system effectively fuses and processes the evidence from these membership values. Through extensive computer simulations using heart disease and diabetes datasets, the proposed scheme was shown to outperform ontology and type-1 fuzzy logic schemes in terms of explainable decision accuracy.

4.3.2. Developing AI systems

The second subcategory examines system development to shed light on the reliability of medical decision support systems. This subcategory includes 5 of 13 articles. AI can assist physicians with the challenging process of risk stratification of patients for therapies, identifying those at the highest risk of decompensation and analysing several small outcomes to optimise overall patient outcomes. The next paradigm shift in medical education will involve the inclusion of physicians in model development and the training of physicians in this field. For instance, the sophistication of AI approaches varies considerably, which affects how easily physicians can comprehend and interpret outcomes. Decision trees are a common tool used by physicians; nevertheless, they are effectively connected to the initial tree structure and are, therefore, fairly static. Conversely, DL models, such as CNNs, are less clearly interpretable and may make establishing a causal relationship more challenging; as a result, the development of such models requires the active cooperation of physicians. Neural networks, which are routinely used to evaluate patient images and their interpretations, frequently require radiologists to curate acceptable imaging data for training. *A priori* interactions between AI developers and medically knowledgeable physicians are essential to determine the accuracy and interpretability thresholds required for each clinical situation [50].

A study developed a diagnostic model that enables clinicians to make accurate decisions and can provide an explanation for each decision [51]. In the first layer, the developed model achieves cross-validation accuracy of 93.95% and an F1 score of 93.94%. In comparison, the second layer obtains cross-validation accuracy of 87.08% and an F1 score of 87.09%. Due to the offered explanations, which are generally compatible with each other and with the literature of Alzheimer's disease (AD), the resulting system is not only accurate but also trustworthy, accountable, and medically useful. The suggested approach can contribute to a better clinical understanding of AD diagnosis and development by offering in-depth insights into the effect of numerous disease risk modalities. Therefore, vital aspects have been suggested to increase trustworthiness in AI systems for ophthalmology, such as reliability, accountability, accuracy, safety, and flexibility [52]. The evolution of decision support systems attempts to formulate development systems to improve the workflow of policymakers in clinical decision making. The optimal treatment and prognosis are determined according to clinical knowledge and examination results.

To ensure the development of appropriate procedures, transparency should be a fundamental aspect of CDSS [52]. CDSS is designed to aid physicians in making clinical decisions by providing recommendations for accurate diagnosis or optimal treatment based on patient-specific evidence, such as examination findings and expert-modelled or machine-learned clinical knowledge. The acceptance and trust in CDSS largely depend on the transparency in the generation of recommendations. Physicians need to understand the key factors and conflicting information that led to the advice provided, as well as the level of certainty and other possible alternatives [53].

The accuracy of diagnosing Autism Spectrum Disorder (ASD) has been improved in this paper using graph GCNs. A deep ASD diagnosing framework, called DeepGCN, has been proposed that integrates ResNet units and the DropEdge strategy to enhance the performance of shallow GCNs. The results indicate that DeepGCN established models with a mean accuracy of 73.7% in classifying ASD and normal controls. A new perspective for the study of early diagnosis of ASD using multi-site and multimodality data is provided, with potential application in other mental illnesses. The support for final decisions and predictive analytics is therefore offered. A multi-level feature fusion technique and IoT system were deployed in [54] for the recognition of multimodal human activities in real-world smart healthcare application scenarios. CNN, Convolution Block Attention Module, and Convolutional Long Short Term Memory (ConvLSTM) were used to handle the time-sensitive multi-source sensor information. The performance of the developed fusion architecture was measured through experiments and evaluations using an open-access multimodal HAR dataset named UP-Fall detection dataset.

4.4. Blockchain

Many practitioners use blockchain in the healthcare sector to maintain patient data and exchange information amongst laboratories, hospitals, and pharmacies. Blockchain technology mitigates errors in the healthcare system. This category includes 7 of 64 articles divided into

two subcategories: data quality and healthcare sustainability.

4.4.1. Data quality

This subcategory includes 3 of 7 articles. Healthcare data must be of high quality to benefit from AI. Moreover, blockchain and sequential distributed learning have led to an increase in healthcare data quality [55]. Similarly, a model was created to incorporate trustworthy marijuana usage information. Data were collected from Twitter, and algorithms were designed to estimate marijuana use for improving drug abuse prevention. The model is used to assist policymakers by optimising the blockchain to exchange information on drug use between health centres [56]. The quality of data produced in the pharmaceutical manufacturing process was ensured in [57] through the use of blockchain technology and smart contracts. This was aimed at ensuring the authenticity, transparency, and immutability of the data. The objective of the EU-funded SPuMoNI project described in the paper was to accomplish this by verifying end-to-end data, evaluating the quality of the data, and implementing intelligent agents for data collection and manipulation. A traceable and auditable system was aimed at for pharmaceutical production in order to comply with regulations and to ensure the well-being of patients.

4.4.2. Healthcare sustainability

This subcategory includes 4 of 7 articles. Information security is a critical issue in the healthcare sector. In this context, the blockchain model must be tested for maintaining secure and trustworthy electronic health records (EHRs). The multi-criteria decision-making approach is used to evaluate and test blockchain technology for maintaining EHRs [58]. Blockchain technology is vital in hospitals for exchanging data and information. However, the COVID-19 pandemic has created the problem of generating such data and misinformation.

The authors of a paper proposed a blockchain-based early warning architecture for medical institutions to crowdsource early warning duties [59]. This paper presented a novel collaborative early warning system for COVID-19 based on blockchain technology and smart contracts. In addition, the authors aimed to reduce the danger of decision-making errors and enhance the performance of early warning by combining distributed surveillance forces and integrating multi-source surveillance resources to their fullest extent. The proposed system enabled two distinct forms of surveillance. It combined its monitoring results on new instances to generate an alarm, including medical federation surveillance based on federated learning (FL) and social collaboration surveillance based on learning markets.

Similarly, a paper presented a framework for healthcare sustainability by incorporating the privacy and reliability of blockchain in the healthcare industry [60]. The proposed framework was intended to facilitate the sharing of high-quality data and information between health centres and to support the internet of health-related devices. Maintaining a sustainable healthcare system while offering high-quality, effective, and safe care is a significant economic and social problem for the healthcare industry and its consumers. Collaboration across the healthcare systems of the EU and a shared vision for future activities would aid in achieving these objectives.

In [61], a Blockchain Decentralised Interoperable Trust (DIT) framework for IoT zones was proposed. A smart contract is guaranteed to authenticate budgets, while an Indirect Trust Inference System reduces semantic gaps and improves the estimation of the trustworthy factor through network nodes and edges. The proposed DIT Internet of Health Things (IoHT) uses a private Blockchain ripple chain to sustainable trustworthy communication by validating nodes based on their interoperable structure, thereby facilitating controlled communication necessary for solving fusion and integration issues across different zones of the IoHT infrastructure.

4.5. Transparency

The issue of transparency is important in healthcare AI applications. This category includes 6 of 64 papers that discuss transparent intelligent systems and develop interface categories.

4.5.1. Transparent intelligent systems

This subcategory includes 3 of 6 articles. A paper discussed the importance of transparency and credibility in an AI system [62]. Furthermore, knowledge from brain science was harnessed to develop smart systems in the healthcare sector. Another study identified the main roles in the deployment of the system to evaluate the AI system from a practice perspective [63]. Regarding robotic surgical training, the paper provided approaches to developing AI applications based on ethical guidelines to be used in surgery [64].

4.5.2. Interface development

This subcategory includes 3 of 6 articles. Many practitioners and academics recommend that the trustworthiness issue should be considered in the development interface for reliable systems. Collaborative design processes have been explored to develop a trustworthy interface in the healthcare sector. The designed interface provides help to physicians as they make judgements on the level of trust they place in patients for their hospital care [65]. In this context, attempts to incorporate transparency and ethical issues into healthcare AI systems have increased exponentially. An ethical healthcare development interface is a critical issue for successfully adopting AI technology in healthcare [66]. Therefore, it was presented on the development of ethical interfaces for technology, the ethics regulation interface was revised, and the ethical regulation interfaces were described. However, one consequence of medical methods has been noted. As a result, the authors of the study in [67] were encouraged to urge the international XAI community to engage in research on multimodal embeddings and interactive explainability in order to construct effective human-AI interfaces. This study determined that graph neural networks (GNNs) are crucial for multimodal causality due to the ability of graph topologies to explicitly establish causal linkages between features.

4.6. Trustworthy digital health

Digital health is a broad concept that includes adopting digital transformation in healthcare by integrating AI tools into the healthcare system. This category includes 11 of 64 contributions explaining health services and the Internet of Medical Things (IoMT).

4.6.1. Health services

Health services-related AI applications aim to explore the linkages between clinical practices and patient outcomes, with the ultimate objective of improving patient care. AI programs are used in a variety of areas, including diagnostics, the development of treatment protocols, drug discovery, personalised medicine, and patient monitoring and care. This subcategory includes 8 of 11 articles. A study investigated the role of FL in digital health to improve health services in a trustworthy manner [68]. The literature confirmed that AI and ML are critical tools in the digital health system. The International Telecommunication Union (ITU) in partnership with World Health Organization (ITU/WHO) focus group on 'AI for Health' is developing validation criteria for health AI that can aid in the transparent evaluation of the quality of powerful but complicated technologies. Standardised benchmarking is particularly useful for determining the advantages and limitations of various health ML/AI models.

The ITU/WHO program has been introduced and contextualised within the broader landscape of digital health and AI standardisation initiatives [69]. Health information technology (HIT) can play a crucial role in establishing national learning health and care systems, which are defined as health and care systems that continually leverage

data-enabled infrastructure to support policymaking, public health, and personalised care. The COVID-19 pandemic provided an opportunity to evaluate the UK's capacity to use HIT and apply the principles of a national learning health and care system in response to a major public health crisis. To avoid impeding progress, it is important for all stakeholders to be responsive to the ethical challenges and unintended consequences of HIT [70].

Regarding kidney disorders, a framework was proposed that was built on digital health services, and it possessed high levels of ethics and governance to support medical judgements and enhance the health service. Several AI/ML-based systems or devices presented the potential to reduce systemic and individual issues associated with kidney disease or to prevent the disease or its progression in at-risk individuals or populations. However, many of these software programs or devices, including wearable health gadgets for dialysis patients that combine telemonitoring and tailored feedback, are still in the early stages of development, and they have not been thoroughly tested to produce beneficial patient outcomes [71]. An intelligent diagnosis decision support system for telemedicine in the Middle East and North Africa (MENA) region was proposed in [72]. The use of a vast health-related dataset curated by the Altibbi company, including unstructured patient questions written in various dialects of Arabic and structured symptoms identified by general practitioners, was incorporated into the proposed system. The aim of the study was to automate the diagnosis process through the provision of an intelligent model, which was intended to assist doctors and clinicians in making accurate decisions and delivering appropriate healthcare services.

In [73], the focus of the study was on the design and deployment of RAI-based solutions for personalised pregnancy healthcare, which includes an exploration of the data types, their sources, and characteristics from a multimodal data fusion perspective. The prevention of pregnancy risks and the improvement of the experience of pregnant women during the process by means of effective computing for detecting emotional distress, in accordance with the biopsychosocial diagnostic model, were the primary objectives. A smart mobile architecture designed for enabling rapid prototyping of personal health monitoring applications with a focus on ensuring the trustworthiness of collected data and provided alerts was presented in [74]. In [75], a healthcare solution based on AI, IoT, and edge computing was proposed for continuously monitoring and providing efficient and reliable healthcare services for the disabled or elderly in pandemic situations in smart cities. The solution used a neural network and was described as being scalable and reliable, with a focus on ensuring the trustworthiness of the system through the use of AI-enabled IoT technology.

4.6.2. The Internet of Medical Things

The IoMT is the collection of medical equipment and applications that link to healthcare IT systems online over computer networks. Machine-to-machine communication, the foundation of the IoMT, is made possible by medical gadgets integrated with Wi-Fi. This subcategory includes 3 of 11 papers. The IoMT contributes to improving the health services provided by maintaining the safety and privacy of patients and enabling doctors and hospitals to provide superior services. However, confidentiality and privacy issues still exist and need to be resolved. Therefore, a paper presented an energy-efficient IoMT model to ensure the confidentiality of the information and to help doctors diagnose diseases [76]. Similarly, it has been proposed that the energy-efficient IoMT model can increase privacy, improve patient health services, and reduce energy consumption and communication expenses [77]. The use of AI in applications such as the IoMT can enhance privacy, provide more accurate disease diagnosis, ensure compliance with ethical and regulatory standards, ease the burden on healthcare providers, improve patient care, and save time and costs.

In [78], a design was presented for the detection of abnormal traffic in the IoMT-Blockchain environment using a deep neural network (DNN). The proposed method first introduced a feature extraction algorithm based on multimodal autoencoders that reduce the complexity of traffic feature information by processing it in groups and extracting fusion features by constructing a multimodal autoencoder. To leverage traffic data information in the detection network, a multi-feature sequence anomaly detection algorithm was also introduced. This algorithm extracts low-level fusion features and high-level temporal features from network traffic and applies them to anomaly detection and classification tasks using residual learning.

4.7. Review

This category provides an insight into earlier works in the literature for a trustworthy AI review and includes 8 of 64 articles. Due to the importance of trustworthy AI, reviews of previous work were conducted in different sectors. For example, a paper provided an overview of the transparency of health informatics processes [13]. Patient trust and ethical requirements were described. In addition, another review was presented in the context of the use and reliable prediction of ML in oncology to improve medical care [18]. The use of AI in the healthcare sector has increased significantly. However, many ethical issues have arisen and need to be resolved. The most influential elements of the ethics of AI were identified by conducting a bibliometric analysis. The ethical categories include defining meta-ethics, medical practice, normative standards, epistemological knowledge, medical and legal concerns, patient rights, ambient intelligence, ethics for robotics, predictive analytics, relationships, and clinician rights [14]. Furthermore, a mini review of XAI in medicine and digital health was conducted. Two XAI methods were proposed that rely on supervised learning for COVID-19 classification [17].

A comprehensive survey on IoMT-based fusion for smart healthcare, a framework that utilises technology for the improvement of healthcare, was conducted in [15]. The use of medical sensors, as well as the importance of multimodal medical signals for disease diagnosis and the requirement for efficient signal fusion, was mentioned in the survey. In addition, a survey of related research in the field, including works published between 2014 and 2020, was conducted, and key challenges and potential future directions were emphasised. In the review in [16], the studies using XAI in AI models for healthcare applications were examined. The authors found three major healthcare datasets, including clinical features, text, and high-dimensional data. The review also included a comparison of the optimal performance of both ML and DL models for XAI. The areas in need of more attention, such as XAI for biosignal abnormalities and clinical note interpretation, were noted. It was concluded that an ideal AI model for healthcare applications should have both high performance and interpretable results. IoT- and IoMT-based edge-intelligent smart health care was surveyed using journal articles from 2014 to 2020 in [79]. Smart healthcare encompasses a wide range of technologies, including IoT, IoMT, medical sensors, AI, edge computing, cloud computing, and next-generation wireless communication technology. Research in this area has primarily focused on exploring the applications and potential of IoT and IoMT, AI, edge and cloud computing, security, and medical signal fusion. Ongoing research is addressing current issues in these areas, while also exploring new directions for future research.

In [80], a comprehensive review of the field of brain disease detection through the fusion of neuroimaging modalities using DL models such as CNNs, recurrent neural networks (RNNs), pretrained generative adversarial networks, and autoencoders was presented. The need for fusion of neuroimaging modalities and the different types of neuroimaging modalities were first discussed. Then the published review papers exploring the use of AI techniques in the field of neuroimaging multimodalities were reviewed. Furthermore, the fusion levels based on DL methods, including input, layer, and decision, and their applications in diagnosing brain diseases, were discussed. This first scientific systematic review on the topic of the trustworthiness of AI in healthcare, was regarded to be of high quality since it offered accurate statistics

from many sources and a huge amount of knowledge.

5. Discussion

This section discusses three important aspects of the literature on trustworthiness in AI: motivations, challenges, and recommendations to mitigate these difficulties with the aim of improving healthcare quality. Before delving into these aspects, it is important to first discuss the findings and analysis of the presented taxonomy in the earlier section regarding the assessment of quality, bias risk, and data fusion, as explained next:

• *Explainable robotics*

In explainable robotics, quality assessment is essential for ensuring that the AI systems developed for human–robot interaction are effective and trustworthy [24]. This is particularly important in the healthcare sector, where errors in AI systems could have serious consequences. Bias risk is another important consideration in explainable robotics. The gap between human-centred and machine-centred disease diagnosis highlights the potential for bias to be introduced into AI models and systems [21]. It is crucial to assess the risk of bias and develop measures to mitigate it. Finally, data fusion can improve the accuracy and reliability of AI systems in healthcare. By combining multiple sources of information, such as graph analysis and feature visualisation, data fusion techniques can help overcome limitations in individual data sources and improve the overall accuracy of AI systems.

• *Hospital death prediction*

In the context of hospital death prediction, invalid and unreliable predictive methods can have significant consequences. Quality assessment is crucial to ensure that the predictive methods used are accurate and effective. This can help prevent errors and improve patient outcomes [26]. Bias risk is another important consideration, as certain populations may be over- or under-represented in the data used to develop the predictive methods, leading to biased results. It is crucial to assess the risk of bias and to develop measures to mitigate it. Finally, data fusion can improve the accuracy and reliability of prediction methods for hospital deaths [27]. By combining multiple sources of information, data fusion techniques can help overcome limitations in individual data sources and improve the overall accuracy of the predictive methods.

• *Trustworthy diagnosis*

The use of computer-aided diagnostic systems is intended to provide transparent and reliable decision-making in the context of disease prediction and detection [28]. Earlier studies have introduced measures to assess the trustworthiness and credibility of workers in diagnosing autism, using controls for neural patterns and behavioural trait scores [29]. Additionally, the evaluation of binary classifiers and diagnostic tests is suggested to be carried out using 'deep ROC analysis', which presents a compromise between AUC and accuracy [35]. This highlights the importance of evaluating the quality of AI-based diagnoses and ensuring that they are free from bias. However, the studies do not explicitly mention the use of data fusion in the context of trustworthy diagnosis. Data fusion has the potential to be used in the healthcare industry to improve the precision of AI-based diagnoses by combining data from various sources, including patient data, medical images, and demographic data. Data fusion could be combined with other fields of studies such as complex networks and their inference, graph causal models, and verification and explainability methods to create more dependable and strong medical AI systems. To prepare for an AI system that is fair, acceptable, and relevant, it is necessary to develop national and international standards and regulations that define the parameters and determine the limits of operation and commitment. These rules must also be successfully implemented, which is a prerequisite for ensuring that AI will be suitable to achieve the aim of serving the public and the global good, and it is imperative that this be done as soon as possible [36].

• *Explainable AI*

In healthcare, trustworthy AI plays a crucial role in supporting and guiding decisions made by biomedical experts. XAI methods, such as layer-wise relevance propagation and CAMs, aid in developing transparent and interpretable AI systems. Several studies have demonstrated the success of XAI in medical applications, such as diagnosing glaucoma and identifying women with GDM who need tailored prenatal treatments [48]. Explanations generated by XAI systems have increased trust and perceived fairness in AI-assisted decision-making scenarios. Bias in AI systems can lead to unfair and unreliable decisions. This highlights the importance of XAI in reducing bias risk in AI systems and ensuring fair decision-making in healthcare [21]. Furthermore, a new data fusion scheme combining type-2 fuzzy logic and DST has been proposed, demonstrating improved accuracy and explainability in IoT-based healthcare systems [49]. In conclusion, XAI methods play an important role in ensuring that AI systems in healthcare are trustworthy and able to support responsible decision-making.

• *Developing AI systems*

The development and utilisation of AI in healthcare require a combination of quality assessment, risk management, and data fusion techniques to ensure that the resulting systems are trustworthy and explainable. By incorporating these factors into the development process, AI can play a crucial role in improving patient outcomes and decision-making in healthcare. The development of a diagnostic model for AD, for example, aims to provide accurate results and explanations for each decision, contributing to a better clinical understanding of the disease [51]. Additionally, physicians must understand the recommendation's certainty and potential alternatives to build confidence in the CDSS in AI systems [52]. The potential for bias in AI systems for healthcare is a concern. It is essential to ensure that the models are developed transparently and that physicians know the primary influencers and contradictory facts leading to certain advices. To minimise the risk of bias, including physicians in model development and training physicians in the field is crucial. Data fusion techniques are used to enhance the performance of AI systems in healthcare by combining multi-source sensor information. Using a combination of CNNs and ConvLSTM [54], the fusion architecture is evaluated using an open-access multimodal dataset to measure its performance. This demonstrates the potential for data fusion in improving the accuracy and reliability of AI systems in healthcare.

• *Data quality*

Quality healthcare data is essential for ensuring that AI can produce accurate results that benefit patients and healthcare providers [56]. Blockchain technology and sequential distributed learning have been identified as effective tools for improving the quality of healthcare data. Furthermore, projects such as the EU-funded SPuMoNI aim to establish auditable and traceable systems for pharmaceutical production to ensure the authenticity and transparency of the data used in AI [57]. The management of bias risk is also critical, as biased data can lead to inaccurate results and perpetuate existing inequalities in the healthcare system. Lastly, effective data fusion can help to improve the overall accuracy of AI results by combining information from multiple sources.

• *Healthcare sustainability*

It is an important aspect of the healthcare industry and is closely related to the quality, reliability, and security of healthcare data. The use of blockchain technology and smart contracts in healthcare can help to maintain a sustainable healthcare system by improving the quality and reliability of EHRs [58], enabling secure and trustworthy data exchange, and facilitating collaboration across healthcare systems. The proposed blockchain-based early warning system for COVID-19 and the framework for healthcare sustainability are examples of how blockchain technology can be utilised to ensure the sustainability of healthcare systems [60]. The proposed DIT framework for IoT zones aims to improve the estimation of trustworthy factors and facilitate controlled communication necessary for solving fusion and integration issues across different zones of the IoHT infrastructure [61].

• *Transparency (Transparent intelligent systems & interface development)*

The studies focus on the importance of transparency and credibility in developing trustworthy AI systems in healthcare. It highlights the need for developing ethical and transparent interfaces, as well as considering the roles of various stakeholders in deploying AI systems [66]. The authors of a study recommend further research in the area of multimodal explainability to improve human–AI interactions in healthcare. The study also highlights the importance of using neural graph networks for multimodal causality in developing effective AI systems [67]. In conclusion, assessing quality, bias risk, and data fusion in trustworthy AI is crucial for successfully adopting AI technology in healthcare.

• *Health services*

Health services-related AI applications investigate the connections between clinical techniques and patient outcomes. AI and ML are considered critical tools in digital health systems. They are used in diagnostics, treatment protocol development, drug development, personalised medicine, and patient monitoring and care [68]. The ITU/WHO focus group on 'AI for Health' is developing validation criteria for health AI to improve the transparent evaluation of their quality [69]. Standardised benchmarking is useful for determining the advantages and limitations of various health ML/AI models. The COVID-19 pandemic showed the potential of using HIT in the UK to create a national learning health and care system. Still, ethical challenges and unintended consequences must be taken into consideration [70]. In kidney disorders, AI/ML-based systems can potentially reduce issues associated with the disease [71]. However, they are still in the early stages of development and have not been thoroughly tested. An intelligent diagnosis decision support system has been proposed for telemedicine in the MENA region to assist doctors and clinicians in making accurate decisions [72]. Some studies focus on the design and deployment of AI-based solutions for personalised pregnancy healthcare, with a focus on ensuring the trustworthiness of collected data. A healthcare solution based on AI, IoT and edge computing has been proposed for continuously monitoring and providing efficient and reliable healthcare services for the disabled or elderly in smart cities, focusing on ensuring the system's trustworthiness through the use of AI-enabled IoT technology.

• *The Internet of Medical Things*

The assessment of quality, bias risk, and data fusion of the IoMT is an ongoing process that must adapt to the changing needs of the healthcare industry. As new medical devices and AI algorithms are developed and integrated into the IoMT, it is important to continuously evaluate and validate their performance to ensure they meet the necessary standards for accuracy and reliability. One way to adapt the studies is to incorporate feedback from healthcare providers and patients into the evaluation process [76]. This will help ensure that the IoMT is designed to meet the needs and preferences of those using it. Additionally, as privacy and security concerns continue to be a major concern in the healthcare industry, it is important to regularly assess and update the measures in place to protect patient information [77]. Another way to adapt the studies is to incorporate emerging technologies into the IoMT. For example, the use of blockchain technology can help to improve the security and privacy of patient information by providing a decentralised, secure platform for data storage and transfer [78]. The use of edge computing and fog computing can also help to reduce the latency and improve the reliability of the IoMT by processing data closer to the source. In conclusion, assessing quality, bias risk, and data fusion of the IoMT is a dynamic process that must adapt to the changing needs of the healthcare industry. By incorporating feedback from healthcare providers and patients, incorporating emerging technologies, and continuously evaluating and validating the performance of the IoMT, it is ensured that it remains effective, reliable, and secure.

5.1. Motivations

Integrating trustworthy AI into healthcare allows for many benefits, which include assisting in the accurate diagnosis of diseases, improving patient care, reducing pressure on healthcare providers, and improving privacy in the healthcare sector. Additionally, it benefits ethical and regulatory values, reducing the cost and duration of treatment by saving time. Finally, it benefits collaboration between parties globally. These categories and more details about these benefits are elaborated in the following subsections (see Fig. 8).

5.1.1. Enhance the clinical understanding of the diagnosis

A diagnosis is based predominantly on laboratory reports or test results rather than the physical examination of a patient. Usually, relying on this traditional method of diagnosis causes an increase in the number of people in queues waiting for the doctor, who, in turn, may miss or misdiagnose something due to the momentum. In contrast, the clinical diagnosis is made based on medical signs and reported symptoms rather than diagnostic tests. The authors of a study ensured that health ML/AI models were trustworthy and would become increasingly significant in healthcare, such as for diagnosis or prognosis reasons, including a variety of pattern recognition tasks [69]. In a separate study [51], a multilayer multimodal detection and prediction model was proposed to improve clinical knowledge of the diagnosis and progression of AD by analysing the effect of different modalities on disease risk. The model was shown to offer benefits such as decentralised decision-making, fairness, auditability, and universality, making it a valuable resource for early detection and prevention of infectious diseases. Another study explored the use of AI in medical assistance and employed a participatory design strategy to develop an explanation interface to help laypeople make informed decisions about trusting AI systems for their healthcare needs [65]. The interface was designed to be trustworthy and easy to understand, even for non-experts in AI.

Due to the vast amount and diversity of clinical and imaging data in ophthalmology, an increasing number of AI systems are being proposed for various stages of patient care, offering potential benefits. There is a significant gap between the development and integration of AI systems in ophthalmic practice even though AI systems can achieve expert-level or even better performance [52]. The authors of a paper suggest that XAI for dermatology would give dermatologists an effective screening tool that they both understand and trust, as an understandable explanation is one of the pillars of any computer-aided design (CAD) system [30]. The medical expert can utilise AI to assist in resolving medical problems and evaluating scientific validity, analytical performance, and clinical performance [44]. Moreover, the study was motivated by the notion that there is significant competition between intelligent systems and human diagnosing capabilities. However, due to XAI solutions, there should be no such dichotomy between datasets, computational tools (e.g., DL and

Fig. 8. Motivations for trustworthy AI categories.

CNN used in this study), and the skills acquired through the experiences of doctors [45]. AI could transform medical practice and redefine the roles of clinicians as a result. AI systems have proven successful in a number of therapeutic activities, including assessing patient prognosis and assisting with surgical procedures. More importantly, the authors recognised that no single algorithm can adapt to or resolve all problems [17]. Success typically depends on the nature of the issue at hand and the available information. Usually, a problem requires a hybrid solution that combines multiple techniques to produce a tangible solution. Each issue is accompanied by a comprehensive examination of what constitutes the optimal algorithm. In addition to training time, it is important to evaluate the transparency of the input size, the capabilities of the DNN, and the efficiency of the network over time [81].

Ethical development of healthcare interfaces and transparent healthcare systems is critical for the successful adoption of AI technology in the field, particularly in areas such as brain sciences. Some scientists discovered that the neurons in the human brain are responsible for thinking/reasoning and have considerably more complicated tree-like structures with deep roots and surface branches, which can be used to mimic computational neurons for implementing deep networks such as CNN. In essential applications such as healthcare and defence, practitioners prefer not to rely on such systems. Although AI systems are frequently built to draw inspiration from the brain, little effort is being made to truly harness brain inputs [62]. CDSS is designed to assist healthcare providers in making informed clinical decisions, by analysing patient-specific evidence items that trigger the inferencing process, such as examination findings and expert-modelled or machine-learned clinical knowledge. The suggestions provided by these systems can help in accurately diagnosing a patient and determining the appropriate course of treatment [53]. To further advance the field of medical diagnosis and improve human health, there is a need to develop AI models that are explainable and less biased and can learn from limited data. It is also important to bridge the gap between research and practical medical applications, especially in the context of future trustworthy medical AI [73]. This requires the establishment of data collection methodologies that promote responsible AI in healthcare, as exemplified by a prospective and qualitative study on pregnancy.

5.1.2. *Improve patient care*

The concept of patient care encompasses the provision of healthcare services by medical professionals to prevent, treat, and manage illnesses while promoting the physical and mental wellness of their patients. In [27], the authors questioned why the task-by-task mechanism is used when assessments and therapies for various illness categories can vary substantially. For example, the diagnosis and treatment of diseases of the pulmonary system and disorders of the genitourinary system are very dissimilar. Therefore, it is difficult to exchange feature spaces between these two circumstances. By adopting this approach, not only the difficulty of training can be reduced but also the induction of noises can be prevented. According to the draft of 'Policy Guidance on Artificial Intelligence for Children' of UNICEF [63], data-driven healthcare therapy or medication for children should not be based on data from adults,

as doing so could pose unforeseen health hazards to children. In [18], the authors evaluated each component of the problem in the context of the trustworthy application of ML approaches in oncology as a representative research example, with the goal of ensuring utility and enhancing patient care in medicine. The high demand for personalised, intelligent solutions in the medical domain has created a need for responsible, trustworthy, useful, and safe AI models that can improve overall patient experience, reduce potential complications, and address the uncertainties that arise in diagnosing COVID-19 [39]. Finally, using timely and accurate AI fall detection through CNN that combines multiple data streams in patient care can improve the delivery of medical services by facilitating quick responses and immediate assistance for patients who have fallen [54].

5.1.3. Healthcare providers

There are several benefits for healthcare providers, e.g., reduced hospital readmissions by accurate diagnosis, reduced inappropriate healthcare interventions, reduced duplication of services, and reduced total healthcare spending. According to the paper in [32], the system plays a significant role in assisting physicians and healthcare providers in having 24/7 remote patient monitoring systems and assisting patients in reducing chronic disease risk factors; the system will be an additional tool for physicians to remotely monitor more patients without increasing their workload. In many cases, people in rural areas who need medical treatment could benefit from high-quality hospital diagnoses aided by ML, which can supplement their experience with expert information from other institutions, ensuring diagnostic consistency to benefit healthcare practitioners. AI has the potential to improve access to healthcare services and specialists, such as specialised treatments, for individuals in low- and middle-income countries where healthcare accessibility is still a global concern [34]. In a study presented in [29], researchers established a set of metrics that were validated through experiments to evaluate the trustworthiness and reliability of crowd workers who were responsible for assigning behavioural feature tags to unstructured videos of children with autism and matched neurotypical controls. This method can help to filter the crowd workforce, making it a cost-effective way to create a reliable and trustworthy diagnostic workforce.

The influence linked communities have on precision health or medicine, and vice versa, provides prospects for any type of study and survey data, such as predicting trends in health-related concerns, for example, drug use behaviour of people. Here, precision medicine changes how the information is handled and achieves a better output to support the choice of the stakeholders. Free online social network analysis appears to be a useful tool for rapidly monitoring population behaviour [56]. Various AI-based models have been developed to diagnose COVID-19. However, only a few models have been able to combine the human-centred approach to the diagnosis of disease with the potential of machine-centred diagnosis. The concept of human–computer interface design empowers clinicians to control the explainability and interpretability of prediction models using decision trees and feature visualisation. By employing this innovative feature selection step, the suggested diagnosis model increased diagnostic precision and lowered execution time [25]. In [33], a study paved the way for a more intelligible time-series analysis in the context of early antimicrobial resistant (AMR) prediction in ICUs and shortened the time required to discover infectious diseases, thus offering the possibility of improved hospital care. Developing accurate, data-driven decision-aided methods that predict diseases early and support healthcare decision makers can benefit healthcare providers by improving diagnostic accuracy, facilitating early interventions, and enabling more efficient use of resources [37].

5.1.4. Protects privacy in the healthcare sector

With respect to healthcare privacy, there exists a federal legislation that grants individuals control over their health data and establishes regulations and restrictions on who can access and receive such information. Blockchain technology is one of the most significant and transformative advancements in the IT sector. It has an important position in the current digital era and has already had a profound impact on human existence. In addition, it is projected that blockchain technology will enhance the existing IT infrastructure in several fields over the next few years. Recent technical advancements have enabled significant progress in the healthcare sector. Information security and accessibility are crucial factors to consider when integrating and connecting with EHR systems to share confidential medical data. In this context, selecting the most effective blockchain model for safe and trustworthy EHRs in the healthcare sector requires a precise system for evaluating the impact of the many accessible blockchain models based on their attributes [58].

In medical and healthcare applications, there should be greater emphasis on security and trust in IoT devices that create an ecosystem to detect the medical conditions of patients, such as blood pressure, oxygen level, heartbeat, and temperature, and take emergency measures [76]. The healthcare-related data of patients is transmitted to remote users and medical centres for subsequent analysis [77]. Transparency in healthcare data processes is a condition to trust in practitioners and patients and for the adoption of digital tools to keep data confidential and not share sensitive information with a provider, which may also induce unethical situations [13]. The AI model proposed in [76] by the authors aimed to increase the maintainability of disease diagnosis systems and support trustworthy communication with the integration of the medical cloud to improve data transfer in medical applications with energy-saving and reliability. Maximum likelihood evidential reasoning (MAKER) is advantageous for pre-processing partial and unclear data [47]. It is possible to translate categories inside an attribute into interpretable numerical features by taking into account ignorance caused due to missing values in input and output attributes.

To enable the UK health and care systems to become data-enabled learning health and care systems, data integration is essential. This will provide all relevant stakeholders with seamless and simultaneous access to health data, which will help to strike the right balance between top-down and bottom-up implementation, enhance usability and interoperability, improve data processing and analysis capabilities, address privacy and security concerns, and promote digital inclusion. These priorities are crucial for strengthening HIT in the UK and the EU [70]. Finally, the IoHT has enabled the universal adoption of the linked health paradigm [60]. The IoHT can deliver linked health monitoring with high-quality service and ultra-low latency due to 5 G support in the healthcare vertical. DL has demonstrated the ability to handle daily huge volumes of IoHT data, automate linked healthcare activities, and aid in decision-making. The objective of authors is to ensure the viability of IoHT-enabled health applications.

5.1.5. Ethical principles and global collaboration

Ethical values provide guidance on acceptable and desirable behaviour in businesses beyond mere compliance with regulations and laws. Despite the presence of learning as a core aspect of a larger Learning Health System strategy, digitalisation in nephrology has been carried out in a fragmented and compartmentalised way across the major clinical application areas. Existing ethical and regulatory processes do not appear to match the demands of AI/ML-based software developers or guarantee public confidence or larger public expectations around data management and security [71]. Putting ethics at the forefront of continuous attempts to create technologies that promise to improve our lives is surely encouraging. It is rumoured that even IT corporations are contemplating the need to employ a chief ethical officer to work alongside legal teams and chief compliance officers. Strong professional regulatory agencies might assist the ongoing improvement of the ethics/regulation interface, which would not only improve the role of technology in society but may also reduce the prevalence of the damaging desire of avoiding political divisions by resorting to ethics

instead [66]. There is increasing interest in ML solutions that address clinical and biological issues. This is a result of the encouraging findings in several research articles, the growing number of AI-based software applications, and the widespread interest in AI as a tool for solving complicated issues. It is essential, therefore, to enhance the output quality of ML and incorporate safeguards to facilitate their adoption [26].

The purpose of the Delphi method is to define the ethical and regulatory implications of AI in robotic surgery training. The Delphi technique has led to international consensus among experts to design and validate material for guidelines on the ethical implications of AI in surgical education [64]. Research has provided policymakers, AI developers, and scientists with a map indicating which aspects of AI-based medical interventions demand stronger rules and guidelines, rigorous ethical design and development, and biotechnology in healthcare [14]. The blockchain architecture was advantageous for tracing data origin and monitoring the training procedure for model degeneration and dishonest conduct. It is believed that the technique will build trust between parties, thereby promoting worldwide collaboration between parties when producing big data-informed AI that is resilient [55]. A paper concluded by introducing a new research field known as Explainable Robotics, which investigates the explicability of human–robot interactions [24]. The focus is on the development of innovative computational models, methodologies, and algorithms for providing explanations that enable robots to function at varying levels of autonomy and interact with people in a trustworthy and human-friendly manner. Individuals may require explanations during human–robot interactions for a variety of context- and human-user-dependant reasons. The integration of AI in various applications emphasises the need for prioritising the development of explainable and robust AI models that can handle complex tasks such as what-if questions and perform well with smaller datasets to minimise the possibility of bias [67]. This need, in turn, can inspire the global XAI community to delve deeper into multimodal embeddings and interactive explainability, setting the groundwork for more effective human–AI interactions in the future.

5.2. Challenges

The principal challenge is making the abstract ideas outlined in the guidelines applicable to each case study. This will definitely include refining the broad notions (e.g., autonomy or privacy) expressed in the principles into concepts applicable to the case study. Even though from a pragmatic standpoint this seems appropriate, due to both time restrictions and the multidisciplinary character of the group of researchers, the philosophical, ethical, and legal scope and depth of the discussion are undoubtedly constrained. These challenges are outlined next.

5.2.1. In terms of legal aspects – respecting all applicable laws and regulations

The ITU/WHO focus group on ‘AI for Health’ is developing validation criteria for health AI that may be used to evaluate the quality of powerful but complicated technologies in a transparent manner [69]. A paper addressed the difficulties with certain techniques and how they may end up consciously or unknowingly impeding the implementation of a legal system equipped to protect fundamental human rights [66]. However, the most significant research obstacles are the storage of medical data on a secure cloud and the improvement of the energy efficiency of illness diagnosis systems [76]. In addition, the rapid expansion of IoMT technology has attracted a large number of cybercriminals who make persistent attempts to hack medical equipment by causing data loss and generating fraudulent certifications. The increase in new technologies for healthcare applications based on the IoMT, the protection of health data, and the provision of trustworthy communication against attackers is attracting considerable research interest. This poses significant difficulties for IVD devices that utilise ML algorithms for data analysis and decision assistance. This can increase the complexity of implementing some of the most effective ML and DL approaches in the biomedical domain due to the failure of manufacturers to provide the necessary explanatory components [44]. Glaucoma is a serious eye condition that can lead to blindness if left untreated for an extended period. As a result, early diagnosis of the disease is critical. Several studies have utilised DL to diagnose glaucoma from colour fundus images. However, it is essential to understand how humans can rely on these models, which are often black box, to make decisions [45]. Addressing data fusion challenges related to uncertainty and lack of essential data for accurate diagnosis involves not only developing a holistic general identification approach but also addressing difficulties in merging data from multiple datasets to create a homogenous dataset suitable for AI and XAI analysis [37]. Respecting all applicable laws and regulations is crucial to ensuring legal compliance and safeguarding data privacy in the development and deployment of AI systems.

An essential part of responding to infectious disease emergencies is early warning. However, most existing early warning systems are centralised and operate in isolation. This can lead to the risk of decisional errors and biases due to reliance on a single source of evidence [59]. CDSS supports clinicians in making clinical decisions [53]. For CDSS, the major challenges have been identified as the presentation and explanation underlying the computation result of the model and its reasoning via a sophisticated human–computer interface. The impact that linked communities have on precision health or medicine, and vice versa, provides potential for any study and survey data, for example, to predict trends in health-related concerns, particularly the drug use behaviour of people. Here, precision medicine changes how information is handled and achieves a more favourable outcome to enhance stakeholder decision-making [56]. When integrating and connecting with EHR systems to share confidential medical information, information security and accessibility are crucial factors to consider [58]. The effectiveness of AI in healthcare largely relies on the quality of the data used to develop models and the trust in the predictions they generate. However, obtaining sufficient amounts of high-quality data to build precise and dependable models remains challenging due to significant legal and ethical restrictions that limit the availability of clinically relevant research data outside the clinical setting [55]. All these references relate to the authors who fulfilled the lawful principle by declaring a partnership such as government collaboration (e.g., the National Key R&D Programme of China, German Federal Ministry of Education and Research). In summary, Fig. 9 provides a visual representation of the studies reviewed in this section, highlighting their main focus on laws and regulations challenges.

5.2.2. In terms of ethics – respecting ethical principles and values

With health informatics becoming more prevalent in healthcare, the ethical considerations surrounding it are also growing in importance, particularly in the context of biomedical informatics [13]. EHRs, the expansion of care teams, and the need for data sharing to ensure continuity of care and support research, as well as the development of AI-powered decision support tools, are all driving the need for new standards of trustworthiness. These standards must address issues such as maintaining patient privacy and confidentiality, preventing security breaches, ensuring accurate patient information, reconciling data sharing and reuse policies, and increasing transparency in the effectiveness of apps. In order to ensure long-term sustainability in healthcare, it is essential to prioritise data security, data privacy, and social acceptance of the data-driven decision-making process [60]. Moreover, integrating AI into robotic surgical training presents numerous opportunities and concerns [64]. Risks associated with the ethical application of AI include data and privacy concerns, lack of transparency, biases, accountability, and liabilities. The ability of AI to deliver on its promises is contingent on the successful resolution of identified ethical and practical concerns, such as explainability and algorithmic bias. Even though such concerns may initially appear purely practical or technical,

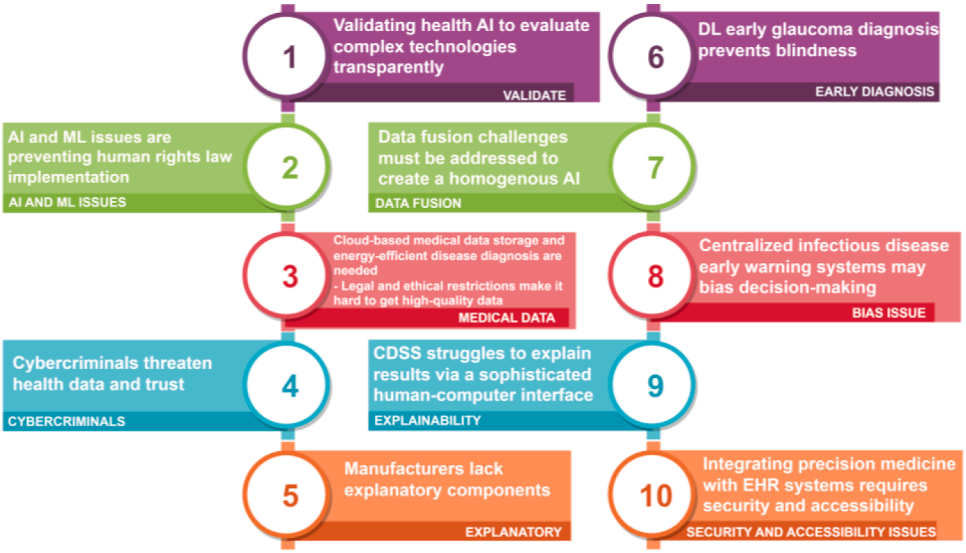

Fig. 9. Legal aspects challenges for trustworthy AI.

a closer examination reveals issues of value, fairness, and trust [34].

Integrating EHRs among healthcare providers, investment in health data science research, generation of real-world data, development of AI and robots, and facilitation of public–private partnerships are current and future issues. When deploying HIT, many ethical problems and unforeseen consequences occur [70]. The key challenge is to find competent, trustworthy and dependable employees through high-fidelity evaluation assignments without disclosing sensitive patient information during the evaluation process [29]. EHRs provide extensive, heterogeneous information concerning the evolution of the health status of patients. However, these data must be meticulously processed to obtain useful information for clinical decision assistance [33]. In an ICU, estimating patient mortality plays a crucial role. Currently, DL models underpin the majority of learning methodologies. However, these mortality prediction methods suffer from two issues: due to the variety of diseases, the specific causes of death are not considered in the learning process, and symptoms are combined without differentiation and localisation, and the mortality prediction learning outcome is not self-explanatory for doctors [27]. In Bio-IoT applications, imbalanced energy consumption between biosensor nodes impedes the timely transmission of patient data to remote centres and affects the medical system. In addition, the sensitive information of patients is transmitted through the insecure internet and is susceptible to security risks. Consequently, data privacy and integrity against hostile traffic are additional problematic research issues for medical application [77]. Existing medical data are not completely used by ML partly since they reside in data silos and access to these data is restricted by privacy concerns [68]. However, without access to sufficient data, ML will not be able to achieve its full potential and, eventually, will not be able to shift from research to clinical practice. Ensuring AI explainability requires addressing not only data limitations but also ethical considerations related to the fusion of diverse data sources to generate accurate analytical results, especially when the goal is to create fewer images that contain all the relevant details. Respecting ethical principles and values is essential to ensuring the development of responsible and trustworthy AI systems [54]. The work presented by [39] discusses the difficulty of creating an effective diagnostic tool for COVID-19 using CT scans and X-rays, while also ensuring the accuracy of the model's predictions. In order to improve the classification of large datasets with various types of images, a DL feature fusion model was identified as an appropriate solution. This highlights the importance of ethical principles and values in the development of diagnostic tools, which must be accurate and efficient to empower healthcare practitioners to provide high-quality care to their patients. In summary, Fig. 10 provides a visual representation of the studies reviewed in this section, highlighting their main focus on ethical principles and values challenges.

5.2.3. *In terms of robustness – from a technical perspective while considering its social environment*

The Continuity of Care Documents CCD system gives healthcare providers a remote patient monitoring system that is accessible 24/7 and enables patients to access medical care round-the-clock. These methods are regarded as further aids for clinicians and patients in monitoring and controlling the disease [32]. The primary research obstacles in Explainable Robotics are identified by the paper in [24]. The primary hurdle is the requirement of novel algorithms that produce justifications, which incorporate earlier encounters, comparisons, and current data and can adjust to specific contexts and objectives. The second challenge for research is to design new computer models of situational and learned trust, as well as techniques for detecting trust in real time. AI, notably in DL/DNN systems, has attained unrivalled performance [62]. However, the majority of these systems cannot explain why a specific decision is taken (black box) and fail badly in situations when other systems would succeed. In important applications such as healthcare and defence, practitioners prefer not to rely on these systems. The majority of AI/ML-based applications are still in the early stages of development. They have not yet been fully validated and have demonstrated that they contribute to favourable patient outcomes, particularly for nephrology patients. In addition, there is no comprehensive or consistent digitalisation strategy, and insufficient data constitute a limiting factor in these sectors [71].

The diagnosis and detection of progression in AD have been extensively researched, but this research has not yet had a significant impact on clinical practice [51]. This is because most studies rely on a single type of data, such as neuroimaging, and because diagnosis and progression detection are often studied as separate problems. Furthermore, recent studies have focused on improving the performance of ML models without considering their explainability, which limits their usefulness in clinical practice. The design of AI explanation interfaces for non-expert users in medical support scenarios remains an open research question. Although earlier research indicates that explanations could aid users in making appropriate trust assessments of AI systems, the design of AI explanation interfaces for non-expert users in medical support scenarios remains a research challenge [65].

The presence of missing values in both input and output attributes is common in certain domains, leading to ambiguity and incompleteness.

Fig. 10. Ethics challenges for trustworthy AI.

Categorical attributes, which differ from numerical attributes in their non-numerical nature, are often incomplete and ambiguous in these datasets, resulting in decision-making uncertainty [47]. As it has not been thoroughly addressed in earlier research, this paper examines the difficulties of handling categorical attributes. In other instances, expert practitioners in healthcare domains are charged with implementing or utilising AI systems while having little or no understanding of the data these complex systems are built on or how they are constructed [63]. AI's use in healthcare presents ethical challenges, and the study in [14] sought to identify the most important aspects of AI ethics research in healthcare by analysing scientific articles using bibliometric, social network, and cluster-based content analysis. Despite continued use of these techniques in the medical sciences, there is delayed acceptance to improve patient care [18]. Significant obstacles have hampered this work, including the availability of curated heterogeneous datasets for model construction, the human-level interpretability of these models, and the clinical repeatability of these methods.

There is a crucial gap between the development and integration of AI systems in ophthalmic practice even though AI systems can achieve expert-level or even better performance. One study focused on the significance of trustworthy AI to close the gap [52]. Although several AI-based models for COVID-19 disease diagnosis have been created [25], few models bridge the gap between traditional human-centred disease diagnosis and the prospective future of machine-centred disease diagnosis, notwithstanding the promise of AI. It is necessary to improve the output quality of ML and provide safeguards to promote their adoption [26]. Detecting whether an ML model cannot generalise to new unknown instances, which may be from a population distinct from the training population or an under-represented subpopulation, is a vital feature in addition to regulatory and logistical techniques. Lack of transparency in decision-making is a major barrier to the successful adoption of AI-based CAD systems in routine clinical workflow. Although regularly employed XAI methods provide an insight into these completely opaque algorithms, such explanations are typically complicated and incomprehensible to all but the most highly trained AI specialists [30]. XAI is an emerging ML research area that aims to deconstruct how black-box decisions of AI systems are formed. This research topic examines the decision-making metrics and models and develops solutions to explain them directly. Many ML algorithms cannot explain why and how a decision was made [17]. In healthcare settings, the utilisation of intelligence analysis tools such as AI and ML has been a hot area of research, and many medical diagnosis cases were highly detected using AI means, especially those related with the fact that AI can explain itself and which aspects are influencing its decisions. Nevertheless, AI and XAI, like any new phenomena, are not perfect and are prone to several challenges whether it's technical, social, etc. A significant XAI fusion challenge was in identifying proper solutions that can address imprecision, missing and inaccurate information, and explain both the result and the procedure of how it was acquired to a medical expert. That is particularly challenging in terms of developing robust, explainable, and less biased models that require fused orchestrated efforts from different frontier research areas [20]. From a technical perspective, AI explainability is also challenged by encountering multimodal different feature representation spaces where the model not only deals with one type of data but is rather more like spanning images, text, genomics data [67], or even deals with the same type of data but has difficulty in fusing them to generate proper data collection and analytical results [73]. These challenges highlight the need for robust AI systems that can handle complex and diverse data in a socially responsible manner. In summary, Fig. 11 provides a visual representation of the studies reviewed in this section, highlighting their main focus on robustness challenges.

5.3. Recommendations

The recommendations focus on six key areas regarding which policy is required to achieve the general vision for trustworthy AI (see Fig. 12), which are outlined next.

5.3.1. Human-centred values and fairness

The Deep Interpretable Mortality Model introduces a novel ICU mortality prediction model [27]. The interpretation is supported with data and intuitive explanations of how the prediction is achieved by combining the evidence-based medicine concept with earlier research. All treatment decisions are based on evidence, and doctors are supplied with a graphical view of the model. This is essential for the clinical cohort, as future medical interventions are based on the trust chains of the entire prediction process, not just a single digit. It is necessary to evaluate each facet of the problem (see Section 4.2) in the context of the trustworthy use of ML approaches in oncology as a representative study

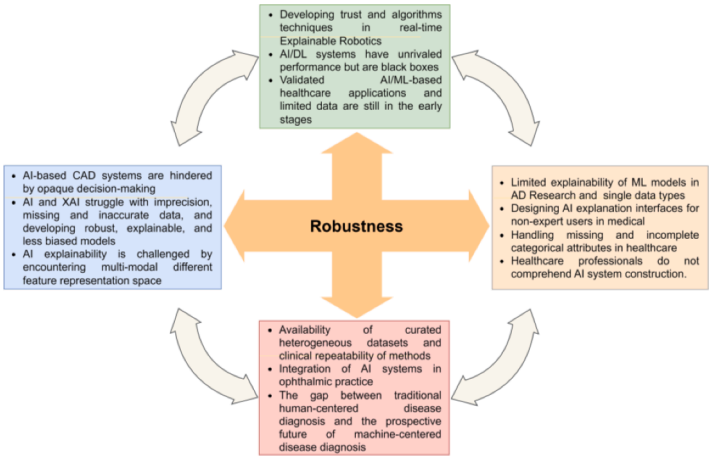

Fig. 11. Robustness challenges for trustworthy AI.

Fig. 12. Recommendation categories of trustworthy AI.

example, with the goal of ensuring utility and enhancing patient care in medicine [18].

5.3.2. Transparency and explainability

The blockchain architecture has been advantageous in tracing data origin and monitoring the training process for model degeneration and dishonest behaviour [55]. The authors believe that this strategy will build confidence between parties, thus fostering worldwide collaboration between parties when delivering robust AI informed by huge amounts of data. The authors of a paper highlighted a crucial component to investigate how FL may provide a future solution for digital health [68]. The ITU and the WHO have launched a global standardisation effort for AI in healthcare, which outlines how standardised benchmarking processes can be used to validate ML/AI models in a consistent and transparent manner [69]. Transparency in healthcare data processes is a condition to trust in healthcare professionals and patients and their adoption of digital tools [13]. A paper recommended a highly accurate and explainable ML model based on a Random Forest (RF) classifier [51]. The authors showed that multimodal RF classifiers could successfully apply to AD detection and progression prediction. Consequently, three strategies for ensuring sustainability in healthcare applications were proposed: privacy, security and explainability of the healthcare data [60].

One study created a User Mental Model, an Expert Mental Model, and a Target Mental Model of explanation, which describe how non-experts and experts comprehend explanations, how their understandings differ, and how they might be integrated [65]. Another paper recommended a novel, XAI strategy that uses graph analysis for feature visualisation and optimisation for COVID-19 blood sample diagnosis [25]. A paper proposed and discussed ideas and provided examples of how explainability and causability are necessary for future AI-based IVD devices to exhibit scientific validity, analytical performance, and clinical performance [44]. In [45], a combination of image processing and DL, supported by XAI, was proposed to ensure reliable glaucoma diagnosis. Another study [17] investigated the current state of XAI and its advancements in healthcare applications. The study presented solutions for XAI that utilise multimodal and multi-centre data fusion, which were then validated in two showcases based on actual clinical scenarios. Quantitative and qualitative analyses can demonstrate the effectiveness of the proposed XAI solutions, indicating their potential for successful implementation in a wider range of clinical inquiries. The use of XAI datasets should be expanded by incorporating additional ones that include multimodal data. Furthermore, modern data fusion techniques should be integrated with XAI to ensure the responsible and ethical utilisation of sensitive data [39].

5.3.3. Technical robustness and safety

A paper presented an efficient recommender system approach for CDD using data mining [32]. In phase two of the CDD strategy, it was suggested that a Unified Collaborative Filtering method be employed to generate accurate medical advice recommendations. The study in [59] suggested a new collaborative early warning system for COVID-19 that utilises blockchain and smart contracts to reduce the risk of decision-making mistakes and improve the effectiveness of early warning [56]. In [53], an interactive multiple-views approach was presented, which outlines the various aspects of the related data, such as therapy guidelines and recommendation systems. It focuses on the relationships between these aspects utilising interaction. The use of trustworthy blockchain applications in healthcare has demonstrated improvements in the quality of information and the security of EHRs, thus reducing errors. Therefore, the authors of a paper suggested that selecting the most effective blockchain model for secure and trustworthy EHRs in the healthcare sector requires a precise mechanism for evaluating the impact of several accessible blockchain models for their features [58]. Moreover, the sensitive data of patients is more prone to potential threats, leading to compromised data security.

In [77], the authors proposed a secure and energy-efficient framework for e-healthcare using IoMT. This framework seeks to reduce energy consumption and ensure timely data delivery to medical experts. From [66], several main contributions can be extracted: (i) construct ethical frameworks for technologies; (ii) examine an effective method for enhancing the link between ethics and legislation; (iii) construct a description of what a more robust ethics-regulation interface may look like, which would facilitate cross-fertilisation amongst political, ethical, and legal approaches; and (iv) examine a 'live' implementation of this ethics–regulation interface, as proposed in Quebec's 'Bill 29′. MAKER is presented for the transformation and imputation of incomplete and ambiguous categorical characteristics [47]. It incorporates the acknowledged uncertainties in the modified input data, which enables a model to understand data constraints and admit questionable predictions throughout the training regime [47].

Another paper recommended building regulatory frameworks for developing, managing, and purchasing AI and HIT systems; establishing public–private collaborations; and ethically and safely using AI in the National Health Service [70]. Additionally, another paper advocated a set of empirically verified parameters for evaluating the trustworthiness and dependability of crowd workers entrusted with assigning behavioural feature tags to unstructured movies of children with autism and matched neurodevelopmental controls [29]. Moreover, a paper promoted developing such multi-stakeholder engagement and the primary action items to be done so that the potential advantages of AI can be realised in real-world ophthalmic situations [52]. In addition, concept activation vectors (CAVs) were used to bridge the gap between human-understandable concepts and those learnt by a DL-based algorithm in its latent space, while Concept Localisation Maps were employed to highlight key concepts in the input space [30].

5.3.4. Accountability

The researchers attained international consensus among specialists to generate and validate material for guidelines on the ethical implications of utilising AI in diverse surgical training environments [64]. This recommendation establishes the groundwork for the deployment of AI applications in surgical training. It should not be up to AI developers, research institutes, or multinational tech businesses to choose how to address these ethical concerns. Relying only on the reliability of businesses and institutions to handle ethical concerns [34]. To promote accountability, it is recommended that cross-domain research teams collaborate to utilise AI for cross-science challenges [20]. Additionally, there is a need for greater cooperation between researchers in AI and medicine globally to ensure that ethical and legal issues are addressed in the development and deployment of AI tools in healthcare. It is also essential to shed more light on ethical and legal issues that arise from using AI, to ensure that the technology is used in an accountable and transparent manner.

5.3.5. Privacy and data governance

The authors suggested in [56] that to improve stakeholder decision-making, data obtained from a connected community, such as Twitter, should be subject to information fusion. This process involves interpreting the data in the context of precision medicine, which can lead to better outcomes. Another paper suggested that a specialised ethical and governance framework for nephrology and AI/ML software should be established [71]. Additionally, the concept of IoT-based e-health that combines AI with homomorphic secret sharing has been advocated [76]. With the use of the medical cloud, the model is intended to improve the maintainability of illness detection systems and facilitate trustworthy communication. AI systems are analysed from the standpoint of a domain practitioner, and essential deployment roles are identified [63]. The unique requirements and duties of each role reveal a tension between openness and privacy that must be resolved so that domain practitioners can intelligently evaluate whether an AI system is suitable for their domain. From the viewpoint of practitioners, the AI

system highlights crucial system deployment responsibilities [14].

It has been suggested that the use of classifier-related measures (such as posterior predicted probability, uncertainty, or conformal prediction) to evaluate pointwise reliability may be biased and misleading [26]. Therefore, it is recommended that the output of the classifier should be evaluated independently of the classifier itself. To further assess this, additional experiments should be conducted on other types of simulated and medical datasets. For multivariate time-series data obtained from the ICU of the University Hospital of Fuenlabrada (Madrid, Spain), interpretable DL and signal processing methods are advocated [33].

5.3.6. *Societal and environmental well-being*

This study field focuses on the development of explanation-providing algorithms that enable robots to converse with people in a trustworthy and human-friendly manner [24]. It is not a simple task to equip a system (e.g., a robot) with rules (data-based or not) to handle all conceivable circumstances for every non-trivial application. Consider a robot that has not been specifically programmed to refrain from making judgements in unexpected settings. In such a scenario, it may act abnormally, perceiving that the robot has taken control of the person [62]. To promote societal and environmental well-being, it is recommended to investigate the performance of applying human segmentation approaches to identify and isolate human beings in images or video footage using fusion techniques [54]. In addition, it is recommended to use different publicly available datasets with XAI to better evaluate the AI method and ensure transparency and fairness in the development of fall detection systems. The use of information fusion can lead to the development of transformative solutions that bridge the gap between research and practical applications in the context of trustworthy medical AI [67].

6. Analysis of characteristics and research gaps

AI can classify skin cancer, identify diabetic eye illness, diagnose chest radiographs, and treat sepsis. Despite promising findings, few clinical AI technologies are implemented in hospitals or utilised by doctors. Current AI techniques are less trustworthy. Existing systems aren't resilient to tiny perturbations or assaults, raising security and privacy problems. Moreover, current methodologies typically favour certain subpopulations or ethnic groupings, and biases may cause unfair forecasts. These issues render present solutions unreliable. Because clinical judgements are so critical, physicians are hesitant to employ these solutions. The most important topics that researchers should investigate are shown in Fig. 13.

This section aims to explore several important points that benefit the scholars by filling the available gaps in future studies. Each subsection highlights a gap and considers the missing piece or pieces in the research literature. Trustworthy AI in healthcare needs to be explored, as it has not been explored sufficiently. However, the following subsections include many significant tables and analyses for outlining the current state-of-the-art trustworthy AI in healthcare literature.

6.1. *The reality of applying trustworthy ai requirements in healthcare studies*

Recently, there have been several discussions on the transparency of AI-based solutions. This makes excellent sense, as when a predictive algorithm detects a high risk for a certain ailment based on the historical data of patients, it is only logical that the patient (and their attending physician) would want to know the basis for this forecast. Some of these algorithms are black boxes, indicating the difficulty in explaining their inner workings. In addition, as this technology relies solely on historical and frequently human-generated data, it may occasionally exhibit various types of prejudice, presenting ethical issues. The concepts of justice, ethics, openness, and dependability are crucial now that AI-based solutions are ubiquitous in MedTech goods. When dealing with the genetic data of patients, this issue might become much more delicate. Before developing completely trustworthy and XAI solutions, it is necessary to identify these characteristics based on the current state of ML systems. The newly published standards of EU on what makes a trustworthy AI technology serve as one evaluation benchmark. According to these recommendations, trustworthy AI should be lawful – according to all applicable rules and regulations; ethical – adhering to ethical principles and values; and resilient – both technically and in relation to the social context of the system. Based on EU rules, there are seven essential elements of trustworthy AI (see Fig. 14), which are as follows:

1) **Human agency and oversight:** which includes the protection of fundamental rights, the involvement of human decision-making, and the need for human oversight
2) **Technical robustness and safety:** which encompasses protection against security threats and the need for backup plans in case of system failure, as well as accuracy, reproducibility, and reliability
3) **Privacy and data governance:** which includes ensuring the quality and integrity of data, respect for privacy, and enabling access to data
4) **Transparency:** which entails the ability to trace and explain how decisions are made and to clearly communicate outcomes to stakeholders.
5) **Diversity, non-discrimination, and fairness:** which includes avoiding unfair bias, ensuring accessibility and universal design, and promoting stakeholder participation
6) **Societal and environmental well-being:** which includes considerations of sustainability and environmental impact, social impact, and democracy

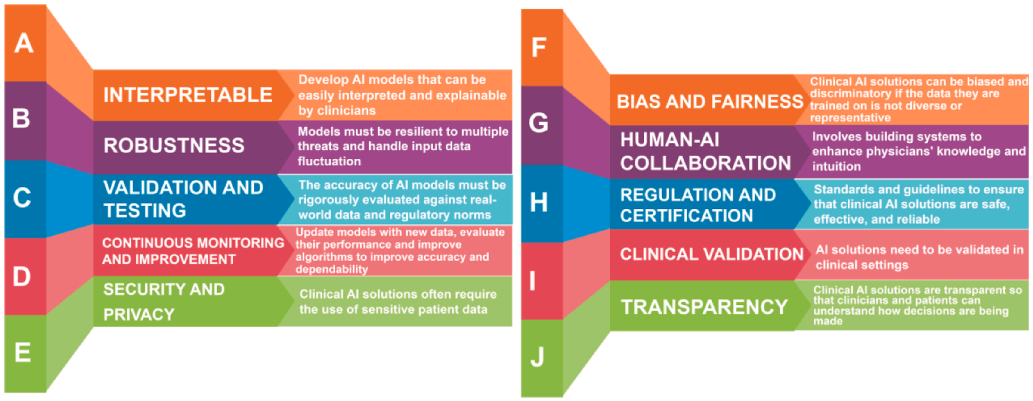

Fig. 13. Topics aimed at addressing the trustworthy issues of clinical AI solutions.

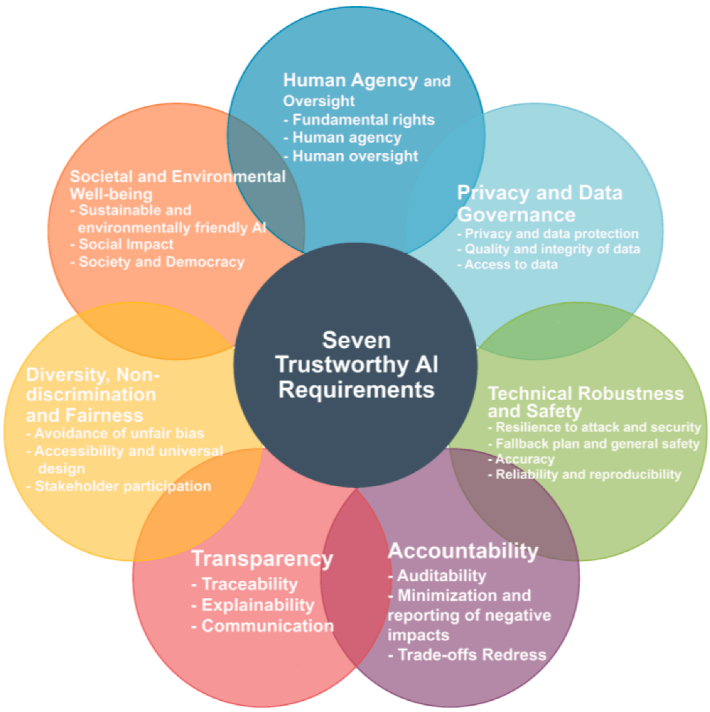

Fig. 14. The seven essential elements of trustworthy AI.

7) **Accountability:** which requires auditability, minimisation and reporting of negative impact, trade-offs, and providing redress when necessary

All seven needs are of similar significance, complement one another, and should be implemented and reviewed throughout the lifespan of the AI system [82]. When implementing criteria across multiple areas and sectors, context and any conflicts between them must be considered. Implementation of these criteria should occur across the full life cycle of an AI system and should be application dependant. While the majority of standards apply to all AI systems, particular attention is paid to those that directly or indirectly impact people. Consequently, some uses (e.g., industrial settings) may be less pertinent. In certain instances, the abovementioned needs are already represented in current legislation. In accordance with the first component of trustworthy AI, it should be repeated that it is the responsibility of AI practitioners to comply with their legal duties in terms of both generally applicable norms and domain-specific legislation.

In addition, it's also worth noting that the term 'trustworthy AI' suggests that the authors are interested in ensuring that the AI used in healthcare is reliable, safe, and ethical. This is an important consideration, given the potential impact of AI on patient outcomes and healthcare practices. Table S1 presents a summary of the frequency of trustworthy AI requirements in healthcare literature. The authors looked at various publications in healthcare literature and identified the most commonly cited requirements for trustworthy AI.

Trustworthy AI is a relatively new area of research, but it has been gaining attention in recent years as the use of AI in healthcare is becoming more prevalent. It is expected that the frequency of research on trustworthy AI requirements in healthcare literature will continue to increase in the future. As a conclusion to Table S1, there are still many gaps in the earlier studies to apply trustworthy requirements. Gaining physicians' confidence is necessary to hasten the use of AI trustworthy requirements in healthcare outcomes. The medical industry will need to develop recommendations for these emerging solutions and reassess the existing regulatory systems that clinicians and patients rely on to ensure that AI trustworthy requirements used in healthcare are evidence-based, accountable, free of bias, and designed to promote equity. This is because technology is constantly evolving, and the medical community must keep up with these advancements, clearly defining roles and responsibilities amongst AI researchers when investigating or developing clinical AI systems. In addition, healthcare organisations and leaders who have the responsibility of validating AI systems in clinical settings should establish trustworthy components towards ethics, evidence, and equity in practice. The successful integration of trustworthy AI into healthcare requires collaboration, and engaging stakeholders early on to address these issues is critical.

6.2. *Healthcare datasets for trustworthy AI*

Datasets play a key role in general AI techniques. The availability of medical datasets that can be applied in trustworthy AI applications is a big issue in the literature. Furthermore, most authors do not publish their dataset and keep it private for various reasons, which has increased the availability issues. Journals continuously insist that dataset statements should be included in the published articles. This means that the published datasets within the articles should comply with the required lawful procedures.

Conversely, the more specific datasets designed for risk bias concerns and solutions are a real concern nowadays. Therefore, more studies should be conducted to make AI techniques more trustworthy based on suitable datasets designed for risk bias concerns and solutions. Table 1 presents important information regarding the dataset availability along with the description, size, or sample of the data and how big data have been considered in studies on AI trustworthiness.

Regarding the dataset description column, the conducted studies paid little attention to the case studies and disease types. The contexts of the datasets used in the literature mostly focused on cell lung cancer, AD, COVID-19, retina, skin cancer, asthma, autism, and fall detection. On comparing the second and fifth columns in Table 1, it can be concluded that some studies considered their datasets to be big data even though they had a small sample size. It is to be noted that datasets with only 422 images, as used in [55], or 804 CT scans, as used in [17], were described as big data. One potential issue is that some researchers

Table 1
Healthcare dataset information used with trustworthy AI.

Ref.	Dataset description	Dataset size	Link of the dataset	Lawfully collected dataset	Considered as big data	Publicly available or private
[32]	The medical dataset has been collected from hospitals in Oman	935 records and 13 attributes	N/A	√	√	Private
[27]	Real-world datasets from MIMIC-III (medical dataset)	Relational database consisting of 26 tables	https://physionet.org/content/mimiciii/1.4/	√	N/A	Public
[55]	422 patients with Non-Small Cell Lung Cancer (NSCLC) had their CT images analysed by an experienced radiologist to link clinical outcome and Gross Tumour Volume (GTV) segmentations with survival.	422 images	https://wiki.cancerimagingarchive.net/display/Public/NSCLC-Radiomics	√	√	Public
[51]	Dataset collected from the Alzheimer's Disease Neuroimaging Initiative (ADNI) utilising multimodal AD	1048 subjects	https://adni.loni.usc.edu/	√	√	Public
[59]	There are 203 positive samples and 406 negative samples in the COVID-19 chest X-ray dataset.	609 images	https://github.com/ieee8023/covid-chestxray-dataset	√	√	Public
[60]	First dataset: It contained a large set of high-resolution retina images taken under a variety of imaging conditions.	Over 25,000 images	https://www.kaggle.com/competitions/diabetic-retinopathy-detection/data	√	√	Public
	Second dataset: Skin Cancer ISIC: The skin cancer dataset included images of malignant and benign oncological diseases, compiled by The International Skin Imaging Collaboration (ISIC), and contains nine classes of skin cancer.	2357 images	https://www.kaggle.com/nodoubttome/skin-cancer9-classesisic	√	√	Public
	Third dataset: It contained a different cough sound diagnosis as a COVID-19 or normal.	Over 1000 records	https://www.medml.net/cough	√	√	Public
	Fourth dataset: It contained a different chest X-ray diagnosis as a COVID-19 or normal.	Over 15,000 images	https://www.medml.net/xray.	√	√	Public
[47]	First dataset: The dataset was utilised to identify possible asthma symptoms in youngsters. It had four categories: sleep disturbance, presence of daytime symptoms, peak expiratory flow rate (the most air that can be expelled from the lungs), and whether physical activity or exercise prompted symptoms.	4827 cases	N/A	√	√	Private
	Second dataset: The information regarding mortgage loans was gathered from Together Financial Services, which is a mortgage lending company located in the UK.	3498 cases	N/A	√	√	Private
[56]	The data was obtained by browsing Chilean Twitter accounts using an algorithm for traversing graph data structures.	1864,788 users from 2008 to 2019	N/A	√	√	Private
[29]	First dataset: YouTube recordings of children with autism and matched neurotypical controls provided a dataset.t	24 publicly accessible videos (6 females with autism, 6 neurotypical females, 6 males with autism, and 6 neurotypical males)	https://doi.org/10.1371/journal.pone.0093533.s001	√	N\A	Public
	Second dataset: It consisted of a diverse representation of worker nationalities compared with Mechanical Turk, which contained workers mostly from the United States and India.	13 multiple choice questions	https://www.microworkers.com/	√	N/A	Public
[25]	This dataset contained de-identified information from hospitalised patients with COVID-19 symptoms who requested the SARS-CoV-2 RT-PCR and further testing while hospitalised.	5644 patients	https://www.kaggle.com/einsteindata4u/covid19	√	√	Public
[26]	Real-world datasets from MIMIC-III (medical dataset)	Relational database consisting of 26 tables	https://physionet.org/content/mimiciii/1.4/	√	N/A	Public
[30]	Three datasets: ISIC2019, Derm7pt, and PH2	6475, 823, 200	https://pubmed.ncbi.nlm.nih.gov/24110966/	√	√	Public
[45]	Three datasets: Drishti-GS, ORIGA−Light, and HRF retinal image datasets	101 retinal image data, 650 retinal images, and 45 retinal images	http://cvit.iiit.ac.in/projects/mip/drishti-gs/mip-dataset2/Home.php and http://www5.informatik.uni-erlangen.de/research/data/fundus-images/	√	√	Public
[33]	This article examined clinical MTS data gathered by ICU patients at Fuenlabrada University Hospital in Madrid, Spain. Data collection spanned 16 years, between 2004 and 2020, with 3470 patients included in the dataset (627 identified as antimicrobial resistant).	3470 patients	N/A	√	√	Private
[17]	Four medical centres in China submitted CT data with identifying information removed to	804 patients	N/A	√	√	Private

(continued on next page)

Table 1 (*continued*)

Ref.	Dataset description	Dataset size	Link of the dataset	Lawfully collected dataset	Considered as big data	Publicly available or private
	safeguard the confidentiality of the information. There were 424 CT volumes from COVID-19-negative patients and 380 CT volumes from COVID-19-positive patients.					
[39]	Two types of input image data were used: CT scan and X-ray images.	Over 20,000 of CT images and 6432 X-ray images	https://www.kaggle.com/datasets/azaemon/preprocessed-ct-scans-for-covid19 and https://www.kaggle.com/datasets/prashant268/chest-xray-covid19-pneumonia	√	√	Public
[54]	The dataset used in the study consisted of a total of 249 videos, comprising 192 videos depicting falls and 57 videos depicting normal activities. The videos encompassed various activities of daily living, such as walking, sitting down, standing up, and squatting.	192 videos	http://le2i.cnrs.fr/Fall-detection-Dataset	√	N/A	Public
[73]	The participants in this study were pregnant women who received prenatal care at two healthcare facilities in Seville, Spain, specifically the University Hospital Virgen del Rocío and the Clinic Victoria Rey. The study was conducted during the period from October to November 2020.	150 women	N/A	√	√	Private

may use the term 'big data' as a buzzword to make their research sound more impressive without meeting the established criteria for big data. This may be a consequence of big data's hype, which has led to confusion about what it really means. Therefore, it is crucial to provide clear justifications for why a dataset is considered big data, taking into account not only the size of the dataset but also the complexity, variety, and velocity. This can help ensure that research studies are trustworthy, accurately labelled, and can help avoid confusion amongst researchers and policymakers about what constitutes big data.

Furthermore, it is commendable that all the authors in the concerned studies mentioned that the datasets used fulfilled the legal requirements. It is crucial that researchers adhere to legal requirements when collecting, processing, and using data, particularly with the growing focus on data privacy and security. This includes compliance with laws and regulations governing data protection, such as the General Data Protection Regulation (GDPR) in the EU, the Health Insurance Portability and Accountability Act in the United States, and similar regulations worldwide.

It is worth noting that some of the datasets in the table are publicly available, such as the NSCLC-Radiomics dataset, the COVID-19 Chest X-Ray dataset, and the Diabetic Retinopathy Detection dataset. Other datasets are identified as not applicable (N/A), which could indicate that the dataset is not available or that it is private and not publicly accessible. While Table 1 provides a list of healthcare datasets for developing AI applications, it is disappointing to see that the number of high-quality, publicly available datasets is limited. Given the potential of AI in improving healthcare outcomes, further investigation and efforts should be made to ensure that more comprehensive and diverse datasets are available to support the development of trustworthy AI in healthcare, especially in areas such as rare diseases or underrepresented populations.

The table highlights the significant role that image categorisation techniques play in medical diagnosis, particularly in the areas of chest and brain imaging using MRI. It is noteworthy that most publications in the table focus on these organs and modalities due to the availability of publicly accessible data. However, it is essential to note that the abundance of data in these areas does not necessarily reflect the diagnostic challenges or the clinical importance of other organs and imaging modalities. Currently, the industry standard for medical imaging involves using classification models and post hoc application of saliency methods for localisation, primarily due to the cost and time required to obtain ground-truth segmentations. This practice highlights the necessity for research into the dependability of such methods in clinical settings, as outlined in [83]. Accordingly, the need for explainability of techniques based on DL is growing in response to the proliferation of methods based on DL, particularly in high-stakes decision-making fields such as healthcare image analysis. For example, saliency approaches were provided to help physicians make more informed diagnoses, which resulted in heat maps highlighting the regions in the medical picture, impacting model prediction [84] (as shown in Fig. 15).

The study mentioned in [84] highlights the importance of evaluating the performance of seven saliency approaches in medical image analysis. The findings suggest that current saliency approaches fall short of human performance and that Grad-CAM shows promising results in localising diseases. This underscores the need for further research and collaboration between medical practitioners, theoretical scientists, and medical imaging specialists to improve the performance of DL approaches in medical image analysis [85]. Additionally, as the use of XAI algorithms becomes more widespread in medical diagnosis, it is important to establish standards that define the minimum sample sizes necessary for specific XAI algorithms. This is crucial given the high cost and time involved in gathering diagnostic imaging datasets and the potential burden that this places on patients. Such standards would ensure that XAI algorithms are effective and efficient in medical diagnosis while minimising the risks and costs associated with gathering and processing large datasets.

6.3. *Mapping the landscape of healthcare perspectives with trustworthy AI*

The most frequent question is: What are the biggest issues and challenges for telemedicine applications when using AI? Researchers agree that they are inadequate telemedicine parity laws. According to the survey conducted by Reach Health, other issues can be considered the top challenges amongst telemedicine programs, such as the lack of EHR integration and medicare reimbursement [58]. However, what is stated in this study about AI trustworthy requirements should also be

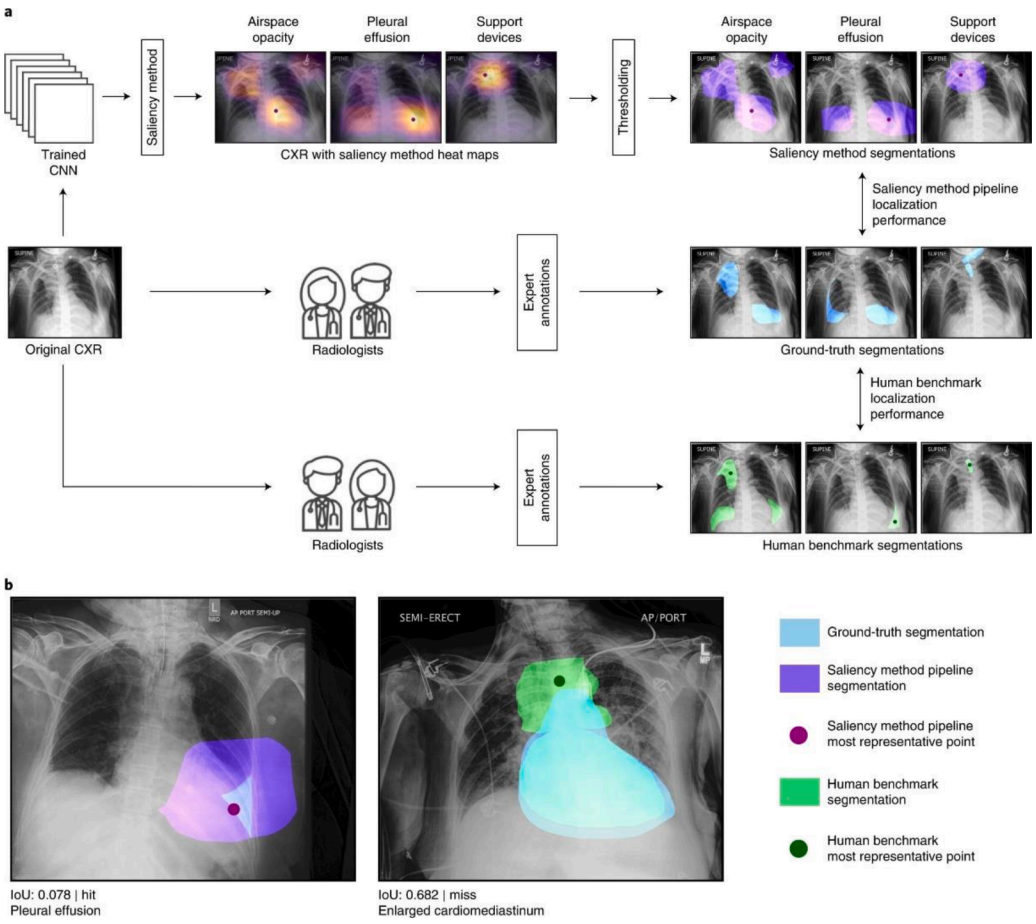

Fig. 15. A conceptual framework for analysing saliency approaches. (**a**) An ensemble CNN trained purely on CXR pictures and pathology task labels is fed a holdout test set CXR image. (**b**) Left: CXR image of pleural effusion with ground-truth and saliency method comments [84].

considered when overcoming the above challenges. Therefore, closing the gap between the telemedicine contexts, diseases. and healthcare perspectives with AI trustworthiness is highly recommended. The authors of this study do not want to be pessimistic, but there is an urgent need for an extraordinary effort on the part of researchers to clarify the picture regarding trustworthy AI in telemedicine.

Furthermore, aging population and rehabilitation are two important terms, and each has a wide field of study. Usually, authors try to apply different AI techniques to support humanity. Additionally, AI trustworthiness has received little attention. To the best of the authors' knowledge, although trustworthy AI has now attracted considerable attention recently, its fulfilment of the requirement is still in the first stages in healthcare systems, which requires careful consideration for developing and enhancing healthcare systems. With a focus on the trustworthiness requirements of AI in healthcare systems, interdisciplinary research for combining AI and IoT perspectives in the context of trustworthy requirements is essential. Several benefits of IoT in healthcare have been reported, such as lower cost, better drug management, and improved patient experience [76]. Conversely, AI in IoT healthcare poses social, ethical, and clinical risks [76]. The lack of

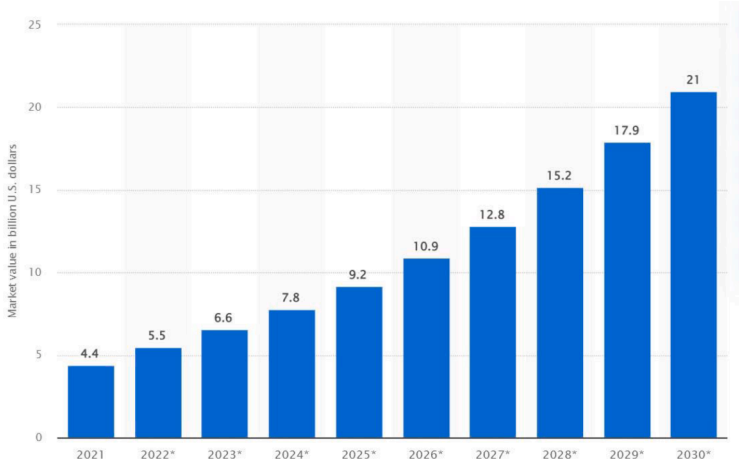

Fig. 16. Global market for XAI in size from 2021 to 2030 [86].

applied trustworthy requirements could lead to the risk of bias and increased health inequalities, as well as vulnerability to hacking and data privacy breaches. Building trusted AI applications using ML systems and the IoT makes smart healthcare services possible. Furthermore, the worldwide market for XAI was estimated to be worth 4.4 billion U.S. dollars in 2021, as reported by NMSC [86] (shown in Fig. 16). It is anticipated that the market will be worth 21 billion dollars in the United States by 2030. In these contexts, XAI methods could potentially benefit from making AI models and aided decision-making systems responsible, which is essential for the applications of these technologies in the healthcare industry.

6.4. AI algorithms for trustworthy AI in healthcare

ML algorithms are one of the most popular procedures of AI in healthcare systems and applications. When accurate ML algorithms are used, while introducing the development model, AI systems can enable fast diagnosis and accurate detection of diseases using medical images and improve patient monitoring. These competencies could mean better, more truthful treatments, better patient outcomes, and lower care costs. The aim of this section is to explore how ML algorithms have been used in the healthcare literature and to determine what metrics have been used. In addition, the activation function in ANNs that determines whether a neuron should be activated or not is a very important feature, and it can determine the output of that node with one input or a set of inputs. This important feature has also been considered with the features used and the number of layers in Table 2.

Table 2 lists various AI applications in the healthcare system, their respective directions (ML, DL, AI), the healthcare system they are applicable to, and the disease type they are focused on. It is worth noting that a number of these AI applications utilise both ML and DL techniques, indicating the importance of combining these two approaches for better results in healthcare. Additionally, the applications are focused on a variety of different disease types. This shows the wide range of areas that AI can be applied to in healthcare and the potential for these technologies to improve patient outcomes and enhance the overall efficiency of the healthcare system [53].

The metrics used to evaluate the performance of ML and DL models vary depending on the application and disease type. A Bayesian network was used by physicians as the metric for the therapy of laryngeal cancer. RF was used for chronic disease diagnosis with metrics including MAE, precision, and recall [32]. For mortality prediction, a combination of logistic regression, RF, XGBoost, and SVM was used, with the metrics including AUROC, AUPRC, ROC, and accuracy [26]. For various diseases such as lung cancer, retinal eye diseases, COVID-19, and skin lesions, CNN or ResNet-18 was used as the method, with accuracy or AUC as the metric. For other diseases such as nephrology, ASD, human lung tumour, mortality of ICU patients, and glaucoma diagnosis, different ML or DL models were used with specific metrics such as accuracy, precision, recall, sensitivity, specificity, F1 score, and AUC. For different applications such as drug use, infectious diseases, and medical data transformation, various methods such as SVM, belief-rule-based models, and wireless body area network were used, with metrics including accuracy, precision, recall, sensitivity, specificity, AUC, and network throughput. The overall picture presented by the table showcases the extensive variety of methods and metrics used in AI-powered healthcare systems for an array of diseases and applications. However, this diversity also highlights the need for continued research in order to enhance and standardise these models to make them more trustworthy and suitable for clinical use. The use of trustworthy AI in healthcare systems is

Table 2
ML healthcare systems-based trustworthy AI techniques.

Ref.	AI direction	Healthcare system	Disease type	Method used	Metrics used
[53]	ML	Clinical decision support system	Therapy of laryngeal cancer	Bayesian network	Objectively judgement by physicians
[32]	Data mining	Recommendation system	Chronic disease diagnosis	Random Forest (RF)	MAE, precision, and recall
[27]	DL	Interpretable prediction system	Mortality prediction	Logistic regression, RF, XGBoost, Support Vector Machine (SVM)	Area Under Receiver Operator Curve (AUROC), Area under Precision-Recall Curve (AUPRC), ROC, accuracy
[55]	ML/DL	Chained distributed ML system	Lung-cancer	CNN/ ResNet-18	Area under curve (AUC)
[71]	ML	Learning Health System initiative	Nephrology	Neural network	Accuracy
[51]	ML	Clinical decision support system	Alzheimer's disease	A two-layer model with RF	Precision, recall, accuracy, F1 score, AUC
[59]	ML	Early warning systems	COVID-19	CNN, LSTM, and BiLSTM	Accuracy
[60]	DL	HIoT-based DL eco-system	E.g. COVID-19	CNN	Accuracy, precision, recall, F1 score, sensitivity, positive predictive value
[52]	DL	Ophthalmic AI system	Retinal eye diseases	CNN/ ResNet-18	Accuracy
[76]	ML	Energy-efficient IoT e-health system	Infectious diseases	Best first search-based AI heuristic algorithm	Energy consumption, packet drop ratio, delivery time, and data leakage
[47]	ML	MAKER system	Asthma	Belief-rule-based models	AUC
[56]	ML	Information fusion system	Drug use	SVM	Accuracy
[58]	Decision-making	Electronic Healthcare Record (HER) system	Any disease with HER	Fuzzy Analytic Analytical Network Process (F-ANP)	Sensitivity Analysis
[29]	ML	ASD system	ASD	Linear regression	Mean average error (MAE)
[25]	ML	XAI	COVID-19	RF	Accuracy, F1 score, sensitivity, specificity, AUROC
[44]	ML/DL	Decision support	human lung tumour	Any ML	Trustworthy criteria
[26]	ML	Prediction	Mortality of ICU Patients	RF	Accuracy, F1 score, Matthews correlation coefficient, Brier score, AUC, AUPRC
[30]	DL	Dermatology system	Skin lesions	DL models	Accuracy, precision, recall, AUC
[45]	DL	Explainable hybrid system	Glaucoma diagnosis	CNN models	Accuracy, sensitivity, recall, specificity, precision, F1 score, AUC
[33]	ML	Early prediction	Antimicrobial multidrug resistance	Recurrent neural networks & linguistic fuzzy systems	Accuracy, specificity, sensitivity, ROC, AUC
[17]	DL	Medical explainable AI	COVID-19	ResNet-50	Accuracy, precision, sensitivity, specificity, AUC
[77]	AI	e-healthcare	Any medical data transformation	Wireless body area network	Network throughput, packets loss, end-to-end delay, energy consumption, link breaches
[48]	ML	Early detection	Gestational Diabetes Mellitus	Logistic regression, SVM, RF, AdaBoost, XGBoost;	AUC, specificity, sensitivity, accuracy

crucial, as medical decisions based on incorrect or unreliable AI results could have severe consequences for patients [17]. Thus, it's important to establish standard metrics and methods for evaluating the accuracy and reliability of AI-based healthcare systems, to ensure that they can be trusted to make informed medical decisions.

Recently, many AI systems have included XAI, a rapidly developing subfield of research within ML that seeks to shed light on the inner workings of AI models and the decision-making processes that occur within black boxes. Its goal is to make AI-based systems more ethical, private, secure, trustworthy, confident, and safe. Conversely, the number of layers of the suggested model is rarely addressed in studies, indicating a significant gap in understanding the mechanism of developing the application and the stages of work of the model. Furthermore, some studies obtain findings but do not analyse the proposed structures in terms of accuracy (i.e., F1 score and other criteria), implying that the researchers do not attain transparency in their work.

6.5. XAI methods in healthcare: theoretical concept and current used

In most cases, model-agnostic explanation techniques can only access the model's output since they approach ML models as if they were black-box functions. Therefore, XAI is an established area in a dynamic community that has created effective techniques to explain and understand sophisticated ML models such as DNNs [87]. The fact that these approaches do not require knowledge about the model's internals, such as the topology, learning parameters (weights, biases), and activation values, enables them to have a wide range of applications and a high degree of flexibility. XAI was invented by the European GDPR (DARPA) to help answer the why questions [88]. This project intends to make it possible for human specialists to comprehend the underlying causes that contribute to a judgement made by AI. Several studies have addressed some important XAI methods in the literature [89,90]. The strategies for explainability are primarily divided into global and local approaches. The global approach illustrates the concept in a general sense, indicating its generic operating principles. The local techniques describe, for each and every data point, how the models are reasoned and the principles that lead to a certain conclusion. Therefore, this section investigates how these XAI methods have been applied in the literature as shown in Table 3. Before that, a brief description of the basic concept for each method has been described in the following:

1. *LIME (Local Interpretable Model Agnostic Explanations)* [91]: The fundamental premise of LIME is that it is possible to clarify a prediction made by a complicated model, such as a DNN, by first fitting a local surrogate model, f_S, whose predictions are straightforward to explain. LIME first creates samples in the neighbourhood N_{xi} of the input of interest x_i and then evaluates those samples using the target model. Finally, LIME attempts to estimate the target model in this immediate vicinity using a straightforward linear function that is straightforward to understand. LIME is applicable locally. The code repository of the LIME can be found at https://github.com/marcotcr.
2. *Anchors* [92]: The fundamental concept behind any black-box classification model is that individual predictions may be explained by locating a decision rule that 'anchors' the prediction to the appropriate degree; this is where the term 'anchors' comes from. The explanations produced are decision rules in the form of IF-THEN statements, which describe areas in the feature space. The predictions are fixed – also known as 'anchored' – to the data piece category that must be explained. Because of this, the categorisation will always be the same regardless of how drastically the other feature values of the data point that are not part of the anchor are altered. Anchors may be utilised locally. For more explanations about the code, the repository can be found at the link https://github.com/marcotcr/anchor.

Table 3
Explainable AI methods and their applications in healthcare.

Ref.	Explainable AI methods	Explanation or justification or aim of those used
[30]	Textual explanations	This study proposes a novel multimodal explanation framework for biomedical image analysis that utilises textual explanations. The classification of dermoscopic images of melanomas is used as a case study to demonstrate the utility of this framework. The proposed framework is evaluated using a variety of datasets, and the study aims to identify human-defined concepts for skin lesion diagnosis within the latent space of the DL model and to locate them on the input image. This allows for providing an intelligible textual explanation of the model's predictions.
[33]	SHAP	This study employs the use of SHAP, a model-agnostic method, in order to develop trustworthy intelligent models for the early prediction of antimicrobial multidrug-resistance. This is achieved through a combination of relevant feature selection and interpretable RNNs. The proposed model is then validated by clinicians, with a focus on the model's understandability and trustworthiness for practical purposes. The generated explanations are automatically constructed from a pool of interpretable rules, which describe the interaction amongst the most relevant features identified by SHAP.
[45]	Class activation mapping (CAM)	The XAI was achieved through the utilisation of CAM, which enables heat map-based explanations for the image processing performed by a CNN, thereby facilitating a reliable diagnosis of glaucoma. The visualisation provided by the CAM-supported CNN model was analysed to determine if the black-box deep learning solution correctly identified locations of disease detection.
[46]	Heatmap, Attribution, Grad-CAM	The authors aimed to achieve targeted goals when utilising XAI techniques, which include trustworthiness and privacy awareness. They also acknowledged the significance of considering the purpose and audience of the model in relation to explainability.
[17]	GAM, LIME, SHAP	Multiple strategies were developed to improve the explainability of artificial intelligence, utilising data from various sources, and validated through two clinical case studies.
[51]	SHAP	The aim of this study is to create an interpretable model that can accurately diagnose Alzheimer's disease and monitor its progression. This approach equips medical practitioners with accurate assessments and a set of justifications for each alternative, thus providing a transparent and understandable decision-making process.
[60]	SHAP, Grad-CAM, LIME	This study suggests that three features of sustainability, namely, security, privacy, and explainability, are crucial for healthcare apps. The effectiveness of these features was evaluated using several IoHT-related linked health applications to demonstrate their usefulness in ensuring the sustainable use of these apps.

3. *GraphLIME* [93]: This approach is similar to LIME in concept; however, it does not follow a linear progression. GNNs are the target of this application. These models, organised in a network structure, can process data that does not conform to Euclidean geometry. The primary responsibilities of GNNs include the categorisation of graphs and nodes, as well as the prediction of links. *GraphLIME* is applicable locally and globally. The experimental code of GraphLIME is located in the following repository https://github.com/WilliamCCHuang/GraphLIME.
4. *LRP (Layer-Wise Relevance Propagation)* [94]: This technique is a propagation-based explanation that needs access to the topology, weights, activations, and so on that are included within the model's internals. In these specific circumstances, the knowledge

discovered about the model enables LRP to simplify to address the explanation issue. LRP does not explain the prediction of a DNN in a single step; instead, it uses the network structure and redistributes the explanatory elements (also referred to as relevance R) layer by layer, beginning with the model's output and moving on to the input variables. In certain orders, LRP is a worldwide regulator, while in others, it is a local regulator. The code of LRP is located in the repositories https://github.com/chr5tphr/zennit and https://github.com/albermax/innvestigate.

5. *Deep Taylor Decomposition (DTD)* [95]: Explaining the decisions made by a neural network through its breakdown is the goal of the DTD methodology, which is a propagation-based explanation technique closely tied to the LRP method. It employs (first-order) Taylor expansion to determine the bottom layer items' percentage or importance. Then the neural network's output is redistributed layer by layer to the input variables. This redistribution of the function value takes place while it is performing the process. DTD is applicable locally and globally. The two repositories that contain DTD code are https://github.com/chr5tphr/zennit and https://github.com/albermax/innvestigate.
6. *Prediction Difference Analysis (PDA)*: PDA is based on the premise that the significance of a feature xi may be approximated by seeing how the outcome of a prediction shifts when the feature in question is concealed from view. The methodology is predicated on the earlier concept introduced by [96], in which, for a certain prediction, each input feature is allotted a significance value about a particular class c. PDA is applicable locally. The following link repository explains the information about the PDA code https://github.com/lmzintgraf/DeepVis-PredDif.
7. *TCAV (Testing with Concept Activation Vectors)* [89]: TCAV is a classification technique using neural networks based on determining how strongly a concept, such as colour, affects categorisation. The TCAV approach is based on CAVs, which explain how brain activations affect a user's idea. In order to compute a CAV of this kind, it is necessary first to gather and merge two sets of data. One set of data will comprise images that are representative of the idea. In contrast, the other data set will consist of images that do not contain this concept. The combined dataset is employed to train a logistic regression model to determine if an image has the concept. User-defined neural network layer activations are categorisation model features. The CAVs are the logistic regression model's coefficients after the model has been applied. TCAV is applicable locally and globally. The following link provides information about the code repository: https://github.com/tensorflow/tcav
8. *XGNN (Explainable Graph Neural Networks)* [89]: Because it is a post-hoc approach that works on the level of the model, the XGNN method does not make an effort to offer specific example-level explanations. XGNN is a classification system that was developed specifically for the purpose of graph tasks. The graphs that were returned are the ones that were the most representative of the GNN decision. Additionally, the returned graphs often contain a certain attribute that is entrenched to make the validation feasible. The only technique that offers explanations at the mode level for GNN designs is XGNN. Since the search for the explanation graph is non-differentiable, RL is a suitable approach for searching. This is an effective method for overcoming the challenge of non-differentiation since the training of GNNs includes aggregations and combinations of data. XGNN is applicable locally and globally. The pseudocode of XGNN can be found in the repository https://github.com/divelab/DIG/tree/dig/benchmarks/xgra%20ph/supp/XGNN.
9. *SHAP (Shapley Values)*: The principle given in this approach applies also to the methods offered in (10, 11, and 12). SHAP is applicable both locally and globally. The explanations for the model f at a specific individual point x are the focus of this strategy. They are founded on a value function known as eS, in which S represents a subset of variable indices such as 1,…, p. This function is often stated as the predicted values of a conditional probability distribution in which all parameters in a subset of S are conditioned. In this case, the subset S contains all the variables. The term 'expected value' is most frequently used for tabular data. The code of the method is located at https://github.com/slundberg/shap.
10. *Asymmetric Shapley Values (ASV)* [97,98]: SHAP values are symmetrical. When it comes to the process of explaining the model, ASV makes it possible to employ extra knowledge about the causal links that exist between variables. A cause–effect relationship that is described in the form of a causal graph makes it possible to redistribute the attribution of variables in such a way that the source variables have a greater attribution, which has an effect not only on the other dependant variables but also on the model predictions. The casual graph is simplified to a collection of vertices not associated with one another in SHAP values, which is a specific case of ASV values: https://github.com/nredell/shapFlex
11. *Break-Down* [99]: Explanations in the form of additive contributions can be generated using methods such as SHAP. On the other hand, these methods are frequently used in analysing complicated models, which are frequently not additive: https://github.com/ModelOriented/DALEX.
12. *Shapley Flow* [100]: In the same way as ASV does, Shapley Flow enables the utilisation of the dependence structure between variables in the explanation process. A causal graph, much like the one used in ASV, is used to describe the link. On the other hand, in contrast to ASV and other approaches, attribution is not given to the nodes (variables) but rather to the edges (relationships between variables). A connection between two nodes in a graph is important if removing that edge would shift the model's predictions. The extra quality that the edge attribution possesses is that the bounds maintain the traditional Shapley values for each possible explanation. The ASV approach is the one that corresponds to the border of explanation that is the most severe. In order to ascertain the attributions for each edge in the causal graph, the Shapley Flow approach is utilised: https://github.com/nathanwang000/Shapley-Flow
13. *Textual Explanations of Visual Models*: Several different ML models are used to generate textual descriptions of images. These models consist of two components: one that analyses the input pictures (usually a CNN) and another that learns a suitable text sequence (typically an RNN). This allows the models to generate textual descriptions of images (RNN). These two components work together to produce picture descriptive phrases based on the assumption that a classification assignment has been completed satisfactorily. The first benchmark datasets that included picture descriptions were already in existence by 2014; these were the Microsoft COCO datasets [101]: https://github.com/LisaAnne/ECCV2016
14. *Integrated Gradients* [102]: This is a common method utilised for understanding DNNs and other models that may be differentiated. It is a method that holds up well in theory and is built on two highly desirable features: sensitivity and implementation invariance. It makes effective use of gradient information at α chosen few spots while being computationally efficient. A model is said to have high sensitivity when it assigns nonzero weights to all of its inputs and baselines that differ in at least one characteristic yet provide distinct predictions. Implementation invariance states that attributions for two models must be the same if those models operate in the same way or are functionally comparable. Even though these two axioms have a very natural sound, it turns out that many different systems of attribution do not have these features. In particular, when a model's predictions have been

flattened for a particular point of interest, the gradient in that point of interest will zero out. As a result, it will not convey any information that helps explain the explanation. Integrated Gradients is applicable locally: https://github.com/ankurtaly/Integrated-Gradients.

15. *Causal Models* [103]: By incorporating counterfactuals into their analysis, structural causal models are sometimes seen to be an extension of Bayesian Models of the RL world. It considers the events that would take place or the states of the environment that would be reached depending on the RL agent's actions. The ultimate objective of every RL agent is to maximise a long-term reward, and the explanation offers a series of causal chains leading up to the condition in which the reward is received. In order to comply with the explanation satisfaction conditions outlined by the Likert scale, the researchers make it a point to maintain the explanations in a minimally complete state by removing some of the intermediate nodes in the causal chains. This is done in order to ensure that the explanations meet the criteria. In order to compute the counterfactual explanation, one must compare different causal chain routes that involve acts that the agent did not choose. The provided counterfactual explanation is composed solely of the disparities between the causal chains. This was done so that the explanation could remain as straightforward as feasible: https://github.com/topics/causal-models.
16. *Meaningful Perturbations* [104]: Meaningful Perturbation is a great adaptability strategy that can apply itself to any ML model immediately. The technique can also be understood regarding rate distortion from a different vantage point. Compared to strategies dependant on propagation, such as LRP, the Meaningful Perturbations method is significantly more demanding in terms of the amount of computing power it requires. Causal Models is applicable locally: https://github.com/topics/causal-models.
17. *EXplainable Neural-Symbolic Learning (X-NeSyL)* [105]: Neuro-Symbolic approaches use human knowledge for tasks such as concept acquisition and provide more interpretable output, such as mathematical equations or domain-specific languages. One application of neuro-symbolic methods is in the ground of AI (DSL). This strategy's approach might be thought of as one that emphasises explainability through design. That implies that every stage of the training process is checked to ensure that the final outcome can be interpreted properly. This is not a haphazard approach; rather, the training incorporates a loss to direct the neural network towards answers that have a structure comparable to that of a human expert. In addition, SHAP values offer intermediate feature relevance findings that are not difficult to comprehend. X-NeSyL is applicable locally and globally. The code explanation for this method can be found at https://github.com/JulesSanchez/MonuMAI-AutomaticStyleClassification and https://github.com/JulesSanchez/X-NeSyL.

As seen in Table 3, there are seven studies that discuss the explainability of XAI in their healthcare applications. These studies utilise various XAI methods, including Textual explanations, SHAP, CAM, Heatmap, Attribution, Grad-CAM, GAM, and LIME. These methods are applied to various healthcare tasks, including Melanoma classification, early prediction of antimicrobial multidrug-resistance, glaucoma diagnosis, AD diagnosis and monitoring, and evaluating the usefulness of sustainability features in healthcare apps. In [30], a multimodal explanation framework was used along with textual explanations to present the classification of Melanoma from dermoscopic images. In [33], SHAP was used to design trustworthy intelligent models for the early prediction of antimicrobial multidrug-resistant and validated by clinicians. In [45], CAM was used to ensure trustworthy decision-making for diagnosis of glaucoma. In [46], it was aimed at improving XAI approaches, including trustworthiness and privacy awareness. In [17], methods to improve XAI using data from multiple sources were developed and validated in two case studies. In [51], it was aimed to create a model that could accurately diagnose AD and be open to interpretation. In [60], the usefulness of three features of sustainability for healthcare apps using several IoHT-related linked health applications was evaluated. To sum up, there is a lack of real contributions in the literature on how XAI interprets the development of models in healthcare.

6.6. Trustworthy AI objectives in healthcare services

Several studies have focused on the aims of trustworthy AI in visible healthcare [106–108]. Therefore, this section focuses on the four main objectives of trustworthy AI in the context of healthcare services:

6.6.1. First objective: to improve the patient experience

Patients' rights are protected, and they have the ability to participate in making decisions on the use of AI in their treatment. Trustworthy AI findings lead to improvements in patient's happiness, quality of life, and clinical outcomes. In addition, patient care is enhanced by accelerating medication development and allowing efficient administration and management of trusted AI healthcare systems.

6.6.2. Second objective: to promote the population's health

AI healthcare tackles high-priority clinical requirements and promotes health equity by minimising health inequalities founded on historical and present injustices and prejudice and by assisting all patients, regardless of characteristics such as identity or socioeconomic standing.

6.6.3. Third objective: to reduce cost

The potential of how AI systems can help and damage the healthcare industry must be considered via oversight and regulatory mechanisms. Pay and cover following applicable rules and regulations give sufficient levels of providing a standard and proof of high quality and work to make the service more affordable and accessible.

6.6.4. Fourth objective: to improve the working lives of healthcare providers

In the field of medicine, doctors are actively engaged in the research, development, and deployment of AI technologies that enhance their capacity to provide patients with scientifically validated, high-quality medical treatment that also boosts the patients' overall health. The obstacles to implementation, such as a lack of knowledge of AI and difficulties around responsibility and payment, should be eliminated and solved.

6.7. Justification aspects for researchers to consider XAI in healthcare

AI algorithms have been widely utilised in the healthcare industry since 2017 and have shown the ability to analyse and understand large amounts of medical data that may be hard or infeasible for humans to understand [1]. One of the key elements of healthcare AI is obtaining informed consent from patients, which involves shared decision-making between doctors and patients, with the final decision being made by the patient. For healthcare AI to be applied, it is important that patients are fully informed about its capabilities and how it works. In order to meet this requirement, research has focused on developing XAI systems for healthcare professionals that can enhance their reasoning and decision-making processes [109]. A new XAI phase (section) should be applied in future studies for estimating feature importance levels on the trained model. Fig. 17 presents the comparison of two models trained for the detection of shoulder abnormalities in this work. Both models correctly predicted the image based on their confidence values. However, the heatmap reveals that the first model is biased and inaccurate, failing to detect the region of interest indicated by the red circle. In contrast, the second model accurately identified the region of interest with a high confidence value. This example highlights the importance of

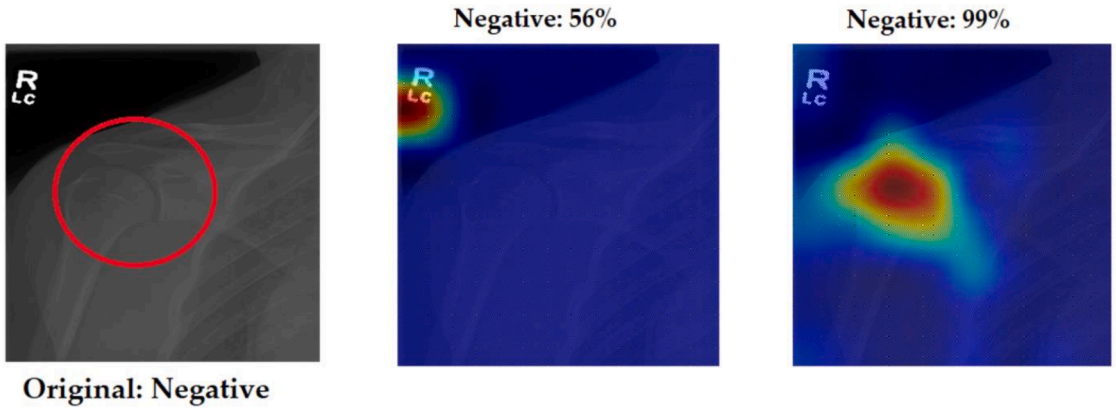

Fig. 17. Grad-CAM visualisation of two DL models.

explainability, as even a model with correct confidence values may not be trusted. By considering explainability, we can improve the results before moving on to the deployment stage. Therefore, several motivational aspects for researchers to apply XAI in future studies are summarised next.

- ✓ **To discover**: Asking for explanations is an effective pedagogical strategy for acquiring knowledge and understanding the underlying task. Explainability is a powerful tool that enables researchers and practitioners to gain new insights and verify the accuracy of an AI system, resulting in more reliable results. It is anticipated in the future XAI algorithms will enable domain experts to identify previously unseen theories and laws in a wide range of fields, such as physics, biology, medicine, and nanoscience. This information can be used to develop a more nuanced understanding of how AI algorithms operate under different learning strategies, data formats, architectural designs, and parameters.
- ✓ **To be reasonable**: There has been much debate in the literature about AI solutions producing biased or unequal results, highlighting the need for explanations to ensure that decisions made by AI algorithms are not biased. Providing explanations for AI-driven results typically means offering reasons or justifications for a particular conclusion, rather than describing the internal mechanisms or thought process behind the decision-making method. In this context, XAI can be used to provide the necessary information to justify results, particularly in cases of unexpected decisions. It also ensures that there is a transparent and verifiable process to support decisions as being reasonable and fair, which can lead to increased trust in the decisions. Additionally, in the future, AI algorithms may be required to provide justifications in order to comply with regulations such as the General Data Protection Regulation's 'right to explanation'.
- ✓ **To control errors**: The purpose of explanations is not just to justify the results of AI algorithms. They can also help prevent errors. In fact, having a better understanding of how a system operates can provide insights into potential vulnerabilities and flaws and enable the rapid identification and correction of mistakes in low-stakes situations. This knowledge can be used to create a set of control rules that can enhance the management of AI-based applications.
- ✓ **To improve the performance**: The necessity to continuously enhance the effectiveness of AI algorithms is another justification for explaining them. XAI algorithms are particularly effective in this regard, as they allow users to understand the reasoning behind the outputs and how to make them better. This can create a collaborative environment between machines and humans, leading to improved learning performance in dynamic and ever-changing environments.

6.8. *Information fusion for healthcare*

In healthcare, information fusion is a technique to combine various types of data to produce reliable and precise decisions. This includes combining data from EHRs, medical devices, research studies, and other resources. The purpose of information fusion is to enhance the outcome for patients by giving healthcare providers a more detailed and precise understanding of the patient's condition [110]. One of the ways to accomplish data fusion is by utilising robust algorithms and methods for ensuring that the data is of high quality and secure. Information fusion is one of the ways to increase the accuracy and reliability of decisions by combining information from multiple modalities. This process can be used to make a final decision more robust and trustworthy for a given case, as shown in Fig. 18. In addition, different fusion techniques are discussed for improving the accuracy and reliability of health data analysis as shown in Table 4.

There are several types of information fusion in healthcare, including the following:

1- **Data Fusion:** This technique is based on fusing data from multiple sources in order to enhance the quality and accuracy of the final decision [111]. Data can be combined from various sources, such as sensors, devices, or other sources. By accomplishing that, the combined data will be a more comprehensive and accurate representation. This technique can be adopted in several applications including healthcare, finance, agriculture, and security. In healthcare, data fusion can be applied in several ways to enhance patient outcomes and decision-making. These include integrating data from different medical devices to produce a more detailed view of a patient's health status.

2- **Feature Fusion**: This technique is based on extracting features using two or more ML algorithms such as CNN and then fusing the extracted features, which will be used to train ML classifiers as shown in Fig. 19.

Feature fusion can be achieved by concatenating, averaging, or combining the features in some other way. Feature fusion aims to produce a more robust and informative representation of the data. By accomplishing that, this technique can enhance the performance of ML classifiers [112]. The feature fusion technique has different approaches including early fusion and late fusion. Early fusion fuses the features before the data is processed by the model. On the other hand, late fusion combines the features after the data has been processed. The choice of approach will rely on the nature of the data and the desired task. In healthcare, feature fusion can be utilised to enhance the performance of ML classifiers and decision-making. The trustworthiness of the final decision can be improved in different ways including (a) combining features from different imaging

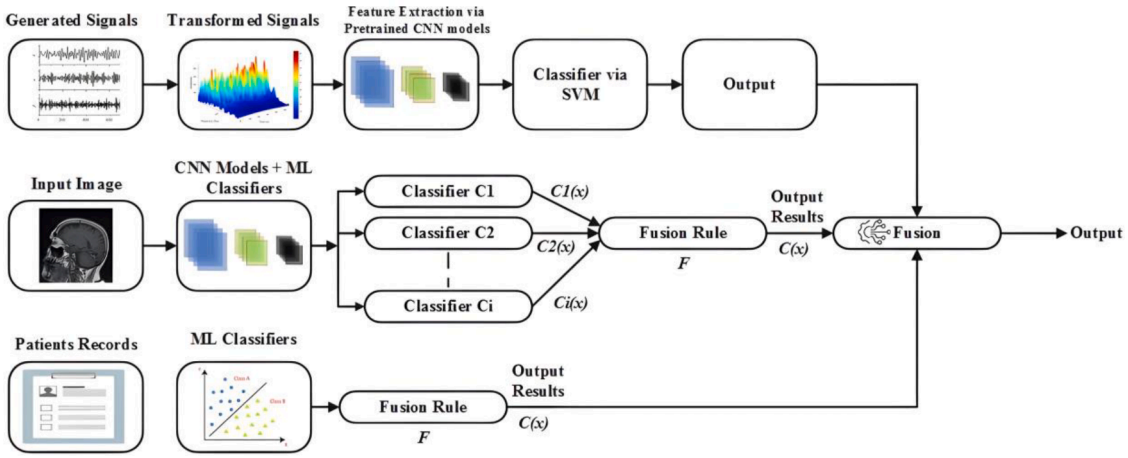

Fig. 18. Example of multimode fusion for a single case.

Table 4
Fusion techniques used in healthcare literature.

Ref.	Application	Source data	Fusion type	Methods used for data fusion	Performance metrics used to evaluate the data fusion method	Results of the data fusion method	Fusion techniques
[25]	Healthcare	Multiple sources	Image fusion	Exploits graph analysis	Improve diagnosis accuracy and decrease the execution time	Performance scores or error rates	Neural networks
[26]	Healthcare	Multiple sources	Data fusion	Maximum value of the probabilities	Develop an anti-diabetic drug failure classifier	Performance scores or error rates	Neural networks
[51]	Healthcare	Alzheimer's Disease Neuroimaging Initiative	Image fusion	Wrapper methods	AUC, precision, recall, accuracy, and F1-score	The system maximises the interpretability of the underlying models and pays special attention to explainability. Multimodal fusion strategies: late fusion and early fusion	Cognitive scores, neuropsychological battery, genetics, lab test, demographics, positron emission tomography, magnetic resonance imaging
[47]	Healthcare and finance	Multiple sources	Data fusion	Dempster-Shafer's method	AUC value	It is applied to combine two or more categorical attribute to reduce the dimensionality of transformed numerical features.	Dempster-Shafer's method
[56]	Social networks	Twitter	Data fusion	Ml	Accuracy, precision, recall	Performance scores or error rates	Neural networks
[59]	Healthcare	Chest X-ray and CT	Image fusion	Neural networks	Accuracy, recall rate, F1	Performance scores or error rates	Medical & Social Fusion Warning Smart Contract

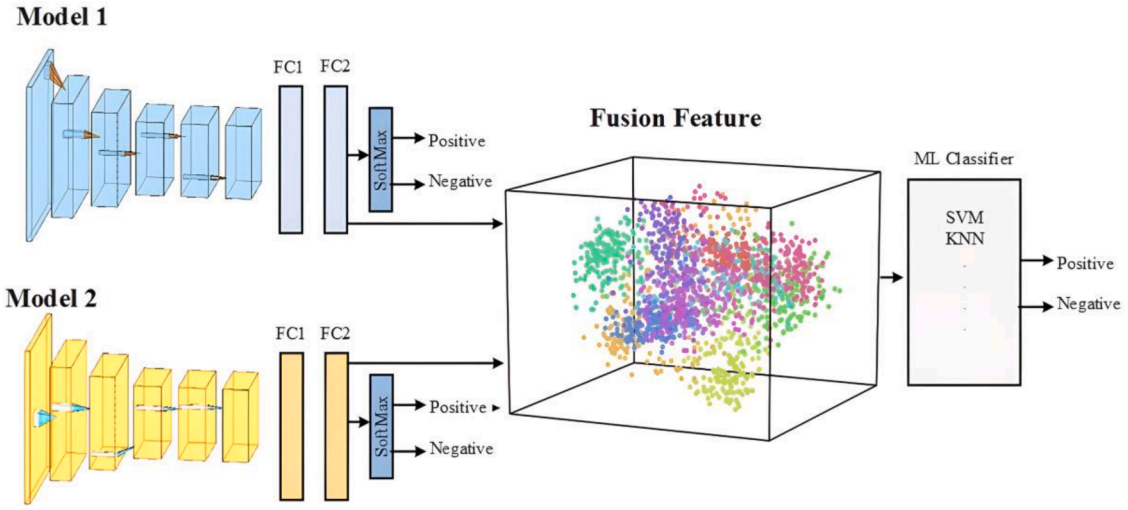

Fig. 19. . Example of feature fusion.

modalities to enhance the accuracy of disease diagnosis, (b) using multi-view fusion to combine features from multiple views of the same data to improve the accuracy of disease diagnosis, and (c) combining features from time-series data to improve the accuracy of disease diagnosis and treatment monitoring.

1- **Image Fusion:** This technique is based on combining multiple images to generate a composite image that is more informative and useful than individual images as shown in Fig. 20. The combined images can be captured from different viewpoints, sensors, or at different times [113]. This technique is necessary for healthcare applications including medical imaging, radiology, and pathology [114]. For example, in medical imaging, image fusion can be utilised to combine data from MRI and CT scans, which can assist in the detection and diagnosis of different diseases including detection of fractures. Another example is where image fusion was used to combine the output of different segmentation algorithms to produce a more informative output [114].
2- **Decision Fusion:** This technique is a way of combining the decisions or outputs of multiple algorithms for a more accurate and reliable final decision [115]. There are several methods to achieving decision fusion such as majority voting, weighted voting, maximum, median, or other statistical methods. In healthcare, decision fusion can be adapted to combine the decisions or outputs of multiple algorithms such as DL models to enhance the robustness of the final decision. Another example is combining the results of multiple diagnostic tests to improve the accuracy of disease diagnoses such as a blood test, a CT scan, and a biopsy to diagnose cancer.
3- **Multimodal Fusion:** This is a technique that combines information from various types of data, which is also known as modalities. This technique helps to produce more accurate and reliable decisions [116]. Multimodal fusion aims to build a stronger and more informative representation of the data, which enhances the performance of ML models and decision-making processes. Several examples of multimodal fusion in healthcare include (i) combining multiple imaging techniques, such as MRI and CT scans, to enhance the accuracy of disease diagnosis, (ii) merging clinical information with MRI to form the final decision, and (iii) combining features from EHRs with patient-reported outcomes to enhance the accuracy of disease diagnosis and treatment planning.
4- **Hybrid Fusion:** This technique is defined as a way of merging multiple images from different sources, such as spatial, spectral, and temporal images [117]. This technique will allow for building a new image that provides more information about the scene than the individual images. This technique can be used in various healthcare tasks, including medical imaging, radiology, and pathology. Hybrid fusion can be utilised to combine data from MRI and PET scans, which can aid in the segmentation and diagnosis of tumours or other diseases.
5- **Temporal Fusion**: This technique combines multiple images of the same scene captured at different points in time [118]. The aim is to generate a new image that contains more information and insights than individual images. This technique can be used in various healthcare tasks, including medical imaging and radiology. An example of this technique is in medical imaging where temporal fusion can be utilised to merge data from dynamic imaging modalities, such as ultrasound and PET scans. By doing so, it can assist in the identification and diagnosis of diseases. In pathology, temporal fusion can be applied to merge data from multiple staining techniques, which can enhance the precision of tissue diagnosis over a period of time.

Table 4 lists various data and image fusion studies, highlighting the diversity of approaches used in this field. The studies listed in the table cover multiple applications and use different data sources, methods, and performance metrics. The results of these studies indicate that other fusion techniques can lead to varying performance outcomes in various applications. For example, in the study in [25], graph analysis for image fusion in healthcare leads to improved diagnostic accuracy and decreased execution time, as indicated by performance scores or error rates. Similarly, in the study identified in [26], data fusion using the maximum value of probabilities developed an anti-diabetic drug failure classifier, with performance scores or error rates used as the performance metrics.

In another study [51], image fusion using wrapper methods in healthcare was used to diagnose AD, maximising interpretability and paying particular attention to explainability. This study evaluated the performance using metrics such as AUC, precision, recall, accuracy, and F1-score. Similarly, in the study in [47], data fusion using Dempster--Shafer's method was applied in healthcare and finance to combine two or more categorical attributes and reduce the dimensionality of transformed numerical features, with the performance evaluated using the AUC value. In conclusion, the results of these studies indicate that different fusion techniques can lead to varying performance outcomes in other applications and that the choice of technique should be based on the application's specific requirements.

7. Conclusion

The use of AI in healthcare has been driven by various economic, political, and ethical considerations, which have helped mitigate AI's potential harms in this field. By addressing these factors, a reasonable level of control has been established over the negative impacts of AI in the healthcare landscape. However, keeping up with the findings of the previous studies on the trustworthiness of AI in healthcare is still a challenge. The coherent analysis outcomes of the previous studies on assessing quality, bias risk, and data fusion have been reorganised into seven categories in one taxonomy. In these contexts, ethical concerns in

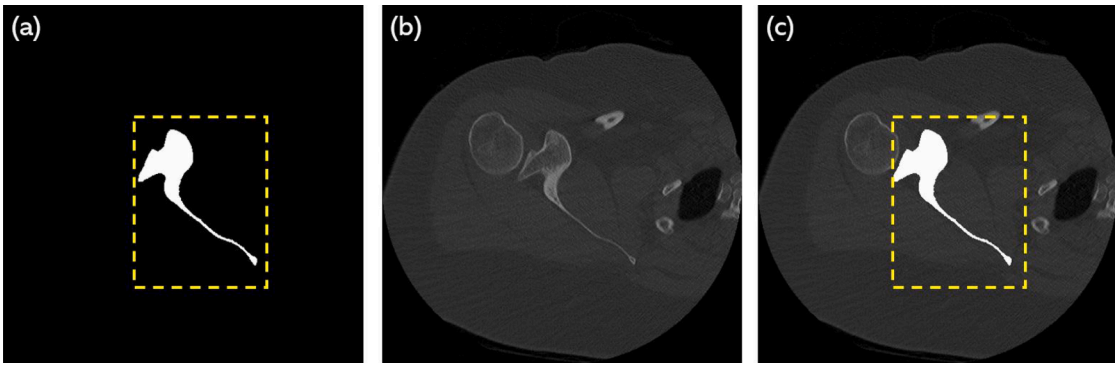

Fig. 20. . Example of image fusion which can make the diagnosis more accurate and efficient: (a) the mask representing scapula segmentation, (b) the original axial CT scan, (c) the scapula mask segmentation fused with the original CT scan.

human–robot interactions have been resolved to make healthcare robotics more trustworthy. Furthermore, previous studies have suggested trustworthy methods for AI applications to increase prediction accuracy for disease diagnosis and treatments. The analysis of earlier studies confirmed that using XAI, such as the 'spread of relevance' at the layer level, helps to highlight and visualise the relevant parts of the inputs and representations in a neural network, leading to a specific outcome. To make informed and justifiable decisions, by using trustworthy and interpretable AI methods, combined with data fusion, is necessary to provide decision support. Furthermore, incorporating digital advancements in the healthcare sector has helped resolve many difficulties and emergencies by incorporating data fusion techniques. This research has shown that the development of trustworthy AI applications, combined with information fusion, will amplify the benefits of digital technologies in healthcare, leading to improved disease diagnosis, more effective treatment identification, and enhanced quality of healthcare services. However, the findings of this study confirm that disregarding the concerns of trustworthiness and bias risk in AI applications leads to a decline in the quality of medical services. This study revealed that many claims open doors for future studies to enrich the literature, encouraging researchers to find more practical implications for practitioners and policymakers regarding the trustworthiness of AI applications in healthcare. Many state-of-the-art theories take into account privacy, safety principles, and ethical issues to enhance the trustworthiness of healthcare technologies by making it legitimate, morally responsible, and robust. This systematic review highlights the opportunities for future studies to expand upon the existing literature and uncover practical implications for using AI in healthcare. The state-of-the-art in this field considers issues such as privacy, safety, and ethics to enhance the trustworthiness of healthcare technologies, making them legitimate, morally responsible, and robust. Therefore, this approach allows new changes in the medical sector by claiming that the quality of information, service, and system is insufficient to accept AI applications in healthcare without trustworthiness. Similarly, normative, behavioural and efficiency beliefs are not considered worthy of investigation while neglecting the legitimacy, ethics, and robustness of AI in healthcare. In practice, it should be mandatory for policyholders to encourage their authorities to test AI applications for trustworthiness prior to their use. Moreover, strict regulations imposed by healthcare decision-makers make it crucial for developers to continually expand and enhance the reliability and robustness of their software. Thus, it should be enforced that practitioners enter into contracts with AI companies that might have a vision of legitimacy, ethics, and robustness in healthcare, which would increase the possibility of using trustworthy AI systems and applications. Future works should aim to evaluate the impact of automated and declarative ML techniques on the quality of medical services and patient outcomes and to investigate the limitations and challenges of implementing these techniques in health data fusion.

Ethics approval

Not applicable.

Consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

Not applicable.

Code availability

Not applicable.

Authors' contributions

The authors contributed equally to this work.

Funding

The authors would like to acknowledge the support received through the following funding schemes of Australian Government: Australian Research Council (ARC) Industrial Transformation Training Centre (ITTC) for Joint Biomechanics under grant IC190100020. The authors also would like to acknowledge the support received through the QUT ECR SCHEME 2022 and the Centre for Data Science First Byte Scheme, The Queensland University of Technology.

Authorship conformation

1) All authors have participated in (a) conception and design, or analysis and interpretation of the data; (b) drafting the article or revising it critically for important intellectual content; and (c) approval of the final version.
2) This manuscript has not been submitted to, nor is under review at, another journal or other publishing venue.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Laith Alzubaidi reports financial support was provided by The Queensland University of Technology.

Data availability

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

References

[1] K.H. Yu, A.L. Beam, I.S. Kohane, Artificial intelligence in healthcare, Nat. Biomed. Eng. 2 (10) (2018) 719–731, https://doi.org/10.1038/s41551-018-0305-z.
[2] A.S. Albahri, et al., IoT-based telemedicine for disease prevention and health promotion: state-of-the-art, J. Netw. Comput. Appl. 173 (2021), 102873, https://doi.org/10.1016/j.jnca.2020.102873. Jan.
[3] G. Rong, A. Mendez, E. Bou Assi, B. Zhao, M. Sawan, Artificial intelligence in healthcare: review and prediction case studies, Engineering 6 (3) (2020) 291–301, https://doi.org/10.1016/j.eng.2019.08.015.
[4] J. Amann, A. Blasimme, E. Vayena, D. Frey, V.I. Madai, Explainability for artificial intelligence in healthcare: a multidisciplinary perspective, BMC Med. Inform. Decis. Mak. 20 (1) (2020) 1–9, https://doi.org/10.1186/s12911-020-01332-6.
[5] E.C. Hayden, The automated lab, Nature 516 (729) (2014) 131–132, https://doi.org/10.1038/516131a.
[6] J. Santamaría, O. Cordón, S. Damas, A comparative study of state-of-the-art evolutionary image registration methods for 3D modeling, Comput. Vis. Image Underst. 115 (9) (2011) 1340–1354, https://doi.org/10.1016/j.cviu.2011.05.006. Sep.
[7] R.C. Deo, Machine learning in medicine, Circulation 132 (20) (2015) 1920–1930, https://doi.org/10.1161/CIRCULATIONAHA.115.001593.
[8] A.F. Markus, J.A. Kors, P.R. Rijnbeek, The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies, J. Biomed. Inform. 113 (2021), 103655, https://doi.org/10.1016/j.jbi.2020.103655.
[9] R.C. Li, S.M. Asch, N.H. Shah, Developing a delivery science for artificial intelligence in healthcare, npj Digit. Med. 3 (1) (2020) 1–3, https://doi.org/10.1038/s41746-020-00318-y.
[10] E.J. Topol, High-performance medicine: the convergence of human and artificial intelligence, Nat. Med. 25 (1) (2019) 44–56, https://doi.org/10.1038/s41591-018-0300-7.

[11] European Commission, Ethics guidelines for trustworthy AI. High-level expert group on artificial intelligence, Eur. Comm. 9 (2019) 1–39.

[12] M. Aria, C. Cuccurullo, bibliometrix: an R-tool for comprehensive science mapping analysis, J. Inform. 11 (4) (2017) 959–975, https://doi.org/10.1016/j.joi.2017.08.007.

[13] B. Séroussi, K.F. Hollis, L.F. Soualmia, Transparency of health informatics processes as the condition of healthcare professionals' and patients' trust and adoption: the rise of ethical requirements, Yearb. Med. Inform. 29 (1) (2020) 7–10, https://doi.org/10.1055/s-0040-1702029.

[14] T. Saheb, T. Saheb, D.O. Carpenter, Mapping research strands of ethics of artificial intelligence in healthcare: a bibliometric and content analysis, Comput. Biol. Med. 135 (2021), 104660, https://doi.org/10.1016/j.compbiomed.2021.104660.

[15] G. Muhammad, F. Alshehri, F. Karray, A. El Saddik, M. Alsulaiman, T.H. Falk, A comprehensive survey on multimodal medical signals fusion for smart healthcare systems, Inf. Fusion 76 (2021) 355–375, https://doi.org/10.1016/j.inffus.2021.06.007.

[16] H.W. Loh, C.P. Ooi, S. Seoni, P.D. Barua, F. Molinari, U.R. Acharya, Application of explainable artificial intelligence for healthcare: a systematic review of the last decade (2011–2022), Comput. Methods Programs Biomed. 226 (2022), 107161, https://doi.org/10.1016/j.cmpb.2022.107161.

[17] G. Yang, Q. Ye, J. Xia, Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: a mini-review, two showcases and beyond, Inf. Fusion 77 (2022) 29–52, https://doi.org/10.1016/j.inffus.2021.07.016.

[18] Y. Balagurunathan, R. Mitchell, I. El Naqa, Requirements and reliability of AI in the medical context, Phys. Medica 83 (2021) 72–78, https://doi.org/10.1016/j.ejmp.2021.02.024.

[19] M.L. Rethlefsen, et al., PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews, J. Med. Libr. Assoc. 109 (2) (2021) 174–200, https://doi.org/10.5195/jmla.2021.962.

[20] A. Holzinger, et al., Information fusion as an integrative cross-cutting enabler to achieve robust, explainable, and trustworthy medical artificial intelligence, Inf. Fusion 79 (2022) 263–278, https://doi.org/10.1016/j.inffus.2021.10.007.

[21] A. Angerschmid, J. Zhou, K. Theuermann, F. Chen, A. Holzinger, Fairness and explanation in AI-informed decision making, Mach. Learn. Knowl. Extr. 4 (2) (2022) 556–579, https://doi.org/10.3390/make4020026.

[22] C. Sohrabi, et al., PRISMA 2020 statement: what's new and the importance of reporting guidelines, Int. J. Surg. 88 (2021), 105918, https://doi.org/10.1016/j.ijsu.2021.105918. Elsevier.

[23] K.W. Khaw, A. Alnoor, H. AL-Abrrow, V. Tiberius, Y. Ganesan, N.A. Atshan, Reactions towards organizational change: a systematic literature review, Curr. Psychol. (2022) 1–24, https://doi.org/10.1007/s12144-022-03070-6.

[24] R. Setchi, M.B. Dehkordi, J.S. Khan, Explainable robotics in human-robot interactions, Procedia Comput. Sci. 176 (2020) 3057–3066, https://doi.org/10.1016/j.procs.2020.09.198.

[25] M. Rostami, M. Oussalah, A novel explainable COVID-19 diagnosis method by integration of feature selection with random forest, Inform. Med. Unlocked 30 (2022), 100941, https://doi.org/10.1016/j.imu.2022.100941.

[26] G. Nicora, M. Rios, A. Abu-Hanna, R. Bellazzi, Evaluating pointwise reliability of machine learning prediction, J. Biomed. Inform. 127 (2022), 103996, https://doi.org/10.1016/j.jbi.2022.103996.

[27] Z. Shi, W. Chen, S. Liang, W. Zuo, L. Yue, and S. Wang, "Deep interpretable mortality model for intensive care unit risk prediction," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11888 LNAI, no. 15th International Conference on Advanced Data Mining and Applications (ADMA). Jilin Univ, Minist Educ, Key Lab Symbol Computat & Knowledge Engn, Changchun 130012, Peoples R China, pp. 617–631, 2019. doi: 10.1007/978-3-030-35231-8_45.

[28] A. Lucieri, M.N. Bajwa, A. Dengel, and S. Ahmed, "Achievements and challenges in explaining deep learning based computer-aided diagnosis systems," *arXiv Prepr. arXiv2011.13169*, Nov. 2020.

[29] P. Washington et al., "Selection of trustworthy crowd workers for telemedical diagnosis of pediatric autism spectrum disorder," *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, vol. 26, no. 26th Pacific Symposium on Biocomputing (PSB). Stanford Univ, Dept Bioengn, Palo Alto, CA 94305 USA, pp. 14–25, 2021. doi: 10.1142/9789811232701_0002.

[30] A. Lucieri, M.N. Bajwa, S.A. Braun, M.I. Malik, A. Dengel, S. Ahmed, ExAID: a multimodal explanation framework for computer-aided diagnosis of skin lesions, Comput. Methods Programs Biomed. 215 (2022), 106620, https://doi.org/10.1016/j.cmpb.2022.106620.

[31] M. Anagnostou, et al., Characteristics and challenges in the industries towards responsible AI: a systematic literature review, Ethics Inf. Technol. 24 (3) (2022) 1–18, https://doi.org/10.1007/s10676-022-09634-1.

[32] A.S. Hussein, W.M. Omar, X. Li, and M. Ati, "Efficient chronic disease diagnosis prediction and recommendation system," in *2012 IEEE-EMBS Conference on Biomedical Engineering and Sciences, IECBES 2012*, 2012, pp. 209–214. doi: 10.1109/IECBES.2012.6498117.

[33] S. Martínez-Agüero, C. Soguero-Ruiz, J.M. Alonso-Moral, I. Mora-Jiménez, J. Álvarez-Rodríguez, A.G. Marques, Interpretable clinical time-series modeling with intelligent feature selection for early prediction of antimicrobial multidrug resistance, Futur. Gener. Comput. Syst. 133 (2022) 68–83, https://doi.org/10.1016/j.future.2022.02.021.

[34] A. Kerasidou, Ethics of artificial intelligence in global health: explainability, algorithmic bias and trust, J. Oral Biol. Craniofacial Res. 11 (4) (2021) 612–614, https://doi.org/10.1016/j.jobcr.2021.09.004.

[35] A.M. Carrington, et al., Deep ROC analysis and AUC as balanced average accuracy, for improved classifier selection, audit and explanation, IEEE Trans. Pattern Anal. Mach. Intell. 45 (1) (2023) 329–341, https://doi.org/10.1109/TPAMI.2022.3145392.

[36] G. Harerimana, J.W. Kim, B. Jang, A deep attention model to forecast the length of stay and the in-hospital mortality right on admission from ICD codes and demographic data, J. Biomed. Inform. 118 (2021), 103778, https://doi.org/10.1016/j.jbi.2021.103778.

[37] N. Ben Yahia, M. Dhiaeddine Kandara, N. Bellamine BenSaoud, Integrating models and fusing data in a deep ensemble learning method for predicting epidemic diseases outbreak, Big Data Res. 27 (2022), 100286, https://doi.org/10.1016/j.bdr.2021.100286.

[38] M.R. Karim, T. Islam, C. Lange, D. Rebholz-Schuhmann, S. Decker, Adversary-aware multimodal neural networks for cancer susceptibility prediction from multiomics data, IEEE Access 10 (2022) 54386–54409, https://doi.org/10.1109/ACCESS.2022.3175816.

[39] M. Abdar, et al., UncertaintyFuseNet: robust uncertainty-aware hierarchical feature fusion model with Ensemble Monte Carlo Dropout for COVID-19 detection, Inf. Fusion 90 (2023) 364–381, https://doi.org/10.1016/j.inffus.2022.09.023.

[40] R.K. Bania, A. Halder, R-HEFS: rough set based heterogeneous ensemble feature selection method for medical data classification, Artif. Intell. Med. 114 (2021), 102049, https://doi.org/10.1016/j.artmed.2021.102049.

[41] M. Loey, S. El-Sappagh, S. Mirjalili, Bayesian-based optimized deep learning model to detect COVID-19 patients using chest X-ray image data, Comput. Biol. Med. 142 (2022), 105213, https://doi.org/10.1016/j.compbiomed.2022.105213.

[42] K.A. Al Mamun, M. Alhussein, K. Sailunaz, M.S. Islam, Cloud based framework for Parkinson's disease diagnosis and monitoring system for remote healthcare applications, Futur. Gener. Comput. Syst. 66 (2017) 36–47, https://doi.org/10.1016/j.future.2015.11.010.

[43] Y.L. Chou, C. Moreira, P. Bruza, C. Ouyang, J. Jorge, Counterfactuals and causability in explainable artificial intelligence: theory, algorithms, and applications, Inf. Fusion 81 (2022) 59–83, https://doi.org/10.1016/j.inffus.2021.11.003. Mar.

[44] H. Müller, A. Holzinger, M. Plass, L. Brcic, C. Stumptner, K. Zatloukal, Explainability and causability for artificial intelligence-supported medical image analysis in the context of the European in vitro diagnostic regulation, N. Biotechnol. 70 (2022) 67–72, https://doi.org/10.1016/j.nbt.2022.05.002.

[45] O. Deperlioglu, U. Kose, D. Gupta, A. Khanna, F. Giampaolo, G. Fortino, Explainable framework for Glaucoma diagnosis by image processing and convolutional neural network synergy: analysis with doctor evaluation, Futur. Gener. Comput. Syst. 129 (2022) 152–169, https://doi.org/10.1016/j.future.2021.11.018.

[46] A. Barredo Arrieta, et al., Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI, Inf. Fusion 58 (2020) 82–115, https://doi.org/10.1016/j.inffus.2019.12.012. Jun.

[47] S. Sachan, F. Almaghrabi, J.B. Yang, D.L. Xu, Evidential reasoning for preprocessing uncertain categorical data for trustworthy decisions: an application on healthcare and finance, Expert Syst. Appl. 185 (2021), 115597, https://doi.org/10.1016/j.eswa.2021.115597.

[48] Y. Du, A.R. Rafferty, F.M. McAuliffe, L. Wei, C. Mooney, An explainable machine learning-based clinical decision support system for prediction of gestational diabetes mellitus, Sci. Rep. 12 (1) (2022) 1170, https://doi.org/10.1038/s41598-022-05112-2.

[49] I. Ullah, H.Y. Youn, Y.H. Han, Integration of type-2 fuzzy logic and Dempster–Shafer theory for accurate inference of IoT-based health-care system, Futur. Gener. Comput. Syst. 124 (2021) 369–380, https://doi.org/10.1016/j.future.2021.06.012.

[50] C. Giordano, M. Brennan, B. Mohamed, P. Rashidi, F. Modave, P. Tighe, Accessing artificial intelligence for clinical decision-making, Front. Digit. Heal. 3 (2021) 65, https://doi.org/10.3389/fdgth.2021.645232.

[51] S. El-Sappagh, J.M. Alonso, S.M.R. Islam, A.M. Sultan, K.S. Kwak, A multilayer multimodal detection and prediction model based on explainable artificial intelligence for Alzheimer's disease, Sci. Rep. 11 (1) (2021), https://doi.org/10.1038/s41598-021-82098-3.

[52] C. González-Gonzalo, et al., Trustworthy AI: closing the gap between development and integration of AI systems in ophthalmic practice, Prog. Retin. Eye Res. (2021), https://doi.org/10.1016/j.preteyeres.2021.101034.

[53] J. Müller, et al., A visual approach to explainable computerized clinical decision support, Comput. Graph. 91 (2020) 1–11, https://doi.org/10.1016/j.cag.2020.06.004.

[54] T. Alanazi, G. Muhammad, Human fall detection using 3D multi-stream convolutional neural networks with fusion, Diagnostics 12 (12) (2022), https://doi.org/10.3390/diagnostics12123060.

[55] F. Zerka, et al., Blockchain for privacy preserving and trustworthy distributed machine learning in multicentric medical imaging (C-DistriM), IEEE Access 8 (2020) 183939–183951, https://doi.org/10.1109/ACCESS.2020.3029445.

[56] M.F. Guiñazú, V. Cortés, C.F. Ibáñez, J.D. Velásquez, Employing online social networks in precision-medicine approach using information fusion predictive model to improve substance use surveillance: a lesson from Twitter and marijuana consumption, Inf. Fusion 55 (2020) 150–163, https://doi.org/10.1016/j.inffus.2019.08.006.

[57] F. Leal, et al., Smart pharmaceutical manufacturing: ensuring end-to-end traceability and data integrity in medicine production, Big Data Res. 24 (2021), 100172, https://doi.org/10.1016/j.bdr.2020.100172.

[58] M. Zarour, et al., Evaluating the impact of blockchain models for secure and trustworthy electronic healthcare records, IEEE Access 8 (2020) 157959–157973, https://doi.org/10.1109/ACCESS.2020.3019829.

[59] L. Ouyang, Y. Yuan, Y. Cao, F.Y. Wang, A novel framework of collaborative early warning for COVID-19 based on blockchain and smart contracts, Inf. Sci. (Ny). 570 (2021) 124–143, https://doi.org/10.1016/j.ins.2021.04.021.

[60] M.A. Rahman, M.S. Hossain, A.J. Showail, N.A. Alrajeh, M.F. Alhamid, A secure, private, and explainable IoHT framework to support sustainable health monitoring in a smart city, Sustain. Cities Soc. 72 (2021), 103083, https://doi.org/10.1016/j.scs.2021.103083.

[61] E.M. Abou-Nassar, A.M. Iliyasu, P.M. El-Kafrawy, O.Y. Song, A.K. Bashir, A.A. A. El-Latif, DITrust chain: towards blockchain-based trust models for sustainable healthcare IoT systems, IEEE Access 8 (2020) 111223–111238, https://doi.org/10.1109/ACCESS.2020.2999468.

[62] N.R. Pal, In search of trustworthy and transparent intelligent systems with human-like cognitive and reasoning capabilities, Front. Robot. AI 7 (2020), https://doi.org/10.3389/frobt.2020.00076.

[63] I. Barclay and W. Abramson, "Identifying roles, requirements and responsibilities in trustworthy AI systems," in *UbiComp/ISWC 2021 - Adjunct Proceedings of the 2021 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2021 ACM International Symposium on Wearable Computers*, 2021, pp. 264–271. doi: 10.1145/3460418.3479344.

[64] J.W. Collins, et al., Ethical implications of AI in robotic surgical training: a Delphi consensus statement, Eur. Urol. Focus (2021), https://doi.org/10.1016/j.euf.2021.04.006.

[65] R. Larasati, A. De Liddo, and E. Motta, "AI healthcare system interface: explanation design for non-expert user trust," in *CEUR Workshop Proceedings*, 2021, vol. 2903.

[66] S. Delacroix, B. Wagner, Constructing a mutually supportive interface between ethics and regulation, Comput. Law Secur. Rev. 40 (2021), 105520, https://doi.org/10.1016/j.clsr.2020.105520.

[67] A. Holzinger, B. Malle, A. Saranti, B. Pfeifer, Towards multi-modal causability with graph neural networks enabling information fusion for explainable AI, Inf. Fusion 71 (2021) 28–37, https://doi.org/10.1016/j.inffus.2021.01.008.

[68] N. Rieke, et al., The future of digital health with federated learning, npj Digit. Med. 3 (1) (2020), https://doi.org/10.1038/s41746-020-00323-1.

[69] M. Wenzel, T. Wiegand, Toward global validation standards for health AI, IEEE Commun. Stand. Mag. 4 (3) (2020) 64–69, https://doi.org/10.1109/MCOMSTD.001.2000006.

[70] A. Sheikh, et al., Health information technology and digital innovation for national learning health and care systems, Lancet Digit. Heal. 3 (6) (2021) e383–e396, https://doi.org/10.1016/S2589-7500(21)00005-4.

[71] C.W.L. Ho, K. Caals, A call for an ethics and governance action plan to harness the power of artificial intelligence and digitalization in nephrology, Semin. Nephrol. 41 (3) (2021) 282–293, https://doi.org/10.1016/j.semnephrol.2021.05.009.

[72] H. Faris, M. Habib, M. Faris, H. Elayan, A. Alomari, An intelligent multimodal medical diagnosis system based on patients' medical questions and structured symptoms for telemedicine, Inform. Med. Unlocked 23 (2021), 100513, https://doi.org/10.1016/j.imu.2021.100513.

[73] A.M. Oprescu, et al., Towards a data collection methodology for responsible artificial intelligence in health: a prospective and qualitative study in pregnancy, Inf. Fusion 83–84 (2022) 53–78, https://doi.org/10.1016/j.inffus.2022.03.011.

[74] M. Esposito, A. Minutolo, R. Megna, M. Forastiere, M. Magliulo, G. De Pietro, A smart mobile, self-configuring, context-aware architecture for personal health monitoring, Eng. Appl. Artif. Intell. 67 (2018) 136–156, https://doi.org/10.1016/j.engappai.2017.09.019.

[75] V.K. Rathi, et al., An edge AI-enabled IoT healthcare monitoring system for smart cities, Comput. Electr. Eng. 96 (2021), 107524, https://doi.org/10.1016/j.compeleceng.2021.107524.

[76] A. Rehman, T. Saba, K. Haseeb, S.L. Marie-sainte, J. Lloret, Energy-efficient iot e-health using artificial intelligence model with homomorphic secret sharing, Energies 14 (19) (2021), https://doi.org/10.3390/en14196414.

[77] T. Saba, K. Haseeb, I. Ahmed, A. Rehman, Secure and energy-efficient framework using Internet of Medical Things for e-healthcare, J. Infect. Public Health 13 (10) (2020) 1567–1575, https://doi.org/10.1016/j.jiph.2020.06.027.

[78] J. Wang, H. Jin, J. Chen, J. Tan, K. Zhong, Anomaly detection in Internet of Medical Things with blockchain from the perspective of deep neural network, Inf. Sci. (Ny). 617 (2022) 133–149, https://doi.org/10.1016/j.ins.2022.10.060.

[79] F. Alshehri, G. Muhammad, A comprehensive survey of the Internet of Things (IoT) and AI-based smart healthcare, IEEE Access 9 (2021) 3660–3678, https://doi.org/10.1109/ACCESS.2020.3047960.

[80] A. Shoeibi, et al., Diagnosis of brain diseases in fusion of neuroimaging modalities using deep learning: a review, Inf. Fusion 93 (2023) 85–117, https://doi.org/10.1016/j.inffus.2022.12.010.

[81] L. Alzubaidi, et al., Review of deep learning: concepts, CNN architectures, challenges, applications, future directions, J. Big Data 8 (1) (2021) 53, https://doi.org/10.1186/s40537-021-00444-8. Dec.

[82] K. Abolfazlian, Trustworthy AI Needs Unbiased Dictators!, in: IFIP Advances in Information and Communication Technology, 584, 2020, pp. 15–23, https://doi.org/10.1007/978-3-030-49186-4_2. IFIP.

[83] M.S. Ayhan, et al., Clinical validation of saliency maps for understanding deep neural networks in ophthalmology, Med. Image Anal. 77 (2022), 102364, https://doi.org/10.1016/j.media.2022.102364. Apr.

[84] A. Saporta, et al., Benchmarking saliency methods for chest X-ray interpretation, Nat. Mach. Intell. 4 (10) (2022) 867–878, https://doi.org/10.1038/s42256-022-00536-x. Oct.

[85] F.L. Fan, J. Xiong, M. Li, G. Wang, On interpretability of artificial neural networks: a survey, IEEE Trans. Radiat. Plasma Med. Sci. 5 (6) (2021) 741–760, https://doi.org/10.1109/TRPMS.2021.3066428.

[86] Bergur Thormundsson, "Global explainable AI market revenues 2021-2030 | Statista.".

[87] L.R. Goldberg, The book of why: the new science of cause and effect, Not. Am. Math. Soc. 66 (07) (2019) 1, https://doi.org/10.1090/noti1912. Aug.

[88] D. Gunning, D.W. Aha, DARPA's explainable artificial intelligence program, AI Mag. 40 (2) (2019) 44–58, https://doi.org/10.1609/aimag.v40i2.2850. Jun.

[89] A. Holzinger, A. Saranti, C. Molnar, P. Biecek, and W. Samek, "Explainable AI methods - a brief overview," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2022, vol. 13200 LNAI, pp. 13–38. doi: 10.1007/978-3-031-04083-2_2.

[90] W. Samek, G. Montavon, S. Lapuschkin, C.J. Anders, and K.R. Müller, "Explaining deep neural networks and beyond: a review of methods and applications," *Proc. IEEE*, vol. 109, no. 3, pp. 247–278, 2021, doi: 10.1109/JPROC.2021.3060483.

[91] M.T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?' Explaining the predictions of any classifier," in *NAACL-HLT 2016 - 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Demonstrations Session*, 2016, pp. 97–101. doi: 10.18653/v1/n16-3020.

[92] M.T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: high-precision model-agnostic explanations," *32nd AAAI Conf. Artif. Intell. AAAI 2018*, vol. 32, no. 1, pp. 1527–1535, Apr. 2018, doi: 10.1609/aaai.v32i1.11491.

[93] Q. Huang, M. Yamada, Y. Tian, D. Singh, Y. Chang, GraphLIME: local interpretable model explanations for graph neural networks, IEEE Trans. Knowl. Data Eng. (2022), https://doi.org/10.1109/TKDE.2022.3187455.

[94] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.R. Müller, W. Samek, On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation, PLoS ONE 10 (7) (2015), e0130140, https://doi.org/10.1371/journal.pone.0130140.

[95] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, K.R. Müller, Explaining nonlinear classification decisions with deep Taylor decomposition, *Pattern Recogn*it. 65 (2017) 211–222, https://doi.org/10.1016/j.patcog.2016.11.008. May.

[96] M. Robnik-Šikonja, I. Kononenko, Explaining classifications for individual instances, IEEE Trans. Knowl. Data Eng. 20 (5) (2008) 589–600, https://doi.org/10.1109/TKDE.2007.190734.

[97] C. Frye, D. de Mijolla, T. Begley, L. Cowton, M. Stanley, and I. Feige, "Shapley explainability on the data manifold," *arXiv Prepr. arXiv2006.01272*, 2020.

[98] C. Frye, C. Rowat, I. Feige, Asymmetric shapley values: incorporating causal knowledge into model-agnostic explainability, Adv. Neural Inf. Process. Syst. 2020 (2020) 1229–1239. -Decem.

[99] P. Biecek, Dalex: explainers for complex predictive models in R, J. Mach. Learn. Res. 19 (1) (2018) 3245–3249.

[100] J. Wang, J. Wiens, and S. Lundberg, "Shapley flow: a graph-based approach to interpreting model predictions," in *International Conference on Artificial Intelligence and Statistics*, 2020, pp. 721–729.

[101] A. Nadeem, A. Jalal, K. Kim, Accurate physical activity recognition using multidimensional features and Markov model for smart health fitness, Symmetry (Basel) 12 (11) (2020) 1–17, https://doi.org/10.3390/sym12111766.

[102] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *34th International Conference on Machine Learning, ICML 2017*, 2017, vol. 7, pp. 5109–5118.

[103] P. Madumal, T. Miller, L. Sonenberg, and F. Vetere, "Explainable reinforcement learning through a causal lens," in *AAAI 2020 - 34th AAAI Conference on Artificial Intelligence*, 2020, vol. 34, no. 03, pp. 2493–2500. doi: 10.1609/aaai.v34i03.5631.

[104] R.C. Fong and A. Vedaldi, "Interpretable explanations of black boxes by meaningful perturbation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, vol. 2017-Octob, pp. 3449–3457. doi: 10.1109/ICCV.2017.371.

[105] N. Díaz-Rodríguez, et al., EXplainable Neural-Symbolic Learning (X-NeSyL) methodology to fuse deep learning representations with expert knowledge graphs: the MonuMAI cultural heritage use case, Inf. Fusion 79 (2022) 58–83, https://doi.org/10.1016/j.inffus.2021.09.022.

[106] E. Crigger, K. Reinbold, C. Hanson, A. Kao, K. Blake, M. Irons, Trustworthy augmented intelligence in health care, J. Med. Syst. 46 (2) (2022) 1–11, https://doi.org/10.1007/s10916-021-01790-z.

[107] F. Gille, A. Jobin, M. Ienca, What we talk about when we talk about trust: theory of trust for AI in healthcare, Intell. Med. 1–2 (2020), https://doi.org/10.1016/j.ibmed.2020.100001.

[108] R. Yang, S. Wibowo, User trust in artificial intelligence: a comprehensive conceptual framework, Electron. Mark (2022) 1–25, https://doi.org/10.1007/s12525-022-00592-6.

[109] K. Stoger, D. Schneeberger, A. Holzinger, Medical artificial intelligence: the European legal perspective, Commun. ACM 64 (11) (2021) 34–36.

[110] F. Ali, et al., A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion, Inf. Fusion 63 (2020) 208–222, https://doi.org/10.1016/j.inffus.2020.06.008.

[111] T. Meng, X. Jing, Z. Yan, W. Pedrycz, A survey on machine learning for data fusion, Inf. Fusion 57 (2020) 115–129, https://doi.org/10.1016/j.inffus.2019.12.001.

[112] N. Zeng, P. Wu, Z. Wang, H. Li, W. Liu, X. Liu, A small-sized object detection oriented multi-scale feature fusion approach with application to defect detection,

IEEE Trans. Instrum. Meas. 71 (2022) 1–14, https://doi.org/10.1109/TIM.2022.3153997.
[113] H. Zhang, H. Xu, X. Tian, J. Jiang, J. Ma, Image fusion meets deep learning: a survey and perspective, Inf. Fusion 76 (2021) 323–336, https://doi.org/10.1016/j.inffus.2021.06.008.
[114] F. Tang, et al., Postoperative glioma segmentation in CT image using deep feature fusion model guided by multi-sequence MRIs, Eur. Radiol. 30 (2) (2020) 823–832, https://doi.org/10.1007/s00330-019-06441-z.
[115] A.H. Al-Timemy, N.H. Ghaeb, Z.M. Mosa, J. Escudero, Deep transfer learning for improved detection of keratoconus using corneal topographic maps, Cognit. Comput. 14 (5) (2022) 1627–1642, https://doi.org/10.1007/s12559-021-09880-3.
[116] J. Li, Q. Wang, Multi-modal bioelectrical signal fusion analysis based on different acquisition devices and scene settings: overview, challenges, and novel orientation, Inf. Fusion 79 (2022) 229–247, https://doi.org/10.1016/j.inffus.2021.10.018.
[117] S.P. Yadav, S. Yadav, Image fusion using hybrid methods in multimodality medical images, Med. Biol. Eng. Comput. 58 (4) (2020) 669–687, https://doi.org/10.1007/s11517-020-02136-6.
[118] Q. Wang, Y. Tang, X. Tong, P.M. Atkinson, Virtual image pair-based spatio-temporal fusion, Remote Sens. Environ. 249 (2020), 112009, https://doi.org/10.1016/j.rse.2020.112009. Nov.