

Explainable Artificial Intelligence in Radiological Imaging: Bridging Computational Explanations and Radiological Reasoning

Kaneeka Vidanage

General Sir John Kotelawala Defense University

K. L.Y.S. Karandana

National Institute of Business Management

Bahasuru Nayanakantha

General Sir John Kotelawala Defense University

Smita Nirkhi

smi.ta.singh@sitnagpur.siu.edu.in

Symbiosis Institute of Technology, Nagpur Campus, Symbiosis International (Deemed University)

Sabyasachi Bhattacharyya

Assam Engineering College

Systematic Review

Keywords: Explainable artificial intelligence, Radiological imaging, Clinical translation, Explainability evaluation, Clinical AI governance

Posted Date: October 22nd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11066503/v1>

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

Additional Declarations: No competing interests reported.

Abstract

Explainable artificial intelligence (XAI) is increasingly applied to radiological imaging to improve the transparency of artificial intelligence systems. However, the clinical usefulness of many XAI approaches remains uncertain because computational explanations are not always aligned with radiological reasoning, user requirements, workflow, or governance needs. This study aimed to map recent XAI research in radiological imaging, identify gaps in the clinical translation of explanations, and develop an integrative framework for clinically oriented XAI. A systematic scoping review was conducted following the Arksey and O'Malley framework and subsequent refinements, with reporting guided by PRISMA-ScR. MEDLINE/PubMed, Scopus, Web of Science Core Collection, IEEE Xplore, ACM Digital Library, and Embase via Ovid were searched for peer-reviewed studies published between January 2020 and June 2026. Eligible studies applied identifiable XAI approaches to radiological or clinically relevant medical imaging and reported information concerning explanations or their evaluation. Evidence was descriptively mapped and thematically synthesized according to XAI methods, clinical applications, explanation evaluation, clinical translation, usability, workflow integration, and governance. The evidence demonstrated substantial methodological diversity across imaging modalities, clinical tasks, AI architectures, and XAI approaches. This includes gradient-based, perturbation and feature-based, attention-based, prototype and concept-based, counterfactual, and inherently interpretable approaches. However, explanation evaluation was predominantly model-centered, with comparatively limited assessment of clinical understanding, relevance, usefulness, and integration into radiological decision-making. A recurring clinical translation gap was identified between computational outputs, such as heatmaps and feature-attribution representations, and the anatomical, morphological, and diagnostic concepts used in radiological reasoning. Workflow, usability, human oversight, and governance were also inconsistently integrated into XAI evaluation. These findings indicate that the principal challenge in radiological XAI is not the absence of explanation techniques but the lack of an integrated process for selecting, translating, evaluating, and documenting explanations according to clinical context. To address this gap, the Explainability–Radiology AI Integration (XRAI) Framework was developed, comprising six linked layers: clinical context profiling, XAI method selection, explanation generation, clinical translation, usability validation and feedback, and regulatory compliance logging. XRAI provides a conceptual and a transparent basis for connecting computational explainability with clinically meaningful and accountable radiological practice. Further empirical evaluation is required to establish its effectiveness across clinical tasks, imaging modalities, user groups, and deployment environments.

1 Introduction

Artificial intelligence (AI) is increasingly used to support the interpretation of medical images [1]. Radiology has been an important area for this development because clinical practice generates large volumes of structured image data from computed tomography (CT), magnetic resonance imaging (MRI), radiography, positron emission tomography (PET), ultrasound, and other modalities [2], [3]. Deep learning models have demonstrated strong performance in several radiological tasks, including lesion detection, disease classification, and image-based risk assessment [4], [5]. These results show the technical potential of AI in diagnostic imaging. However, high predictive performance alone does not ensure safe or effective clinical use [6].

A major challenge is that many high-performing AI models are difficult to interpret. Deep neural networks learn complex relationships through multiple nonlinear layers, making it difficult for clinicians to determine why a particular prediction was produced [7], [8]. This creates an important problem in radiology. A radiologist may receive a correct AI prediction but still need to know which image findings contributed to that prediction, whether those findings are clinically plausible, and whether the model may have relied on irrelevant features [9]. This information is particularly important when AI is used in situations where an incorrect result can lead to delayed diagnosis, unnecessary investigation, or inappropriate treatment [10], [11]. Explainable artificial intelligence (XAI) has therefore emerged as an important approach for improving the transparency of medical AI systems.

A wide range of XAI methods have been applied to medical imaging. Gradient-based approaches, including Gradient-weighted Class Activation Mapping (Grad-CAM) and related methods, are commonly used to identify image regions associated with a model prediction [12]. Perturbation- and feature-based methods such as Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP) provide alternative forms of attribution. Attention-based methods, prototype-based approaches, concept-based explanations, and counterfactual methods have also been explored [13], [14]. These techniques can provide useful information about model behavior, but the presence of an explanation does not necessarily mean that the explanation is clinically useful [15].

The distinction between technical explanation quality and clinical usefulness is particularly important in radiology [16]. Many XAI studies evaluate explanations using measures related to model behavior or visual characteristics. In contrast, fewer studies examine whether radiologists can understand the explanation, whether it corresponds to clinically meaningful findings, or whether it improves decision-making [17], [18]. A visually plausible heatmap, for example, may show where a model responded strongly without explaining why that region matters clinically. Similarly, a feature-importance score may identify influential image features without indicating how those features relate to

the anatomical and diagnostic concepts used in routine radiological reporting [19]. Thus, an explanation can be technically faithful to a model while remaining difficult to interpret or act upon in clinical practice [20].

This problem reflects a broader clinical translation gap between XAI outputs and radiological reasoning. Radiologists communicate findings using structured clinical concepts that include anatomical location, density or signal characteristics, lesion margins, enhancement patterns, distribution, and differential diagnosis [10], [18]. In contrast, many XAI systems produce heatmaps, activation maps, feature-importance rankings, attention scores, or other computational artefacts [21]. These outputs are valuable for analyzing model behavior, but they are not, by themselves, clinical statements. A heatmap indicating activation in the left perihilar region, for example, does not directly communicate whether the model is responding to consolidation, a mass, vascular anatomy, or another feature relevant to diagnosis [19], [21]. The gap between these two forms of information has received limited systematic attention in existing literature [22], [23].

Clinical deployment also requires consideration of workflow, human oversight, safety, and governance [9]. The evaluation of healthcare AI therefore needs to extend beyond model accuracy and include clinical context, stakeholder involvement, usability, and evidence of safety and utility [24], [25]. Regulatory developments have further increased the need for transparent and auditable AI systems [26]. Frameworks and principles addressing responsible AI, including the World Health Organization guidance on ethics and governance of AI for health and emerging regulatory requirements such as the European Union Artificial Intelligence Act, emphasize transparency, human oversight, accountability, and appropriate risk management [24], [27]. Yet regulatory and governance considerations remain poorly integrated into much of the XAI literature in diagnostic imaging.

Taken together, these issues suggest that the main challenge is no longer simply how to generate an explanation. The more important question is how to select, generate, translate, evaluate, and document an explanation that is appropriate for a specific radiological task and its intended user [28]. Current research tends to address individual components of this problem rather than providing an integrated approach that connects XAI methods with clinical context, radiological language, usability, and governance [9], [21], [23].

To address this gap, the present study conducted a systematic scoping review of peer-reviewed research on XAI in radiological imaging published between January 2020 and June 2026. The review maps the use of XAI methods across imaging modalities and clinical tasks and examines how explanations have been evaluated.

Attention is given to the relationship between technical explanation measures and clinical evaluation, the translation of XAI outputs into radiological concepts, and the consideration of regulatory requirements. The synthesis is then used to identify structural gaps that limit the clinical integration of XAI. Based on these findings, we propose the Explainability–Radiology AI Integration (XRAI) Framework, a context-aware framework designed to connect computational explanations with clinical radiological reasoning. XRAI consists of six linked layers. Clinical context profiling, XAI method selection, explanation generation, clinical translation, usability validation and feedback, and regulatory compliance logging. The framework is intended to move XAI from the production of isolated explanation artefacts toward a more structured process in which explanations are selected according to clinical context, translated into clinically meaningful information, evaluated by intended users, and documented for accountability.

The contribution of this study is therefore threefold. First, it provides a cross-method and cross-modality map of recent XAI research in diagnostic imaging. Second, it identifies the clinical translation gap as a central limitation in moving XAI outputs from technical research settings into radiological practice. Third, it introduces and expert-evaluates the XRAI Framework as a structured approach for addressing this gap while incorporating usability and governance considerations. The framework is intended not as a replacement for individual XAI techniques, but as an integration layer that helps determine when and how different explanation approaches should be used in clinical radiology.

2 Methods

2.1 Study design

A scoping review was conducted to map the current evidence on explainable artificial intelligence (XAI) in radiological imaging and to identify gaps relevant to clinical translation. A scoping review was selected because the literature spans computer science, biomedical engineering, clinical radiology, and health informatics and includes heterogeneous study designs, XAI methods, imaging modalities, and evaluation approaches [29], [30].

The review followed the methodological framework of Arksey and O'Malley, with subsequent refinements proposed by Levac et al. [31], [32]. Reporting was guided by the Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) [33], with additional reporting elements informed by PRISMA 2020 [34].

The review protocol was not prospectively registered. However, the review question, eligibility criteria, database selection, data extraction categories, and synthesis strategy were defined before the literature search. No patient-level data or clinical records were used.

2.2 Review question and scope

The review was structured using the Population–Concept–Context (PCC) framework recommended for scoping reviews [35].

Population

Studies using radiological or medical imaging data from any patient population or demographic group has been considered. Eligible modalities included radiography, computed tomography (CT), magnetic resonance imaging (MRI), positron emission tomography (PET), single-photon emission computed tomography (SPECT), and diagnostic ultrasound.

Concept

The use of XAI to explain predictions or behavior of artificial intelligence and deep learning systems applied to radiological images. XAI approaches included gradient-based methods such as Grad-CAM and its variants, perturbation and feature-based methods such as LIME and SHAP, attention-based methods, prototype- and concept-based approaches, counterfactual explanations, and inherently interpretable model architectures [18, 20].

Context

Clinical and research applications of radiological imaging, including diagnosis, screening, disease characterization, prognosis, and clinical decision support.

The review addressed three questions:

1. Which XAI methods have been applied to radiological imaging between 2020 and 2026?
2. How have the quality and clinical relevance of XAI explanations been evaluated?
3. What gaps exist between current XAI practices and the practical requirements of clinical radiology?

These questions were used to guide both the descriptive mapping of the literature and the subsequent development of the XRAI framework.

2.3 Information sources and search strategy

A systematic literature search was performed in six databases: MEDLINE/PubMed, Scopus, Web of Science Core Collection, IEEE Xplore, ACM Digital Library, and Embase via Ovid. The databases were selected to provide coverage across clinical medicine, biomedical engineering, computer science, and health informatics.

2.3.1 Search strategy and keywords

The search strategy was developed around three main concept blocks: explainable artificial intelligence, medical or radiological imaging, and clinical or diagnostic application. Each block included related terms and commonly used alternative expressions. Examples of search terms included explainable artificial intelligence, XAI, explainability, interpretability, interpretable AI, radiology, diagnostic imaging, medical imaging, computed tomography, magnetic resonance imaging, radiography, mammography, PET, and SPECT.

Within each concept block, related terms were combined using the Boolean operator OR. The major concept blocks were then combined using AND. This approach was used to broaden retrieval within each concept while maintaining the required relationship between explainability, imaging, and clinical application.

Search strings were developed iteratively using known-item testing against relevant publications identified before the search. For MEDLINE/PubMed and other indexed databases, controlled vocabulary and Medical Subject Headings (MeSH) were incorporated where applicable. Truncation, phrase searching, and database-specific field tags were used according to the syntax supported by each database. The search strategy was adapted for each database rather than applying a single unmodified query across all sources.

The search strategy was structured around three main concepts: XAI, medical imaging, and clinical application. Figure 1 shows how these concepts were combined across the databases. The database-specific structure was used to maintain the same overall scope while adapting the search terms to each database.

Figure 1 Full search strings deployed across respective databases.

2.4 Eligibility criteria

Studies were included when they met all the following criteria:

- published as a peer-reviewed journal article between 1 January 2020 and 30 June 2026;
- available in English with accessible full text;
- applied at least one identifiable XAI method, either post hoc or inherent to the model;
- used radiological or medical imaging data from a modality relevant to clinical radiology; and
- reported information about the explanation, its behavior, or its evaluation rather than reporting predictive performance alone.

Studies were excluded if they were conference abstracts, posters, editorials, commentaries, opinion articles, book chapters, or letters. Used only non-imaging clinical data; applied AI to radiological images without addressing interpretability or explainability; were available only as preprints; or lacked accessible as full text.

2.5 Study selection

All retrieved records were imported into Rayyan for screening and duplicate management [36]. Duplicate records identified automatically were manually checked before screening.

Study selection was conducted in two stages. First, two reviewers independently screened titles and abstracts against the predefined eligibility criteria. Records that could not be clearly excluded at this stage were assessed in full text. Second, the same reviewers independently evaluated the full texts of potentially eligible studies.

Disagreements were resolved through discussion. Where agreement could not be reached, a third reviewer was available to make the final decision. Inter-rater agreement was assessed using Cohen’s kappa (κ), with $\kappa \geq 0.70$ considered acceptable [37]. The screening process resulted in the final set of studies included in the review. Figure 2 summarizes the identification, screening, eligibility, and inclusion stages used to obtain the final evidence base. The final included studies were then used for structured data extraction and subsequent descriptive and thematic synthesis.

Figure 2 PRISMA-ScR flow diagram detailing the study selection process

2.6 Data extraction and charting

A structured data extraction form was developed and pilot-tested using ten randomly selected included studies. One reviewer extracted the data, and a second reviewer independently verified the extracted information. Differences were resolved through discussion.

The following information was extracted:

- study characteristics, including publication year, country, and study design;
- imaging modality and anatomical region;
- clinical task, including classification, detection, localization, segmentation, prognosis, and risk prediction;
- AI model architecture and training approach;
- XAI method and method category;
- explanation type and evaluation approach;
- dataset characteristics and source;
- main findings related to XAI;
- reported limitations relevant to explainability or clinical translation; and
- explicit statements concerning clinical applicability or deployment.

The extraction categories are summarized in Table 1.

Table 1
Data extraction categories and variables charted from the included studies.

Field Category	Variables Extracted
Study identification	Authors, year, journal, country
Study design	Type; dataset origin (retrospective/prospective/public benchmark)
Imaging characteristics	Modality; anatomical region; acquisition parameters if reported
Clinical task	Classification, detection, segmentation, prognosis, risk prediction
AI model	Architecture type; pre-trained or de novo training
XAI method	Specific technique(s); category (ante-hoc/post-hoc; local/global; agnostic/specific)
Explanation evaluation	Assessment method (quantitative metrics, radiologist review, user study, none)
Dataset	Public or private; approximate size; geographic source
Key findings	Principal reported outcomes related to XAI
Limitations	Author-reported limitations relevant to explainability or clinical translation
Clinical orientation	Explicit statement of suitability or intent for clinical deployment

Table 2 summarizes the characteristics of the included studies by imaging modality. Figure 3 presents the same modality distribution visually, making the relative contribution of each imaging area easier to compare. Chest imaging and neurological imaging represented the largest groups in the included evidence base.

Table 2
Dataset Characteristics of Included Studies by Imaging Modality

Imaging modality	Studies, n (%)	Primary clinical tasks	Common benchmark/reference datasets	Dataset source	Typical dataset size	Geographic representation
Chest (X-ray and thoracic CT)	71 (38%)	Classification, detection/localization	ChestX-ray14; RSNA Pneumonia Detection Challenge; NIH CXR; COVID-19 Radiography Database; The Cancer Imaging Archive (TCIA)	Predominantly public	1,000–112,120 images	USA, China, Europe, India
Neurological (brain MRI)	52 (28%)	Classification, segmentation, prognosis	BraTS; ADNI; IXI; OASIS; private institutional cohorts	Public and private (mixed)	500–20,000 scans	USA, Germany, Netherlands, China
Breast (mammography/MRI)	26 (14%)	Classification, detection	INbreast; VinDr-Mammo; CBIS-DDSM; RSNA Screening Mammography; private screening datasets	Public and private (mixed)	1,000–54,000 examinations	USA, UK, Vietnam, EU
PET/SPECT (functional imaging)	6 (3%)	Classification, prognosis	ADNI-PET; private nuclear medicine institutional archives	Predominantly private	100–2,500 scans	USA, Germany, South Korea
Musculoskeletal (X-ray/MRI)	8 (4%)	Classification, detection, segmentation	MURA; private orthopedic cohorts	Public and private (mixed)	500–40,000 radiographs	USA, China, Australia
Abdominal/pelvic (CT/MRI)	14 (7%)	Classification, segmentation, detection	LiTS; Kits; Medical Segmentation Decathlon; private institutional cohorts	Public and private (mixed)	200–10,000 scans	USA, Europe, China
Cardiac imaging (echo/CT/MRI)	5 (3%)	Classification, segmentation	ACDC; private echocardiography archives	Public and private (mixed)	150–5,000 studies	France, USA, UK
Other/multimodal (ultrasound, hybrid)	5 (3%)	Classification, detection	BUSI; multimodal private cohorts	Predominantly private	100–3,000 images	Mixed/multi-site
Total	187 (100%)	—	—	—	—	—

Figure 3 depicts the visualized distribution of imaging modalities to get much firm insight.

The included studies also differed in their clinical objectives. Figure 4 shows the distribution of studies across the main clinical tasks. Classification, detection and localization formed the largest task groups, while segmentation, prognosis and risk prediction were less frequently represented.

This distribution provides the clinical context for interpreting the XAI methods identified in the next stage of the synthesis. No formal risk-of-bias or methodological quality assessment was performed because the purpose of the review was to map the evidence base and identify patterns across a heterogeneous body of literature rather than estimate a pooled effect [30], [38].

2.7 Data synthesis

The evidence was synthesized in two stages.

First, descriptive mapping was used to summarize the distribution of studies by publication year, geographic location, imaging modality, clinical tasks, AI architecture, XAI method, and explanation evaluation approach. This provided an overview of how XAI has been applied across radiological imaging. Second, qualitative thematic synthesis was used to identify recurring limitations and gaps related to explanation evaluation, clinical relevance, usability, workflow integration, and regulatory considerations. Reported study limitations and descriptions of clinical use were coded and grouped into broader themes. The synthesis followed the general principles of thematic analysis described by Thomas and Harden, including coding, development of descriptive themes, and generation of higher-level interpretations [39]. The two stages were then integrated to identify areas where the technical capabilities of existing XAI methods did not adequately address

the requirements of clinical radiology. Attention was given to four dimensions: method selection, explanation evaluation, clinical translation, and governance.

The resulting evidence synthesis was used to inform the design of the Explainability–Radiology AI Integration (XRAI) Framework. XRAI was developed as a transparent framework rather than as a new XAI algorithm. Its purpose is to provide a structured pathway linking clinical context to XAI method selection, explanation generation, translation into clinically meaningful information, usability evaluation, and regulatory documentation.

3 Results

The evidence synthesis identified a heterogeneous body of research examining explainability in artificial intelligence systems applied to radiological imaging. The discussion literature varied substantially in imaging modality, clinical task, AI architecture, XAI technique, explanation format, and approach used to evaluate explanations. This heterogeneity reflects the multidisciplinary nature of radiological XAI research but also limits direct comparison between individual approaches. The findings were therefore synthesized across four main dimensions as application of XAI methods, evaluation of explanations, clinical relevance and translation, and workflow and governance considerations.

3.1 Characteristics of the evidence base

The included studies covered a broad range of radiological applications, including classification, detection, localization, segmentation, prognosis, and risk prediction. CT, MRI, radiography, and other diagnostic imaging modalities were represented across the literature. The clinical objectives also varied from automated identification of specific abnormalities to broader disease characterization and prediction tasks. Most studies evaluated explainability in association with deep learning models, particularly models designed for image classification, detection, or segmentation [15], [19], [38]. The use of XAI was generally positioned as an additional interpretability component rather than as an integral part of the model design [40]. Consequently, many studies first optimized predictive performance and subsequently applied an XAI technique to provide an explanation of the resulting prediction.

The evidence also demonstrated considerable variation in the purpose assigned to XAI. In some studies, explanations were primarily used to inspect model behavior and identify image regions associated with predictions. Other studies used explanations to investigate potential model errors, assess whether the model attended clinically relevant structures, or provide additional information intended to support clinical interpretation. This variation indicates that explainability is being used for multiple purposes that are not necessarily equivalent and may require different forms of explanation and evaluation.

3.2 XAI methods and explanation approaches

Gradient-based methods represented an important category of approaches used in radiological imaging. Grad-CAM and related techniques were commonly used to generate spatial heatmaps highlighting image regions associated with a model prediction [41]. These approaches were particularly suitable for image-based classification and localization tasks because they provided a visual representation that could be superimposed on the original image. Perturbation- and feature-based approaches, including LIME and SHAP, provided alternative forms of attribution by estimating the contribution of image regions or features to individual predictions. These methods offered a more explicit representation of feature importance, but their clinical interpretation depended on how the identified features were defined and presented [42]. Attention-based approaches were also used to provide information about regions receiving greater model attention. Prototype-based and concept-based approaches provided alternative mechanisms for connecting model predictions with representative examples or higher-level concepts. Counterfactual approaches further extended explainability by examining how changes to an input might alter a model prediction. Inherently interpretable architecture represented a different direction in which interpretability was incorporated into the model rather than generated entirely as a post-hoc explanation.

Overall, the literature indicates that radiological XAI remains strongly oriented toward post-hoc explanation techniques. Heatmaps, activation maps, feature-attribution representations, and attention visualizations were frequently used to communicate model behavior. Although these approaches can improve transparency at the computational level, their outputs do not necessarily correspond directly to the anatomical, morphological, or diagnostic concepts used in radiological reasoning. The distribution of XAI methods also varied across imaging modalities and clinical tasks. Table 3 summarizes these patterns and shows how the main XAI method families were used across the included studies.

Table 3
XAI method families **Part A. XAI Method Family × Imaging Modality**

XAI method family	Chest	Neurological	Breast	PET/SPECT	Musculoskeletal	Other	Row total n (%)
Gradient-based (Grad-CAM, Score-CAM, Grad-CAM++, LayerCAM)	38 (42%)	24 (27%)	14 (16%)	2 (2%)	5 (6%)	7 (8%)	90 (48.1%)
SHAP (Shapley additive explanations)	11 (32%)	12 (35%)	5 (15%)	2 (6%)	2 (6%)	2 (6%)	34 (18.2%)
Transformer attention (attention rollout, cross-attention visualisation)	11 (39%)	13 (46%)	2 (7%)	1 (4%)	0 (0%)	1 (4%)	28 (15.0%)
LIME (local interpretable model-agnostic explanations)	9 (43%)	2 (10%)	5 (24%)	1 (5%)	1 (5%)	3 (14%)	21 (11.2%)
Other (counterfactual, prototype-based, concept-based, inherently interpretable)	2 (14%)	1 (7%)	0 (0%)	0 (0%)	0 (0%)	11 (79%)	14 (7.5%)
Column total n (%)	71 (38.0%)	52 (27.8%)	26 (13.9%)	6 (3.2%)	8 (4.3%)	24 (12.8%)	187 (100%)

Part B. XAI Method Family × Clinical Task Type

XAI method family	Classification	Detection and localization	Segmentation	Prognosis and risk stratification	Other	Row total n (%)
Gradient-based (Grad-CAM, Score-CAM, Grad-CAM++, LayerCAM)	52 (58%)	22 (24%)	10 (11%)	4 (4%)	2 (2%)	90 (48.1%)
SHAP (Shapley additive explanations)	16 (47%)	7 (21%)	4 (12%)	5 (15%)	2 (6%)	34 (18.2%)
Transformer attention (attention rollout, cross-attention visualisation)	16 (57%)	9 (32%)	3 (11%)	0 (0%)	0 (0%)	28 (15.0%)
LIME (local interpretable model-agnostic explanations)	10 (48%)	5 (24%)	4 (19%)	2 (10%)	0 (0%)	21 (11.2%)
Other (counterfactual, prototype-based, concept-based, inherently interpretable)	3 (21%)	2 (14%)	5 (36%)	2 (14%)	2 (14%)	14 (7.5%)
Column total n (%)	97 (51.9%)	45 (24.1%)	26 (13.9%)	13 (7.0%)	6 (3.2%)	187 (100%)

Gradient-based methods were the most frequently represented method family, followed by SHAP, transformer attention, LIME, and other approaches. The distribution also varied by clinical task, indicating that XAI use was not uniform across radiological applications.

3.3 Evaluation of explanations

Substantial variation was observed in how explanations were evaluated. Several studies focused primarily on properties of the generated explanation or its relationship to model behavior. Such evaluations included visual inspection, attribution-based measures, localization characteristics, perturbation analysis, and comparisons between explanation outputs and image regions associated with known findings. These approaches provide important information about whether an explanation reflects aspects of model behavior. However, technical evaluation does not necessarily establish whether an explanation is understandable, clinically meaningful, or useful to a radiologist. An explanation may accurately represent the internal behavior of a model while still failing to communicate information in a form that supports clinical reasoning. A smaller but important body of research considered human or expert evaluation of explanations. These studies provided evidence that clinical usefulness depends not only on whether an explanation identifies an influential image region but also on whether the identified region corresponds to a meaningful radiological finding. The distinction between model-focused evaluation and user-focused evaluation therefore emerged as an important feature of the evidence base.

The findings suggest that explanation quality should not be treated as a single technical property. An explanation may possess computational validity while having limited clinical interpretability, usability, or decision-support value. This indicates the need for evaluation approaches that consider both the behavior of the underlying model and the requirements of its intended clinical users. The differences in

evaluation approaches are important because they show whether XAI was assessed mainly from a technical perspective or also from a clinical perspective. Figure 5 summarizes the evaluation approaches used across the included studies and provides a visual view of this distribution.

The distribution shows the greater use of model-centered evaluation compared with human or clinical evaluation. This finding supports the need to assess not only whether an explanation reflects model behavior, but also whether it is understandable and useful to clinical users.

3.4 Clinical translation gap

A recurring finding across the evidence was the limited connection between computational XAI outputs and the structured language of radiological interpretation. Radiologists generally describe findings using concepts related to anatomy, morphology, density or signal characteristics, lesion margins, distribution, enhancement, and diagnostic interpretation. Many XAI methods, in contrast, produce heatmaps, activation maps, attention scores, or feature-attribution values. These outputs can indicate where or which features influenced a prediction but do not necessarily explain the clinical significance of those features. For example, highlighting a region within an image may demonstrate that the model considered the region important without identifying the corresponding anatomical structure or pathological finding. The explanation therefore remains at the level of computational attribution rather than being translated into a clinically meaningful statement.

This distinction represents an important gap between explainability and clinical translation. The generation of an explanation does not by itself establish that the explanation can be incorporated into radiological reasoning. The evidence therefore indicates a need for an intermediate translation process through which computational explanations can be related to clinically recognizable radiological concepts.

3.5 Workflow, usability, and governance considerations

The literature also demonstrated limited and inconsistent consideration of the wider clinical environment in which XAI systems would operate. While model performance and explanation generation were frequently examined. Fewer studies provided detailed consideration of workflow integration, intended users, usability, human oversight, or the role of explanations in clinical decision-making. This is particularly important because the value of an explanation depends on its intended purpose and user [43]. An explanation designed for model debugging may differ substantially from one intended to support a radiologist during diagnostic interpretation. Similarly, an explanation that is useful during retrospective research evaluation may not be appropriate for real-time clinical workflow. Governance and accountability considerations were also not consistently integrated with XAI evaluation. Transparent documentation of how an explanation was generated, why a particular XAI method was selected, and how the explanation was evaluated can become important when AI systems are used in clinical environments. Regulatory and responsible AI principles therefore require consideration alongside technical explainability rather than being treated as separate concerns.

3.6 Structural gaps identified from the synthesis

The combined descriptive and thematic synthesis identified four interconnected structural gaps in current radiological XAI research. First, XAI method selection is often insufficiently connected to clinical context. Different radiological tasks, imaging modalities, model architectures, and intended users may require different explanation approaches, yet method selection is frequently presented primarily as a technical decision. Second, explanation evaluation remains predominantly model centered. Technical measures can determine whether an explanation reflects model behavior but provide limited evidence regarding clinical understanding, usefulness, or impact on decision-making. Third, translation from computational explanation to radiological reasoning remains underdeveloped. Many explanations identify influential regions or feature without converting them into clinically meaningful anatomical, morphological, or diagnostic concepts. Fourth, usability, workflow integration, and governance are insufficiently integrated with explanation generation and evaluation.

This creates a separation between the technical production of explanations and the practical requirements of safe clinical deployment. Taken together, these findings indicate that the central limitation of current radiological XAI research is not the absence of explanation techniques. Rather, the limitation lies in the lack of an integrated process for determining which explanation is appropriate for a particular clinical context, translating the resulting explanation into clinically meaningful information, evaluating its usefulness with intended users, and maintaining appropriate documentation for accountability. These structural gaps provided the basis for the development of the Explainability–Radiology AI Integration (XRAI) Framework. Rather than introducing another XAI algorithm, the framework addresses the integration problem by connecting clinical context profiling, XAI method selection, explanation generation, clinical translation, usability validation, feedback, and regulatory compliance within a single structured pathway.

4 Explainability–Radiology AI Integration (XRAI) Framework

The findings of the evidence synthesis indicate that the principal challenge in radiological XAI is not the availability of explanation techniques but the absence of a structured process for integrating those techniques with clinical requirements. Existing approaches frequently address explanation generation as an isolated technical activity, while clinical context, radiological reasoning, usability, workflow integration, and governance are considered separately or inconsistently. In response to these gaps, the Explainability–Radiology AI Integration (XRAI) Framework was developed as a context-aware integration framework for connecting computational explainability with clinical radiological reasoning. XRAI is not proposed as a new XAI algorithm and does not replace existing explanation techniques. Instead, it provides an integration layer through which an appropriate XAI approach can be selected, generated, translated, evaluated, and documented according to the clinical context in which an AI system is intended to operate. The framework consists of six linked layers: clinical context profiling, XAI method selection, explanation generation, clinical translation, usability validation and feedback, and regulatory compliance logging. The six layers are connected to form a single workflow rather than six separate components. Figure 6 presents the overall structure of the XRAI Framework and shows how the process moves from clinical context to explanation selection, explanation generation, clinical translation, user evaluation, and regulatory documentation.

The Fig. 6 shows that clinical context is placed at the beginning of the process. The computational explanation is then translated into clinically meaningful information before it is evaluated by the intended users. Feedback from this evaluation can be used to refine earlier stages. Regulatory documentation runs across the process to support traceability and accountability. The following sections describe each layer of the framework and explain its role in the overall workflow.

4.1 Clinical context profiling

The first layer establishes the context in which the AI system and its explanation will be used. This layer recognizes that explainability requirements are not uniform across radiological applications. The appropriate form of explanation may depend on the imaging modality, anatomical region, clinical task, AI model, intended user, stage of the clinical workflow, and purpose for which the explanation is required. Clinical context profiling therefore defines the characteristics of the task before an XAI method is selected. Relevant factors include whether the system performs classification, detection, localization, segmentation, prognosis, or risk prediction; whether the output is intended for diagnostic support, screening, research, quality assurance, or model inspection; and whether the primary user is a radiologist or another clinical or technical stakeholder.

The purpose of this layer is to establish a clear relationship between the clinical problem and the explainability requirement. Rather than treating explanation generation as a generic post-processing step, XRAI positions the intended clinical context as the starting point of the explainability process.

4.2 XAI method selection

Following context profiling, the second layer determines the most appropriate explanation approach for the identified requirements. The framework accommodates gradient-based, perturbation- and feature-based, attention-based, prototype-based, concept-based, counterfactual, and inherently interpretable approaches. Method selection is therefore treated as a context-dependent decision rather than a purely technical preference. Characteristics of the underlying AI model, imaging modality, clinical task, expected explanation type, and intended user are considered when selecting an approach. For example, spatial localization may be particularly relevant when the clinical objective requires identification of an abnormal region, whereas feature- or concept-based explanations may be more appropriate when the objective requires understanding of higher-level diagnostic characteristics. The framework does not prescribe a single XAI technique for a particular task. Instead, it establishes a structured decision point at which the suitability of an explanation method can be considered against the requirements established during clinical context profiling. This approach also allows different XAI methods to be used for different clinical contexts without assuming that one method provides universally adequate explanations.

4.3 Explanation generation

The third layer concerns the generation of the computational explanation. At this stage, the selected XAI method is applied to the AI model and the relevant radiological input to produce an explanation associated with the model's prediction or output. Depending on the selected approach, the resulting explanation may take the form of a heatmap, activation map, feature attribution, attention representation, prototype comparison, concept representation, or counterfactual output. The explanation should retain an explicit association with the underlying model prediction so that the relationship between the prediction and the explanation can be examined. An important feature of XRAI is that explanation generation is not considered the final output of the process. The generated explanation represents an intermediate computational representation that must subsequently be interpreted in relation to the clinical context. This distinction addresses a recurring limitation identified in the reviewed literature, where the presence of an explanation is sometimes treated as sufficient evidence of explainability.

4.4 Clinical translation

The fourth layer represents the central integration component of the XRAI Framework. It addresses the gap between computational explanation outputs and the concepts used in radiological reasoning. At this stage, the computational explanation is related to clinically meaningful information such as anatomical location, morphological characteristics, distribution, density or signal properties, lesion characteristics, and other relevant radiological concepts. The objective is not simply to identify where a model responded strongly but to establish what the highlighted or attributed information represents in clinical terms. Clinical translation therefore converts an explanation from a model-oriented representation into information that can be interpreted within the relevant radiological context. For example, an activation region may be associated with a particular anatomical structure or radiological finding rather than being presented only as a colored region on an image. The resulting representation can then be considered in relation to the clinical task and the model's prediction. This layer is particularly important because a technically faithful explanation may still have limited clinical value if its meaning is not understandable to the intended user. XRAI therefore treats clinical translation as a distinct stage rather than if computational attribution automatically constitutes clinical explanation.

4.5 Usability validation and feedback

The fifth layer evaluates whether the translated explanation is understandable, clinically relevant, and useful to its intended users. Evaluation is therefore extended beyond model-centered explanation metrics to include human and clinical considerations. The evaluation may consider whether the explanation identifies clinically meaningful information, whether users can correctly understand its relationship to the AI prediction, whether it supports rather than unnecessarily complicates clinical reasoning, and whether it provides sufficient information for the intended decision-support context. The specific evaluation criteria may vary according to the clinical task and intended users.

Feedback obtained during this stage is incorporated into the framework as an iterative mechanism. If users identify that an explanation is difficult to interpret, clinically incomplete, misleading, or poorly aligned with workflow requirements, the findings can inform revision of the explanation method, translation process, or contextual requirements. This feedback mechanism recognizes that explainability is not necessarily achieved through a single design decision. The suitability of an explanation may become clearer only when it is evaluated by the users for whom it is intended.

4.6 Regulatory compliance logging

The sixth layer incorporates documentation and accountability requirements into the explainability process. It recognizes that clinical AI systems require more than an explanation presented to an end user. There is also a need for traceable information concerning how the explanation was produced, the context in which it was used, and how its suitability was assessed. The compliance logging layer therefore provides a structured record of relevant information, including the clinical context, selected XAI method, explanation generation process, translation approach, evaluation outcomes, user feedback, and relevant governance decisions. Such documentation can support transparency, auditability, human oversight, and accountability throughout the AI lifecycle. The purpose of this layer is not to define a specific regulatory framework or impose a single compliance mechanism. Rather, it provides a location within the overall XAI process where governance and documentation requirements can be incorporated. This distinguishes XRAI from approaches that treat regulatory considerations as external to technical explainability.

4.7 Integrated XRAI workflow

As shown in Fig. 6, the six layers operate as an interconnected pathway rather than as independent components. Clinical context profiling establishes the requirements for explanation, which informs XAI method selection. The six layers operate as an interconnected pathway rather than as independent components. Clinical context profiling establishes the requirements for explanation, which informs XAI method selection. The selected method generates a computational explanation, which is subsequently translated into clinically meaningful information. The translated explanation is evaluated by intended users, and the resulting feedback can inform refinement of earlier stages. Throughout this process, relevant decisions and outcomes are documented through the regulatory compliance layer.

The resulting workflow can therefore be represented as:

Clinical context → XAI method selection → Explanation generation → Clinical translation → Usability validation and feedback → Regulatory compliance logging

The workflow is iterative rather than strictly linear. Findings from usability validation may require modification of the selected XAI method or the clinical translation process, while governance requirements may influence the information collected at earlier stages. The framework consequently supports refinement as the AI system progresses from research development toward clinical use.

The principal distinction between XRAI and conventional XAI implementation is therefore the position of clinical context and translation within the overall process. Rather than moving directly from an AI prediction to a computational explanation, XRAI introduces explicit stages

for determining why an explanation is required, selecting an appropriate method, translating the output into clinically meaningful information, evaluating its usefulness, and documenting the resulting process.

4.8 Framework contribution

The XRAI Framework addresses the structural gaps identified through the evidence synthesis by integrating four requirements that are frequently separated in existing research: context-sensitive method selection, explanation evaluation, clinical translation, and governance. Its contribution is consequently methodological and integrative rather than algorithmic. The framework provides a structured pathway for moving from technically generated explanations toward explanations that are better aligned with the reasoning, workflow, and accountability requirements of radiological practice. It also provides a common structure through which different XAI techniques can be evaluated without assuming that any individual method is universally appropriate. By positioning clinical context as the starting point and usability and governance as integral components of the process, XRAI reframes explainability as an end-to-end clinical integration problem rather than solely as a model transparency problem. The framework therefore provides a basis for systematic evaluation of whether an explanation is not only computationally informative but also clinically meaningful, usable, and appropriately documented for its intended application.

5 Discussion

5.1 Principal Findings

This scoping review demonstrates that explainability in diagnostic radiology has progressed substantially at the methodological level, but its translation into clinically meaningful and operationally useful information remains comparatively underdeveloped. The reviewed literature encompasses a broad range of XAI approaches, including gradient-based methods, perturbation and feature-attribution techniques, attention mechanisms, prototype- and concept-based approaches, counterfactual explanations, and inherently interpretable models. However, the selection and evaluation of these approaches are frequently centered on model behavior rather than on the specific requirements of the radiological task, intended user, and clinical workflow. The distribution of imaging modalities and clinical tasks shown in Figs. 3 and 4 also indicates that radiological XAI is being studied across a wide range of applications. This variation makes a single explanation strategy difficult to apply across all clinical settings.

A central finding is therefore that explainability should not be regarded solely as a property of an AI model or an explanation algorithm. In radiology, the usefulness of an explanation depends on whether it can be interpreted within the clinical context in which the prediction is generated and used [44]. An explanation that successfully identifies image regions associated with a model prediction may demonstrate technical interpretability, but this does not necessarily indicate that the explanation provides actionable information to a radiologist. The distinction between computational explanation and clinical explanation represents an important structural gap within the current literature. The findings further indicate that explanation evaluation remains heterogeneous. Many studies assess explanations through technical or model-centered criteria, while comparatively less attention is given to whether clinicians understand the explanation, consider it relevant, or can incorporate it into their decision-making process. Explainability evaluation should therefore extend beyond demonstrating that an XAI method produces an interpretable output and should consider the relationship between explanation, prediction, clinical reasoning, and workflow [45].

5.2 Relationship to Existing XAI Research

The findings are consistent with the broader development of XAI, in which considerable methodological attention has been directed toward generating explanations for increasingly complex machine-learning models. Post-hoc approaches are particularly attractive in medical imaging because they can be applied to existing high-performing models without requiring major changes to the underlying architecture [46]. Nevertheless, the flexibility of post-hoc techniques also creates challenges concerning the stability, faithfulness, and clinical interpretation of their outputs. The reviewed evidence suggests that no individual XAI method can be considered universally appropriate for diagnostic radiology. Different clinical tasks generate different explanation requirements [47]. For example, lesion detection and localization may benefit from spatial explanations, whereas risk prediction or disease characterization may require feature-level, concept-level, or counterfactual information. Similarly, the explanation required by a radiologist may differ from that required by a clinician reviewing an AI-assisted decision or by a technical stakeholder auditing model behavior [19]. This reinforces the need to move from a method-first approach toward a context-first approach. Rather than asking which XAI technique should be applied to a particular model, the more clinically relevant question is which form of explanation is appropriate for a specific task, user, and decision context. The XRAI Framework addresses this requirement by positioning clinical context profiling before XAI method selection.

5.3 Clinical Translation as the Central Challenge

The most significant implication of the review is the need for an explicit translation layer between computational explanations and radiological reasoning. Radiologists generally interpret images through structured clinical concepts involving anatomy, morphology, density

or signal characteristics, lesion distribution, enhancement, relationships between structures, and differential diagnostic reasoning [48]. In contrast, many XAI systems communicate model behavior through heat maps, activation maps, attention scores, or feature-importance representations [49].

The evaluation pattern shown in Fig. 5 further supports this concern. Technical assessment is more common than direct assessment of clinical understanding and usefulness. This difference helps explain why computational explanations may not translate easily into radiological reasoning. These outputs may provide valuable evidence about where or which features influenced a model, but they do not inherently communicate why that information is clinically meaningful. Consequently, the presence of an explanation should not be equated with clinical interpretability. A heatmap may indicate a relevant anatomical region without explaining whether the model responded to a lesion, an imaging artefact, surrounding tissue, or another unintended characteristic. The XRAI Framework therefore treats clinical translation as a central integration layer rather than an optional post-processing step. This layer provides a conceptual mechanism for connecting computational explanation outputs with clinically meaningful radiological concepts. Such translation can potentially improve the relevance of explanations while also providing a clearer basis for evaluating whether explanations support or interfere with clinical reasoning [50].

5.4 Implications for Clinical AI Development

The workflow shown in Fig. 6 addresses this issue by placing clinical context and clinical translation within the explainability process. This allows the explanation to be considered in relation to the task and the intended user before it is used for clinical support. The findings have implications for the design and evaluation of future radiological AI systems. Explainability should be considered during system development rather than added only after model training. Early identification of the intended clinical task, user, workflow position, and explanation purpose can guide the selection of appropriate XAI techniques and reduce the risk of generating explanations that are technically valid but clinically unhelpful [51]. Evaluation should similarly incorporate multiple dimensions. Technical assessment remains important for understanding model behavior, but it should be complemented by human-centered assessment of comprehensibility, clinical relevance, usefulness, and consistency with expert reasoning. Such evaluation is particularly important because an explanation can increase user confidence without necessarily improving decision quality [52]. The objective should therefore be meaningful clinical support rather than simply greater user trust.

The framework also supports iterative development. Feedback obtained from intended users can be used to refine the selected explanation method, explanation representation, and clinical translation process. This creates a feedback-oriented approach in which explainability is treated as an evolving component of clinical AI rather than as a fixed characteristic of a deployed model.

5.5 Governance, Transparency and Accountability

Explainability also has implications beyond direct clinical interpretation [53]. As AI systems become integrated into clinical environments, there is a growing need to document how explanations are generated, interpreted, evaluated, and used [54]. Governance processes should be able to establish what model and explanation method were used, within which clinical context, for which intended purpose, and under what human oversight arrangements [55]. The regulatory compliance logging layer of XRAI addresses this requirement by treating explanation-related information as part of the broader accountability record [47]. Such documentation can support auditability, monitoring, and review when AI systems are modified or deployed across different clinical environments [56]. Importantly, logging should not be interpreted as a substitute for human oversight [47], [56]. Rather, it provides supporting evidence for responsible governance and facilitates reconstruction of explanation-related decisions when required.

5.6 Strengths and Limitations

A key strength of this review is its multidisciplinary focus across XAI methods, radiological applications, explanation evaluation, clinical translation, usability, and governance. Rather than evaluating individual explanation algorithms in isolation, the review identifies structural relationships between these components and uses them to develop the XRAI Framework as an integrative perspective. Several limitations should nevertheless be considered. The review was restricted to English-language, peer-reviewed, full-text literature published between January 2020 and June 2026 and therefore may not capture relevant evidence outside these criteria. The heterogeneous nature of the included literature also limits direct comparison across studies and does not support quantitative pooling of explanation performance. In addition, a formal risk of bias assessment was not undertaken because the primary objective was evidence mapping rather than estimation of a pooled effect. The protocol was not prospectively registered, which should also be considered when interpreting the methodological scope of the review.

Most importantly, XRAI should be regarded as a conceptual integration framework rather than a validated clinical decision-support algorithm. Its practical value requires further empirical investigation across imaging modalities, clinical tasks, user groups, and deployment

environments. Future studies should therefore examine whether context-sensitive explanation selection and explicit clinical translation improve explanation usefulness, clinician performance, and safe interaction with AI systems [57].

5.7 Future Research Directions

Future research should move toward standardized and clinically grounded evaluation of XAI in radiology [58]. This includes developing evaluation criteria that jointly consider technical explanation quality, clinical relevance, user understanding, decision support, and workflow compatibility. Comparative studies involving different XAI approaches within the same clinical task could also clarify which explanation forms are most appropriate for radiological contexts [59].

Further work is also required to operate the clinical translation layer proposed by XRAI. This could involve structured mappings between computational explanation outputs and radiological concepts, together with expert defined criteria for assessing whether such mappings accurately represent clinically meaningful information. Longitudinal evaluation in realistic clinical environments will additionally be necessary to determine whether explanations remain useful when integrated into routine workflows [60]. Overall, the evidence suggests that the next stage of radiological XAI research should move beyond the question of whether an AI model can be explained toward the broader question of whether its behavior can be explained **appropriately, meaningfully, and responsibly within clinical practice**. The XRAI Framework provides a structured basis for this transition by connecting clinical context, XAI selection, explanation generation, clinical translation, usability evaluation, and governance within a single integrative process.

6 Conclusion

This scoping review demonstrates that explainable artificial intelligence in radiological imaging has developed rapidly at the methodological level, but its translation into clinically meaningful, usable, and accountable information remains limited. A broad range of XAI approaches has been applied across imaging modalities and clinical tasks [61]. However, explanation generation and evaluation remain predominantly focused on model behavior [18]. The evidence indicates that a technically informative explanation does not necessarily provide information that can be readily interpreted or incorporated into radiological reasoning. The central gap identified in this review is therefore not the absence of XAI methods, but the limited integration of computational explanations with clinical context, radiological concepts, user needs, workflow requirements, and governance. Explanation selection should be guided by the clinical task, intended user, imaging context, and purpose of the explanation rather than by the availability or popularity of a particular XAI technique [62]. Similarly, evaluation should extend beyond technical properties of explanations to include clinical relevance, user understanding, usefulness, and compatibility with clinical decision-making [63].

To address these gaps, the Explainability–Radiology AI Integration (XRAI) Framework provides a structured, context-aware pathway linking clinical context profiling, XAI method selection, explanation generation, clinical translation, usability validation and feedback, and regulatory compliance logging. The framework positions clinical translation as an explicit integration layer between computational attribution and radiological reasoning and incorporates user evaluation and governance within the same explainability process. In this way, XRAI provides a common structure through which different XAI approaches can be considered without assuming that any individual method is universally appropriate. XRAI should be regarded as a conceptual integration framework rather than a validated clinical decision-support system. Its practical effectiveness therefore requires prospective empirical evaluation across imaging modalities, clinical tasks, user groups, and deployment environments.

Future studies should determine whether context-sensitive explanation selection and explicit clinical translation improve explanation quality, clinician understanding, decision support, workflow compatibility, and safe interaction with AI systems [13]. Overall, the findings support a shift in radiological XAI research from asking whether an AI system can be explained toward determining whether its behavior can be explained appropriately, meaningfully, and responsibly within clinical practice. The integration of technical explainability with clinical reasoning, usability, and governance may provide a more appropriate foundation for the responsible translation of AI-assisted radiological systems into clinical care [64].

Declarations

Competing interests -

- The authors declare that they have no competing interests

Consent to Publish declaration

- Not applicable

Ethics approval and consent to participate –

- Not applicable, as this is a scoping review.

Funding –

Not applicable.

Author Contribution

Kaneeka Vidanage - Concept and drafting of the entire paper.K.L.Y.S. Karandana - Drawing of diagrams / figures.Bahasuru Nayanakantha - Writing : Review & Editing. Smita Nirkhi - Proofing the entire manuscript.Sabyasachi Bhattacharyya - Proofing the entire manuscript.

Data Availability

Paper list, utilized for the review has been provided as an annexure with this submission.

References

1. Patrício C, Neves JC, Teixeira LF. Explainable Deep Learning Methods in Medical Image Classification: A Survey. *ACM Comput Surv. Apr.* 2024;56(4):1–41. 10.1145/3625287.
2. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med.* Jan. 2019;25(1):44–56. 10.1038/s41591-018-0300-7.
3. Litjens G, et al. A survey on deep learning in medical image analysis. *Med Image Anal.* Dec. 2017;42:60–88. 10.1016/j.media.2017.07.005.
4. McKinney SM, et al. International evaluation of an AI system for breast cancer screening. *Nature.* Jan. 2020;577(7788):89–94. 10.1038/s41586-019-1799-6.
5. Fouad S et al. Dec., Explained Deep Learning Framework for COVID-19 Detection in Volumetric CT Images Aligned with the British Society of Thoracic Imaging Reporting Guidance: A Pilot Study, *J Digit Imaging. Inform. med.*, vol. 38, no. 6, pp. 3634–3651, 2025, 10.1007/s10278-025-01444-3
6. Saw SN, Yan YY, Ng KH. Current status and future directions of explainable artificial intelligence in medical imaging. *Eur J Radiol.* Feb. 2025;183:111884. 10.1016/j.ejrad.2024.111884.
7. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell.* May 2019;1(5):206–15. 10.1038/s42256-019-0048-x.
8. Yang Y, Ying R, Tong C. ProtoMM: Interpretable Prototype-Based Multimodal Model for Brain Cancer Survival Prediction. *J Digit Imaging Inf med.* Aug. 2026;39(4):3006–19. 10.1007/s10278-025-01713-1.
9. Haupt M, Maurer MH. Explainable artificial intelligence in radiology: Methods, clinical applications, limitations, and future directions. *Eur J Radiol Artif Intell.* Jun. 2026;6:100098. 10.1016/j.ejrai.2026.100098.
10. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med.* Oct. 2019;17(1):195. 10.1186/s12916-019-1426-2.
11. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit Health.* Nov. 2021;3(11):e745–50. 10.1016/S2589-7500(21)00208-9.
12. Haupt M, Maurer MH, Thomas RP. Explainable Artificial Intelligence in Radiological Cardiovascular Imaging—A Systematic Review. *Diagnostics.* May 2025;15(11):1399. 10.3390/diagnostics15111399.
13. Barredo Arrieta A et al. Jun., Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, *Information Fusion*, vol. 58, pp. 82–115, 2020, 10.1016/j.inffus.2019.12.012
14. Lundberg SM, Lee S-I. A Unified Approach to Interpreting Model Predictions, in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2017. Accessed: Sep. 16, 2026. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html
15. De Vries BM, Zwezerijnen GJC, Burchell GL, Van Velden FHP, Van Oordt, Boellaard R. Explainable artificial intelligence (XAI) in radiology and nuclear medicine: a literature review. *Front Med.* May 2023;10:1180773. 10.3389/fmed.2023.1180773.

16. Jin W, Li X, Fatehi M, Hamarneh G. Guidelines and evaluation of clinical explainable AI in medical image analysis. *Med Image Anal.* Feb. 2023;84:102684. 10.1016/j.media.2022.102684.
17. Reyes M et al. May., On the Interpretability of Artificial Intelligence in Radiology: Challenges and Opportunities, *Radiology: Artificial Intelligence*, vol. 2, no. 3, p. e190043, 2020, 10.1148/ryai.2020190043
18. van der Velden BHM, Kuijff HJ, Gilhuijs KGA, Viergever MA. Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. *Med Image Anal.* Jul. 2022;79:102470. 10.1016/j.media.2022.102470.
19. Ihongbe IE, Fouad S, Mahmoud TF, Rajasekaran A, Bhatia B. Evaluating Explainable Artificial Intelligence (XAI) techniques in chest radiology imaging through a human-centered Lens. *PLoS ONE.* Oct. 2024;19(10):e0308758. 10.1371/journal.pone.0308758.
20. Houssein EH, Gamal AM, Younis EMG, Mohamed E. Explainable artificial intelligence for medical imaging systems using deep learning: a comprehensive review. *Cluster Comput.* Sep. 2025;28(7):469. 10.1007/s10586-025-05281-5.
21. Ahmed F, Naz NS, Khan S, Rehman AU, Ismael WM, Khan MA. Explainable artificial intelligence (XAI) in medical imaging: a systematic review of techniques, applications, and challenges. *BMC Med Imaging.* Jan. 2026;26(1):37. 10.1186/s12880-025-02118-w.
22. Chaddad A, Peng J, Xu J, Bouridane A. Survey of Explainable AI Techniques in Healthcare, *Sensors*, vol. 23, no. 2, p. 634, Jan. 2023, 10.3390/s23020634
23. Ray PP. A review on explainable artificial intelligence in radiomics: State-of-the-art tools, prospective use cases, challenges and future directions. *Eur J Radiol Artif Intell.* Mar. 2026;5:100069. 10.1016/j.ejrai.2025.100069.
24. Chang Y, Hsieh M-H, Ju P-C, Liu Y-C, Chang C-C. Clinical Plausibility in Large Language Model Robustness Testing for Medicine: A Scoping Review. *J Med Syst.* May 2026;50(1):77. 10.1007/s10916-026-02405-1.
25. Crigger E, Reinbold K, Hanson C, Kao A, Blake K, Irons M. Trustworthy Augmented Intelligence in Health Care, *J Med Syst*, vol. 46, no. 2, p. 12, Jan. 2022, 10.1007/s10916-021-01790-z
26. Champendal M, Müller H, Prior JO, Dos Reis CS. A scoping review of interpretability and explainability concerning artificial intelligence methods in medical imaging. *Eur J Radiol.* Dec. 2023;169:111159. 10.1016/j.ejrad.2023.111159.
27. *Ethics and Governance of Artificial Intelligence for Health: WHO Guidance*, 1st ed. Geneva: World Health Organization, 2021.
28. Van Der Velden BHM, Kuijff HJ, Gilhuijs KGA, Viergever MA. Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. *Med Image Anal.* Jul. 2022;79:102470. 10.1016/j.media.2022.102470.
29. Arksey H, O'Malley L. Scoping studies: towards a methodological framework, *International Journal of Social Research Methodology*, vol. 8, no. 1, pp. 19–32, Feb. 2005, 10.1080/1364557032000119616
30. Grant MJ, Booth A. A typology of reviews: an analysis of 14 review types and associated methodologies. *Health Info Libr J.* Jun. 2009;26(2):91–108. 10.1111/j.1471-1842.2009.00848.x.
31. Aviso M, et al. Multi-sectoral governance for health: a scoping review and typology development. *Health Policy Plan.* Sep. 2026;106:czag. 10.1093/heapol/czag106.
32. Levac D, Colquhoun H, O'Brien KK. Scoping studies: advancing the methodology. *Implement Sci.* Sep. 2010;5(1):69. 10.1186/1748-5908-5-69.
33. Tricco AC, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. *Ann Intern Med.* Oct. 2018;169(7):467–73. 10.7326/M18-0850.
34. Page MJ et al. Mar., The PRISMA 2020 statement: an updated guideline for reporting systematic reviews, *BMJ*, p. n71, 2021, 10.1136/bmj.n71
35. Bouck Z, Straus SE, Tricco AC. Systematic Versus Rapid Versus Scoping Reviews Scoping reviews. In: Evangelou E, Veroniki AA, editors. *Meta-Research: Methods and Protocols*. New York, NY: Springer US; 2022. pp. 103–19. 10.1007/978-1-0716-1566-9_6.
36. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan—a web and mobile app for systematic reviews. *Syst Rev.* Dec. 2016;5(1):210. 10.1186/s13643-016-0384-4.
37. Landis JR, Koch GG. The Measurement of Observer Agreement for Categorical Data. *Biometrics.* 1977;33(1):159–74. 10.2307/2529310.
38. Munn Z, Peters MDJ, Stern C, Tufanaru C, McArthur A, Aromataris E. Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. *BMC Med Res Methodol.* Nov. 2018;18(1):143. 10.1186/s12874-018-0611-x.
39. Li X, Razak RBA. From generative to agentic AI: qualitative thematic synthesis of cognitive and motivational impacts on personalized higher education pathways, *Interactive Learning Environments*, vol. 0, no. 0, pp. 1–15, Sep. 2026, 10.1080/10494820.2026.2729788
40. Knapič S, Malhi A, Saluja R, Främpling K. Explainable Artificial Intelligence for Human Decision Support System in the Medical Domain, *MAKE*, vol. 3, no. 3, pp. 740–770, Sep. 2021, 10.3390/make3030037

41. Watanabe A, Ketabi S, Namdar K, Khalvati F. Improving disease classification performance and explainability of deep learning models in radiology with heatmap generators. *Front Radio*. Oct. 2022;2:991683. 10.3389/fradi.2022.991683.
42. G. M., A. V., M.-M. S., and M. H., Concept attribution: Explaining CNN decisions to physicians. *Comput Biol Med*, 123, p. 103865, Aug. 2020, 10.1016/j.combiomed.2020.103865
43. Răz T, De Mortanges AP, Reyes M. Explainable AI in medicine: challenges of integrating XAI into the future clinical routine. *Front Radiol*. Aug. 2025;5:1627169. 10.3389/fradi.2025.1627169.
44. Fouad S, Hakobyan L, Ihongbe IE, Kavakli-Thorne M, Atkins S, Bhatia B. Human-Centered User Interface Design for Explainable AI in Chest Radiology: A Multi-Phase Co-Design Approach. *IEEE Access*. 2026;14:12498–513. 10.1109/ACCESS.2026.3653233.
45. Brandenburg JM, Müller-Stich BP, Wagner M, Schaar MVD. Can surgeons trust AI? Perspectives on machine learning in surgery and the importance of eXplainable Artificial Intelligence (XAI). *Langenbecks Arch Surg*. Jan. 2025;410(1):53. 10.1007/s00423-025-03626-7.
46. Singh Y et al. Aug., Beyond Post hoc Explanations: A Comprehensive Framework for Accountable AI in Medical Imaging Through Transparency, Interpretability, and Explainability, *Bioengineering*, vol. 12, no. 8, p. 879, 2025, 10.3390/bioengineering12080879
47. Neri E, et al. Explainable AI in radiology: a white paper of the Italian Society of Medical and Interventional Radiology. *Radiol med*. May 2023;128(6):755–64. 10.1007/s11547-023-01634-5.
48. Zamir MT, Khan SU, Gelbukh A, Felipe EM, Riverón, Gelbukh I. Explainable AI-Driven Analysis of Radiology Reports Using Text and Image Data: Experimental Study. *JMIR Form Res*. Oct. 2025;9:e77482. 10.2196/77482.
49. Ghauth KI, Kustiawan YA. Explainable Artificial Intelligence (XAI) for Cancer Classification in Medical Imaging: A Systematic Review. *MAKE*. May 2026;8(5):134. 10.3390/make8050134.
50. Fahad N, Sadib RJ, Sajib RH, Morol MK, Nandi D, Liew TH. Responsible artificial intelligence in medical imaging: a systematic review, *Front. Digit. Health*, vol. 8, p. 1884692, Jul. 2026, 10.3389/fdgth.2026.1884692
51. Famiglini L, Campagner A, Barandas M, La Maida GA, Gallazzi E, Cabitza F. Evidence-based XAI: An empirical approach to design more effective and explainable decision support systems. *Comput Biol Med*. Mar. 2024;170:108042. 10.1016/j.combiomed.2024.108042.
52. Salih A, et al. Explainable Artificial Intelligence and Cardiac Imaging: Toward More Interpretable Models. *Circ: Cardiovasc Imaging*. Apr. 2023;16(4). 10.1161/CIRCIMAGING.122.014519.
53. Kandala R, Moharir AK, Nayak DA. From Explainability to Action: A Generative Operational Framework for Integrating XAI in Clinical Mental Health Screening, in *2025 IEEE International Conference on Data Mining Workshops (ICDMW)*, Washington, DC, USA: IEEE, Nov. 2025, pp. 1316–1325. 10.1109/ICDMW69685.2025.00156
54. Hafeez Y, Memon K, AL-Quraishi MS, Yahya N, Elferik S, Ali SSA. Explainable AI in Diagnostic Radiology for Neurological Disorders: A Systematic Review, and What Doctors Think About It. *Diagnostics*. Jan. 2025;15(2):168. 10.3390/diagnostics15020168.
55. Yang G, Ye Q, Xia J. Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond. *Inform Fusion*. Jan. 2022;77:29–52. 10.1016/j.inffus.2021.07.016.
56. Stogiannos N et al. Oct., Black box no more: a scoping review of AI governance frameworks to guide procurement and adoption of AI in medical imaging and radiotherapy in the UK, *BJR*, p. 20221157, 2023, 10.1259/bjr.20221157
57. Dwivedi R et al. Sep., Explainable AI (XAI): Core Ideas, Techniques, and Solutions, *ACM Comput. Surv.*, vol. 55, no. 9, pp. 1–33, 2023, 10.1145/3561048
58. Liu Y, Liu C, Zheng J, Xu C, Wang D. Improving Explainability and Integrability of Medical AI to Promote Health Care Professional Acceptance and Use: Mixed Systematic Review. *J Med Internet Res*. Aug. 2025;27:e73374. 10.2196/73374.
59. Pal M, Saha HN, Chakrabarti A. The Trust-Aware XAI (TAXAI) framework: a quantitative model for interpretable and reliable clinical AI systems. *Sci Rep*. Apr. 2026;16(1):15455. 10.1038/s41598-026-44167-3.
60. Groen AM, Kraan R, Amirkhan SF, Daams JG, Maas M. A systematic review on the use of explainability in deep learning systems for computer aided diagnosis in radiology: Limited use of explainable AI? *Eur J Radiol*. Dec. 2022;157:110592. 10.1016/j.ejrad.2022.110592.
61. Cheddi F, Habbani A, Nait-Charif H. Explainable Approach Using Semantic-Guided Alignment for Radiology Imaging Diagnosis, *ijacsa*, vol. 16, no. 7, 2025, 10.14569/IJACSA.2025.0160761
62. Saurabh M, Nehra A, Mishra A, Sharma D, Solanki, Gaurav, Explainable Artificial Intelligence for Healthcare: Enhancing Clinical Trust and Decision-Making in Radiology, in. *2026 IEEE 6th International Conference on Computing, Power, and Communication Technologies (IC2PCT)*, Greater Noida, India: IEEE, May 2026, pp. 915–919. 10.1109/IC2PCT68894.2026.11584427
63. Ali S, et al. Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Inform Fusion*. Nov. 2023;99:101805. 10.1016/j.inffus.2023.101805.

Figures

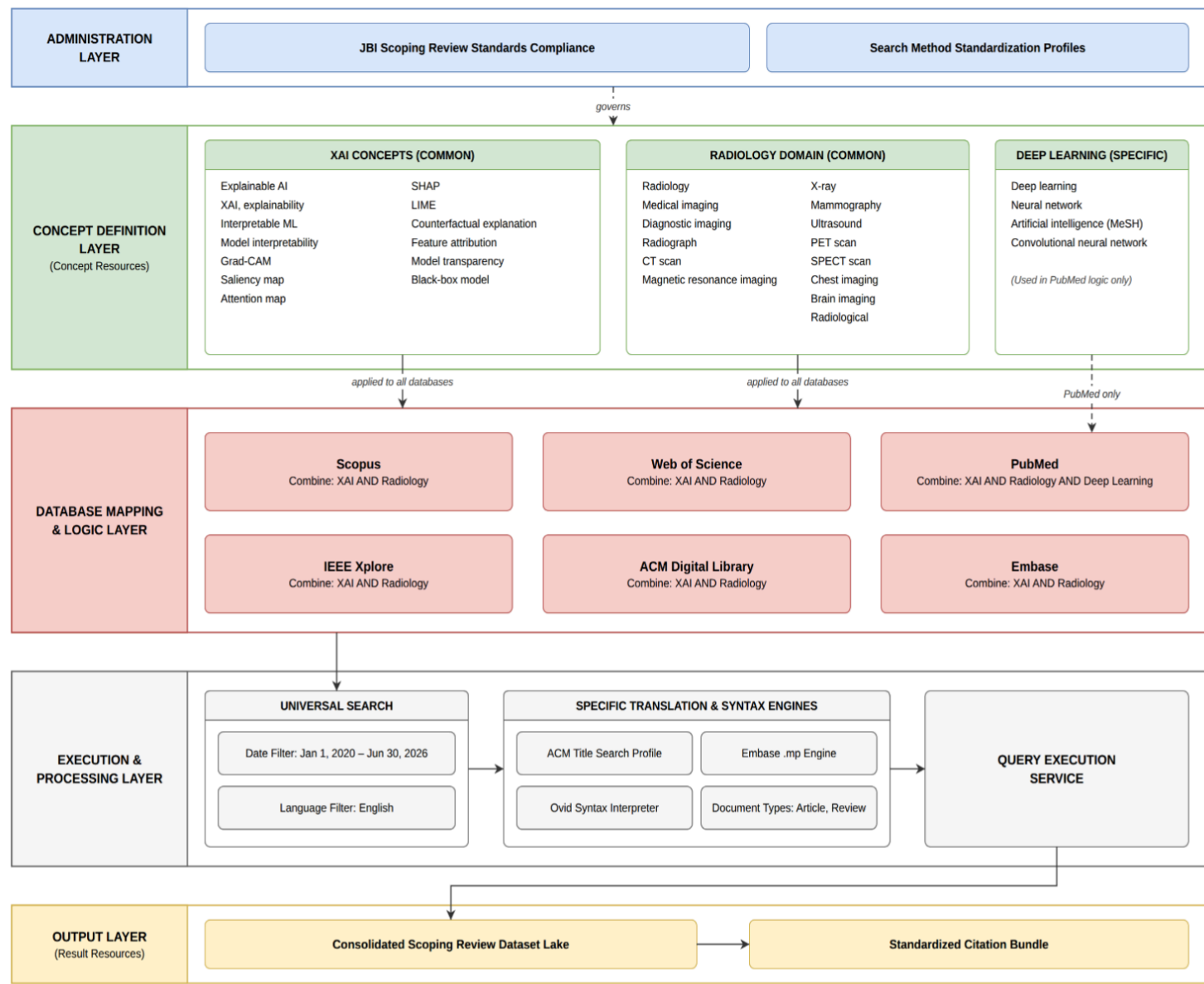


Figure 1

Full search strings deployed across respective databases

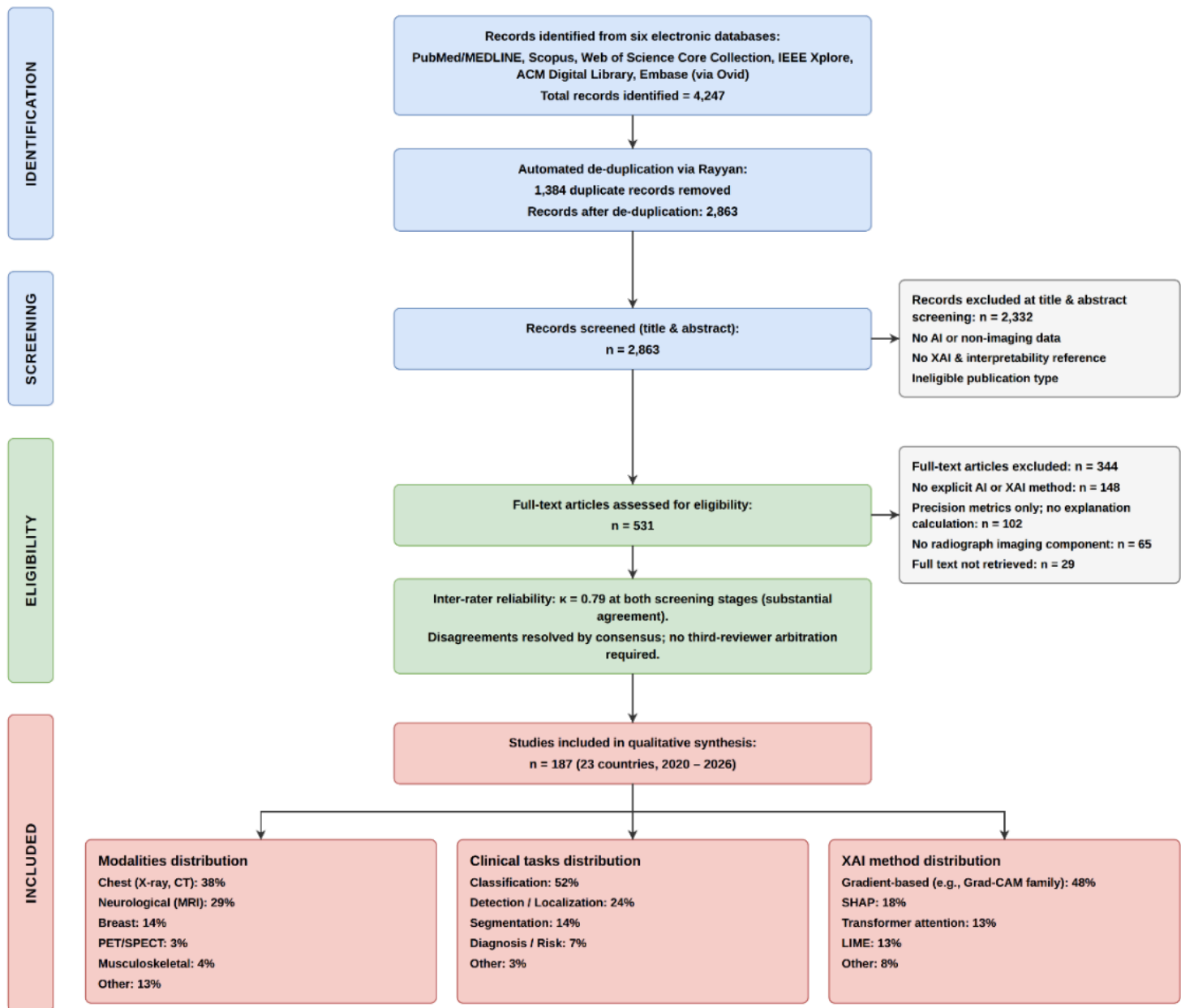


Figure 2

PRISMA-ScR flow diagram detailing the study selection process

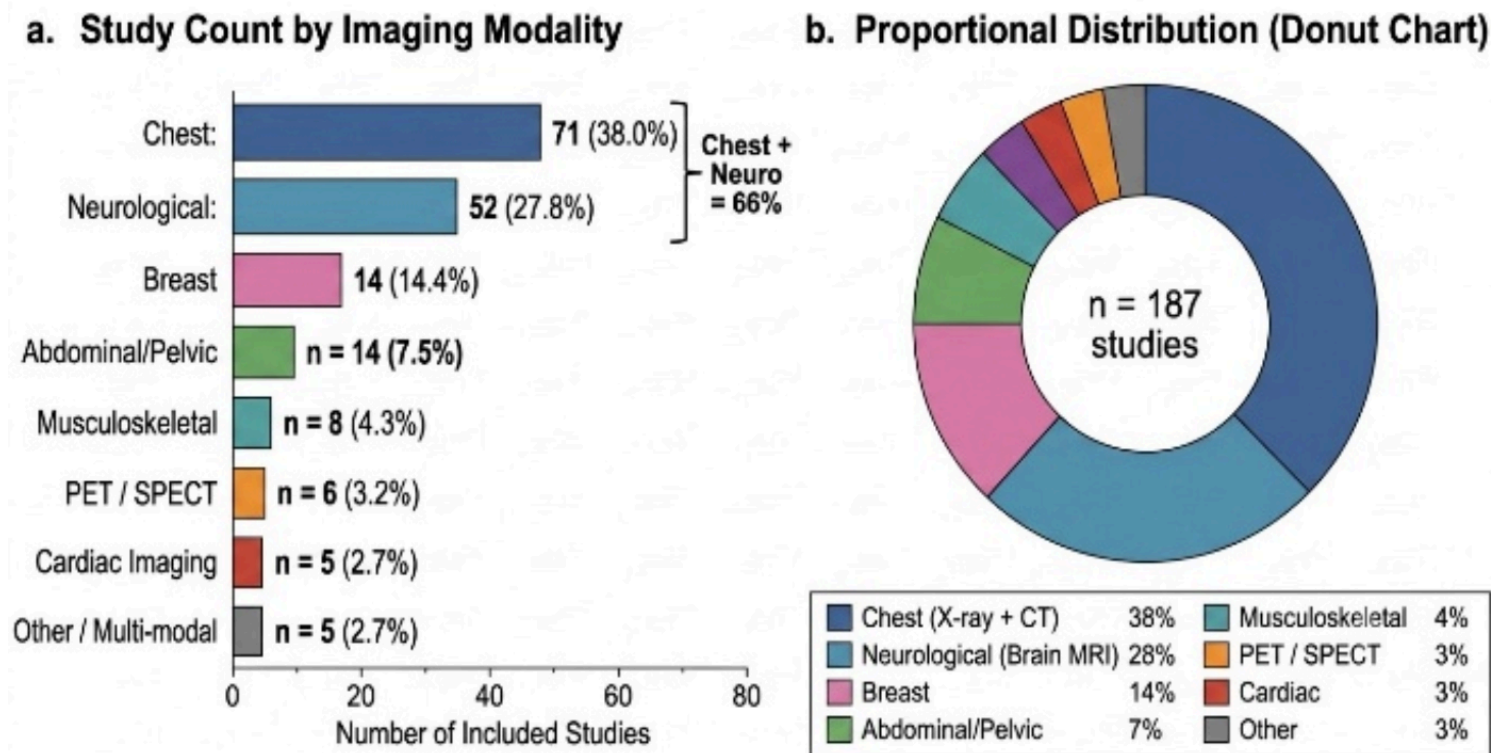


Figure 3

Distribution of included studies across imaging modalities (n = 187).

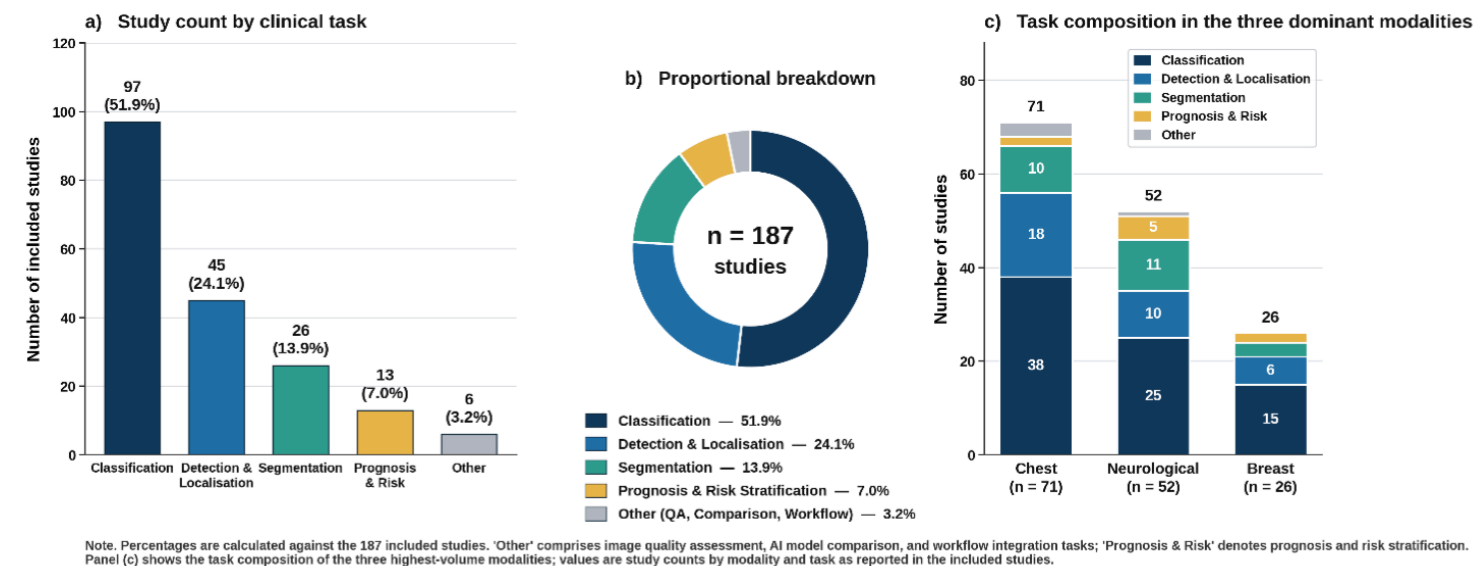


Figure 4

Distribution of Clinical Tasks Across Included Studies (n = 187). Panel (a) shows the study count by clinical task. Panel (b) provides a proportional breakdown. Panel (c) illustrates the task composition for the three dominant modalities.

Figure 6

XRAI Framework showing the six linked stages from clinical context profiling to regulatory compliance logging.