BMC Medical Informatics and Decision Making

SYSTEMATIC REVIEW

Open Access

Contrastive explainable AI in healthcare: a systematic review of trends, benefits, research gaps, and future directions

Samuele Pe[1*†], Antonino Angi[2*†], Cristian Izquierdo[2], Enea Parimbelli[1,3], Machteld Johanna Boonstra[4,5], Folkert Wouter Asselbergs[5,6,7] and Karim Lekadir[2,8]

Abstract

Background Contrastive eXplainable Artificial Intelligence (cXAI) has emerged as a promising approach to align machine learning explanations with human reasoning. Unlike classical XAI, which addresses questions such as "Why P?", contrastive explanations explicitly consider alternatives, answering "Why P, rather than Q?". By making the implicit comparison explicit and focusing only on features that distinguish between outcomes, cXAI has been argued to provide more cognitively aligned, less ambiguous, and more actionable explanations than classical XAI. Moreover, in the context of healthcare, this comparative structure closely mirrors clinicians' abductive reasoning in differential diagnosis and treatment selection. Our purpose is to systematically examine how contrastive explanations have been developed and applied in healthcare; emphasize the related benefits; identify current trends, research gaps, and future directions.

Methods We conducted a systematic literature review of studies published from 2018 onward, following the PRISMA 2020 guidelines. Five scientific databases (IEEE Xplore, Scopus, PubMed, ACM Digital Library, and SpringerLink) were queried using structured search strings, complemented by an AI-assisted search using the Elicit tool. Eligible studies focused on the development or application of contrastive explanation techniques in healthcare-related machine learning tasks. Extracted information included data modalities, AI models, contrastive methods, visualization techniques, and validation strategies.

Results Only 18 studies met the eligibility criteria, reflecting the relative novelty of the field, while also indicating that the translation of cXAI into healthcare applications is still at an early stage. Most works focused on tabular data and medical imaging, while physiological signals and video/volumetric data were notably underrepresented. Contrastive Explanation Method (CEM) was the most frequently employed approach. Visualization methods for explanations predominantly relied on heatmaps or feature-importance displays, with textual explanations mainly used in studies

[†]Samuele Pe and Antonino Angi contributed equally to this work.

*Correspondence:
Samuele Pe
samuele.pe01@universitadipavia.it
Antonino Angi
antonino.angi@ub.edu

Full list of author information is available at the end of the article

© The Author(s) 2026. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

involving patients or lay users. Importantly, only a small number of studies reported empirical evaluations involving end users, indicating a lack of rigorous validation.

Conclusions cXAI offers a clinically aligned and cognitively coherent approach to explainability in healthcare by explicitly framing decisions in relation to meaningful alternatives. Despite its strong conceptual suitability, its adoption in real-world medical settings remains limited, with underexplored modalities and scarce user-centered validation. Advancing cXAI in healthcare will require *(i)* rigorous clinical assessment of explanations and systematic collection of user feedback and *(ii)* the development of multimodal, human-centered explanation frameworks. Through these efforts, cXAI can evolve from a promising theoretical paradigm into a practical tool for clinically aligned decision support.

Keywords Clinical decision support, Contrastive explanations, Explainable artificial intelligence (XAI), Healthcare informatics

Background

Contrastive explanations have recently gained increasing attention within the field of eXplainable Artificial Intelligence (XAI) [1], particularly in high-stakes domains such as healthcare. Their theoretical foundations lie in philosophy and cognitive science, and they offer a framework that differs conceptually from traditional explanatory paradigms. In this section, we introduce contrastive explanations and clarify their distinctions from related concepts. We also delineate the benefits of contrastive explanations in healthcare, especially in comparison to classical XAI, thus providing the necessary background and motives for the subsequent systematic review.

Introduction to contrastive explanations

Within the broader field of XAI, contrastive explanations have emerged as an innovative paradigm, capable of representing the reasoning process of machine learning (ML) models in a more intuitive and human-aligned manner compared to classical XAI methods. While traditional XAI techniques typically address the question "Why P?", where *P* denotes the model's predicted output, contrastive XAI (cXAI) introduces a comparison dimension "Why P, rather than Q?" [2]. This mirrors the structure of human-generated explanations, frequently comparative in nature [3], thereby reducing users' cognitive load by making explanations more concise, unambiguous, and pragmatically oriented.

In general, a contrastive explanation does not attempt to list every reason behind an outcome like classical XAI would do, but rather points only to what distinguishes the observed outcome P from an alternative Q. Conceptually, this intuition has a long history in philosophy: for example, an early account by Lewis [4] stated that to explain *why P, rather than Q*, one looks for a cause that is present when P occurs but would have been absent if Q had occurred instead. Lipton [5] later refined this view through what he called the *difference condition*, which further narrows the search of causal factors. Miller [6] placed these ideas on a more formal footing, expressing them in terms compatible with modern causal and computational models. This account distinguishes explanations based on what the alternative Q represents and, thus, what is the phenomenon to be explained (termed *explanandum*). Miller talks about counterfactual explananda, when Q is a fully hypothetical alternative that did not occur, usually called the "foil"; the explanation concerns why P happened instead of this unrealized possibility. The second category is the bi-factual explananda, when Q - termed here as the "surrogate" - is an outcome that did occur, but under different circumstances. To further clarify these distinctions, we may refer to Van Bouwel et al. [7], who identified four types of explanatory questions:

- **Plain fact**: "Why does object X have property P?", the classical XAI approach.
- **P-contrast**: "Why does object X have property P, rather than property Q?", counterfactual explanandum.
- **O-contrast**: "Why does object X have property P, but object Y has property Q?", bi-factual explanandum between objects.
- **T-contrast**: "Why does object X have property P at time t_1, but property Q at time t_2?", bi-factual explanandum across time.

From this definition and formalization, contrastive explanations can be more clearly distinguished from another important XAI paradigm, *counterfactual* explanations (or, simply, "counterfactuals"), despite their frequent conflation in the literature [8] and the terminology similarity (e.g. in "counterfactual explananda"). In simple terms, a contrastive explanation describes why an AI system predicted one outcome rather than a specific alternative (Q) by identifying the features that distinguish the two outcomes according to the model's internal representation. For example, a lesion may be classified as "malignant" rather than "benign" because of its irregular borders. A counterfactual explanation, instead, describes what

would need to change in the input for the prediction to switch to an alternative outcome. In the same example, the explanation might state that if the lesion had regular borders, the prediction would change from "malignant" to "benign".

More formally, while cXAI addresses the question "Why P, rather than Q?" by identifying the features that differentiate two outcomes, counterfactual explanations instead answer "What should change for Q to occur instead of P?" by searching for minimal modifications to the input that allow for this class change [8, 9]. In practice, as described by AlRegib and Prabhushankar [10], counterfactuals are inherently intervention-based, focusing on hypothetical alterations such as "What if the tumor were absent?" or "What if the mass had a different shape?" and discussing their outcomes. To further clarify these distinctions and the related nomenclature, we refer to Table 1.

Despite these conceptual differences, the distinction is not always clear in practice. Many contrastive explanation methods are attribution-based (e.g., Prabhushankar et al. [11]), whereas some approaches, such as the Contrastive Explanation Method (CEM) [12], rely on input perturbations, a mechanism also central to counterfactual generation. It is not a surprise that while some works explicitly distinguish the two concepts [9–11, 13], others use the terms interchangeably [14–18] and others, finally, consider counterfactuals as a subsample of contrastive explanations, by highlighting their inherent contrastive nature and, implicitly, identifying them as an example of bi-factual cXAI [19]. Nevertheless, in line with established works in the context of explainable AI [8, 10, 11], we treat counterfactuals as conceptually distinct from contrastive explanations.

The benefits of cXAI on medical data

This alignment with human reasoning characterizing cXAI becomes even more significant in domains where comparison is intrinsic to decision-making. Among these, healthcare stands out: contrastive explanations are particularly relevant in this high-stakes domain, where the diagnosis process is fundamentally differential [20]. Clinicians formulate multiple diagnostic hypotheses that may explain the observed clinical evidence (symptoms and signs), *compare* them, and select the most plausible and simple one. To do so, they seek evidence that distinguishes one candidate diagnosis (e.g., pneumonia) from either its negation ("not pneumonia") or from competing alternatives (*e.g.,* influenza or bronchitis) [12]. It is therefore unsurprising that clinicians tend to prefer contrastive explanations over classical ones [21].

Surpassing the limits of classical XAI

The central objective of eXplainable AI is to make the decision-making process of black-box ML models more transparent and accessible to human users [22]. Traditional XAI methods aim to answer the absolute question "Why P?", describing the set of factors leading to the predicted outcome. However, explanations cannot provide true causal inference [23], and are therefore better understood as the representation of *observed correlations* [10]. Nevertheless, this attempt to highlight all the factors influencing the results often produces exhaustive, feature-heavy explanations that may overwhelm users and hinder their ability to draw actionable insights from explanations [24, 25].

Miller [3], through an extensive, interdisciplinary review of literature spanning philosophy, psychology, and cognitive science, demonstrated that human reasoning is inherently contrastive: when confronted with the question "Why P?", people instinctively introduce a term of comparison and reformulate the question as "Why P, rather than not-P?" or "Why P, rather than Q?". In classical XAI, however, this comparison term (the foil) is typically left implicit, making it impossible for users to find a suitable foil to understand the explanation. As noted by

Table 1 Conceptual comparison between contrastive explanations (with counterfactual or bi-factual explananda) and counterfactual explanations

	cXAI with counterfactual explananda	cXAI with bi-factual explananda	Counterfactual explanations
Guiding question	"Why P, rather than Q?"	"Why P in the current situation, but Q in another situation?"	"What should change for Q to occur instead of P?"
Mechanism	Often observational; identifies the features that differentiate the actual outcome P from the foil Q.	Often observational; identifies the features across two factual instances (or time points) that differentiate between P and the surrogate Q.	Intervention-based; searches for minimal input modifications that would change the prediction from P to Q.
Output	A set of discriminative features (or a visual representation thereof) that separate P from Q.	Paired sets of discriminative features (or a visual representations thereof) contrasting two real cases (or temporal states), explaining why P holds in one case and Q in the other.	A modified input (or feature change) that would produce Q instead of P.
Clinical example	"The lesion is classified as malignant rather than benign due to its irregular borders."	"Lesion X is malignant because it shows irregular borders, lesion Y is benign because it has smooth borders."	"If the lesion had regular borders, the model would classify it as benign."

Gouveia and Malík in the context of medical AI [26], current XAI methods identify which features are important but rarely clarify *why* they matter, generating an *interpretability gap*. Users may therefore assume a foil that differs from the one implicitly assumed by the method, leading to misinterpretations. This mismatch can trigger harmful cognitive biases: clinicians may selectively interpret explanations to confirm a preferred decision (*confirmation bias* [27]) or, conversely, simply rely on automated recommendations (*automation bias* [28]) without fully understanding them. Making the foil explicit addresses this ambiguity by directly supporting the intended comparison, reducing cognitive effort and better aligning explanations with users' reasoning processes [25].

Moreover, contrastive XAI avoids the need to account for all relevant features. As discussed by Jacovi et al. [25], the contrastive formulation prunes the space of potential causes to only those that differentiate between outcomes. For instance, explaining "Why is this image a beetle?" requires listing many anatomical characteristics (e.g., number of legs, wings, body segmentation). Instead, the contrastive question "Why a beetle, rather than a spider?" can be answered simply by noting the presence of six legs in the input image [3].

These considerations have contributed to growing regulatory and ethical support for contrastive approaches. Institutions such as the European Data Protection Supervisor (TechDispatch on XAI [29]) and the Council of Europe [30] emphasize the importance of explanations that are meaningful, contestable, and aligned with human reasoning. In the clinical domain itself, theoretical and early empirical work suggests that contrastive explanations may reduce cognitive load, improve interpretability, and facilitate more informed and confident decision-making [21, 31, 32].

Fostering abductive and differential reasoning

The integration of AI into decision-making has sparked extensive debate due to the fundamental differences between human and machine reasoning. As an example, Prabhushankar and AlRegib [10, 33] show, in the context of COVID-19 detection from CT scans, how models can achieve good classification performances focusing on wrong, disease-unrelated, or simply not human-interpretable patterns in the image. One well-known consequence is *algorithm aversion* [34], where users reject AI recommendations due to concerns about nonsensical or harmful errors, often driven by the black-box nature of ML models. Early attempts to mitigate this issue focused on improving model accuracy and enhancing interpretability through classical XAI approaches, as well as promoting human-like outputs to strengthen user trust [35]. However, such improvements have contributed to the opposite effect: *automation bias* [28], in which users over-rely on AI-generated suggestions and undervalue their own professional judgment. This phenomenon generally manifests as automation complacency [36] and loss of human agency [37]. Nevertheless, automation bias is especially problematic in high-stakes domains, such as healthcare itself, where uncritical dependence on AI can contribute to clinician deskilling [38], lack of skills development [39], medical de-responsibilization [40], and defensive medicine [41].

Generally, transparent frameworks in which users can monitor, interact with [3], and overrule AI suggestions [42] have proven especially valuable in decision support, but, with the introduction of the concept of "contrast", baseline requirements for decision-support should be rethought. Several lines of work in healthcare have already shown that explanations relying on comparisons (rather than relying on single unilateral explanations) foster effective and synergistic human–AI collaboration. For example, research on Frictional AI [43] and Judicial AI [44] demonstrate that grounding clinical recommendations in XAI-based "juxtaposed evidence" improves diagnostic accuracy and confidence, particularly in complex cases or for more experienced users, while simultaneously reducing automation bias and preserving human agency. Buçinca et al. [45] showed that explaining the difference between AI prescriptions and human reasoning, via contrastive explanations, helps clinicians in detecting AI mistakes or misaligned assumptions, and not only limits the risk of deskilling, but also promotes human decision-making skills.

In addition to this, a central limitation of machine learning models is that they operate through *induction*: they extract patterns and correlations from data but do not articulate explicit causal rules. XAI methods attempt to provide interpretable representations of these inductively derived patterns. However, recent perspectives emphasize the value of XAI not only as an explanation device but also as a form of *reliability metadata* that stimulate reflective human engagement and supports safer, more deliberate decision-making [46]. This aligns with a shift from merely accepting AI-generated hypotheses, toward using AI as a tool to generate and support candidate hypotheses for the observed phenomena (*abductive reasoning* [47]) and this transition has gained momentum in XAI research [3]. For instance, AlRegib and Prabhushankar [10] argue that classical, counterfactual, and contrastive explanations can together generate a set of hypotheses describing the input (either in terms of presence or absence of a given property), enabling abductive inference about the underlying data. They also show that the combination of these three explanatory paradigms is sufficient to fully characterize all the relevant information in a model, thereby achieving *complete explanations* [48].

Other studies highlight the importance of abductive reasoning paradigms in medical AI. For example, Lim et al. [21], by developing a system for generating clinically relevant, contrastive explanations, demonstrate that abductive explanations are preferred over classical saliency maps in clinical diagnosis. Gouveia and Malík [26] further argue that clinicians routinely face uncertainty and must justify why one diagnosis fits better than others (differential diagnosis [20]); explanations should therefore be both "abductive and social", supporting hypothesis generation and selection based on patient evidence. Contrastive XAI aligns with this diagnostic reasoning structure, fostering trust and comprehension in AI-assisted decisions. Moreover, it can enhance patient-centered communication, by allowing clinicians to answer questions such as "Why this treatment instead of another?" or "What would happen if we chose a different approach?". This aligns with the principles of informed consent, in which patients must be able to weigh the benefits, risks, and uncertainties of different options before agreeing to a course of action [49]. By framing AI outputs in a contrastive and abductive manner, cXAI not only supports the cognitive processes of clinicians but also directly contributes to ethical, patient-centered care by embedding transparency and reasoning clarity into the informed consent process.

Motivations

Despite the promising potential of contrastive explainable artificial intelligence in medicine, as discussed in the previous sections, the field is currently at a critical stage of development. While more studies are proposing contrastive explanation methods, the concept of contrastive explanations remains relatively new in healthcare and is frequently conflated with partially related or distinct approaches, such as counterfactual explanations or contrastive learning. This lack of conceptual clarity makes relevant contributions difficult to identify, interpret, and compare.

At the same time, the rapid expansion of artificial intelligence in clinical contexts has created an urgent need for explanation methods that are not only technically robust but also aligned with clinical reasoning, fostering trust and supporting (rather than overriding) clinical decision-making. However systematic comparisons of existing cXAI techniques are still lacking in clinical settings.

Table 2 positions the present review within the existing literature on contrastive and counterfactual explanations. We decided to consider both reviews of contrastive explanations and reviews of counterfactual explanations because of the widespread confusion between these two concepts and their terminology (see Sect. 1.1). Most previous reviews have focused on counterfactual explanations and have adopted a cross-domain perspective rather than a healthcare-specific one. Furthermore, not all reviews followed a systematic methodology: for example, although the review by Stepin et al. [8] was conducted systematically, it followed the outdated Systematic Literature Review (SLR) guidelines proposed by Kitchenham et al. [56] rather than the PRISMA 2020 framework. For each review, we report its primary focus, application domain, number of databases or search engines used (DBs), and number of included studies. To the best of our knowledge, the present work is the first PRISMA-based systematic review specifically dedicated to contrastive explanations in healthcare.

These factors highlight why a systematic review is particularly needed at this point in time: the field has reached sufficient maturity to warrant consolidation, yet remains too disorganized to effectively guide researchers and practitioners. Without a structured synthesis, the risk is continued conceptual ambiguity, duplication of efforts, and slow translation into clinical practice. Therefore, beyond clarifying why contrastive explanations represent a powerful resource for healthcare, this work aims to provide a comprehensive and timely overview of the current state of the art. We present the first systematic review of cXAI in healthcare, with the goal of identifying trends, highlighting research gaps, and outlining future

Table 2 Existing reviews and surveys on contrastive and counterfactual explainability

Review	Year	Explanation type	Domain	Systematic (methodology)	DBs	Included studies
Chou et al. [50]	2021	Counterfactual	General	Yes (PRISMA)	3	125
Stepin et al. [8]	2021	Contrastive and counterfactual	General	Yes (SLR)	2	113
Karimi et al. [17]	2022	Contrastive and counterfactual	General	No	Not available	60
Zhang et al. [51]	2023	Counterfactual	General	No	Not available	Not available
Guidotti [52]	2024	Counterfactual	General	No	Not available	59
Verma et al. [53]	2024	Counterfactual	General	No	Not available	>350
Jiang et al. [54]	2024	Counterfactual	General	No	Not available	35
Buñay-Guisñan et al. [55]	2026	Counterfactual	Education	Yes (PRISMA)	4	29
This review	**2026**	**Contrastive**	**Healthcare**	**Yes (PRISMA)**	**5 + Elicit**	**18**

Note: SLR = Systematic Literature Review guidelines. DBs = number of databases or search engines used. 'Not available" indicates that the information could not be retrieved from the article or its supplementary materials, or that the review itself was not publicly accessible

directions for the field. The scope of this review is limited to peer-reviewed studies on the use or development of contrastive explanations in any healthcare application. The review was conducted by querying five established scientific repositories and performing additional screening using the Elicit tool [57].

The contributions of our work can be summarized as follows:

- **Discussion of the benefits:** A review of general-purpose, cXAI-related literature, provided in Sect. 1.2, summarized the main characteristics of contrastive explanations, especially in relation to classical XAI, highlighting those features that make cXAI a breakthrough resource in the healthcare domain.
- **Identification of trends:** The core of our paper consisted of a systematic review of the literature concerning the development and application of cXAI in healthcare, resulting in the evaluation of 18 peer-reviewed studies (Sect. 3).
- **Research gaps:** Based on the results of the systematic review, several research gaps were identified and discussed (Sect. 4.1).
- **Possible future directions:** We conclude the discussion by outlining future directions for cXAI that may help bridge the previously identified gaps and support the translation of this innovative framework into clinical decision-making processes (Sect. 4.2).

Briefly, our findings indicate that the field is still in an early stage of development, showcasing a restricted case-mix of applications and limited empirical validation in clinical settings. We also acknowledge important limitations of the review itself: in particular, the small number of included studies and our deliberately focused scope excluded works that adopt contrastive reasoning without explicitly framing it as such. The patterns presented here should therefore be read as descriptive observations on an emerging area of research, rather than as generalizable conclusions about cXAI in healthcare. Within these limits, this work provides a structured map of methods, data modalities, target users, and validation practices, and outlines research gaps and future directions to support the further development of cXAI in clinical settings.

Methods

This systematic review of the literature on the development or application of contrastive explanation techniques in healthcare was designed and reported following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 guidelines [58], which provide a standardized framework for ensuring transparency, methodological rigor, and replicability in the conduct and reporting of systematic reviews. We limited our search to articles published from 2018 onward, following the release of Miller's influential review advocating for contrastive explainability [3], which first appeared on *arXiv* [59] in 2018. Our methodological approach is organized around four main phases: *(i)* identification of relevant studies through structured database queries and complementary AI-assisted retrieval; *(ii)* screening of records against pre-specified eligibility criteria; *(iii)* full-text eligibility assessment; and *(iv)* inclusion of the selected studies in the qualitative synthesis. The remainder of this section details the rationale for each methodological choice, and the full identification, screening, and selection process is summarized in the PRISMA flow diagram (Fig. 1).

The identification phase relied on structured queries across five well-established scientific repositories: *IEEE Xplore* [60], *Scopus* [61], *PubMed* [62], *ACM Digital Library* [63], and *SpringerLink* [64]. These databases were selected to provide complementary coverage of the publication venues most relevant to cXAI in healthcare: PubMed indexes the biomedical and clinical literature; IEEE Xplore and the ACM Digital Library cover the computer-science and engineering venues where most methodological work on XAI is published; SpringerLink provides access to the substantial body of medical-informatics research disseminated through Springer journals and conference proceedings; and Scopus serves as a broad multidisciplinary indexer capturing cross-domain publications that bridge AI and healthcare. As done in previous systematic reviews of AI and XAI in healthcare [65, 66], this combination of repositories ensures coverage across biomedical, computer-science, and multidisciplinary venues. We did not separately query other repositories (e.g., *Web of Science*, *Embase*), as their indexed content either overlaps substantially with the selected databases or lies outside our inclusion criteria, focusing on domains such as pharmacology or clinical trials that are tangential to the field of healthcare. Preprint repositories such as *arXiv* were also not queried directly, in line with our eligibility criteria, which restrict inclusion to peer-reviewed articles to ensure methodological reliability. To further mitigate the risk of database-specific indexing gaps, we complemented the structured queries with a search using the *Elicit* tool [57], which retrieves articles independently of database-specific syntax and indexing conventions, thereby ensuring more comprehensive coverage of the literature on healthcare-oriented cXAI.

Table 3 summarizes the search queries used across the different databases. The queries were intentionally designed to balance specificity and recall. In line with previous works [8], the use of the wildcard expression

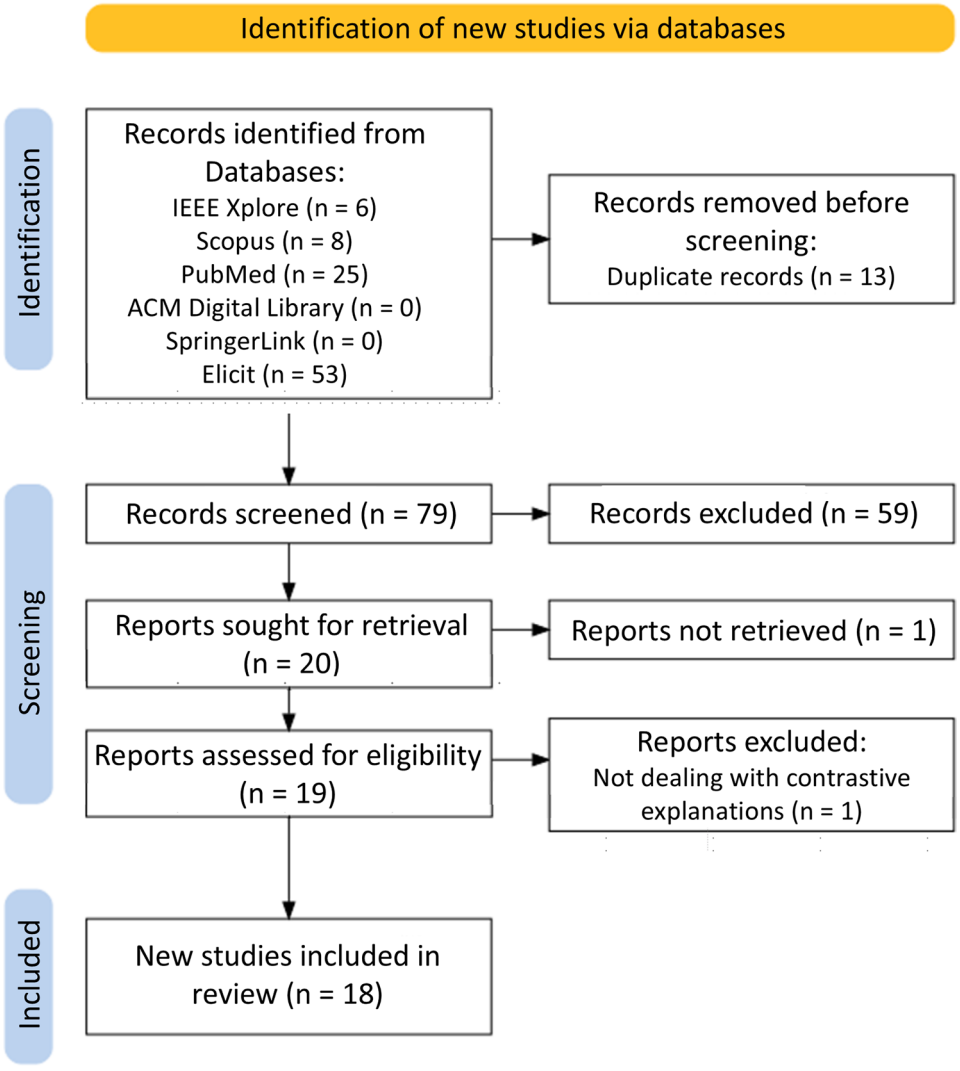


Fig. 1 PRISMA flow diagram illustrating the systematic literature review process. The diagram illustrates the number of records identified, screened, excluded, and included at each stage, following the PRISMA 2020 guidelines [58]

Table 3 Search queries for each database and the research information for Elicit

Database / Research Engine	Query String / Research Information
IEEE Xplore	("contrastive expla*") AND ("health*" OR "medic*" OR "clinic*" OR "patient*" OR "diagnos*" OR "treat*" OR "patient*")
Scopus	TITLE-ABS-KEY("contrastive expla*") AND TITLE-ABS-KEY("health*" OR "medic*" OR "clinic*" OR "patient*" OR "diagnos*" OR "treat*")
PubMed	("contrastive explanation"[Title/Abstract] OR "contrastive explainability"[Title/Abstract]) AND ("health"[Title/Abstract] OR "medical"[Title/Abstract] OR "clinic"[Title/Abstract] OR "patient"[Title/Abstract] OR "diagnos"[Title/Abstract] OR "treat"[Title/Abstract])
ACM Digital Library	(acmdlTitle:"contrastive expla*" OR recordAbstract:"contrastive expla*") AND (acmdlTitle:("health*" OR "medic*" OR "clinic*" OR "patient*" OR "diagnos*" OR "treat*") OR recordAbstract:("health*" OR "medic*" OR "clinic*" OR "patient*" OR "diagnos*" OR "treat*"))
SpringerLink	"contrastive expla*" AND ("health*" OR "medic*" OR "clinic*" OR "patient*" OR "diagnos*" OR "treat*")
Elicit	*Research statement:* Contrastive Explanations in Healthcare. *Filters:* • Focus on contrastive explanations (*Why P, rather than Q?*) • Conducted in healthcare or clinical contexts *Selection threshold:* 3.5

"contrastive expla*" was adopted to capture multiple terminological variants commonly used in the literature, including "contrastive explanation", "contrastive explanations", and "contrastive explainability". While this choice may appear restrictive, it reflects the explicit linguistic framing used by authors who intentionally adopt a contrastive explanatory perspective, as opposed to works that implicitly rely on contrast without formalizing it as such. Domain-related keywords (e.g., "health*", "clinic*", "diagnos*") were included to ensure relevance to healthcare contexts while allowing flexibility across medical subfields. Queries were first refined within IEEE Xplore to establish a baseline strategy and subsequently adapted to the syntax and indexing conventions of the remaining databases, ensuring consistency across sources while respecting their structural differences.

Unlike conventional databases, Elicit does not rely on Boolean keyword queries. The tool draws largely on the Semantic Scholar corpus, which currently indexes over 200 million academic publications across disciplines, and uses language-model-based semantic similarity to retrieve articles relevant to a natural-language research question, even when their wording does not exactly match that of the query [67, 68]. Additional filters can then be applied in the form of specific questions, with the answers used to characterize and score each retrieved article. To complete the selection process, a relevance score threshold must be chosen. Various thresholds were tested, and a score of *3.5* was selected, as it provided the best balance between the number of retrieved studies and their relevance to the research question. Elicit was used solely as a candidate-retrieval mechanism: all records returned by the tool were exported and subjected to the same PRISMA-compliant eligibility screening applied to the records retrieved from the five structured databases, with no inclusion decision delegated to the tool itself. This protocol is consistent with the findings of Bernard et al. [69], who evaluated Elicit against a traditional umbrella-review workflow and concluded that AI-assisted retrieval may serve as a useful complementary strategy, provided that retrieved records are subject to human oversight and screened against the review's original inclusion and exclusion criteria. According to this, Elicit was used to broaden the search and reduce the risk of database-specific indexing gaps, while manual eligibility assessment remained the basis for all inclusion decisions.

Screening and selection were performed according to specifically designed eligibility criteria, which were chosen to ensure both conceptual clarity and practical relevance:

- **Inclusion criteria**:
  - Articles focusing on the use or development of contrastive explanations in any healthcare application;
  - Peer-reviewed articles published from 2018 onwards.
- **Exclusion criteria**:
  - Reviews, extended abstracts, and book chapters;
  - Studies not involving machine learning or deep learning models for decision support;
  - Studies focusing on contrastive explanations in non-medical domains (e.g., chemistry, philosophy, cognitive sciences) without empirical application to healthcare or clinical decision-making;
  - Studies exclusively concerning *contrastive learning* methods;
  - Studies exclusively concerning *counterfactual explanations*.

We focused exclusively on peer-reviewed studies that propose or apply contrastive explanations in healthcare settings, in order to capture methods explicitly intended to support medical decision-making. Reviews and book chapters were excluded for two reasons. First, this review focuses on characterizing novel methodological developments in cXAI and their applicability to clinical tasks, which requires primary methodological or empirical contributions. Reviews generally synthesize existing studies rather than report original methods or clinical applications, while book chapters are often expository in nature. Second, because cXAI in healthcare is still a small and emerging field, including both reviews and the primary studies they summarize could count the same work twice and make the field appear more widely developed or clinically adopted than it really is. Furthermore, studies exclusively addressing *contrastive learning* were excluded because such approaches focus on representation-learning objectives rather than on explanation generation. Nevertheless, to ensure the collection of all relevant papers, the exclusion of contrastive learning-related works was not injected in the search queries, but was rather an objective of the subsequent screening procedures. Similarly, studies solely concerned with *counterfactual explanations* were omitted in the screening phases to avoid conceptual overlap and terminological ambiguity, as discussed in Sect. 1.1, and to preserve a clear focus on explicitly contrastive explanatory frameworks.

A detailed overview of the identification, screening, and selection phases is provided in the PRISMA flow diagram (Fig. 1). The diagram was realized leveraging *PRISMA2020* R package [70] and then adapted for visualization purposes.

To complement the PRISMA flow diagram, which documents the attrition of records across the four review phases, a UML activity diagram of the operational workflow is also provided (Fig. 2). The activity diagram maps the sequential actions and decision nodes followed during the review, in order to support exact replication of the procedure.

Results

This section presents the plain results of the systematic review, synthesizing the characteristics, methodological choices, and application contexts of the studies that met the eligibility criteria. We first provide an overview of the included articles, highlighting publication trends, data types, clinical tasks, and cXAI methods adopted. We then focus on how contrastive explanations are displayed and validated, examining visualization strategies, intended target users, and the extent to which empirical user evaluations have been conducted. Together, these results offer a structured perspective on the current status and trends of cXAI applications in healthcare and help identify prevailing patterns and gaps in the literature.

Overview of identified approaches

Table 4 displays in chronological order (with ties ordered by authors' names) the studies included in our review, briefly describing their properties. Among the 18 included works, tabular data (10/18, 56%) and medical imaging (6/18, 33%) are the most common modalities; the Contrastive Explanation Method (CEM) [12] is the most frequently adopted contrastive technique (7/18, 39%); and only 2/18 (11%) studies report empirical user evaluations.

Despite searching across the five selected databases and complementing them with an Elicit-assisted search, only 18 studies met the eligibility criteria. This likely reflects both the relative novelty of the field and the fact that the translation of cXAI into healthcare applications is still at an early stage, as we further discuss in Sect. 4.1.1. To complement this concern, in four cases (Dhurandhar et al. [12, 80], Alvarez Melis et al. [31], Luss et al. [74]) the reported experiments serve primarily as illustrative demonstrations on open-source data of the proposed methodologies rather than as evaluations conducted within real clinical workflows. Furthermore, several papers share common authors (e.g., Dhurandhar et al. [12, 80], Luss et al. [74]), suggesting that relatively few research groups are actively pursuing cXAI in medical settings. We also remark that many works are recent, with only 3 out of 18 (17%) published before 2021, reflecting a growing interest in this area.

The distribution of data types across the 18 studies is heterogeneous: ten focus on tabular data, six on medical images, only two on signal data, and none concern video or raw volumetric data. This aligns with trends in the broader cXAI field, which also favors tabular and image-based tasks over time-varying data [84]. Nevertheless, we identify this as a major flaw of current literature, emphasizing the importance of temporal and 3D data in clinical AI as we properly discuss in Sect. 4.1.2. Notably, another underrepresented data type is natural language, despite outside the field of healthcare several cXAI works address text processing [25, 85].

A prominent technique in the reviewed studies is the *CEM* (Contrastive Explanation Method) [12], one of the earliest cXAI methods, which was employed or adapted in 7/18 papers. CEM identifies *pertinent positives* (PP), features that must be present for the model to predict a given class P, and *pertinent negatives* (PN), features that are absent but their presence would change the prediction to class Q. CEM has been applied in contexts such as disease detection/diagnosis [72, 83] and emotion recognition [74, 75]. The prevalence of CEM across the included studies is not justified through head-to-head comparisons within the corpus itself, but several factors plausibly contribute to its popularity, as discussed in Sect. 4.1.2.

To further characterize this finding, Table 5 reports for each of the included studies using CEM the validation strategy adopted, together with the main findings supporting the suitability, robustness, and practical usefulness of CEM in healthcare-related applications of contrastive explainability. Overall, the available evidence consistently supports the effectiveness of CEM across a variety of application domains and data modalities. Whenever quantitative comparisons were performed, CEM consistently outperformed traditional attribution methods such as LIME in generating valid explanations, while qualitative evaluations highlighted its ability to produce clinically meaningful and user-appreciated explanations by jointly identifying PPs and PNs. Collectively, these findings confirm that CEM provides a robust and practically useful framework for generating faithful, interpretable, and actionable contrastive explanations.

Another remarkable cXAI method example is the work of Fahad et al. [82], which builds upon the technique originally introduced by Jacovi et al. [25], one of the most influential contributions in the cXAI literature. This model-agnostic approach projects embedding vectors into a latent space that maximizes the distinction between classes P and Q, and leverages the perturbation of latent features along with a contrastive behavior metric to get the explanation. A final interesting technique is *Contrastive Grad-CAM* [10], used in the work of Prabhushankar and AlRegib for CT-based COVID-19 detection [33]: it consists in an extension of Grad-CAM [86] for contrastive scenarios, and, exactly like Grad-CAM, it is potentially adaptable to any convolutional-based task

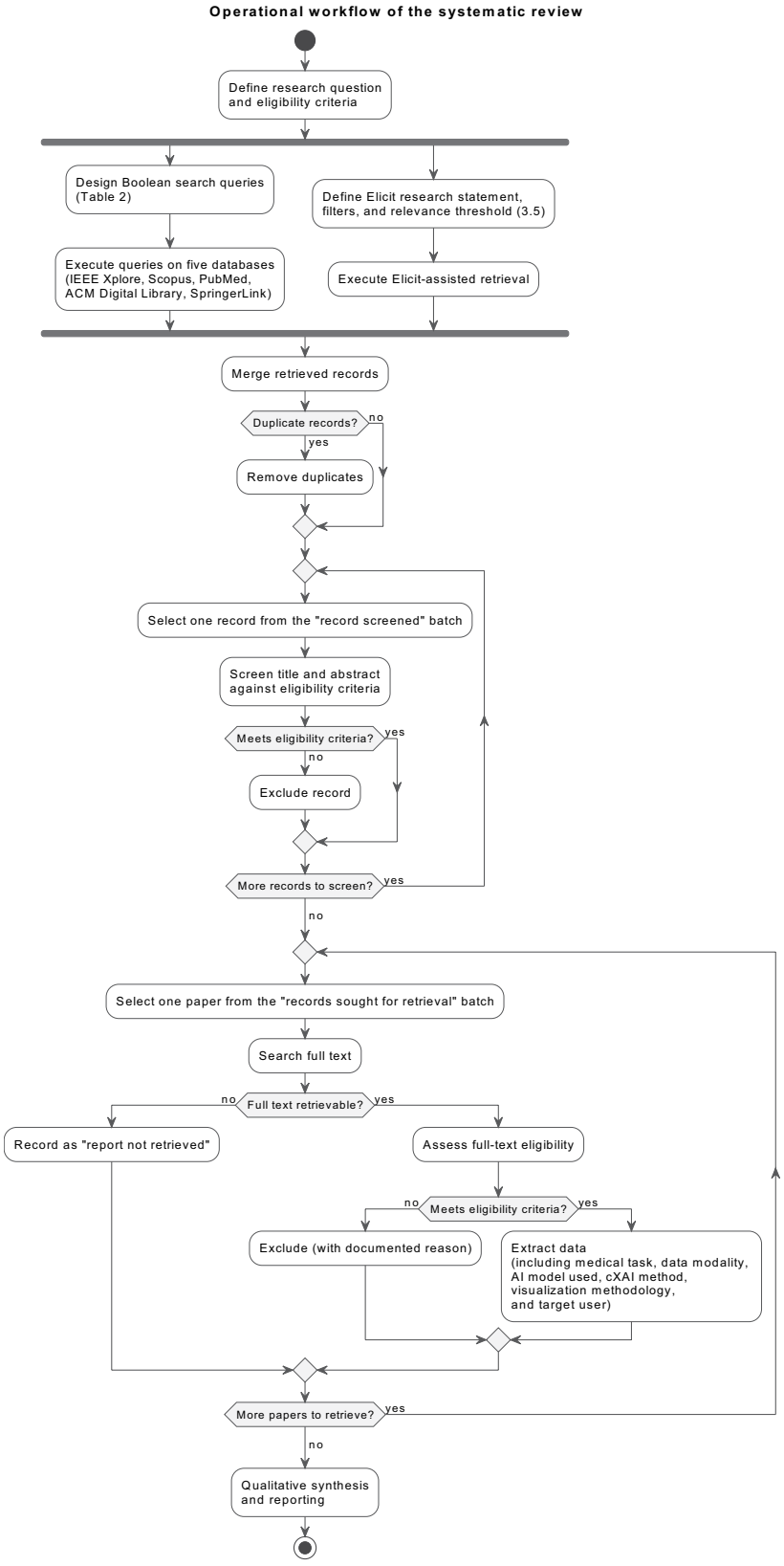


Fig. 2 UML activity diagram of the systematic-review workflow

Table 4 Summary of the articles included in our systematic review

Ref.	Task(s)	Data	AI method(s)	cXAI method	Visualization	Target user(s)
Dhurandhar et al. [12]	Autism diagnosis	Image	Multilayer perceptron	CEM	Heatmap	Not reported
Neerincx et al. [71]	Activity completion prediction	Tabular	Not reported	PeCoX framework	Textual	Clinicians, young patients, parents
Roberts and Tsiligkaridis [72]	COVID-19 detection	Image	Convolutional neural network	CEM	Contrastive examples	Not reported
Alvarez Melis et al. [31]	Diabetes, breast cancer, heart disease detection	Tabular	Naïve bayes	Weight of Evidence	Bar plot	Not reported
Banerjee et al. [73]	Mortality prediction	Tabular	Random forest, logistic regression	Class-contrastive heatmap	Heatmap	Clinicians, patients, information-technology professionals
Luss et al. [74]	Skin lesions, emotion recognition	Image	Convolutional neural network	CEM	Heatmap, contrastive examples	Information-technology professionals
Mishra et al. [75]	Mental health prediction	Tabular	Decision tree, random forest, multilayer perceptron	CEM	Highlight features	Not reported
Prabhushankar and Alregib [33]	COVID-19 detection	Image	Convolutional neural network	Contrastive Grad-CAM	Heatmap	Not reported
Finzel et al. [76]	Emotion recognition	Signal	Rule-based learning	Prototypes	Contrastive examples	Not reported
Jacobsen et al. [77]	Polyp segmentation	Image	Convolutional neural network	Uncertainty estimation	Heatmap	Not reported
Tahirovic and Krivic [78]	Breast cancer detection	Tabular	Logistic regression	Prototypes	Textual	Clinicians
Banerjee [79]	Mortality prediction	Tabular	Random forest, logistic regression	Class-contrastive heatmap	Heatmap	Clinicians, patients, information-technology professionals
Dhurandhar et al. [80]	Diabetes, vertebral column disease detection	Tabular	Decision tree, random forest, multilayer perceptron	CEM	Textual	Not reported
Balikci Cicek et al. [81]	Parkinson's diagnosis	Tabular	Multilayer perceptron	CEM	Heatmap, highlight features	Not reported
Buçinca et al. [45]	Health-related exercise selection	Tabular	Support vector machines	Embedding differences	Textual	Clinicians
Fahad et al. [82]	Gastro-intestinal disease detection	Image	Convolutional neural network	Contrastive projection	Highlight features	Not reported
Lim et al. [21]	Heart disease detection	Signal	Convolutional neural network	DiagramNet	Contrastive examples	Clinicians
Kucukakcali et al. [83]	Ovarian cancer diagnosis	Tabular	Multilayer perceptron	CEM	Highlight features	Not reported

Note: "Not reported" indicates that the paper did not explicitly identify any target user for the explanations

[87]. Collectively, all these examples highlight the strong potential for extending established, state-of-the-art cXAI methodologies to the healthcare domain.

Our review also identifies innovative, case-specific techniques that may inspire new methodological developments or that can be further applied in medical contexts. Alvarez-Melis et al. [31], for instance, utilize the Weight of Evidence (WoE) from information theory [88] to generate importance scores comparing the predicted class to a foil. Banerjee et al. [73, 79] propose a perturbation-based method for feature importance scoring, leading to the construction of *class-contrastive heatmaps*. Other examples feature prototype-based methods [76, 78], which aim to find solutions to identify exemplar instances that help users understand why one prediction was made instead of another. This variety of techniques illustrate the diversity of contrastive approaches and their applicability to medical decision-support tasks.

Table 6 and Fig. 3 summarize the key characteristics of the main contrastive explainability approaches identified. The table reports, in alphabetical order, the methods that emerged as the most prominent and potentially reusable across a wide range of healthcare applications. For each method, we outline its main data and model requirements (if applicable), provide a concise description of its functioning and the mechanism through which

Table 5 Summary of representative studies evaluating the contrastive explanation method (CEM) and its validation

Ref.	Usage of CEM and validation strategy	Main findings
Dhurandhar et al. [12]	Evaluated on MNIST, procurement fraud, and brain activity datasets. Compared qualitatively with LIME and LRP and quantitatively assessed whether PPs preserve the original prediction and PNs change the prediction.	CEM was the only method capable of identifying both PPs and PNs. Quantitative evaluation showed that PPs preserved the original class and PNs changed the prediction in all evaluated datasets (100% success). Furthermore, on the brain connectivity dataset, CEM identified clinically meaningful autism-related connectivity alterations, whereas LRP highlighted the relevant visual-network regions in only about 40% of autistic subjects.
Roberts and Tsiligkaridis [72]	A perturbation-based CEM-inspired framework was evaluated on COVID-19 ultrasound images by comparing standard and adversarially robust CNNs. Evaluation focused on robustness and visual interpretability.	Robust CNNs produced more stable and clinically meaningful contrastive explanations than standard CNNs. Perturbation-based explanations were better aligned with human visual perception, although no quantitative comparison with other XAI methods or user study was performed.
Luss et al. [74]	The proposed latent-feature CEM was compared with Grad-CAM and LIME using three image datasets (MNIST, CalebA, ISIC Lesion). The comparison evaluated the ability of PPs (or equivalent explanations for the competing methods) to preserve the original prediction. Additionally, a user study involving 200 participants on an emotion recognition task assessed the perceived usefulness of each method individually and when sequentially combined with another explanation technique.	CEM was the only method achieving a 100% success rate in preserving the original prediction. It also received significantly higher usefulness ratings than Grad-CAM, with statistically significant pairwise improvements. Participants particularly appreciated the additional information provided by PNs, which complemented conventional attribution maps.
Mishra et al. [75]	CEM was integrated into an explainable reasoning engine together with feature importance and counterfactual explanations for psychological disorder prediction but no real evaluation of CEM was performed.	CEM evaluation absent.
Dhurandhar et al. [80]	MACEM (a model-agnostic extension of CEM) was quantitatively compared with LIME and SHAP. Evaluation measured the ability of PPs to preserve the original prediction, PNs to change the prediction, and Correct Feature Ranking - which assesses how accurately the explanations reflect the contribution of features to the model's confidence - across multiple tabular datasets (German Credit, FICO, Vertebral Column …) and classifiers (decision tree and frandom forest).	MACEM achieved a 100% success rate in both preserving the original prediction with PPs and changing it with PNs across all evaluated datasets, consistently outperforming LIME and SHAP in generating valid PPs and PNs. CEM generally achieved higher Correct Feature Ranking scores than LIME, while SHAP occasionally obtained higher scores because its optimization objective closely matches the evaluation metric.
Balikci Cicek et al. [81]	CEM and LIME were jointly applied to explain a deep neural network trained on volumetric brain MRI features for Parkinson's disease diagnosis. Validation relied on consistency between explanation methods.	CEM feature importances were consistent with the established algorithm LIME and neuroanatomical evidence, supporting the robustness of the CEM explanations.
Kucukakcali et al. [83]	CEM was applied to explain a deep neural network for ovarian cancer diagnosis. Validation was based on clinical interpretation of PPs and PNs.	The authors concluded that the main advantage of CEM lies in its ability to generate patient-specific contrastive explanations, revealing not only why a prediction was made, as traditional attribution methods do, but also how the diagnosis would change following minimal modifications to the input. According to the authors, this additional counterfactual information improves model transparency and supports clinician trust.

contrastivity is achieved, and describe the resulting form of contrastive explanation. Figure 3 complements this overview by providing a high-level schematic representation of each method: it visually illustrates the underlying workflow, highlights the mechanism through which contrastivity is achieved, and depicts the form of explanation generated by each approach.

To conclude, current research is characterized by a limited number of studies, often produced by a small group of contributors, and frequently validated in controlled or illustrative settings rather than real clinical environments. Methodologically, there is a clear reliance on a few established approaches, most notably CEM and its variants, alongside the adaptation of cXAI methodologies developed outside the field of healthcare. At the same time, the heterogeneity of proposed solutions, including perturbation-based and prototype-based methods, highlights ongoing experimentation and conceptual diversity within the field. From a data perspective, the predominance of tabular and image-based applications, coupled with the scarcity of work on temporal, volumetric, and textual data, reveals a significant gap between current research efforts and the complexity of real-world clinical data.

Explanation display and validation

In the context of cXAI, explanation display plays a crucial role, as contrastive reasoning inherently relies on

Table 6 Overview of the principal contrastive explainability methods identified in the reviewed studies

cXAI method	Included studies	Explicit restrictions	Output	Functioning
Class-contrastive heatmap (Fig. 3a)	Banerjee et al. [73], Banerjee [79]	Introduced for a binary problem but extensible to a multiclass problem.	*Feature importance*: a heatmap highlighting for each subject if each perturbation move the prediction toward P or toward Q.	*Perturbation-based.* Modify one or more input features and measure how the predicted probability changes for class P (and class Q).
Contrastive Explanation Method (Fig. 3b)	Dhurandhar et al. [12, 80], Roberts and Tsiligkaridis [72], Luss et al. [74], Mishra et al. [75], Balikci Cicek et al. [81], Kucukakcali et al. [83]	None, it is framework-agnostic.	*Important features*: PP, features whose presence is minimally sufficient to justify P, and PNs, features whose presence would justify Q.	*Perturbation-based.* Compute PPs by identifying the input perturbation δ^{pos} that minimises a specifically designed cost function, e.g., as defined in Eq. 3 of Dhurandhar et al. [12]. Compute PNs by identifying the input perturbation δ^{neg} that minimises a specifically designed cost function, e.g., as defined in Eq. 1 of Dhurandhar et al. [12].
Contrastive Grad-CAM (Fig. 3c)	Prabhushankar and Alregib [33]	Introduced for an image-based problem but extensible to signals and video/volume data. Only applicable to convolutional-based models.	*Feature importance*: a heatmap highlighting the features that contribute more strongly to class P than to class Q.	*Attribution-based.* The method is described in Prabhushankar et al. [11]. Define a contrastive objective function between the predicted class and the foil class (e.g., $L(P,Q) = logit_P - logit_Q$). Compute contrastive gradients w.r.t. the parameters of a chosen layer and apply Global Average Pooling: $\alpha_k = \text{GAP}(\frac{\partial L(P,Q)}{\partial A^k})$ Finally aggregate the activations of the chosen layer with the computed weights: $map = \text{ReLU}\left(\sum_k \alpha_k A^k\right)$
Contrastive projection (Fig. 3d)	Fahad et al. [82]	None, it is framework-agnostic.	*Feature importance*: a score for each feature (or embedding dimension) quantifying its contribution to distinguishing class P from class Q. Positive values support P, whereas negative values support Q.	*Perturbation-based.* The method is inspired from the work of Jacovi et al. [25]. Perturb input features either in the original feature space or in a latent embedding space and quantify the effect of the perturbation using a contrastive behavior score: $\delta^{\text{contr}}_{p,q} = \frac{p_P}{p_P+p_Q} - \frac{q_P}{q_P+q_Q}$,
Prototypes (Fig. 3e)	Finzel et al. [76], Tahirovic and Krivic [78]	None, it is framework-agnostic.	*Contrastive examples*: used to highlight discriminative differences for an input x_i from x_q*.	*Prototype-based.* The method is based on the near miss selection method from Winston [89]). Define a similarity metric and, after considering an item belonging to class P, find the most similar example belonging to class Q.
Weight of Evidence (Fig. 3f)	Alvarez Melis et al. [31]	None, it is framework-agnostic.	*Feature importance*: positive values support class P, negative values support class Q.	*Attribution-based.* The method is based on the Weight of Evidence score from Good [88]. For each feature, compute the WoE value: $woe(P/Q : x_i) = \log \frac{p(x_i\|P)}{p(x_i\|Q)}$

comparison, differentiation, and cognitive alignment with the user's decision process. Visualization and presentation choices directly influence how effectively users can interpret contrasts between alternatives, assess their relevance, and integrate them into clinical reasoning. Another issue that is frequently neglected in literature is validation: without empirical evaluation involving end users, it remains unclear whether contrastive explanations are presented appropriately and whether they truly improve understanding, trust calibration, or decision quality in practice.

In this regard, in Table 4 we provide, for each included article, the visualization technique adopted, along with information about the intended users. An interesting point in this small set of results is the high variability of visualization techniques used for cXAI explanations. Most of the retrieved papers (10/18, 56%) employ importance-score visualization modalities. These techniques vary in complexity, ranging from simple mentions (or display) of the most relevant features [75, 82] to an exhaustive listing of features along with their corresponding importance values [81, 83]; from the use of bar plots (or tornado-like plots [31]) to heatmaps, which can serve either as a quick comprehensive summarization/comparison tool [73, 79] or be superimposed on the original data in the case of images [33].

In prototype or perturbation-based methods, the tendency is to directly show examples (4/18 papers, 22%): either exemplar items from the dataset in the former case, or perturbed items in the latter. Perturbed items

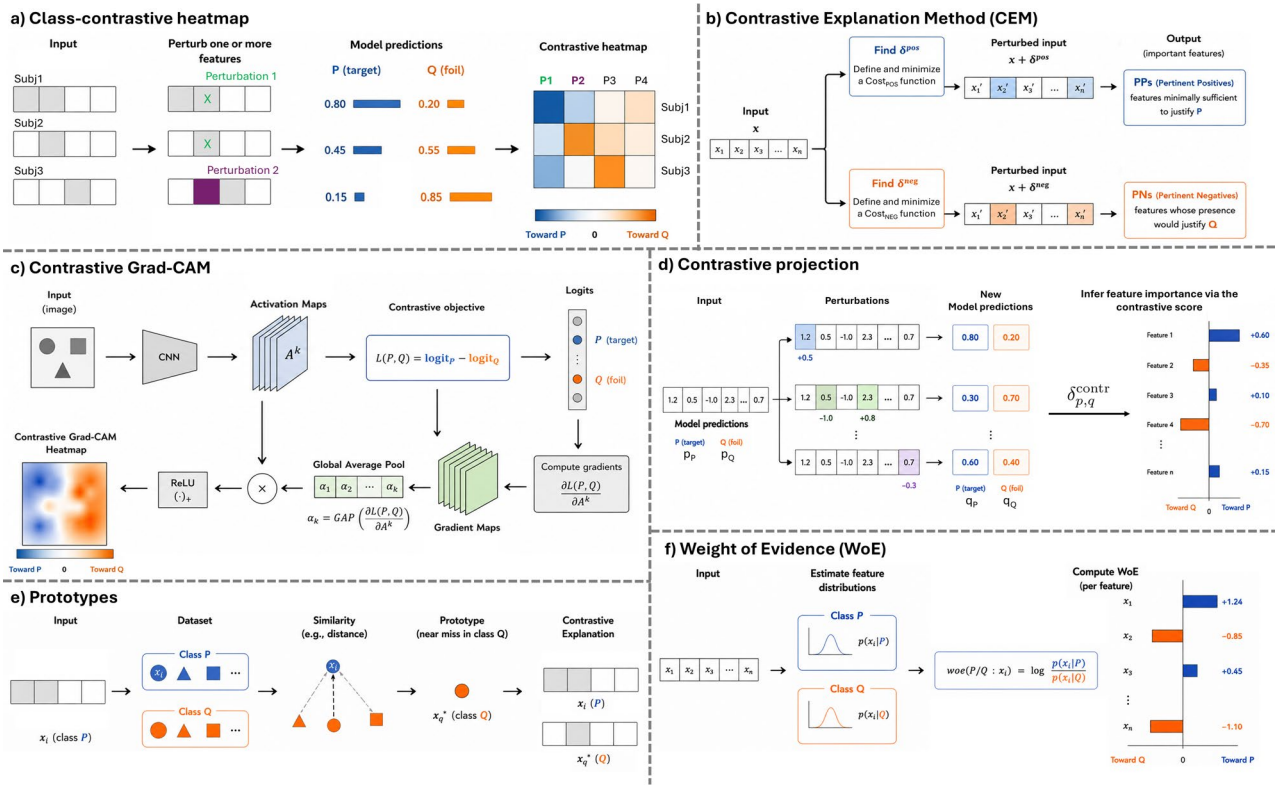


Fig. 3 Visual representation of the principal contrastive explainability methods identified in the reviewed studies: **a**) class-contrastive heatmap, **b**) contrastive explanation method, **c**) contrastive Grad-CAM, **d**) contrastive projection, **e**) prototypes, and **f**) Weight of evidence

can also be processed through Generative AI algorithms (e.g., variational autoencoders) to enforce realism and credibility [74]. Finally, 4 out of 18 papers (22%) employ natural language processing techniques, for example [78], or complex large language model (LLM) solutions [45] to convert explanations into textual form. These strategies are consistent with broader trends in contemporary AI and represent a promising direction for healthcare applications, as they enable the generation of hybrid, user-centered explanations that can be adapted to the needs of diverse stakeholders, including clinicians and patients. These aspects are further explored in Sect. 4.2.

We note that almost every article explicitly mentioning the target user refers to clinicians. Articles that do not clearly state the intended user (10 out of 18, 56%) are also implicitly designed for clinicians, given the type of task, mostly diagnosis. Nevertheless, three articles (17% of the papers) [71, 73, 79] consider patients (or their parents, in the case of young subjects) as targets for explanations. This is clearly related to the nature of the study: Neerincx et al. [71], for example, describe a basic chatbot intended for young diabetes patients to monitor self-management activities. It is worth noting that heatmap-based and textual explanations are the preferred visualization methods in these three articles, generally due to their high understandability. This strong predominance of clinicians as target users likely reflects the diagnostic focus of most included studies, but also indicates that the explanation design choices observed in this body of work are largely shaped around expert reasoning, with comparatively little attention to patient-facing contexts. The implications of this imbalance for explanation design are discussed in Sect. 4.1.2. Notably, two works (11%), by Banerjee et al. [73, 79] and Luss et al. [74], also mention information technology professionals as potential users or target audiences for their research.

A key finding of our review concerns the limited extent of empirical validation. While many of the included articles discuss the utility and adaptability of contrastive explanations either in general [12, 31] or specifically within the healthcare domain [21], only 2 of the 18 included studies (11%) conducted full user evaluations to empirically assess the efficacy, usability and practical utility of cXAI methods. The remaining studies relied on illustrative examples, qualitative plausibility arguments, or comparisons with established clinical literature. We return to the possible reasons for, and the implications of, this scarcity of end-user evaluations in Sect. 4.1.1.

Luss et al. [74] compared their CEM-based approach (CEM-MAF) with LIME [90] and Grad-CAM [86], investigating whether explanations were useful in isolation and how user understanding changed when an additional

explanation was provided after an initial one. Their results showed that cXAI was perceived as the most useful and outperformed other XAI methods, largely due to the use of pertinent negatives, which offered very intuitive insights. They also demonstrated that pertinent positives can help identify classifiers that should not be trusted, specifically those requiring too few PPs. Moreover, introducing a contrastive explanation as additional complementary resource significantly improved user understanding. Overall, user feedback suggested that contrastive explanations excel when clarity and minimal cognitive effort are needed.

Buçinca et al. [45] experimented comparing different human–AI interaction protocols: no AI support, classical XAI support and multiple cXAI supports with different foils, including an ML-inferred human prediction, the user's unsupported prediction and a random foil. They showed that explaining the difference between AI recommendations and inferred human decisions through contrastive explanations helped clinicians detect AI errors or misaligned assumptions. This approach not only mitigated the risk of deskilling [38] but also promoted human decision-making skills. While cXAI performed comparably to classical explanations, it did not fully eliminate over-reliance on AI; nevertheless, it significantly enhanced users' sense of agency [37] and autonomy, with "open-minded" users benefiting the most from contrastive explanations.

Nonetheless, even in the absence of formal user-study evaluations, some studies attempt to conclude the medical significance of their results. For instance, Balikci Cicek et al. [81] derived clinical implications by interpreting CEM scores in light of established literature on Parkinson's disease neuroanatomy. Other works focused on methodological insights: Neerincx et al. [71] found that different user groups, defined by age and role, preferred different explanation modalities, while Jacobsen et al. [77] compared uncertainty estimation techniques for the generation of cXAI, showing, for example, that Monte Carlo dropout produced highly correlated predictions with weaker contrasts, making its uncertainty estimates less informative as contrastive explanations compared to ensemble-based methods.

Overall, the analysis of explanation display and validation practices reveals a field that is methodologically diverse but still immature in terms of user-centered design and empirical grounding. Visualization strategies vary widely, with a predominance of importance-based representations complemented by example-driven and hybrid textual approaches, reflecting an effort to balance fidelity and interpretability. However, these design choices are largely oriented toward expert users, with clinicians implicitly or explicitly considered the primary audience, and only limited attention given to patient-facing explanations or multi-stakeholder needs. Most critically, the scarcity of rigorous user evaluations highlights a significant gap between methodological development and real-world applicability: while many studies argue for the potential benefits of contrastive explanations, only a small fraction provide empirical evidence of their effectiveness in practice. We hope that our observations encourage further research on cXAI in healthcare, particularly toward rigorous validation studies that assess not only methodological soundness but also real-world usefulness, impact, and adoption in clinical practice.

Discussion

This section provides an in-depth discussion of the results of this review. Drawing on the analysis of the studies included in our systematic review, we identify the main methodological and practical gaps that emerge and outline how cXAI can be further leveraged to address explainability requirements and enhance clinical decision support systems, in line with the benefits identified in the previous sections.

Research gaps and challenges

Despite strong conceptual alignment with clinical reasoning, contrastive explainability remains underexplored in real-world medical settings. Existing studies often focus on methodological development rather than validation, and the adoption of cXAI across different data modalities, user groups, and clinical tasks is uneven. Understanding these shortcomings is critical to identifying research priorities, addressing methodological limitations, and ultimately enabling cXAI to fulfill its promise as a tool for improving explainability, trust, and decision-making in clinical workflows.

The lack of research

A limited evidence base Our systematic review highlights that cXAI remains only sparsely adopted in healthcare research and practice, as evidenced by the limited number of included articles (18). This gap is striking given that many clinical processes, such as differential diagnosis, treatment selection and risk stratification, are inherently contrastive. Most medical AI systems still rely on classical XAI paradigms that answer absolute "Why?" questions, even though these are often misaligned with how clinicians naturally reason and justify decisions. This suggests that the limited uptake of cXAI is not due to a lack of theoretical suitability, but rather to methodological inertia, limited tooling, and insufficient empirical evidence demonstrating its added value in real clinical workflows.

Lack of user-centered validation The vast majority of reviewed studies apply cXAI methods without explicitly involving clinicians, patients, or other end-users in the design or evaluation loop. Only a small fraction of works (2/18) conduct user studies to assess whether contrastive explanations actually improve diagnostic accuracy, calibrate clinician confidence, support error detection, or mitigate automation bias and deskilling. As a result, many claims about the clinical usefulness of cXAI remain speculative in these works. We consider this lack of validation a major concern, although this phenomenon is common across several clinical domains, for example, a recent systematic review of AI in pediatric oncology imaging reports a reliance on classical attribution methods and scarcity of user-centered validation comparable to those we witnessed [91].

Two main factors help explain why only few studies include end-user evaluations. First, a substantial part of the corpus consists of methodological contributions, in which the proposed cXAI technique is illustrated on publicly available clinical datasets rather than deployed in a clinical workflow: in four of the included works, the medical use case is presented explicitly as a demonstration [12, 31, 74, 80]. Second, the cXAI literature in healthcare is still young and methodologically heterogeneous, with no established evaluation protocols or shared benchmarks for contrastive explanations, which may discourage systematic empirical assessment. These observations are consistent with broader concerns raised about the XAI literature, where evaluation has often relied on proxy measures of interpretability rather than on user-based studies in realistic decision contexts [25, 27].

Given this validation gap, user studies should be regarded as a central and strongly recommended instrument for validating and benchmarking both the cXAI explanation techniques and human-AI interaction protocols based on it, by comparing them to either no-support scenarios or to other well-established human-AI interaction settings. Beyond measuring conventional performance indicators such as diagnostic accuracy and diagnostic confidence, user studies enable the assessment of human-centered dimensions, including understandability, perceived usefulness, cognitive support, and overall trust in the system [45]. These dimensions are particularly important in healthcare, where the effectiveness of an AI system depends not only on predictive performance but also on its ability to support clinical reasoning and facilitate appropriate reliance. A further critical aspect concerns the assessment of the interaction dynamics between clinicians and decision support systems. As discussed in Sect. 1.2.1, automation bias and algorithmic aversion are central to characterizing the human–AI relationship in clinical decision support. Quantifying them is straightforward and simply requires comparing user performance before and after exposure to AI assistance, for example in terms of diagnostic accuracy [92].

Need for comprehensive validation Despite the utmost importance of user-centered validation, general evaluation methods for explainable AI should also be adapted to characterize contrastive explanations and to compare them to their classical counterpart. Common metrics include fidelity, which measures how accurately an explanation reflects the underlying model behavior; sparsity or minimality, which quantify whether the explanation focuses on a small and interpretable set of features; and stability/robustness, which assesses whether similar inputs produce similar explanations under small perturbations. For example, fidelity can be reformulated to assess whether the explanation preserves the model's discriminative reasoning between two competing diagnoses, while sparsity can indicate whether the explanation isolates only the clinically relevant differences between them. Finally, recent clinical XAI studies stress the importance of complementing interpretability assessment with calibration and subgroup fairness evaluation [93]. One last interesting approach for validating, or at least characterizing, contrastive explanations was proposed by Jacovi et al. [25], who introduced a metric for measuring the "contrastive behavior" within an interventionist framework. Their method evaluates the extent to which an intervention changes the model's preference between a fact class P and a foil class Q. Let p and q denote the model output probabilities before and after the intervention, respectively. The contrastive behavior score is defined as:

$$\delta_{p,q}^{\text{contr}} = \frac{p_P}{p_P + p_Q} - \frac{q_P}{q_P + q_Q}, \tag{1}$$

The metric is bounded within the interval [−1,1], with larger absolute values indicating a stronger contrastive influence. Intuitively, by perturbing one given factor, the score quantifies how much that factor contributes to preferring class P over class Q; intuitively, negative values would suggest a preference toward class Q. Jacovi et al. employ this metric in two complementary ways: *(i)* to rank explanatory factors (i.e., input features, like pixels in an image) according to their usefulness in distinguishing a fixed foil, and *(ii)* to rank foil classes according to how strongly a given factor differentiates the fact (P) from alternative predictions (candidate foils). The application of this score is straightforward for intervention-based approaches, such as CEM, but it can be adapted to evaluate a multitude of use cases.

Methodological gaps

Methodological polarization toward CEM A prominent methodological pattern in the reviewed corpus is the prevalence of CEM across the included studies (7/18 papers, 39%), which is not justified through head-to-head comparisons within the corpus itself, but several factors plausibly contribute to its diffusion. First, CEM is among the earliest formalized cXAI methods [12], it is very simple to implement, and it is explicitly aligned with Miller's account of contrastive human reasoning [3], which has been influential in shaping the cXAI literature. Second, it accommodates both tabular and image inputs, the two dominant modalities in our review (16/18 studies, 89%). Third, CEM is available as ready-to-use open-source software in the Alibi Explain library [94], which lowers the barrier to adoption.

These three points are best read as plausible contributing factors rather than as evidence of superiority, nevertheless there are several observations emerging from the reviewed studies and summarized in Table 5 that may further justify the popularity of this method. First, whenever direct comparisons with alternative explainability methods were performed, CEM consistently demonstrated competitive or superior performance. In particular, both the original CEM and its model-agnostic extension (MACEM) achieved perfect success rates in generating valid pertinent positives and pertinent negatives - i.e., preserving or changing model predictions appropriately - consistently outperforming LIME. Other benchmark studies conducted outside the healthcare domain indicate that CEM performs competitively with respect to explanation plausibility and realism. For instance, Guidotti [52] reported favorable implausibility scores for CEM compared with several alternative perturbation-based methods.

Similarly, the only rigorous user-centered comparison included in the present review, conducted by Luss et al. [74], reported significantly higher perceived usefulness than both LIME and Grad-CAM, an advantage largely attributed to the additional information provided by PNs. Although these findings should not be interpreted as conclusive evidence of overall superiority, they consistently suggest that the explicit modeling of contrastive evidence represents a practical advantage over conventional attribution methods. Interestingly, similar observations were reported in other works in the cXAI literature; e.g,. Torres et al. [95] found that CEM was preferable to LIME as explanations were based on a reduced set of salient features, although no user study was conducted.

Second, the reviewed studies repeatedly reported that CEM explanations were consistent with established domain knowledge. In healthcare applications, CEM identified clinically meaningful brain regions associated with Parkinson's disease [81], plausible hematological biomarkers for ovarian cancer [83], and clinically interpretable ultrasound patterns for COVID-19 diagnosis [72]. Similar observations were reported outside medicine, where, for example, CEM recovered biologically meaningful toxicophores in a computational toxicology study [96]. Collectively, these findings suggest that CEM not only produces faithful contrastive explanations but also generates explanations that are readily interpretable from a domain-expert perspective.

Finally, unlike conventional attribution methods that only indicate which features support the current prediction, CEM consistently provides both the minimal evidence supporting the predicted class (pertinent positives) and the minimal changes required to obtain an alternative outcome (pertinent negatives). This dual perspective intersects the advantages of classical XAI feature-attribution methodologies with the advantages of counterfactual explanations. This feature was repeatedly identified as one of the principal strengths of the method and it likely contributes to its widespread adoption in healthcare-oriented contrastive explainability research [74, 83].

Collectively, these observations suggest that CEM combines conceptual simplicity, clinically interpretable explanations, and competitive explanatory performance, which may partially explain its widespread adoption in healthcare-oriented cXAI studies.

Limited coverage of data modalities Our review highlights a pronounced lack of attention to physiological or kinematic signals, such as ECG, EEG, and skeleton data, and video or volumetric data (e.g., echocardiography videos and computed tomography scans), despite their central role in clinical practice. This gap is of primary importance because many clinical AI tasks rely on temporal or 3D data, including arrhythmia detection, seizure localization, ICU deterioration prediction, and tumor characterization. However, few cXAI methods are designed for these data types, which limits contrastive reasoning to a relatively narrow part of clinical AI.

Extending cXAI to temporal and volumetric data is possible, but it brings several challenges. First, perturbation-based methods such as CEM depend on finding small but meaningful changes to the input. Although time series, videos, and volumetric scans can, in principle, be perturbed, the perturbations must preserve temporal, physiological, or anatomical plausibility. For time series, for example, prior work has shown that naive perturbations can generate out-of-distribution or clinically implausible signals [97]. Second, explanations should also be able to specify where the relevant contrast is expressed, for example in a particular lead, temporal segment, or waveform pattern [98] without altering the visualization which makes these explanations hard to

effectively display. Static heatmaps may not be enough for time-varying signals or volumetric images, which may require time-aligned attribution plots, animated explanations or paired multi-planar views [99]. The recent work on gradient-based attribution for signal and video/volume data [87, 100] shows that progress is possible, but further developing cXAI methods for temporal and volumetric data remains an important direction for cardiology, neurology, intensive care, radiology and wearable health monitoring.

Visualization and target users From this review, it is also clear that another underexplored dimension is explanation modality. Most existing cXAI approaches rely heavily on visual encodings, such as heatmaps, bar plots, or highlighted features. While often effective for expert users because they are simple, direct and aligned with established diagnostic practices, these representations can still impose a non-trivial cognitive load and may be suboptimal for lay or non-expert users [45].

The strong predominance of clinicians as target users, observed in Sect. 3.2, has direct implications for explanation design. Clinicians and patients differ in background knowledge, vocabulary and information needs, and an explanation that is effective for one group may be inappropriate for the other. Clinicians generally benefit from explanations that align with diagnostic workflows, use clinical terminology, and support comparison across differential hypotheses. Patients, in contrast, typically require explanations expressed in accessible language, contextualized within their personal situation, and framed to support shared decision-making rather than technical interpretation. The few patient-oriented studies in our review [71, 73, 79] appear consistent with this distinction, as they favor textual and heatmap-based formats chosen for their interpretability by non-expert users. The current scarcity of patient-facing cXAI works therefore represents a notable gap, particularly given the role of contrastive reasoning in supporting informed consent and shared decision-making.

Textual contrastive explanations represent a promising modality, particularly for communicating model reasoning in a concise and clinically meaningful form (e.g., "this diagnosis was favored over the alternative because feature A was present while feature B was absent"). Such textual summaries can complement visual explanations, reduce cognitive load, and facilitate integration into clinical narratives, reports, and patient-centered communication. Nevertheless, textual contrastive explanations remain surprisingly rare among the papers included in this study, despite their potential to improve accessibility and communicability, and their increasing feasibility, as cXAI outputs (e.g., contrastive features, pertinent positives/negatives, or heatmaps) can be directly translated into structured natural language using LLMs. Moreover, as already described in the XAI literature [99], future systems should adopt multimodal explanation strategies, enabling end users to flexibly access visual, textual, or combined representations depending on task complexity, expertise level, and clinical context.

Future directions for medical cXAI

Several future directions for medical cXAI emerge from the benefits highlighted in Sect. 1.2 and the gaps discussed in the preceding sections. On one hand, contrastive explanations can be further leveraged as a form of clinical metadata, enriching electronic health records, decision support systems, and patient-centered reports with structured, interpretable insights that explicitly highlight differentiating features between competing hypotheses. Equally important is the need to address the urgent methodological and practical gaps identified in this review. These include the under-representation of temporal, volumetric and multimodal data, the scarcity of textual and hybrid explanation modalities, and the lack of robust validation in real-world clinical settings. Addressing these gaps will be essential to ensure that cXAI fulfills its potential in supporting safe, reliable, and transparent clinical decision-making.

In this context, our group is currently advancing the translation of visual contrastive methods (Grad-CAM, in particular) into practice through the extension of our SignalGrad-CAM Python package [87, 100]. The updated framework now supports the generation of contrastive class activation maps for images, signals, and video/volumetric data, applicable to any convolutional-based networks, across different deep learning ecosystems (PyTorch and TensorFlow/Keras), and for both classification and regression tasks.

With the present work, we also advocate for a trustworthy, safer, and user-oriented use of contrastive explanations, which can be achieved by integrating them with state-of-the-art multimodal methodologies in artificial intelligence. For example, recent advances in Generative AI (GenAI) and Agentic AI systems, including LLM-based agents and chain-of-thought or other reasoning-oriented architectures, reflect a growing interest in making AI decision processes more understandable and communicable. Recent studies have highlighted the potential of domain-adapted GenAI systems to assist clinical decision-making, medical reporting, and multimodal data interpretation, while also emphasizing the need to address challenges such as data heterogeneity, hallucination risks, privacy preservation, workflow integration, interpretability, and user trust [101].

Furthermore, while cXAI focuses on explaining why a specific decision or action was selected over relevant alternatives, Agentic AI approaches primarily aim to

make intermediate reasoning, planning or task-execution processes more explicit: integrating these perspectives may further enhance transparency and trust in high-stakes or human-facing decision scenarios by bridging internal reasoning mechanisms with externally meaningful, contrastive explanations. In practice, an LLM-based diagnostic agent could pair its stepwise reasoning trace with a contrastive output, presenting the recommended diagnosis (e.g., pneumonia) alongside a clinically plausible foil (e.g., congestive heart failure) and the discriminating, clinical-meaningful features. Such a configuration would allow the clinician to inspect both how the agent reasoned and why one diagnosis was preferred over the alternative, rather than receiving only a final recommendation. This capability can be further enhanced through the deployment of both XAI-based (typically visual) outputs and textual explanations, enabling less experienced or non-expert users to access and interpret relevant clinical evidence more effectively.

We also aim to spur research on the comparison and benchmarking of cXAI-based paradigms against well-established or promising human–AI interaction protocols. This includes fostering the creation of a body of cXAI literature grounded in rigorous validation, supported by robust metrics that capture both system performance and user experience. Such efforts represent a promising research direction, particularly in light of the potential of cXAI to support clinical reasoning, trust calibration, and decision accountability.

As part of future work building on the findings of this systematic review and motivated by the identified need for more interactive, clinician-centered, and comparative explanation paradigms in healthcare, we plan to conduct a systematic analysis of the conceptual and practical relationships between contrastive explainability and another contrast-based explanation/decision-support tool: the Judicial AI framework [44]. Although both approaches rely on comparisons between alternative outcomes, they differ epistemically and operationally in how such comparisons are constructed and communicated, and a comparison between them would enhance the general understanding and assessment of both these innovative decision-support aids.

In parallel, we intend to investigate how contrastive explanations can support structured, side-by-side reasoning across two or more plausible alternatives, by adopting principles inspired by the Judicial AI perspective with explanations framed as "evidence" supporting two competing sides through the use of juxtaposed (in other words, "contrastive") visualizations, such as paired heatmaps. For example, in an emergency department triage scenario where the system must decide between admission to the ICU and admission to a general ward, a Judicial-cXAI framework would present, instead of a simple AI-predicted label, the following juxtaposed explanations: "Why admit to the ICU, rather than to the ward?", in the form of heatmaps over the patient's vital-sign trajectories and key laboratory values, alongside "Why admit to the ward, rather than to the ICU", grounded in the features that argue against intensive care. By exposing both sides of the comparison, this configuration conveys considerably more context-specific information than a single one-sided explanation and supports the clinician's own deliberation between the two alternatives.

Taken together, the directions outlined above share a common premise: advancing cXAI in healthcare requires moving from isolated method development toward clinically validated, user-centered frameworks that handle complex clinical data (temporal, volumetric, multimodal) and that can communicate contrasts through visual, textual, or hybrid means. Whether implemented through, or in association with, GenAI, Agentic AI, Judicial AI, or other emerging paradigms, the practical value of cXAI will ultimately depend on rigorous user studies and evaluation in realistic clinical workflows.

Limitations of this review

Several limitations of the present review should be acknowledged. First, the number of included studies is small (18), reflecting both the novelty of the field and our deliberately focused scope. As a result, the patterns described in this work should be read as descriptive observations on an emerging area of research, rather than as generalizable conclusions about cXAI in healthcare.

Second, our search strategy, although complemented by an Elicit-assisted retrieval process, may have missed relevant contributions. The wildcard expressions, such as "contrastive expla*", used across all five databases capture studies that explicitly frame their contribution under the cXAI terminology, but do not retrieve works that adopt contrastive reasoning implicitly without using proper terms. Our eligibility criteria additionally exclude unpublished clinical deployments, gray literature, and non-English publications. These limitations are consistent with known challenges in systematic reviews of rapidly emerging AI subfields, where publication bias and heterogeneous terminology can affect coverage.

Third, the methodological heterogeneity of the included studies, including different data modalities, AI models, cXAI methods, evaluation protocols, and clinical tasks, precludes a formal quantitative meta-analysis. The synthesis presented here is therefore qualitative and descriptive in nature, and direct cross-study comparisons of method effectiveness are not possible within the corpus.

Finally, this review is intentionally restricted to healthcare applications, and therefore does not examine the use of contrastive explanations in other domains. Nevertheless,

most of the cXAI methods identified in this review are not inherently domain-specific. Many were originally proposed outside healthcare; for example, Dhurandhar et al. [12] introduced the Contrastive Explanation Method using a handwritten digits-recognition tasks, while Jacovi et al. [25] developed latent-space contrastive analyses in the context of natural language processing. Furthermore, several methods are largely agnostic to the underlying data modality and predictive model, making them potentially applicable across a wide range of domains (see Table 6). For instance, Nogueira et al. [102] applied contrastive Grad-CAM to object detection in autonomous-driving systems, a field far removed from the image-classification settings in which the method was originally demonstrated, including everyday-image classification [10] and healthcare image classification [33].

As a consequence we acknowledge that the patterns identified in this review are likely to recur in other high-stakes settings where decisions are inherently comparative and not just in the context of medicine and differential diagnosis. In finance, for example, contrastive explanations could clarify why one credit or fraud decision was reached rather than a plausible alternative, an aim closely related to existing work on algorithmic recourse [17]. In education, contrastive and counterfactual explanations have been examined as a means of helping learners understand why one recommendation or assessment outcome was favored [55]. More broadly, we stress on the fact that the perturbation-based, attribution-based, and prototype-based methods summarized in Table 6 do not depend on clinical data and could in principle support comparative decision-making in sectors such as law, human resources, and quality control. We present these as possibilities rather than as systematic findings as a systematic examination of cross-domain applications would require different eligibility criteria and research questions and therefore falls outside the scope of the present review.

All the limitations above define the boundaries within which the present findings should be interpreted, and they motivate several of the future-research directions outlined in Sect. 4.2.

Conclusions

This systematic review examined how cXAI has been developed and applied in healthcare. Across the five queried databases and the complementary Elicit-assisted search, 18 studies met our eligibility criteria, suggesting that cXAI in medicine remains a small and emerging area despite its conceptual fit with differential and abductive clinical reasoning. The included studies focused mainly on tabular and image data, relied predominantly on a limited set of methods, and rarely included end-user evaluations: only 2 of the 18 studies reported user studies. As a consequence, claims about the practical clinical value of cXAI, including its potential effects on cognitive load, trust calibration, and automation bias, are supported primarily by theoretical arguments and by adjacent literature, rather than by direct evidence within the reviewed corpus.

We conclude our objective evaluation by noting that the findings of this review may benefit several stakeholder groups. Researchers may use the identified methodological gaps and future directions to guide the development and validation of new cXAI approaches: developers and Human-AI interaction designers, for example, may leverage the reported evidence to select explanation techniques that better align with clinical reasoning processes and may use contrastive explanations as a foundation for designing interaction protocols that promote critical reflection, preserve user agency, and mitigate automation bias. More broadly, clinicians may benefit from the future availability of decision-support systems capable of providing more meaningful and differential explanations, while patients and lay users may benefit from the development of more transparent, understandable, and human-like forms of AI support.

In short, realizing the potential of cXAI as a clinically aligned approach to explainability will require user-centered validation in realistic clinical settings, the extension of cXAI methods to underrepresented data modalities (particularly physiological signals and volumetric data), and the development of textual and multimodal explanation strategies tailored to clinicians and patients. Addressing these gaps is necessary before the conceptual promise of cXAI can be evaluated as a practical contribution to clinical decision support.

A realistic pathway toward the widespread adoption of cXAI in healthcare is likely to occur through a series of incremental milestones. In the short term (1–3 years), efforts should focus on improving the accessibility of cXAI methods through open-source softwares for easily applicable cXAI methods and standardized implementations, and methodological benchmarking studies. During this phase, systematic comparisons between alternative cXAI approaches, as well as comparisons with classical XAI techniques, are needed to establish best practices and identify the most suitable methods for different clinical tasks.

In this context, our research group has contributed to the dissemination of cXAI through the present systematic review and educational initiatives such as a demonstration of cXAI implementation in cardiology at the EFMI MIE2026 conference [103]. We are also developing and extending an open-source Python package for contrastive Grad-CAM explanations [87] and preparing a clinician-centered evaluation study within the AI for Heart Failure (AI4HF) project [104], aimed at comparing cXAI- and Judicial AI-based decision-support systems.

In the medium term (3–5 years), greater emphasis should be placed on in-depth user-centered validation. Large-scale studies involving clinicians, patients, and other stakeholders are needed to evaluate not only technical properties of explanations but also their effects on diagnostic accuracy, confidence calibration, trust, cognitive workload, automation bias, and overall clinical decision making. Such studies are essential to determine whether the theoretical benefits of contrastive explanations translate into measurable improvements in practice.

Finally, in the longer term (5+ years), cXAI methods should be integrated into clinical decision-support systems and evaluated in real-world clinical environments. Prospective studies assessing workflow integration, user acceptance, patient communication, and clinical outcomes will be necessary to establish the practical value of cXAI and support its routine adoption in healthcare.

Abbreviations

AI	Artificial Intelligence
CEM	Contrastive Explanation Method
CT	Computed Tomography
DB	DataBase
cXAI	contrastive eXplainable Artificial Intelligence
GenAI	Generative Artificial Intelligence
LLM	Large Language Model
ML	Machine Learning
PN	Pertinent Negatives
PP	Pertinent Positives
WoE	Weight of Evidence
XAI	eXplainable Artificial Intelligence

Acknowledgements
SP is a PhD student enrolled in the National PhD program in Artificial Intelligence, XXXIX cycle, course on Health and life sciences, organized by Università Campus Bio-Medico di Roma. This work was supported by the Italian Ministry of Research, under the complementary actions to the NRRP "Fit4MedRob - Fit for Medical Robotics" Grant (# PNC0000007). EP acknowledges funding support provided by the Italian project PRIN PNRR 2022 InXAID - Interaction with eXplainable Artificial Intelligence in (medical) Decision-making. CUP: H53D23008090001 funded by the European Union - Next Generation EU. ChatGPT-5.5 has been used for language editing and small adjustments in some of the figures presented, but all content has been reviewed and approved by all the authors.

Author contributions
S. P. and A. A. as equal first authors carried out the document selection, the systematic review process, and the writing of the manuscript. M. J. B. and F. W. A. provided extensive intellectual feedback, critically revised the manuscript for important content, and proofread the text. C. I., E. P., and K. L. supervised the process. All authors read and approved the final manuscript.

Funding
This work received funding from the European Union's Horizon Europe research and innovation program under Grant Agreement No. 101,080,430 (AI4HF project).

Data availability
No datasets were generated or analysed during the current study.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]Department of Electrical, Computer, and Biomedical Engineering, University of Pavia, Pavia 27100, Italy
[2]Artificial Intelligence in Medicine Lab (BCN-AIM), Departament de 9 10 Matemàtiques i Informática, Universitat de Barcelona, Barcelona 08007, Spain
[3]Telfer School of Management, University of Ottawa, Ottawa, Ontario K1N 6N5, Canada
[4]Department of Medical Informatics, Amsterdam UMC Location University of Amsterdam, Amsterdam 1105 AZ, The Netherlands
[5]Department of Cardiology, Amsterdam Cardiovascular Sciences, Amsterdam University Medical Centers, University of Amsterdam, Amsterdam 1105 AZ, The Netherlands
[6]Institute of Health Informatics, University College London, London NW1 2DA, United Kingdom
[7]National Institute for Health Research, University College London Hospitals, Biomedical Research Centre, University College London, London NW1 2PG, United Kingdom
[8]Institucio´ Catalana de Recerca i Estudis Avanc¸ats (ICREA), Barcelona 08010, Spain

Received: 24 February 2026 / Accepted: 3 July 2026
Published online: 15 July 2026

References

1. Gohel P, Singh P, Mohanty M. Explainable AI: current status and future directions. 2021 Jul. https://doi.org/10.48550/arXiv.2107.07045.
2. Hilton DJ. Logic and causal attribution. In: Contemporary science and natural explanation: commonsense conceptions of causality. New York, NY, US: New York University Press; 1988. p. 33–65.
3. Miller T. Explanation in artificial intelligence: insights from the social sciences. Artif Intell. 2019, Feb;267:1–38. https://doi.org/10.1016/j.artint.2018.07.007.
4. Lewis D. Causal explanation. In: Lewis D, editor. Philosophical papers, volume II. OUP Usa; 1986. p. 214–40.
5. Lipton P. Contrastive explanation. Roy I Ph S. 1990;27:247–66: Cambridge University Press. doi: https://doi.org/10.1017/s1358246100005130.
6. Miller T. Contrastive explanation: a structural-model approach. Knowl Eng Rev. 2021, Jan;36(e14). https://doi.org/10.1017/S0269888921000102.
7. Van Bouwel J, Weber E. Remote causes, bad explanations? J Theory Soc Behaviour. 2002, Nov;32(4):437–49. https://doi.org/10.1111/1468-5914.00197.
8. Stepin I, Alonso JM, Catala A, Pereira-Fariña M. A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. IEEE Access. 2021;9:11974–2001. https://doi.org/10.1109/ACCESS.2021.3051315.
9. McGill AL, Klein JG. Contrastive and counterfactual reasoning in causal judgment. J Pers Soc Psychol. 1993;64(6):897–905. US, American Psychological Association. doi: https://doi.org/10.1037/0022-3514.64.6.897.
10. AlRegib G, Prabhushankar M. Explanatory paradigms in neural networks: towards relevant and contextual explanations. IEEE Signal Process Mag. 2022;39(4):59–72. https://doi.org/10.48550/arXiv.2202.11838.
11. Prabhushankar M, Kwon G, Temel D, AlRegib G. Contrastive explanations in neural networks. 2020 IEEE International Conference on Image Processing (ICIP). 2020. p. 3289–93. ISSN: 2381-8549. Available from: https://doi.org/10.1109/ICIP40778.2020.9190927.
12. Dhurandhar A, Chen PY, Luss R, Tu CC, Ting P, Shanmugam K, et al. Explanations based on the missing: towards contrastive explanations with pertinent negatives. Adv Neural Inf Process Syst. 2018;31. https://doi.org/10.48550/arXiv.1802.07623.
13. Castelnovo A, Crupi R, Mombelli N, Nanino G, Regoli D. Evaluative item-contrastive explanations in rankings. Cognit Computation. 2024, Nov;16(6):3035–50. https://doi.org/10.1007/s12559-024-10311-2.
14. Le T, Wang S, Lee DG. Generating concise and informative contrastive sample to explain neural network Model's prediction. Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining.

KDD'20. New York, NY, USA: Association for Computing Machinery; 2020. p. 238–48. Available from: https://doi.org/10.1145/3394486.3403066.
15. Poyiadzi R, Sokol K, Santos-Rodriguez R, De Bie T, Flach P. FACE: feasible and actionable counterfactual explanations. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. AIES'20. New York, NY, USA: Association for Computing Machinery; 2020. p. 344–50. Available from: https://doi.org/10.1145/3375627.3375850.
16. Sokol K, Flach P. One Explanation Does Not Fit All. KI - Künstliche Intelligenz. 2020, Jun;34(2):235–50. https://doi.org/10.1007/s13218-020-00637-y.
17. Karimi AH, Barthe G, Schölkopf B, Valera I. A survey of algorithmic recourse: contrastive explanations and consequential recommendations. ACM Comput Surv. 2022;55(5):1–29. https://doi.org/10.1145/3527848.
18. Shang R, Feng KJK, Shah C. Why Am I not seeing it? Understanding users' needs for counterfactual explanations in everyday recommendations. Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. FAccT'22. New York, NY, USA: Association for Computing Machinery; 2022. p. 1330–40. Available from: https://doi.org/10.1145/3531146.3533189.
19. Lombrozo T. Explanation and abductive inference. Oxford Handb Thinking Reasonin. 2012, Nov.260–76. https://doi.org/10.1093/oxfordhb/9780199734689.013.0014.
20. Jain B. The key role of differential diagnosis in diagnosis. Diagnosis (Berlin, Ger). 2017, Nov;4(4):239–40. https://doi.org/10.1515/dx-2017-0005.
21. Lim B, Cahaly J, Sng C, Chew A. Diagrammatization and abduction to improve AI interpretability with domain-aligned explanations for medical diagnosis. CHI'25: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems. 2025. p. 25. Available from: https://doi.org/10.1145/3706598.3714058.
22. 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, Aug;51(5):93:1–93:42. https://doi.org/10.1145/3236009.
23. Causality PJ. 2nd. Cambridge University Press; 2009.
24. Jacovi A, Goldberg Y. Towards faithfully interpretable NLP systems: how should we define and evaluate faithfulness? Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics. 2020. p. 4198–205. Available from: https://doi.org/10.18653/v1/2020.acl-main.386.
25. Jacovi A, Swayamdipta S, Ravfogel S, Elazar Y, Choi Y, Goldberg Y. In: Moens MF, Huang X, Specia L, Swt Y, editors. Contrastive explanations for model interpretability. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana. Dominican Republic: Association for Computational Linguistics; 2021. p. 1597–611. Available from: https://doi.org/10.18653/v1/2021.emnlp-main.120.
26. Gouveia SS, Malík J. Crossing the trust gap in medical AI: building an abductive bridge for xAI. Philos Technol. 2024, Aug;37(3):105. https://doi.org/10.1007/s13347-024-00790-4.
27. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit Health. 2021, Nov;3(11):e745–50. https://doi.org/10.1016/S2589-7500(21)00208-9.
28. Lyell D, Coiera E. Automation bias and verification complexity: a systematic review. J Am Med Inf Assoc: JAMIA. 2017, Mar;24(2):423–31. https://doi.org/10.1093/jamia/ocw105.
29. EDPS. EDPS, editor. TechDispatch #2/2023 - explainable artificial intelligence | European data protection Supervisor. Accessed 9 Dec 2025. Available from: https://www.edps.europa.eu/data-protection/our-work/publications/techdispatch/2023-11-16-techdispatch-22023-explainable-artificial-intelligence.
30. COE. VI. Recommendations for common ethical standards for trustworthy AI - Human rights and biomedicine - COE Accessed 9 December 2025. Available from: www.coe.int. https://www.coe.int/en/web/human-rights-and-biomedicine/recommendations-for-common-ethical-standards-for-trustworthy-ai.
31. Alvarez Melis D, Kaur H, Daumé Iii H, Wallach H, Wortman Vaughan J. From human explanation to model interpretability: a framework based on Weight of evidence. Proceedings of the AAAI Conference on Human Computation and Crowdsourcing. vol. 9. 2021. p. 35–47. ISSN: 2769-1349, 2769-1330. Available from: https://doi.org/10.1609/hcomp.v9i1.18938.
32. Jacobs M, Pradier MF, McCoy T, Perlis R, Doshi Velez F, Gajos K. How machine-learning recommendations influence clinician treatment selections: the example of the antidepressant selection. Transl Psychiatry. 2021, Feb, 11;11(1). https://doi.org/10.1038/s41398-021-01224-x.
33. Prabhushankar M, Alregib G. Extracting causal visual features for limited label classification. Conference: 2021 IEEE International Conference on Image Processing (ICIP). 2021. p. 3701. Available from: https://doi.org/10.1109/ICIP42928.2021.9506393.
34. Dietvorst BJ, Simmons JP, Massey C. Algorithm aversion: people erroneously avoid algorithms after seeing them err. J Exp Psychol Gen. 2015, Feb;144(1):114–26. https://doi.org/10.1037/xge0000033.
35. Sun Y, Sheng D, Zhou Z, Wu Y. AI hallucination: towards a comprehensive classification of distorted information in artificial intelligence-generated content. Humanit Soc Sci Commun. 2024;11(1):1–14. https://doi.org/10.1057/s41599-024-03811-x.
36. Parasuraman R, Manzey DH. Complacency and bias in human use of automation: an attentional integration. Hum Factors. 2010, Jun;52(3):381–410. https://doi.org/10.1177/0018720810376055.
37. Moore JW. What is the sense of agency and Why does it matter? Front Psychol. 2016, Aug;7:1272. https://doi.org/10.3389/fpsyg.2016.01272.
38. Aquino YSJ, Rogers WA, Braunack-Mayer A, Frazer H, Win KT, Houssami N, et al. Utopia versus dystopia: professional perspectives on the impact of healthcare artificial intelligence on clinical roles and skills. Int J Med Inf. 2023, Jan;169:104903. https://doi.org/10.1016/j.ijmedinf.2022.104903.
39. Benítez TM, Xu Y, Boudreau JD, Kow AWC, Bello F, Van Phuoc L, et al. Harnessing the potential of large language models in medical education: promise and pitfalls. J Am Med Inf Assoc: JAMIA. 2024, Feb;31(3):776–83. https://doi.org/10.1093/jamia/ocad252.
40. Sureau C. Medical deresponsibilization. J Assisted Reproduct Genet. 1995, Sep;12(8):552–58. https://doi.org/10.1007/BF02212919.
41. Grote T, Berens P. How competitors become collaborators—bridging the gap(s) between machine learning algorithms and clinicians. Bioethics. 2022;36(2):134–42. https://doi.org/10.1111/bioe.12957.
42. Lekadir K, Frangi AF, Porras AR, Glocker B, Cintas C, Langlotz CP, et al. FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. BMJ. 2025, Feb;388:e081554. British Medical Journal Publishing Group Section: Research Methods & Reporting. https://doi.org/10.1136/bmj-2024-081554.
43. Cabitza F, Natali C, Famiglini L, Campagner A, Caccavella V, Gallazzi E. Never tell me the odds: investigating pro-hoc explanations in medical decision making. Artif Intell Med. 2024, Apr;150:102819. https://doi.org/10.1016/j.artmed.2024.102819.
44. Cabitza F, Famiglini L, Fregosi C, Pe S, Parimbelli E, La Maida GA, et al. From oracular to Judicial: enhancing clinical decision making through contrasting explanations and a novel interaction protocol. Proceedings of the 30th International Conference on Intelligent User Interfaces. 2025. p. 745–54. Available from: https://doi.org/10.1145/3708359.3712157.
45. Buçinca Z, Swaroop S, Paluch AE, Doshi-Velez F, Gajos KZ. Contrastive explanations that anticipate human misconceptions can improve human decision-making skills. Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems. 2025. p. 1–25. Available from: https://doi.org/10.1145/3706598.3713229.
46. Cabitza F, Parimbelli E. Let XAI generate reliability metadata, not medical explanations. Comput Methods Programs Biomed. 2026, Jan;273:109090. https://doi.org/10.1016/j.cmpb.2025.109090.
47. Peirce CS. Harvard lectures on pragmatism: lecture II. In: Collected papers of Charles Sanders Peirce. Vol. 5. Hartshorne, Charles and Weiss, Paul; 1903. p. 64. Available from: https://www.scirp.org/reference/referencespapers?referenceid=2899941.
48. Gilpin LH, Bau D, Yuan BZ, Bajwa A, Specter M, Kagal L. Explaining explanations: an overview of interpretability of machine learning. 2018 IEEE 5th International Conference on data science and advanced analytics (DSAA). IEEE; 2018. p. 80–89. Available from: https://doi.org/10.1109/DSAA.2018.00018.
49. EU. Regulation (EU) 2016/679 of the European Parliament and of the council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/EC (General data protection regulation) (text with EEA relevance). 2016. http://data.europa.eu/eli/reg/2016/679/oj. EU. Accessed 10 December 2025. Available from.
50. Chou YL, Moreira C, Bruza P, Ouyang C, Jorge J. Counterfactuals and causability in explainable artificial intelligence: theory, algorithms, and applications. arXiv. ArXiv:2103.04244 [cs.AI]. 2021 Available from: http://arxiv.org/abs/2103.04244.
51. Zhang X, Dai L, Peng Q, Tang R, Li X. A survey of counterfactual explanations: definition, evaluation, algorithms, and applications. In: Xiong N, Li M, Li K, Xiao Z, Liao L, Wang L, editors. Advances in natural computation, fuzzy systems and knowledge discovery. Cham: Springer International Publishing; 2023. p. 905–12.

52. Guidotti R. Counterfactual explanations and how to find them: literature review and benchmarking. Data Min Knowl Discov. 2024;38(5):2770–824. https://doi.org/10.1007/s10618-022-00831-6.
53. Verma S, Boonsanong V, Hoang M, Hines K, Dickerson J, Shah C. Counterfactual explanations and algorithmic recourses for machine learning: a review. ACM Comput Surv. 2024, Oct;56(12):1–42. https://doi.org/10.1145/3677119.
54. Jiang J, Leofante F, Rago A, Toni F. Robust counterfactual explanations in machine learning: a survey. IJCAI'24: Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence. 2024. https://doi.org/10.24963/ijcai.2024/894.
55. Buñay-Guisñan P, Lara JA, Romero C. Counterfactual explanations in education: a systematic review. WIREs Data Min Knowl Discov. 2026;16(1):e70060 E70060 DMKD-00842.R1. https://doi.org/10.1002/widm.70060. https://wires.onlinelibrary.wiley.com/doi/pdf/10.1002/widm.70060.
56. Kitchenham B, Charters S. Guidelines for performing systematic literature reviews in software engineering. Elsevier; 2007 Jan 2.
57. Elicit. Elicit: the AI research Assistant. Elicit. Accessed 10 Dec 2025. Available from: https://elicit.com/.
58. PRISMA. PRISMA statement. PRISMA. Accessed 10 Dec 2025. Available from: https://www.prisma-statement.org/.
59. arXiv. arXiv.Org e-print archive. arXiv. Accessed Dec 10 2025. Available from: https://arxiv.org/.
60. IEEE. IEEE Xplore. IEEE. Accessed 10 December 2025. Available from: https://ieeexplore.ieee.org/Xplore/guesthome.jsp.
61. Scopus. Scopus preview - Scopus - welcome to Scopus. Scopus Accessed 10 December 2025. Available from: https://www.scopus.com/home.uri?zone=header%26origin=.
62. PubMed. PubMed. Accessed 10 December 2025. Available from: https://pubmed.ncbi.nlm.nih.gov/.
63. ACM. ACM digital library. Accessed 10 December 2025. Available from: https://dl.acm.org/.
64. Springer. SpringerLink. SpringerLink. Accessed 10 December 2025. Available from: https://link.springer.com/.
65. Donoso-Guzmán I, Kacafírková KS, Szymanski M, Jacobs A, Parra D, Verbert K. A systematic review of user-centred evaluation of explainable AI in healthcare. arXiv Prepr arXiv: 250613904. 2025. https://doi.org/10.48550/arXiv.2506.13904.
66. Liu Y, Liu C, Zheng J, Xu C, Wang D. Improving explainability and integrability of medical AI to promote health care professional acceptance and use: mixed systematic review. J Med Internet Res. 2025;27:e73374. https://doi.org/10.2196/73374.
67. Kung J. Elicit (product review). J Can Health Libraries Assoc / J de l'Assoc des bibliothèques de la santé du Canada. 2023, Apr;44(1). https://doi.org/10.29173/jchla29657.
68. OECD. Artificial intelligence in science: challenges, opportunities and the future of research. Paris: OECD Publishing; 2023. Available from: https://doi.org/10.1787/a8d820bd-en.
69. Bernard N, Sagawa JY, Bier N, Lihoreau T, Pazart L, Tannou T. Using artificial intelligence for systematic review: the example of elicit. BMC Med Res Methodol. 2025;25(1):75. https://doi.org/10.1186/s12874-025-02528-y.
70. Haddaway NR, Page MJ, Pritchard CC, McGuinness LA. PRISMA2020: an R package and shiny app for producing PRISMA 2020-compliant flow diagrams, with interactivity for optimised digital transparency and open synthesis. Campbell Systematic Rev. 2022;18(2):e1230. https://doi.org/10.1002/cl2.1230.
71. Neerincx MA, van der Waa J, Kaptein F, van Diggelen J. Using perceptual and cognitive explanations for enhanced human-agent team performance. Engineering Psychology and Cognitive Ergonomics: 15th International Conference, EPCE 2018, Held as Part of HCI International 2018, Las Vegas, NV, USA, July 15-20, 2018, Proceedings. Berlin, Heidelberg: Springer-Verlag; 2018. p. 204–14. Available from: https://doi.org/10.1007/978-3-319-91122-9_18.
72. Roberts J, Tsiligkaridis T. Ultrasound diagnosis of COVID-19: robustness and explainability. arXiv preprint arXiv:201201145. 2020. https://doi.org/10.48550/arXiv.2012.01145.
73. Banerjee S, Lio P, Jones PB, Cardinal RN. A class-contrastive human-interpretable machine learning approach to predict mortality in severe mental illness. npj Schizophr. 2021;7(1):60. https://doi.org/10.1038/s41537-021-00191-y.
74. Luss R, Chen PY, Dhurandhar A, Sattigeri P, Zhang Y, Shanmugam K, et al. Leveraging latent features for local explanations. Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. KDD'21. New York, NY, USA: Association for Computing Machinery; 2021. p. 1139–49. Available from: https://doi.org/10.1145/3447548.3467265.
75. Mishra S, Tripathy HK, Kumar Thakkar H, Garg D, Kotecha K, Pandya S. An explainable intelligence driven query prioritization using balanced decision tree approach for multi-level psychological disorders assessment. Front Public Health. 2021;9:795007. https://doi.org/10.3389/fpubh.2021.795007.
76. Finzel B, Kuhn SP, Tafler DE, Schmid U. Explaining with attribute-based and relational near misses: an interpretable approach to distinguishing facial expressions of pain and disgust. Inductive Logic Programming: 31st International Conference, ILP 2022, Windsor Great Park, UK, September 28–30, 2022, Proceedings. Berlin, Heidelberg: Springer-Verlag; 2022. p. 40–51. Available from: https://doi.org/10.1007/978-3-031-55630-2_4.
77. Jacobsen FL, Hicks SA, Halvorsen P, Riegler MA. Estimating predictive uncertainty in gastrointestinal polyp segmentation. 2022 IEEE 35th International Symposium on Computer-Based Medical Systems (CBMS). 2022. p. 44–49. ISSN: 2372-9198. Available from: https://ieeexplore.ieee.org/document/9867012.
78. Tahirovic E, Krivic S. Interpretability and explainability of logistic regression model for breast cancer detection. Icaart (3). 2023;161–68. Available from: https://doi.org/10.5220/0011627600003393.
79. Banerjee S. Generating complex explanations for artificial intelligence models: an application to clinical data on severe mental illness. Life. 2024;14(7):807. https://doi.org/10.3390/life14070807.
80. Dhurandhar A, Pedapati T, Balakrishnan A, Chen PY, Shanmugam K, Puri R. Model agnostic contrastive explanations for classification models. IEEE J Emerg Sel Top Circuits Syst. 2024, Dec;14(4):789–98. https://doi.org/10.1109/JETCAS.2024.3486114.
81. Balikci Cicek I, Kucukakcali Z, Deniz B, Algül FE. An explainable approach to Parkinson's diagnosis using the contrastive explanation method-CEM. Diagn (Basel, Switz). 2025, Aug;15(16):2069. https://doi.org/10.3390/diagnostics15162069.
82. Fahad M, Hassan S, Cheikh FA, Ullah M. Enhancing interpretability in gastrointestinal disease detection: a comparative analysis of deep learning architectures and explainable AI. In: AI Machot F, Horsch MT, Scholze S, editors. Designing the conceptual landscape for a XAIR validation infrastructure. Cham: Springer Nature Switzerland; 2025. p. 133–48. Available from: https://doi.org/10.1007/978-3-031-89274-5_10.
83. Kucukakcali Z, Cicek IB, Akbulut S. Explainable machine learning for the early clinical detection of ovarian cancer using contrastive explanations. J Clin Med. 2025, Sep;14(17):6201. https://doi.org/10.3390/jcm14176201.
84. Labaien J, Zugasti E, De Carlos X. Contrastive explanations for a deep learning model on time-series data. International Conference on Big Data Analytics and Knowledge Discovery. Springer; 2020. p. 235–44. Available from: https://doi.org/10.1007/978-3-030-59065-9_19.
85. Yin K, Neubig G. Interpreting language models with contrastive explanations. Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Abu Dhabi. United Arab Emirates: Association for Computational Linguistics; 2022. p. 184–98. Available from: https://doi.org/10.18653/v1/2022.emnlp-main.14.
86. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. Int J Comput Vision. 2020, Feb;128(2):336–59. https://doi.org/10.1007/s11263-019-01228-7.
87. Pe S. bmi-labmedinfo/signal_grad_cam. BMI "Mario Stefanelli" lab - UNIPV. Accessed 21 January 2026. Available from: https://github.com/bmi-labmedinfo/signal_grad_cam.
88. Good I. Weight of evidence: a brief survey. Bayesian Stat. 1985;2:249–70.
89. Winston PH. Learning structural descriptions from examples. 1970. Available from: http://hdl.handle.net/1721.1/6884.
90. Ribeiro M, Singh S, Guestrin C. "Why should I trust you?": explaining the predictions of any classifier. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations. San Diego, California: Association for Computational Linguistics; 2016. p. 97–101. Available from: https://doi.org/10.18653/v1/N16-3020.
91. Shafighi Shahri E, Ehsasatvatan A, Salavatizadeh M, Talaeepur S. Artificial intelligence applications in pediatric oncology imaging: a systematic review of diagnostic, prognostic, and therapeutic evaluation tools. Infosci Trends. 2025;2(4):36–47. https://doi.org/10.61186/ist.202502.03.04.
92. Cabitza F, Campagner A, Angius R, Natali C, Reverberi C. AI shall have No dominion: on how to measure technology dominance in AI-supported human decision-making. Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. CHI'23. New York, NY, USA: Association for Computing Machinery; 2023. p. 1–20. Available from: https://doi.org/10.1145/3544548.3581095.

93. Alisoufi M, Dehvari S, Heidari S, Mohebbi N, Jafarpour Sadegh A, Eshragh Abad Shapori P. A cost-aware and calibrated explainable AI framework for equitable coronary artery disease risk stratification using minimal clinical data. Infosci Trends. 2026;3(11):80–104. https://doi.org/10.66224/ist.202603.11.06.
94. Seldon Technologies. Contrastive explanation method | alibi explain. Accessed: May 03, 2026. https://docs.seldon.ai/alibi-explain/explanations/examples/contrastive-explanation-method.
95. Del Castillo Torres G, Roig-Maimó MF, Mascaró-Oliver M, Amengual-Alcover E, Mas-Sansó R. Understanding how CNNs recognize facial expressions: a case study with LIME and CEM. Sensors. 2023, Jan;23(1):131. https://doi.org/10.3390/s23010131.
96. Lim KW, Sharma B, Das P, Chenthamarakshan V, Dordick JS. Explaining chemical toxicity using missing features. ICML Workshop on Extending Explainable AI Beyond Deep Models and Classifiers. 2020.
97. Delaney E, Greene D, Keane MT. Instance-based counterfactual explanations for time series classification. International conference on case-based reasoning. Springer; 2021. p. 32–47. Available from: https://doi.org/10.1007/978-3-030-86957-1_3.
98. Liu Z, Zhang Y, Wang T, Wang Z, Luo D, Du M, et al. Explaining time series via contrastive and locally sparse perturbations. arXiv preprint arXiv:240108552. 2024. https://doi.org/10.48550/arXiv.2401.08552.
99. Kayser M, Menzat B, Emde C, Bercean B, Novak A, Espinosa A, et al.: Fool me Once? Contrasting textual and visual explanations in a clinical decision-support setting. 2024 arXiv. Available from: http://arxiv.org/abs/2410.12284.
100. Pe S, Buonocore TM, Giovanna N, Enea PS-C. Beyond image explanation. Joint Proceedings of the xAI 2025 Late-breaking Work, Demos and Doctoral Consortium co-located with the 3rd World Conference on eXplainable Artificial Intelligence (xAI 2025), Istanbul, Turkey, July 9-11, 2025. vol. 4017. 2025. p. 209–16. Available from: https://ceur-ws.org/Vol-4017/paper_27.pdf.
101. Yang W, Lu Y, Yeom S, Herbert D. Adapting GenAI strategies: understanding models, aims, and challenges in different targeted data and domains. IEEE Access. 2025;13:185181–217. https://doi.org/10.1109/ACCESS.2025.3622002.
102. Nogueira C, Fernandes L, Fernandes JND, Cardoso JS. Explaining bounding boxes in deep object detectors using post hoc methods for autonomous driving systems. Sensors (Basel, Switz). 2024, Jan;24(2):516. https://doi.org/10.3390/s24020516.
103. MIE. EFMI MIE. 2026. Available from: https://mie2026.efmi.org/.
104. AI4HF. AI4HF. 2026. Available from: https://www.ai4hf.com.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.