

Human-Centered Pathways to Trustworthy AI in Healthcare: A Comparative Analysis of Explainable AI, Human-in-the-Loop, Hybrid AI, and Uncertainty Quantification Techniques

Ali Kohan

Fasa University <https://orcid.org/0009-0007-7691-9629>

Junjie Xu

East China Normal University

Luwei Xiao

National University of Singapore

Xingjiao Wu

East China Normal University

Ashima Kukkar

Chitkara University

Sadiq Hussain

Dibrugarh University

Mohamad Roshanzamir

Fasa University

Roohallah Alizadehsani

r.alizadehsani@deakin.edu.au

Deakin University

U. Rajendra Acharya

University of Southern Queensland

Systematic Review

Keywords: Artificial Intelligence in Medicine, Clinical Decision Support Systems, AI Interpretability, AI Ethics, Responsible AI, Neuro-Symbolic Integration

Posted Date: February 27th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-8976235/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Abstract

Despite its transformative potential in healthcare, the adoption of artificial intelligence (AI) in clinical practice remains constrained by a persistent trust deficit among clinicians and patients. To address this, we conducted a systematic comparative review of 112 peer-reviewed studies published between 2015 and 2025, following the PRISMA guidelines for study selection. Articles were sourced from major scientific databases, focusing on methodological innovations and clinical evaluations to enhance AI trustworthiness. Using a novel Composite Human-Centered Trustworthiness Score (HCTS), we systematically evaluated and compared the contributions of relevant studies. Our analysis identified four human-centered pathways: explainable AI (XAI), comprising intrinsic interpretable models and post-hoc techniques (e.g., SHAP, LIME) to support error analysis and stakeholder communication; human-in-the-loop (HITL) frameworks that leverage clinician expertise via active learning and interactive visualization to improve model reliability and usability; hybrid neuro-symbolic architectures that integrate symbolic reasoning with deep learning to achieve robustness in complex or data-sparse settings; and uncertainty quantification (UQ) methods (e.g., Bayesian inference, Monte Carlo dropout, and ensemble techniques) that provide confidence estimates that are critical for high-stakes clinical decisions. We found that integrated strategies, including XAI-driven HITL loops and XAI + UQ frameworks, yield the greatest gains in transparency, human oversight, and computational capability. Addressing technical challenges (data heterogeneity, system interoperability), ethical and regulatory imperatives (fairness, accountability), and advancing multimodal and continual-learning paradigms are essential for ensuring the safe, transparent, and sustainable deployment of AI in clinical practice.

1. Introduction

The introduction of Artificial Intelligence (AI) into healthcare marks a transformative era, unlocking new possibilities in diagnosis, prognosis, treatment optimization, and healthcare system management [1]. Whether it is deep learning algorithms that perform radiological scans with a high level of accuracy or predictive models that intercept the earliest signs of a disease, AI has already shown the potential to transform clinical processes and enhance patient outcomes [2]. Nevertheless, AI in healthcare is currently suffering a basic trust gap, even as healthcare facilities increase its reliance and its technical prowess advances. Clinical practice settings are dynamic, unpredictable and the stakes are high; any choice will have a significant impact on human life and dignity [3]. Accuracy is not enough in such an environment. Instead, trustworthy is the missing piece: a complex of transparency, accountability, robustness, and contextual sensitivity that makes AI systems serve, not stand in place of, human care [4].

The transition to an anthropocentric emphasis on trustworthy AI addresses an important fact: healthcare AI tools are not isolated systems; they exist within socio-technical environments encompassing clinicians, patients, institutions, and regulatory organizations. Thus, responsible AI should not only help to make the right prediction, but also inform about the process of making it; how confident the system is; in what ways human operators could disagree with its results, or suggest some contextual ideas [5]. This involves a foundational shift, departing from the conventional algorithm-focused design and adopting frameworks that explicitly model the human user in the learning loop, uncertainty, and consider both data-driven learning and domain knowledge [6, 7].

To operationalize this paradigm, four converging pathways have emerged as critical enablers, as shown in Fig. 1.

- **Explainable AI (XAI):** Deep neural networks (DNN) that form the basis of AI models can be viewed as black box models - in other words, they are tricky and challenging to comprehend or audit. XAI aims to achieve this by enhancing transparency by interpreting the models' decision-making logic in a manner understandable to end users. In medicine, it could include emphasizing which symptoms, characteristics, or regions of an image most influenced a diagnosis. XAI is necessary to establish trust in the clinician and make it possible to have actual validation and regulatory approval [8–12].
- **Human-in-the-Loop (HITL):** Systems of AI operated without human supervision can introduce a conflict between AI values and clinical decisions or ethics. The HITL frameworks elicit human knowledge at significant points of the AI

lifecycle, including data labelling, model training, instant validation, in decision feedback points. Such a collaborative dynamic generates repeated learning, constant enhancement of the models, and greater accountability, especially in dynamic or ambiguous situations [13–16].

- **Hybrid AI:** In data-driven machine learning, although it is good at learning patterns, it is usually not endowed with the ability of logic when faced with sparse data, unlikely situations, or built-in ethics. This has been resolved in hybrid AI systems by combining symbolic AI (e.g., ontologies, rules, knowledge graphs) with statistical models. In healthcare, these systems can potentially integrate the ability to cope with guidelines and empirical learning in an optimized manner, as they inherently provide the greatest strengths of interpretability and flexibility [17, 18].
- **Uncertainty Quantification (UQ):** AI systems tend to be overconfident, particularly when they are wrong, which can be problematic in clinical practice. The goal of UQ techniques is to estimate the model's confidence in its outputs by raising predictions that fall outside the possible range of the training distribution or contain ambiguous inputs. This allows clinicians to view AI outputs as probabilistic recommendations, not deterministic conclusions, enhancing shared decision-making and risk evaluation [19, 20].

Although people have focused on these pathways in isolation, achieving an inclusive convergence of these pathways is critical to developing a distinctively trustworthy AI technology in healthcare. The proposed pathways focus on distinct aspects of trustworthiness: interpretability with the XAI paradigm, adaptability and control with HITL, contextual reasoning with Hybrid AI, and safety in high-uncertainty scenarios with UQ. Combined thoughtfully, they establish a comprehensive philosophy of human-centered AI design.

This conceptual and practical gap has motivated the present paper to provide a broader analysis and synthesis of these four pathways, into a coherent system of trustworthy AI in healthcare. In particular, this paper will address the following objectives:

- Map the landscape of current applications and research in XAI, HITL, Hybrid AI, and UQ within the healthcare domain;
- Analyze how these approaches contribute individually and collectively to the goals of interpretability, safety, and clinician collaboration;
- Propose guidelines for choosing among these methods in particular scenarios;
- Highlight open challenges, including data limitations, regulatory constraints, usability barriers, and ethical tensions;
- Outline directions for future research, especially in developing multimodal, continuously learning, and policy-aligned systems.

To address these goals, we conducted a structured literature review using methodological inclusion and exclusion criteria, emphasizing empirical research and field implementations. The rest of the paper is organized as follows: Section 2 presents the methodology used to select and analyze the studies. Other sections 3–6 in detail explain the individual approaches comprising definitions, technical approaches, case studies and constraints. Section 7 provides a comparative analysis and identifies areas of convergence and synergy. Section 8 discusses the general issues and directions that must be addressed to advance the agenda of human-centered AI trustworthiness. Lastly, Section 9 summarizes the paper, presenting the main findings and recommendations for research, policy, and practice.

In a world where artificial intelligence is becoming deeply enmeshed in life-and-death decisions, developing systems that are reliable not just in conception but also in interaction and interpretation is not a luxury but a necessity. This work is intended to offer a clear, specific and interdisciplinary view on the development of engineering AI that is respectful of, supportive of, and augmentative of the human aspect of healthcare.

1.1 Comparative Scope and Contribution

Table 1 positions our review within the evolving landscape of trustworthy AI literature by contrasting its scope, methodological approach, and thematic coverage with those of seven of the most relevant prior works. These reviews were

selected using the keywords outlined in Section 2.1 to identify existing review papers on these topics. The majority of prior systematic reviews ([9], [21], [22]) rightly identify XAI as a primary factor in building trust, but largely treat HITL and Hybrid AI with only secondary or implicit coverage, focusing instead on psychological aspects of clinician trust rather than on the technical and procedural design of human oversight or the integration of human expertise [22]. While several review papers address UQ in healthcare [19, 23–25], few analyze or compare it with other trustworthiness mechanisms. Moreover, most existing reviews lack a healthcare-specific focus. Collectively, these limitations reveal a significant gap: no single comparative analysis integrates all four essential human-centered pillars—XAI, HITL, Hybrid AI, and UQ—within a unified, cross-paradigm framework explicitly tailored to the high-stakes environment of healthcare. This review moves beyond fragmented analyses toward an integrative roadmap for designing, evaluating, and deploying trustworthy AI systems that are not only technically sound but also meaningfully aligned with human needs, clinical workflows, and ethical values in medicine.

Table 1

Comparison of prior reviews in the field and the unique positioning of the current review. ✓ = primary focus; △ = secondary or partial coverage; ☒ = not addressed.

Ref	Year	Title	XAI	HITL	Hybrid AI	UQ	Methodology	Limitation
[9]	2021	A Systematic Review of Human–Computer Interaction and Explainable Artificial Intelligence in Healthcare with Artificial Intelligence Techniques	✓	△	☒	☒	Systematic Review	Focuses primarily on XAI and HCI in healthcare; does not comparatively analyze HITL, Hybrid AI, or UQ as distinct trust-building pathways.
[26]	2023	A Review of Trustworthy and Explainable Artificial Intelligence (XAI)	✓	△	☒	△	Narrative Review	Broad overview of trustworthy AI components; lacks granular comparative analysis of human-centered pathways or hybrid integrations; not specific to healthcare.
[21]	2023	Towards Risk-Free Trustworthy Artificial Intelligence: Significance and Requirements	✓	△	☒	☒	Systematic Review	Comprehensive coverage of trustworthy AI requirements (e.g., explainability, fairness, privacy) but does not comparatively analyze HITL, Hybrid AI, or UQ as human-centered pathways in healthcare; lacks structured cross-paradigm comparison.
[27]	2025	Toward Trustworthy Artificial Intelligence (TAI) in the Context of Explainability and Robustness	✓	△	☒	△	Narrative Review	Broad scope not specific to healthcare; lacks focus on Hybrid AI and detailed HITL.
[22]	2025	Trust in Artificial Intelligence–Based Clinical Decision Support Systems Among Health Care Workers: Systematic Review	✓	△	☒	☒	Systematic Review	Focuses on clinician trust factors (e.g., transparency, usability) but does not comparatively analyze HITL, Hybrid AI, or UQ as integrated technical pathways to trustworthy AI; lacks structured cross-paradigm analysis.
[28]	2025	A Roadmap Toward Neurosymbolic Approaches in AI Design	△	☒	✓	☒	Systematic Review	Lacks specific focus on healthcare workflows and no coverage of HITL or UQ.
[25]	2025	Explainability and uncertainty: Two sides of the same coin for enhancing the interpretability of deep learning models in healthcare	✓	☒	☒	✓	Position Paper	Focuses exclusively on the integration of XAI and UQ in healthcare deep learning models; lacks a comparative analysis of all four human-centered trust pathways.
Current Review	2026	Human-Centered Pathways to Trustworthy AI in Healthcare: A	✓	✓	✓	✓	Systematic Review / Comparative Analysis	Focused on four specific Trustworthy AI pathways; intentionally excludes other trustworthy pillars

Ref	Year	Title	XAI	HITL	Hybrid AI	UQ	Methodology	Limitation
		Comparative Analysis of Explainable AI (XAI), Human-in-the-Loop, Hybrid AI, and uncertainty quantification						(e.g., fairness, privacy, security).

2. Methodology

The systematic review method was used to identify, appraise, and synthesize literature on Explainable AI (XAI), Human-in-the-Loop (HITL), Hybrid AI, and Uncertainty Quantification (UQ) integration in healthcare AI. The aim was to identify trends, methods, gaps, and synergies that enable the creation of trustworthy, human-friendly AI tools in healthcare.

2.1 Search Strategy

The literature review was conducted using primary scientific databases, including IEEE Xplore, PubMed, Scopus, Web of Science, ACM Digital Library, and Google Scholar. Several Boolean combinations of the following search terms were applied:

- "Explainable AI" OR "XAI" OR "Interpretability"
- "Human-in-the-Loop" OR "HITL"
- "Hybrid AI" OR "Neurosymbolic"
- "Uncertainty Quantification" OR "UQ"
- "Healthcare" OR "Clinical decision support" OR "Medical AI"
- "Cardiovascular" OR "Neurological" OR "Oncology" OR "Critical Care" OR "Cancer"
- "Trustworthy AI" OR "Reliable AI" OR "Transparent AI"

Only peer-reviewed articles published in English were searched, and they needed to be relevant to the healthcare practice of AI.

2.2 Inclusion and Exclusion Criteria

The inclusion and exclusion criteria used to ensure the rigor and relevance of the studies are summarized in Table 2, with the corresponding flowchart shown in Fig. 2.

Table 2
Inclusion and exclusion criteria

Inclusion Criteria	IC1	Articles that describe empirical or methodological contributions related to XAI, HITL, Hybrid AI, or UQ in healthcare evaluated using the trustworthy score described in section 2.4.
	IC2	Research versions that either provide technology assessment results (e.g., performance reporting, model fidelity) or human data (e.g., clinician trust, interpretability).
	IC3	Studies using AI in actual clinical practice or verified datasets.
Exclusion Criteria	EC1	The scoping reviews were concentrated on non-clinical AI.
	EC2	Theoretical articles that are not implemented, simulated, or validated.
	EC3	Duplications of records, editorial, whitepapers, or unreviewed material.

2.3 Study Selection Process

Several research articles were obtained from different research databases. The next step was to identify the relevant articles to ensure an efficient, focused review. Accordingly, articles related to XAI, HITL, Hybrid AI, or UQ in healthcare were considered. Figure 4 shows the publication trends for the four pathways. The selection of the studies was conducted according to the PRISMA (Preferred Reporting Items of Systematic Reviews and Meta-Analyses) guidelines, as presented in Fig. 3:

- Identification: A total of 2,725 articles were identified through initial database searches.
- Screening: After removing duplicates and non-English articles, 1,975 records remained.
- Eligibility: Titles and abstracts were screened, reducing the pool to 234 potentially relevant papers.
- Inclusion: After full-text review, 112 studies were selected for final inclusion based on the trustworthy score described in section 2.4.

2.4 Composite Human-Centered Trustworthy Score (HCTS)

A qualitative thematic synthesis method was used, involving both categorization and mapping, used to compare quantitatively:

- The studies were categorized using the coded focus areas (XAI, HITL, Hybrid AI, UQ).
- Methods were additionally sub-categorized by clinical domain (e.g., imaging, diagnostics, risk scoring).
- Human-centered trustworthiness was scored in terms of:
 1. Transparency (T)
 2. Interaction level (I)
 3. Contextual reasoning (C)
 4. Uncertainty management (U)
- A Composite Human-Centered Trustworthy Score (HCTS) would be defined as:

$$HCTS = \frac{T + I + C + U}{4}$$

Where:

- $T \in [0,1]$: Degree of Explainability
- $I \in [0,1]$: Level of human integration
- $C \in [0,1]$: Hybrid reasoning capability
- $U \in [0,1]$: Ability to quantify and communicate uncertainty

This score enabled comparison of studies based on the extent to which they address the four pillars of trustworthy AI.

2.4.1 Scoring System

A criterion-based guideline with explicit anchor points (Table 3) was developed through consensus workshops with two domain experts (one clinician and one AI researcher). Each dimension (T, I, C, U) was scored on a 4-point scale [0, 0.33, 0.67, 1.0]. Three independent raters (biomedical informaticians with expertise in trustworthy AI) received standardized training from the domain experts and independently scored all 234 studies using the rubric instructions in Table 3. Final scores represent consensus after resolving discrepancies (> 0.33 difference) with a senior reviewer. To ensure the review focused on studies that substantively addressed the core principles of trustworthy AI, an eligibility threshold was applied: only papers with at least one dimension score ≥ 0.67 or an average HCTS of ≥ 0.33 were included in the analysis resulting in 112 papers.

Table 3
HCTS Scoring Rubric with Anchor Examples

Dimension	Score 0 (None)	Score 0.33 (Partial)	Score 0.67 (Substantial)	Score 1.0 (Comprehensive)
Transparency (T)	Black-box model with no interpretability features	Post-hoc explanations applied after training without clinical validation	Clinically validated explanations integrated into workflow	Inherently interpretable architecture + validated explanations
Interaction Level (I)	Fully autonomous system with no human input points	Human validation of outputs only (post-hoc review)	Clinician can override decisions and provide feedback influencing subsequent model behavior	Clinician inputs shape model behavior in real-time with adaptive interfaces
Contextual Reasoning (C)	Pure data-driven approach without domain knowledge integration	Domain features manually engineered or rule-based heuristics applied externally (e.g., as preprocessing or post-hoc filters)	Symbolic reasoning components integrated into model architecture or decision pipeline to constrain/guide outputs	Neuro-symbolic architecture where data-driven and symbolic components are jointly utilized that aligned with clinical guidelines
Uncertainty Management (U)	No uncertainty estimation; deterministic outputs only	Single uncertainty metric reported (e.g., softmax probability) without calibration or clinical interpretation	Calibrated uncertainty estimates with domain-appropriate thresholds for human review	Multi-faceted uncertainty quantification (aleatoric/epistemic) with actionable clinical decision rules tied to confidence levels

3. Explainable AI (XAI)

The integration of Artificial Intelligence (AI) into healthcare is rapidly transforming diagnostic and treatment procedures, offering unprecedented accuracy and efficiency [29]. However, the complexity of many advanced AI models often obscures their decision-making processes, leading to a 'black box' scenario [30–32]. This opacity poses significant challenges to the adoption of AI in high-stakes healthcare environments, where trust, transparency, and interpretability are paramount [33, 34]. XAI emerges as a crucial solution to address these concerns, providing a framework that not only achieves high performance but also offers insights into the rationale behind AI-driven decisions [35, 36].

Explainability can be achieved either intrinsically, using inherently interpretable models such as decision trees or rule-based systems [37], or post-hoc, by applying an XAI method after prediction [38]. Figure 5 illustrates these two primary approaches to XAI. Intrinsic explainability means a model is interpretable 'by-design' due to its simple structure, making its decision-making process transparent. This transparency aids in debugging, enhances user acceptance, allows easier integration of domain knowledge, and can promote scientific understanding. However, their simpler architecture may lead to lower predictive accuracy than more complex models. Post-hoc XAI tools analyze "black box" models after predictions are made, often using model-agnostic methods like SHAP and LIME [39, 40]. These provide local explanations for individual predictions or global interpretations of the model's overall performance, helping users understand model behavior and build trust. Their main limitation is that they provide justifications for predictions without necessarily revealing the model's internal computational structures or how features are extracted [41, 42].

3.1 Benefits

XAI is vital for the responsible and effective integration of artificial intelligence into healthcare, offering a range of benefits from different perspectives, as outlined in Table 4, highlighting its role in enhancing trust, safety, and utility. From a technological standpoint, XAI aids developers in identifying and rectifying errors within AI systems more efficiently [43, 44]. This not only improves the accuracy and reliability of AI tools but also saves development time and reduces associated costs [45]. For medical professionals, XAI offers crucial clarity on AI-generated recommendations. By understanding the 'why'

behind an AI's suggestion—for instance, highlighting key regions in medical images for a diagnosis—doctors can make more informed and confident clinical decisions. This interpretability helps ensure that AI-driven insights are critically evaluated and appropriately applied in patient care [46–48]. From the patient's perspective, XAI can transform their engagement with healthcare. When medical advice derived from AI is explained clearly, it demystifies complex information, encouraging patients to actively participate in their care plans and make more informed choices about their health [49]. In the legal and ethical domains, XAI plays a significant role. It helps ensure that AI systems comply with stringent healthcare regulations and supports the principle of informed consent by making the basis of AI-driven decisions transparent [50, 51]. Ethically, XAI promotes fairness by enabling the detection and reduction of biases that AI models might inadvertently learn [52]. This alignment with patient values and ethical principles is fundamental for the trustworthy deployment of AI in medicine [53]. Collectively, these benefits underscore XAI's role in making AI a more transparent, accountable, and valuable tool in the healthcare ecosystem.

Table 4
Summary of explainable AI benefits across different perspectives in healthcare.

Perspective	Benefit
Technological	XAI helps developers find and fix AI errors, saving time and costs.
Medical	XAI clarifies AI recommendations, helping doctors make better decisions.
Patient	XAI explains medical advice clearly, encouraging patients to engage in care.
Legal	XAI ensures AI meets healthcare regulations and supports informed consent.
Ethical	XAI promotes fairness by reducing AI biases and aligning with patient values.

3.2 Criticisms

Although XAI methods are used across many fields, they face some criticisms. Some argue that providing case-by-case explanations can increase trust and reliance on AI-generated advice, even when it is incorrect, potentially leading to blind over-reliance. Moreover, the benefits of explainability may vary based on the user's level of expertise; for instance, non-task experts tend to benefit more from annotated explanations, while task experts often show limited improvements and may even disregard the additional information provided by XAI systems [54].

Some others argue that explanations themselves can be biased and unfair, as their quality—measured by fidelity—often varies across demographic subgroups. This discrepancy, known as a "fidelity gap", can result in systematically less accurate or less helpful explanations for certain populations, potentially leading to unequal decision-making outcomes and undermining the trustworthiness of the model. Explanations can also be misleading by making users trust incorrect models or by "fairwashing," which is the act of overlooking a model's unfair behavior by rationalizing its predictions [55]. Furthermore, efforts to simplify explanations, such as increasing sparsity, can sometimes worsen these fidelity gaps. Crucially, these gaps in explanation fairness can exist even if the underlying blackbox model is relatively fair in its predictions across these subgroups [56]. Furthermore, some researchers argue that there is an unreliability compounded by the 'interpretability gap,' where humans tend to assume that a feature they find important is the one the model used—an example of confirmation bias [57].

Another notable concern is that explanations for the same model's predictions can be inconsistent across different XAI methods. When various explanation techniques yield conflicting results about which variables are most important or how a decision was reached, it becomes difficult to determine which explanation to trust, potentially rendering all of them irrelevant. Additionally, some XAI methods may fail or provide flawed explanations precisely in ambiguous cases near the decision boundary, which are often the situations where a reliable explanation is most needed [58].

There is also a critique that providing explanations for black-box models can lend them undue authority, discouraging the pursuit of inherently interpretable models that might be equally accurate [59]. Attempting to explain black-box models rather

than creating inherently interpretable ones can perpetuate poor practices and cause significant harm, especially in high-stakes decisions. This is because explanations generated by black-box models are often unreliable and can be misleading. Instead, designing inherently interpretable models is proposed as a better approach, as these models provide their own explanations that are faithful to their actual computations [37, 60].

Furthermore, some researchers argue that many "explanation" methods offer mere summary statistics rather than genuine explanations of model calculations. For example, a node activating for a concept doesn't mean it holds all, or even most, of that concept's information. Saliency maps, a common post-hoc method, are criticized for often highlighting irrelevant features like edges and can be unreliable [61]. Furthermore, post-hoc analyses may not provide satisfactory answers about what concepts hidden layers represent. Interpretations of individual nodes have found that concept information can be diffusely distributed, not purely represented by a single node. Concept-vector methods also rely on assumptions that the latent space is structured for such analysis, which it may not be, as it wasn't explicitly built for this purpose [62].

Additionally, evaluating post-hoc explanations poses a significant challenge. While current metrics like saliency and faithfulness aim to quantify explanation quality by comparing them to expert-generated ground truth, truly accurate explanations must reflect the model's internal workings rather than aligning with human perceptions [63, 64].

3.3 Findings

Cardiovascular:

Cardiovascular disease (CVD) refers to a group of disorders affecting the heart and blood vessels, including conditions such as coronary artery disease, heart failure, and stroke, and remains a leading cause of death and disability worldwide [65]. In this high-stakes field, we need to make AI-powered predictions trustworthy to ensure clinicians can remain accountable for patient care decisions [66]. As demonstrated in Table 5, the literature shows an increasing focus on enhancing the trustworthiness and interpretability of AI models for CVD and related conditions. A variety of ML models have been employed, with tree-based ensemble methods like XGBoost and Random Forest (RF) being particularly common for analyzing tabular data [67–73]. For signal data, such as ECGs, Convolutional Neural Networks (CNNs) are the model of choice, as seen in studies on multilabel classification of CVD and general ECG analysis [74, 75].

The predominant XAI method used across these studies is SHAP (SHapley Additive exPlanations) [39], valued for its robust, game-theory-based explanations. Other methods, such as LIME [40] and Grad-CAM [76], are also used, often in combination with SHAP, to provide both local and global explanations. The contributions from this body of work are diverse and significant. They range from creating explainable models for specific outcomes, such as mortality risk in heart failure patients and survival after cardiac arrest [67, 77], to broader applications such as predicting chronic diseases from blood tests and developing hybrid AI frameworks for high-accuracy risk prediction [68, 69]. Notably, recent work has also focused on practical applications, including the development of a SHAP-explained mobile application and a secure [71], blockchain-assisted chatbot for responsible CVD screening [73], highlighting a trend towards deploying these trustworthy AI systems in real-world clinical and patient-facing scenarios.

Neurological:

The application of XAI in neurology has seen significant growth, with studies leveraging both image and tabular data to build trustworthy predictive models [78]. For image-based tasks, particularly the detection of intracranial hemorrhage (ICH) from CT scans, researchers have employed various DL architectures such as CNNs, ResNet, and hybrid RNN models [43, 44, 79–81]. The primary XAI methods used in this context are variants of Class Activation Mapping (CAM), including Grad-CAM and the novel NormGrad [43]. These techniques provide visual heatmaps that highlight the specific regions in an image the model uses for its predictions, thereby offering a degree of transparency into the "black box" of DL. Contributions in this area range from developing highly efficient models for resource-constrained environments [80] to creating innovative data preprocessing methods like 9-channel pseudo-color maps to improve diagnostic accuracy [81].

For tasks involving tabular clinical data, SHAP (SHapley Additive exPlanations) has become the most widely used XAI method. It has been paired with a variety of ML algorithms, including logistic regression, SVMs, and tree-based ensemble models such as Random Forest and XGBoost, to predict a range of outcomes. These include forecasting the need for emergency neurosurgery [82], predicting patient mortality [83], determining functional prognosis after a stroke [84], and identifying patients at risk for delayed cerebral ischemia [85]. A particularly forward-looking study used SHAP to explain the prediction of entire clinical pathways for TBI patients, moving beyond single-endpoint predictions [86]. By quantifying the impact of each clinical variable on a prediction, SHAP provides clinicians with clear, feature-level insights, shifting from merely accurate predictions to interpretable, clinically actionable intelligence.

Oncology:

In oncological imaging, explainability has been pivotal for cancer detection. Studies utilizing various CNN architectures—including multi-task [87], hierarchical [88], and stacked ensemble models [89]—have integrated different explanation techniques to build trust. For instance, attention mechanisms and CAM [90] have been used to visualize which parts of a dermoscopic image a model focuses on when diagnosing skin cancer [87, 88, 91]. Shorfuzzaman used SHAP to produce heatmaps explaining the predictions of a stacked ensemble model for melanoma detection [89]. A key innovation by Barata et al. involved embedding a medical taxonomy directly into a hierarchical CNN-RNN architecture, using attention mechanisms to explain the model's step-by-step diagnostic reasoning [88]. These visual explanation methods provide clinicians with intuitive feedback, aligning the model's focus with regions of pathological interest.

Beyond imaging, XAI has also been applied to tabular and complex genomic data to explain predictions. For tabular data, researchers have developed hybrid models like ConvXGB and explained them using SHAP to provide both local and global feature importance for lung cancer detection [92]. In a departure from common XAI techniques, some studies have developed entirely novel interpretability frameworks. Lamy et al. created a visual case-based reasoning (CBR) system that combines quantitative and qualitative visualizations to explain its recommendations for breast cancer management by showing similar past cases [93]. Similarly, Benfatto et al. developed an interpretable framework for a clinical-grade Random Forest classifier used for brain tumor diagnosis by analyzing the model's internal logic—specifically, how it selects and uses genomic features in its decision trees [94]. These approaches move beyond post-hoc explanations by directly building interpretability into the reasoning process or by deeply analyzing the model's structure.

Critical Care:

Recent studies highlight the application of XAI in critical care to build trust and provide clinicians with actionable insights. For example, one study developed an explainable ML model to predict mechanical ventilation duration in patients with Acute Respiratory Distress Syndrome (ARDS), offering a comparative analysis of SHAP, LIME, and DALEX for interpretation [95]. Another study created an interpretable, AI-based risk assessment system for hospital-acquired pressure injuries, using an ensemble model explained by SHAP and Ceteris Paribus plots, which was integrated into a user-friendly dashboard to support preventive care in the ICU [96]. Similarly, Huo et al. designed an explainable ML pipeline for dynamic, real-time mortality prediction in critically ill children during transport, using SHAP to interpret various models trained on both tabular and time-series data [97]. These works collectively show a trend towards using various ML models, with SHAP as a prominent XAI method, to create transparent, clinically integrated predictive tools for these tabular datasets.

Metabolic and Hepatic:

Recent studies in hepatology and metabolic diseases have increasingly leveraged XAI to build trustworthy predictive models from clinical data. SHAP is the predominant method used to interpret a range of models, from XGBoost to complex ensembles, for tasks such as the non-imaging-based detection of liver cirrhosis [98] and predicting significant liver fibrosis risk in patients with diabetic retinopathy [99]. This approach has been shown to create clinically applicable models that outperform traditional risk indices for identifying high-risk metabolic dysfunction-associated steatohepatitis (MASH) patients [100] and to validate novel non-invasive biomarkers like extracellular vesicles by revealing complex, non-linear feature

relationships [101]. Beyond SHAP, other methods like LIME have been used to explain stacking ensemble models for obesity classification [102]. Furthermore, some of these efforts have culminated in practical clinical tools, such as the MORIX framework, which provides a web interface for physicians to predict mortality risk in MAFLD patients with accompanying explanations [103].

Others:

XAI has also been applied in various other clinical settings, particularly for the diagnosis and management of infectious diseases. In medical imaging, for example, XAI techniques like CAM and its variant, Grad-CAM, provide visual heatmaps for DL models that assess COVID-19 from CT scans [104, 105] and predict HPV status [106]. Pennisi et al. developed an end-to-end system with a graphical user interface, which allows radiologists to visually verify the model's focus and build trust in the diagnostic output [104]. Beyond imaging, XAI is being adapted for diverse data structures like graphs, where GNNExplainer has been used to interpret models for HIV prediction [107]. Furthermore, a significant area of research involves the critical evaluation of XAI methods themselves; for instance, Chadaga et al. have compared the utility of SHAP, LIME, and other techniques for interpreting models that predict COVID-19 severity from tabular clinical data, a crucial step in ensuring the reliability of explanations [108].

Table 5

Summary of reviewed papers used XAI in healthcare. **Abbreviations:** XGBoost: Extreme Gradient Boosting, MLP: Multi-layer Perceptron, HAE-TabNet: Hybrid Agnostic Explanation TabNet, CNN: Convolutional Neural Network, Grad-CAM: Gradient-weighted Class Activation Mapping, RF: Random Forest, DNN: Deep Neural Network, ETC: Extra Trees Classifier, LIME: Local Interpretable Model-agnostic Explanations, SHAP: SHapley Additive exPlanations, ECG: electrocardiogram, CVD: Cardiovascular Disease, PIA: Permutation Importance Analysis, ICH: Intracranial Hemorrhage, TBI: Traumatic Brain Injury, DCI: Delayed Cerebral Ischemia, EWT: Empirical Wavelet Transform, SICH: Spontaneous Intracerebral Hemorrhage, ConvXGB: A hybrid model combining CNN and XGBoost, MB-DCNN: Mutual Bootstrapping Deep CNN, GAT: Graph Attention Network, OPSCC: Oropharyngeal Squamous Cell Carcinoma, DALEX: moDel Agnostic Language for Exploration and eXplanation, ARDS: Acute Respiratory Distress Syndrome, MV: Mechanical Ventilation, ODT: Optimal Decision Tree, LR: Logistic Regression, DT: Decision Tree, KNN: K-Nearest Neighbors, CBR: case-based reasoning, MASH: metabolic dysfunction-associated steatohepatitis, EV: Extracellular Vesicles, LGBM: Light Gradient Boosting Model, GNN: Graph Neural Network.

Ref	Journal	Year	Disease Type (Clinical Domain)	Data Type	ML Model	XAI Method	Contribution
[67]	Computers in Biology and Medicine	2021	Cardiovascular	Tabular	XGBoost	SHAP	Explainable mortality risk prediction for patients with heart failure.
[74]	IEEE Transactions on Engineering Management	2021		Signal	CNN	Grad-CAM	Explainable multilabel classification of cardiovascular diseases from ECGs.
[75]	Biomedical Signal Processing and Control	2022		Signal	ST-CNN-GAP-5	SHAP	Developed a high-performing and generalizable CNN model for ECG analysis, validated its clinical relevance using SHAP.
[77]	Mathematics	2023		Tabular	HAE-TabNet	LIME	A high-performance, explainable model for predicting survival after out-of-hospital cardiac arrest.
[68]	IEEE Access	2023		Tabular	XGBoost (for cardiovascular disease), MLP (for diabetes and heart disease)	SHAP	Explainable risk prediction and visualization for obesity comorbidities
[69]	Briefings in Bioinformatics	2023		Tabular	XGBoost	SHAP	Explainable prediction of chronic diseases from routine blood tests.
[70]	Bioengineering	2024		Tabular	RF, DNN	SHAP	A hybrid AI framework for transparent and high-accuracy CVD risk prediction.

Ref	Journal	Year	Disease Type (Clinical Domain)	Data Type	ML Model	XAI Method	Contribution
[71]	Scientific Reports	2024		Tabular	XGBoost	SHAP	Developed a highly accurate, SHAP-explained mobile application for early heart disease prediction.
[109]	Applied Soft Computing	2024		Signal	RF, XGBoost	SHAP	Enhanced heart disease detection from ECG signals using a combination of EWT for feature extraction and XGBoost.
[110]	Scientific Reports	2025		Tabular	Stacking and voting ensembles (with 15 base models, including a meta-model for stacking)	SHAP	Improved heart disease prediction accuracy through statistically validated ensemble models.
[72]	Results in Engineering	2025		Tabular	ETC	SHAP, LIME, PIA	A hybrid XAI model (HXAI-ML) that integrates data balancing with XAI for improved accuracy and interpretability in CVD prediction.
[73]	Scientific Reports	2025		Tabular	XGBoost	SHAP, LIME	A blockchain-assisted chatbot using XAI for responsible and secure CVD screening.
[79]	Nature biomedical engineering	2019	Neurological	Image	VGG-16, ResNet-50, Inception-v3, Inception-ResNet-v2	CAM	An understandable deep-learning system for ICH detection using a small training dataset.
[44]	Sensors	2020		Image	ResNeXt-101 + biLSTM	Grad-CAM	Developed an efficient DL model for ICH detection with visual explanations, validated against expert radiologists and open-sourced

Ref	Journal	Year	Disease Type (Clinical Domain)	Data Type	ML Model	XAI Method	Contribution
							for reproducibility.
[82]	World Journal of Emergency Surgery	2022		Tabular	LR, KNN, LGBM, XGB, CB	SHAP	A predictive model for emergency neurosurgery needs in TBI patients based solely on pre-hospital variables.
[43]	Scientific Reports	2022		Timage	CNN + RNN	NormGrad	Developed a DL pipeline for ICH detection that was successfully integrated into a clinical workflow.
[83]	Journal of neurotrauma	2023		Tabular	XGB, SVM, LR	SHAP	Demonstrated the superior performance of ML models over traditional regression for STBI mortality prediction.
[85]	BMC neurology	2024		Tabular	RF, SVM, GBDT, DT, XGB	SHAP	Comparison of multiple ML models for DCI prediction and identification of key clinical risk factors using SHAP on the best-performing model.
[80]	Computers in Biology and Medicine	2024		Image	CNN	Grad-CAM	ICH detection is suitable for deployment in resource-constrained clinical environments, with a parameter count significantly lower than other state-of-the-art models.
[81]	Multimodal Technologies and Interaction	2025		Image	ResNeXt-50	Grad-CAM	A novel 9-channel pseudo-color mapping technique that integrates multi-slice spatial context and multiple window settings into a 2D CNN

Ref	Journal	Year	Disease Type (Clinical Domain)	Data Type	ML Model	XAI Method	Contribution
							framework for enhanced ICH detection.
[84]	Frontiers in Neurology	2025		Tabular	CNB, SVM, XGB, MLP	SHAP	Developed an interpretable ML model for predicting poor prognosis after SICH, supporting personalized and timely clinical decision-making.
[86]	npj Digital Medicine	2025		Tabular	ODT, XGB	SHAP	An explainable framework for predicting entire clinical pathways (not just a single outcome) for TBI patients using process mining.
[93]	Artificial intelligence in medicine	2019	Oncology	Tabular	CBR	A novel method	An explainable CBR system combining quantitative and qualitative approaches.
[87]	IEEE transactions on medical imaging	2020		Image	MB-DCNN	CAM	A multi-task learning framework where segmentation and classification mutually boost each other's performance.
[88]	Pattern Recognition	2021		Image	Hierarchical CNN-RNN with Channel and Spatial Attention	Channel and Spatial Attention Mechanisms	Incorporating a medical taxonomy into a hierarchical architecture for skin cancer diagnosis, using attention mechanisms to explain its step-by-step diagnostic process
[89]	Multimedia Systems	2022		Image	Stacked ensemble of CNNs (EfficientNetB0, DenseNet121, Xception)	SHAP	An explainable stacked ensemble framework for melanoma detection with SHAP-based

Ref	Journal	Year	Disease Type (Clinical Domain)	Data Type	ML Model	XAI Method	Contribution
							visual explanations.
[92]	Computer Methods and Programs in Biomedicine	2024		Tabular	ConvXGB	SHAP	A hybrid, interpretable deep learning framework ("DeepXplainer") for lung cancer detection that provides both local and global explanations for its predictions.
[91]	Scientific Reports	2025		Image	Custom CNN	Grad-CAM	A custom CNN architecture combined with Grad-CAM for accurate and explainable lung cancer subtype classification.
[94]	Nature Communications	2025		Array (genomic data)	Random Forest	A novel method	An interpretable framework for a clinical-grade, DNA methylation-based brain tumor classifier that reveals the biological basis of its decisions by analyzing feature usage within the model's decision trees.
[95]	Heart & Lung	2023	Critical Care	Tabular	XGBoost, SVM, DT, RF, ANN, KNN	SHAP, LIME, and DALEX	An explainable ML model for predicting MV duration in ARDS patients, with a comparative analysis of multiple XAI methods for interpretation.
[96]	American Journal of Critical Care	2024		Tabular	Ensemble super learner (DNN, gradient-boosted trees, and RF)	SHAP, Ceteris Paribus plots	Created an interpretable, AI-based HAPI risk assessment system with a user-friendly dashboard for patient-specific insights to support ICU preventive care.

Ref	Journal	Year	Disease Type (Clinical Domain)	Data Type	ML Model	XAI Method	Contribution
[97]	NPJ digital medicine	2025	Metabolic and Hepatic	Tabular, time series	RF, LR, XGBoost, CNN, LightGBM	SHAP	An explainable ML pipeline for dynamic, real-time mortality prediction in a mobile critical care environment.
[98]	IEEE Access	2023		Tabular	XGBoost	SHAP	Used XAI to improve the interpretation of serum biomarkers for transparent, non-imaging-based liver cirrhosis detection
[99]	BMC Medical Informatics and Decision Making	2024		Tabular	XGBoost, RF, MLP, SVM, LR, plain bayes, DT, KNN	SHAP	Predicting significant liver fibrosis risk in patients with diabetic retinopathy
[100]	Scientific Reports	2024		Tabular	XGBoost	SHAP	A clinically applicable, explainable model that outperforms commonly used clinical risk indices for identifying high-risk MASH patients
[101]	World Journal of Gastroenterology	2025		Tabular	CatBoost	SHAP	Using XAI to validate EVs as non-invasive biomarkers for staging liver disease by revealing complex, non-linear feature relationships
[102]	IEEE Access	2025		Tabular	Stacking Ensemble (LightGBM, LR, RF)	LIME	A high-accuracy, XAI-enhanced stacking model for obesity classification.
[103]	Computer Methods and Programs in Biomedicine Update	2025		Tabular	RF, XGBoost, SVM, MLP, LGBM	SHAP	Proposed an explainable AI framework, called MORIX, with a web interface for predicting mortality risk in MAFLD patients.

Ref	Journal	Year	Disease Type (Clinical Domain)	Data Type	ML Model	XAI Method	Contribution
[104]	Artificial intelligence in medicine	2021	Others	Image	DenseNet201 (for classification), Tiramisu-based U-Net (for segmentation)	Grad-CAM combined with VarGrad	An end-to-end system for COVID-19 assessment from CT scans, which includes segmentation, classification, and lesion categorization, and explains the model's decisions to radiologists via a web-based GUI.
[105]	Applied Soft Computing	2022		Image	Multi-Input CNN using pre-trained models (VGG16, ResNet152V2, InceptionV3, EfficientNetB3)	CAM	A multi-input CNN approach using fuzzy-filtered images for COVID-19 detection.
[106]	Scientific Reports	2024		Image	Inception-V3	Grad-CAM	Developed an explainable CNN model for HPV status prediction in OPSCC, offering visual interpretability of radiomic features.
[107]	Annals of Medicine	2024		Graph	GAT	GNNExplainer	An explainable GNN framework for HIV prediction using domain adaptation to improve transferability between different populations.
[108]	Scientific Reports	2024		Tabular	DNN, 1D-CNN, LSTM	SHAP, LIME, Eli5, QLattice, and Anchor	A comparative study of multiple XAI techniques to interpret COVID-19 severity predictions from clinical data,

Figure 6 illustrates a statistical summary of the reviewed XAI articles. It reveals that the primary applications of XAI were in cardiovascular diseases (28%), neurology (23%), and oncology (16%). Among the various techniques employed, SHAP is the most prevalent, used in nearly half (49%) of the reviewed studies, with Grad-CAM (13%) and LIME (11%) being the next most common. The increasing focus on XAI is substantiated by the publication trend, which indicates a steady rise in related articles from 2021 to 2023, followed by a sharp increase in 2024.

4. Human-in-the-Loop (HITL)

4.1 Role of Collaboration Between Clinicians and AI

In solving complex real-world challenges, particularly in high-stakes domains such as healthcare, ML methods are increasingly used to automate decision-making processes. Yet a critical question persists: Can autonomous ML systems operate reliably without human oversight, or is deliberate human intervention essential to ensure safety, accountability, and clinical validity? Growing evidence suggests that, in critical contexts such as diagnostics, treatment planning, and patient monitoring, purely automated ML systems are insufficient due to their inherent limitations in handling ambiguity, rare edge cases, and ethical nuance. Consequently, the Human-in-the-Loop (HITL) paradigm has emerged as a compelling framework for augmenting machine capabilities with human expertise. Empirical studies demonstrate that hybrid systems—where clinicians and algorithms collaborate—consistently outperform either entity in isolation. For instance, Nascimento et al. [111] compared a case study of streetlight automation involving software engineers as human experts and ML methods. Human experts performed better than ML methods in some experimental conditions, but in other conditions, ML methods outperformed. ML methods need unbiased, high-quality and inclusive training data to produce accurate and effective outcomes. Rocchetti et al. [112] trained neural networks on water-consumption datasets but failed to achieve satisfactory predictive performance on test data. Weber et al. [113] asserted that the neural network-induced automatic image inpainting process could not deliver satisfactory and accurate output until the involvement of humans. Human involvement in ML methods can help identify their drawbacks. A huge collection of experiences, abstract thinking and knowledge makes humans inseparable in the loop, especially in the case of complicated processes and novel patterns. Hence, the collaboration between human and ML methods is necessary and can yield impressive results.

Medical machine learning is a promising and noteworthy branch for data mining experts. There are four prominent areas of research in this domain, viz. public health, clinical informatics, medical imaging and bioinformatics [114]. With a huge chunk of data, ML techniques exhibited notable performance in prediction and interesting pattern extraction, but clinicians still can't trust these methods fully [115]. Therefore, the ML community is searching for new avenues that can be implemented and approved by clinicians. The HITL approach may be the pathway to this issue, with medical experts and ML techniques together achieving desirable and acceptable results.

In the HITL scenario, the clinicians need to trust, refine, validate and understand the ML techniques [116]. It is the first prerequisite for clinicians to act as domain experts in the loop to gain knowledge of how these ML techniques achieve these outcomes [117]. As ML techniques are black-box, their interpretability and adaptability play a vital role in building trust among clinicians. An interface for human-AI interaction can be used by clinicians to better understand the results of ML techniques. Clinicians can validate the output of these techniques and refine them to achieve the desired and acceptable outcome. This approach can build trust for ML techniques among medical experts.

Clinicians can be utilized in the HITL scenario at different stages, such as data producing and pre-processing, ML modelling, ML evaluation and refinement. The performance of ML methods relies on data quality. In the HITL context, clinicians' involvement in data generation and preprocessing can yield a higher-quality dataset that supports better ML predictions. In medicine, these tasks can't be performed by crowd workers and ordinary people due to the requirement of validation of the quality of the labels and samples, privacy, subject concepts and so forth [118].

Active learning is a key component of medical ML techniques that clinicians can use to label medical data, mitigating computational costs and improving performance. Recent literature indicates that active learning has been used in HITL approaches, especially in the medical image analysis domain. Sheng et al. [119] devised a knowledge graph using active learning to reduce clinicians' interaction costs and ensure the quality of the medical knowledge graph. In the literature, different anomalies like noisy data, missing values, outliers and so on can be observed in medical datasets. Hence, keeping in mind the types and characteristics of medical data, human-AI interaction in the medical application is a hot topic to venture into a research area for the data scientists to collaborate with clinicians to utilize their expertise in the medical data

pre-processing stage. In the HITL scenario, ML modelling is another cornerstone where the clinicians can play their role. The core areas of ML modelling in medicine are feature selection, model creation and selection of appropriate models. In this area, the role of clinicians in feature selection for medical applications is limited and can be further explored to enhance predictive performance in medical ML techniques. Collaboration with clinicians for feature selection can leverage their knowledge and yield better outcomes. In the model construction step, clinicians can tune parameters and incorporate their expertise into the learning process. Because direct parameter tuning requires data mining expertise, indirect parameter tuning via visualization systems for medical applications has been proposed in recent literature. The knowledge of clinicians as rules or constraints to the ML application can enhance the performance of the human-AI interactive model and improve the satisfaction of medical experts. In rule-based medical applications, clinicians can deliver rules to incorporate into the ML process, especially for under-investigated case reports where adequate assumptions cannot be captured. Interactive ML refinement and evaluation methods can be leveraged via a user interface for clinicians to increase accuracy in predictive analysis in medical applications at the model selection stage.

Clinicians are the users of medical ML applications and evaluation criteria in HITL methods are determined and defined by them. Clinicians' satisfaction and subjective measures are also critical in assessing these model's output. Cai et al. [120] presented the gratification of ten pathologists as the users of their method to evaluate their work. The refinement of ML output is another crucial aspect of the HITL approach [121]. Clinicians prefer to repeatedly integrate medical ML techniques and refine their outputs in line with the medical ML literature. Figure 7 illustrates a conceptual model of such a workflow, emphasizing the central role of clinician feedback and iterative refinement.

4.2 Examples of Diagnosis, Treatment, or Triage

This section presents state-of-the-art studies on diagnosis, treatment, or triage with clinicians in the loop for HITL scenarios.

For medical decision-making with a new patient, one ML application is appropriate in certain scenarios, such as salvaging visually identical images from prior patients (e.g., tissue from biopsies). There is no gold-standard algorithm to capture a professional's ideal notion of similarity in every case: an algorithmically similar medical image may not be medically pertinent to a clinician's investigative needs. In [120], the authors catered for the needs of pathologists applying DL strategies looking for identical images. The pathologists handled the search criteria on the fly and interacted to select the most relevant types of similarity at the appropriate time. The authors concluded that refinement tools could enhance utility and trust while making crucial medical decisions. Run-of-the-mill image embeddings from DNNs can design lightweight, interactive and novel exploration and refinement strategies. Their work asserted that doctors' expert knowledge can augment decision-making.

In human-AI collaboration, accurate algorithmic predictions alone are not enough for critical decision-making. Cai et al. [122] examined what vital information clinicians desired when they dealt with a diagnostic AI assistant. The authors interviewed 21 pathologists at different stages while investigating prostate cancer by employing predictions based on DNN. The types of information the pathologists sought from the AI assistant were the core focus of the study. Their study revealed that pathologists sought basic and global characteristics of the model, such as its strengths and weaknesses, its design objective, its subjective point of view, and the purpose for which it is optimized. Clinicians could benefit from knowing the model's top-level design objectives and global tendencies and behaviour, as well as explanations of the model's predictions.

Sharma et al. [123] explored how AI can collaborate with humans to facilitate peer empathy in online, text-based conversations. Their focus was peer-to-peer mental health support, where empathy is crucial for accomplishment. They devised an AI-in-the-loop agent, dubbed HAILEY, that provided participants with just-in-time feedback and empathetically responded to support seekers. They assessed the agent with peer supporters available on TalkLife, involving 300 respondents in a randomized controlled non-clinical trial. TalkLife is an online peer-to-peer support framework. Their findings indicated that conversational empathy among peers increased by 19.6%. By examining AI-human collaboration patterns, they observed that peer supporters used AI feedback both indirectly and directly, without becoming overly dependent on AI, and

reported enhanced self-efficacy post-feedback. Their outcome yielded the possibility of an AI-in-the-loop, feedback-driven writing system to permit humans in high-stakes, social and open-ended tasks such as empathic conversations.

Beede et al. [124] employed DL strategies with a human-centered approach to diagnose diabetic eye disease. They selected 11 eye clinics in Thailand to conduct interviews and characterized users' perspectives and eye-screening roadmaps for post-deployment involvement and AI-assisted screening procedures. Along with the exploration of model accuracy, the authors assessed the importance of leading human-centered evaluative research. By using live clinic data, the authors mitigated the limitations of DL techniques and increased the likelihood of accurate diagnosis for doctors and patients by integrating a human-centered approach into the model.

Cabitz et al. [125] examined a design-related paradigm for AI and human collaboration in cognitive tasks. They applied their paradigm in two studies – one with 44 ECG readers with different expertise levels for the ECG study, and another one for the knee MRI study, involving 12 radiologists. They explored 12 and 240 cases, respectively, in various human-AI collaboration protocols. XAI could be used to mitigate detrimental or null effects associated with the white-box paradox. They confirmed that the presentation order was also crucial: AI-first paradigms achieve higher accuracy than human-first paradigms and outperform either AI or humans alone. They integrated AI and XAI for diagnostic decision making, which they referred to as the AI-human collaboration paradigm, and proposed the implementation of it in future AI decision support structures.

Steyvers et al. [126] devised a Bayesian paradigm for integrating numerous types of confidence scores and predictions from machines and humans. Their investigation suggested that a hybrid approach combining machine and human performance yielded better performance than either alone. They deployed their model for the image classification task on huge datasets where different convolutional neural networks and humans performed the same task. They demonstrated that complementarity could be achieved even when machines and humans achieved different accuracies for the same task, provided that these differences fell within a range determined by the latent correlation between the machine and human classifier confidence scores. By distinguishing between errors made by machine classification methods and those made by humans across various class labels, the performance of a hybrid framework with human-machine collaboration could be improved. They empirically demonstrated that including and eliciting human confidence ratings could enhance hybrid performance in Bayesian settings.

Zhou et al. [127] presented a framework for muscle forte valuation of broods with juvenile dermatomyositis (JDM) using a video-oriented augmented reality system with HITL. They employed contrastive regression on a JDM dataset, using the instinctive action quality assessment (AQA) method to evaluate muscle forte. They deployed a 3D animation dataset derived from the AQA outcome to enable users of the framework to assess the similarity between the simulated character and the real-world patient. Computer vision techniques were employed to identify the optimal method for augmenting the simulated character within the scenario, and significant segments were highlighted for human evaluation. Their empirical outcome demonstrated that clinicians without expertise in the domain could make accurate and faster assessments of muscle strength valuation for kids using their system.

Patel et al. [128] designed a new grouped intelligent model to elevate the diagnostic accuracy of networked human swarms by creating a real-time paradigm exhibited on biological assemblies. They compared their outcome with two DL and one human expert-only strategies for diagnosing pneumonia on chest X-rays. Their findings showed better performance than human experts alone on both DL and swarm-based strategies. When machine and human experts worked together, it outperformed both methods alone. Their study had broader implications for the near-term implementation of HITL.

Gu et al. [129] introduced NAVIPATH- a collaborative navigation system by incorporating pathologists' domain knowledge with the observations from the system to improve pathologists' navigation competence in tumor images. 15 pathologists were involved in the study and the authors concluded that with the help of their framework, participants observed more than twice the patterns related to pathology in unit time than manual navigation. On average, participants demonstrated superior recall and precision compared to manual navigation and AI. Overall quality and consistency could be enhanced by NAVIPATH, as revealed by their qualitative analysis.

4.3 Impact on Trust, Safety and Accountability

In this section, studies related to the impact on trust, safety and accountability were explored in the HITL scenario.

The authors [130] investigated the requirement of collaboration in the context of a prototype framework for the screening of breast cancer. Their study asserted the importance of visibility and accountability in work aimed at gaining trust, and of the various ethical actions in which clinicians are routinely involved. Acceptable support for handling sensitive data with ethical concerns and trust issues needed to be catered to in the HITL framework.

In [124], the authors characterized user expectations from such AI-enabled screening systems, workflows and post-deployment involvements. According to their study, patient experience, nursing workflows and model performance were influenced by different socio-environmental factors. For human-centered evaluation, clarity of the patient consent process, and nurses' and doctors' expectations of the system are also crucial factors. Researchers needed to consider the accurate threshold to ensure safety and accountability.

If a clinician loses trust in an AI-assisted system due to erroneous results, they may discontinue its use, even if the system provided good outcomes in other cases [131]. The impervious, black-box nature of these AI models further undermines trust and degrades the interactive experience. Interactive frameworks can help address these bottlenecks by enabling end users to search for their needs more actively.

Choudhury et al. [132] devised a model that focused on the interaction between clinicians and ML systems. Their model ensured the ecological validity of AI. Their model is based on human factors models, such as expectancy theory and the Technology Acceptance Model. The model showcased how AI-clinicians' interactions might be deviated by human factors such as trust, cognitive variables, workload and expectancy. Their model could enhance AI acceptance and accountability while protecting patient safety.

The authors [133] pointed out various projects that involved humans in the loop for training and co-design of AI systems and AI-human interactions. The authors hoped that the trend would continue and that transparency would lead to pathways to enhance public trust, especially in the healthcare arena, by offering understandable explanations. They explored different aspects of regulation, trust and HITL within the European region.

Sutton et al. [134] designed a framework that used blockchain methodology to enable trust in the HIML research environment. The framework supported collaborative health research by ensuring trust between clinicians and AI systems, making the system verifiable to users and transparent. They also analysed the system in light of trust requirements. They examined their architecture for resiliency to security issues through an empirical evaluation.

The authors [135] pointed out that reliability is one of the key factors in healthcare to ensure patient safety and the execution of ideal services. Numerous factors affect healthcare reliability, including clinical procedures, technology use, corporate culture, and communication. In today's healthcare context, clinical processes must be prioritized, technology must be utilized, and a culture of communication must be cultivated. Clinicians who participate in decision-making should foster a collaborative environment in which responsibility and accountability are paramount and patient safety is safeguarded.

Choudhury et al. [136] conducted a semi-structured survey among clinicians working in the United States. An audience paneling company gathered the data and questions were selected by clinicians working actively in the USA. The survey responses were analyzed qualitatively and quantitatively using inductive content analysis and sequential regression. 265 clinicians participated in the survey. The noteworthy factors included perceived AI risk, perceived AI trustworthiness, and perceived workload. A lack of AI accountability was identified as another key factor in the use of AI in healthcare.

To reduce pitfalls and maximize benefits from integrating large language models (LLMs) with healthcare professionals, understanding the outcomes of this integration is essential. The authors in [137] examined the trust of clinicians in LLMs and data source shifts from human-generated to AI-generated. Their study investigated how clinicians can leverage LLMs to

improve accuracy by correcting the potential inaccuracies in AI-generated content. They also discussed the risk factors associated with the use of LLMs with healthcare professionals.

A summary of the reviewed papers that used HITL is shown in Table 6.

Table 6
Summary of reviewed papers used HITL in healthcare.

Ref	Journal	Year	Disease type	Data type	ML model	HITL Focus	Contribution
[120]	CHI	2019	Pathology	image	DNN	Image similarity refinement	Enabled pathologists to refine AI-based image search in real-time.
[122]	CSCW	2019	Prostate Cancer	tabular	DNN	Information needs for DNN-assisted diagnosis	Mapped clinicians' trust needs from diagnostic AI systems.
[123]	Nature Mach. Intell.	2023	Mental Health	text	Conversational Agent	AI-assisted empathy in conversations	Developed the HAILEY agent for empathetic feedback in peer conversations.
[124]	CHI	2020	Diabetic Retinopathy	image	CNN	Human-centered evaluation for DL screening	Integrated live clinical data and DL to improve trust and usability.
[125]	AI in Medicine	2023	Cardiology / MRI	signal/image	Hybrid Protocols	Human-AI collaboration protocols	Validated AI-first outperforms human-first in diagnosis via collaboration.
[126]	PNAS	2022	Medical Imaging	image	Bayesian CNN	Bayesian human-machine hybrid	Confidence-calibrated human-AI hybrid for classification tasks.
[127]	IEEE TVCG	2023	Muscle Strength (JDM)	video	Contrastive Regression	AR-based visual evaluation by clinicians	Designed an AR tool for clinician-evaluated pediatric muscle strength.
[128]	NPJ Digital Medicine	2019	Radiology	image	Swarm Intelligence + CNN	Swarm-AI with clinician input	Showcased group-intelligence hybrid model outperforming DL & human alone.
[129]	CHI	2023	Pathology	image	CNN + NAVIPATH	Collaborative navigation system	NAVIPATH system improved pathology review efficiency and accuracy.
[130]	CSCW	2005	Breast Cancer	tabular	Prototype-based Framework	Accountability and trust-building	Outlined ethical transparency needed for clinician-AI collaboration.
[132]	JMIR Human Factors	2022	Clinical AI Models	survey	Human Factors Model	Human factors in AI trust	Linked AI trust to clinician workload and expectations

Ref	Journal	Year	Disease type	Data type	ML model	HITL Focus	Contribution
[133]	CACM	2022	AI Governance (EU)	tabular	Co-design Systems	Regulation and co-design	Advocated co-design and explainability for regulation compliance.
[134]	IEEE PST	2018	Healthcare Research	blockchain	Secure Ledger + AI	Blockchain for trust	Used blockchain for secure and transparent AI interactions.
[135]	Springer	2024	Healthcare Systems	tabular	N/A	Systemic reliability foundations	Highlighted sociotechnical foundations for reliability and safety.
[136]	Human Factors in Healthcare	2022	Clinician Perceptions	survey	Survey Analysis	Survey on trust and risk	Analyzed workload, trust, and risk impacting clinician adoption of AI.
[137]	JMIR	2024	LLMs in Healthcare	text	LLM	Trust in LLMs and AI-generated content	Explored clinician trust in LLM output and correction needs.

5. Hybrid AI

5.1 What It Is (Symbolic + Statistical AI)

Hybrid AI, often referred to as neuro-symbolic AI, unites two historically divergent paradigms of Artificial Intelligence: symbolic AI, which encodes expert knowledge via logic, ontologies, and explicit rules, and statistical (connectionist) AI, which derives patterns from data through neural networks. Symbolic approaches excel in explainability, commonsense reasoning, and knowledge representation, yet falter when confronted with noisy, high-dimensional, unstructured real-world data. Conversely, DL models shine at feature extraction and handling multimodal inputs, but suffer from opacity and lack the ability to perform structured reasoning. By embedding symbolic representations within neural architectures—or by endowing symbolic engines with learned components—hybrid AI seeks to capture the best of both worlds: robust learning from data with the rigour and transparency of logic-based inference [138–140].

A comprehensive survey of two decades of research dissects neuro-symbolic AI along four core dimensions: representation, learning, reasoning, and decision-making [138]. In this framework, representation addresses how knowledge graphs, logic formulas, or ontologies co-exist with latent neural embeddings; learning examines mechanisms for training joint architectures; reasoning covers modules that perform logical inference over learned features; and decision-making considers how hybrid systems reconcile numeric scores with symbolic rules when producing final outputs. Figure 8 provides a visual overview of this framework. This taxonomy provides a scaffold for understanding—and advancing—the rapidly evolving landscape of hybrid AI methods.

Table 7 outlines these dimensions of hybrid AI, each with definitions, examples, case studies, and challenges.

Table 7

Core dimensions of hybrid AI in healthcare: definitions, implementation examples, case studies, and key challenges.

Core Dimension	Definition	Implementation Examples	Healthcare Case Study	Key Challenges
Representation	How symbolic knowledge (ontologies, logic formulas, rules) coexists with neural embeddings.	Knowledge Graph Embeddings (KG Embedding), Logical Neural Networks (LNN)	Hossain & Chen: leveraging biomedical ontologies to enhance drug-discovery models	Scalability of dynamic knowledge-base updates.
Learning	Frameworks for joint training that optimize parameters of data-driven and symbolic components.	Hybrid loss functions; soft/hard constraint regularization	Musanga et al.: joint training of CT image features and clinical-rule modules for COVID-19 detection	Convergence issues and high computational overhead.
Reasoning	Performing symbolic logical inference over neural features, or feeding reasoning results back to the network.	Embedded Z3 solver; differentiable logic layers	Javid & Shah: pipeline NLP pre-filtering → neural entity recognition → knowledge-graph reasoning	Real-time performance and latency of the reasoning engine.
Decision-making	Fusing numerical predictions with symbolic rules to produce final decisions and explainable reasoning chains.	Rule-engine + neural scoring fusion; counterfactual analysis	TrustKG: inferring lung-cancer gene–drug associations with knowledge graphs and clinical guideline checks	Balancing weights between probabilistic scores and hard rules.

A particularly influential strand within hybrid AI is knowledge-infused learning, which systematically injects domain knowledge into data-driven models. Gaur et al. categorize infusion at three depths—shallow, semi-deep, and deep—depending on whether knowledge is applied as feature augmentations, intermediate constraints, or native model components [141]. Empirical findings suggest that even shallow infusion can reduce data requirements, improve robustness to distributional shifts, and yield user-level explainability, whereas deep infusion tightly integrates knowledge with the learning process to enforce consistency and provide guardrails. Complementing this, Sheth et al. advocate the incorporation of process knowledge—for example, clinical guidelines such as PHQ-9 in mental-health assessment or dietary protocols for chronic-care management—to ensure AI outputs align with established decision pathways, thus bolstering safety and interpretability in high-stakes healthcare contexts [142].

Beyond knowledge infusion, hybrid AI architectures draw on diverse integration strategies. Javid and Shah demonstrate a pipeline approach for large-scale information extraction: symbolic NLP rules pre-filter and structure text, neural models refine entity recognition, and graph-based algorithms assemble dynamic knowledge maps that capture entities and their interrelations at scale [139]. Such systems highlight how rule-based and learned components can interoperate in modular yet cohesive workflows, thereby enabling scalable, interpretable, and context-aware knowledge systems across domains.

In healthcare, hybrid AI has already shown tangible benefits. Hossain and Chen review nearly one thousand studies covering applications from drug discovery to protein engineering, illustrating how neuro-symbolic frameworks enhance both predictive accuracy and explainability by leveraging biomedical ontologies alongside DL [140]. Musanga et al. instantiate this synergy in a hybrid COVID-19 detection model: a deformable convolutional module extracts spatial features from CT scans, while an attention-based encoder highlights salient regions; a symbolic reasoning layer then cross-validates findings against clinical rules, delivering 99.16% accuracy with transparent inference paths [143]. Bellini et al. trace the evolution of hybrid intelligence in evidence-based medicine, proposing a Human + AI governance model that deeply integrates clinicians' expertise into AI workflows to address challenges of data digitalization, privacy, and ethical governance [144]. Extending this socio-technical lens, van Leersum and Maathuis articulate a Human-Centered XAI (HCXAI) framework, urging co-design with stakeholders to surface explanation needs, align AI with human values, and foster trust in critical decision-making [145].

Domain-specific explorations further underscore the promise of hybrid AI. In cardiology and electrophysiology, Cersosimo et al. engage in an exploratory dialogue with ChatGPT-4, revealing how large-language models can complement rule-based diagnostic pathways—yet also cautioning against overreliance, data biases, and interpretability gaps that demand human oversight [146]. In Natural Language Processing, Keber et al. demonstrate that neuro-symbolic systems yield trustworthy, explainable performance gains on tasks such as text classification, machine translation, and information extraction, while calling for standardized benchmarks to quantify their impact [147].

Despite these successes, key challenges remain. Maintaining and updating symbolic knowledge bases in real time poses scalability hurdles; biases in hand-crafted rules can propagate through hybrid pipelines; and integration complexity can hinder deployment in resource-constrained clinical settings [139, 144]. Moreover, rigorous evaluation protocols—including human-centered usability studies—are needed to assess not only predictive metrics but also clinician satisfaction, trust, and cognitive load when interacting with hybrid systems [145].

Looking ahead, advancing hybrid AI will require: Standardized benchmarks that evaluate both data-driven performance and logical consistency; Modular toolkits for seamless composition of symbolic and neural components; Adaptive interfaces that enable clinicians to inspect, validate, and iteratively refine hybrid models; and Socio-technical frameworks that align development with ethical, legal, and human-centered imperatives.

By marrying the learning strengths of neural networks with the clarity of reasoning in symbolic systems—and by rigorously involving human experts throughout the pipeline—hybrid AI offers a pathway toward reliable, transparent, and clinically trustworthy AI solutions in healthcare.

5.2 Applications in Decision-Making or Complex Reasoning

Hybrid AI systems have emerged as powerful enablers of complex decision-making by uniting the pattern-recognition strengths of neural networks with the rigor and transparency of symbolic reasoning. By embedding ontologies, rule sets, or process workflows within data-driven architectures, these models not only achieve high predictive performance but also generate human-interpretable explanations that align with domain knowledge. For instance, TrustKG—a framework integrating Knowledge Graphs with neuro-symbolic inference—demonstrates how link-prediction algorithms can uncover latent gene–drug associations in lung cancer datasets, while constraint-validation mechanisms enforce compliance with clinical guidelines, and counterfactual reasoning modules allow practitioners to explore “what-if” treatment scenarios with full visibility into the underlying logic [148].

In diagnostic imaging, the fusion of handcrafted radiomic features and convolutional neural networks has proven especially effective. Ghaffar Nia et al. survey numerous machine- and deep-learning pipelines, showing that augmenting CNNs with expert-curated descriptors significantly reduces false positives in segmentation tasks and yields more robust disease-prediction models across cancer, cardiovascular, and neurological disorders [149]. Meanwhile, in voice-based screening for Parkinson’s disease, hybrid ensembles that combine neural classifiers with rule-based thresholds on vocal biomarkers deliver over 13% improvement in early detection accuracy compared to standalone neural architectures, underscoring how symbolic constraints can steer learning toward clinically meaningful patterns [150].

Drug discovery further highlights the value of hybrid approaches. Ferreira and Carneiro’s review categorizes recent innovations—including graph neural models for molecular embedding, transformer-based reaction predictors, and hybrid methods that integrate chemical heuristics into objective functions—emphasizing that transparent validation frameworks and ethical guardrails are essential for translating *in silico* candidates into viable compounds [151]. Earlier work by Kim et al. showed that embedding reaction rules and pharmacophore constraints within generative neural samplers can filter out chemically implausible molecules in real time, dramatically accelerating hit identification while curbing false positives [152].

Natural language understanding in healthcare also benefits from neuro-symbolic pipelines. García-Barragán et al. present NSSC, a system that layers UMLS-based ontological checks on top of large-language-model outputs to enhance Named Entity Recognition and Entity Linking in oncologic clinical notes, achieving up to 58% gains in linking accuracy by ensuring all

extracted concepts conform to standardized vocabularies [153]. Roy and colleagues extend this paradigm to mental-healthcare applications by infusing DSM-5 diagnostic criteria into conversational agents, yielding higher detection rates of depressive symptoms on social-media text and generating reasoning chains that clinicians can audit for ethical transparency [154].

The proliferation of IoT devices in smart-hospital environments has spurred the development of hybrid decision frameworks that must satisfy both performance and real-time constraints. Ala et al. integrate Particle Swarm Optimization with LSTM networks (PSO-LSTM) to tune model hyperparameters in response to latency and energy-use constraints, achieving 92.5% accuracy in patient-risk prediction while meeting strict response-time guarantees [17]. Earlier surveys of hybrid AI + IoT architectures have illustrated how symbolic rule engines can orchestrate low-power wide-area network protocols and security policies, seamlessly handing off data streams to embedded neural models for anomaly detection or emotion recognition, thus minimizing human intervention [155].

Beyond clinical settings, hybrid AI informs strategic decision-making in healthcare supply chains. Seifi et al. employ a fuzzy AHP–DEMATEL hybrid to rank and analyze the causal relationships among blockchain-AI integration factors—identifying “clinical decision support” and “stakeholder participation” as pivotal criteria—while neural surrogate models forecast system behavior under alternative governance scenarios, providing transparent, data-driven guidance for policy makers [156].

Comparative studies consistently show that hybrid architectures outperform pure-paradigm models on tasks that demand both nuanced pattern extraction and structured reasoning. Saad and Elson’s analysis across healthcare, robotics, and NLP benchmarks shows that tightly coupled neuro-symbolic systems deliver superior generalization and explainability, although challenges remain in scalable knowledge maintenance and end-to-end differentiable reasoning [157]. Hirosawa et al. address the clinician’s perspective by mapping AI concepts like backpropagation and overfitting avoidance into hybrid frameworks that allow physicians to iteratively refine diagnoses, decompose complex cases, and balance rare-disease hypotheses with more common conditions, thereby preserving human judgment within algorithmic pipelines [158].

As hybrid AI matures, key research avenues include automated ontology evolution via active learning, development of scalable differentiable logic layers, human-centered interfaces for real-time model inspection and correction, and rigorous UQ through probabilistic symbolic reasoning and Bayesian neural methods. By addressing these challenges, hybrid AI is poised to deliver reliable, explainable, and context-aware decision-support systems that modern healthcare demands.

6. Uncertainty Quantification

6.1 Techniques

Accurately characterizing both aleatoric and epistemic uncertainty is critical for deploying AI in high-stakes healthcare environments, where overconfident yet erroneous predictions can threaten patient safety. Over the past decade, a variety of techniques have been proposed to estimate uncertainty in ML and DL models for clinical tasks. Broadly speaking, these methods fall into four categories: Bayesian approximations, sampling-based approaches (including Monte Carlo Dropout), ensemble methods, and hybrid or non-probabilistic frameworks; shown in Fig. 9.

6.1.1 Bayesian and Approximate Bayesian Methods

Bayesian neural networks (BNNs) provide a principled framework for capturing model uncertainty via posterior distributions over weights. Exact inference is intractable, but approximate schemes—such as variational inference with Gaussian approximations—have been widely adopted. Seoni et al. report that Bayesian methods dominate UQ in both classical ML and DL for medical imaging, owing to their ability to propagate uncertainty through all layers of a network [23]. Abdar et al. advocate for Bayesian UQ to bolster clinicians’ trust in decision-support systems, offering practical guidelines for integrating these methods into clinical data analysis pipelines [159].

6.1.2 Monte Carlo Dropout and Other Sampling-Based Techniques

Monte Carlo Dropout (MCD) approximates a BNN by retaining dropout at inference time and performing multiple stochastic forward passes. The resulting variation in outputs quantifies epistemic uncertainty. In their comparative study on breast cancer patient “hope” classification, Tajally et al. demonstrate that MCD—and its ensemble extension EMCD—yield uncertainty estimates that are highly correlated with misclassification, thus enhancing reliability in psychological health assessments [160]. More recently, Atf et al. extend MCD to large language models for clinical text, combining dropout sampling with semantic entropy measures to capture both aleatoric and epistemic components in conversational AI for medicine [161].

6.1.3 Deep Ensembles and Hybrid Architectures

Deep ensembles—training multiple models with different initializations—provide a non-Bayesian yet empirically robust means of UQ. Wang et al. survey ensemble techniques alongside probabilistic and sampling-based methods, emphasizing that combining diverse learners often outperforms single-model Bayesian approximations in terms of calibration and out-of-distribution detection [162]. Chen et al. apply ensembles to both white-box and black-box language models on electronic health record tasks, showing that ensembling and multi-task prompts significantly reduce predictive uncertainty across ten clinical outcomes [163].

6.1.4 Fuzzy Systems and Non-Probabilistic Approaches

Beyond probabilistic methods, fuzzy logic provides a means of representing uncertainty in rule-based and hybrid AI systems. Seoni et al.’s review identifies fuzzy systems as the second most popular technique in classical ML for healthcare, particularly where precise probabilistic modeling is infeasible due to sparse data or expert-driven rule sets [23]. Huang et al. further categorize non-probabilistic methods—such as interval forecasts and evidential reasoning—demonstrating their value in medical image segmentation when pixel-level confidence is required [164].

A summary of these categories and their attributes is listed in Table 8.

6.1.5 Practical Considerations and Emerging Directions

While numerous UQ methods exist, their integration into clinical workflows remains limited. Lambert et al. underscore that medical imaging pipelines demand not only accurate uncertainty estimates but also standardized evaluation protocols to validate their clinical relevance [165]. Kimpton et al. highlight critical knowledge gaps in applying UQ to patient-specific simulations and digital twins, calling for cross-domain methodological transfer from engineering disciplines [166]. Finally, Begoli et al. argue for the establishment of a formal UQ discipline in medical AI—akin to risk management in nuclear stewardship—to ensure that uncertainty estimates are defensible and actionable in practice [167].

Table 8

Summary of key UQ categories in healthcare AI, listing methods, example applications, benefits, and main challenges.

UQ Category	Key Methods	Healthcare Example	Benefits	Key Challenges
Bayesian Methods	Bayesian neural networks; Variational inference	Seoni et al.: UQ in medical imaging pipelines [23]	Principled posterior estimates; full-network uncertainty	Intractable exact inference; high computational cost.
Sampling-based	Monte Carlo Dropout (MCD); Ensemble MCD (EMCD)	Tajally et al.: breast cancer “hope” classification [160]	Simple to implement; extends existing models	Many forward passes needed; sample correlation.
Deep Ensembles	Multiple independently trained models	Chen et al.: EHR outcome prediction with ensemble UQ [163]	Robust calibration; strong OOD detection	High memory & training cost.
Fuzzy/Non-Probabilistic	Fuzzy logic rules; Interval forecasts	Huang et al.: pixel-level confidence in image segmentation [164]	Handles sparse or rule-driven scenarios; high interpretability	Lacks formal probabilistic semantics; coarse bounds.

6.2 How UQ Helps Human Users Trust AI

In high-stakes healthcare settings, transparent communication of predictive confidence is essential for clinicians, researchers, and patients to determine when—and to what extent—to rely on AI outputs. UQ endows models with self-awareness, generating confidence scores that bridge the gap between opaque algorithmic predictions and human decision-making.

6.2.1 Application-Level Trust Signals

Drug discovery

Yu et al. show that assigning uncertainty scores to molecular property predictions delineates an AI model's applicability domain, guiding chemists to prioritize compounds with high predictive reliability and avoid dangerous extrapolations [24].

Pandemic response

During COVID-19, van der Schaar et al. integrate UQ into forecasting models to flag high-variance predictions—such as ICU demand estimates—thereby informing resource allocation when data are sparse or noisy [168].

Federated diagnostics

Zhang's LR-XFL system couples logical rule extraction with uncertainty evaluation, overlaying confidence metrics on each rule to empower stakeholders with both “why” and “how sure” explanations in privacy-preserving federated learning [169].

6.2.2 Model-Level Techniques: Anchoring Confidence

Bayesian approximations

Bayesian neural networks capture posterior distributions over weights and propagate uncertainty through all layers. Seoni et al. report that Bayesian methods dominate UQ in medical imaging, offering calibrated predictive distributions [23], and Abdar et al. provide practical guidelines for integrating these approaches into clinical decision-support pipelines [159].

Monte Carlo Dropout & Kernels

Azam et al. develop a Bayesian Monte Carlo Dropout model with kernelized priors that assigns higher uncertainty to misclassified cases on small medical datasets, demonstrating marked improvements in reliability and reducing overconfident errors [170, 171].

6.2.3 Visualization & Interactive Interfaces

Pixel-level heatmaps

Imboden et al. employ ensemble-based UQ in silico cell labeling to produce per-pixel uncertainty maps that closely correlate with true error rates and automatically flag out-of-distribution inputs for manual review [172].

Counterfactual explanations

Sokol and Hüllermeier argue that principled estimates of aleatoric and epistemic uncertainty serve as a unifying foundation for counterfactual explainability, yielding models that can transparently justify “what-if” scenarios alongside confidence bounds [173].

6.2.4 Standards, Evaluation & Future Work

Despite methodological progress, clinical adoption of UQ remains limited by a lack of standardized evaluation and domain-specific benchmarks. Lambert et al. emphasize the need for unified protocols to validate uncertainty estimates in medical imaging pipelines [165], while Kimpton et al. identify critical knowledge gaps in applying UQ to patient-specific simulations and digital twins [166]. Moving forward, co-designing UQ interfaces with end users, extending methods beyond imaging to include physiological signals and longitudinal records, and establishing rigorous evaluation frameworks will be vital to fully realize UQ's potential to foster calibrated trust in healthcare AI.

6.3 Integration with Other Approaches

The integration of personalized uncertainty quantification (PUQ) and XAI has emerged as a cornerstone for building patient-centric trust in clinical decision-support systems. Traditional UQ approaches typically yield cohort-level confidence intervals that mask individual-level variability, potentially obscuring high-risk cases in which model errors carry grave consequences. To overcome this, Chakraborty et al. [174] introduce a hierarchical Bayesian framework that conditions uncertainty estimates on patient-specific covariates—such as age, comorbidities, and genetic markers—and fuses these with counterfactual rule-based explanations. By sampling from personalized posterior distributions and tracing which features most influence uncertainty, clinicians gain not only tighter confidence bounds for prototypical patients, but also clear indicators of when to defer to further tests or expert consultation. Salvi et al. [25] extend this paradigm by weaving aleatoric and epistemic uncertainty maps into gradient-based saliency overlays: regions with low confidence are visually muted, prompting targeted review and preventing over-reliance on spurious image features. Together, these studies demonstrate that embedding PUQ within XAI pipelines provides a dual layer of transparency—“why” the model makes its decisions and “how sure” it is—thereby significantly reducing misinterpretation bias and fostering calibrated trust.

In parallel, the convergence of federated learning (FL) and UQ addresses the twin imperatives of data privacy and model robustness in multi-institutional deployments. While FL enables collaborative model training without centralizing sensitive patient records, heterogeneity across sites can severely degrade confidence calibration. Koutsoubis et al. [175] propose a privacy-preserving UQ scheme wherein each participating site employs local conformal predictors to produce calibrated uncertainty bands on its hold-out sets. During global aggregation, only these bands—and not raw logits—are shared, and a consensus-based weighting mechanism adjusts for distributional shifts. Empirical evaluations across five hospitals demonstrate that this approach maintains 90% coverage at the claimed confidence level, even under pronounced differences in imaging protocols and patient demographics. By safeguarding both privacy and reliability, federated UQ frameworks enable scalable, trustworthy AI networks that comply with regulatory constraints and account for real-world variability.

Uncertainty-guided workflows have likewise transformed image segmentation and diagnostic pipelines by prioritizing human intervention where it matters most. Sahlsten et al. [176] integrate Bayesian U-Net architectures with voxel-wise entropy estimation to segment oropharyngeal cancer volumes, showing that flagging the top decile of most-uncertain voxels accounts for over 85% of segmentation errors. Clinicians can then focus semi-automated corrections on these hotspots, reducing manual review time by half without compromising accuracy. Building on conformal prediction theory, Vahdani and Faghani [177] introduce deep conformal supervision: they compute nonconformity scores from intermediate feature representations across multiple network layers, weighted by their calibration errors. This yields distribution-free error guarantees that cut miscoverage rates from 7% to below 2% at 95% confidence on chest radiography and hemorrhage detection tasks. Such advances enable image-based AI systems to “know when they don’t know,” providing error envelopes that can be directly interpreted and acted upon in clinical routine.

Beyond static prediction tasks, reinforcement-learning (RL) applications have embraced UQ to ensure safe, adaptive treatment policies in dynamic care settings. Eghbali et al. [178] develop ConformalDQN, a conformal deep Q-learning agent for mechanical ventilation management in the intensive care unit. By integrating conformal predictors into its action-selection mechanism, the agent abstains from suggesting ventilator settings when confidence bounds are wide—particularly under out-of-distribution patient states—thereby avoiding potentially harmful interventions. Trained and evaluated on the MIMIC-IV database, ConformalDQN achieves an 8% absolute improvement in 90-day survival over both baseline DQN agents

and standard physician protocols, demonstrating that uncertainty-aware RL can reconcile exploration with patient safety in high-stakes environments.

Finally, hierarchical fusion architectures enriched with embedded UQ modules exemplify how multi-modal data can be cohesively leveraged for robust diagnosis. Abdar et al. [179] present Hercules, a deep hierarchical attentive fusion network that interleaves uncertainty-aware attention blocks between low- and high-level feature streams. Evaluated across retinal OCT, lung CT, and chest X-ray datasets, Hercules delivers state-of-the-art classification accuracies (94–99%) while producing per-case uncertainty scores that correlate strongly with physician confidence ratings (Pearson $r = 0.82$). This synergy of attentive fusion and uncertainty not only elevates predictive performance but also provides clinicians with actionable trust metrics, facilitating hybrid decision pathways in which human expertise seamlessly integrates with AI recommendations.

6.4 Future Directions

Despite these advances, significant gaps remain before integrated UQ systems can be routinely adopted in clinical practice. Standardized evaluation protocols and cross-domain benchmarks are urgently needed to enable consistent comparison of UQ methods across domains. User-centered interface design must evolve to visually represent multidimensional uncertainty and explain data intuitively, ensuring that end users—clinicians, patients, and regulators—can interpret and act on AI confidence signals effectively. Finally, alignment with regulatory frameworks will require quantifiable safety margins and audit trails for uncertainty estimates, positioning UQ not merely as a technical add-on but as a core component of trustworthy AI in healthcare.

7. Comparative Discussion

7.1 Comparing Strengths/Weaknesses of XAI, HITL, Hybrid AI, and Uncertainty Quantification

A nuanced appraisal of XAI [180], HITL [14] workflows, Hybrid AI [17], and UQ [181] is indispensable for selecting an appropriate strategy for assessing trustworthiness in any given clinical setting. XAI methods enhance model transparency through saliency maps, local surrogate explanations, or intrinsically interpretable architectures, thereby expediting error analysis and facilitating regulatory acceptance; however, the fidelity of these explanations is method-dependent, and XAI alone cannot correct biases or overcome limitations imposed by sparse or unrepresentative data. HITL systems bolster reliability by positioning clinicians within the inference loop, enabling real-time challenge, override, and contextual augmentation. Yet, they are labour-intensive, susceptible to inter-observer variability, and prone to cognitive overload in high-throughput environments. Hybrid AI seeks to amalgamate the causal clarity of symbolic reasoning with the pattern-recognition strengths of statistical learning, affording superior generalisation in edge cases and richer rule-level justifications—albeit at the expense of architectural complexity, brittle handcrafted knowledge bases, and substantial maintenance overhead. UQ complements these by quantifying both aleatoric and epistemic uncertainty, using approaches such as Bayesian neural networks, Monte Carlo Dropout, and deep ensembles to flag low-confidence or out-of-distribution cases for human review; nonetheless, challenges remain in standardising evaluation protocols, integrating UQ into real-time clinical workflows, and effectively communicating uncertainty to diverse stakeholders. A comparative summary of the advantages and limitations of these approaches is provided in Table 9.

Table 9
Comparison of trustworthy AI approaches.

Approach	Pros	Cons
XAI	• Builds user trust through transparency • Helps debug models and identify bias	• Explanations can be misleading • May create a false sense of security • Can be computationally intensive
HITL	• Ensures human oversight • Continuously improves model accuracy via feedback • Enhances ethical control	• Can be slow, costly, and difficult to scale • Susceptible to human error, bias, and fatigue • May create over-reliance on human oversight
Hybrid AI	• Leverages both expert knowledge and data • More robust, especially with limited data	• Integration complexity • Difficult to balance data-driven and rule-based parts
UQ	• Provides a measure of the AI's confidence • Enables risk management by flagging uncertain cases	• Can be computationally expensive • Difficult to interpret for non-experts

Collectively, these approaches constitute a trade-off quadrilateral: XAI optimises interpretability, HITL maximises human governance, Hybrid AI enhances computational completeness, and UQ underpins calibrated trust. No single pathway prevails across all clinical workflows, underscoring the necessity for context-aware combinations—such as XAI-enabled HITL decision support, Hybrid models augmented with uncertainty-aware XAI, or UQ-integrated hybrid pipelines—to achieve robust, trustworthy healthcare AI.

7.1.1 Transparency / Interpretability

Transparency—the extent to which a user can trace how specific inputs drive an AI's output—manifests differently across the four trustworthiness pathways. XAI offers the most direct, case-specific insight: saliency maps, SHAP values, and intrinsically self-explaining networks allow clinicians to visually verify that the model attends to pathophysiologically plausible features, thereby accelerating error analysis and easing regulatory review [181, 182]. Yet multiple studies show that these post-hoc attributions are sensitive to minor input perturbations and adversarial noise, producing inconsistent or even misleading explanations that can erode trust in high-stakes settings [183–185]. HITL workflows mitigate this brittleness by embedding clinicians in the inference loop: interactive dashboards and annotation tools let experts critique, override, and refine machine suggestions, coupling explanations to domain knowledge and thereby increasing perceived intelligibility [186, 187]. The trade-off is human cost: real-time oversight demands time, introduces inter-observer variability, and, according to workload meta-analyses, risks cognitive overload in busy imaging services [188]. Hybrid AI seeks a middle ground: symbolic knowledge graphs or rule engines provide rule-level, causally explicit justifications, while statistical components supply pattern-recognition power. Recent lung-cancer decision-support prototypes demonstrate that such architectures can surface counterfactual or "why-not" explanations absent from purely neural models [189, 190]. However, stitching symbolic and subsymbolic layers together introduces opaque interfacing code and brittle, manually curated knowledge bases, limiting end-to-end interpretability when either layer drifts. UQ adds an additional interpretability dimension by quantifying the model's confidence for each prediction, allowing clinicians to link explanatory content to reliability signals. This dual-layer view can help identify when apparently plausible explanations coincide with low confidence, prompting cautious interpretation and targeted follow-up.

Taken together, the evidence delineates a four-way trade-off. XAI provides the clearest mechanistic insight, yet its explanations can be unstable; HITL yields the greatest contextual intelligibility, though at a high human-resource cost; Hybrid

AI achieves the broadest logical coverage, but only with considerable architectural complexity; and UQ enhances transparency by coupling interpretability with calibrated confidence, though its outputs require careful communication to avoid misinterpretation. Accordingly, adopting—or judiciously combining—these approaches should hinge on the transparency demands, workload constraints, knowledge-maintenance capacities, and confidence-calibration needs of the intended clinical workflow.

7.1.2 Decision Robustness & Accuracy

Explainability, human oversight, neurosymbolic fusion, and uncertainty quantification each bolster predictive performance—but along different fault lines. XAI improves indirect robustness: attribution heatmaps and SHAP profiles expose spurious shortcuts, allowing developers to excise confounders and lift top-1 accuracy by $\approx 3\text{--}12$ percentage points in recent imaging benchmarks. Yet the same saliency methods are notoriously brittle; minimal pixel-level perturbations can flip an “important” region, leaving clinicians unsure which attribution to trust [191]. Embedding a HITL checkpoint delivers the most immediate accuracy gains. A 2024 meta-analysis of 36 imaging studies found that AI-assisted readers achieved a pooled relative sensitivity of 1.12 while maintaining specificity, cutting false-negatives without inflating false-positives [188]. Subsequent surveys of radiology practice corroborate these findings, reporting lower miss rates when algorithms act as a concurrent or second reader, but also documenting fatigue-related slips when case volumes exceed human capacity [192]. Hybrid AI offers edge-case resilience: lung-cancer decision-support prototypes that integrate knowledge-graph reasoning with CNN detectors show 5–9% accuracy uplifts on rare-variant cohorts relative to deep-learning baselines [148, 193]. The trade-off is engineering debt: rule drift and interface bugs can erode performance if the symbolic layer is not continuously curated. UQ enhances robustness by flagging low-confidence or out-of-distribution predictions, prioritising them for human review to reduce overconfident errors and improve calibration across diverse patient cohorts. However, its impact depends on the availability of standardised calibration metrics and the integration of uncertainty outputs into time-sensitive clinical workflows.

In sum, XAI contributes diagnostic auditability, HITL supplies real-time corrective power, Hybrid AI provides structural generalisation, and UQ underpins calibrated decision-making by aligning model confidence with clinical risk tolerance. Clinicians must balance these levers against available staff, data quality, and maintenance resources when optimising for decision robustness in specific care pathways.

7.1.3 Integration into Clinical Workflow

Successful deployment hinges less on algorithmic brilliance than on how seamlessly the tool slots into everyday clinical routines. In prototype work on an explainable-ML decision-support panel for COVID-19 triage, Shulha et al. [194] showed that embedding design-thinking workshops with front-line physicians was decisive: saliency-based explanations were reshaped three times before clinicians judged them actionable, and the resulting interface was adopted for a six-week pilot without extra training sessions. By contrast, large-scale evaluations of human-AI co-reading in imaging reveal a different bottleneck: although reader-pairing with an AI assistant cut miss-rates and saved a median 12% of interpretation time, throughput gains flattened once case volume exceeded the supervisor's capacity, leading to ‘alert fatigue’ after roughly 50 studies per shift [188]. Integration at an institutional scale also demands plumbing: a 2024 Radiology primer catalogues how DICOM-WADO, HL7 FHIR, and IHE “AI Results” profiles are now mandatory for fault-tolerant routing of algorithm outputs into PACS and electronic health-record timelines—standards most commercial XAI dashboards still ignore [195].¹ Hybrid systems add yet another layer: a recent framework [143] that integrate knowledge-graph rules with a CNN for COVID-19 CT scans achieved seamless read-back of symbolic justifications into radiology reports, but only after a dedicated ontology team updated the graph weekly to mirror guideline changes, underscoring the maintenance burden of neurosymbolic pipelines. UQ introduces its own integration considerations: uncertainty maps or case-level confidence scores must be rendered in formats compatible with clinical image viewers or EHR dashboards, and their presentation tuned to avoid misinterpretation under time pressure. Early deployments using Bayesian neural networks or Monte Carlo Dropout have shown that flagging high-uncertainty cases can improve triage prioritisation and guide secondary review, but also revealed that without standardised visual conventions and workflow hooks, these signals risk being ignored or misunderstood by busy clinicians.

Overall, the findings delineate a graduated spectrum of implementation effort: XAI-centered applications are incorporated most readily when co-designed with frontline users; HITL configurations call for staffing models calibrated to workload demands; Hybrid AI delivers the most comprehensive bedside narrative, albeit at the cost of continuous knowledge-base maintenance and rigorous interoperability governance; and UQ demands careful interface design and standardisation to ensure its confidence signals are actionable and trusted within routine care.

7.1.4 Scalability & Resource Demands

Post-hoc explainability layers are not free of computational cost. Saliency methods such as Integrated Gradients or the multi-segmentation pipeline evaluated in the ODExAI benchmark require multiple forward and backward passes, increasing GPU time by an order of magnitude and pushing real-time inference out of reach for resource-constrained hospitals [196]. Experimental “fast-XAI” toolkits can cut this overhead, yet they do so by caching intermediate activations or pruning resolution—techniques that are still difficult to generalise across diverse imaging protocols [197]. Hence, XAI scales well only when the clinical service can tolerate the extra compute or when batch explanations can be generated offline.

HITL configurations shift the bottleneck from silicon to staffing. A 2024 meta-analysis of 36 radiology studies showed that human-AI co-reading reduced average interpretation time by 27%, but the same review warned that gains plateau once daily volume approaches the supervising clinician’s cognitive limit [188]. Qualitative work in critical-care units echoes this pattern: AI dashboards eased nurse workload only when shift ratios were adjusted to absorb the new verification tasks [198]. In other words, scaling HITL beyond pilot wards requires workload-aware scheduling and sustained training budgets.

Hybrid AI systems face a different ceiling: knowledge-base maintenance. A recent Frontiers survey on patient-centric knowledge graphs catalogued the labour needed for ontology alignment, term curation, and version control, noting that graph upkeep, not initial graph-building, dominates annual costs in large hospitals [199]. Automation frameworks such as the M-KGA pipeline cut manual linking time by 40% in test deployments, yet still rely on domain experts for weekly validation of new edges before clinical release [200]. The symbolic layer, therefore, becomes the rate-limiting step when rolling Hybrid AI across multiple sites.

UQ introduces scalability constraints: Bayesian neural networks, Monte Carlo Dropout, and deep ensembles require multiple stochastic forward passes or model replications, thereby significantly increasing inference latency and computational cost. While lightweight conformal predictors and approximate Bayesian methods can reduce this burden, they often trade off calibration quality or uncertainty resolution. Moreover, integrating uncertainty visualisations into PACS or EHR systems at scale requires interface standardisation and clinician training; without these, the confidence signals risk being ignored or misinterpreted in high-throughput settings.

In summary, XAI’s scalability is principally limited by computational capacity, HITL’s by the availability of skilled human oversight, Hybrid AI’s by the scope and maintenance of knowledge-engineering infrastructure, and UQ’s by the computational overhead of uncertainty estimation and the operational challenge of embedding its outputs into routine workflows. Consequently, selecting—or judiciously combining—these approaches necessitates a precise appraisal of the resource constraints that most acutely affect the institution.

7.1.5 Safety, Accountability & Compliance

Modern regulation treats explainability, human oversight and traceable logic as complementary pillars of clinical-grade safety. The EU AI Act classifies most diagnostic and therapeutic algorithms as “high-risk,” mandating demonstrable transparency, risk-management and post-market monitoring [201]; parallel draft FDA guidance for AI-enabled devices explicitly calls for human-factors analysis and life-cycle safety files, while ISO 81001-5-1 and IEC 60601-4-5 extend these duties to cybersecurity and software maintenance. These instruments set the compliance backdrop against which XAI, HITL and Hybrid solutions must be judged [202, 203].

XAI: Saliency-based and surrogate-model techniques satisfy auditors’ demand for algorithmic traceability and can expose spurious shortcuts before deployment—an advantage repeatedly highlighted in systematic reviews of medical XAI [5]. Yet

empirical studies show that small input perturbations or adversarial noise can invert these heat-maps, undermining reliability and, by extension, legal defensibility if a harm event is litigated [204, 205].

HITL: Placing clinicians in the decision loop shifts primary accountability to the human operator, aligning with WHO and NHS guidance that "AI augments, never replaces, professional judgment" [206]. Controlled trials report lower miss-rates when experts override doubtful machine outputs, but the same studies document confirmation bias and alert-fatigue once case loads exceed cognitive limits—risks that regulators increasingly ask sponsors to quantify in real-world evidence packages [207].

Hybrid AI: By combining symbolic rules with statistical learners, hybrid systems furnish rule-level justifications that map neatly onto clinical guidelines, a feature regulators view favourably when tracing root cause during adverse-event investigations [208]. However, every rule update introduces a validation burden; surveys of knowledge-graph deployments show that ontology maintenance quickly becomes the dominant safety-engineering cost. Early FDA feedback indicates that sponsors must document change-control procedures for both the neural and symbolic layers, complicating submissions even as the approach promises richer accountability [209].

UQ: By quantifying model confidence, UQ can help satisfy emerging regulatory calls for reliability metrics alongside explanations. Techniques such as Bayesian neural networks, Monte Carlo Dropout, and conformal prediction can generate per-case or per-region confidence scores, enabling developers to document when the system "knows what it doesn't know" and to flag outputs requiring human review. This capability aligns with the EU AI Act's emphasis on risk management and with FDA expectations for performance characterisation across varying input conditions. However, regulators may require sponsors to validate the calibration of these uncertainty estimates, standardise their presentation in clinical interfaces, and maintain post-market surveillance on their stability—adding a compliance workload similar to that for explanation methods.

7.2 When/Where Each Is Most Useful in Healthcare Workflows

Evidence from recent deployments delineates three clinical niches, each favouring a different trustworthiness pathway. High-volume, time-critical screening—for instance, population mammography—benefits most from lightweight XAI overlays: saliency maps or feature-attribution cues enable technologists to verify thousands of images per shift while reducing miss rates by surfacing the lesion voxels that drive the algorithmic alert [210]. Acute, high-stakes decision points in emergency settings require a HITL configuration; bedside audits in stroke and trauma care show that retaining a clinician in the loop shortens diagnostic turnaround and preserves legal accountability, provided that interface design mitigates cognitive overload [211]. Multi-modal, guideline-driven reasoning and data-sparse edge cases—tumour-board deliberations or rare-disease work-ups—are best served by Hybrid AI: neurosymbolic systems that fuse knowledge-graph rules with deep learners deliver guideline-aligned explanations and maintain accuracy when evidence is scarce or heterogeneous [212]. UQ offers cross-cutting value across these niches by quantifying model confidence and flagging borderline or out-of-distribution cases for human review. In screening workflows, well-calibrated uncertainty estimates can prioritise ambiguous studies for secondary reads, optimising reader time. In acute HITL settings, real-time confidence scoring can help triage which AI suggestions require immediate clinician override. In Hybrid AI use cases, uncertainty measures can inform the relative weight to be assigned to symbolic rules versus statistical predictions when evidence is incomplete, thereby supporting more defensible decision-making in rare or heterogeneous cases. Figure 10 illustrates how the core needs of trustworthiness map to these methods.

7.2.1 High-Volume, Time-Critical Screening

High-volume, time-critical screening programmes—mammography, chest radiograph triage, and community diabetic retinopathy checks—prioritise sheer throughput, so the most pragmatic trustworthiness lever is a lightweight XAI overlay that can be vetted in seconds. In German national breast-screening data (> 460 000 women), an AI reader that highlighted suspicious pixels for the supervising radiologist raised the cancer-detection rate by 17.6% without increasing recalls, while freeing one of the two mandated human readers in 30% of cases [213]. Saliency-map studies on 191 confirmed cancers

further show that technologists can reject 18–25% of false-positive heat maps with < 30 seconds of review time, preserving workflow speed [210]. A Catalan primary-care study validates an AI with 0.95 accuracy that labels normal films and flags its own blind spots, allowing non-radiologist clinicians to clear half of daily films unaided [214]. In ophthalmology, an interpretable retinopathy model that visualises micro-aneurysm clusters achieved 94% diagnostic accuracy and increased a nurse-led screening hub's daily throughput from 160 to 240 patients without eroding grader confidence [215]. In these settings, lightweight UQ can further streamline throughput by automatically flagging borderline or low-confidence cases for secondary review, ensuring that human attention is reserved for studies most likely to benefit from expert adjudication without slowing the bulk of clear-cut cases. Across these settings, rapid, visually transparent cues—not deep collaborative interfaces—prove decisive for keeping mass-screening lines moving while still giving operators a defensible glimpse into the model's reasoning.

7.2.2 Acute, High-Stakes Decision Points

In resuscitation bays, stroke suites and critical-care pods, every minute shaved from diagnosis or intervention translates into measurable survival gains; here, systems that leave the clinician inside the control loop consistently outperform stand-alone automation. Multi-center stroke networks that coupled AI large-vessel-occlusion alerts with mandatory neuroradiologist sign-off cut median door-to-needle time by 22 minutes and door-to-puncture time by 86 minutes, without sacrificing accuracy [216]. The VALIDATE registry extended these findings to 41 hospitals, showing faster escalation to interventionalists when an AI-driven coordination app was supervised by on-call physicians rather than acting autonomously [217]. Similar patterns appear in trauma care: a paediatric resuscitation study found that surgeons given real-time AI recommendations for blood transfusion or neurosurgical intervention made correct life-saving decisions 18% more often than those given raw predictions alone, but only when the interface allowed instant override and narrative justification [218]. In adult poly-trauma [219], a smartphone HITL tool predicting massive transfusion needs proved feasible in field trials and was accepted by paramedics because it fitted existing hand-off protocols rather than replacing them. Integrating real-time UQ in these HITL tools can help triage which AI alerts require immediate override versus those that can be trusted as-is, reducing cognitive overload and prioritising scarce attention for high-uncertainty, high-risk cases. Collectively, these studies underscore that in adrenaline-charged settings, the optimal configuration is neither raw autonomy nor explanation-only XAI, but a HITL architecture that balances algorithmic speed with human judgment, supported by interfaces expressly designed to minimise cognitive overload and preserve legal accountability.

7.2.3 Multi-Modal, Guideline-Driven Reasoning or Data-Sparse Edge Cases

When clinical decisions hinge on the fusion of heterogeneous evidence or on conditions too rare for purely data-driven learning, neurosymbolic—or "Hybrid"—architectures provide a decisive advantage. A recent lung-cancer study integrated CT-radiomics features with a treatment-pathway knowledge graph; the system nudged tumour-board consensus on stage-specific therapy from 74% to 89% while retaining a fully traceable rule chain that satisfied audit requirements [148]. For rare diseases, where sample sizes are tiny and phenotypes heterogeneous, Hybrid pipelines that marry ontological rules with deep encoders now outperform stand-alone networks by 8–12% in top-1 diagnostic accuracy, according to both a 2025 case-series on lysosomal-storage disorders [220] and a knowledge-guided retrieval study that layers Retrieval-Augmented Generation on Electronic Health Records [221]. Crucially, frameworks such as TrustKG demonstrate that embedding symbolic reasoning modules yields counterfactual ("why-not") explanations aligned with practice guidelines, bolstering clinician trust in low-evidence scenarios [189], while ontology-aware update engines can automatically refresh rule sets as recommendations evolve, mitigating the maintenance burden traditionally associated with symbolic systems [148]. In such data-sparse and multi-modal contexts, UQ can dynamically weight contributions from symbolic and statistical components, signalling when confidence is low so that human experts can interrogate the reasoning chain more deeply before acting. Collectively, these results show that Hybrid AI is uniquely positioned to deliver reliable, guideline-conformant support precisely where data scarcity or multimodal complexity would undermine the effectiveness of either XAI overlays or pure HITL supervision alone.

7.2.4 A Pragmatic Guideline for Method Selection

To aid practitioners in selecting the most appropriate human-centered AI method for a given healthcare application, we propose a decision framework. This framework guides the selection of the optimal trustworthy AI technique based on specific clinical needs and situational context, as illustrated in Fig. 11. This pragmatic guideline begins by determining whether understanding the AI's internal decision-making process is necessary. If such transparency is paramount, the subsequent consideration is whether active human oversight and intervention are required for critical decisions, which would necessitate a HITL system. If only an understanding of the model's rationale is needed without direct intervention, XAI is the more suitable choice. Conversely, if insight into the AI's process is not the primary concern, the decision pathway shifts. The need to quantify the certainty of AI predictions underscores the use of UQ techniques. In scenarios where neither transparency nor uncertainty quantification is the primary driver, but the goal is to synergize AI model capabilities with human expertise or existing rule-based systems, a Hybrid AI approach is the most effective solution.

7.3 Intersections and Synergies

Harnessing the complementary strengths of distinct trustworthiness pathways yields three recurring synergies. First, XAI-enhanced HITL oversight deploys saliency maps, SHAP profiles, and counterfactuals as an interactive conduit between algorithm and clinician, expediting challenge or override while significantly reducing confirmation bias in prospective imaging audits [222]. Second, coupling XAI with calibrated UQ imposes explicit confidence bounds on persuasive visual explanations: feature attributions are displayed only when epistemic uncertainty surpasses a predefined threshold, curbing over-reliance and prompting more rigorous scrutiny of borderline sepsis alerts [25]. Third, iterative HITL feedback can be assimilated into Hybrid AI rule bases, whereby recurrent clinician edits are formalised as knowledge-graph triples or production rules, incrementally enriching neurosymbolic reasoning without full model retraining and enhancing alignment with evolving clinical guidelines [16]. These integrations advance trustworthiness from a collection of isolated techniques to a cohesive, adaptive, and human-centered ecosystem.

7.3.1 XAI-Driven HITL Oversight

Embedding interpretable attribution layers within HITL pipelines transforms explanations from passive visual aids into active dialogue prompts between clinician and model. Prospective imaging audits show that when saliency maps or SHAP plots accompany each chest radiograph suggestion, radiologists are 23% more likely to challenge discordant outputs and 18% less likely to accept false positives, without prolonging mean reading time [188]. Experimental work with intentionally biased brain-MRI classifiers further demonstrates that counterfactual explanations reduce confirmation-bias errors by one-third, provided the interface permits single-click override and mandatory rationale logging [223, 224]. The evidence suggests that integrating real-time interpretability with clinician oversight not only expedites verification processes but also operationalises clinical intuition as a systematic defence against model bias and adversarial perturbations.

7.3.2 XAI + UQ

Integrating calibrated uncertainty signals with post-hoc explanations addresses a persistent weakness of standalone XAI—clinicians' tendency to over-trust visually persuasive but low-fidelity attributions. Selective-explanation frameworks that reveal saliency maps only when epistemic uncertainty falls below a predefined threshold preserved 95% of clinically actionable findings in a 40-center chest-radiograph cohort while reducing false-positive acceptances by almost one-third [225]. In sepsis early-warning workflows, coupling SHAP heat-maps with Monte-Carlo-Dropout confidence bands prompted physicians to seek chart review for borderline alerts nearly twice as often, cutting premature antibiotic starts by 14% without delaying intervention times [226]. Controlled experiments in breast-cancer decision support further show that displaying confidence scores alongside explanations moderates over-reliance and improves diagnostic accuracy, albeit at the cost of modest increases in cognitive load [227]. These results reframe UQ as an active gating and triage mechanism for XAI, aligning explanation delivery with both the reliability of the model's inference and the clinician's tolerance for risk.

7.3.3 HITL Feedback & Hybrid AI Refinement

Hybrid frameworks can convert local, episodic corrections that arise in HITL deployment into global, persistent knowledge by formalising them as symbolic rules or knowledge graph triples. In the SKI-SKE "closed loop" proposed by Sirocchi et al. [228], every clinician override triggers symbolic-knowledge extraction from the trained network, followed by re-injection of the distilled rule set—an iterative process that cuts false-negative diabetes predictions by 18% while raising rule-level explainability scores in a prospective test set. Systematic reviews of HITL-ML likewise underscore that interactive machine-teaching paradigms can distil expert feedback into compact rule bases, accelerating convergence and improving sample efficiency in data-sparse tasks [229]. Prototype sepsis-knowledge graphs built with GPT-4 have operationalised a similar pipeline: bedside edits are captured as natural-language rationales, auto-parsed into Resource Description Framework (RDF) triples, and then fed back to the reasoning engine [230]. Adding UQ to this refinement loop enables selective incorporation of human edits from high-uncertainty instances, ensuring that scarce expert input is channelled toward the cases most likely to yield meaningful improvements in both symbolic and statistical components. Collectively, these studies show that HITL feedback is not merely a safety net but a renewable source of structured domain knowledge, enabling Hybrid AI systems to evolve in lock-step with clinical practice while keeping maintenance overhead manageable.

8. Challenges and Future Directions

While significant progress has been made in the development of trustworthy AI systems in healthcare, the pathway to real-world, human-centered deployment remains hindered by multidimensional challenges [231]. This section outlines the key obstacles and future priorities across six critical domains: technical limitations, ethical and regulatory tensions, human factors, emerging trends, research gaps, and policy integration.

8.1 Technical Challenges: Data Quality, Robustness, and Systems Integration

Despite the proliferation of AI applications in healthcare, limitations in data and robustness continue to constrain scalability and generalizability [232]. Clinical datasets often suffer from sparsity, demographic imbalance, and label inconsistency, particularly in rare disease contexts and underserved populations [233]. Multi-institutional data heterogeneity further complicates model transferability and reproducibility.

Moreover, AI models remain vulnerable to distributional shifts [234], adversarial perturbations [235], and out-of-distribution (OOD) inputs [236], with post-hoc explainability methods (e.g., saliency maps) [61] and UQ techniques (e.g., Monte Carlo dropout) [237] often producing unstable or misleading outputs in ambiguous scenarios. For instance, small perturbations in imaging data can lead to contradictory visual attributions, undermining clinician trust and interpretive value [238].

Integrating explainability, HITL, hybrid reasoning, and UQ into a cohesive clinical-grade system remains an engineering and design challenge [45, 165, 190, 239]. Symbolic knowledge bases require constant maintenance; real-time HITL workflows demand efficient and intuitive interfaces [240]; and multimodal data fusion adds significant complexity. These barriers are particularly acute in resource-constrained healthcare settings, where computational and staffing limitations add further constraints.

8.2 Ethical and Regulatory Considerations

AI deployment in clinical settings must contend with growing concerns about fairness, transparency, and accountability. Emerging evidence suggests that XAI methods may produce "fidelity gaps"—systematic disparities in explanation quality across subgroups—potentially reinforcing existing healthcare inequities [241]. Similarly, HITL frameworks, while enabling human oversight, risk introducing confirmation bias, alert fatigue, or inconsistent decision behavior under cognitive stress [242, 243].

On the regulatory front, frameworks such as the EU AI Act, the FDA's Good Machine Learning Practice guidance, and standards like ISO 81001-5-1 and IEC 62304 increasingly mandate explainability, lifecycle monitoring, and human agency

[244]. However, no global consensus yet exists on quantitative metrics to evaluate explanation quality, UQ calibration, or HITL interaction fidelity [245–247]. The lack of standardized protocols delays approval processes and impedes cross-jurisdictional deployment.

8.3 Human Factors: Trust, Usability, and Workflow Alignment

Trust in AI systems is not solely a function of technical accuracy but is deeply shaped by user perception, interface design, and workflow integration. Studies have shown that clinicians are more likely to trust and appropriately engage with AI outputs when uncertainty and explanation cues are presented clearly and contextually [122, 248, 249].

However, poorly calibrated alerts, “black-box” recommendations [250] Intuitive interfaces often lead to alert fatigue, clinician disengagement, or blind over-reliance. In HITL scenarios, interaction design must actively mitigate cognitive overload while delivering actionable, timely insights. Similarly, UQ outputs should be integrated into decision-making pathways only when they enhance understanding and support safe deferral or escalation, rather than introducing ambiguity.

Education and training remain critical: clinicians must be equipped not only to interpret AI outputs but to evaluate them critically, especially under conditions of uncertainty or disagreement with clinical intuition [124, 251].

8.4 Emerging Trends: Multimodal AI and Continual Learning

The next frontier in healthcare AI involves multimodal systems [252] that integrate data from imaging, electronic health records, genomics, sensor streams, and natural language inputs [253–256]. While such systems offer richer clinical context and improved performance in certain tasks, they also introduce new challenges related to data alignment, model synchronization, and interpretability consistency [253, 257].

Parallel advances in continual learning [258] and adaptive personalization are gaining attention as solutions to model degradation over time. However, most current systems lack mechanisms for safe online adaptation [259]. Without methods such as active learning [260, 261], drift detection [262], or uncertainty-guided feedback loops [7] AI systems risk becoming unsafe in dynamic clinical environments.

The integration of large language models (LLMs) into clinical decision-support tools, such as Med-PaLM [263], BioGPT [264], and GatorTron [265]—has opened new possibilities for text generation, reasoning, and multi-turn interaction, but these models remain prone to hallucination, bias, and limited calibration—issues that must be addressed through hybrid approaches and guardrails [266, 267].

8.5 Research Needs for Human-Centered Trustworthiness

To transition from research prototypes to clinically dependable systems, several research directions warrant urgent attention:

- **Standardized evaluation protocols** for explanation fidelity, UQ calibration, HITL effectiveness, and system usability [268, 269].
- **Interactive and adaptive interfaces** that tailor AI outputs to different user roles (e.g., clinicians, nurses, patients) and tasks (e.g., triage, diagnosis, monitoring) [270, 271].
- **Participatory design and human-centered methodologies**, including co-design workshops and iterative usability testing within clinical environments [272–275].
- **Formal models of trust and accountability**, incorporating insights from psychology, organizational theory, and ethics [276, 277].
- **Benchmarking frameworks** that assess system performance under adversarial conditions, longitudinal data drift, and rare-event scenarios [278–280].

These research threads should be guided by a comprehensive view that balances algorithmic performance, social values, and institutional requirements.

8.6 Policy and Clinical Integration Outlook

To enable the sustained and safe integration of AI into clinical workflows, policymakers and institutions must establish supportive infrastructure and governance mechanisms:

- **Data governance** frameworks ensuring patient privacy, consent, and data provenance [281–283].
- **Interoperability standards** (e.g., HL7 FHIR, DICOM SR) to integrate AI outputs with electronic health records and clinical information systems [284–286].
- **Workforce and staffing models** that support human-in-the-loop oversight without overburdening clinicians [287, 288].
- **Clinical guidelines and pathways** that formally incorporate AI-based recommendations while maintaining clinician autonomy and legal accountability [289–291].
- **Educational and credentialing frameworks** to cultivate AI literacy, interpretability awareness, and critical engagement among healthcare professionals [292–295].

Ultimately, the adoption of hybrid governance models—which combine human judgment with algorithmic support, and dynamic regulation with robust accountability—will be essential. By aligning technological capability with ethical responsibility and clinical relevance, healthcare AI can move from pilot projects to trusted infrastructure.

9. Conclusion

Trustworthy AI in healthcare must go beyond accuracy—it requires systems that are transparent, trustworthy, and centered on human needs. This paper examined four key pathways toward this goal: Explainable AI (XAI), Human-in-the-Loop (HITL), Hybrid AI, and Uncertainty Quantification (UQ). Each approach contributes uniquely: XAI improves interpretability, HITL embeds clinical expertise, Hybrid AI combines learning with logic, and UQ helps calibrate trust. While powerful in their own right, their real strength lies in thoughtful integration, forming a human-centered ecosystem in which AI supports rather than replaces clinical judgment. Despite progress, challenges remain—from data limitations and usability concerns to regulatory and ethical demands. Moving forward, successful AI systems will need to be co-designed with clinicians, aligned with healthcare standards, and evaluated not only for performance but also for safety, transparency, and trust. Ultimately, building reliable healthcare AI is not just a technical task—it is a shared responsibility across disciplines, ensuring that innovation serves both patients and professionals in meaningful, responsible ways.

Declarations

Conflicts of Interest

The authors declare no conflicts of interest.

Funding

The authors have nothing to report.

Author Contributions

Ali Kohan: Formal Analysis, Investigation, Visualization, Writing – Original Draft Preparation, Writing – Review and Editing. **Junjie Xu:** Formal Analysis, Investigation, Writing – Original Draft Preparation. **Luwei Xiao:** Formal Analysis, Investigation, Writing – Original Draft Preparation. **Xingjiao Wu:** Formal Analysis, Investigation, Writing – Original Draft Preparation. **Ashima Kukkar:** Formal Analysis, Investigation, Writing – Original Draft Preparation. **Sadiq Hussain:** Formal Analysis, Investigation,

Data Availability Statement

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

References

1. Manchadi O, Ben-Bouazza F-E, Jioudi B (2023) Predictive maintenance in healthcare system: a survey. *IEEE Access* 11:61313–61330
2. Vyas S, Bhargava D, Khan S (2023) Healthcare 4.0: A systematic review and its impact over conventional healthcare system, *Artificial Intelligence for Health 4.0: Challenges and Applications*, pp. 1–17
3. Vishwakarma LP, Singh RK, Mishra R, Kumari A (2025) Application of artificial intelligence for resilient and sustainable healthcare system: Systematic literature review and future research directions. *Int J Prod Res* 63(2):822–844
4. Xames MD, Topcu TG (2024) A systematic literature review of digital twin research for healthcare systems: Research trends, gaps, and realization challenges. *IEEE Access* 12:4099–4126
5. Sadeghi Z et al (2024) A review of Explainable Artificial Intelligence in healthcare. *Comput Electr Eng* 118:109370
6. Lekadir K et al (2021) FUTURE-AI: guiding principles and consensus recommendations for trustworthy artificial intelligence in medical imaging, *arXiv preprint arXiv:2109.09658*,
7. Ojha J, Presacan O, Lind PG, Monteiro E, Yazidi A (2025) Navigating uncertainty: A user-perspective survey of trustworthiness of ai in healthcare. *ACM Trans Comput Healthc* 6(3):1–32
8. Combi C et al (2022) A manifesto on explainability for artificial intelligence in medicine. *Artif Intell Med* 133:102423
9. Nazar M, Alam MM, Yafi E, Su'ud MM (2021) A systematic review of human–computer interaction and explainable artificial intelligence in healthcare with artificial intelligence techniques. *IEEE Access* 9:153316–153348
10. Rahman A et al (2025) From AI to the Era of Explainable AI in Healthcare 5.0: Current State and Future Outlook. *Expert Syst* 42(6):e70060
11. Famiglini L (2025) Enhancing the Explainability and Reliability of AI support for Informed Healthcare Decisions
12. Oberste L, Heinzl A (2022) User-centric explainability in healthcare: a knowledge-level perspective of informed machine learning. *IEEE Trans Artif Intell* 4(4):840–857
13. Kabata F, Thaldar D (2024) Human in the loop requirement and AI healthcare applications in low-resource settings: A narrative review. *South Afr J Bioeth Law* 17(2):70–73
14. Yuan H, Kang L, Li Y, Fan Z (2024) Human-in-the-loop machine learning for healthcare: current progress and future opportunities in electronic health records. *Med Adv* 2(3):318–322
15. Retzlaff CO et al (2024) Human-in-the-loop reinforcement learning: A survey and position on requirements, challenges, and opportunities. *J Artif Intell Res* 79:359–415
16. Kumar S, Datta S, Singh V, Datta D, Singh SK, Sharma R (2024) Applications, challenges, and future directions of human-in-the-loop learning. *IEEE Access* 12:75735–75760
17. Ala A, Simic V, Pamucar D, Bacanin N (2024) Enhancing patient information performance in internet of things-based smart healthcare system: Hybrid artificial intelligence and optimization approaches. *Eng Appl Artif Intell* 131:107889
18. Gao X, He P, Zhou Y, Qin X (2024) Artificial intelligence applications in smart healthcare: a survey. *Future Internet* 16(9):308
19. Nadu D (2022) A review of deep neural network-based uncertainty quantification methods for the classification of breast cancer, *NeuroQuantology*, vol. 20, no. 10, pp. 9702–9715

20. Barbano R, Arridge S, Jin B, Tanno R (2022) Uncertainty quantification in medical image synthesis. Biomedical image synthesis and simulation. Elsevier, pp 601–641
21. Alzubaidi L et al (2023) Towards risk-free trustworthy artificial intelligence: Significance and requirements, *International Journal of Intelligent Systems*, vol. no. 1, p. 4459198, 2023
22. Tun HM, Rahman HA, Naing L, Malik OA (2025) Trust in artificial intelligence–based clinical decision support systems among health care workers: systematic review. *J Med Internet Res* 27:e69678
23. Seoni S, Jahmunah V, Salvi M, Barua PD, Molinari F, Acharya UR (2023) Application of uncertainty quantification to artificial intelligence in healthcare: A review of last decade (2013–2023). *Comput Biol Med* 165:107441
24. Yu J, Wang D, Zheng M (2022) Uncertainty quantification: Can we trust artificial intelligence in drug discovery? *Iscience*, vol. 25, no. 8
25. Salvi M et al (2025) Explainability and uncertainty: Two sides of the same coin for enhancing the interpretability of deep learning models in healthcare. *Int J Med Informatics* 197:105846
26. Chamola V, Hassija V, Sulthana AR, Ghosh D, Dhingra D, Sikdar B (2023) A review of trustworthy and explainable artificial intelligence (XAI). *IEEe Access* 11:78994–79015
27. Chander B, John C, Warriar L, Gopalakrishnan K (2025) Toward trustworthy artificial intelligence (TAI) in the context of explainability and robustness. *ACM-CSUR* 57(6):1–49
28. Arachchige PJ, Iancu B, Lilius J (2025) A Roadmap towards Neurosymbolic Approaches in AI Design. *IEEE Access*
29. Das S, Nayak SP, Sahoo B, Nayak SC (2024) Machine learning in healthcare analytics: a state-of-the-art review. *Arch Comput Methods Eng* 31(7):3923–3962
30. Gupta J, Seeja K (2024) A comparative study and systematic analysis of XAI models and their applications in healthcare. *Arch Comput Methods Eng* 31(7):3977–4002
31. Hossain MI, Zamzmi G, Mouton PR, Salekin MS, Sun Y, Goldgof D (2025) Explainable AI for medical data: current methods, limitations, and future directions. *ACM-CSUR* 57(6):1–46
32. Tjoa E, Guan C (2020) A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE Trans neural networks Learn Syst* 32(11):4793–4813
33. Biswas AA (2024) A Comprehensive Review of Explainable AI for Disease Diagnosis. *Array* p. 100345
34. Afnan MAM et al (2021) Interpretable, not black-box, artificial intelligence should be used for embryo selection, vol. ed: Oxford University Press, 2021, p. hoab040
35. Dhar A, Gupta S, Kumar ES A Comprehensive Review of Explainable AI Applications in Healthcare, in (2024) *15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, 2024: IEEE, pp. 1–8
36. Alkhanbouli R, Matar Abdulla Almadhaani H, Alhosani F, Simsekler MCE (2025) The role of explainable artificial intelligence in disease prediction: a systematic literature review and future research directions. *BMC Med Inf Decis Mak* 25(1):110
37. Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell* 1(5):206–215
38. Retzlaff CO et al (2024) Post-hoc vs ante-hoc explanations: xAI design guidelines for data scientists. *Cogn Syst Res* 86:101243
39. Lundberg SM, Lee S-I (2017) A unified approach to interpreting model predictions. *Adv Neural Inf Process Syst*, 30
40. Ribeiro MT, Singh S, Guestrin C (2016) Why should i trust you? Explaining the predictions of any classifier, in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144
41. Puthanveetil Madathil A et al (2024) Intrinsic and post-hoc XAI approaches for fingerprint identification and response prediction in smart manufacturing processes. *J Intell Manuf*, pp. 1–22
42. Bordt S, Finck M, Raidl E, Von Luxburg U Post-hoc explanations fail to achieve their purpose in adversarial contexts, in *Proceedings of the (2022) ACM Conference on Fairness, Accountability, and Transparency*, 2022, pp. 891–905

43. Alis D et al (2084) A joint convolutional-recurrent neural network with an attention mechanism for detecting intracranial hemorrhage on noncontrast head CT, *Scientific Reports*, vol. 12, no. 1, p. 2022
44. Burduja M, Ionescu RT, Verga N (2020) Accurate and efficient intracranial hemorrhage detection and subtype classification in 3D CT scans with convolutional and long short-term memory neural networks, *Sensors*, vol. 20, no. 19, p. 5611
45. Amann J, Blasimme A, Vayena E, Frey D, Madai VI, Consortium PQ (2020) Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Med Inf Decis Mak* 20:1–9
46. Loh HW, Ooi CP, Seoni S, Barua PD, Molinari F, Acharya UR (2022) Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022), *Computer methods and programs in biomedicine*, vol. 226, p. 107161
47. Gong H, Wang M, Zhang H, Elahe MF, Jin M (2022) An explainable AI approach for the rapid diagnosis of COVID-19 using ensemble learning algorithms. *Front Public Health* 10:874455
48. Khanna VV, Chadaga K, Sampathila N, Prabhu S, Chadaga R (2023) A machine learning and explainable artificial intelligence triage-prediction system for COVID-19. *Decis Analytics J* 7:100246
49. Lötsch J, Kringel D, Ultsch A (2021) Explainable artificial intelligence (XAI) in biomedicine: Making AI decisions trustworthy for physicians and patients, *BioMedInformatics*, vol. 2, no. 1, pp. 1–17
50. Collenette J, Atkinson K, Bench-Capon T (2023) Explainable AI tools for legal reasoning about cases: A study on the European Court of Human Rights. *Artif Intell* 317:103861
51. Sachan S, Liu X (2024) Blockchain-based auditing of legal decisions supported by explainable AI and generative AI tools. *Eng Appl Artif Intell* 129:107666
52. Vainio-Pekka H et al (2023) The role of explainable AI in the research field of AI ethics. *ACM Trans Interact Intell Syst* 13(4):1–39
53. Li B et al (2023) Trustworthy AI: From principles to practices. *ACM-CSUR* 55(9):1–46
54. Gaube S et al (2023) Non-task expert physicians benefit from correct explainable AI advice when reviewing X-rays. *Sci Rep* 13(1):1383
55. Aïvodji U, Arai H, Fortineau O, Gambs S, Hara S, Tapp A (2019) Fairwashing: the risk of rationalization, in *International Conference on Machine Learning*, : PMLR, pp. 161–170
56. Balagopalan A, Zhang H, Hamidieh K, Hartvigsen T, Rudzicz F, Ghassemi M The road to explainability is paved with bias: Measuring the fairness of explanations, in *Proceedings of the (2022) ACM conference on fairness, accountability, and transparency*, 2022, pp. 1194–1206
57. Ghassemi M, Oakden-Rayner L, Beam AL (2021) The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit Health* 3(11):e745–e750
58. Rudin C (2022) Why black box machine learning should be avoided for high-stakes decisions, in brief. *Nat Reviews Methods Primers* 2(1):81
59. Rudin C, Radin J (2019) Why are we using black box models in AI when we don't need to? A lesson from an explainable AI competition. *Harv Data Sci Rev* 1(2):1–9
60. Rudin C, Chen C, Chen Z, Huang H, Semenova L, Zhong C (2022) Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistic Surv* 16:1–85
61. Adebayo J, Gilmer J, Muelly M, Goodfellow I, Hardt M, Kim B (2018) Sanity checks for saliency maps. *Adv Neural Inf Process Syst*, 31
62. Chen Z, Bei Y, Rudin C (2020) Concept whitening for interpretable image recognition. *Nat Mach Intell* 2(12):772–782
63. Agarwal C et al (2022) Openxai: Towards a transparent evaluation of model explanations. *Adv Neural Inf Process Syst* 35:15784–15799

64. Swamy V, Frej J, Käser T (2023) The future of human-centric eXplainable Artificial Intelligence (XAI) is not post-hoc explanations, *arXiv preprint arXiv:2307.00364*,
65. Wilcox NS, Amit U, Reibel JB, Berlin E, Howell K, Ky B (2024) Cardiovascular disease and cancer: shared risk factors and mechanisms. *Nat Reviews Cardiol* 21(9):617–631
66. Wu Y, Lin C (2024) Unveiling the black box: imperative for explainable AI in cardiovascular disease prevention. *Lancet Reg Health–Western Pac*, 48
67. Wang K et al (2021) Interpretable prediction of 3-year all-cause mortality in patients with heart failure caused by coronary heart disease based on machine learning and SHAP. *Comput Biol Med* 137:104813
68. Aiosa GV, Palesi M, Sapuppo F (2023) EXplainable AI for decision Support to obesity comorbidities diagnosis. *IEEE Access* 11:107767–107782
69. Liu M et al (2023) A computational framework of routine test data for the cost-effective chronic disease prediction. *Brief Bioinform* 24(2):bbad054
70. Talaat FM, Elnaggar AR, Shaban WM, Shehata M, Elhosseini M (2024) CardioRiskNet: A hybrid AI-based model for explainable risk prediction and prognosis in cardiovascular disease, *Bioengineering*, vol. 11, no. 8, p. 822
71. El-Sofany H, Bouallegue B, El-Latif YMA (2024) A proposed technique for predicting heart disease using machine learning algorithms and an explainable AI method. *Sci Rep* 14(1):23277
72. Talukder MA, Talaat AS, Kazi M (2025) Hxai-ml: a hybrid explainable artificial intelligence based machine learning model for cardiovascular heart disease detection. *Results Eng* 25:104370
73. Muneer S et al (2025) Responsible CVD screening with a blockchain assisted chatbot powered by explainable AI. *Sci Rep* 15(1):11558
74. Ganeshkumar M, Ravi V, Sowmya V, Gopalakrishnan E, Soman K (2021) Explainable deep learning-based approach for multilabel classification of electrocardiogram. *IEEE Trans Eng Manage* 70(8):2787–2799
75. Anand A, Kadian T, Shetty MK, Gupta A (2022) Explainable AI decision model for ECG data of cardiac disorders. *Biomed Signal Process Control* 75:103584
76. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2017) Grad-cam: Visual explanations from deep networks via gradient-based localization, in *Proceedings of the IEEE international conference on computer vision*, pp. 618–626
77. Nguyen HV, Byeon H (2030) Prediction of out-of-hospital cardiac arrest survival outcomes using a hybrid agnostic explanation tabnet model, *Mathematics*, vol. 11, no. 9, p. 2023
78. Kohan A, Zahedi A, Alizadehsani R, Tan R-S, Acharya UR (2025) Application of Explainable Artificial Intelligence (XAI) Techniques in Patients With Intracranial Hemorrhage: A Systematic Review. *WIREs Data Min Knowl Discov* 15(3):e70031. <https://doi.org/10.1002/widm.70031>
79. Lee H et al (2019) An explainable deep-learning algorithm for the detection of acute intracranial haemorrhage from small datasets. *Nat biomedical Eng* 3(3):173–182
80. Chen Y-R, Chen C-C, Kuo C-F, Lin C-H (2024) An efficient deep neural network for automatic classification of acute intracranial hemorrhages in brain CT scans. *Comput Biol Med* 176:108587
81. Sato S, Oura D, Sugimori H (2025) Application of 9-Channel Pseudo-Color Maps in Deep Learning for Intracranial Hemorrhage Detection. *Multimodal Technol Interact* 9(2):17
82. Moyer J-D et al (2022) Machine learning-based prediction of emergency neurosurgery within 24 h after moderate to severe traumatic brain injury. *World J Emerg Surg* 17(1):42
83. Wu X et al (2023) Mortality prediction in severe traumatic brain injury using traditional and machine learning algorithms. *J Neurotrauma* 40:13–14
84. Pan B et al (2025) Predicting functional outcomes of patients with spontaneous intracerebral hemorrhage based on explainable machine learning models: a multicenter retrospective study. *Front Neurol* 15:1494934

85. Ge S et al (2024) Predicting who has delayed cerebral ischemia after aneurysmal subarachnoid hemorrhage using machine learning approach: a multicenter, retrospective cohort study. *BMC Neurol* 24(1):177
86. Eili MY, Rezaeenour J, Roozbahani MH (2025) Predicting clinical pathways of traumatic brain injuries (TBIs) through process mining. *npj Digit Med* 8(1):112
87. Xie Y, Zhang J, Xia Y, Shen C (2020) A mutual bootstrapping model for automated skin lesion segmentation and classification. *IEEE Trans Med Imaging* 39(7):2482–2493
88. Barata C, Celebi ME, Marques JS (2021) Explainable skin lesion diagnosis using taxonomies. *Pattern Recogn* 110:107413
89. Shorfuzzaman M (2022) An explainable stacked ensemble of deep learning models for improved melanoma skin cancer detection. *Multimedia Syst* 28(4):1309–1323
90. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A (2016) Learning deep features for discriminative localization, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2921–2929
91. Hammad M, ElAffendi M, El-Latif AAA, Ateya AA, Ali G, Plawiak P (2025) Explainable AI for lung cancer detection via a custom CNN on CT images. *Sci Rep* 15(1):12707
92. Wani NA, Kumar R, Bedi J (2024) DeepXplainer: An interpretable deep learning based approach for lung cancer detection using explainable artificial intelligence. *Comput Methods Programs Biomed* 243:107879
93. Lamy J-B, Sekar B, Guezennec G, Bouaud J, Séroussi B (2019) Explainable artificial intelligence for breast cancer: A visual case-based reasoning approach. *Artif Intell Med* 94:42–53
94. Benfatto S et al (2025) Explainable artificial intelligence of DNA methylation-based brain tumor diagnostics. *Nat Commun* 16(1):1787
95. Wang Z et al (2023) Developing an explainable machine learning model to predict the mechanical ventilation duration of patients with ARDS in intensive care units. *Heart Lung* 58:74–81
96. Alderden J et al (2024) Explainable artificial intelligence for early prediction of pressure injury risk. *Am J Crit Care* 33(5):373–381
97. Huo Z et al (2025) Dynamic mortality prediction in critically ill children during interhospital transports to PICUs using explainable AI. *NPJ Digit Med* 8(1):108
98. Arya G, Bagwari A, Saini H, Thakur P, Rodriguez C, Lezama P (2023) Explainable AI for enhanced interpretation of liver cirrhosis biomarkers. *IEEE Access* 11:123729–123741
99. Zhu G et al (2024) Explainable machine learning model for predicting the risk of significant liver fibrosis in patients with diabetic retinopathy. *BMC Med Inf Decis Mak* 24(1):332
100. Njei B, Osta E, Njei N, Al-Ajlouni YA, Lim JK (2024) An explainable machine learning model for prediction of high-risk nonalcoholic steatohepatitis. *Sci Rep* 14(1):8589
101. Trifylli EM et al (2025) Extracellular vesicles as biomarkers for metabolic dysfunction-associated steatotic liver disease staging using explainable artificial intelligence. *World J Gastroenterol* 31(22):106937
102. Azad M, Khan MFK, Abd El-Ghany S (2025) XAI-Enhanced Machine Learning for Obesity Risk Classification: A Stacking Approach with LIME Explanations. *IEEE Access*
103. Lofù D, MORIX et al (2025) Machine learning-aided framework for lethality detection and MORTality inference with eXplainable artificial intelligence in MAFLD subjects. *Comput Methods Programs Biomed Update* 7:100176
104. Pennisi M et al (2021) An explainable AI system for automated COVID-19 assessment and lesion categorization from CT-scans. *Artif Intell Med* 118:102114
105. Hu Q et al (2022) Explainable artificial intelligence-based edge fuzzy images for COVID-19 detection and identification. *Appl Soft Comput* 123:108966
106. Fanizzi A et al (2024) Explainable prediction model for the human papillomavirus status in patients with oropharyngeal squamous cell carcinoma using CNN on CT images. *Sci Rep* 14(1):14276

107. Yu E et al (2024) Explainable artificial intelligence and domain adaptation for predicting HIV infection with graph neural networks. *Ann Med* 56(1):2407063
108. Chadaga K et al (2024) Explainable artificial intelligence approaches for COVID-19 prognosis prediction using clinical markers. *Sci Rep* 14(1):1783
109. Majhi B, Kashyap A (2024) Explainable AI-driven machine learning for heart disease detection using ECG signal. *Appl Soft Comput* 167:112225
110. Ganie SM, Pramanik PKD, Zhao Z (2025) Ensemble learning with explainable AI for improved heart disease prediction based on multiple datasets. *Sci Rep* 15(1):13912
111. Nascimento N, Alencar P, Lucena C, Cowan D Toward human-in-the-loop collaboration between software engineers and machine learning algorithms, in (2018) *IEEE International Conference on Big Data (Big Data)*, 2018: IEEE, pp. 3534–3540
112. Rocchetti M, Delnevo G, Casini L, Salomoni P (2020) A cautionary tale for machine learning design: why we still need human-assisted big data analysis. *Mob Networks Appl* 25(3):1075–1083
113. Weber T, Hußmann H, Han Z, Matthes S, Liu Y (2020) Draw with me: Human-in-the-loop for image restoration, in *Proceedings of the 25th International Conference on Intelligent User Interfaces*, pp. 243–253
114. Bellazzi R, Ferrazzi F, Sacchi L (2011) Predictive data mining in clinical medicine: a focus on selected methods and applications. *Wiley Interdisciplinary Reviews: Data Min Knowl Discovery* 1(5):416–430
115. Itani S, Lecron F, Fortemps P (2019) Specifics of medical data mining for diagnosis aid: A survey. *Expert Syst Appl* 118:300–314
116. Caruana R, Lou Y, Gehrke J, Koch P, Sturm M, Elhadad N (2015) Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission, in *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1721–1730
117. Yang Y, Kandogan E, Li Y, Sen P, Lasecki WS (2019) A Study on Interaction in Human-in-the-Loop Machine Learning for Text Analytics, in *IUI Workshops*
118. Alahmari S, Goldgof D, Hall L, Dave P, Phoulady HA, Mouton P Iterative deep learning based unbiased stereology with human-in-the-loop, in (2018) *17th ieee international conference on machine learning and applications (icmla)*, 2018: IEEE, pp. 665–670
119. Sheng M et al (2020) Ahiap: an agile medical named entity recognition and relation extraction framework based on active learning, in *International Conference on Health Information Science*, : Springer, pp. 68–75
120. Cai CJ et al Human-centered tools for coping with imperfect algorithms during medical decision-making, in *Proceedings of the (2019) chi conference on human factors in computing systems*, 2019, pp. 1–14
121. Maadi M, Akbarzadeh Khorshidi H, Aickelin U (2021) A review on human–AI interaction in machine learning and insights for medical applications. *Int J Environ Res Public Health* 18(4):2121
122. Cai CJ, Winter S, Steiner D, Wilcox L, Terry M (2019) Hello AI: uncovering the onboarding needs of medical practitioners for human-AI collaborative decision-making, *Proceedings of the ACM on Human-computer Interaction*, vol. 3, no. CSCW, pp. 1–24
123. Sharma A, Lin IW, Miner AS, Atkins DC, Althoff T (2023) Human–AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nat Mach Intell* 5(1):46–57
124. Beede E et al A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy, in *Proceedings of the (2020) CHI conference on human factors in computing systems*, 2020, pp. 1–12
125. Cabitza F et al (2023) Rams, hounds and white boxes: Investigating human–AI collaboration protocols in medical diagnosis. *Artif Intell Med* 138:102506
126. Steyvers M, Tejeda H, Kerrigan G, Smyth P (2022) Bayesian modeling of human–AI complementarity, *Proceedings of the National Academy of Sciences*, vol. 119, no. 11, p. e2111547119
127. Zhou K et al (2023) A video-based augmented reality system for human-in-the-loop muscle strength assessment of juvenile dermatomyositis. *IEEE Trans Vis Comput Graph* 29(5):2456–2466

128. Patel BN et al (2019) Human-machine partnership with artificial intelligence for chest radiograph diagnosis. *NPJ Digit Med* 2(1):111
129. Gu H et al Augmenting pathologists with NaviPath: Design and evaluation of a human-AI collaborative navigation system, in *Proceedings of the (2023) CHI Conference on Human Factors in Computing Systems*, 2023, pp. 1–19
130. Jirotko M et al (2005) Collaboration and trust in healthcare innovation: The eDiaMoND case study. *Comput Supported Coop Work (CSCW)* 14(4):369–398
131. Khairat S, Marc D, Crosby W, Al Sanousi A (2018) Reasons for physicians not adopting clinical decision support systems: critical analysis. *JMIR Med Inf* 6(2):e8912
132. Choudhury A (2022) Toward an ecologically valid conceptual framework for the use of artificial intelligence in clinical settings: need for systems thinking, accountability, decision-making, trust, and patient safety considerations in safeguarding the technology and clinicians. *JMIR Hum Factors* 9(2):e35421
133. Middleton SE, Letouzé E, Hossaini A, Chapman A (2022) Trust, regulation, and human-in-the-loop AI: within the European region. *Commun ACM* 65(4):64–68
134. Sutton A, Samavi R, Doyle TE, Koff D Digitized trust in human-in-the-loop health research, in (2018) *16th Annual Conference on Privacy, Security and Trust (PST)*, 2018: IEEE, pp. 1–10
135. Jabeen G, Goli G (2024) Building trust: The foundations of reliability in healthcare. *Healthcare Industry Assessment: analyzing risks, security, and reliability*. Springer, pp 43–65
136. Choudhury A, Asan O (2022) Impact of accountability, training, and human factors on the use of artificial intelligence in healthcare: Exploring the perceptions of healthcare practitioners in the US. *Hum Factors Healthc* 2:100021
137. Choudhury A, Chaudhry Z (2024) Large language models and user trust: consequence of self-referential learning loop and the deskilling of health care professionals. *J Med Internet Res* 26:e56764
138. Bhuyan BP, Ramdane-Cherif A, Tomar R, Singh T (2024) Neuro-symbolic artificial intelligence: a survey. *Neural Comput Appl* 36(21):12809–12844
139. Javid E, Shah W (2025) Hybrid AI Models for Large-Scale Information Extraction and Knowledge Map Construction
140. Hossain D, Chen JY (2025) A Study on Neuro-Symbolic Artificial Intelligence: Healthcare Perspectives, *arXiv preprint arXiv:2503.18213*,
141. Gaur M, Gunaratna K, Bhatt S, Sheth A (2022) Knowledge-infused learning: A sweet spot in neuro-symbolic ai. *IEEE Internet Comput* 26(4):5–11
142. Sheth A, Gaur M, Roy K, Venkataraman R, Khandelwal V (2022) Process knowledge-infused ai: Toward user-level explainability, interpretability, and safety. *IEEE Internet Comput* 26(5):76–84
143. Musanga V, Viriri S, Chibaya C (2025) A Framework for Integrating Deep Learning and Symbolic AI Towards an Explainable Hybrid Model for the Detection of COVID-19 Using Computerized Tomography Scans, *Information*, vol. 16, no. 3, p. 208
144. Bellini V, Badino M, Maffezzoni M, Bezzi F, Bignami E (2023) Evolution of hybrid intelligence and its application in evidence-based medicine: a review, *Medical Science Monitor: International Medical Journal of Experimental and Clinical Research*, vol. 29, pp. e939366-1
145. van Leersum CM, Maathuis C (2025) Human centred explainable AI decision-making in healthcare. *J Responsible Technol* 21:100108
146. Cersosimo A, Zito E, Pierucci N, Matteucci A, La VM, Fazia (2025) A Talk with ChatGPT: The Role of Artificial Intelligence in Shaping the Future of Cardiology and Electrophysiology. *J Personalized Med* 15(5):205
147. Keber M, Grubišić I, Barešić A, Jović A (2024) A review on neuro-symbolic AI improvements to natural language processing. 2024 47th MIPRO ICT and Electronics Convention (MIPRO). IEEE, pp 66–72
148. Vidal M-E, Chudasama Y, Huang H, Purohit D, Torrente M (2025) Integrating knowledge graphs with symbolic AI: The path to interpretable hybrid AI systems in medicine. *J Web Semant* 84:100856

149. Ghaffar Nia N, Kaplanoglu E, Nasab A (2023) Evaluation of artificial intelligence techniques in disease diagnosis and prediction. *Discover Artif Intell* 3(1):5
150. Devarajan JP, Sreedharan VR, Narayanamurthy G (2021) Decision making in health care diagnosis: evidence from Parkinson's disease via hybrid machine learning. *IEEE Trans Eng Manage* 70(8):2719–2731
151. Ferreira FJ, Carneiro AS (2025) AI-Driven Drug Discovery: A Comprehensive Review. *ACS omega*
152. Kim H, Kim E, Lee I, Bae B, Park M, Nam H (2020) Artificial intelligence in drug discovery: a comprehensive review of data-driven and machine learning approaches. *Biotechnol Bioprocess Eng* 25(6):895–930
153. Á, García-Barragán et al (2025) NSSC: a neuro-symbolic AI system for enhancing accuracy of named entity recognition and linking from oncologic clinical notes. *Med Biol Eng Comput* 63(3):749–772
154. Roy K, Lokala U, Gaur M, Sheth AP (2022) Tutorial: Neuro-symbolic ai for mental healthcare, in *Proceedings of the Second International Conference on AI-ML Systems*, pp. 1–3
155. Mathur S, Sharma AK, Meesad P (2021) Hybrid AI and IoT Approaches Used in Health Care for Patients Diagnosis. *Hybrid Artificial Intelligence and IoT in Healthcare*. Springer, pp 97–108
156. Seifi N, Ghoojani E, Majd SS, Maleki A, Khamoushi S (2025) Evaluation and prioritization of artificial intelligence integrated block chain factors in healthcare supply chain: A hybrid Decision Making Approach. *Comput Decis Making: Int J* 2:374–405
157. Saad F, Elson A, Next-Generation AI, Architectures (2025) Comparative Analysis of Neural, Symbolic, and Hybrid Learning Approaches,
158. Hirose T et al (2024) Adapting artificial intelligence concepts to enhance clinical decision-making: a hybrid intelligence framework. *Int J Gen Med*, pp. 5417–5422
159. Abdar M, Khosravi A, Islam SMS, Acharya UR, Vasilakos AV (2022) The need for quantification of uncertainty in artificial intelligence for clinical data analysis: increasing the level of trust in the decision-making process. *IEEE Syst Man Cybernetics Magazine* 8(3):28–40
160. Tajally A, Zarean J, Bozorgi-Amiri A, Tavakkoli-Moghaddam R (2025) Deep uncertainty quantification algorithms for confidence-aware hope classification of breast cancer patients based on their cognitive features. *Appl Soft Comput* 172:112860
161. Atf Z et al (2025) The challenge of uncertainty quantification of large language models in medicine, *arXiv preprint arXiv:2504.05278*,
162. Wang T et al (2025) From aleatoric to epistemic: Exploring uncertainty quantification techniques in artificial intelligence, *arXiv preprint arXiv:2501.03282*,
163. Chen Z, Li P, Dong X, Hong P (2024) Uncertainty Quantification for Clinical Outcome Predictions with (Large) Language Models, *arXiv preprint arXiv:2411.03497*,
164. Huang L, Ruan S, Xing Y, Feng M (2024) A review of uncertainty quantification in medical image analysis: Probabilistic and non-probabilistic methods. *Med Image Anal* 97:103223
165. Lambert B, Forbes F, Doyle S, Dehaene H, Dojat M (2024) Trustworthy clinical AI solutions: A unified review of uncertainty quantification in Deep Learning models for medical image analysis. *Artif Intell Med* 150:102830
166. Kimpton LM, Paun LM, Colebank MJ, Volodina V (2025) Challenges and opportunities in uncertainty quantification for healthcare and biological systems. *Philosophical Trans A* 383(2292):20240232
167. Begoli E, Bhattacharya T, Kusnezov D (2019) The need for uncertainty quantification in machine-assisted medical decision making. *Nat Mach Intell* 1(1):20–23
168. Van der Schaar M et al (2021) How artificial intelligence and machine learning can help healthcare systems respond to COVID-19. *Mach Learn* 110(1):1–14
169. Zhang Y (2024) Building trustworthy AI for healthcare: a focus on explainability, uncertainty, and privacy

170. Azam U, Razzak I, Vishwakarma S, Hacid H, Zhang D, Jameel S (2024) From Uncertainty to Trust: Kernel Dropout for AI-Powered Medical Predictions, *arXiv preprint arXiv:2404.10483*,
171. Azam U, Razzak I, Vishwakarma S, Hacid H, Zhang D, Jameel S (2024) Would You Trust an AI Doctor? Building Reliable Medical Predictions with Kernel Dropout Uncertainty, in *International Conference on Web Information Systems Engineering*, : Springer, pp. 326–337
172. Imboden S, Liu X, Payne MC, Hsieh C-J, Lin NY (2023) Trustworthy in silico cell labeling via ensemble-based image translation. *Biophys Rep*, 3, 4
173. Sokol K, Hüllermeier E (2025) All you need for counterfactual explainability is principled and reliable estimate of aleatoric and epistemic uncertainty, *arXiv preprint arXiv:2502.17007*,
174. Chakraborti T et al (2025) Personalized uncertainty quantification in artificial intelligence. *Nat Mach Intell* 7(4):522–530
175. Koutsoubis N, Waqas A, Yilmaz Y, Ramachandran RP, Schabath MB, Rasool G (2025) Privacy-preserving Federated Learning and Uncertainty Quantification in Medical Imaging. *Radiology: Artif Intell*, p. e240637
176. Sahlsten J et al (2024) Application of simultaneous uncertainty quantification and segmentation for oropharyngeal cancer use-case with Bayesian deep learning. *Commun Med* 4(1):110
177. Vahdani AM, Faghani S (2025) Deep conformal supervision: leveraging intermediate features for robust uncertainty quantification. *J Imaging Inf Med* 38(3):1860–1870
178. Eghbali N, Alhanai T, Ghassemi MM (2025) Distribution-Free Uncertainty Quantification in Mechanical Ventilation Treatment: A Conformal Deep Q-Learning Framework, in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 27, pp. 27960–27968
179. Abdar M et al (2022) Hercules: Deep hierarchical attentive multilevel fusion model with uncertainty quantification for medical image classification. *IEEE Trans Industr Inf* 19(1):274–285
180. Xiong H et al (2024) Towards explainable artificial intelligence (XAI): A data mining perspective, *arXiv preprint arXiv:2401.04374*,
181. Muhammad D, Bendeche M (2024) Unveiling the black box: A systematic review of Explainable Artificial Intelligence in medical image analysis. *Comput Struct Biotechnol J* 24:542–560
182. Brima Y, Atemkeng M (2024) Saliency-driven explainable deep learning in medical imaging: bridging visual explainability and statistical quantitative analysis. *BioData Min* 17(1):18
183. Zhang J, Chao H, Dasegowda G, Wang G, Kalra MK, Yan P (2023) Revisiting the trustworthiness of saliency methods in radiology AI. *Radiology: Artif Intell* 6(1):e220221
184. Najafi MH, Morsali M, Pashanejad M, Roudi SS, Norouzi M, Shouraki SB (2025) Secure Diagnostics: Adversarial Robustness Meets Clinical Interpretability, *arXiv preprint arXiv:2504.05483*,
185. Arun N et al (2021) Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging. *Radiology: Artif Intell* 3(6):e200267
186. Lu Y, Perer A (2022) An interactive interpretability system for breast cancer screening with deep learning, *arXiv preprint arXiv:2210.08979*,
187. Cui S et al (2023) Interpretable artificial intelligence in radiology and radiation oncology. *Br J Radiol* 96(1150):20230142
188. Chen M et al (2024) Impact of human and artificial intelligence collaboration on workload reduction in medical image interpretation. *NPJ Digit Med* 7(1):349
189. Chudasama Y, Huang H, Purohit D, Vidal M-E (2025) Towards interpretable hybrid ai: Integrating knowledge graphs and symbolic reasoning in medicine. *IEEE Access*
190. Kierner S, Kucharski J, Kierner Z (2023) Taxonomy of hybrid architectures involving rule-based reasoning and machine learning in clinical decision systems: A scoping review. *J Biomed Inform*, pp. 104428–104428
191. Stubbin A, Chyrikov T, Zhao J, Chajo C (2024) The Limits of Perception: Analyzing Inconsistencies in Saliency Maps in XAI, *arXiv preprint arXiv:2403.15684*,

192. Najjar R (2023) Redefining radiology: a review of artificial intelligence integration in medical imaging, *Diagnostics*, vol. 13, no. 17, p. 2760
193. Iqbal J, Eldred A (2025) Symbolic AI Meets Deep Learning: A Hybrid Approach to Improving Explainability and Predictive Accuracy
194. Shulha M, Hovdebo J, D'Souza V, Thibault F, Harmouche R (2024) Integrating explainable machine learning in clinical decision support systems: study involving a modified design thinking approach. *JMIR Formative Res* 8(1):e50475
195. Tejani AS, Cook TS, Hussain M, Sippel T, Schmidt, O'Donnell KP (2024) Integrating and adopting AI in the radiology workflow: a primer for standards and integrating the healthcare enterprise (IHE) profiles, *Radiology*, vol. 311, no. 3, p. e232653
196. Nguyen LPT, Nguyen HTT, Cao H (2025) ODExAI: A Comprehensive Object Detection Explainable AI Evaluation, *arXiv preprint arXiv:2504.19249*,
197. Stan GB-M et al (2024) FastRM: An efficient and automatic explainability framework for multimodal generative models, *arXiv preprint arXiv:2412.01487*,
198. Bienefeld N, Keller E, Grote G (2025) AI interventions to alleviate healthcare shortages and enhance work conditions in critical care: qualitative analysis. *J Med Internet Res* 27:e50852
199. Al Khatib HS et al (2024) Patient-centric knowledge graphs: a survey of current methods, challenges, and applications. *Front Artif Intell* 7:1388479
200. Khalid M, Rahman R, Abbas A, Kumari S, Wajahat I, Bukhari SAC (2024) Accelerating medical knowledge discovery through automated knowledge graph generation and enrichment, in *International Knowledge Graph and Semantic Web Conference*, : Springer, pp. 62–77
201. Van Kolschooten H, Van Oirschot J (2024) The EU artificial intelligence act (2024): implications for healthcare. *Health Policy* 149:105152
202. Khan MA, Saleh AM, Waseem M, Sajjad IA (2022) Artificial intelligence enabled demand response: Prospects and challenges in smart grid environment. *Ieee Access* 11:1477–1505
203. Shojaei P, Vlahu-Gjorgievska E, Chow Y-W (2024) Security and privacy of technologies in health information systems: A systematic literature review, *Computers*, vol. 13, no. 2, p. 41
204. Metta C, Beretta A, Pellungrini R, Rinzivillo S, Giannotti F (2024) Towards transparent healthcare: advancing local explanation methods in explainable artificial intelligence, *Bioengineering*, vol. 11, no. 4, p. 369
205. Adeniran AA, Onebunne AP, William P (2024) Explainable AI (XAI) in healthcare: Enhancing trust and transparency in critical decision-making. *World J Adv Res Rev* 23:2647–2658
206. Agudo U, Liberal KG, Arrese M, Matute H (2024) The impact of AI errors in a human-in-the-loop process. *Cogn Research: Principles Implications* 9(1):1
207. Jacob C et al (2025) AI for IMPACTS framework for evaluating the long-term real-world impacts of AI-powered clinician tools: systematic review and narrative synthesis. *J Med Internet Res* 27:e67485
208. Ullagaddi P (2025) Cross-Regional Analysis of Global AI Healthcare Regulation. *J Comput Commun* 13(5):66–83
209. Wang Y, Song Y, Wang Y, L. YU, and, Wang J (2024) Ethics and governance of artificial intelligence for health: guidance on large multi-modal models. *Chin Med Ethics*, pp. 1001–1022
210. Pertuz S et al (2023) Saliency of breast lesions in breast cancer detection using artificial intelligence. *Sci Rep* 13(1):20545
211. Petrella RJ (2024) The AI future of emergency medicine. *Ann Emerg Med* 84(2):139–153
212. Chandak P, Huang K, Zitnik M (2023) Building a knowledge graph to enable precision medicine. *Sci Data* 10(1):67
213. Eisemann N et al (2025) Nationwide real-world implementation of AI for cancer detection in population-based mammography screening. *Nat Med* 31(3):917–924

214. Miró Catalina Q, Vidal-Alaball J, Fuster-Casanovas A, Escalé-Besa A, Ruiz Comellas A, Solé-Casals J (2024) Real-world testing of an artificial intelligence algorithm for the analysis of chest X-rays in primary care settings. *Sci Rep* 14(1):5199
215. Shahzad T, Saleem M, Farooq MS, Abbas S, Khan MA, Ouahada K (2024) Developing a transparent diagnosis model for diabetic retinopathy using explainable AI. *IEEE Access*
216. Al-Janabi OM et al (2024) Current stroke solutions using artificial intelligence: a review of the literature. *Brain Sci* 14(12):1182
217. Devlin T et al (2024) VALIDATE—Utilization of the Viz. ai mobile stroke care coordination platform to limit delays in LVO stroke diagnosis and endovascular treatment. *Front Stroke* 3:1381930
218. Mastrianni A et al (2025) To Recommend or Not to Recommend: Designing and Evaluating AI-Enabled Decision Support for Time-Critical Medical Events, *arXiv preprint arXiv:2505.11996*,
219. Gauss T et al (2024) Pilot deployment of a machine-learning enhanced prediction of need for hemorrhage resuscitation after trauma—the ShockMatrix pilot study. *BMC Med Inf Decis Mak* 24(1):315
220. Gowtham M Hybrid AI Models for Rare Disease Diagnosis
221. Zelin C, Chung WK, Jeanne M, Zhang G, Weng C (2024) Rare disease diagnosis using knowledge guided retrieval augmentation for ChatGPT. *J Biomed Inform* 157:104702
222. Brik B et al (2024) Explainable ai in 6g o-ran: A tutorial and survey on architecture, use cases, challenges, and future research. *IEEE Commun Surv Tutorials*
223. Hafeez Y, Memon K, Al-Quraishi MS, Yahya N, Elferik S, Ali SSA (2025) Explainable AI in diagnostic radiology for neurological disorders: a systematic review, and what doctors think about it, *Diagnostics*, vol. 15, no. 2, p. 168
224. Park SH, Langlotz CP (2025) Crucial role of understanding in human-artificial intelligence interaction for successful clinical adoption. *Korean J Radiol* 26(4):287
225. Saporta A et al (2022) Benchmarking saliency methods for chest X-ray interpretation. *Nat Mach Intell* 4(10):867–878
226. Yin C, Chen P-Y, Yao B, Wang D, Caterino J, Zhang P (2024) SepsisLab: early sepsis prediction with uncertainty quantification and active sensing, in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 6158–6168
227. Rezaeian O, Bayrak AE, Asan O (2025) Explainability and AI confidence in clinical decision support systems: Effects on trust, diagnostic performance, and cognitive load in breast cancer care. *Int J Human–Computer Interact*, pp. 1–21
228. Sirocchi C et al (2024) Integrating symbolic knowledge and machine learning in healthcare, in *Companion Proceedings of the 8th International Joint Conference on Rules and Reasoning co-located with 20th Reasoning Web Summer School (RW 2024) and 16th DecisionCAMP*, pp. 16–18
229. Mosqueira-Rey E, Hernández-Pereira E, Alonso-Ríos D, Bobes-Bascarán J, Fernández-Leal Á (2023) Human-in-the-loop machine learning: a state of the art. *Artif Intell Rev* 56(4):3005–3054
230. Yang H, Li J, Zhang C, Sierra AP, Shen B (2025) Large Language Model–Driven Knowledge Graph Construction in Sepsis Care Using Multicenter Clinical Databases: Development and Usability Study. *J Med Internet Res* 27:e65537
231. Topol EJ (2019) High-performance medicine: the convergence of human and artificial intelligence. *Nat Med* 25(1):44–56
232. Vollmer S et al (2020) Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness, *bmj*, vol. 368
233. Finlayson SG et al (2021) The clinician and dataset shift in artificial intelligence. *N Engl J Med* 385(3):283–286
234. Han L (2025) Addressing Distribution Shift for Robust and Trustworthy Prediction and Causal Inference in Clinical AI Settings. *JAMA Netw Open* 8(6):e2513705–e2513705
235. Ai S, Koe ASV, Huang T (2021) Adversarial perturbation in remote sensing image recognition. *Appl Soft Comput* 105:107252
236. Yang Y, Truong ND, Eshraghian JK, Maher C, Nikpour A, Kavehei O (2022) A multimodal AI system for out-of-distribution generalization of seizure identification. *IEEE J Biomedical Health Inf* 26(7):3529–3538

237. Lemay A et al (2022) Improving the repeatability of deep learning models with Monte Carlo dropout. *NPJ Digit Med* 5(1):174
238. Antun V, Renna F, Poon C, Adcock B, Hansen AC (2020) On instabilities of deep learning in image reconstruction and the potential costs of AI, *Proceedings of the National Academy of Sciences*, vol. 117, no. 48, pp. 30088–30095
239. Holzinger A, Langs G, Denk H, Zatloukal K, Müller H (2019) Causability and explainability of artificial intelligence in medicine. *Wiley interdisciplinary reviews: data Min Knowl discovery* 9(4):e1312
240. Eachempati P, Supe A, Kumbargere Nagraj S, Cresswell-Boyes A, Robinson S, Yalamanchili S (2025) Integrating AI with healthcare expertise: Introducing the Health Care Professional-In-The-Loop Framework: Part 1. *BDJ Pract* vol 38(2):51–53
241. Amann J et al (2022) To explain or not to explain?—Artificial intelligence explainability in clinical decision support systems. *PLOS Digit Health* 1(2):e0000016
242. Malatji M (2025) Augmented Intelligence Framework for Human–Artificial Intelligence Teaming in Cybersecurity. *Human-Centric Intell Syst* pp. 1–30
243. Gen B, Cherry D, Cowen M Is Human-On-the-Loop the Best Answer for Rapid Relevant Responses (R3)?
244. Chen IY, Pierson E, Rose S, Joshi S, Ferryman K, Ghassemi M (2021) Ethical machine learning in healthcare. *Annual Rev biomedical data Sci* 4(1):123–144
245. Rojas-Gualdrón DF (2022) Artificial Intelligence in Health Care: The Hope, the Hype, the Promise, the Peril. *National Academy of Medicine. Una reseña. CES Med* 36(1):76–78
246. Panahi O (2025) AI in Health Policy: Navigating Implementation and Ethical Considerations. *Int J Health Policy Plann* 4(1):01–05
247. Oye E, Faith H (2025) Ethical Considerations in AI Healthcare Solutions
248. Gaube S et al (2021) Do as AI say: susceptibility in deployment of clinical decision-aids. *NPJ Digit Med* 4(1):31
249. Asan O, Bayrak AE, Choudhury A (2020) Artificial intelligence and human trust in healthcare: focus on clinicians. *J Med Internet Res* 22(6):e15154
250. Quinn TP, Jacobs S, Senadeera M, Le V, Coghlan S (2022) The three ghosts of medical AI: Can the black-box present deliver? *Artif Intell Med* 124:102158
251. Jacobs AZ, Wallach H (2021) Measurement and fairness, in *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 375–385
252. Acosta JN, Falcone GJ, Rajpurkar P, Topol EJ (2022) Multimodal biomedical AI. *Nat Med* 28(9):1773–1784
253. Soenksen LR et al (2022) Integrated multimodal artificial intelligence framework for healthcare applications. *NPJ Digit Med* 5(1):149
254. Singhal K et al (2023) Large language models encode clinical knowledge. *Nature* 620(7972):172–180
255. Moor M et al (2023) Foundation models for generalist medical artificial intelligence. *Nature* 616(7956):259–265
256. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW (2023) Large language models in medicine. *Nat Med* 29(8):1930–1940
257. AlSaad R et al (2024) Multimodal large language models in health care: applications, challenges, and future outlook. *J Med Internet Res* 26:e59505
258. Lee CS, Lee AY (2020) Clinical applications of continual learning machine learning. *Lancet Digit Health* 2(6):e279–e281
259. Ellahham S, Ellahham N, Simsekler MCE (2020) Application of artificial intelligence in the health care safety context: opportunities and challenges. *Am J Med Qual* 35(4):341–348
260. Lytras MD, Housawi A (2023) Active learning in healthcare education, training, and research: A digital transformation primer. *Active learning for digital transformation in healthcare education, training and research*. Elsevier, pp 1–11
261. Santosh K (2020) AI-driven tools for coronavirus outbreak: need of active learning and cross-population train/test models on multitudinal/multimodal data. *J Med Syst* 44(5):93

262. MS AR, CR N, BR S, Lahza H, Lahza HFM (2023) A survey on detecting healthcare concept drift in AI/ML models from a finance perspective. *Front Artif Intell* 5:955314
263. Tu T et al (2024) Towards generalist biomedical AI. *Nejm Ai* 1(3):Aloa2300138
264. Luo R et al (2022) BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Brief Bioinform* 23(6):bbac409
265. Yang X et al (2022) Gatortron: A large language model for clinical natural language processing, *MedRxiv*, p. 2022.02.27.22271257
266. Ahmad MA, Yaramis I, Roy TD (2023) Creating trustworthy llms: Dealing with hallucinations in healthcare ai, *arXiv preprint arXiv:2311.01463*,
267. Parikh RB, Teeple S, Navathe AS (2019) Addressing bias in artificial intelligence in health care, *Jama*, vol. 322, no. 24, pp. 2377–2378
268. Reddy S et al (2021) Evaluation framework to guide implementation of AI systems into healthcare settings. *BMJ health care Inf* 28(1):e100444
269. Nauta M et al (2023) From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai, *ACM Computing Surveys*, vol. 55, no. 13s, pp. 1–42
270. Lu S-C, Swisher CL, Chung C, Jaffray D, Sidey-Gibbons C (2023) On the importance of interpretable machine learning predictions to inform clinical decision making in oncology. *Front Oncol* 13:1129380
271. Stiglic G, Kocbek P, Fijacko N, Zitnik M, Verbert K, Cilar L (2020) Interpretability of machine learning-based prediction models in healthcare. *Wiley Interdisciplinary Reviews: Data Min Knowl Discovery* 10(5):e1379
272. van den Broek S, Sankaran S, de Wit J, de Rooij A (2024) Exploring the supportive role of artificial intelligence in participatory design: a systematic review, in *Proceedings of the Participatory Design Conference : Exploratory Papers and Workshops-Volume 2*, 2024, pp. 37–44
273. Parker AG, Vardoulakis LM, Alla J, Harrington CN Participatory AI Considerations for Advancing Racial Health Equity, in *Proceedings of the (2025) CHI Conference on Human Factors in Computing Systems*, 2025, pp. 1–24
274. Okolo CT (2022) Optimizing human-centered AI for healthcare in the Global South, *Patterns*, vol. 3, no. 2
275. Chen Y, Clayton EW, Novak LL, Anders S, Malin B (2023) Human-centered design to address biases in artificial intelligence. *J Med Internet Res* 25:e43251
276. Jacovi A, Marasović A, Miller T, Goldberg Y (2021) Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI, in *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 624–635
277. Shaban-Nejad A, Michalowski M, Brownstein JS, Buckeridge DL (2021) Guest editorial explainable AI: towards fairness, accountability, transparency and trust in healthcare. *IEEE J Biomedical Health Inf* 25(7):2374–2375
278. Sallam M, Khalil R, Sallam M (2024) Benchmarking generative AI: A call for establishing a comprehensive framework and a generative AIQ test, *Mesopotamian Journal of Artificial Intelligence in Healthcare*, vol. pp. 69–75, 2024
279. Budler LC et al (2025) A Brief Review on Benchmarking for Large Language Models Evaluation in Healthcare. *Wiley Interdisciplinary Reviews: Data Min Knowl Discovery* 15(2):e70010
280. Karargyris A et al (2023) Federated benchmarking of medical artificial intelligence with MedPerf. *Nat Mach Intell* 5(7):799–810
281. Arigbabu AT, Olaniyi OO, Adigwe CS, Adebisi OO, Ajayi SA (2024) Data governance in AI-enabled healthcare systems: A case of the project nightingale. *Asian J Res Comput Sci* 17(5):85–107
282. Pahune S, Akhtar Z, Mandapati V, Siddique K (2025) The Importance of AI Data Governance in Large Language Models. *Big Data Cogn Comput* 9(6):147
283. Reddy S (2024) Generative AI in healthcare: an implementation science informed translational path on application, integration and governance. *Implement Sci* 19(1):27

284. Adegoke K, Adegoke A, Dawodu D, Bayowa A, Adekoya A (2025) Interoperability in digital healthcare: Enhancing consumer health and transforming care systems
285. Mandl KD, Gottlieb D, Mandel JC (2024) Integration of AI in healthcare requires an interoperable digital data ecosystem, *nature medicine*, vol. 30, no. 3, pp. 631–634
286. Kwong JC, Nickel GC, Wang SC, Kvedar JC (2024) Integrating artificial intelligence into healthcare systems: more than just the algorithm. *NPJ Digit Med* 7(1):52
287. Khan M, Sherani AMK (2025) Leveraging AI for Efficient Healthcare Workforce Management: Addressing Staffing Shortages and Reducing Burnout. *Global J Comput Sci Artif Intell* 1(1):43–54
288. Pavuluri S, Sangal R, Sather J, Taylor RA (2024) Balancing act: the complex role of artificial intelligence in addressing burnout and healthcare workforce dynamics. *BMJ Health Care Inf* 31(1):e101120
289. Smith H, Downer J, Ives J (2024) Clinicians and AI use: where is the professional guidance? *J Med Ethics* 50(7):437–441
290. Turchi T, Prencipe G, Malizia A, Filogna S, Latrofa F, Sgandurra G (2024) Pathways to democratized healthcare: Envisioning human-centered AI-as-a-service for customized diagnosis and rehabilitation. *Artif Intell Med* 151:102850
291. Raza MM, Venkatesh KP, Kvedar JC (2024) Generative AI and large language models in health care: pathways to implementation. *npj Digit Med* 7(1):62
292. Hua D, Petrina N, Young N, Cho J-G, Poon SK (2024) Understanding the factors influencing acceptability of AI in medical imaging domains among healthcare professionals: A scoping review. *Artif Intell Med* 147:102698
293. Ayorinde A et al (2024) Health care professionals' experience of using AI: systematic review with narrative synthesis. *J Med Internet Res* 26:e55766
294. Mucci A, Green WM, Hill LH (2024) Incorporation of artificial intelligence in healthcare professions and patient education for fostering effective patient care, *New Directions for Adult and Continuing Education*, vol. no. 181, pp. 51–62, 2024
295. Božić V (2024) Artificial Intelligence in nurse education. *Engineering applications of artificial intelligence*. Springer, pp 143–172

Footnotes

1. Abbreviations: DICOM-WADO = Digital Imaging and Communications in Medicine – Web Access to DICOM Objects; HL7 FHIR = Health Level Seven Fast Healthcare Interoperability Resources; IHE = Integrating the Healthcare Enterprise; PACS = Picture Archiving and Communication System.

Figures

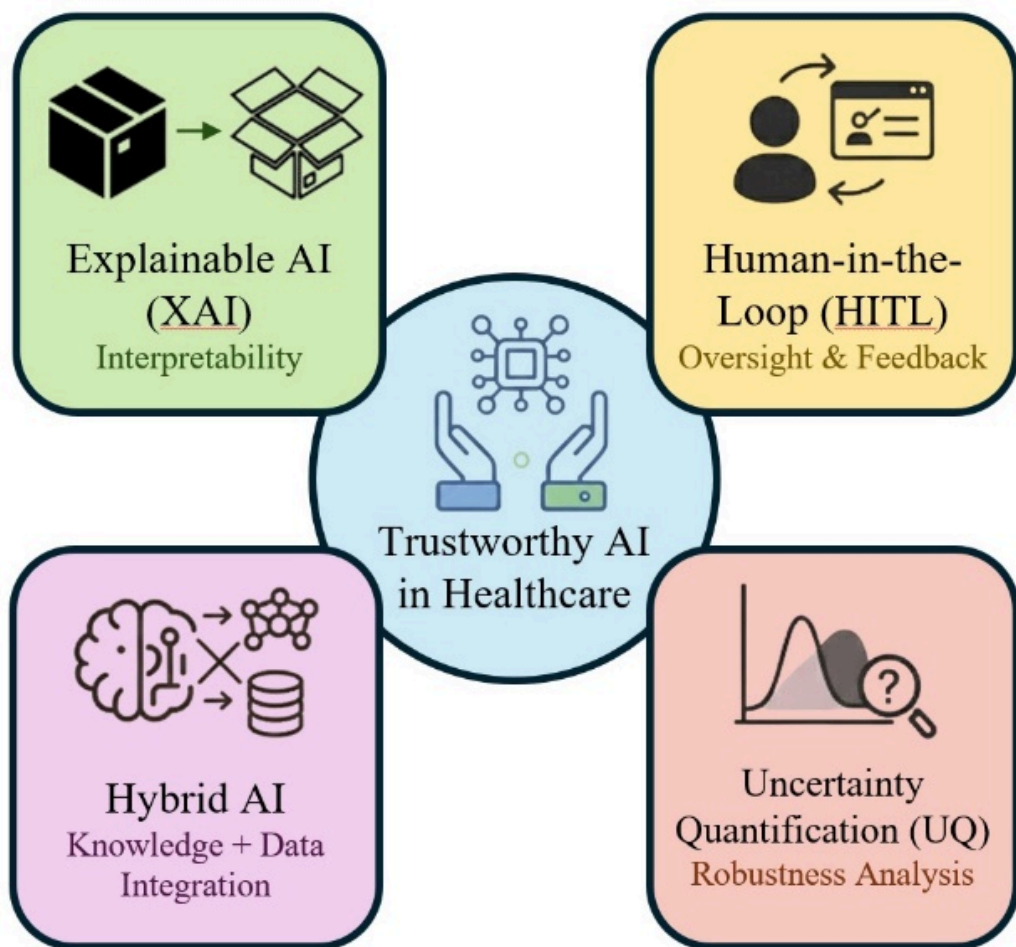


Figure 1

Human-Centered Pathways to Trustworthy AI in Healthcare.

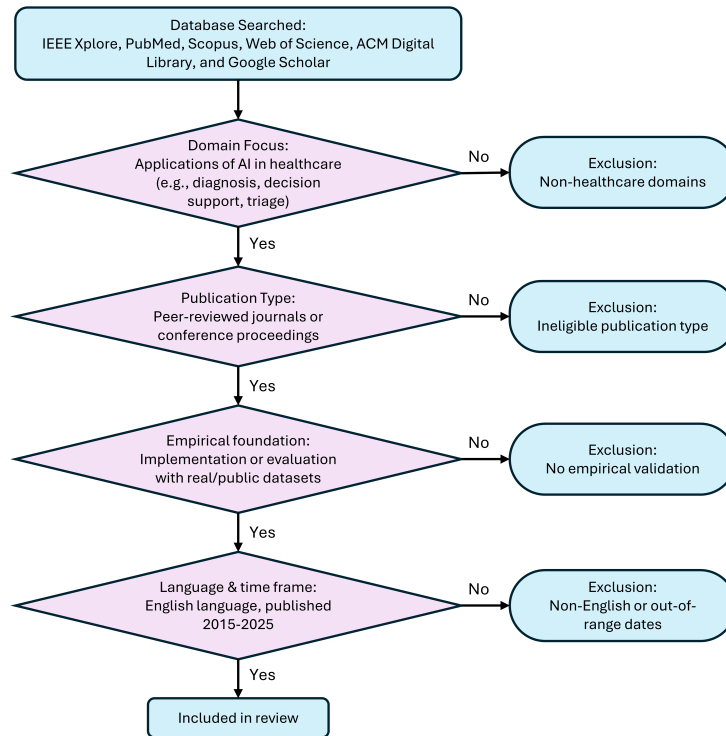


Figure 2

Systematic study of inclusion and exclusion of research articles.

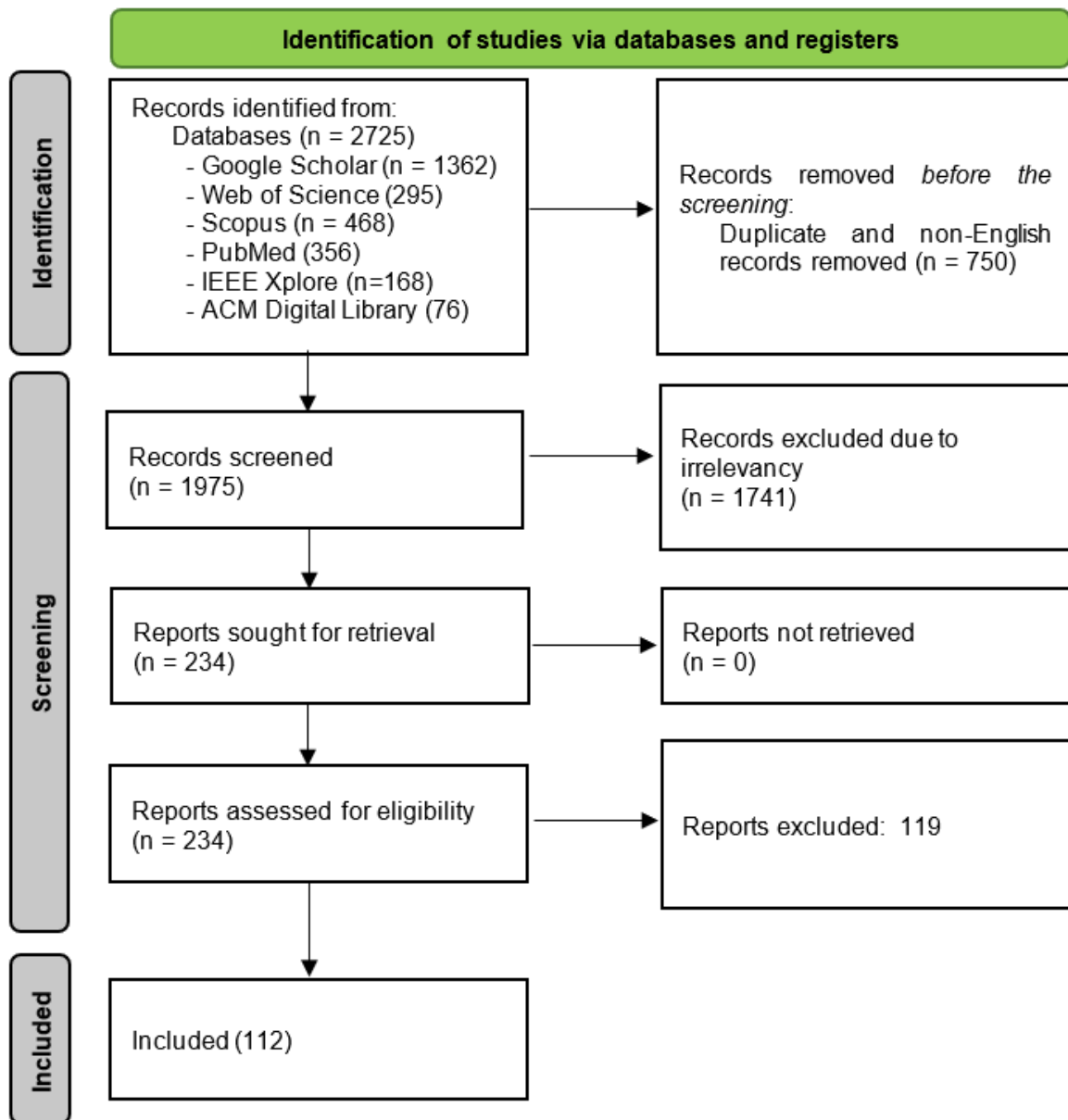
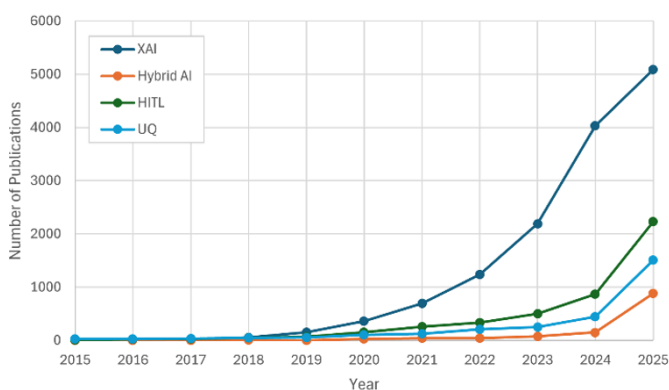
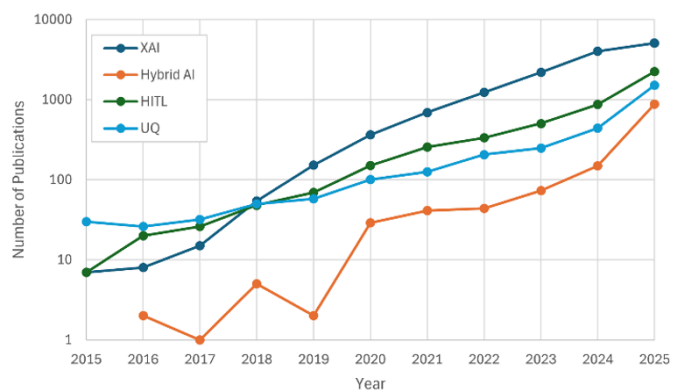


Figure 3

Study selection process flow diagram.



(a)



(b)

Figure 4

Publication trend chart (2015–2025) for four pathways based on keyword searches. The annual number of included studies is displayed on (a) a linear scale and (b) a logarithmic scale to highlight comparative growth.

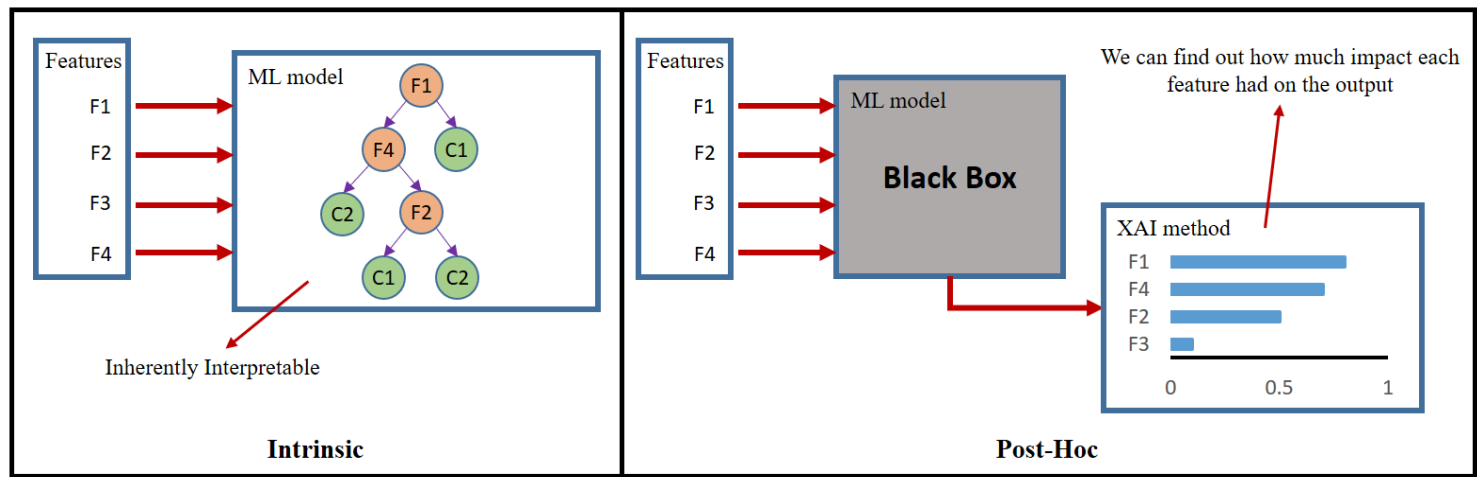


Figure 5

The difference between an intrinsically explainable model (e.g., a decision tree) and a post-hoc interpretable model. "F" stands for feature, and "C" stands for class.

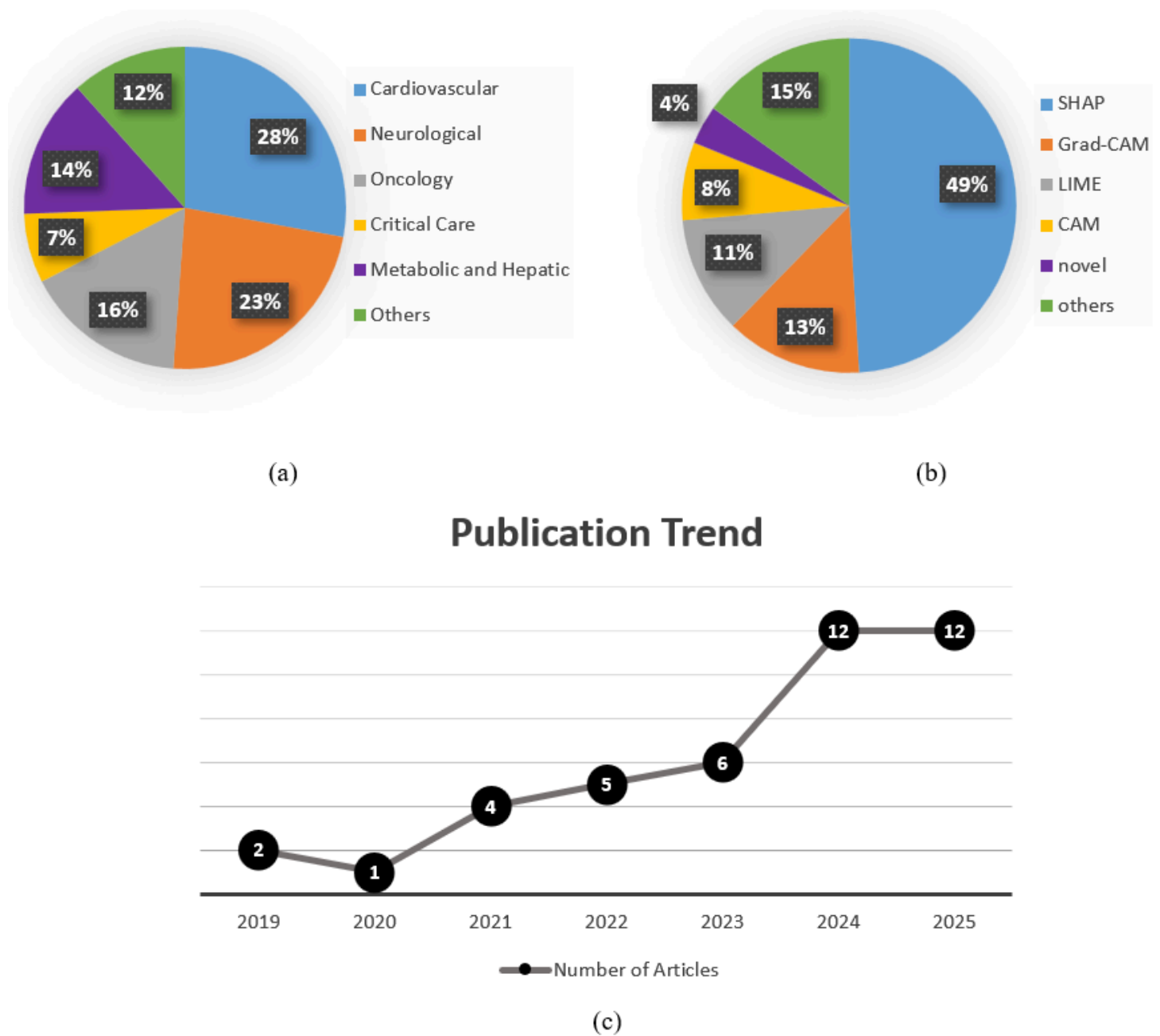


Figure 6

A statistical summary of the reviewed literature on XAI in healthcare.

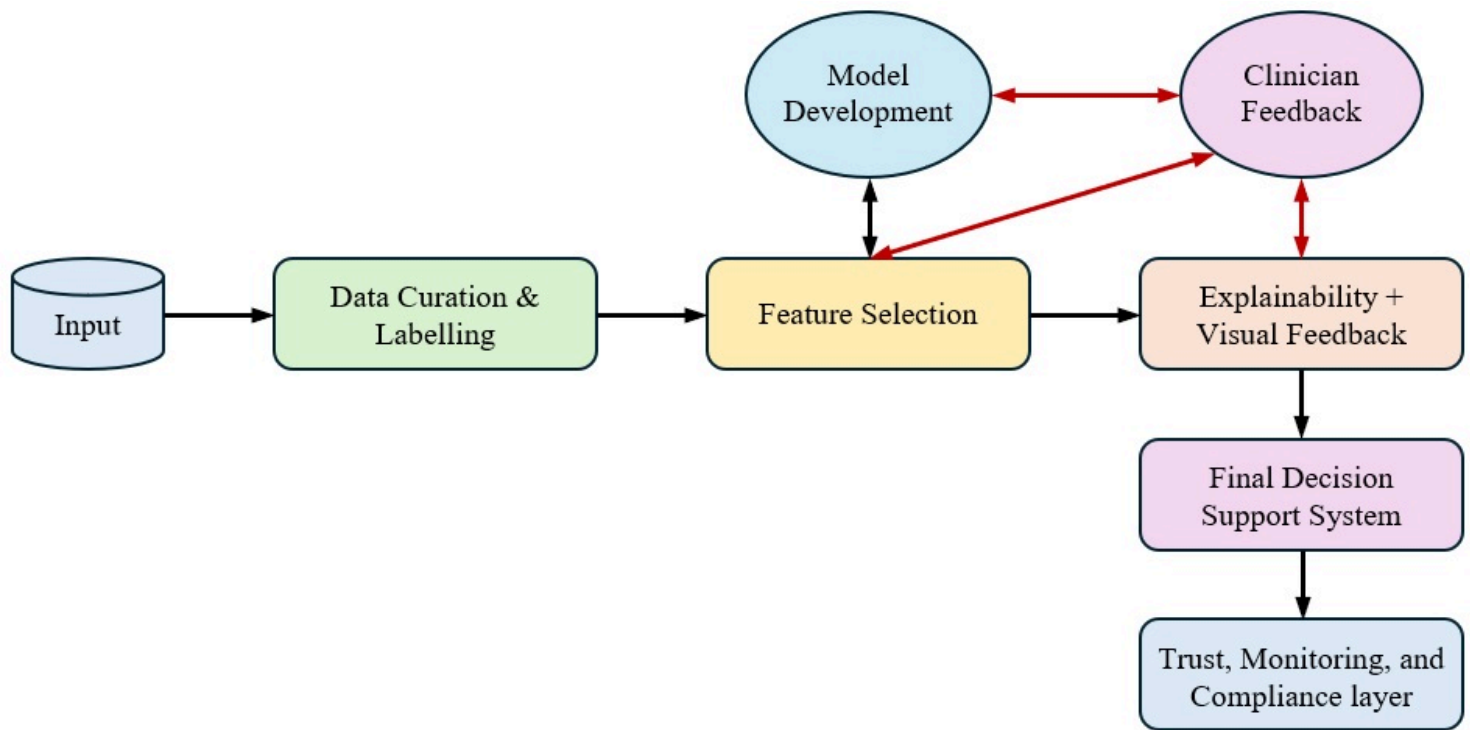


Figure 7

A conceptual model of the HITL workflow for developing a clinical decision support system.

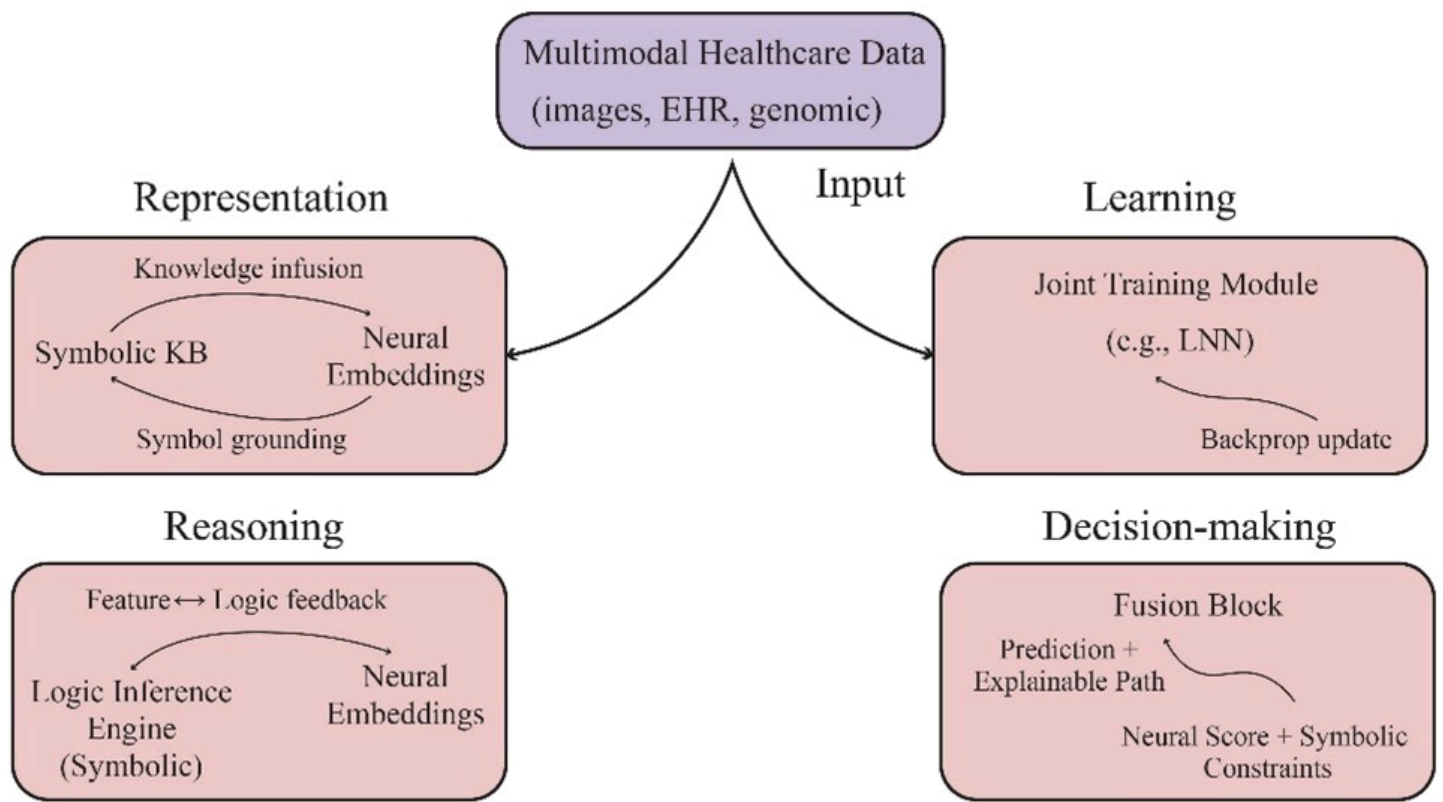


Figure 8

Hybrid AI workflow: multimodal healthcare data pass through Representation (symbolic↔neural), Learning (joint training), Reasoning (bidirectional inference), and Decision-making (fusion for explainable outputs).

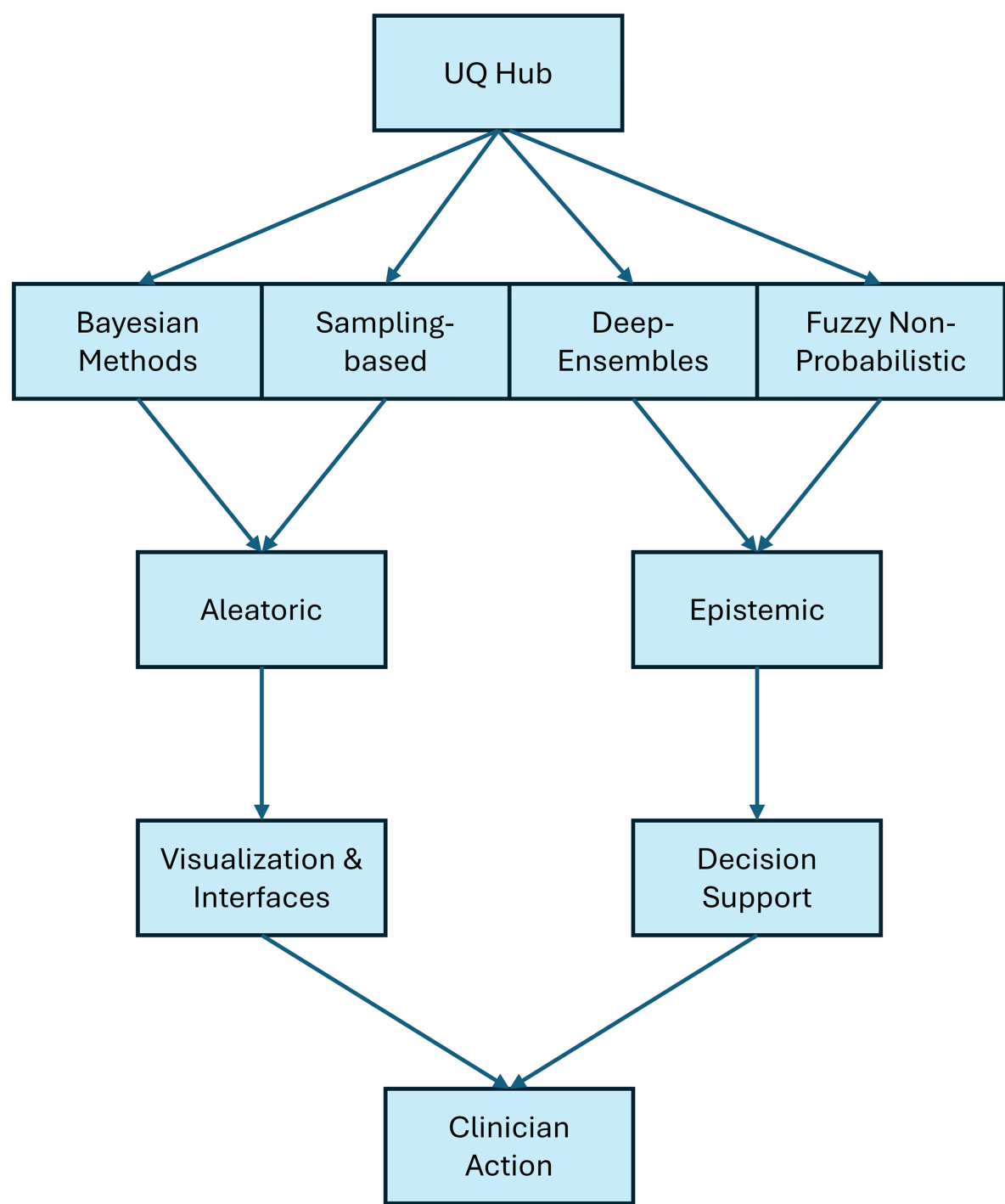


Figure 9

Simplified UQ workflow for healthcare AI: a central UQ Hub feeds four method categories.