Discover Artificial Intelligence

Review

Explainable and interpretable artificial intelligence in medicine: a systematic bibliometric review

Maria Frasca[1] · Davide La Torre[2] · Gabriella Pravettoni[1,3] · Ilaria Cutica[1]

Received: 31 October 2023 / Accepted: 23 February 2024
Published online: 27 February 2024

Abstract
This review aims to explore the growing impact of machine learning and deep learning algorithms in the medical field, with a specific focus on the critical issues of explainability and interpretability associated with black-box algorithms. While machine learning algorithms are increasingly employed for medical analysis and diagnosis, their complexity underscores the importance of understanding how these algorithms explain and interpret data to take informed decisions. This review comprehensively analyzes challenges and solutions presented in the literature, offering an overview of the most recent techniques utilized in this field. It also provides precise definitions of interpretability and explainability, aiming to clarify the distinctions between these concepts and their implications for the decision-making process. Our analysis, based on 448 articles and addressing seven research questions, reveals an exponential growth in this field over the last decade. The psychological dimensions of public perception underscore the necessity for effective communication regarding the capabilities and limitations of artificial intelligence. Researchers are actively developing techniques to enhance interpretability, employing visualization methods and reducing model complexity. However, the persistent challenge lies in finding the delicate balance between achieving high performance and maintaining interpretability. Acknowledging the growing significance of artificial intelligence in aiding medical diagnosis and therapy, and the creation of interpretable artificial intelligence models is considered essential. In this dynamic context, an unwavering commitment to transparency, ethical considerations, and interdisciplinary collaboration is imperative to ensure the responsible use of artificial intelligence. This collective commitment is vital for establishing enduring trust between clinicians and patients, addressing emerging challenges, and facilitating the informed adoption of these advanced technologies in medicine.

1 Introduction

The rise of Artificial Intelligence (AI) led to a revolutionary transformation in the medical field, redefining how diagnostic and therapeutic challenges are addressed. This synergy is revolutionizing the concept of personalized patient care, opening up new perspectives in prevention, therapy, and optimization of medical resources. In this evolving scenario, AI algorithms allow us to analyze complex biomedical data, identify and extract hidden patterns, and drive the decision-making process. The ability to process large amounts of information quickly and efficiently allows for earlier diagnoses, targeted treatments and more efficient management of medical conditions [1].

✉ Davide La Torre, davide.latorre@skema.edu; Maria Frasca, maria.frasca@unimi.it; Gabriella Pravettoni, gabriella.pravettoni@unimi.it; gabriella.pravettoni@ieo.it; Ilaria Cutica, ilaria.cutica@unimi.it | [1]Department of Oncology and Hemato-Oncology, University of Milan, Milan, Italy. [2]SKEMA Business School, Université Côte d'Azur, Sophia Antipolis, Nice, France. [3]Applied Research Division for Cognitive and Psychological Science, IEO, European Institute of Oncology, Milano, Italy.

Unfortunately, understanding how these systems make decisions remains a critical issue for clinicians, professionals, patients and stakeholders involved in the process. There is an ongoing debate on the transparency of decision-making processes, the interpretability of algorithms, and ethical issues related to the adoption of automated systems in the clinical context [2].

Although current research results indicate that AI algorithms can outperform humans in certain analytical tasks, the lack of interpretability and explainability limit the adoption of AI-based solutions in the medical context, as it raises legal and ethical concerns, potentially hindering progress and preventing new technologies from realizing their full potential in improving health care.

In AI systems theory, we can identify two distinct methods for categorizing models and algorithms: Interpretable (i.e. white-box) and non-interpretable (i.e. black-box) [3]. This differentiation is based on the clarity of the relationship between the input–output data and the outcomes generated by the model. White-box models have recognizable and understandable characteristics that help explaining the influence of variables on predicting outcomes, for instance linear regression models and decision trees belong to this family. On the other hand, black-box models are based on highly complex structures where the processes, parameters and predictions are unknown, for instance, in the case of deep learning algorithms and random forest models. Black-box models might incorporate harmful biases [4, 5] and this could hurt clinicians' confidence and trust. Indeed, if an algorithm is trained on data reflecting cultural or social biases, it can result in discriminatory decisions that exacerbate inequalities rather than reduce them. This raises important ethical questions about fairness and justice towards the individuals involved.

Finally, while white-box models feature easier-to-understand processes and results, black-box models show better performance and accuracy. This can be mainly attributed to the intrinsic ability of black-box models to learn complex and non-linear representations of data. Their more complex and flexible structure allows to capture intricate details and hidden relationships in the data, improving the model's ability to make more accurate predictions. While white-box models often rely on simpler and more interpretable relationships [3].

It has also to be considered the rising of ethical concerns related to the opacity of algorithms. Regulatory frameworks such as the General Data Protection Regulation (GDPR) [6], a European legislative framework, defines the legal requirements for acquisition, storage, transfer, processing and analysis of health data [7]. The GDPR also emphasizes the "right to explanation", which gives individuals the right to understand how automated decisions can affect them. It also highlights the concept of accountability, placing the responsibility on organizations to demonstrate compliance with regulations. This implies greater transparency in data processing and the need to clarify how algorithms operate when involving personal data.

In summary, this paper aims to explore the interpretability and explainability of Machine Learning (ML) and Deep Learning (DL) algorithms within the medical domain. We will investigate the particular challenges highlighted in existing literature and analyze the ramifications of algorithmic opacity. We will also examine specific case studies that show how interpretability and explainability can make a difference in daily clinical practice, and explore the most relevant technical approaches to generate easy-to-understand explanations.

The paper is structured as follows: Sect. 1 presents an introduction to the topic of explainability and interpretability of AI in the medical field. In Sect. 2 we provide a brief discussion on explainability and interpretability in artificial intelligence. Section 3 presents the investigations already available in the literature on the interpretability and explainability of AI algorithms in the medical field.

Section 4 defines the inclusion and exclusion criteria for the construction of our dataset. In section 5 we describe the results obtained from the Scopus and Web of Science databases and propose seven research questions. We highlight the results obtained by comparing the two datasets, considering different levels of detail. In particular, in subsections 5.1, 5.3, 5.3, we provide a quantitative analysis of the state of the art relating to the application of interpretability and explainability techniques in AI in the medical field, in particular by focusing on the main channels used for publication and in which countries the most active are located research centres. In Sect. 6 we proceed with the analysis of the 10 papers chosen for the analysis and in subsections 6.1, 6.2, 6.3, we analyze the application domains of the proposed techniques, provide technical insights on the formalization of the problems and discuss the performance metrics used for the evaluation and finally we consider the challenges faced by each paper. In Sect. 7 we discuss the selected papers and compare them with our definition of explainability and interpretability of AI algorithms in the medical field. Finally in Sect. 8 we provide conclusions of the proposed study.

2 Explainability and interpretability

AI algorithms' interpretability and explainability represent one of the fundamental pillars in the research and development of advanced AI systems [8].

Explainability is often confused with interpretability when interpretability is a prerequisite for explainability [9]. Furthermore, interpretability is sometimes defined as a part of explainability [10]. Ambiguity can arise when, in contexts of discussion about AI, the two words are used interchangeably, without precise distinction between the two concepts. However, the difference is important and can influence how we evaluate and implement AI algorithms.

The Explainability property of a model includes the ability to provide detailed and understandable reasoning for specific decisions. Explainability usually involves the capability to identify and understand the parameters within a system, understand nodes, architectures, and computational units that process and transmit information, and understand the significance of each component Within a system.

Interpretability, on the other hand, is the ability to understand the general behaviour of a model without necessarily delving into each decision. Thus, interpretability can be seen as a broader component of explainability, providing a general, high-level view of the model without going into specific details. Interpretability pertains to the extent to which a system's cause-and-effect relationships can be understood. It also involves not only understanding but also describing how a system operates or behaves. In general AI systems may vary in terms of their interpretability, with some being more transparent and interpretable than others. This is also related to the notion of degree of interpretability. To summarize, we might say that:

- *Interpretability.* It is about the degree to which a cause and effect can be seen inside a system and refers to a certain way something is understood or described. As a result, it can identify the cause of a problem and foresee what will happen if the input or computational parameters change [10].
- *Explainability,* It refers to the ability to provide explicit and understandable justifications or reasoning for model-specific decisions, it is the ability to recognize the parameters and understand what a node stands for and its significance within a system [10].

Existing methods to support explainability and interpretability of AI systems are often divided into a priori and a posteriori approaches. *A priori* explainability and interpretability techniques refer to concepts known or presumed before experience or observation of specific data. In this context, a priori explainability and interpretability refer to integrated techniques or design considerations applied during the creation and design of an AI model. They can be summarized as follows:

- The architecture simplicity can enhance interpretability by employing models with simpler topologies, such as linear models or shallow neural networks [11];
- Feature engineering involves selecting or creating features that reflect understandable or well-known concepts in a given domain. Regularization techniques, such as penalizing large weights to encourage sparsity and enhance interpretability, can be employed [12];
- Regularization algorithms, including L1 and L2 regularization, can be utilized to promote sparsity and improve interpretability [13];
- Implementing a depth limitation can prevent the development of overly complex systems, contributing to the interpretability of the model.

A posteriori explainability and interpretability techniques refer to concepts derived or deduced from experience or observation of specific data. In this context, a posteriori explainability and interpretability refer to techniques applied after the model has been trained to understand the specific decisions made by the model. They include:

- LIME (Local Interpretable Model-agnostic Explanations), which provides local explanations for certain input instances and offers a post-hoc comprehension of model decisions, is one technique that serves as an example [14];
- SHAP (SHapley Additive exPlanations), which uses the Shapley values idea to assign feature contributions to each input variable and explains predictions made by the model in the past [15];

- Feature visualizations, which use specific data instances to build retrospective visualizations of the most significant aspects;
- A sensitivity analysis to assess how slight changes in the inputs might have affected the model's predictions in the past [16].

Usually a posteriori approach is not considered during system design and it is concerned with extracting explanatory information from existing systems [17] (typically based on the black-box approach).

Developing an explainable and interpretable AI system is usually challenging and variable because its complexity will depend on the main purpose of the model as well as on its related variables and features. There are many reasons why explainability and interpretability can be desirable or necessary in AI systems [18] among which are ethical, legal, and practical. The following are some of the main obstacles to their implementation:

- *Opacity of complex models:* Deep learning techniques, such as deep neural networks, can build highly complex models with millions of parameters. Even the model's developers themselves may find it challenging to comprehend how the model comes to its conclusions due to its intricacy.
- *Trade-off between performance and interpretability:* It is frequently required to utilize more complex models to gain higher performance in terms of accuracy and generalization. But more intricate models typically have a harder time being understood. A significant problem is striking the ideal balance between performance and interoperability.
- *Bias in training data:* If a model is trained on biased data, it may carry over these biases and produce discriminating or inaccurate decisions. It's critical to comprehend how the model uses the data and find any unintentional biases.
- *Interpretability of features:* A key component of interpretability is knowing which features are pertinent to model decisions. In many situations, particularly with complicated models like deep neural networks, it can be challenging to pinpoint which parts impact the model.
- *Scalability:* Because large-scale deep learning models contain many more parameters and demand more computer resources for interpretive analysis, interpretability can be more challenging to achieve in these models.
- *Changes in model behaviour:* A machine learning algorithm's behaviour may alter over time due to modifications to the input data or the training it receives. It can be difficult to maintain an interpretable model in the face of such changes.
- *Social acceptance:* Even if a model may be interpreted in theory, it may be challenging to convince users to do so, especially if the justifications offered do not line up with their intuitions.

To address these challenges, scientists and engineers are developing various techniques and approaches to improve the interpretability and explainability of AI algorithms. This includes using visualization techniques, interpreting features, generating textual explanations, and reducing model complexity.

Finding a balance between high performance and interpretability is an ongoing challenge, but essential to ensuring that AI can be used ethically and responsibly for the benefit of society as a whole [19]. Advances in this field are critical to shaping the future of artificial intelligence.

3 Related work

In the extant literature, there are already some review articles that present and discuss trends and challenges related to the integration of the notions of interpretability and explainability into AI systems. We do not include them in our literature analysis, so this section provides an overview of the most significant contributions.

Biran et al. [20] look at various approaches for making machine learning models understandable. The authors investigate methods for enhancing user confidence and boosting real-world adoption by making complex models accessible and intelligible. The authors categorize the many interpretation techniques they look at into several groups. Rule-based approaches, global and local procedures, visualization techniques, and model representation-based approaches are some of these categories. Additionally, they demonstrate the practical importance of these interpretation strategies by highlighting how they are used in particular fields like finance, marketing, and health.

Guidotti et al. [21] provides a comprehensive overview of the methods proposed in the literature to explain decision-making systems based on opaque and dark machine learning models. They identified several explainability problems and provided a formal definition for each. Identified "black box" problems include the model explainability problem, the outcome explainability problem, the model inspection problem, and the transparent box design problem. The analysis of the literature led to the conclusion that, although many approaches have been proposed to explain "black boxes", some important scientific questions remain unresolved.

By incorporating viewpoints from the social sciences, Miller et al. [22] explore the function of explanations in AI in this paper. The authors emphasize the need to make algorithm conclusions more user-acceptable and understandable while discussing the significance of explanations in artificial intelligence systems. They look at the many difficulties in giving convincing explanations, taking into account variations in how explanations are perceived on a cultural, social, and psychological level. The authors stress how crucial it is to involve users in the design of explanations to make sure they are useful and well-received. To increase the comprehension and social acceptance of intelligent technology, the paper proposes an interdisciplinary method to address the complexity of explanations in AI.

Arrieta et al. [19] investigate the concepts, taxonomies, prospects, and difficulties within Explainable Artificial Intelligence (XAI) to address the moral and societal ramifications associated with the secrecy of AI algorithms. The authors stress the significance of creating more understandable models to encourage the responsible use of AI and the significance of comprehending and interpreting algorithmic judgments to guarantee responsible governance. Based on factors including the type of explanation, the level of granularity, and the manner of application, the authors suggest various taxonomies to categorize XAI techniques. The authors also discuss difficulties in implementing XAI, such as the necessity to manage the trade-off between precision and interpretability and the role of privacy in connection to model explainability. Other difficulties include the need to balance explainability and complexity.

Tjia et al. [23] discuss the importance of interpretability in black x machine decisions in a medical context. They provide an overview of the interpretations proposed in different studies and classify them according to their clarity. In the medical field, such explanations are essential to justify the reliability of algorithmic decisions. However, this article highlights challenges such as risks associated with manipulating explanations and the quality of training data. It also highlights the importance of specialized training to correctly interpret algorithm descriptions in a medical context. Finally, this article calls for a critical approach to the use of algorithmic interpretation, which should be seen as a complementary support to medical decisions until a more robust approach to interpretability is developed.

Stiglic et al. [24] emphasize the significance of interpretability in machine learning (ML) models in the context of healthcare. The authors divided interpretability approaches into two main categories: one centred on personalized (local) interpretation, which emphasizes thorough justifications at the individual level, and the other concerned with the synthesis of prediction models on a population level (global), useful for getting a broad overview of trends. Additionally, they divided interpretability techniques into two groups: model-specific procedures and model-neutral strategies. The former analyses predictions made by a particular ML model, like a neural network. On the other hand, model-agnostic approaches offer clear justifications for any ML model's predictions, regardless of its architecture.

Amann et al. [7] discuss the issue of interpretability in the use of artificial intelligence (AI) in the healthcare industry, highlighting that while AI-based systems have demonstrated superior performance over humans in specific analytical tasks, the lack of interpretability has drawn criticism. The authors use the case of AI-based clinical decision support systems as a starting point for their multidisciplinary analysis of the applicability of interpretability for medical AI, taking into account the perspectives of technology, law, medicine, and patients. Based on the findings of this conceptual study, an ethical evaluation of the "Principles of Biomedical Ethics" by Beauchamp and Childress (autonomy, beneficence, nonmaleficence, and justice) is carried out to ascertain the necessity of interpretability in medical AI. Each domain draws attention to a distinct collection of factors and ideals crucial to comprehending the function of interpretability in clinical practice. The importance of taking into account the interaction between human actors and medical AI was emphasized from both a medical and patient perspective. The absence of interpretability in clinical decision support systems poses a threat to fundamental medical ethical principles and may have unfavourable effects on both individual and public health.

In the paper [25] Mehrabi et al conduct in-depth research on bias and bias in machine learning models. The authors look at the causes of bias, and how it manifests in models and solutions to these issues. They identified the main causes of bias in machine learning models, including characteristics included in the models, training data, and the learning process itself. The authors investigated different metrics to evaluate bias in models, such as group fairness, individual fairness, and fairness of opportunity, and they analyzed several strategies to mitigate bias in machine learning models, including gathering balanced data, adjusting model weights, and implementing fairness metrics to evaluate model performance. They examine the difficulties brought about by these measurements as well.

In this work, Chakrobartty et al. [26] focus on the recent advancement of XAI in the setting of medicine to offer a comprehensive overview of XAI approaches and techniques noted in the literature. The article addresses XAI approaches and techniques utilized in ML systems in the medical industry through a thorough literature review. The given conceptual framework for categorizing XAI approaches and techniques aids in the organization and discussion of the available literature. The balance between interpretability and accuracy is emphasized in the paper as a major subject in the literature, with some studies emphasizing interpretability in addition to accuracy.

In [2] the authors use the systematic mapping procedure to review the literature on interpretability strategies utilized in the medical area. The following factors were taken into account: the locations and years of publications; the types of contributions; the medical and ML disciplines; the ML objectives; the interpretation of "black box" ML techniques; the examination of interpretability techniques; the performance of the techniques; the best techniques; and, lastly, the datasets used in the evaluations of interpretability techniques. The results show an increase in the number of interpretability studies over time, with a predominance of solution proposals and empirical types based on experiments, after selecting 179 articles (1994–2020) from six digital libraries (ScienceDirect, IEEE Xplore, ACM Digital Library, SpringerLink, Wiley, and Google Scholar). The most common medical activities, fields, and ML aims were discovered to be classification, oncology, and diagnosis. The most popular "black box" ML approaches for interpretability are artificial neural networks. Accuracy, integrity, and the number of rules were other criteria that were frequently employed to gauge interpretability.

In [27] the authors consider the problem of interpretability. They emphasize that while AI systems have displayed outstanding performance in numerous clinical activities, efforts to make the AI more "interpretable" or explicable have been prompted by the lack of transparency of some of its black boxes. The paper argues that clinicians may favour interpretable systems even at the expense of maximum accuracy, defending the importance of interpretability. This inclination is supported by the fact that to get the intended benefits, doctors must employ AI. The authors make the point that giving accuracy priority over interpretability could be a "lethal bias," reducing the advantages of AI for patients.

Combi et al. look into the application of explainable artificial intelligence (XAI) in biomedical settings in [17],. The authors identify five key topics that demand more study. The first point is the importance of bridging the symbolic and sub-symbolic machine learning methodologies. Engineering explainability in intelligent systems is another major problem, and overcoming it calls for a thorough investigation of the structural, functional, and behavioural traits of various intelligent systems as well as the requirements of their users. The assessment and enhancement of the results of explainable elements and methodologies is the third key part. It is emphasized that the study must examine the effects on users' beliefs, attitudes, and behaviour and how accurately intelligent systems make decisions. Determining whether explainability is required also becomes a concern. Finally, the need to look into user-centred explainability artefact design becomes apparent. For XAI, user-centred design is essential.

Farah et al. [28] review key ideas for creating medical devices using AI and emphasize the value of algorithm performance, interpretability, and explainability. According to the literature review, three crucial criteria-performance, interpretability, and explainability-have been highlighted by health technology assessment organizations as being crucial for establishing trust in AI-based medical devices and are therefore essential for their evaluation. Based on the model's structure and the data at hand, suggestions were given for how and when to evaluate performance. Furthermore, methods for supporting their evaluation have been developed, taking into account the fact that interpretability and explainability can be challenging concepts to define mathematically. An estimated regulatory requirements flowchart for the development and assessment of AI-based medical devices has been made available.

In [29] Ali et al. give a summary of recent developments and trends in the field of the explainability and interpretability of AI algorithms. Using a hierarchical categorization method, the authors categorize the XAI techniques into four categories: (i) data explainability, (ii) model explainability, (iii) post-hoc explainability, and (iv) evaluation of explanations. They also provide information on existing evaluation measures, open-source software, and datasets with potential future study topics. They also discussed explainability's significance in terms of legal constraints, user viewpoints, and application orientation, which they refer to as XAI issues. The authors reviewed 410 critical publications that were published between January 2016 and October 2022 to assess XAI approaches and evaluations. The proposed framework for the end-to-end implementation of an XAI system combines evaluation approaches with design objectives, among them XAI considerations.

Finally, in [30] Band et al. looked into the uses of explainable artificial intelligence (XAI) in the healthcare industry. XAI aims to make the outcomes of artificial intelligence (AI) and machine learning (ML) algorithms in decision-making systems transparent, fair, accurate, general, and understandable. In this article, a critical evaluation of earlier research on the interpretability of ML and AI techniques in medical systems is conducted. The article also covers the potential impact of AI and ML on healthcare services.

4 Research methodology

This section describes the research methodology used in this study. It consists of five phases, which can be summarized as follows: (i) Definition of the research question, ii) Preliminary data analysis, iii) Definition of inclusion and exclusion criteria and iv) Identification of relevant studies based on inclusion and exclusion criteria v) Data Extraction and analysis. For each document analyzed, we considered: the problem addressed, the formalization of the problem, the approach used, and the challenges faced.

The first step is to define the research questions. Specifically, the following questions were considered:

- *RQ1:* How many scientific studies have been published between 2013 and 2023 regarding the interpretability and explainability of ML and DL algorithms?
- *RQ2:* What are the most relevant publication channels?
- *RQ3:* Which countries had the most active research centres?
- *RQ4:* What application areas and methods were used?
- *RQ5:* What are the most interpretability and explainability algorithms?
- *RQ6:* What metrics were used to evaluate performance?
- *RQ7:* What were the challenges addressed?

The databases we use to collect papers are those of the search engines Scopus and Web of Science.

Scopus is a comprehensive academic research database, offering abstracts and citations across various disciplines. Authors, connections, and citation trends are all covered. For transdisciplinary research, citation analysis, and evaluating the significance of articles, researchers use Scopus. The database has options for setting up alerts, extensive search capabilities, and journal analytics. Institutions often grant access, and Scopus is frequently used for literature reviews and keeping up with the most recent research [31]. Web of Science is an academic research database that provides access to a wide range of scientific and academic information. It uses information from scientific journals, conferences, patents, and other sources. It is renowned for its multidisciplinary reporting. Users can investigate relationships between scientific papers using Web of Science's sophisticated search and analysis features. It is frequently used to assess the influence of academic work, spot trends in the field, and find academic partners. Libraries or academic institutions are often the only ways to access the Web of Science [32].

To limit the scope of our research to the notions of interpretability and explainability of ML and DL algorithms, we defined the following search string: *"((explainable OR interpretable OR interpretability OR explainability) AND ((machine AND learning) OR (deep AND learning) OR (artificial AND intelligence)))"*. The search produced 26,951 results for the Scopus database and 21,633 results for the Web Of Science database. To refine the results of our analysis, we used the following inclusion and exclusion criteria.

Inclusion criteria:

- Written in English;
- Papers that had been written between 2013 and 2023;
- The study must be either a journal article, or a proceeding paper, or a book chapter.
- The focus is clearly on the interpretability and explainability of the ML and DL algorithms;
- If there are duplicate articles, the most recent version is included.

Exclusion criteria:

- If there are duplicate articles, the most recent version will be included.

Our analysis produced 23,805 results for the database Scopus and 19,709 for the database Web of Science. Subsequently, we investigated how many documents there are by subject area, as shown in Figs. 1 and 2 for the Scopus and WOS databases, respectively.

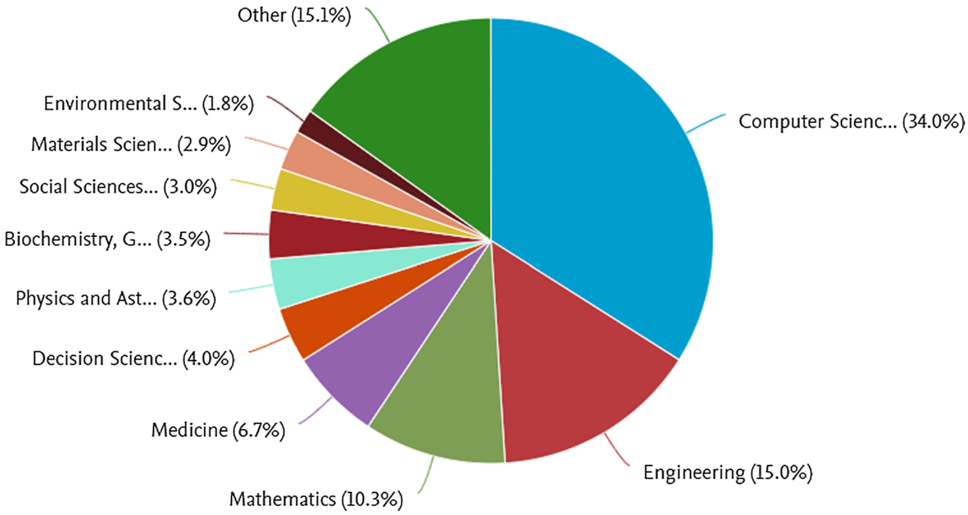

Fig. 1 Documents by subject area for the Scopus database

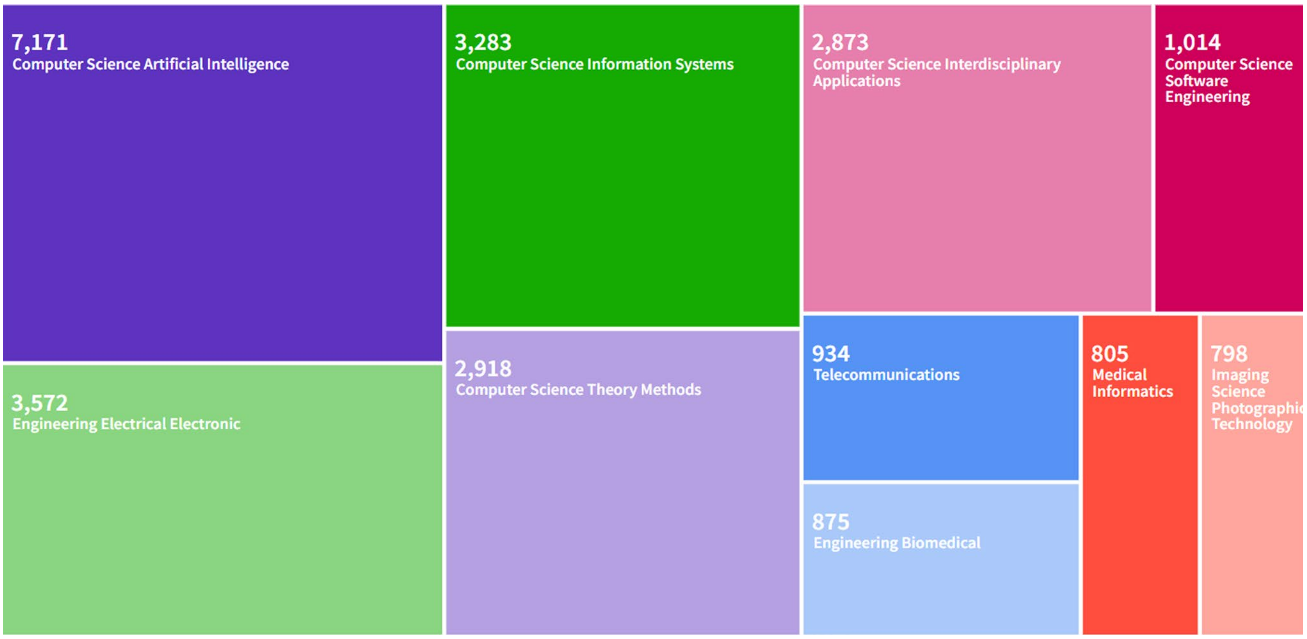

Fig. 2 Documents by subject area for the Web of Science dataset

Then we selected only those papers related to the medical area and identified 3,178 documents for the database Scopus and 805 documents for the database WOS. We compared the two datasets and took into consideration only those papers listed in both datasets. Our final dataset was composed of 448 papers.

Then we combined information on the index keywords of these documents with the number of citations and the year of publication. Specifically, we calculated the frequency of each keyword to identify the most commonly used applications and methods in the literature. For this purpose, we standardized the keywords to avoid spelling inconsistencies. These values were combined with citation count and publication year to identify the most recent relevant studies. If index keywords were missing, we used the author's keywords. For documents without authors or index keywords, we used the title as the relevant keyword. To identify the documents to be included in our analysis, we applied the following criteria:

- Occurrence of keywords (i.e., most common keywords in the papers in the dataset);
- Year of publication;
- Number of citations based on the year of publication.

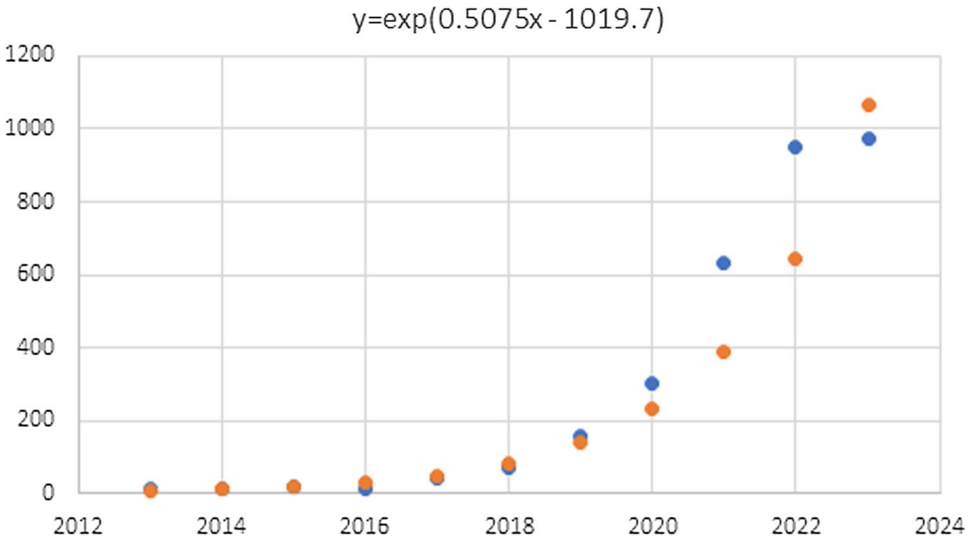

Fig. 3 Academic studies published from 2013 to 2023 in Scopus database (in blue the original data, in red the estimated values). From this analysis, we can see that the annual growth rate in Scopus is 0.5075

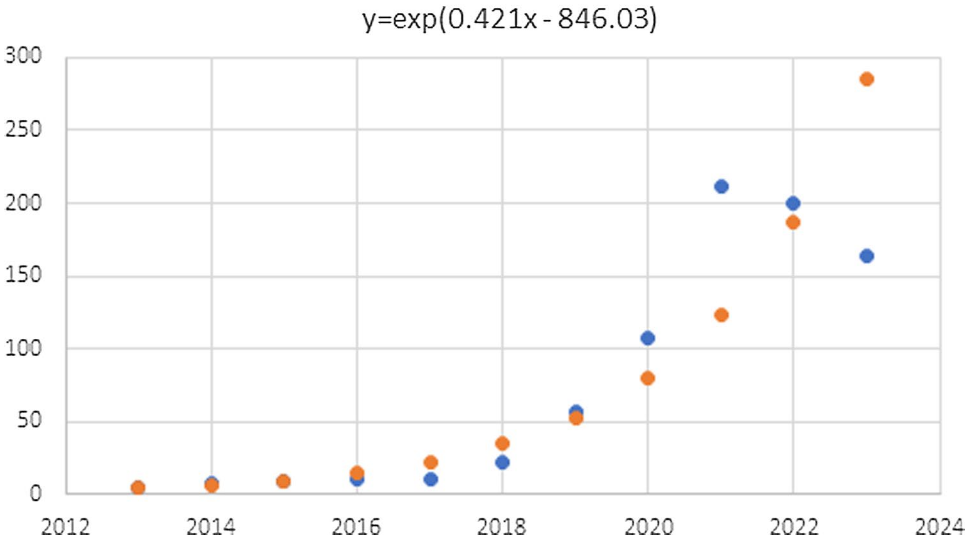

Fig. 4 Academic studies published from 2013 to 2023 in the WOS database (in blue the original data, in red the estimated values). From this analysis we can see that the annual growth rate in WoS is 0.421

Following these criteria, we identified and thoroughly examined 10 papers that addressed the research questions outlined previously. In the subsequent sections, we begin by examining the initial documents retrieved from Scopus and Web of Science (WOS) using the predefined search strings and applying specific inclusion/exclusion criteria.

5 Results

The analysis of the initial set of papers, selected based on the search strings outlined in Sect. 4, was used to address the first four research questions, namely RQ1, RQ2, RQ3, and RQ4.

5.1 RQ1: How many scientific studies have been published between 2013 and 2023 regarding the interpretability and explainability of ML and DL algorithms?

This research question aims to quantify the interest of the international scientific community in the application of AI interpretability and explainability methods in the medical field over the last 10 years. As shown in Figs. 3 and 4, the number of publications remained relatively low until 2018, with the number of publications each year being less than eighty. There has been a rapid growth of interest in this topic since 2019, reaching 974 articles for the Scopus database and 164 articles for the WOS database in 2023, demonstrating the growing interest in this topic in recent years. It is important to note that the data for the year 2023 is current as of October 2023.

We then compared the data from the two databases and extracted only the papers common to both, obtaining 448 scientific studies, and we will only consider these papers for the following analyses.

5.2 RQ2: What are the most relevant publication channels?

With this research question, we aim to show which are the main channels used for disseminating research in the application of explainability and interpretability techniques in AI. The Table 1 shows the results of our analysis.

5.3 RQ3: In which country were located the most active research centres?

This research question focuses on countries whose research centres contribute to the study of the explainability and interpretability of AI in the medical field. In Fig. 5 we only focused on countries with at least 5 publications and, as we can see, the largest number of articles came from research centres located in the United States (114 articles), followed by China (80 articles), Italy (27 articles), Spain (16 articles), United Kingdom (16 articles), South Korea (13 articles), Canada (12 articles)2, France (10 articles), Taiwan (10 articles), Netherland (8 articles), Austria (7 articles), Germany (7 articles), Singapore (7 articles), India (6 articles) and Japan (5 articles). It is important to note the presence of articles jointly written by research centres located in multiple countries. To show the relationships between co-authors, in Fig. 6, we represent the countries with at least 5 occurrences among the analyzed documents. As we can see, the countries with the most connections are the United States and China (9 connections).

5.4 RQ4: What application areas and methods were used?

This research question aims to analyze the application domains and techniques used for the explainability and interpretability of AI in the medical field. To do this, we analyzed the keywords in the index of the 448 articles that were not excluded. Figure 7 shows the application domain. We grouped keywords into macro areas and, as you can see from the image, most of the articles were classified in one of them.

Regarding the proposed approaches, we followed the same procedure as previously described for application domains grouping keywords that refer to the same method, shown in Fig. 8.

Furthermore, we performed a bibliographic analysis of the co-occurrence of the index keywords using VOSviewer [33] as shown in Figure 9. Having a co-occurrence means that 2 keywords occur in the same work. After a data cleaning process, VOSviewer detected 10 clusters by considering keywords with 3 occurrences at least. In the diagram, each cluster corresponds to a colour, and each element within a cluster corresponds to a colour. The size of the circle and the label of the circle depend on the number of occurrences of the related keyword. The lines between elements describe the co-occurrences of keywords in an article. Each cluster groups together keywords identifying an application domain and/or the approaches used to address problems related to the explainability and interpretability of AI algorithms in the medical field.

6 Analysis of the main papers

In this section, we focus on the 10 documents chosen using the selection criteria for the main documents (see Sect. 5). First, we provide a high-level analysis of the application of the domain and interpretability and explainability approaches in AI used in the medical field. Then, we give an overview of the formalization of the problem of interpretability and explainability of algorithms used in the AI field (i.e. methodologies used, type of research) (research question RQ5). Next, we analyze the performance measures used for the evaluation of the results (research question RQ6). Finally, we evaluate the main challenges faced (research question RQ7). In Table 2, we indicate for each article (first column) the year of publication (second column), the number of citations (third column), the method underlying the proposed technique (fourth column), and the dataset used for the analysis (fifth column).

The most important application domains are shown in Fig. 7. In particular, they include COVID-19, Alzheimer's disease, cardiac disease, electrocardiograms, brain and breast cancer.

Table 1 The main channels used for disseminating research in the application of explainability and interpretability techniques

Source	Number of Pubblications
IEEE Journal of Biomedical and Health Informatics	65
Computer Methods and Programs in Biomedicine	60
BMC Medical Informatics and Decision Making	56
Artificial Intelligence in Medicine	47
Journal of Biomedical Informatics	43
npj Digital Medicine	26
Journal of the American Medical Informatics Association	15
International Journal of Medical Informatics	14
Journal of Medical Internet Research	12
Frontiers in Digital Health	9
Proceedings—2021 IEEE 9th International Conference on Healthcare Informatics, ISCHI 2021	8
JAMIA Open	7
Digital Health	7
Statistics in Medicine	6
Proceedings - IEEE 20th International Conference on Bioinformatics and Bioengineering, BIBE 2020	6
The Lancet Digital Health	5
Proceedings—2022 IEEE 10th International Conference on Healthcare Informatics, ICHI 2022	5
Medical Decision Making	5
Journal of Medical Systems	5
JMIR Formative Research	5
Progress in Biomedical Optics and Imaging - Proceedings of SPIE	4
Journal of Healthcare Informatics Research	4
JMIR Medical Informatics	4
BIBE 2021–21st IEEE International Conference on BioInformatics and BioEngineering, Proceedings	4
Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, BCB 2020	3
Proceedings—2023 IEEE International Conference on Digital Health, ICDH 2023	3
2022 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology, CIBCB 2022	3
JMIR mHealth and uHealth	2
Health Information Science and Systems	2
Health Informatics Journal	2
BMJ Health and Care Informatics	2
2020 IEEE International Conference on Healthcare Informatics, ICHI 2020	2
Statistical Methods in Medical Research	1
Proceedings—2023 IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies, CHASE 2023	1
Journal of Evaluation in Clinical Practice	1
Clinical and Experimental Emergency Medicine	1
Applied Clinical Informatics	1
ACM-BCB 2018—Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics	1
2019 IEEE International Conference on Healthcare Informatics, ICHI 2019	1

In Table 2 we also show that the main techniques used, that are neural networks [34–37] (such as CNN, LSTM and Boltzmann machine) [38, 39] followed by other machine learning algorithms (such as K-nearest neighbours, Logistic Regression, Naïve Bayes, Random Forest and Support Vector Machines).

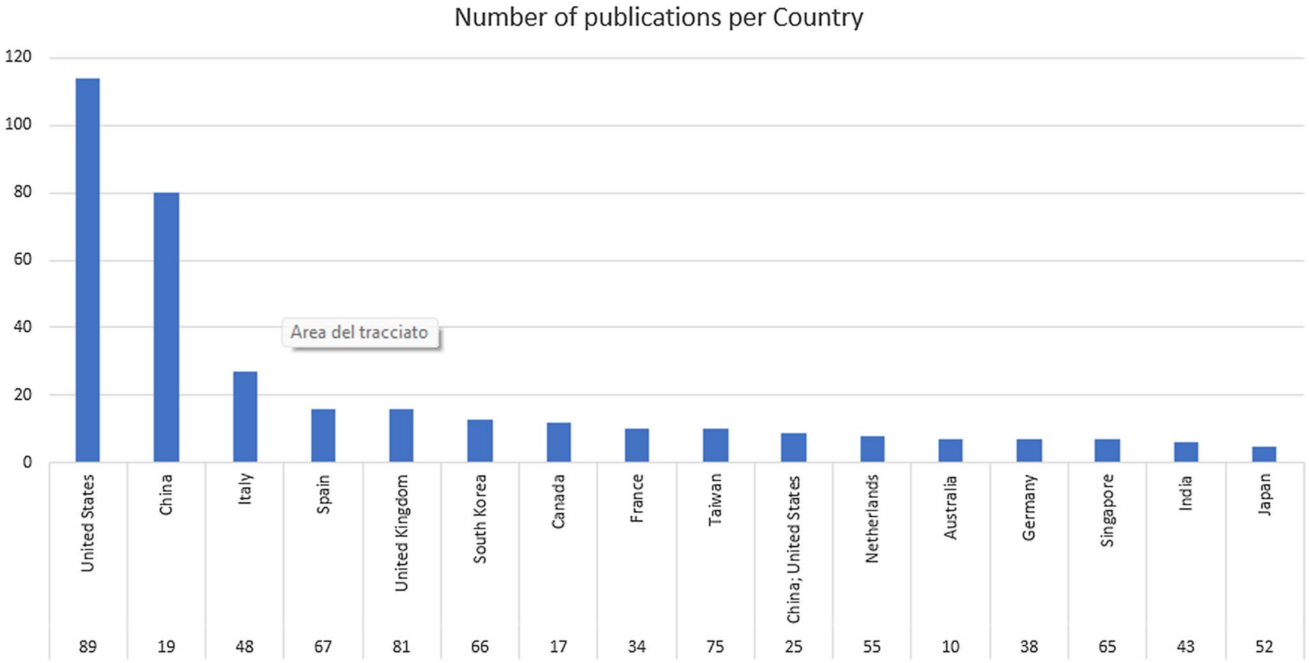

Fig. 5 Number of publications per Country on explainability and interpretability of AI in the medical field

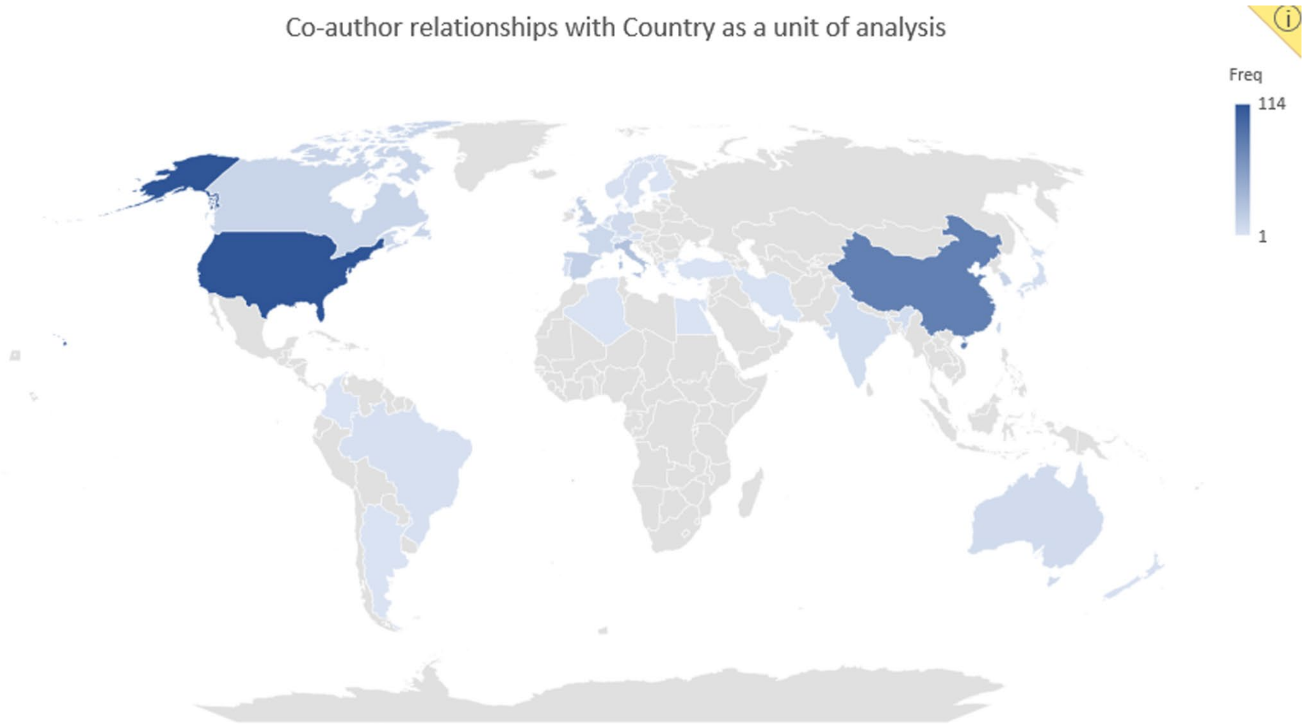

Fig. 6 Co-author relationships with country as a unit of analysis

6.1 RQ5: What are the most interpretability and explainability algorithms?

The focus of this research question is on technical aspects that may prove beneficial for practitioners in gaining insight into the environment explored by the authors. The table encapsulates the problem formulation details extracted from

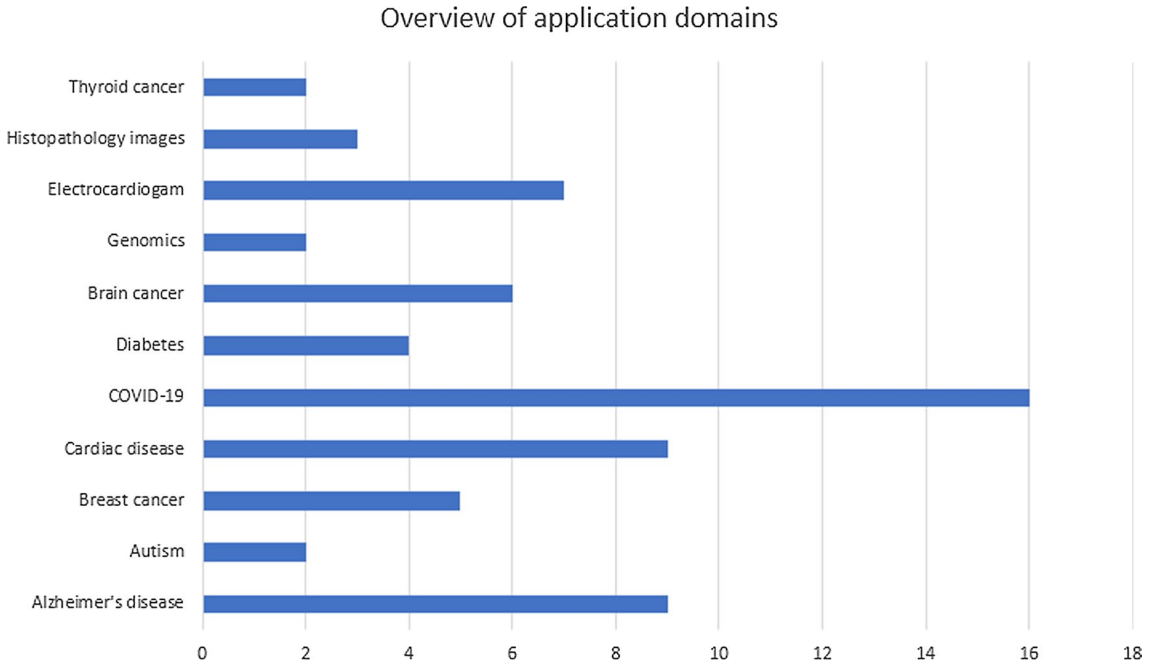

Fig. 7 Overview of the application domains

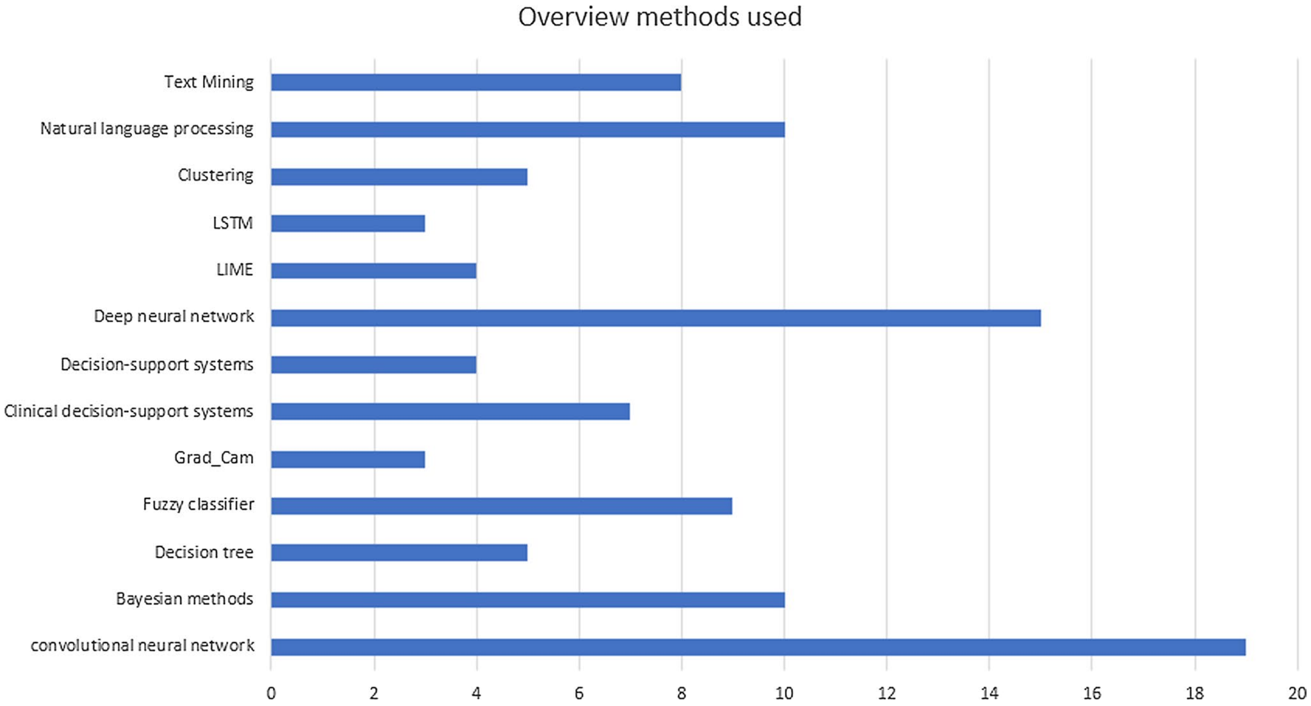

Fig. 8 Overview of the methods used

the chosen articles. Each paper is scrutinized to discern the interpretability and explainability techniques employed. Additionally, insights are furnished regarding the datasets utilized in the experiments, delineating between real and synthetic data. It is important to note that not all documents explicitly provide this information. Consequently, instances where the dataset information is unspecified are denoted with "N/A." The most used techniques are:

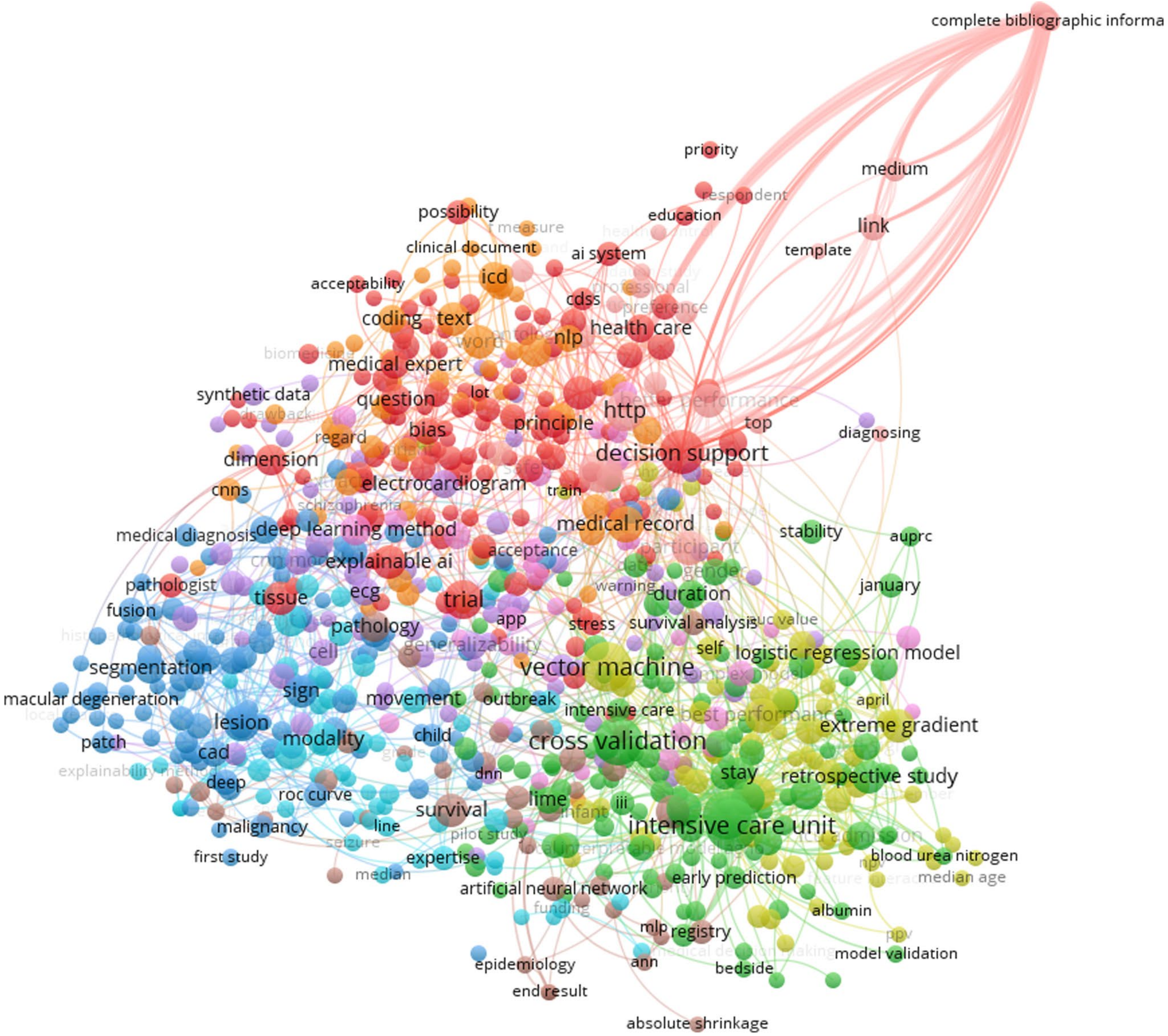

Fig. 9 Bibliometrics analysis on the co-occurrence of index keywords

- *Decision support systems:* or DSS are capable of converting the output of these algorithms into comprehensible graphics. The aspects that have the greatest impact on decisions are highlighted using graphs, heat maps, and other visual representations. A well-designed DSS also provides a user interface that is simple to use even for those without a thorough understanding of machine learning [7, 38].
- *Gradient-weighted Class Activation Mapping:* or Grad-Cam useful for highlighting the parts of an image given as input to CNN that have most influenced the network's decision in recognizing a specific class of objects. It evaluates the gradients of the output level of the network concerning the class we are examining. These gradients tell us how much each part of the image contributed to the decision. Then, these gradients are used to weigh various aspects of the image, particularly those of the last convolutional layer of the network. The result is an activation map that can be overlaid on the original image. This map visually tells us which regions of the image were crucial for the network in making its decision [34].
- *Shapley Additive Expansion* or SHAP is an advanced explainability technique that helps us decompose and understand complex model decisions, it is based on cooperative game theory. To evaluate the importance of a variable, SHAP performs random permutations of the variables, evaluating how the model's predictions change compared to the original input. Each value represents how much each variable contributes on average to the model's predictions. The

Table 2 Technical information about the selected works

Paper	Year	Citations	Method	Dataset
Explainable Deep Learning for Pulmonary Disease and Coronavirus COVID-19 Detection from X-rays [34]	2020	375	VGG-16 CNN and Grad-CAM, Activation map	[52–54]
Explainability for artificial intelligence in healthcare: a multidisciplinary perspective [7]	2020	308	Clinical decision support systems	N/A
Deep learning interpretation of echocardiograms [35]	2020	201	EchoNet CNN	EchoNet-Dynamic Cardiac Ultrasound
Detection of COVID-19 Infection from Routine Blood Exams with Machine Learning: A Feasibility Study [38]	2020	194	Decision Tree, Extremely Randomized Trees, K-nearest neighbours, Logistic Regression, Naïve Bayes, Random Forest, Support Vector Machines, Decision support systems	IRCCS Ospedale San Raffaele
Explainable artificial intelligence for breast cancer: A visual case-based reasoning approach [39]	2019	156	kNN, Rainbow boxes-inspired algorithm, Distance-weighted kNN, Case-Based Reasoning, MDS	Breast Cancer Wisconsin
Dynamic and explainable machine learning prediction of mortality in patients in the intensive care unit: a retrospective study of high-frequency data in electronic patient records [36]	2020	130	LSTM, Shapley Additive Expansion	Electronic patient records (EPRs), the Central Person Registry (CPR), and the Danish National Patient Registry (DNPR)
On the interpretability of machine learning-based model for predicting hypertension [40]	2019	124	Feature Importance, Partial Dependence Plot, Individual Conditional Expectation, Feature Interaction, Global Surrogate Models, Local Surrogate Models, Shapley Value	Henry Ford Health Systems
Learning vector representation of medical objects via EMR-driven nonnegative restricted Boltzmann machines (eNRBM) [55]	2015	109	Restricted Boltzmann machine, EMR-driven nonnegative RBM	[37, 56]
Electronic medical record phenotyping using the anchor and learn framework [45]	2016	98	Anchor and learn framework	Trauma centre and tertiary academic teaching hospital
A new concordant partial AUC and partial c statistic for imbalanced data in the evaluation of machine learning algorithms [46]	2020	90	AUC	Breast cancer the Wisconsin and Ljubljana data sets

final result is a SHAP graph, which shows, for each prediction of the model, how much each variable influenced the decision [36, 40].

- *Local Interpretable Model-agnostic Explanations* or LIME to explain a specific prediction of the model select a sample of data similar to the one we are examining. Then, it introduces some perturbations, or small random changes, to these characteristics to create a "perturbed" data set. It then uses the machine learning model to make predictions about these perturbed instances, and the resulting predictions are weighted based on how similar each perturbed instance is to the original [40].

6.2 RQ6: What metrics were used to evaluate performance?

This research question aims to provide an overview of the metrics used to measure the performance of the algorithms utilized in the 10 selected papers. In the second column of the Table, we report information on the metrics found in the articles. As we can see in Table 3, performance measures vary widely depending on the application domain and the objective of the method proposed in each article. But the most used metrics are:

- *Accuracy:* is an indicator of how well the model can correctly classify instances in the dataset and is the ratio of correct predictions to total predictions. In ML and DL algorithms, accuracy indicates how close measurements are to the true value [41].
- *Precision:* is a measure that provides information on the quality of positive predictions made by a model. Thus, precision is calculated as the ratio of true positives to the sum of true positives and false positives, that is, how precise the model is in declaring objects positive. [42].
- *Recall:* indicates the ability of a model to correctly identify all the positive instances present in a dataset. Recall is obtained by dividing the number of true positives by the sum of true positives and false negatives. In essence, recall gives us an idea of how sensitive the model is in detecting positive examples while trying to minimize false negatives [43].
- *AUC:* evaluate the discrimination capacity of a classification model. The AUC represents the area under this ROC curve. It measures how well the model can distinguish between positive and negative classes [44].

6.3 RQ7: What were the challenges addressed?

Artificial intelligence (AI) in the medical field has revolutionized the approach to the diagnosis, treatment and management of pathologies, posing new challenges in personalizing care. This innovative panorama, as we already specified in the previous sections, raises crucial questions about the transparency, interpretability and ethical executability of AI predictive and decision-making models. To address these challenges, we have identified and reviewed 10 papers in the literature that employ cutting-edge methodologies. From rapidly identifying COVID-19 using X-ray images to enhancing the explainability of AI-driven clinical decision support systems, these studies provide a comprehensive overview of the advancements and obstacles in the integration of artificial intelligence in health and medicine. In the third column of Table 3, we summarize the challenges discussed in these papers. Similar to performance metrics, Table 3 illustrates the wide range of challenges encountered, influenced by the specific application context and objectives of each method proposed in the literature. Below we explore in detail all the papers taken into consideration.

Brunese et al. [34] address their challenge by proposing the use of deep learning for the automatic and rapid diagnosis of coronavirus disease (COVID-19) using chest X-rays. The proposed methodology consists of three phases. The first phase is used to detect pneumonia, the second phase is used to distinguish between pneumonia and COVID-19, and the last phase is used to identify areas with COVID-19 symptoms on the x-ray used. Experimental results from 6,523 chest x-rays showed effectiveness with an average time to detect COVID-19 infection of 2.5 s and an average accuracy of 97%. The approach uses transfer learning with the VGG-16 model and has 96% accuracy in differentiating healthy patients from those with common lung diseases, and 98% accuracy has been achieved in detecting COVID-19.

Amann et al. [7] consider the problem of explainability in the field of AI in the healthcare sector, focusing on AI-based clinical decision support systems. This study takes a multidisciplinary approach by approaching from technical, legal, medical and patient perspectives and analyzes the importance of the explainability of AI in medicine. The findings highlight the risks to individual and public health when explainability is omitted from clinical decision support systems.

Ghorbani et al. [35] propose the application of convolutional neural networks (CNNs) to echocardiographic images for cardiac analysis. The deep learning model used called EchoNet identifies local cardiac structures, estimates cardiac

Table 3 Performance measures used by the authors to evaluate the approaches challenges faced

Paper	Metrics	Challenges
Explainable Deep Learning for Pulmonary Disease and Coronavirus COVID-19 Detection from X-rays [34]	Sensitivity, Specificity, F-Measure, Accuracy, Time	Provide fully automatic and faster diagnosis by adopting deep learning for COVID-19 detection from X-rays
Explainability for artificial intelligence in healthcare: a multidisciplinary perspective [7]	Qualitative evaluation	Provide a comprehensive assessment of the role of explainability in medical AI and make an ethical assessment of what explainability means for the adoption of AI tools in clinical practice
Deep learning interpretation of echocardiograms [35]	Accuracy and AUC	Provide preliminary interpretation in areas with insufficient numbers of trained cardiologists to predict difficult phenotypes by quantifying cardiac function
Detection of COVID-19 Infection from Routine Blood Exams with Machine Learning: A Feasibility Study [38]	Accuracy, Balanced accuracy, Positive Predictive Value, Sensitivity, Specificity, AUC, Decision tree	Demonstrate the feasibility and clinical robustness of using blood test analysis and machine learning as an alternative to rRT-PCR to identify COVID-19 positive patients
JExplainable artificial intelligence for breast cancer: A visual case-based reasoning approach [39]	Accuracy	Propose a CBR method that can be automatically executed as an algorithm and presented visually in a user interface to provide visual explanations or for visual reasoning
Dynamic and explainable machine learning prediction of mortality in patients in the intensive care unit: a retrospective study of high-frequency data in electronic patient records [36]	MCC, Positive likelihood contribution	LSTM, Shapley Additive Expansion
Dynamic and explainable machine learning prediction of mortality in patients in the intensive care unit: a retrospective study of high-frequency data in electronic patient records [36]	MCC, Positive likelihood contribution, Negative likelihood contribution	LTo test whether machine learning methods using time series data analyzes improved mortality prognosis for ICU patients by providing real-time predictions of 90-day mortality
On the interpretability of machine learning-based model for predicting hypertension [40]	Accuracy	FDemonstrate the utility of various model-agnostic explainability techniques of machine learning models to analyze random forest model results to predict individuals' risk of developing hypertension based on cardiorespiratory fitness data
Learning vector representation of medical objects via EMR-driven nonnegative restricted Boltzmann machines (eNRBM) [55]	Recall, Precision, F-measure	Present a computational framework to exploit EMR with minimal human supervision via a bounded Boltzmann machine
Electronic medical record phenotyping using the anchor and learn framework [45]	AUC	Develop a phenotype library that uses structured and unstructured data from the EMR to represent patients for real-time clinical decision support
A new concordant partial AUC and partial c statistic for imbalanced data in the evaluation of machine learning algorithms [46]	AUC	Propose the alternatives AUC, such as the partial AUC and the area under the precision-recall curve

performance indicators and predicts systemic phenotypes such as age, gender, weight and height that can influence cardiovascular risk. In this study, This study suggests that integrating interpretive frameworks can help identify regions of interest, contribute to a better understanding of normal echocardiogram variations, and reveal features missed by clinicians.

Brinati et al. [38] suggest a novel strategy for identifying COVID-19-positive patients using machine learning models based on common blood tests. The authors created two machine-learning models that consider standard data from blood tests and elements like age and gender. The algorithms' accuracy ranged from 82% to 86%, while their sensitivity ranged from 92% to 95%. The model's interpretability is based on a decision tree that also received medical approval, demonstrating the reliability of the chosen characteristics.

Lamy et al. [39] present a technique for breast cancer that uses "Case-Based Reasoning" and functions as both an automatically executable algorithm and a graphical user interface for explanations. CBR allows for easily justified results using analogous cases as examples, in contrast to "black box" methods. A scatterplot based on multidimensional dimension reduction is used in the visual interface. The visual interface combines polar-MDS scatterplots with "rainbow boxes" to transform the CBR problem into a "colour dominance" challenge.

Thorsen et al. [36] use machine learning models based on temporal data, enabling real-time predictions, to enhance the 90-day death prognosis for ICU patients. One hour of time resolution is used to train an LSTM neural network model. With a Matthews correlation coefficient and area under the ROC curve increasing from 0.29 and 0.73 at admission to 0.57 and 0.88 at discharge, the results demonstrate that predictive performance increases over time. Input from a fifth hospital's data is used to externally validate the model. To explain predictions, the Shapley algorithm is used to pinpoint the traits that influence predictions at various time steps.

Elshawi et al. [40] discuss how to improve the interpretability of machine learning models, particularly those that use data from cardiorespiratory fitness to predict the risk of developing hypertension using a random forest model. Different model-agnostic explanation strategies, categorized as global and local interpretation strategies, are used. They pointed out that although local interpretations concentrate on particular cases, global interpretations aid clinicians in comprehending the overall conditional distribution described by the response function. The authors draw attention to the fact that, depending on the demands of the application, both global and local techniques may be appropriate. LIME offers local explanations based on outdated data points and regional regression models. Even though it allows doctors to make judgments about how the patient's characteristics have changed over time, LIME is criticized for its instability and the addition of linearity in the local model. Shapley's value prediction, on the other hand, divides the difference between the median prediction and the estimated distance between the characteristic values. Despite providing an equal distribution of contributions, the model is computationally expensive and requires access to the addenda used to add the model.

Tran et al. [37] describe a computational framework that makes use of a non-negative restrictive Boltzmann machine to utilize electronic medical record data with little to no human oversight. By embedding medical objects in a low-dimensional vector space, this framework produces a new representation of those objects. This model enables algebraic and statistical operations including item grouping, risk stratification, and projection onto a 2D plane. Two requirements are added to the model parameters: (a) non-negative coefficients, and (b) structural regularity, to enhance the model's interpretability. The generated representation aids in short-term risk classification and displays clusters of traits that are clinically significant.

Halpern et al. [45] concentrated on the value of Electronic Medical Records (EMR) in identifying the best methods for patient care. The suggested approach intends to effectively acquire statistically driven phenotypes with little manual assistance. To support clinical choices in real-time, a phenotype library was created that represents patients using structured and unstructured data from EMRs. Concerning prospectively acquired baseline data, eight of these phenotypes were assessed using retrospective data on emergency room patients. The findings demonstrate that the resulting phenotypes are interpretable, quick to construct, and perform as well as phenotypes learnt statistically from a large number of manual labelling.

The Receiver Operating Characteristic (ROC) curves and the area under the ROC curve (AUC) are used by Carrington et al. [46] to evaluate the effectiveness of classification models and diagnostic tests. We focus in particular on the issue of employing unbalanced data, when positive and negative classes are not equally represented. For ROC data, the authors suggest a new concordant partial AUC and a new partial c statistic, which are crucial metrics and approaches for comprehending and interpreting certain ROC curves and AUC regions. These new partial measures are validated for their equivalence and are obtained from the AUC and c statistics, respectively. They are examined using two authentic breast cancer datasets as well as a traditional ROC example.

7 Discussion

This paper aims to evaluate the scientific community's interest in applying AI interpretability and explainability methods to the medical field over the past decade. After analyzing publication trends from 2013 to 2023, we have observed a significant increase in interest, especially from 2019, with 974 articles indexed by Scopus and 164 indexed by WOS in 2023. Our study was based on 448 common articles. Our analysis included trends in publications, dissemination channels, contributing countries and collaboration networks. Significant contributors were identified, with the United States leading the list (114 articles), followed by China, Italy, and others.

To sharpen our conclusions, we focused our analysis on 10 selected documents. Major application domains include COVID-19, Alzheimer's disease, cardiac disease, electrocardiograms, and brain and breast cancer, reflecting the impact of these topics on society. We also delved into the most used AI algorithms such as CNN, LSTM and the Boltzmann machine and the most used techniques for explainability and interpretability such as decision support systems, Grad-Cam, SHAP and LIME.

According to our research findings, the rapid integration of Machine Learning (ML) and Deep Learning (DL) techniques has led to growing concerns about these methodologies. There is an increasing apprehension regarding the complexity of AI algorithms and the absence of transparency in data mining and decision-making processes.

The literature analysis conducted in this study reveals that the majority of research on interpretability and explainability in AI algorithms within the medical field focuses on neural networks and machine learning algorithms. Many of the examined papers employ interpretability techniques for AI algorithms, with several of them emphasizing the close connection between interpretability and explainability [10].

Indeed, in certain instances, interpretability and explainability are frequently treated as interchangeable terms. The scientific community needs to establish clear definitions and unambiguous vocabulary to facilitate the transfer of results and information more effectively. AI models currently lack a formal structure that would assist users in gaining confidence in the employed techniques.

To the best of our knowledge, to date, to improve explainability and interpretability algorithms, the scientific community is working on:

- *Built-in interpretability:* ML and DL algorithms are designed so that they are inherently more interpretable and understandable. The goal is to make clear the reasons behind the decisions made by a model, thus reducing the opacity of complex models. We also try to design these algorithms seeking a balance between interpretability and performance [47].
- *Self-interpretable models:* involves the creation of architectures or models that are understandable in themselves, without the need for post hoc techniques. These models aim to provide transparency in their decisions right from the design and training phase [48] [49].
- *Post hoc explainability methods:* are techniques used to interpret and explain already trained AI models, which may be complex and difficult to understand. These methods attempt to provide retroactive explanations about the decisions made by a model, without affecting its original learning process [50].
- *Stakeholder involvement:* To ensure greater acceptance and understanding, there was growing interest in involving doctors, patients and other stakeholders in the design and interpretation of models [51].

8 Conclusion

Medical diagnosis, therapy, and decision-making rely heavily on ML and DL algorithms, yet the increasing complexity of these algorithms presents challenges, particularly in balancing high performance with interpretability. Transparent decision-making, ethical and regulatory compliance, and trust from healthcare and medical providers are essential for deploying AI models in these domains. Accuracy is paramount, prompting careful consideration of whether to prioritize accuracy or explainability and interpretability given the contrast between interpretable white-box and non-interpretable black-box techniques. Interpretability and explainability are crucial for establishing user trust, ensuring legal and ethical compliance, and fostering broader social acceptance. They transcend technical concerns, becoming moral and practical imperatives in the medical environment, where algorithmic decisions have significant effects. Assessing prior research highlights the demand for explainable and interpretable AI algorithms in medicine. Despite progress in tools such as visualization, feature interpretation, and model complexity reduction, challenges persist. These include the opacity of large models, trade-offs between performance and interpretability, and biases in training data. The ethical significance of AI in healthcare and medicine cannot be ignored. Addressing concerns such as bias, privacy, security, and legal liability is essential to ensure the ethical use of AI. On the other side, the psychological impact of algorithmic judgments on users, especially in healthcare and medicine, is linked to clear communication about the capabilities and limitations of artificial intelligence systems. Takeaway lessons include the need for a strong commitment to transparency, fairness, and patient-centred design of AI models in health and medicine as well as establishing precise criteria and standards to enhance trust among users and stakeholders.

Author contributions The authors contributed equally to this paper. All authors reviewed the manuscript.

Data availability Our data have been freely downloaded from Scopus and Web of Science using the keywords mentioned in the paper. The dataset used in our analysis can be easily replicated.

Declarations

Competing interests The authors declare no competing interests.

References

1. London AJ. Artificial intelligence and black-box medical decisions: accuracy versus explainability. Hastings Cent Rep. 2019;49(1):15–21.
2. Hakkoum H, Abnane I, Idri A. Interpretability in the medical field: a systematic mapping and review study. Appl Soft Comput. 2022;117: 108391.
3. Loyola-Gonzalez O. Black-box vs. white-box: understanding their advantages and weaknesses from a practical point of view. IEEE Access. 2019;7:154096–113.
4. Kolasinska A, Lauriola I, Quadrio G. Do people believe in artificial intelligence? a cross-topic multicultural study. In Proceedings of the 5th EAI International Conference on Smart Objects and Technologies for Social Good, 2019:31–6.
5. Gilvary C, Madhukar N, Elkhader J, Elemento O. The missing pieces of artificial intelligence in medicine. Trends Pharmacol Sci. 2019;40(8):555–64.
6. General Data Protection Regulation. General data protection regulation (GDPR). Intersoft Consulting. Accessed in October, 2018;24(1).
7. Amann J, Blasimme A, Vayena E, Frey D, Madai VI. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Med Inform Decis Mak. 2020;20(1):1–9.
8. Phillips PJ, Hahn CA, Fontana PC, Broniatowski DA, Przybocki MA. Four principles of explainable artificial intelligence, vol. 18. Gaithersburg: National Institute of Standards and Technology; 2020.
9. Nassih Rym, Berrado Abdelaziz. State of the art of fairness, interpretability and explainability in machine learning: Case of prim. In Proceedings of the 13th International Conference on Intelligent Systems: Theories and Applications, 2020:1–5.
10. Linardatos P, Papastefanopoulos V, Kotsiantis S. Explainable AI: a review of machine learning interpretability methods. Entropy. 2020;23(1):18.
11. Alicioglu G, Sun B. A survey of visual analytics for explainable artificial intelligence methods. Comput Graph. 2022;102:502–20.
12. Shickel B, Tighe PJ, Bihorac A, Rashidi P. Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. IEEE J Biomed Health Inform. 2017;22(5):1589–604.
13. Shashanka M, Raj B, Smaragdis P. Sparse overcomplete latent variable decomposition of counts data. Adv Neural Inform Process Syst, 2007;20.
14. Ribeiro MT, Singh S, Guestrin C. "why should i trust you?" explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016:1135–44.
15. Langer M, Oster D, Speith T, Hermanns H, Kästner L, Schmidt E, Sesing A, Baum K. What do we want from explainable artificial intelligence (XAI)?—a stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. Artif Intell. 2021;296: 103473.
16. Murdoch WJ, Singh C, Kumbier K, Abbasi-Asl R, Yu B. Definitions, methods, and applications in interpretable machine learning. Proc Natl Acad Sci. 2019;116(44):22071–80.
17. Combi C, Amico B, Bellazzi R, Holzinger A, Moore JH, Zitnik M, Holmes JH. A manifesto on explainability for artificial intelligence in medicine. Artif Intell Med. 2022;133: 102423.
18. von Eschenbach WJ. Transparency and the black box problem: why we do not trust ai. Philos Technol. 2021;34(4):1607–22.
19. Arrieta AB, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, García S, Gil-López S, Molina D, Benjamins R, et al. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible ai. Inf Fus. 2020;58:82–115.
20. Biran O, Cotton C. Explanation and justification in machine learning: a survey. In IJCAI-17 workshop on explainable AI (XAI), 2017;8:8–13.
21. Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D. A survey of methods for explaining black box models. ACM Comput Surv (CSUR). 2018;51(5):1–42.
22. Miller T. Explanation in artificial intelligence: insights from the social sciences. Artif Intell. 2019;267:1–38.

23. Tjoa E, Guan C. A survey on explainable artificial intelligence (XAI): Toward medical XAI. IEEE Trans Neural Netw Learn Syst. 2020;32(11):4793–813.
24. Stiglic G, Kocbek P, Fijacko N, Zitnik M, Verbert K, Cilar L. Interpretability of machine learning-based prediction models in healthcare. Wiley Interdiscip Rev Data Min Knowl Discov. 2020;10(5): e1379.
25. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A survey on bias and fairness in machine learning. ACM Comput Surv(CSUR). 2021;54(6):1–35.
26. Chakrobartty S, El-Gayar O. Explainable artificial intelligence in the medical domain: a systematic review. 2021.
27. Hatherley J, Sparrow R, Howard M. The virtues of interpretable medical artificial intelligence. Camb Q Healthc Ethics, 2022:1–10.
28. Farah L, Murris JM, Borget I, Guilloux A, Martelli NM, Katsahian SIM. Assessment of performance, interpretability, and explainability in artificial intelligence-based health technologies: what healthcare stakeholders need to know. Mayo Clin Proc. 2023;1(2):120–38.
29. Ali S, Abuhmed T, El-Sappagh S, Muhammad K, Alonso-Moral JM, Confalonieri R, Guidotti R, Del Ser J, Díaz-Rodríguez N, Herrera F. Explainable artificial intelligence (XAI): What we know and what is left to attain trustworthy artificial intelligence. Inf Fusion. 2023;99: 101805.
30. Band SS, Yarahmadi A, Hsu C-C, Biyari M, Sookhak M, Ameri R, Dehzangi I, Chronopoulos AT, Liang H-W. Application of explainable artificial intelligence in medical health: A systematic review of interpretability methods. Inform Med Unlocked. 2023;40: 101286.
31. Ballew BS. Elsevier's scopus® database. J Electron Resour Med Libr. 2009;6(3):245–52.
32. Drake M. Encyclopedia of library and information science, vol. 1. Boca Raton: CRC Press; 2003.
33. Van Eck N, Waltman L. Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics. 2010;84(2):523–38.
34. Brunese L, Mercaldo F, Reginelli A, Santone A. Explainable deep learning for pulmonary disease and coronavirus covid-19 detection from x-rays. Comput Methods Programs Biomed. 2020;196: 105608.
35. Ghorbani A, Ouyang D, Abid A, He B, Chen JH, Harrington RA, Liang DH, Ashley EA, Zou JY. Deep learning interpretation of echocardiograms. NPJ Digit Med. 2020;3(1):10.
36. Thorsen-Meyer H-C, Nielsen AB, Nielsen AP, Kaas-Hansen BS, Toft P, Schierbeck J, Strøm T, Chmura PJ, Heimann M, Dybdahl L, et al. Dynamic and explainable machine learning prediction of mortality in patients in the intensive care unit: a retrospective study of high-frequency data in electronic patient records. Lancet Digit Health. 2020;2(4):e179–91.
37. Tran T, Luo W, Phung D, Harvey R, Berk M, Kennedy RL, Venkatesh S. Risk stratification using data from electronic medical records better predicts suicide risks than clinician assessments. BMC Psychiatry. 2014;14(1):76.
38. Brinati D, Campagner A, Ferrari D, Locatelli M, Banfi G, Cabitza F. Detection of covid-19 infection from routine blood exams with machine learning: a feasibility study. J Med Syst. 2020;44:1–12.
39. Lamy J-B, Sekar B, Guezennec G, Bouaud J, Séroussi B. Explainable artificial intelligence for breast cancer: a visual case-based reasoning approach. Artif Intell Med. 2019;94:42–53.
40. Elshawi R, Al-Mallah MH, Sakr S. On the interpretability of machine learning-based model for predicting hypertension. BMC Med Inform Decis Mak. 2019;19(1):1–32.
41. Menditto A, Patriarca M, Magnusson B. Understanding the meaning of accuracy, trueness and precision. Accredit Qual Assur. 2007;12:45–7.
42. Prenesti E, Gosmaro F. Trueness, precision and accuracy: a critical overview of the concepts as well as proposals for revision. Accredit Qual Assur. 2015;20:33–40.
43. Buckland M, Gey F. The relationship between recall and precision. J Am Soc Inform Sci. 1994;45(1):12–9.
44. Huang J, Ling CX. Using AUC and accuracy in evaluating learning algorithms. IEEE Trans Knowl Data Eng. 2005;17(3):299–310.
45. Halpern Y, Horng S, Choi Y, Sontag D. Electronic medical record phenotyping using the anchor and learn framework. J Am Med Inform Assoc. 2016;23(4):731–40.
46. Carrington AM, Fieguth PW, Qazi H, Holzinger A, Chen HH, Mayr F, Manuel DG. A new concordant partial AUC and partial c statistic for imbalanced data in the evaluation of machine learning algorithms. BMC Med Inform Decis Mak. 2020;20:1–12.
47. Mariotti E, Moral JMA, Gatt A. Exploring the balance between interpretability and performance with carefully designed constrainable neural additive models. Inf Fus. 2023;99: 101882.
48. Ashwath VA, Sikha OK, Benitez R. TS-CNN: a three-tier self-interpretable CNN for multi-region medical image classification. IEEE Access; 2023.
49. La Rosa B, Capobianco R, Nardi D. A self-interpretable module for deep image classification on small data. Appl Intell. 2023;53(8):9115–47.
50. Dwivedi R, Dave D, Naik H, Singhal S, Omer R, Patel P, Qian B, Wen Z, Shah T, Morgan G, et al. Explainable AI (XAI): core ideas, techniques, and solutions. ACM Comput Surv. 2023;55(9):1–33.
51. Anwar SM. Expert systems for interpretable decisions in the clinical domain. In: Byrne MF, Parsa N, Greenhill AT, Chahal D, Ahmad O, Bagci U, editors. AI in clinical medicine: a practical guide for healthcare professionals. Hoboken: Wiley Online Library; 2023. p. 66–72.
52. Cho B-J, Choi YJ, Lee M-J, Kim JH, Son G-H, Park S-H, Kim H-B, Joo Y-J, Cho H-Y, Kyung MS, et al. Classification of cervical neoplasms on colposcopic photography using deep learning. Sci Rep. 2020;10(1):1–10.
53. Ozturk T, Talo M, Yildirim EA, Baloglu UB, Yildirim O, Acharya UR. Automated detection of covid-19 cases using deep neural networks with x-ray images. Comput Biol Med. 2020;121: 103792.
54. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. Chestx-ray8: hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2017:2097–106.
55. Tran T, Nguyen TD, Phung D, Venkatesh S. Learning vector representation of medical objects via EMR-driven nonnegative restricted Boltzmann machines (eNRBM). J Biomed Inform. 2015;54:96–105.
56. Tran T, Phung D, Luo W, Venkatesh S. Stabilized sparse ordinal regression for medical risk stratification. Knowl Inform Syst. 2015;43:555–82.