

ORIGINAL RESEARCH

Open access

A Decade of Explainable and Interpretable Artificial Intelligence in Healthcare: A Bibliometric Perspective

Charlotte Dupuis^{1*}, Hugo Perrin¹, Elise Morel², Thomas Martin³

Abstract

This review investigates the expanding influence of machine learning and deep learning methodologies within the healthcare sector, concentrating specifically on the vital challenges of interpretability and explainability inherent in black-box models. Although machine learning frameworks are increasingly deployed to assist in medical analytics and diagnostics, their intricate nature emphasizes the necessity of comprehending how these systems process and interpret data to facilitate well-informed clinical choices. This paper offers a thorough examination of the obstacles and potential solutions documented in existing scholarship, delivering a synthesis of contemporary techniques in this domain. Furthermore, it establishes precise conceptual boundaries for interpretability and explainability, seeking to untangle the differences between these notions and outline their consequences for clinical decision-making. Utilizing a dataset of 448 publications to address seven distinct research questions, our investigation demonstrates that this research area has experienced exponential growth over the past ten years. Additionally, the psychological factors governing public opinion highlight the vital need for transparent communication regarding the capabilities and constraints of artificial intelligence. To mitigate opacity, investigators are actively engineering strategies to lower model complexity and implement visual explanation tools. Nevertheless, maintaining a delicate equilibrium between high predictive performance and adequate model interpretability remains an ongoing challenge. Given the expanding role of artificial intelligence in supporting clinical diagnostics and therapeutics, the development of inherently interpretable artificial intelligence systems is critical. Within this rapidly changing paradigm, a steadfast dedication to transparency, ethical standards, and cross-disciplinary partnerships is essential to guarantee the conscientious deployment of artificial intelligence. This shared dedication is foundational to cementing long-term trust between medical professionals and patients, resolving novel problems, and promoting the educated integration of these sophisticated technologies into medical practices.

Keywords Automated vehicles, Multi-modal trajectory forecasting, Trustworthy Artificial Intelligence, Trajectory forecasting

*Correspondence:

Charlotte Dupuis
charlotte.dupuis@univ-lille.fr

¹ Department of Bibliometric Analysis and XAI in Healthcare, Faculty of Engineering, University of Lille, Lille, France

² Department of Interpretable AI and Medical Informatics, Faculty of Engineering, University of Montpellier, Montpellier, France

³ Department of Healthcare AI and Bibliometric Review, Faculty of Engineering, University of Nantes, Nantes, France

Introduction

The integration of Artificial Intelligence (AI) has initiated a groundbreaking shift in medicine, fundamentally reshaping the methodologies used to navigate diagnostic and therapeutic complexities. This collaborative intersection is transforming the paradigm of customized patient care, introducing novel opportunities for disease prevention, targeted therapy, and the strategic allocation of healthcare assets. Within this transitioning landscape, AI architectures empower researchers and

practitioners to evaluate intricate biological data, uncover hidden patterns, and guide clinical strategies. The capacity to manage vast quantities of data rapidly and systematically enables accelerated diagnostic timelines, tailored therapeutic interventions, and highly optimized disease management [1].

Regrettably, deciphering the exact mechanisms behind the conclusions reached by these models continues to be a major obstacle for doctors, healthcare professionals, patients, and administrators. A continuous debate persists regarding the transparency of these automated processes, the legibility of the underlying algorithms, and the ethical dilemmas associated with incorporating automated frameworks into clinical workflows [2].

Even though contemporary empirical evidence demonstrates that AI systems can surpass human experts in specialized analytical tasks, their lack of transparency and interpretability restricts their widespread adoption in clinical settings. This opacity introduces legal and moral uncertainties that can impede technological momentum and prevent these innovative systems from achieving their full capacity to advance healthcare delivery.

Within the theoretical framework of AI, models and algorithms are typically classified into two primary categories: interpretable models (referred to as white-box) and opaque models (referred to as black-box) [3].

This classification depends on how transparently the relationship between input variables, intermediate calculations, and final system outputs can be observed. White-box architectures possess discernible and comprehensible traits that allow observers to trace how individual variables affect a given prediction; standard linear regression and decision trees serve as typical examples of this category. Conversely, black-box systems are characterized by highly intricate architectures where the internal processing paths, parameter weights, and foundational predictions remain hidden from the user, as seen in deep learning frameworks and random forest architectures. Furthermore, black-box frameworks risk incorporating detrimental biases [4, 5], which can undermine the confidence and reliance of medical practitioners. Specifically, if a model is calibrated on datasets that mirror existing social or cultural prejudices, it may generate discriminatory outcomes that widen rather than close societal divides. This introduces critical ethical dilemmas regarding equity and fairness for the populations affected by these technologies.

Ultimately, while white-box architectures provide more accessible pathways and outcomes, black-box designs consistently deliver superior predictive accuracy and performance metrics. This disparity is primarily driven by the fundamental capacity of black-box models to autonomously learn highly sophisticated, non-linear data representations. Their intricate and adaptable configurations allow them to isolate subtle nuances and covert interdependencies within datasets, thereby maximizing predictive precision, whereas white-box frameworks generally depend on more straightforward, easily traceable interactions [3].

Attention must also be directed toward the rising ethical anxieties caused by algorithmic opacity. Regulatory structures, most notably the European Union's General Data Protection Regulation (GDPR) [6], mandate strict legal frameworks governing the collection, archiving, transmission, processing, and examination of health-related information [7]. The GDPR explicitly champions the "right to explanation," empowering users to comprehend the logic underlying automated decisions that impact them. It likewise stresses the principle of accountability, legally requiring institutions to verify their adherence to these guidelines. Consequently, this demands heightened transparency in data handling workflows and underscores the need to clarify algorithmic mechanics when handling private individual data.

In conclusion, this systematic investigation is designed to analyze the interpretability and explainability of Machine Learning (ML) and Deep Learning (DL) methodologies within medical contexts. We analyze the explicit constraints identified in current literature and evaluate the broader ramifications of algorithmic non-transparency. Moreover, we examine concrete clinical case studies illustrating how explainability and interpretability can profoundly influence day-to-day medical operations, while exploring the primary engineering methodologies used to produce easily digestible explanations.

Explainability and Interpretability

The properties of interpretability and explainability in AI algorithms constitute core foundational pillars within the research, development, and engineering of sophisticated AI architectures [8]. Explainability is frequently conflated with interpretability, despite the fact that interpretability serves as a necessary baseline requirement for explainability [9]. Conversely, some frameworks define interpretability as a subordinate element of explainability [10]. This ambiguity frequently arises when these two expressions are utilized interchangeably within AI discourse without drawing a clear boundary between their conceptual definitions. Nonetheless, establishing this distinction remains vital, as it directly impacts how AI frameworks are evaluated and deployed.

The explainability profile of an architecture encompasses its capacity to offer explicit, comprehensive, and intelligible rationales for individual outputs. Characterizing an AI system's explainability typically requires the ability to identify and map internal system parameters, decode specific architectural configurations, computational nodes, and processing units that route information, and comprehend the exact function of each constituent part within the broader framework.

In contrast, interpretability denotes the capacity to understand the overarching behavior of an algorithmic model without the necessity of dissecting every discrete output. Thus, interpretability can be understood as a macro-component of explainability, offering a high-level, holistic perspective of the model's operations rather than parsing fine-grained technical details. Interpretability focuses on the degree to which an observer can trace the internal cause-and-effect mechanisms of a system. It demands not only an intrinsic understanding but also the ability to articulate how a system operates or reacts. Broadly speaking, artificial intelligence architectures exhibit varying thresholds of transparency, a phenomenon closely linked to the concept of degrees of interpretability. To summarize these concepts:

- **Interpretability** relates to the extent to which causal relationships can be discerned within an internal architecture and describes a specific methodology through which a system is understood or characterized. Consequently, it allows observers to diagnose the root cause of systemic errors and project how modifications to input variables or computational parameters will alter subsequent outcomes [10].
- **Explainability** concerns the capacity to supply direct, comprehensible justifications or structural reasoning for specific algorithmic outputs; it entails the ability to isolate internal parameters and decipher the exact meaning and functional role of a given node within the network [10].

Contemporary methodologies designed to facilitate the explainability and interpretability of AI systems are conventionally bifurcated into *a priori* and *a posteriori* paradigms. *A priori* explainability and interpretability techniques involve principles that are established or assumed prior to empirical experience or the observation of specific data points. Within this framework, *a priori* methods represent embedded strategies or architectural considerations integrated directly during the initial engineering and design phases of an AI model. These can be categorized as follows:

- Structural simplicity can optimize interpretability through the utilization of models featuring less complex topologies, including linear models or shallow neural network frameworks [11];
- Feature engineering entails the selection or development of input variables that correspond directly to clear, well-established concepts within a specific domain. Concurrently, regularization methods may be applied to penalize excessive parameter weights, thereby promoting sparsity and augmenting model legibility [12];
- Regularization algorithms, such as L_1 and L_2 mathematical constraints, can be systematically deployed to induce model sparsity and improve general interpretability [13];
- Restricting the maximum depth of a model can halt the growth of hyper-complex architectures, thereby supporting the overall interpretability of the framework.

Conversely, *a posteriori* explainability and interpretability methodologies involve insights derived or inferred through empirical observation and the assessment of specific datasets. In this context, *a posteriori* techniques are implemented

post-training to decode the specific inferences generated by a finalized model. Examples of these post-hoc techniques include:

- LIME (Local Interpretable Model-agnostic Explanations), which generates localized justifications for individual input instances and provides an after-the-fact comprehension of distinct algorithmic decisions [14];
- SHAP (SHapley Additive exPlanations), which leverages foundational principles from Shapley value theory to allocate feature contribution metrics to each input variable, thereby clarifying historical predictions generated by a model [15];
- Feature visualization techniques, which utilize discrete data examples to construct retrospective graphical mappings of the most mathematically influential attributes;
- Sensitivity analysis, which evaluates how minor variations introduced into input vectors would have altered the historical predictions of the model [16].

As a general rule, *a posteriori* methodologies are omitted during initial system configuration and instead focus on mining explanatory insights from pre-existing architectures [17] (which typically rely on a black-box design paradigm).

Constructing an inherently explainable and interpretable AI system is normally an intricate and variable task, given that its overall complexity depends directly on the core objective of the model along with its corresponding variables and data attributes. There are numerous ethical, legal, and operational considerations that make interpretability and explainability highly advantageous or mandatory within AI applications [18]. The primary barriers to their execution include:

- **Opacity of complex models:** Deep learning methodologies, such as deep neural networks, routinely construct highly intricate models consisting of millions of distinct parameters. Due to this extreme architectural density, even the system developers themselves may struggle to decipher the exact mathematical path a model uses to reach its conclusions.
- **Trade-off between performance and interpretability:** Developers are frequently compelled to deploy more complex architectures to maximize performance metrics regarding generalization and accuracy. However, highly intricate models are inherently more difficult to comprehend. Achieving the optimal balance between predictive accuracy and model legibility remains a fundamental challenge.
- **Bias in training data:** If a model is trained on skewed or non-representative data, it will likely perpetuate these underlying disparities, yielding discriminatory or flawed inferences. Consequently, it is vital to audit how a model processes information to detect and isolate unintentional biases.
- **Interpretability of features:** An essential component of a transparent architecture involves identifying which specific variables drive algorithmic conclusions. In many scenarios, particularly when utilizing complex deep neural networks, pinpointing the precise components that influence model outputs can be exceptionally difficult.
- **Scalability:** Achieving structural interpretability becomes significantly more challenging in massive deep learning models, as they integrate vastly more parameters and demand extensive computational infrastructure for post-hoc interpretive processing.
- **Changes in model behaviour:** The operational behavior of a machine learning model can shift over time due to modifications in incoming data streams or continuous training updates. Ensuring that a model remains interpretable amidst these dynamic behavioral shifts is an ongoing obstacle.
- **Social acceptance:** Even when a model is theoretically transparent, encouraging user adoption can be difficult, particularly if the generated algorithmic rationales contradict human intuition.

To surmount these barriers, researchers and engineers are pioneering a variety of tools and methodologies aimed at refining the clarity and legibility of AI frameworks. These initiatives focus on the utilization of advanced visualization routines, localized feature interpretations, automated textual explanations, and the structural reduction of model complexity.

Navigating the friction between peak model performance and internal interpretability represents a continuous challenge; yet, resolving this tension is critical to guaranteeing that artificial intelligence is deployed ethically, responsibly, and for the collective benefit of society [19]. Continuing breakthroughs in this domain remain vital to determining the trajectory of future artificial intelligence architectures.

Related Work

The existing literature contains several review papers that examine and analyze the trends and obstacles associated with embedding the principles of interpretability and explainability into artificial intelligence systems. These prior reviews are omitted from our primary dataset analysis; consequently, this section delivers a synthesis of the most impactful contributions from those authors.

Biran *et al.* [20] examine diverse methodologies aimed at rendering machine learning frameworks comprehensible. The investigators explore techniques for bolstering user confidence and accelerating practical integration by transforming intricate models into accessible and logical formats. The authors sort the surveyed interpretation methodologies into distinct groups, which include rule-based schemes, global and localized operations, visualization methods, and techniques structured around model representation. Furthermore, they emphasize the functional relevance of these interpretive strategies by detailing their deployment within specialized fields such as marketing, finance, and healthcare.

Guidotti *et al.* [21] offer an extensive evaluation of the frameworks found in existing literature for demystifying decision-making architectures that rely on obscure and dark machine learning models. They distinguish multiple challenges related to explainability and establish a rigorous definition for each. These classified "black box" problems encompass the model explainability dilemma, the outcome explainability dilemma, the model inspection dilemma, and the transparent box design challenge. Their literature evaluation concludes that while numerous methodologies have been engineered to decipher "black boxes," several vital scientific questions remain unresolved.

By incorporating theoretical perspectives from the social sciences, Miller *et al.* [22] analyze the functional role of explanations within artificial intelligence. The investigators emphasize the necessity of rendering algorithmic outcomes more palatable and intelligible to users while exploring the core value of explanations in intelligent frameworks. They address the diverse hurdles encountered when delivering persuasive arguments, accounting for variations in how justifications are processed from cultural, sociological, and psychological viewpoints. The authors underscore the critical importance of engaging end-users during the structuring of explanations to ensure they are both actionable and positively received. To navigate the intricate nature of algorithmic explanations and improve the comprehension and public acceptance of intelligent technologies, the paper advocates for an interdisciplinary approach.

Arrieta *et al.* [19] review the core ideas, classifications, future outlooks, and challenges within Explainable Artificial Intelligence (XAI) to confront the moral and societal implications of algorithmic non-transparency. The authors stress the necessity of engineering more transparent models to advocate for the ethical deployment of AI, alongside the importance of decoding and interpreting algorithmic choices to guarantee responsible governance. Based on variables such as explanation type, granularity level, and deployment method, the researchers introduce several taxonomies to classify XAI methods. They also examine implementation barriers in XAI, such as managing the friction between precision and interpretability, the relationship between privacy and model explainability, and the challenge of balancing explainability against architectural complexity.

Tjia *et al.* [23] evaluate the critical nature of interpretability for black-box machine inferences within clinical environments. They provide a synthesis of the interpretive strategies introduced across various studies, organizing them according to their level of clarity. Within healthcare, such justifications are mandatory to validate the dependability of automated decisions. However, this study outlines significant risks, including the potential manipulation of explanations and variations in training data quality. It likewise highlights the necessity of specialized professional training to correctly analyze algorithmic outputs in medical settings. Ultimately, this work advocates for a critical approach to algorithmic

interpretation, suggesting it be treated as a auxiliary support mechanism for clinical choices until more validated interpretability methods are developed.

Stiglic *et al.* [24] highlight the vital role of interpretability within machine learning (ML) models deployed in healthcare fields. The authors categorize interpretability methods into two primary divisions: the first centers on customized (local) interpretation, which highlights granular justifications at the individual patient level, while the second concerns the synthesis of predictive models at a population (global) level, which is useful for identifying overarching systemic trends. Additionally, they distinguish interpretability strategies into model-specific routines and model-neutral approaches. The former parses the inferences generated by a unique ML architecture, such as a neural network, whereas model-agnostic strategies deliver transparent rationales for predictions from any ML framework, irrespective of its underlying configuration.

Amann *et al.* [7] address the challenge of interpretability when utilizing artificial intelligence (AI) in the healthcare sector, noting that while AI-driven systems have displayed superior performance over human practitioners in specific analytical workflows, their lack of transparency has drawn scrutiny. The researchers use AI-based clinical decision support systems as a foundational benchmark for their multidisciplinary evaluation of interpretability's relevance to medical AI, integrating viewpoints from technology, law, medicine, and patient advocacy. Utilizing the insights from this conceptual analysis, they perform an ethical assessment anchored in Beauchamp and Childress's "Principles of Biomedical Ethics" (autonomy, beneficence, nonmaleficence, and justice) to determine the necessity of interpretability in healthcare AI. Each professional domain highlights a unique set of variables and ideals essential for understanding the function of interpretability in clinical practice. Both medical and patient perspectives emphasize the need to carefully consider the interplay between human actors and medical AI, warning that a lack of interpretability in decision support frameworks threatens fundamental biomedical ethics and could negatively affect individual and public health.

In their publication [25], Mehrabi *et al.* execute an in-depth investigation regarding bias and fairness within machine learning architectures. The authors analyze the root causes of bias, its manifestation in algorithmic models, and potential remedies for these issues. They isolate the primary drivers of bias in machine learning, identifying variables embedded within the models, systemic flaws in training datasets, and the intrinsic learning process itself. The investigators evaluate diverse metrics for auditing model fairness, including group fairness, individual fairness, and fairness of opportunity, while analyzing multiple mitigation strategies such as collecting balanced datasets, tuning internal model weights, and embedding fairness metrics into performance evaluations. They further examine the systemic complications introduced by these measurement tools.

In this study, Chakrobarty *et al.* [26] concentrate on contemporary breakthroughs in medical XAI to deliver a thorough synthesis of the explainability strategies documented in current scholarship. Through a comprehensive literature review, the article examines the XAI methodologies and techniques used in medical machine learning systems. The authors introduce a conceptual framework for categorizing these XAI approaches, which helps organize and analyze the available literature. Balancing interpretability against accuracy emerges as a prominent theme in the reviewed literature, with several referenced studies prioritizing interpretability alongside predictive precision.

In [2], the researchers utilize a systematic mapping protocol to evaluate existing scholarship regarding interpretability techniques applied in the medical sector. They account for variables including publication venues and timelines, contribution types, the specific medical and ML sub-disciplines involved, the machine learning objectives, the chosen methodologies for interpreting "black box" models, the validation frameworks for interpretability, the performance of these techniques, the most effective strategies, and the datasets used for evaluation. After filtering 179 publications (spanning 1994–2020) across six digital libraries (ScienceDirect, IEEE Xplore, ACM Digital Library, SpringerLink, Wiley, and Google Scholar), their findings illustrate an upward trajectory in interpretability research over time, with a clear dominance of empirical studies and proposed solution frameworks. The data reveal that classification tasks, oncology, and diagnostic applications represent the most common medical actions, clinical fields, and machine learning goals. Artificial neural

networks constitute the most frequent "black box" architectures targeted for interpretability, while metrics such as accuracy, integrity, and the total number of rules are routinely used to gauge model interpretability.

In [27], the issue of model transparency is examined. The investigators point out that despite the remarkable success of artificial intelligence across various medical tasks, its lack of openness has created a pressing need for more "interpretable" or understandable systems. This paper contends that medical practitioners often prefer models they can understand, even if it means sacrificing some degree of predictive accuracy, thereby underscoring the vital role of interpretability. This preference stems from the fact that clinicians must actively utilize these tools to achieve better patient outcomes. Consequently, the authors suggest that prioritizing raw performance over comprehensibility could represent a "lethal bias" that ultimately reduces the clinical utility of AI for patients.

Combi et al. analyze how explainable artificial intelligence (XAI) is integrated into biomedical domains in [17]. They outline five primary areas that require deeper investigation. First, they highlight the necessity of connecting symbolic machine learning with sub-symbolic approaches. Second, building explainability into intelligent systems remains a major obstacle, requiring a detailed evaluation of how different systems behave, function, and interface with their users. The third core element involves measuring and optimizing the performance of explainable models and tools. The authors stress that future studies must track how these systems alter the beliefs, perspectives, and actions of users, alongside monitoring overall decision-making accuracy. Fourth, determining whether explainability is even required for a given task is raised as a critical question. Lastly, they emphasize the importance of designing explainability tools around the end-user, noting that user-centered design is a foundational requirement for XAI.

Farah et al. [28] analyze the core principles behind engineering AI-driven medical software, focusing on the interplay between algorithmic power, interpretability, and explainability. Their review of existing literature indicates that healthcare technology evaluation bodies view these three elements as crucial pillars for building clinical trust, making them mandatory components of the appraisal process. They offer specific guidance on how and when to benchmark performance based on available datasets and the architecture of the model. Furthermore, because interpretability and explainability are notoriously difficult to define mathematically, specific framework tools have been created to aid their assessment. The authors also supply a projected regulatory roadmap flowchart to guide the design and verification of AI-based clinical devices.

In [29], Ali et al. map out the latest milestones and emerging patterns within the domain of AI interpretability and explainability. By applying a structured, hierarchical taxonomy, they split XAI methodologies into four distinct branches: (i) data explainability, (ii) model explainability, (iii) post-hoc explainability, and (iv) evaluation of explanations. Their work provides an overview of current performance metrics, available open-source software, and benchmarking datasets, while charting future research directions. They also address how explainability intersects with statutory regulations, user expectations, and specific target industries—factors they classify as core XAI considerations. To evaluate these methodologies, the authors conducted a meta-analysis of 410 foundational papers published from January 2016 through October 2022. The resulting end-to-end deployment framework couples system design goals directly with evaluation techniques and these broader XAI considerations.

Finally, in [30], Band et al. explore how explainable artificial intelligence (XAI) is utilized throughout the medical sector. The overarching goal of XAI is to ensure that the choices generated by machine learning (ML) and artificial intelligence (AI) algorithms remain fair, open, precise, generalizable, and easy to comprehend. This paper delivers a rigorous assessment of past literature tracking ML and AI interpretability within clinical settings. Additionally, it explores the broader transformations that AI and ML technologies are poised to bring to healthcare delivery systems.

Materials and Methods

This section outlines the specific research design implemented for this study. The framework is divided into five sequential phases, which can be summarized as follows: (i) Formulation of the core research questions, (ii) Initial data assessment,

(iii) Establishment of inclusion and exclusion criteria, (iv) Selection of target studies matching those criteria, and (v) Data extraction and subsequent analysis. Every retrieved document was analyzed across four dimensions: the core problem targeted, how the problem was formalized, the specific methodology applied, and the hurdles encountered.

The initial phase focuses on establishing the core research questions. The study was designed to answer the following specific inquiries:

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?

Literature collection was conducted by searching the indexing databases of Scopus and Web of Science.

Scopus serves as an extensive multidisciplinary database that indexes abstracts and citations across numerous scientific fields. It tracks author profiles, institutional connections, and citation metrics. Scholars rely on Scopus to conduct cross-disciplinary research, perform citation mappings, and gauge the academic reach of published papers. The platform features customizable search alerts, advanced querying tools, and journal impact analytics. Typically accessible via institutional subscriptions, Scopus is a standard tool for compiling literature reviews and monitoring current scientific breakthroughs [31].

Web of Science is an academic literature database providing access to a diverse array of global scientific data. It aggregates literature from peer-reviewed journals, symposium proceedings, patents, and related academic records, and is highly regarded for its wide-ranging multidisciplinary index. Its advanced querying and analytics tools allow researchers to map out citation networks and connections between papers. It is routinely used to evaluate the scholarly impact of specific publications, detect emerging industry shifts, and identify prospective research partners. Access to Web of Science is generally limited to university libraries and academic research institutions [32].

To isolate our investigation to the specific domains of machine learning and deep learning algorithm interpretability and explainability, we formulated the following search syntax: "((explainable OR interpretable OR interpretability OR explainability) AND ((machine AND learning) OR (deep AND learning) OR (artificial AND intelligence)))". This initial query retrieved 26,951 records from the Scopus index and 21,633 records from the Web of Science database. To refine this raw output into a targeted sample, we implemented a structured set of inclusion and exclusion benchmarks.

Inclusion criteria:

- Disseminated exclusively in the English language;
- Published within the decade spanning 2013 to 2023;
- Categorized as a peer-reviewed journal article, conference proceeding paper, or book chapter;
- Centered primarily on the explicability and interpretability of machine and deep learning architectures;

- Retained only the most recent iteration in cases of duplicate publications.

Exclusion criteria:

- Retained only the most recent iteration in cases of duplicate publications.

Enforcing these screening protocols narrowed the volume to 23,805 records within Scopus and 19,709 records within Web of Science. Next, we mapped the institutional output across diverse disciplinary fields, as illustrated in **Figures 1 and 2** for the Scopus and WOS indexes, respectively.

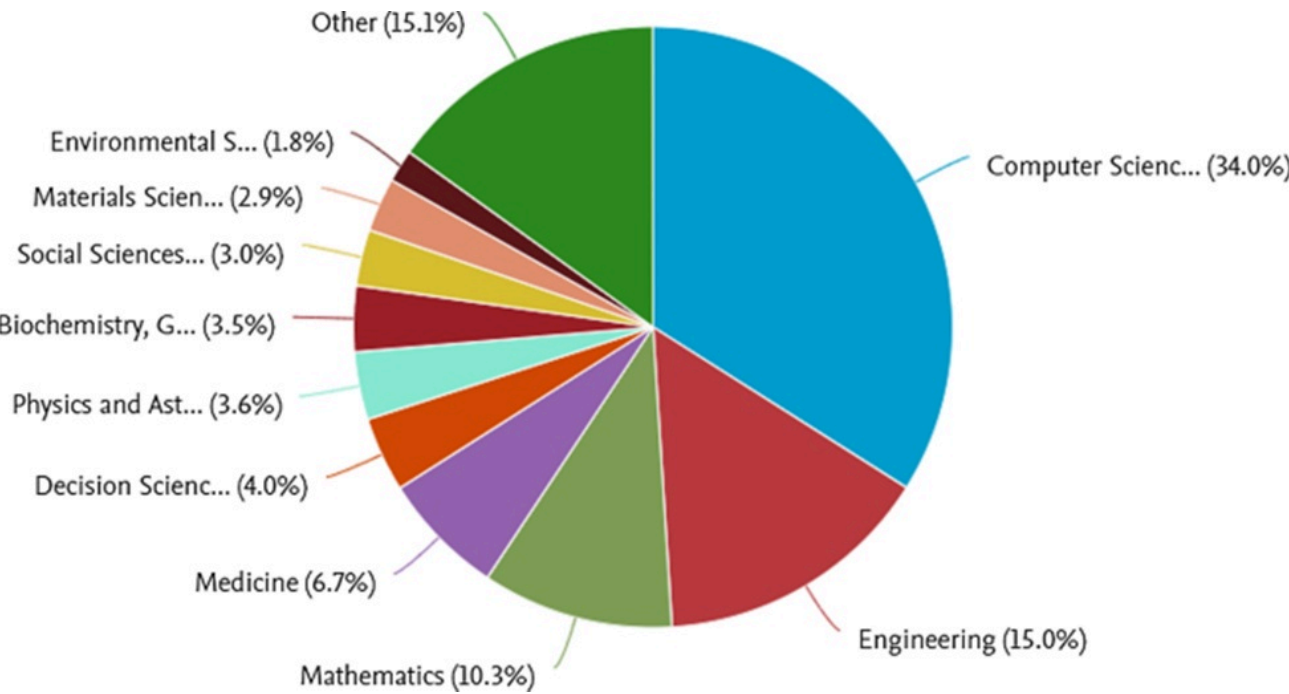


Figure 1. Documents by subject area for the Scopus database

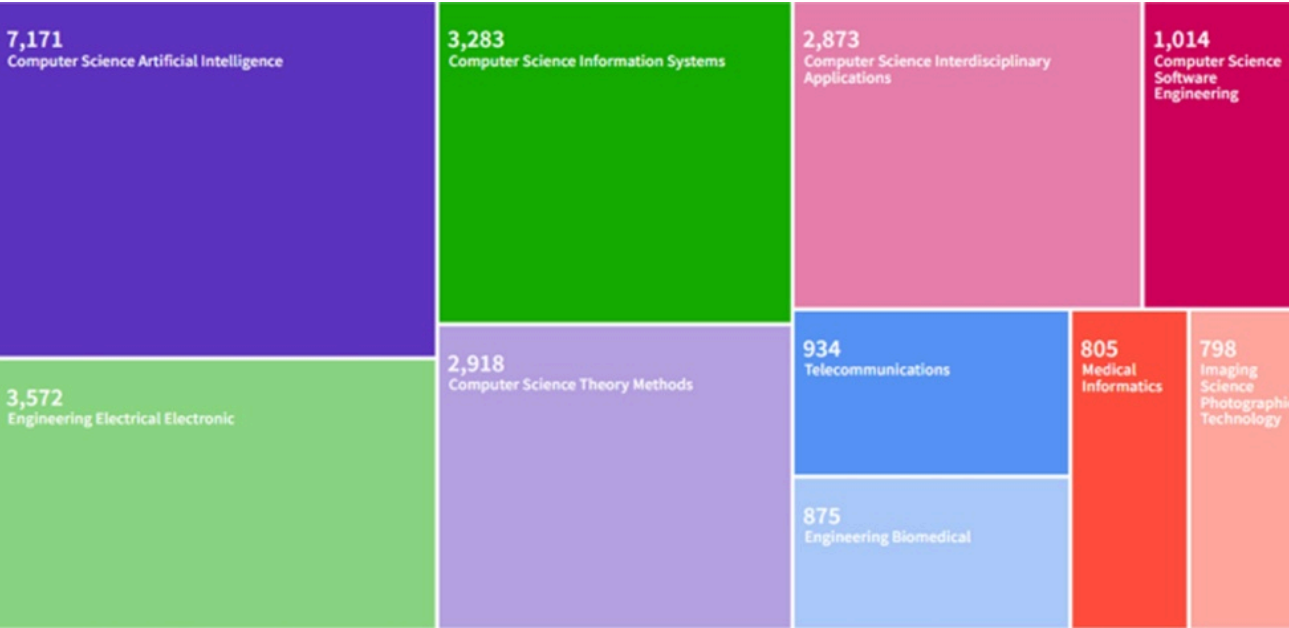


Figure 2. Documents by subject area for the Web of Science dataset

We subsequently narrowed our scope to the clinical sector, filtering out 3,178 documents from Scopus and 805 from WOS. By cross-referencing these two cohorts and isolating only the overlapping literature cataloged simultaneously in both systems, we arrived at a final corpus of 448 papers.

We then synthesized the indexed keywords of these remaining papers with their respective citation counts and publication timelines. Specifically, we computed the frequency metrics for every keyword to map out the dominant technical methodologies and application landscapes within the literature. To ensure data consistency and eliminate spelling variations, a keyword standardization pipeline was executed. These standardized values were evaluated alongside annual citation frequencies to highlight the most impactful contemporary studies. Where index keywords were omitted by the database, we extracted author-supplied terms; for papers lacking both, keywords were derived from the publication titles. To pinpoint the definitive subset for final evaluation, we applied the following parameters:

- Keyword occurrence density (i.e., identifying the most prominent terms within the dataset);
- Chronological publication year;
- Citation volume normalized against the year of release.

Based on these benchmarks, we isolated and scrutinized 10 core papers that directly address our pre-established research questions. The upcoming sections trace the trajectory of this literature from the initial search queries executed in Scopus and Web of Science (WOS) through the rigorous application of our inclusion and exclusion filters.

Results and Discussion

An analysis of this foundational paper cohort, curated via the search parameters was utilized to systematically resolve the first four research inquiries: RQ1, RQ2, RQ3, and RQ4.

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

This question serves to quantify the shifting focus of the global scientific community toward clinical AI explainability and interpretability frameworks over the last ten years. As demonstrated in **Figures 3 and 4**, the annual publication volume remained marginal until 2018, failing to exceed eighty entries per annum. However, research momentum shifted drastically in 2019, accelerating to a peak of 974 papers in Scopus and 164 papers in WOS by 2023, which underscores the intensifying focus on this domain. It is vital to state that the metrics for the year 2023 reflect the data compiled up to October 2023.

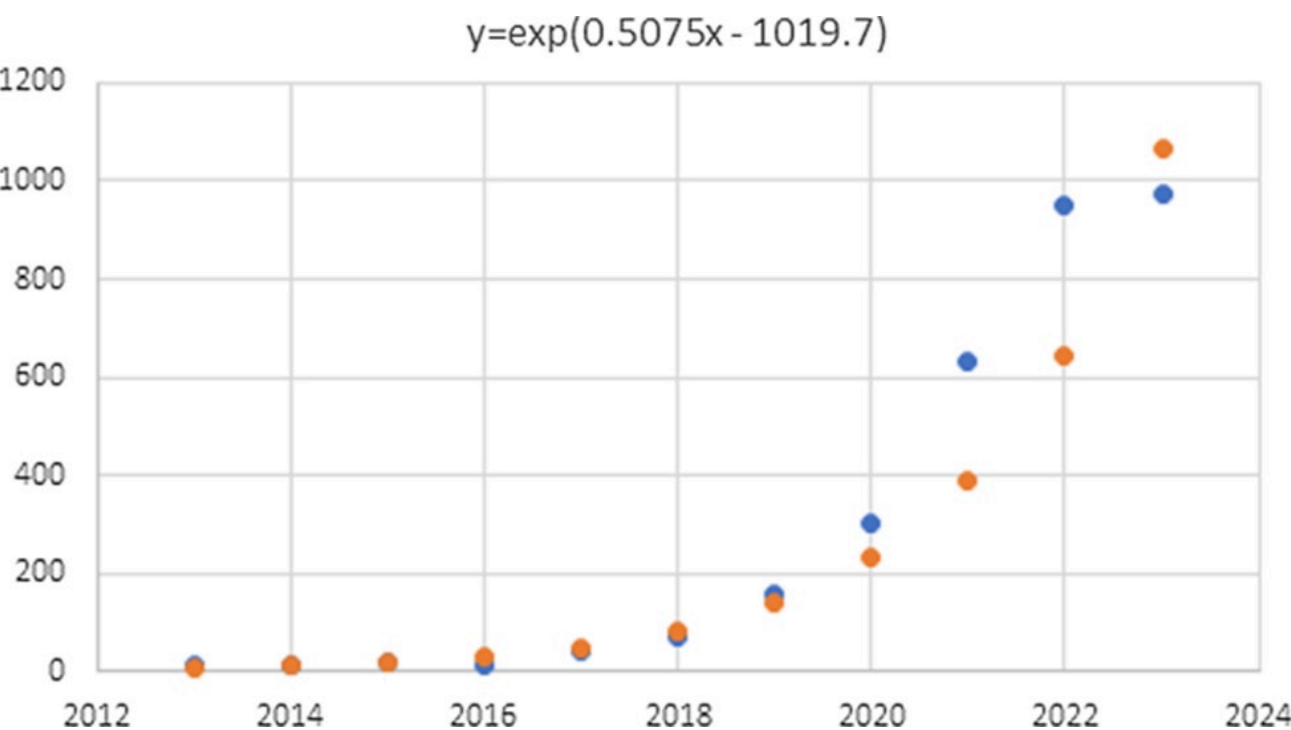


Figure 3. Academic studies published from 2013 to 2023 in Scopus database (in blue the original data, in red the estimated values). From this analysis, we can see that the annual growth rate in Scopus is 0.5075

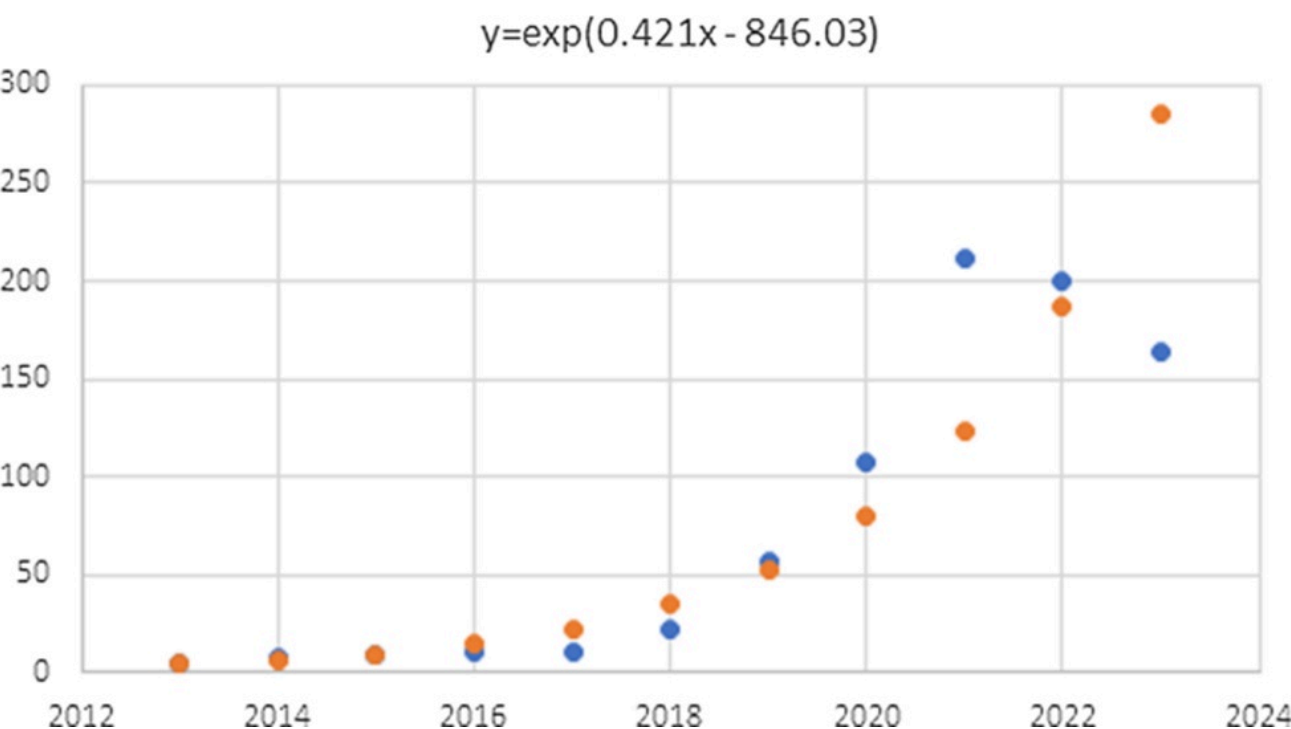


Figure 4. Academic studies published from 2013 to 2023 in the WOS database (in blue the original data, in red the estimated values). From this analysis we can see that the annual growth rate in WoS is 0.421

By cross-referencing the data from both systems, we isolated a core group of 448 mutually indexed studies, which form the exclusive basis for all subsequent thematic profiles.

RQ2: What are the most relevant publication channels?

Through this research question, we map out the primary dissemination avenues and scholarly journals utilized to publish research regarding AI explainability and interpretability methodologies. **Table 1** catalogs the results of this distribution analysis.

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

Publication Source	Total Publications
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 of the 2021 IEEE 9th International Conference on Healthcare Informatics (ICHI 2021)	8
JAMIA Open	7
Digital Health	7
Statistics in Medicine	6
Proceedings of the IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE 2020)	6
The Lancet Digital Health	5
Proceedings of the 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 – SPIE Proceedings	4
Journal of Healthcare Informatics Research	4
JMIR Medical Informatics	4

Proceedings of the 21st IEEE International Conference on Bioinformatics and BioEngineering (BIBE 2021)	4
Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (BCB 2020)	3
Proceedings of the 2023 IEEE International Conference on Digital Health (ICDH 2023)	3
Proceedings of the 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
Proceedings of the 2020 IEEE International Conference on Healthcare Informatics (ICHI 2020)	2
Statistical Methods in Medical Research	1
Proceedings of the 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
Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics (ACM-BCB 2018)	1
Proceedings of the 2019 IEEE International Conference on Healthcare Informatics (ICHI 2019)	1

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

This inquiry highlights the geographical distribution of the research hubs advancing the study of medical AI explainability and interpretability. In **Figure 5**, our visualization isolates nations associated with at least 5 publications. The analysis reveals that the largest share of literature emerged from research institutions within the United States (114 papers), followed closely by China (80 papers). The remaining active centers include Italy (27 papers), Spain (16 papers), the United Kingdom (16 papers), South Korea (13 papers), Canada (12 papers), France (10 papers), Taiwan (10 papers), the Netherlands (8 papers), Austria (7 papers), Germany (7 papers), Singapore (7 papers), India (6 papers), and Japan (5 papers). Notably, the dataset includes collaborative papers co-authored across multiple international boundaries. To trace these multinational co-authorship networks, **Figure 6** maps countries sharing a minimum threshold of 5 co-occurrences within the corpus. The strongest collaborative link occurs between the United States and China, which share 9 connections.

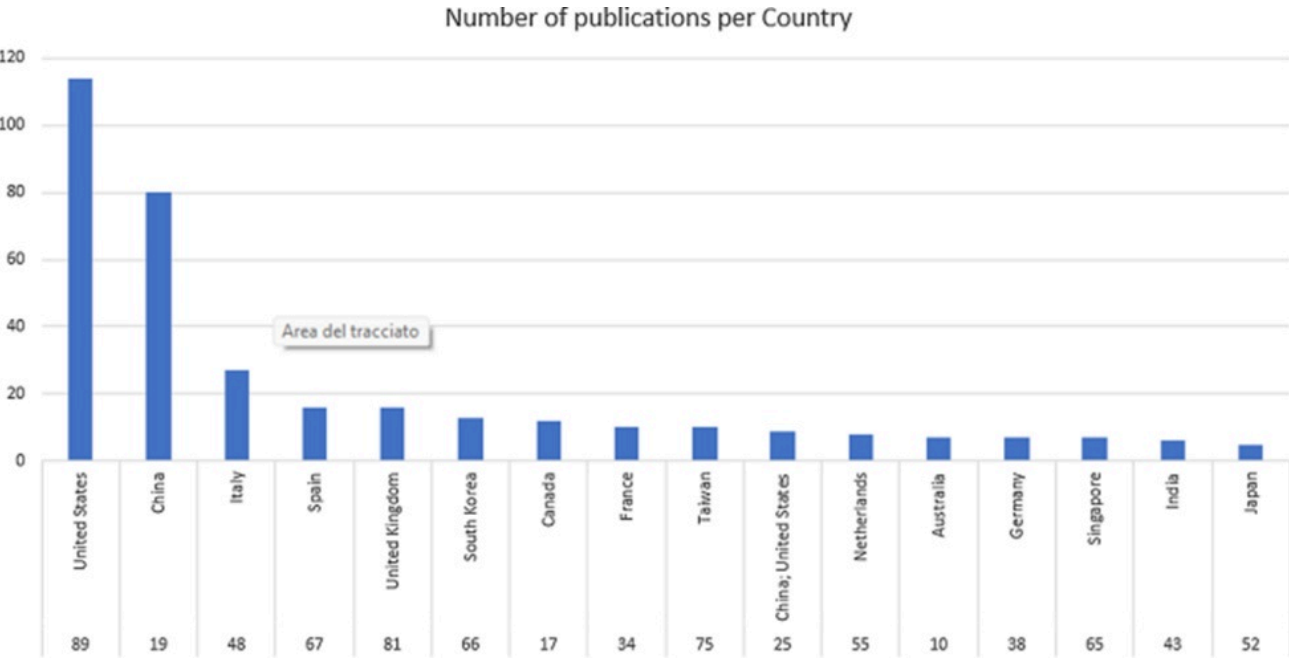


Figure 5. Number of publications per Country on explainability and interpretability of AI in the medical field

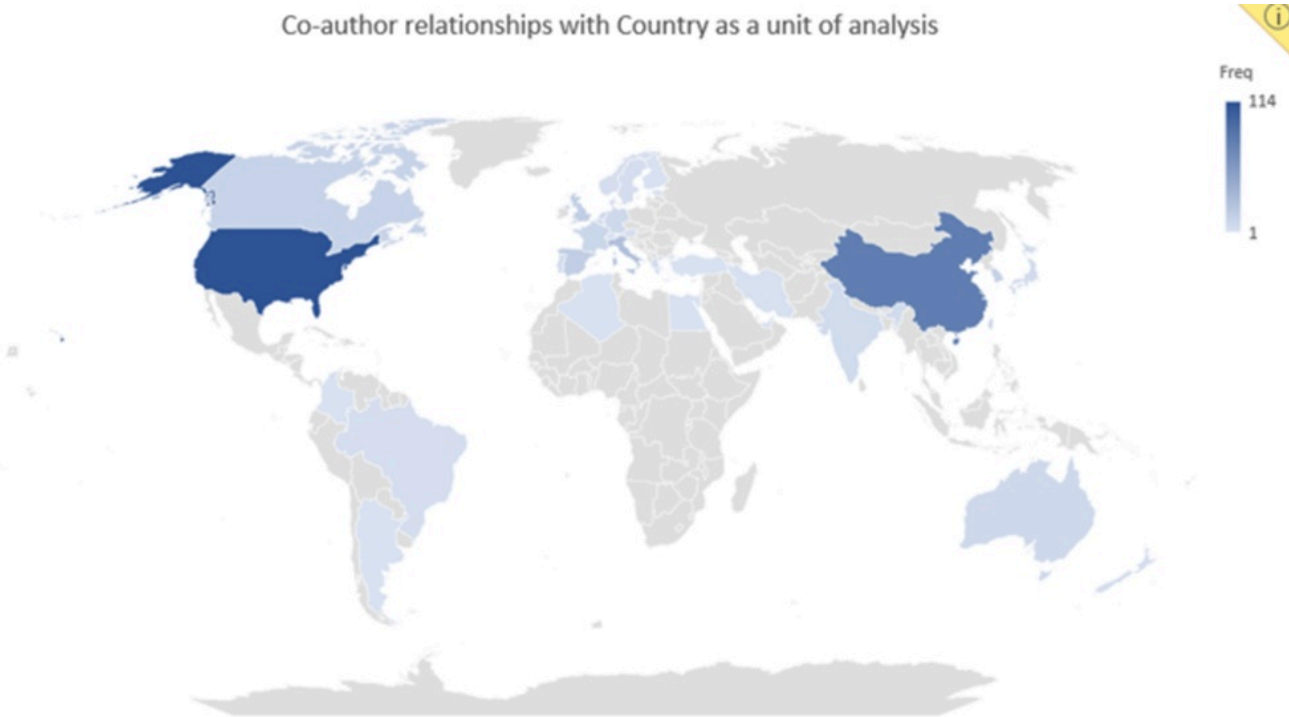


Figure 6. Co-author relationships with country as a unit of analysis

RQ4: What application areas and methods were used?

The goal of this inquiry is to catalog the specific clinical sub-domains and algorithmic methods deployed to achieve interpretable and explainable medical AI. To accomplish this, we deconstructed the index keywords from the 448 included

papers. **Figure 7** models the targeted clinical domains. Keywords were aggregated into high-level macro categories, with the vast majority of papers falling squarely into one of these defined groups.

Regarding the technical approaches, we implemented the same thematic aggregation technique used for the clinical fields, clustering terms that describe identical underlying methodologies, as visualized in **Figure 8**.

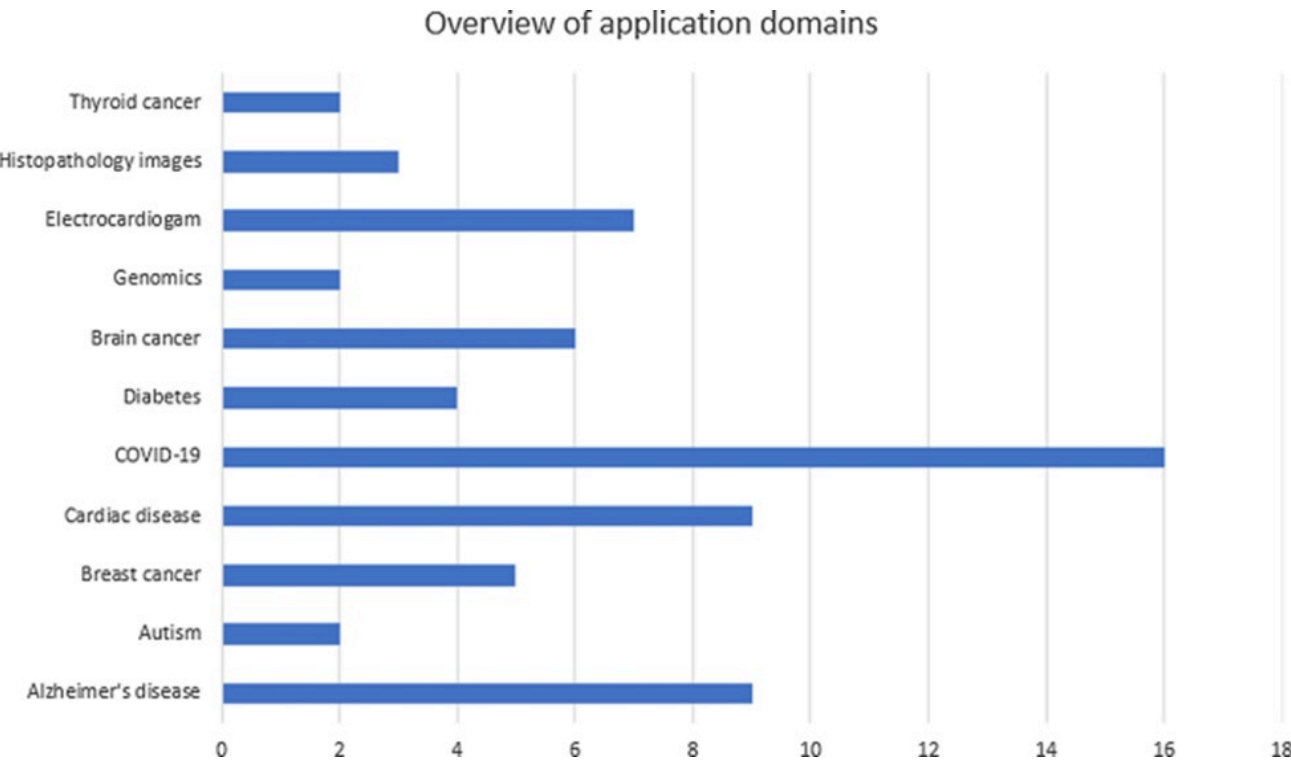


Figure 7. Overview of the application domains

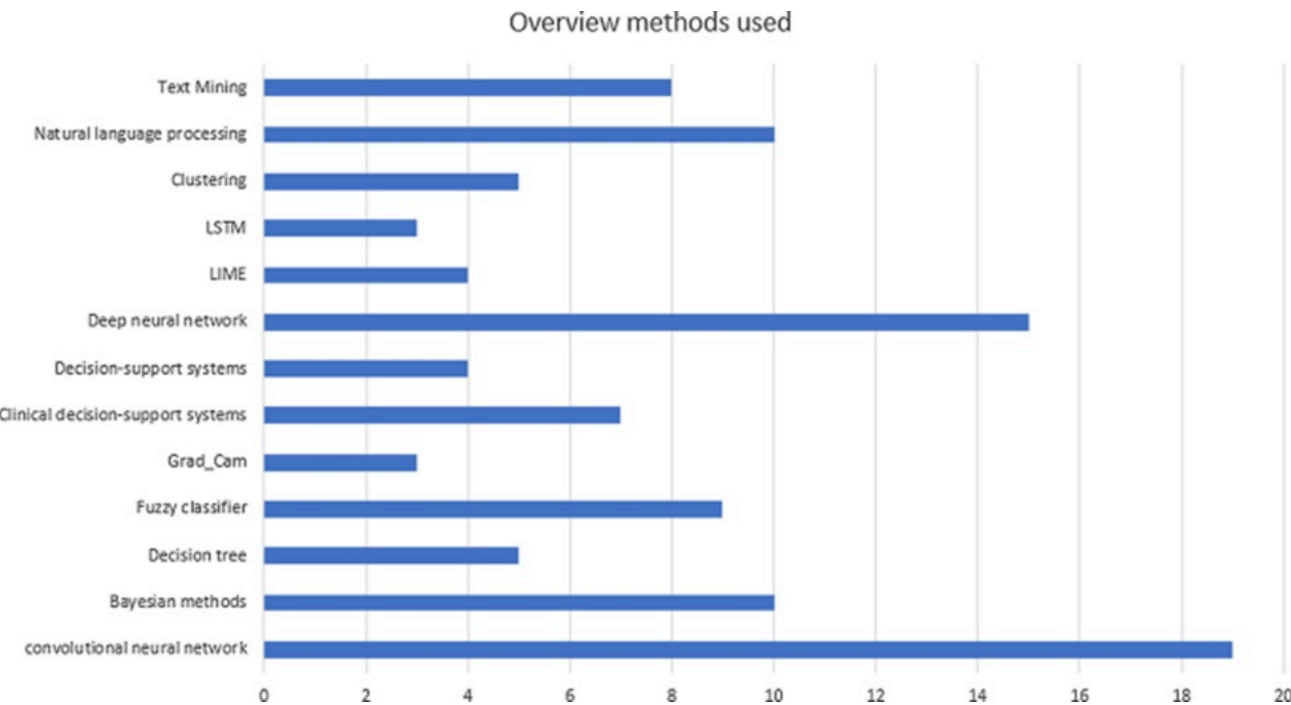


Figure 8. Overview of the methods used

Additionally, we mapped out a bibliometric co-occurrence network of the indexed terms using VOSviewer [33], which is displayed in **Figure 9**. A co-occurrence relationship signifies that two distinct keywords are featured within the same publication. Following an initial data-refinement stage, VOSviewer mapped 10 separate thematic clusters by filtering for terms with at least 3 occurrences. Within this network topology, distinct colors denote separate clusters and their constituent elements. The scale of both the circular nodes and text labels is determined by the total appearance frequency of the corresponding keyword. The intersecting vector lines trace keyword co-occurrences within individual articles. Every isolated cluster aggregates keywords that define either a clinical application domain, a technical strategy, or both, aimed at overcoming the challenges of interpretability and explainability in healthcare AI algorithms.

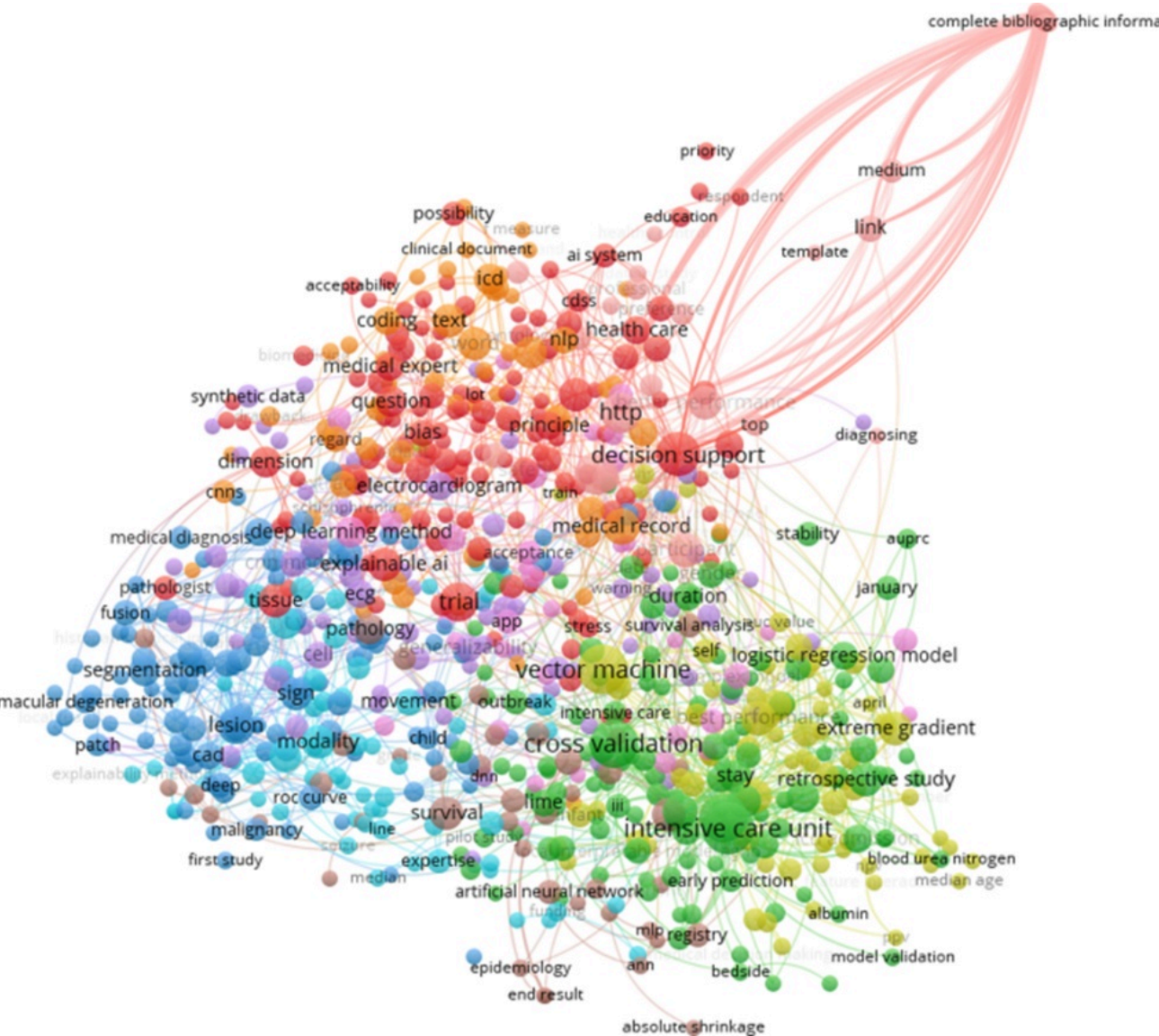


Figure 9. Bibliometrics analysis on the co-occurrence of index keywords

Analysis of the Main Papers

This section provides a detailed evaluation of the 10 core documents selected via the screening protocols. We begin with a macro-level investigation into how distinct clinical domains utilize artificial intelligence interpretability and explainability methods. Following this, we outline how the problem of algorithmic interpretability and explainability is formalized within the AI field, focusing on research types and technical methodologies (resolving research question RQ5). We then examine the specific performance metrics deployed to evaluate these systems (resolving research question RQ6). Lastly, we synthesize the primary obstacles encountered in the literature (resolving research question RQ7). **Table 2** categorizes each investigation by publication year (second column), citation count (third column), the core algorithmic approach driving the technology (fourth column), and the evaluation dataset utilized (fifth column).

The dominant clinical application domains are mapped out in **Figure 7**. These notably encompass COVID-19, Alzheimer’s disease, cardiac pathologies, electrocardiogram analysis, as well as brain and breast malignancies.

As detailed in **Table 2**, the primary technical frameworks consist of neural network architectures [34–37] (including CNN, LSTM, and Boltzmann machine models) [38, 39], which are frequently supplemented by traditional machine learning algorithms (such as K-nearest neighbors, Logistic Regression, Naïve Bayes, Random Forest, and Support Vector Machines).

Table 2. Technical information about the selected works

Study	Publication Year	Citation Count	Applied Methodology	Dataset / Data Source
Explainable Deep Learning for Pulmonary Disease and Coronavirus COVID-19 Detection from X-rays [34]	2020	375	VGG-16-based CNN with Grad-CAM and activation map visualization	[40–42]
Explainability for Artificial Intelligence in Healthcare: A Multidisciplinary Perspective [7]	2020	308	Clinical Decision Support Systems (CDSS)	N/A
Deep Learning Interpretation of Echocardiograms [35]	2020	201	EchoNet convolutional neural network (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 Neighbors, Logistic Regression, Naïve Bayes, Random Forest, Support Vector Machines, and Decision Support Systems	IRCCS Ospedale San Raffaele
Explainable Artificial Intelligence for Breast Cancer: A Visual Case-Based Reasoning Approach [39]	2019	156	kNN, Distance-Weighted kNN, Rainbow Boxes-Inspired Algorithm, Case-Based Reasoning, and Multidimensional Scaling (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	Long Short-Term Memory (LSTM) network and Shapley Additive Explanations	Electronic Patient Records (EPRs), Central Person Registry (CPR), and Danish National Patient Registry (DNPR)
On the Interpretability of Machine Learning-Based Models for Predicting Hypertension [43]	2019	124	Feature Importance, Partial Dependence Plots, Individual Conditional Expectation, Feature Interaction Analysis, Global Surrogate Models, Local Surrogate Models, and Shapley Values	Henry Ford Health Systems
Learning Vector Representation of Medical Objects via EMR-Driven Nonnegative Restricted Boltzmann Machines (eNRBM) [44]	2015	109	Restricted Boltzmann Machine (RBM) and EMR-driven Nonnegative RBM (eNRBM)	[37, 45]
Electronic Medical Record Phenotyping Using the Anchor and Learn Framework [46]	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 [47]	2020	90	Area Under the Curve (AUC)-based Evaluation Method	Wisconsin and Ljubljana Breast Cancer Datasets

RQ5: What are the most interpretability and explainability algorithms?

This research question examines the technical dimensions that assist domain practitioners in understanding the operational environments mapped by the authors. The accompanying table summarizes the problem formulation parameters extracted from the chosen literature. Each publication is deconstructed to isolate the exact interpretability and explainability frameworks deployed. Additionally, we classify the experimental datasets based on whether they utilize real-world clinical data or synthetic data. Because certain documents omit these specific parameters, instances with unspecified dataset characteristics are designated as "N/A." The most prevalent methodologies include:

- **Decision support systems:** These frameworks, or DSS, translate the complex outputs of black-box algorithms into interpretable graphical displays. By utilizing charts, heat maps, and alternative data visualizations, they highlight the specific features that exert the greatest leverage on clinical decisions. A robustly engineered DSS also features an intuitive user interface, ensuring accessibility for medical professionals who lack extensive expertise in machine learning [7, 38].
- **Gradient-weighted Class Activation Mapping:** Commonly known as Grad-Cam, this approach highlights the spatial regions within an input image that most heavily directed a CNN's classification choice toward a target category. It operates by computing the gradients of the network's output layer relative to the specific class under review. These gradients quantify the exact decision-making contribution of various image regions and are

subsequently used to weight the feature maps of the final convolutional layer. This process generates a coarse localized activation map that can be superimposed onto the source image, visually exposing the critical zones that governed the network's prediction [34].

- **Shapley Additive Expansion:** Based on cooperative game theory, SHAP represents an advanced explainability technique designed to deconstruct and decode the decisions of highly complex models. To measure the importance of an individual variable, SHAP executes random permutations across the feature space, measuring how the model's output shifts relative to the baseline input. The resulting values reflect the average marginal contribution of each variable to the final prediction. These outputs are compiled into a SHAP plot, which explicitly details how much each feature shifted the model's decision-making trajectory for any given instance [36, 43].
- **Local Interpretable Model-agnostic Explanations:** To clarify an individual prediction, LIME isolates a localized data sample closely resembling the instance under evaluation. It then applies perturbations—introducing minor, randomized variations to these features—to build a synthetic, perturbed dataset. The baseline machine learning model then generates predictions for these perturbed entries, and the resulting outputs are mathematically weighted based on the spatial proximity and similarity between each perturbed instance and the original observation [43].

RQ6: What metrics were used to evaluate performance?

This inquiry serves to synthesize the benchmarking metrics utilized to quantify algorithmic performance across the 10 selected publications. The second column of the table catalogs the specific metrics extracted from each study. As demonstrated in Table 3, performance frameworks differ substantially based on the clinical target and the objective of the proposed methodology. However, the most universally deployed metrics include:

- **Accuracy:** This metric measures a model's capacity to correctly categorize instances across a dataset, defined mathematically as the ratio of correct classifications to total predictions. Within machine learning and deep learning contexts, accuracy defines how closely a set of empirical measurements aligns with the true reference values [48].
- **Precision:** This metric evaluates the quality and reliability of a model's positive predictions. It is calculated by dividing the number of true positive instances by the sum of true positives and false positives, thereby quantifying the model's exactness when designating an object as positive [49].
- **Recall:** This represents a model's capacity to successfully identify the complete set of positive instances contained within a dataset. Recall is computed as the ratio of true positives to the sum of true positives and false negatives. Fundamentally, recall reflects the model's sensitivity in capturing positive targets while minimizing the occurrence of false negatives [50].
- **AUC:** This metric measures the overall discriminative capacity of a classification model. The AUC represents the definitive area under the Receiver Operating Characteristic (ROC) curve, providing a unified scalar value that reflects how effectively a model separates positive and negative target classes [51].

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

Study	Evaluation Metrics	Research Challenge / Objective
Explainable Deep Learning for Pulmonary Disease and Coronavirus COVID-19 Detection from X-rays [34]	Sensitivity, Specificity, F-Measure, Accuracy, Time	Develop a fully automated and rapid diagnostic framework by leveraging deep learning techniques for COVID-19 detection from chest X-ray images.
Explainability for Artificial Intelligence in Healthcare: A Multidisciplinary Perspective [7]	Qualitative Evaluation	Deliver a broad examination of explainability in healthcare AI and explore its ethical implications for integrating AI-based tools into clinical workflows.

Deep Learning Interpretation of Echocardiograms [35]	Accuracy, AUC	Enable preliminary cardiac assessment in regions with limited access to trained cardiologists by using deep learning to quantify cardiac function and identify complex phenotypes.
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	Assess the practicality and clinical reliability of combining blood-test data with machine learning methods as an alternative to rRT-PCR for identifying COVID-19-positive patients.
Explainable Artificial Intelligence for Breast Cancer: A Visual Case-Based Reasoning Approach [39]	Accuracy	Introduce a Case-Based Reasoning (CBR) framework that can be automatically executed and visually presented through a user interface to support explainable reasoning and decision-making.
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	Investigate explainable mortality prediction in ICU settings using machine learning techniques and interpretable contribution measures.
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	Evaluate whether machine learning models utilizing time-series patient data can improve mortality forecasting for ICU patients through real-time prediction of 90-day mortality risk.
On the Interpretability of Machine Learning-Based Models for Predicting Hypertension [43]	Accuracy	Demonstrate the effectiveness of model-agnostic explainability approaches for interpreting random forest predictions of hypertension risk based on cardiorespiratory fitness information.
Learning Vector Representation of Medical Objects via EMR-Driven Nonnegative Restricted Boltzmann Machines (eNRBM) [44]	Recall, Precision, F-Measure	Propose a computational framework that exploits electronic medical records (EMRs) with limited human supervision using a constrained Boltzmann machine architecture.
Electronic Medical Record Phenotyping Using the Anchor and Learn Framework [46]	AUC	Create a phenotype repository that combines structured and unstructured EMR data to represent patients and support real-time clinical decision-making.
A New Concordant Partial AUC and Partial C Statistic for Imbalanced Data in the Evaluation of Machine Learning Algorithms [47]	AUC	Introduce alternatives to the traditional AUC metric, including partial AUC and the area under the precision–recall curve, for evaluating models on imbalanced datasets.

RQ7: What were the challenges addressed?

Artificial intelligence (AI) within the medical sector has fundamentally transformed approaches to diagnosing, treating, and managing clinical conditions, while introducing fresh challenges regarding the personalization of patient care. As previously highlighted in earlier sections, this rapidly evolving paradigm provokes essential inquiries concerning the transparency, interpretability, and ethical implementation of AI-driven predictive and decision-making systems. To evaluate these issues, we identified and analyzed 10 academic papers from the literature that utilize state-of-the-art methodologies. Ranging from the rapid identification of COVID-19 via chest X-ray imagery to improving the interpretability of AI-based clinical decision support tools, these investigations offer a thorough perspective on both the breakthroughs and barriers associated with embedding artificial intelligence into clinical medicine. The third column of **Table 3** outlines the specific challenges investigated across these publications. Correspondingly to the performance metrics, **Table 3** reflects the diverse array of difficulties encountered, which are shaped by the distinct application environments and goals of each methodology found in the literature. A granular examination of each selected paper is detailed below.

Brunese *et al.* [34] tackle their targeted problem by introducing a deep learning framework designed for the automated and accelerated diagnosis of coronavirus disease (COVID-19) utilizing chest X-rays. Their recommended protocol operates across three distinct stages. The initial phase detects the presence of pneumonia, the subsequent phase differentiates between standard pneumonia and COVID-19, and the final phase isolates the specific regions displaying COVID-19 manifestations on the processed X-ray. Evaluation outcomes involving 6,523 chest X-rays demonstrated high efficacy, yielding an average detection time for COVID-19 infection of 2.5 s and a mean accuracy of 97%. This technique leverages transfer learning via the VGG-16 architecture, achieving a 96% accuracy rate when distinguishing healthy individuals from patients suffering from routine pulmonary diseases, alongside a 98% accuracy rate specifically for COVID-19 detection.

Amann *et al.* [7] investigate the challenge of explainability within medical AI applications, centering their analysis on AI-powered clinical decision support systems. Utilizing a multidisciplinary framework, this investigation examines the issue through technical, legal, medical, and patient-centered lenses to critique the necessity of explainability in clinical settings. Their conclusions emphasize the potential hazards posed to both individual patient safety and broader public health networks when explainability mechanisms are absent from clinical decision support architectures.

Ghorbani *et al.* [35] propose implementing convolutional neural networks (CNNs) on echocardiographic imagery to conduct cardiac assessments. Their deep learning architecture, designated as EchoNet, delineates localized cardiac structures, calculates indicators of cardiac performance, and forecasts systemic phenotypes—such as age, gender, weight, and height—that can impact overall cardiovascular risk. This investigation indicates that incorporating interpretive frameworks can assist in pinpointing regions of interest, foster a deeper comprehension of typical variations in echocardiograms, and uncover clinical markers that might otherwise be overlooked by medical practitioners.

Brinati *et al.* [38] introduce an alternative approach for diagnosing COVID-19-positive individuals by applying machine learning configurations to routine blood panels. The researchers developed two distinct machine learning frameworks that evaluate standard blood test variables alongside demographics like age and gender. The diagnostic accuracy of these models spanned from 82% to 86%, while their sensitivity metrics ranged between 92% and 95%. The interpretability of this framework relies on a decision tree structure that obtained clinical validation, confirming the dependability of the selected diagnostic features.

Lamy *et al.* [39] introduce a breast cancer diagnostic methodology driven by "Case-Based Reasoning" (CBR), which operates simultaneously as an automated executable algorithm and a graphical user interface tailored for explanations. In contrast to opaque "black box" models, CBR yields easily verifiable outcomes by utilizing comparable historical cases as illustrative benchmarks. The visual layout incorporates a scatterplot derived from multidimensional dimension reduction. This graphical interface merges polar-MDS scatterplots with "rainbow boxes" to reframe the CBR problem as a "colour dominance" task.

Thorsen *et al.* [36] employ machine learning architectures trained on temporal data to facilitate real-time prognoses, aiming to improve the 90-day mortality prediction for patients admitted to intensive care units. An LSTM neural network framework is trained utilizing a one-hour temporal resolution. The outcomes indicate that predictive capacity grows stronger over time, as evidenced by the Matthews correlation coefficient and the area under the ROC curve rising from 0.29 and 0.73 at the time of admission to 0.57 and 0.88 at the point of discharge. External validation of the model was performed using data supplied by a fifth independent hospital. To demystify the generated predictions, the Shapley algorithm is applied to isolate the specific patient characteristics driving the forecasts at various temporal intervals.

Elshawi *et al.* [43] analyze methods for maximizing the interpretability of machine learning architectures, particularly models utilizing cardiorespiratory fitness metrics to forecast hypertension development risks via a random forest approach. Diverse model-agnostic explanation techniques, separated into global and local interpretive categories, are deployed. The authors observe that while local explanations emphasize individual patient instances, global interpretations assist medical professionals in understanding the overarching conditional distribution outlined by the response function. The researchers emphasize that both global and local strategies can be highly relevant depending on the particular demands of the clinical application. LIME delivers localized explanations established on historical data points and regional regression setups. Although LIME permits practitioners to make deductions regarding chronological shifts in patient attributes, it faces criticism due to its inherent instability and the introduction of linearity within localized modeling. Conversely, Shapley value estimation apportions the variance between the median forecast and the calculated distance across feature values. Even though this model ensures an equitable allocation of feature contributions, it is computationally intensive and necessitates access to the underlying training data used to construct the model.

Tran *et al.* [37] detail a computational architecture that leverages a non-negative restricted Boltzmann machine to exploit electronic medical record data with minimal to no manual intervention. By projecting medical data points into a low-dimensional vector space, this framework constructs an alternative representation of those medical features. The resulting model facilitates algebraic and statistical functions, such as feature clustering, risk stratification, and 2D plane projection. To optimize model interpretability, two constraints are integrated into the model parameters: (a) non-negative coefficients, and (b) structural regularity. The compiled representation facilitates short-term risk classification and highlights clusters of attributes that hold clinical significance.

Halpern *et al.* [46] focus on the utility of Electronic Medical Records (EMR) for determining optimal workflows in patient care. The recommended framework is designed to seamlessly extract statistically validated phenotypes with minimal manual configuration. To assist real-time clinical decision-making, a phenotype repository was established to represent patient states using both structured and unstructured data derived from EMRs. Evaluating prospectively collected baseline data, eight of these phenotypes were tested using retrospective records from emergency department patients. The data reveal that the generated phenotypes are easily interpretable, rapidly constructed, and perform on par with phenotypes derived statistically through intensive manual annotation.

Carrington *et al.* [47] utilize Receiver Operating Characteristic (ROC) curves and the area under the ROC curve (AUC) to assess the performance of classification configurations and diagnostic assessments. Particular attention is directed toward the complications introduced by imbalanced datasets, where positive and negative classes suffer from unequal representation. For ROC analysis, the researchers introduce a novel concordant partial AUC alongside a new partial c statistic, which serve as vital metrics and techniques for understanding and decoding specific ROC tracks and regional AUC zones. These newly formulated partial metrics are verified for mathematical equivalence and are derived directly from standard AUC and c statistics, respectively. Their utility is tested using a conventional ROC baseline alongside two authentic breast cancer datasets.

Discussion

This study intends to assess the academic community's focus on implementing AI interpretability and explainability techniques within the clinical domain over the last decade. Following an examination of publication trajectories spanning

from 2013 to 2023, a pronounced upward trend in research interest was detected, particularly from 2019 onward, culminating in 974 documents indexed by Scopus and 164 indexed by WOS in 2023. The foundation of our research comprised 448 overlapping articles. The scope of our evaluation encompassed publication trajectories, dissemination platforms, contributing nations, and international collaboration structures. Prominent contributors were mapped, with the United States positioning at the forefront (114 articles), followed sequentially by China, Italy, and other nations.

To refine the granularity of our deductions, we concentrated our review on 10 targeted papers. Core clinical application domains comprised COVID-19, Alzheimer's disease, cardiovascular disorders, electrocardiograms, along with brain and breast malignancies, highlighting the societal relevance of these medical areas. Furthermore, we investigated the most widely deployed AI architectures, including CNN, LSTM, and the Boltzmann machine, alongside the predominant methodologies leveraged for explainability and interpretability, such as clinical decision support configurations, Grad-Cam, SHAP, and LIME.

Reflected in our research observations, the swift integration of Machine Learning (ML) and Deep Learning (DL) frameworks has catalyzed expanding anxieties regarding these methodologies. Growing apprehension exists concerning the technical complexity of AI models and the deficit of transparency characterizing both data mining operations and automated decision pathways.

The comprehensive literature synthesis executed in this research demonstrates that the bulk of academic work addressing interpretability and explainability within medical AI architectures centers primarily on neural networks and machine learning models. A significant portion of the analyzed papers deploy interpretability tools for AI systems, with multiple studies accentuating the intimate relationship connecting interpretability and explainability [10].

Concurrently, in specific contexts, interpretability and explainability are regularly utilized as interchangeable concepts. The academic community requires the formulation of standardized definitions and precise terminology to optimize the transmission of findings and clinical insights. Present AI architectures still lack a unified formal framework that could assist end-users in developing confidence in the applied methodologies.

To the best of our knowledge, the contemporary academic community is actively concentrating on the following areas to advance explainability and interpretability algorithms:

- **Built-in interpretability:** Designing ML and DL architectures to ensure they remain inherently more interpretable and transparent. The objective centers on clarifying the underlying logic dictating a model's outputs, thereby mitigating the opacity associated with intricate architectures. Researchers simultaneously seek to engineer these algorithms to strike an optimal equilibrium between interpretability and operational capacity [52].
- **Self-interpretable models:** Engineering architectural frameworks or models that are intrinsically comprehensible on their own, eliminating the necessity for post hoc diagnostics. These configurations aim to guarantee procedural transparency directly from the initial engineering and optimization phases [53, 54].
- **Post hoc explainability methods:** Deploying analytical techniques to decode and clarify pre-trained AI frameworks that are inherently opaque and complex. These approaches function by delivering retroactive justifications for the determinations produced by a model without modifying its foundational optimization or learning mechanisms [55].
- **Stakeholder involvement:** Cultivating greater clinical integration and comprehension by fostering a rising commitment to engaging clinicians, patients, and allied stakeholders directly in the layout and interpretability assessment of these models [56].

Conclusion

Clinical diagnostics, therapeutic interventions, and medical decision-making are increasingly dependent upon ML and DL frameworks; however, the compounding complexity of these models introduces pronounced difficulties, particularly when

negotiating high operational capacity against systemic interpretability. Achieving clear decision-making pathways, maintaining ethical and regulatory alignment, and securing the confidence of healthcare operators and medical institutions remain critical prerequisites for embedding AI tools in these specific sectors. Because predictive accuracy is vital, researchers must carefully weigh whether to favor diagnostic accuracy or explainability and interpretability, given the distinct divide separating transparent white-box models from opaque black-box frameworks. Interpretability and explainability remain vital for cultivating user trust, maintaining regulatory and ethical integrity, and driving broader societal integration. These elements move beyond purely technical considerations, manifesting as ethical and practical obligations within clinical spaces where algorithmic deductions yield major consequences. Appraising historical investigations underscores the pressing demand for transparent and interpretable AI frameworks in medicine. Although advancements have been realized in visualization instruments, feature attribution analysis, and structural complexity reduction, fundamental problems persist. These include the inherent opacity of large-scale architectures, the operational trade-offs connecting capacity and clarity, and structural biases embedded within optimization datasets. The ethical weight of AI deployments in healthcare and clinical medicine remains undeniable. Resolving challenges linked to algorithmic bias, data confidentiality, system security, and statutory liability is vital to safeguard ethical AI execution. Conversely, the psychological ramifications of algorithmic conclusions on users—especially in clinical environments—are closely tied to transparent communication regarding the scope and boundaries of artificial intelligence systems. Fundamental takeaways encompass the necessity for a steadfast commitment to transparency, algorithmic equity, and patient-centric development of AI solutions in medicine, alongside the formulation of rigorous benchmarks and guidelines to consolidate trust among end-users and organizational stakeholders.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 06 Oct 2023

Revised: 01 Dec 2023

Accepted: 08 Jan 2024

Published online: 15 March 2024

Rights and permissions

Open Access The author(s) retain copyright. This article is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. It may be shared and adapted for non-commercial purposes with appropriate attribution, an indication of changes, and distribution of adaptations under the same license. Third-party material may be subject to separate terms identified in its credit line. View the license at <https://creativecommons.org/licenses/by-nc-sa/4.0/>.

References

- London AJ. Artificial intelligence and black-box medical decisions: accuracy versus explainability. *Hastings Cent Rep*. 2019;49(1):15-21.
- Hakkoum H, Abnane I, Idri A. Interpretability in the medical field: a systematic mapping and review study. *Appl Soft Comput*. 2022;117:108391.
- Loyola-Gonzalez O. Black-box vs. white-box: understanding their advantages and weaknesses from a practical point of view. *IEEE Access*. 2019;7:154096-154113.
- 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. p.31-36.
- Gilvary C, Madhukar N, Elkhader J, Elemento O. The missing pieces of artificial intelligence in medicine. *Trends Pharmacol Sci*. 2019;40(8):555-564.
- General Data Protection Regulation. General data protection regulation (GDPR) [Internet]. Intersoft Consulting; 2018 [cited 2018 Oct]. Available from: Intersoft Consulting website.
- 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.
- Phillips PJ, Hahn CA, Fontana PC, Broniatowski DA, Przybocki MA. Four principles of explainable artificial intelligence. Gaithersburg (MD): National Institute of Standards and Technology; 2020. 18 p.
- Nassih R, Berrado A. 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. p.1-5.
- Linardatos P, Papastefanopoulos V, Kotsiantis S. Explainable AI: a review of machine learning interpretability methods. *Entropy*. 2021;23(1):18.
- Alicioglu G, Sun B. A survey of visual analytics for explainable artificial intelligence methods. *Comput Graph*. 2022;102:502-520.
- 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*. 2018;22(5):1589-1604.
- Shashanka M, Raj B, Smaragdis P. Sparse overcomplete latent variable decomposition of counts data. In: *Advances in Neural Information Processing Systems* 20. 2007.
- 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. p.1135-1144.
- Langer M, Oster D, Speith T, Hermanns H, Kästner L, Schmidt E, et al. 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.
- Murdoch WJ, Singh C, Kumbier K, Abbasi-Asl R, Yu B. Definitions, methods, and applications in interpretable machine learning. *Proc Natl Acad Sci U S A*. 2019;116(44):22071-22080.
- Combi C, Amico B, Bellazzi R, Holzinger A, Moore JH, Zitnik M, et al. A manifesto on explainability for artificial intelligence in medicine. *Artif Intell Med*. 2022;133:102423.
- von Eschenbach WJ. Transparency and the black box problem: why we do not trust AI. *Philos Technol*. 2021;34(4):1607-1622.

- Arrieta AB, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, et al. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf Fusion*. 2020;58:82-115.
- Biran O, Cotton C. Explanation and justification in machine learning: a survey. In: *IJCAI-17 Workshop on Explainable AI (XAI)*; 2017. p.8-13.
- Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D. A survey of methods for explaining black box models. *ACM Comput Surv*. 2018;51(5):1-42.
- Miller T. Explanation in artificial intelligence: insights from the social sciences. *Artif Intell*. 2019;267:1-38.
- Tjoa E, Guan C. A survey on explainable artificial intelligence (XAI): toward medical XAI. *IEEE Trans Neural Netw Learn Syst*. 2020;32(11):4793-4813.
- 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.
- Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A survey on bias and fairness in machine learning. *ACM Comput Surv*. 2021;54(6):1-35.
- Chakrobartty S, El-Gayar O. Explainable artificial intelligence in the medical domain: a systematic review. 2021.
- Hatherley J, Sparrow R, Howard M. The virtues of interpretable medical artificial intelligence. *Camb Q Healthc Ethics*. 2022;1-10.
- 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):120-138.
- Ali S, Abuhmed T, El-Sappagh S, Muhammad K, Alonso-Moral JM, Confalonieri R, et al. Explainable artificial intelligence (XAI): what we know and what is left to attain trustworthy artificial intelligence. *Inf Fusion*. 2023;99:101805.
- Band SS, Yarahmadi A, Hsu CC, Biyari M, Sookhak M, Ameri R, et al. Application of explainable artificial intelligence in medical health: a systematic review of interpretability methods. *Inform Med Unlocked*. 2023;40:101286.
- Ballew BS. Elsevier's Scopus database. *J Electron Resour Med Libr*. 2009;6(3):245-252.
- Drake M. *Encyclopedia of library and information science*. Boca Raton (FL): CRC Press; 2003. Vol. 1.
- Van Eck N, Waltman L. Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*. 2010;84(2):523-538.
- 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.
- Ghorbani A, Ouyang D, Abid A, He B, Chen JH, Harrington RA, et al. Deep learning interpretation of echocardiograms. *NPJ Digit Med*. 2020;3(1):10.
- Thorsen-Meyer HC, Nielsen AB, Nielsen AP, Kaas-Hansen BS, Toft P, Schierbeck J, 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-e191.
- Tran T, Luo W, Phung D, Harvey R, Berk M, Kennedy RL, et al. Risk stratification using data from electronic medical records better predicts suicide risks than clinician assessments. *BMC Psychiatry*. 2014;14(1):76.

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.

Lamy JB, 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.

Cho B-J, Choi YJ, Lee M-J, Kim JH, Son G-H, Park S-H, et al. Classification of cervical neoplasms on colposcopic photography using deep learning. *Sci Rep*. 2020;10(1):1-10.

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.

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.

Eishawi 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.

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.

Tran T, Phung D, Luo W, Venkatesh S. Stabilized sparse ordinal regression for medical risk stratification. *Knowl Inform Syst*. 2015;43:555-82.

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.

Carrington AM, Fieguth PW, Qazi H, Holzinger A, Chen HH, Mayr F, et al. 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.

Menditto A, Patriarca M, Magnusson B. Understanding the meaning of accuracy, trueness and precision. *Accredit Qual Assur*. 2007;12:45-7.

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.

Buckland M, Gey F. The relationship between recall and precision. *J Am Soc Inform Sci*. 1994;45(1):12-9.

Huang J, Ling CX. Using AUC and accuracy in evaluating learning algorithms. *IEEE Trans Knowl Data Eng*. 2005;17(3):299-310.

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.

Ashwath VA, Sikha OK, Benitez R. TS-CNN: a three-tier self-interpretable CNN for multi-region medical image classification. *IEEE Access*; 2023.

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.

Dwivedi R, Dave D, Naik H, Singhal S, Omer R, Patel P, et al. Explainable AI (XAI): core ideas, techniques, and solutions. *ACM Comput Surv*. 2023;55(9):1-33.

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.