BMC Medical Imaging

RESEARCH

Open Access

Explainable artificial intelligence (XAI) in medical imaging: a systematic review of techniques, applications, and challenges

Fahad Ahmed[1], Naila Sammar Naz[1], Sunawar Khan[1], Ateeq Ur Rehman[2,3,4], Waleed M. Ismael[5*] and Muhammad Adnan Khan[6*]

Abstract

Explainable Artificial Intelligence (XAI) is crucial for enhancing transparency and trustworthiness, as well as for developing AI-based diagnostic systems in medical imaging to achieve clinical acceptability and reliability. This synthesis aligns with the current context of XAI in medical imaging on four key dimensions: trends in techniques, their application to clinical use cases or with human subjects, and the associated problems. In radiology and pathology, we report on image analysis based on current literature from credible databases. This review extends on the existing surveys by explicitly addressing the focus on feature selection (FS), graph neural networks (GNNs), and multimodal transformers and combining them to a cohesive XAI taxonomy and matches the techniques to the specific impact points in the clinical workflow in radiology and pathology. We present saliency maps, attention mechanisms, and gradient-based and rule-based elucidations of deep learning (DL) models in today's healthcare environment. Finally, the results show that XAI significantly enhances overall clinical decision-making by making the high-level reasoning of the model readily available to users, thereby increasing their confidence in clinical decisions despite the remaining obstacles, including standardization, interpretability, data bias, and complicated data integration. From 980 records, 289 duplicates were removed, 691 screened, 209 excluded, 482 full texts assessed, 263 excluded with 219 yielding 133 included studies. Exploring this third dimension within an XAI environment identifies research gaps and paves the way for robust medical imaging XAI solutions.

Clinical trial number Not applicable.

Keywords Explainable artificial intelligence (XAI), Medical imaging, Deep learning (DL), Radiology, Pathology, Saliency maps, Model interpretability

*Correspondence:
Waleed M. Ismael
Waleed.m@auhd.edu.ye
Muhammad Adnan Khan
adnan@gachon.ac.kr
[1]School of Computer Science, National College of Business Administration and Economics, Lahore 54000, Pakistan
[2]Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai, Tamil Nadu, India
[3]Applied Science Research Center, Applied Science Private University, Amman, Jordan
[4]Chitkara University Institute of Engineering and Technology, Chitkara University, Rajpura, Punjab, India
[5]Department of Information Technology, Faculty of Engineering, Azal University for Human Development, Sanaa, Yemen
[6]Department of Software, Faculty of Artificial Intelligence and Software, Gachon University, Seongnam-si 13557, Republic of Korea

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

Introduction

In the past few years, Artificial Intelligence (AI) has made great strides in the world of medical imaging and medical diagnostics, due to the advances in computational capabilities that have led to better clinical outcomes. This section provides the premise of a systematic literature review (SLR) detailing how the integration of AI in medical imaging, the need for XAI, and the rationale behind a standalone individual subsection. A linear diagram is used in a healthcare context to record important dates relating to AI and XAI that is turned into a discussion about the chronological development of it. In addition, the explicit identification of critical research gaps, issues and objectives of the current review and the contribution of the present review to the existing knowledge are provided. Formulated research questions are also presented.

Over the past 10 years, AI using machine learning (ML) and DL has shown to be useful for diagnosing breast cancer and brain cancer, detecting retinal disease, and segmenting medical images [1–4]. Using the framework, it was discovered that there is a low-dimensional embedding of deep neural network (DNN) of semantic and visual dimensions. Nonetheless, when human and DNN representation was put directly in a comparison, there were significant differences in image processing [5]. The recent developments in the field of artificial intelligence have given rise to the creation of XAI, which helps in increasing transparency by exposing the inner workings of artificial intelligence algorithms. However, there are challenges to overcome, such as balancing performance with explainability and security and policy issues [6]. XAI makes the algorithms used in healthcare understandable, thus making the decision-making processes of physicians more confident. The Defense Advanced Research Projects Agency (DARPA) says that the goal of XAI is to produce accurate and interpretable models [6]. The efficacy of learning increases the understanding, trust, and oversight of AI partners. Although a lot of applicability is offered by XAI, it is especially important in critical clinical decision making where errors can have important consequences. Algorithmic decision-making should be transparent before it is used in patient care as required by the EU General Data Protection Regulation (GDPR) [7]. The final recommendation given by the U.S. Department of Health and Human Services requires that for users of Clinical Decision Support Systems (CDSS) to use them efficiently and safely, regulatory standards understanding is paramount. Explainability makes XAI reliable and transparent, making it easier to make clinical judgement and adhere to regulatory standards.

AI is computer methods to generalize decisions, to recognize patterns, and learn from data. AI in medical imaging mainly uses the clinical imaging technologies like ML, DL, and CV to interpret the radiological images and detect the abnormality present in the images which helps us in a diagnostic process [8]. AI was first used in the healthcare industry, with straightforward algorithms that tried to identify patterns with the help of pattern matching. This approach was continued and we are now seeing more advanced architectures of neural networks including Convolutional Neural Networks (CNNs) which have caused a dramatic increase in the accuracy of diagnosis [9]. The integration of AI in medical imaging has significantly enhanced healthcare by making the process of analyzing images more accurate and efficient. New technologies like DL and CNNs enable quick identification of abnormalities and, thereby, support early diagnosis of the disease and personalized forms of treatment to achieve better patient results [10]. Due to further development automated segmentation and detection of lesions and disease classification and prognosis prediction can be followed and this can help radiologist to increase the accuracy and working efficiency. In turn, increased computational resources, new algorithm policy, and access to a wealth of medical data allows AI to revolutionize the approach to the analysis of medical images.

In addition to this, XAI is a set of techniques that allows users to understand how AI systems work or operate. Information on how the systems make decisions is critical to clinicians. In contrast to this, traditional black box AI models, which don't offer explanations on their decision-making processes are mainly used to generate predictions [11]. The recent developments in AI and DL have achieved remarkable success in fields such as image and language processing. However, their black box nature can be challenging, particularly in healthcare, where transparency is essential for making sound decisions. Efforts are underway to enhance interpretability via approaches such as saliency maps and signal extraction and the focus is on mathematical structures and sensitivity analysis for better applicability of AI in medical settings [12]. Then, such a decision-making process could be explained or predicted by using techniques like Gradient-weighted Class Activation Mapping (Grad-CAM), Attention mechanisms, Local Interpretable Model-Agnostic Explanations (LIME) and Shapley Additive Explanations (SHAP) [13]. As a good case in point, it has been proved that it is a critical area where transparency is very crucial as the decisions taken have a significant impact on the patient's outcomes. XAI lowers clinical liability through grounding an automated diagnostic tool to an ethical framework and clinical guidelines [14].

The combination of DL and medical imaging techniques has improved healthcare diagnosis and prognosis. However, these models are complex, and this presents problems to their decision-making processes. To tackle this, visualization methods and interpretability have become relevant in making the whole approach

less opaque thus establishing confidence in their predictions. In this paper, a summary of various interpretation and visualization techniques will be provided, accompanied by demonstrations on their use and analyses of their effectiveness in improving the interpretability, reliability, and clinical importance of DL models in medical image analysis [15]. Medical image analysis has been greatly affected by the development of DL. Classification, segmentation and picture synthesis has been done using CNN, Recurrent Neural Networks (RNNs), Autoencoders, Generative Adversarial Networks (GANs), U-Net models and Vision Transformers (ViTs). Despite their accomplishment, there are still significant challenges. These include data availability, model interpretability, overfitting and computational complexity. These problems, in addition to multimodal integration, computing efficiency, and explainability, need to be addressed in order to bring these models to clinical use. DL research will enhance its applicability for real-time applications, enhance the intelligibility of the outcomes, and assure technology interoperability with healthcare systems [16].

Black Box AI is accurate, scalable, and opaque, exposing invisible biases, responsibilities, and ethics. Instead, XAI provides transparent decision details to build confidence, accountability, and justice in sensitive applications. Most ML and DL models are black boxes, restricting their use in healthcare, banking, and public safety. XAI improves interpretability, reduces errors, and builds confidence. This study examines cutting-edge XAI systems, their performance, and future challenges to increase transparency and projections [17].

Furthermore, XAI achieves these features by providing comfort of thought for patients' minds, enhancing patient safety, and eradicating prejudices from the provided data and incorrect predictions, thereby being significant within the clinical context [18]. Therefore, decision-making among an increasing number of medical practitioners is based on XAI. For this reason, the fundamental value of XAI in medical imaging operations should not be overlooked these days. XAI and Black Box AI explanations are presented through the processes, outputs, and applications. XAI employs an explainable process that yields clear, explainable outcomes, making it applicable in matters that require reliability, such as healthcare, financial, and legal systems. On the contrary, Black Box AI operates using an opaque mechanism, which ensures high accuracy but produces incomprehensible output, and is best suited for applications where performance is prioritized over explainability, such as in cases of facial recognition, speech recognition, and recommendation systems. Fig. 1 illustrates the comparison between these two.

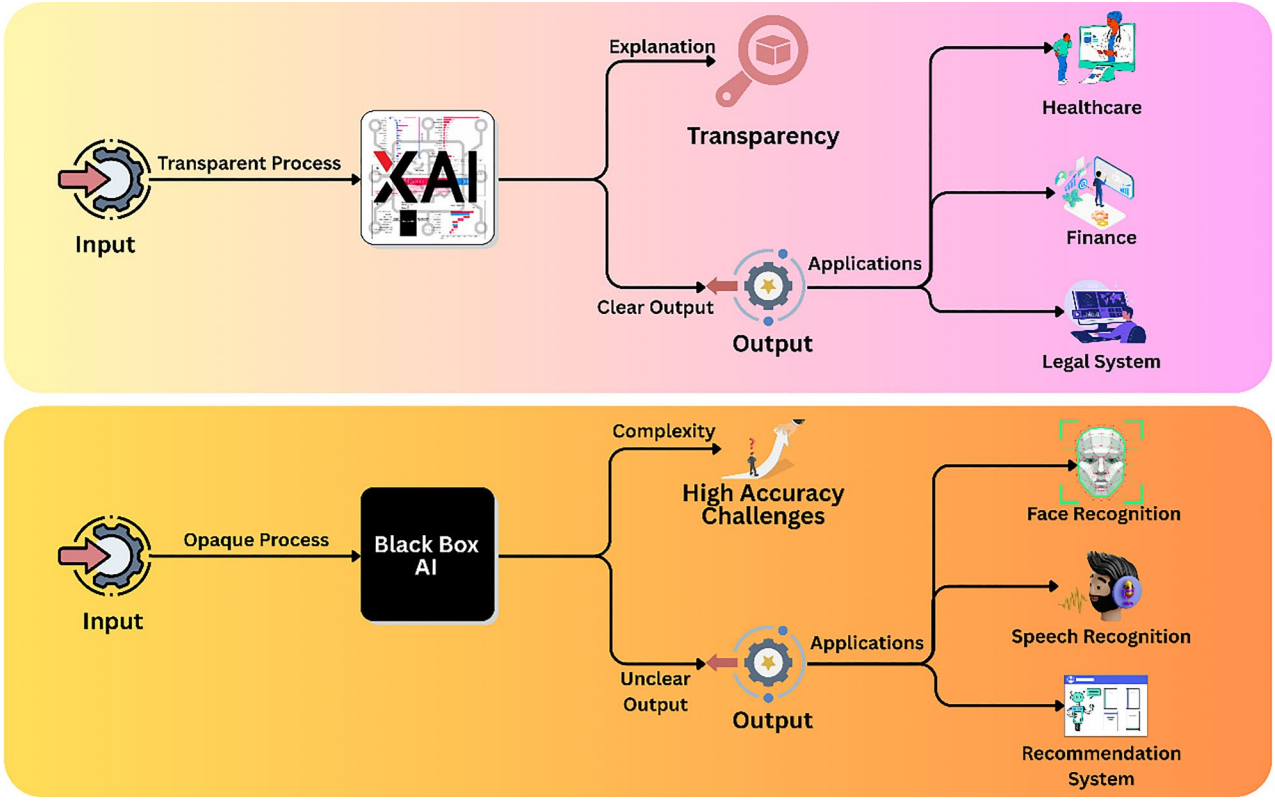

Fig. 1 Comparison between XAI and Black box AI

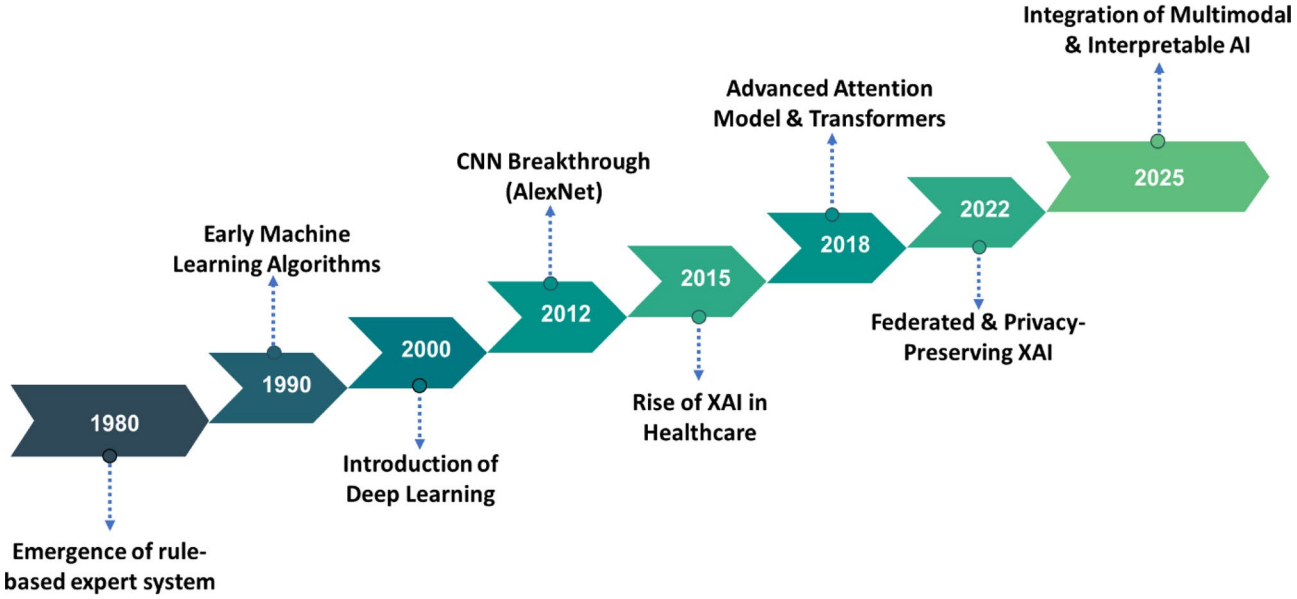

Fig. 2 Evolution of AI and XAI in medical imaging

Figure 2 illustrates the history of AI, highlighting its key achievements that contributed to the emergence of XAI. Since the introduction of rule-based expert systems in the 1980s, the field has progressed to ML algorithms in the 1990s and DL in the 2000s, which offer better performance but poorer interpretability. The 2012 CNN breakthrough catalyzed the acceleration of black-box models, which need transparency, particularly in the world of healthcare, by 2015. Transformers in 2018 and federated privacy-preserving XAI in 2022 have led the discipline to the next phase of multimodal and interpretable AI, which is likely to emerge in 2025, characterized by a persistent trade-off between accuracy and explainability.

Figure 2 illustrates the development of AI and XAI, as well as their application in biomedical imaging, from early rule-based expert systems to present-day Multimodal interpretive diagnosis tools. The early motivation for decision support was a simple, rule-based algorithm in the 1980s. Before that, we could switch from early ML algorithms in the early 1990s to learn model patterns from image datasets. In the early 2000s, DL emerged, leading to unprecedented accuracy in results and paving the way for the development of CNN architectures, such as AlexNet, in 2012. As of 2015, interpretability has become a focus for gaining more detailed insights, which has given rise to the field of independent XAI research. Recent advancements in AI for diagnostic systems include federated learning, attention mechanisms, and transformer-based architectures that provide explainability and interpretability in AI-driven diagnostic systems. It effectively shows stakeholders how shifting transparency needs translate to contemporary XAI research in healthcare through a timeline, major dates, technological breakthroughs in its field, and trends.

This systematic review is added value in the increasing literature about XAI in medical imaging. First, it suggests a formal taxonomy according to which not only methods are classified but outlines Feature Selection (FS) and model-agnostic methods and includes interpretability methods based on DLs. Second, we concentrate on more recent architectures, such as Graph Neural Networks (GNNs) and multimodal transformers, which have not been represented in earlier surveys but are an ideal point of real-life clinical usage. Third, this review classifies all families of methods by clinical workflow impact points, which can be used to elaborate on where and how it can raise trust and enhance decision-making in radiology and pathology. Fourth, we bring together ethical, regulatory, and accountability issues into concrete clinical vignettes, which will offer practical advice on how to apply these issues in practice. Lastly, the study is put into perspective about how it compares to the current reviews in Table 1, its significance to the literature that is already established.

The review [26] gives an insightful summary of AI technologies in medical imaging applications in terms of their role in solving specific tasks. It deals with innovations in DL such as GANs, transfer learning, and federated learning, and their application for several tasks that require image segmentation, classification, and reconstruction in different modalities such as X-ray, Computed Tomography (CT), Magnetic Resonance Imaging (MRI), Positron Emission Tomography (PET), and Ultrasound (US). The importance of interpretability, robustness and generalizability in clinical practice and the ethical considerations of data privacy and bias is emphasized in the review.

Table 1 Comparative overview of related surveys on XAI in medical imaging (2020–2025)

Reference	Modalities covered	XAI families discussed	Human-in-the-loop/clinical workflow	Regulatory & ethics	Distinctive contributions of this review
[12]	MRI, CT, X-ray, pathology, fundus	model-agnostic, intrinsic	Minimal – only high-level need	Very low (brief GDPR mention)	First unified taxonomy integrating Feature Selection, GNNs, and multimodal Transformers; explicit workflow impact mapping; detailed regulatory analysis
[19]	Medical imaging, disease classification, segmentation	Local & global methods, knowledge base, distillation algorithms, interpretable ML	Emphasis on human–machine collaboration	Not specified	Introduction of a novel XAI recommendation and scoring system for human-enforced explanation feedback
[20]	Medical imaging and related healthcare tasks	Saliency maps, prototype methods, and XAI categorize model interpretation.	XAI bridges AI models and clinician interpretation for acceptance.	Challenges include opaque AI decisions and transparency in healthcare.	Categorizes XAI methods, highlights pros/cons, challenges, and future directions.
[21]	Medical imaging, disease classification, segmentation	Not specified	Not specified	Not specified	Focus on hybrid ML/DL approaches, challenges in data availability, interpretability, and model generalization
[22]	Medical imaging, disease classification, segmentation	Not specified	Not specified	Not specified	Focus on hybrid ML/DL approaches, challenges in data availability, interpretability, and model generalization
[23]	CT, MRI, X-ray, US, PET	3–4 (saliency, attention, perturbation, surrogate)	Partial – anatomical location only	Low (basic ethical notes)	First unified taxonomy integrating Feature Selection, GNNs, and multimodal Transformers; explicit workflow impact mapping; detailed regulatory analysis
[24]	Medical imaging	Not specified	Not specified	Not specified	Emphasis on human-centered design principles for transparent ML in medical image analysis
[25]	Medical imaging, disease classification, segmentation	Not specified	Not specified	Not specified	Emphasis on human-centered design principles for transparent ML in medical image analysis
Proposed Study	Radiology (X-ray, CT, MRI), Pathology (digital & WSI), Multimodal (PET-CT, MRI-PET, etc.)	4 major families + Feature Selection + Graph Neural Networks + Multimodal Transformers	High – every XAI family explicitly mapped to clinical workflow impact points + synthesis of clinician perceptions	High – detailed FDA clearances, EU AI Act & MDR analysis, liability vignettes, proposal for standardized XAI reporting framework	This review consolidates feature selection (FS), graph neural networks (GNNs), and multimodal transformers into a XAI taxonomy with explicit links between XAI methods and specific points in the radiology and pathology clinical workflow, a detailed analysis of rules and ethics including clinical vignettes, and a suggestion for a regular XAI reporting framework for future research.

The progress is monumental, but a significant gap remains in the knowledge of applying XAI methods systematically in day-to-day clinical practice. Compared to most of the available research, which focuses on particular algorithms in isolation or with a specific clinical application, it does not provide a comprehensive interpretability metric or an inter-comparison of different aspects. The current literature does not address methodological standardization or the comparative effectiveness of various XAI methods. Furthermore, little attention has been given to the study of clinicians' perceptions or the pragmatic clinical impact of XAI-based tools, which have not been adequately examined. Promising results have been achieved, but further work is needed to research the proper fitting of data bias to a particular distribution, standardize interpretability to expose inherent bias regardless of our efforts, and address the intrinsic challenges introduced by multimodal instance contexts that clinical adoption brings. However, the gaps can be mitigated through a rigorous framework for reviewing available data, well-defined research opportunities, and recommendations for developing future research that highlight the successful clinical aspects of implementing XAI advancements.

This study outlines the primary objectives, which primarily involve synthesizing current XAI knowledge, applying XAI to clinical use cases, and addressing the challenges of interpreting medical images.

- Categorize and evaluate current XAI methodologies, document clinical applications in radiology and pathology imaging, identify limitations and

interpretability challenges, and propose future research directions centered on gaps.
- The review evaluates the efficacy and clinical effects of XAI procedures in the literature to provide evidence-based recommendations. Understanding physicians' views on XAI technologies, their real-world usability, and the integration of these technologies into clinical system is also essential.

Ultimately, achieving these objectives provides clear and actionable insights, fostering the effective translation of XAI advancements into practical medical imaging practices.

- This systematic study improves medical imaging XAI understanding with its complete synthesis.
- The review presents distinct value in consolidating dispersed studies methodically, clearly noting paramount techniques, and critically analyzing interpretability difficulties. The comparison structure of XAI methods and the identification of consistently effective techniques facilitate informed decision-making for the clinical execution of these methods.
- The peer-reviewed evaluation of the practical applicability and integration issues provided by the review contains actionable information valuable to clinicians, healthcare institutions, researchers, and policymakers.
- Well-defined recommendations are distilled from detected research gaps and endorse targeted, impactful forthcoming research endeavors, meaning XAI enhancement in robustness and clinical utility.
- Moreover, by explicating methodological limitations and focusing on standardized interpretability metrics, the review contributes to methodological rigor, enhancing the quality of future XAI research.

Thus, systematic review significantly advances XAI's theoretical understanding and practical applicability in medical imaging.

To methodically guide the systematic review, the following research questions are formulated:

- Which XAI methodologies are predominantly utilized in current medical imaging literature, and how do they differ in effectiveness and interpretability?
- What are the primary clinical applications of XAI reported across radiology and pathology imaging studies?
- What is the major interpretability, standardization, and clinical integration challenges associated with current XAI techniques in medical imaging?
- How do clinicians perceive XAI-based diagnostic tools' utility, effectiveness, and usability in real-world clinical settings?

These clearly defined questions systematically structure the literature search, analysis, and synthesis processes, ensuring rigorous evaluation of available evidence, clear identification of practical implications, and comprehensive reporting of findings.

The organization of this systematic literature review paper is outlined as follows. The paper commences with an Introduction that delineates the advent of AI in medical imaging, the pressing need for explainability, the historical evolution of XAI, the unique contributions of this review relative to previous surveys, the formulation of four research questions, and a succinct outline of the document. It is then succeeded by the Methodology of Systematic Review that describes the PRISMA 2020-compliant methodology, search strategy of major databases (2015–2025), duplication, inclusion/exclusion criteria, screening, full-text assessment, inter-rater consensus, and final selection of 133 studies. XAI Techniques in Medical Imaging is the third large section, which introduces a new taxonomy according to which existing methods of XAI can be divided into four families (saliency maps and heatmaps, attention mechanisms, model-agnostic methods, and rule-based/symbolic methods such as GNNs) with each method being described, mathematically formulated, and provided with strengths and limitations. The following paper Applications of XAI in Medical Imaging will cover actual, real-world clinical application in three areas: radiology imaging (X-ray, CT, MRI), pathology imaging (digital and whole-slide images), and multimodal medical imaging like PET-CT, MRI-PET with concrete examples given of the applications and the available summary tables of the studies included. The Clinical Impact of XAI in Healthcare section examines how XAI enhances a more accurate diagnosis, aids in clinical decision-making, and raises the degree of trust and usage by clinicians. This is then succeeded by Challenges and Limitse of XAI in Medical Imaging which critically discusses the lack of standardized interpretability measures, the problems of data bias and generalizability, challenges of clinical workflow integration, and the issue of ethics/legal concerns. An XAI section on Role of XAI on Medical Imaging helps illuminate differences in concepts (interpretability vs. explainability, black-box vs. white-box, post-hoc vs. ante-hoc) and comprehend why the concept of explainability has now become uncompromising in high-stakes medical imaging settings. Its Discussion discusses the new trends (transformers, federated XAI, causal and symbolic reasoning), elaborates on a detailed standardized reporting framework of future XAI research, and provides the

Table 2 Table of abbreviations

Abbreviation	Full Form	Abbreviation	Full Form
AI	Artificial Intelligence	LRP	Layer-wise Relevance Propagation
CV	Computer Vision	LIME	Local Interpretable Model-Agnostic Explanations
CNN	Convolutional Neural Network	ML	Machine Learning
CDSS	Clinical Decision Support Systems	MRI	Magnetic Resonance Imaging
CT	Computed Tomography	MHA	Multi-Head Attention
DL	Deep Learning	NLP	Natural Language Processing
DNN	Deep Neural Network	PET	Positron Emission Tomography
DT	Decision Tree	PCA	Principal Component Analysis
DARPA	Defense Advanced Research Projects Agency	PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
DeepLIFT	Deep Learning Important Features	ReLU	Rectified Linear Unit
FS	Feature Selection	SHAP	Shapley Additive Explanations
FL	Federated Learning		
GA	Genetic Algorithm	TL	Transfer Learning
GAN	Generative Adversarial Network	US	Ultrasound
GRU	Gated Recurrent Unit	SSL	Self-Supervised Learning
GNN	Graph Neural Network	ViT	Vision Transformer
Grad-CAM	Gradient-weighted Class Activation Mapping	XAI	Explainable Artificial Intelligence
HIPAA	Health Insurance Portability and Accountability Act		

future research directions necessary to overcome the existing barriers. Regulatory Implications of XAI in Medical Imaging is the next one, which considers the current (2025) regulatory environment, such as FDA clearances, the EU AI Act and Medical Device Regulation, and the importance of explainability in market authorization. Lastly, the paper has a Conclusion and Future Work section that summarizes the most important findings, restates the transformational power of XAI when appropriately applied, and recommends future interdisciplinary collaboration to reach the goal of robust, trustful, and clinically accepted XAI systems in medical imaging. The list of abbreviations used in the current study have been listed in Table 2.

Methodology of systematic review

This systematic review follows PRISMA guidelines to ensure methodological transparency and clarity in evaluating studies on XAI in medical imaging.

Literature search strategy

Databases searched: PubMed, IEEE Xplore, Springer, Elsevier, Wiley, and ACM using keywords like "XAI", "medical imaging", "radiology", "pathology", and "interpretability. The search string used across all databases was: ('explainable AI' OR XAI OR 'interpretability' OR 'saliency map' OR Grad-CAM OR LIME OR SHAP) AND ('medical imaging' OR radiology OR pathology OR 'MRI' OR 'CT' OR 'X-ray'), Filters: English language, years 2015–2025.

One of the records was uploaded to application and the initial automated step of deduplication was performed with the usage of the title, author, and the publication year in which duplicates were identified. The remainder of the records were then exported to EndNote X9 where additional manual duplication was performed, and additional duplicates were discovered. This was done in two steps to ensure that no duplicates were obtained, and 691 different records were ready to screen. This represents in Table 3 in detail step by step.

Table 3 Step-by-step deduplication protocol

Step	Action	Tool & Settings	Output
1. Database Export	All 980 records exported from databases	Native exports from each platform	
2. Import into EndNote	Created a **new EndNote library**	EndNote X9 → *File* → *Import* → *File* → Import Option: **Reference Manager (RIS)**	980 references
3. Automated Deduplication	Ran EndNote's built-in duplicate finder.	*References* → *Find Duplicates*	691

Inclusion criteria

- Scholarly articles subjected to peer review (2015–2025)
- Studies focusing explicitly on XAI methods
- Clinical applications in medical imaging (radiology/pathology)
- Systematic reviews, original research, and case studies
- Conference Papers and conference Abstracts

Exclusion criteria

- Articles published before 2015
- Non-English language articles
- Editorials, opinion pieces, and conference abstracts
- Studies unrelated to medical imaging

Table 4 outlines the inclusion and exclusion criteria used in the study to determine which studies were considered for the systematic review.

Table 4 Inclusion and exclusion criteria

Inclusion Criteria	Exclusion Criteria
Scholarly articles subjected to peer review (2015–2025)	Articles published before 2015
Studies focusing explicitly on XAI methods	Non-English articles
Clinical applications in medical imaging	Articles unrelated to medical imaging
Systematic reviews, original research, and case studies, conference papers and conference abstracts	Editorials, opinion papers

Screening and selection

Data extraction and screening were done in both phases with the differences coming to a consensus or adjudication. Title/abstract screening was done in Rayyan, and decisions were made to include blinded. The same was done with full-text assessment. Disagreements in terms of title/abstract screening and full-text screening were addressed through consensus/adjudication. The data were extracted using a piloted form in Microsoft Excel to make sure that data extraction was conducted twice to reduce the risk of selection bias and to increase reliability.

Cohen kappa and percent agreement were used to measure inter-rater agreement. Title/Abstract screening and full-text screening and data extraction had 0.82 (95% CI: 0.78–0.86) and 0.79 (95% CI: 0.74–0.84), individually, suggesting that there is significant and almost perfect agreement. Conflict resolution was done by deliberation and adjudication. Table 5 indicates the measurement of the observer agreement in categorical data expressed in terms of inter-rater agreements at different steps of review.

Titles and abstracts were evaluated, followed by a comprehensive evaluation of the full texts, with discrepancies addressed using a consensus-driven method. Figure 3 depicts the PRISMA flow diagram of the literature selection process, delineating the phases of identification, screening, and inclusion of studies in the systematic review.

According to the PRISMA 2020 [27] protocol, our systematic search was conducted in the following published databases: Springer, PubMed, IEEE Xplore, Elsevier, Wiley, and ACM yielding They located 980 records by systematic searches conducted by following the elimination of 289 duplicates and evidently irrelevant records, 691 records were left in title and abstract screening with 209 of them eliminated. The rest 482 articles were obtained in full texts and evaluated against the inclusion criteria; 263 articles would not satisfy the inclusion criteria and were excluded. This gave a provisional inclusion of 219 studies. In the process of final data extraction and quality appraisal, 86 more articles were eliminated due to the following reasons: not explicitly applied XAI methodology ($n = 46$), only conference abstract or poster ($n = 21$), no obtainable full text despite considerable efforts ($n = 13$), due to redundancy or duplication of results ($n = 6$). Finally, 133 articles were incorporated in the qualitative synthesis and end systemic review. This methodical methodology guaranteed the inclusion of only the most pertinent and high-caliber publications in the research. The new PRISMA flowchart (Fig. 3) provides a comprehensive overview of this process.

Data extraction and synthesis

The mined data was through the assistance of structured study identifiers, methodology, XAI techniques, clinical application and outcomes. The form was put through 10 tests and refined. Data were extracted using a standardized form covering authorship, year, imaging modality, XAI techniques, applications, and outcomes. Analysis was done through comparative tables and thematic mapping, with the two key fields showing a 0.85 correlation.

Explainable artificial intelligence techniques in medical imaging

Advancements in medical imaging have been substantially enhanced by the evolution of artificial intelligence models, especially DL algorithms. Nonetheless, the opaque characteristics of these models hinder clinical adoption owing to insufficient openness and trust. Investigating XAI is essential for tackling these difficulties and improving the transparency and interpretability of machine learning models. This section categorizes popular XAI methods into four groups. Activation maps, heatmaps, attention mechanisms, model-agnostic explanations, rule-based symbolic approaches, basic ideas,

Table 5 Inter-rater agreement (Cohen's κ) across review stages

Stage	κ (Cohen's)	95% CI	% Agreement	Interpretation
Title/Abstract Screening	**0.82**	(0.78–0.86)	91.20%	**Almost Perfect**
Full-Text Screening	**0.79**	(0.74–0.84)	88.70%	**Substantial**
Data Extraction (Key Fields)	**0.85**	(0.81–0.89)	93.40%	**Almost Perfect**

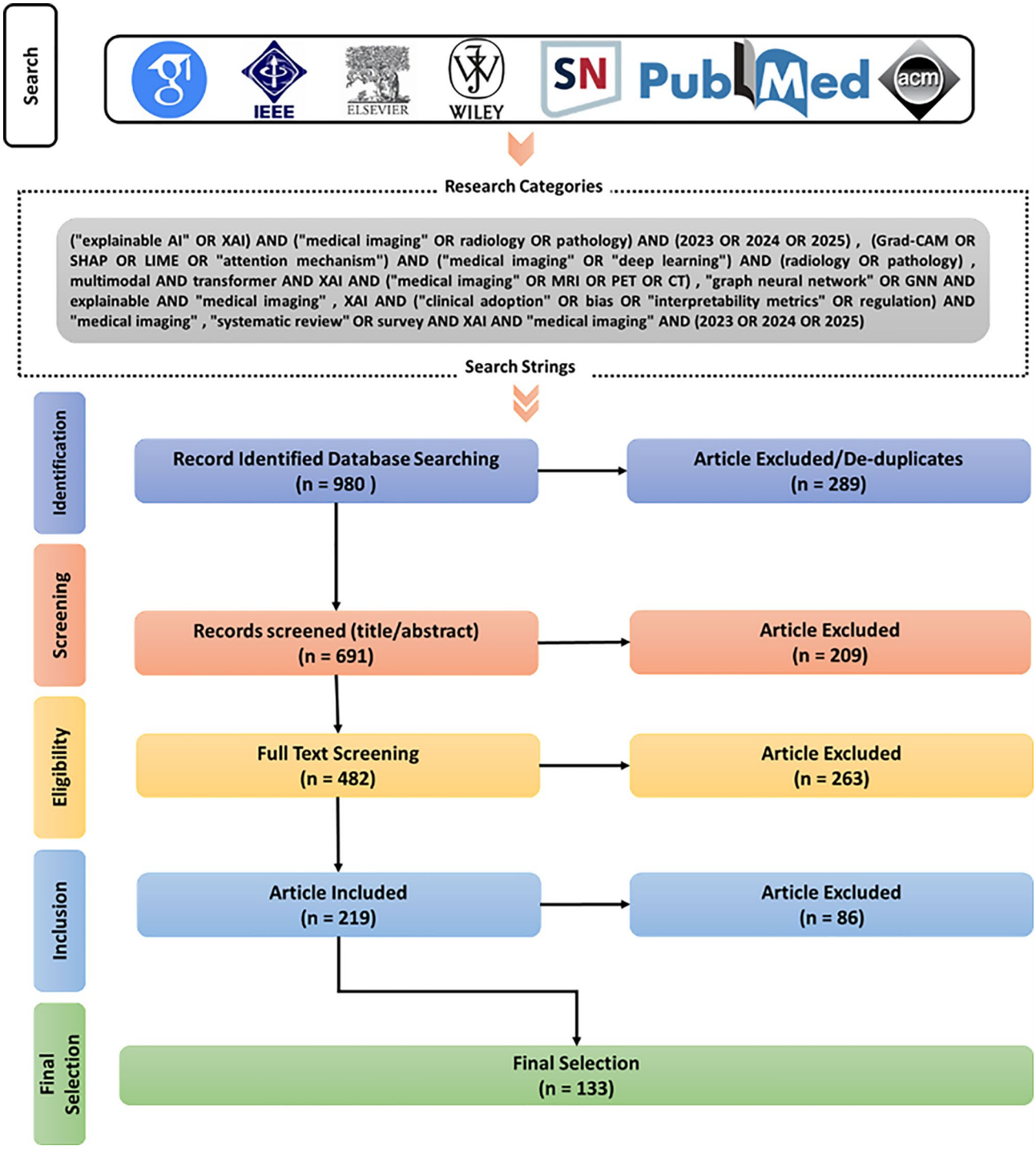

Fig. 3 PRISMA flow diagram of literature selection

mathematical and practical implementations, strengths, and shortcomings. Figure 4 illustrates the taxonomy of explainable medical imaging techniques employed to enhance interpretability in AI-driven medical technology. As the results demonstrate, XAI enhances the interpretability of models, promoting model trust and clinical validation, especially in healthcare systems [28].

Saliency maps and heatmap-based methods

Saliency maps are graphical representations highlighting regions within images that significantly influence model predictions. These methods are beneficial for medical imaging because they visually convey what features contribute most prominently to specific diagnostic outcomes.

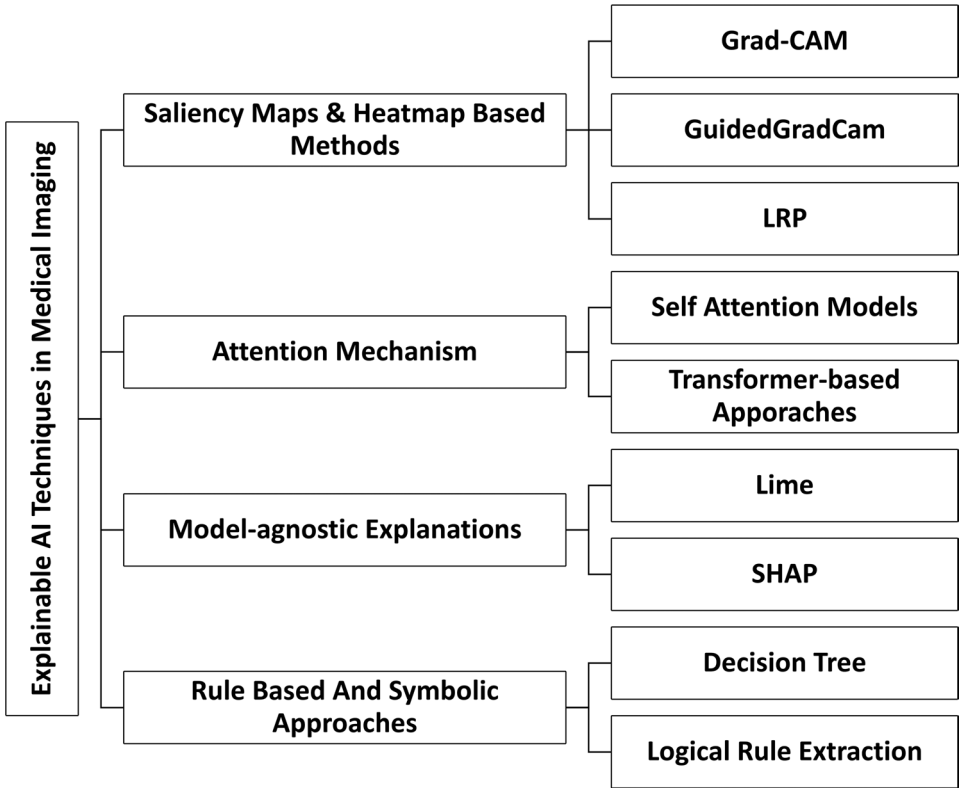

Fig. 4 Taxonomy of Explainable techniques in medical imaging

Gradient-weighted Class Activation Mapping (Grad-CAM)
Grad-CAM [29] is a popular method that enables the identification of key areas in an image and the prediction of a CNN. In contrast to the comparatively straightforward reliance on feature maps at the last convolutional layer apparent in the traditional Class Activation Mapping (CAM) [30] approach, Grad-CAM leverages the gradient that traverses down to the final convolutional layer to create coarse yet useful localization maps. These maps help bring to the model's attention areas of the image that are crucial for making a decision, thereby enhancing the interpretation.

In Grad-CAM, the significance weights α_k^c of every feature map A_{ij}^k In the last convolutional layer, the gradients are calculated by taking the gradient of the class score. y^c Concerning each feature map. In detail, the contribution of each attribute map to the predicted class score for a known class c is computed. The value for each feature map is weighted with the dependence on the following Equation (A1) in Appendix A:

Here, α_k^c represents the significance weight of the kth feature map for class c, A_{ij}^krefers to the activation value of the kth feature map at the spatial location $i, j$and y^c is the predicted output score for the class c. The sums are over all spatial positions in the feature map, and Z is a standardization factor that is usually the total number of spatial places in the array.

After computing these important weights, the final class activation map, Guided-Grad-CAM [29], is generated by summing the weighted feature maps. This process highlights the areas of the image that are most influential for class prediction. The resulting Grad-CAM can be shown as follows in Equation (A2) in Appendix A:

This output displays a heatmap over the input image, highlighting the areas of focus for the model when predicting a class c. In medical imaging, Grad-CAM enables clinicians to identify specific and pathological regions, thereby enhancing diagnostic confidence and improving the interpretability of the model.

In this equation, A_k represents the feature map of the convolutional layer. Four feature maps are attributed to four gradient-based weights. The weight provides a heatmap of what areas in the image are critical to the decision of the model giving clinicians a visual depiction of what in an image led the model to its decision as illustrated in the Equation (A3) in Appendix A:

In this case, y_c represents the score of the category c in the classes and Z represents the count of the pixels within the feature map. The procedure is useful in CNN-based FS models of medical imaging because it enables clinicians to understand the rationale behind the selection of specific features within a DL model.

The role in enhancing the interpretability of AI-based FS is also the enhancement of the interpretability of AI-based Hybrid FS techniques, which are the combinations of classical FS techniques and the DL models. As examples, Principal Component Analysis (PCA) can be used to reduce the dimensionality of high-dimensional medical images and CNN can be used to find features. Equation (A4) in the appendix A is the PCA equation:

The second matrix W of eigenvectors (principal components) of the original dataset X produces the transformed dataset X. Through this hybrid style, we strike the balance between standard statistical capabilities and recent DL capabilities to choose such capabilities in an efficient and interpretable manner. Also, since there is a Hybrid FS, it is possible to use the Genetic Algorithms (GAs) to reach such optimization, modeling natural evolutionary processes. Thus, the GA fitness value can be formulated as follows in form of Equation (A5) in Appendix A:

In the case of $f(S)$ is the fitness of the selected feature subset S, and accuracy (S_i) is the performance of a model that has been trained on the feature subset S_i This approach can be used to identify the most important features, enhance the interpretability of a model, and maintain high performance.

In the finality, AI-based FS methods are required to be interpretable and explainable in order to be used in clinical settings. The SHAP, Grad-CAM, and Hybrid FS techniques and the XAI are also promising methods of enhancing transparency. This will assist the clinicians in understanding when to rely on and apply the AI model when making decisions. Even though other methods including SHAP are still in their active developmental phase and are yet to be put into practice as a normal practice in clinical FS, they can still enhance reliability and clinician confidence. It can be interpreted, and clinicians are able to know when to rely on AI models to make decisions. With the use of these two methods, an AI-based FS system will be capable of becoming not only high-performing but also clinically viable, which is why it will pave the way to its broader use in healthcare. Table 6 compares different interpretability approaches in FS and indicates their advantages and disadvantages.

Table 6 Comparison of interpretability methods in FS

Method	Description	Advantages	Limitations
XAI methods [31]	Transparency, trust transparent.	interpretability, accountability	complexity, scalability. explain models.
Grad-CAM [32]	Highlights important image regions in CNN decisions.	Provides visual explanations.	Limited to CNNs, only works on images.
Feature extraction & selection [33]	Hybrid genetic algorithm for optimal selection.	Reduces dimensionality, boosts efficiency, captures diverse data	Computationally heavy, prone to overfitting, depends on tuning.

Medical practitioners must meticulously examine patient symptoms before rendering a diagnosis, and AI would need to emulate this process to gain widespread acceptance. To make DL, ML, and Natural Language Processing (NLP) models more transparent, XAI provides visibility into their decision-making data. Recently developed XAI methods for medical imaging are classified and described in this chapter. As XAI ideas are applied to medical images and text data analysis, DL systems and their explanations can be enhanced [34].

Guided backpropagation

Guided backpropagation [35] is an attribution method tested for robustness in classifying structural MRI data, revealing that some widely used methods produce inconsistent outcomes. This improves visualization clarity and lucidity, focusing on pixel-level information that is relevant. The guided Backpropagation changes gradients as seen in Equation (A6) in Appendix A:

Where $R^{(l)}$ represents modified gradients at the layer l, $f^{(l)}$ characterizes the activations, and L represents the loss function. The $\odot$ symbol denotes element-wise multiplication. This selective gradient propagation forms visually coherent saliency maps, which are invaluable for detecting lesions or abnormalities in small–scale medical images.

Layer-wise relevance propagation (LRP)

LRP [36] creates an explicit understanding of neural network decisions by distributing relevance scores to inputs and backpropagating the scores through the network layers. It enables us to identify precisely which input features contribute most significantly to model predictions. LRP relevance is also calculated recursively by the following Eq. A7 in Appendix A:

Here, $R_i^{(l)}$ is the relevance of neuron i at layer l, z_{ij} denotes weighted activations from a neuron i to neuron j, and $\in$ ensures numerical stability. This approach is particularly efficient in medical imaging due to its transparency and the accuracy of pathological marker localization.

Deep learning important features (DeepLIFT)

DeepLIFT calculates the difference between the model's output prediction and a reference point, propagating the importance of each feature in contributing to the final prediction [37]. DeepLIFT is highly effective for explaining models that classify medical images, such as MRI or X-ray scans. It highlights key areas, such as specific tissue regions that have a foremost influence on the model's

likelihood, offering transparency for diagnosis assistance [38]. The basic equation for DeepLIFT is illustrated in Eq. A8 in Appendix A:

In this equation, y_i represents the model's output for the ith class. The term $f_i(x_j)$ refers to the output for the input x_j, while $f_i(x_0)$ denotes the output for the reference point x_0. The change in the output with respect to featurex_j is represented by Δf_j. The summation $\Sigma_{j=1}^{n}$ aggregates these differences across all features, calculating how the model's output for the i^{th}class is affected by variations in the features.

Integrated Gradients (IG) sum up gradients across the baseline to the input and have been applied in tumor detection on both CT and MRI scans [39], as presented in the Equation (A9) in Appendix A.

This equation defines the integrated gradients' attribution for feature x_i, measuring its contribution to the model output $f(x)$ relative to a baseline x'.

Occlusion Analysis (OA) systematically covers pieces of the input image as a method of testing prediction consistency; this has been applied to detecting lung nodules in CT [40]. To achieve robust saliency maps, RISE randomly perturbs regions of an image and has been used in tumor segmentation [41].

Attention mechanisms

Attention mechanisms are beneficial because they enable the explicit modelling of dependencies within a dataset. By examining the beginning or end of the dataset, we can identify how the model, during the decision-making process, focuses on particular parts of the data.

Self-attention models

The self-attention mechanism [42] enables the model to attribute varying significance to distinctive regions or parts of an input image. The attention mechanism works with the following calculation, as expressed in Equation (A10) in Appendix A:

Q, K and V characterize Query, Key, and Value matrices obtained from feature maps, and d_k indicates dimensionality. Self-attention in medical imaging enables models to emphasize clinically significant areas of the image, improving accuracy in lesion detection and disease classification.

Transformer-based approaches

Transformers [43] rely on the self-attention mechanism across multiple heads, enabling parallel computation and modelling complex relationships. The computation of attention across various heads in a transformer-based model is indicated by Eq. A11 in Appendix A:

Each attention head is computed individually, as defined in Equation (A12) in Appendix A:

Transformer models have revolutionized the landscape of medical imaging tasks, particularly in scenarios where achieving global and local interactions in images is crucial, such as tumor segmentation and disease staging.

This technique extends attention by utilizing multiple attention heads, enabling the model to focus on distinct parts of the image concurrently. In breast cancer detection, Multi-Head Attention (MHA) can focus on both the texture of tissue and the contours of masses in mammograms to increase accuracy. MHA applies the attention procedure in parallel throughout separate heads, enabling the model to focus on multiple parts of the image at once.

Model-agnostic explanations

Valuable for their flexibility, model-agnostic methods enable an understanding of any predictive model without restrictive conditions on its internal structure.

Local interpretable model-agnostic explanations (lime)

LIME [44] constructs local approximations of model predictions by using simple explainable models such as linear regression to explicate the composite model's predictions in the surrounding neighborhood of the instance of interest, as demonstrated in Equation (A13) in Appendix A:

Here, f represents the primary model, f is the interpretable alternate, L is the loss function for local fidelity, π_x is the locality measure, and $\Omega(g)$ penalizes complexity. The AI-driven diagnostic rationale of LIME has been successfully used in medical imaging to improve clinicians' understanding of AI-driven diagnostic studies.

Shapley Additive explanations (SHAP)

XAI is an important issue in improving the transparency, interpretability, and clinical trust of DL models in breast cancer (BC) histopathology. Such XAI methods as Grad-CAM, SHAP, and LIME are beneficial to attempt to reveal the decision-making process of the AI models, which would make sure that clinicians know how decisions are explained. This will be critical to the incorporation of AI in clinical practices, as it will establish trust in AI-based decisions. XAI enhances diagnostic precision through the provision of understandable visual information on the explanation of tumor classification and segmentation, which enable clinicians to warrant and rely on AI outcomes. Moreover, XAI methods help to curb the threat of biases by determining and fixing discriminative patterns in data. In general, XAI will enable the wider use of DL models in the clinical environment because it ensures model transparency, lowers the risks of errors, and enhances the level of equity of diagnostic output in a wide range of populations [45].

SHAP [31] appoints importance scores to the input features, providing model-agnostic yet exact feature contributions; however, it is too computationally resource-intensive to be visualized directly. SHAP offers a theoretically grounded interpretability metric reflecting Shapley values, indicating several contributions of each feature as depicted in Equation (A14) in Appendix A:

In this equation S is a subgroup of all features, eliminating the feature i, M is the total number of features, and $f_x(S)$ denotes prediction using attributes in S. SHAP values offer consistent explanations widely accepted in medical diagnostics.

Anchors (anchors: high-precision model-agnostic explanations)

Anchors explain model decisions by discovering small sets of features (anchors) that are good predictors to the model's output. In order to predict the commencing of diabetic retinopathy, Anchors explain model decisions by illuminating important areas, such as blood vessels and lesions. In predicting diabetic retinopathy from retinal images, Anchors mark very important parts of the retina, such as blood vessels and lesions. By emphasizing these areas, the decisions made by the model are explained in terms that are consistent with medical knowledge, which facilitates making the model more transparent as explained in Equation (A15) in Appendix A:

In the present context, where A is the set of selected features or can be referred to as anchor, and 1 is the set of features that are considered as irrelevant. The term P(A) denotes the probability of the output y with anchor features A. The objective is to identify an anchor that maintains the predictive behavior of a model, so that the model does not get affected by changes in the irrelevant features. This approach is focused on developing a stable and strong set of features, which maximizes the accuracy of the predictions, even if we have irrelevant features.

Feature selection is a preprocessing method that is used to find important features in applications such as medical imaging, biomedical signals, and DNA analysis. This method is useful to reduce dimensionality and improve the understanding of the causes of the disease [46] in one case study using the real data of a patient. Counterfactual Explanations to generate what-if scenarios by making small changes in the input features thus affecting what predictions are used, in case of breast cancer detection in mammograms [47].

Perturbation-based Importance (PI) measures which are performed via random perturbation of the qualities of inputs and the degradations of performances are observed, such as in MRI-based Alzheimer diagnosis [48]. Such approaches are complementary to gradients and attention-based approaches, and provide more options for clinical interpretation in different modalities.

Rule-based and symbolic approaches

Decision Trees (DTs)

DTs [38, 49] explicitly represent decision-making processes through hierarchical rules as illustrated in Equation (A16) in Appendix A:

Those rules are transparent and interpretable, making them easily understood by clinical practitioners and thus enabling simple validation and direct implementation into diagnostic procedures.

Logical rule extraction

Extracts logical rules [50] from model's forecasts. For example, it could form if-then regulations which match the model's decision boundaries as shown in Equation (A17) in Appendix A:

Each predicate p_i represents logical conditions on features. Extracted rules offer significant clinical utility due to their transparency, accountability, and ease of integration into clinical practice guidelines.

This work demonstrates that with a powerful tool, GNNs, such relationships in data can be effectively modeled. Projects such as modeling relationships in multimodal medical images have shown GNNs to be powerful. GNNs are suitable for multimodal data because each modality can be considered a node in the graph, and the relationships between different features across modalities can be represented as edges within these nodes. In general, GNNs are applied for learning the representation of nodes to capture both local and global dependencies by aggregating information from neighboring nodes. This is the equation corresponding to the graph convolution operation in GNN, expressed in Equation (A18) in Appendix A:

Where:

- $h_v^{(k+1)}$ is the updated feature representation for node v at the $k+1-th$ iteration,
- $N(v)$ represents the neighbors of node v,
- $W^{(k)}$ is the weight matrix for the kth layer,
- $b^{(k)}$ is the bias term, and
- σ is an activation function such that Rectified Linear Unit (ReLU)

Considering the multi-modal FS, GNNs is able to map to integrate multi-modal features across modalities that are variable. Dependence is captured in between the features with different resolutions or modalities missing. GNNs lend themselves more to the fusing of information better and this is important to deal with complex medical image analysis tasks.

Bayesian Fusion is another method of solving the resolution and missing modality problem in multimodal FS. This approach considers FS as a probabilistic model and computes the probability of various feature subsets

based on the available data on all the modalities. Bayesian fusion assists in combining the uncertainty and variance of the data across modalities. The general formulation of Bayesian fusion models is shown in Equation (A19) in Appendix A:

Where:

- $p(M)$ is the posterior probability of the feature set F given the modality set M,
- $p(M)\,p(F)$ is the likelihood of observing the modalities M given the features F,
- $p(F)$ is the prior probability of the feature set,
- $p(M)$is the marginal likelihood of the modalities.

Self-supervised learning (SSL) can also be applied to address the challenges of scarce labeled data and missing or incomplete modalities. Proceeding on supervised, in contrast to SSL, uses the inherent structure in the data to generate sensible feature representations with less dependence on annotation. More specifically, reconstruction-based SSL methods seek to reduce the distance between original and reconstruction data, which can be formulated as demonstrated in Equation (A20) in Appendix A:

Where:

- x_i is the original data point like a modality,
- $\widehat{x_i}$ is the predicted data point after applying the SSL model,

In the scope of multi-modal FS, SSL can be used to extract representations from incomplete data, to infer the missing information from one modality using the other available media. The main strength of such an approach is that it works well if there is a lack of training data and imaging modalities are not available. Table 7 summarizes different XAI approaches, showing some of their main characteristics and limitations.

Table 7 Summary of XAI techniques with key features and limitations

XAI Technique	Key Features	Limitations
Grad-CAM [51]	Visual heatmaps, easy interpretability	Low spatial resolution; CNN-specific
Guided Back-propagation [52]	Clear pixel-level attribution	May produce noisy explanations
LRP [53]	Detailed explanations layer-by-layer	Computationally intensive
Self-attention Models [53]	Contextual relevance, adaptable	Complexity in training, potential bias
LIME & SHAP [54]	Model-agnostic; flexible	Approximation-based; not always accurate
DTs/Rule-based [55]	Highly interpretable, human-readable	Limited complexity handling
DeepLIFT [56]	Attribution based on reference point	May not work well with highly non-linear models
Transformers [57]	Efficient multi-head attention, captures long-range dependencies	Requires large datasets for training
GNNs [58]	Excellent for multimodal data fusion, graph-based relationships	Computationally expensive, complex models

C4.5 algorithm (improved dt algorithm)

C4.5 is an enhancement of traditional DTs, introducing pruning and handling both continuous and categorical data. C4.5 can be used for medical support, such as diagnosing bone fractures from X-rays, where it splits features based on severity and fracture type. C4.5 is particularly useful in medical decision support systems, such as diagnosing bone fractures from X-rays. The algorithm divides data based on fracture type and severity, producing a DT that clearly outlines the decision-making process, making it easier for healthcare professionals to interpret and use in Equation (A21) in Appendix A.

In this context, S represents the current dataset, while S_v refers to the subset of data corresponding to a particular value v. The term $Entropy(S)$ measures disorder or uncertainty within the dataset. The information gain is determined by subtracting the weighted summation of the entropies of the subsets S_v from the entropy of the original dataset S. This measure helps to identify the feature that provides the most significant reduction in uncertainty, making it valuable for deciding tree algorithms in ML. Table 8 presents a summary of the attribute attribution techniques employed in XAI.

Comparative analysis of method families

Although each of the separate models—gradient-based, perturbation approach-based, attention-based, and rule-based—has its own view on model interpretability, clinicians and developers may find general directions in which one is best suited to facilitate certain choices. Different families of XAI methods differ not only in terms of their computational expenses but also in their ability to explain themselves on a clinical level, remain stable under high-stakes scenarios, and suit specific modalities of imaging.

These comparative dimensions are summarized in Table 9. Gradient produces highly efficient and intuitive heatmaps, which are frequently used; however, it can bring into sharp focus non-relevant structures and have a non-global rank. Although computationally expensive, perturbation methods can provide more model-agnostic" what-if" answers that are valuable for auditing and quality assurance. The attention-based mechanisms are effective in handling large and complex inputs, such as whole-slide pathology images, but their interpretations do not necessarily correspond to clinical reasoning and require large datasets. Model-agnostic attribution methods, such as LIME and SHAP, are not architecture-specific; however,

Table 8 Summary of feature attribution techniques in XAI

XAI Method	Description	Imaging Modality	Clinical Use Case	Applications
IG [39]	Integrates baseline-to-input gradients to predict characteristics.	CT	Tumor Detection in CT and MRI	Tumor detection in CT and MRI scans.
OA [40]	Blocks image parts to test predictions.	CT	Lung Nodule Detection	CT lung nodule detection.
RISE [41]	Randomly selecting image portions yields robust explanations.	CT	Tumor Segmentation	Tumor segmentation in radiology.
Counterfactual Explanations [47]	Changes input features and observes predictions to create "what-if" scenarios.	Mammograms	Breast Cancer Detection	Breast cancer detection in mammograms.
LIME [59]	Approximates large models with simpler models around specific predictions for localized explanations.	Retinal Fundus	Optical Coherence Tomography, Diabetic Retinopathy, Breast Cancer, Ocular Illness Predictions	Optical Coherence Tomography, Diabetic retinopathy, breast cancer, and ocular illness predictions using fundus pictures.
SHAP [60]	Determines prediction feature contribution scores using cooperative game theory Shapley values.	Chest X-ray	Pneumonia Detection, Lung Cancer Detection	Pneumonia detection in chest X-rays, lung cancer detection in CT scans.
CAM [61]	Creates heatmaps to emphasize model-predictive image regions.	MRI	Tumor Localization	Tumor localization in MRI.
Grad-CAM [62]	Highlights key locations using output class gradients relative to convolutional feature maps.	MRI	Tumor Segmentation, COVID-19 Detection	Tumor segmentation, COVID-19 detection in chest X-rays.
GradCAM [63]	ResNet50 and Grad-CAM for interpretable brain tumor detection.	MRI	Interpretable brain tumor detection using MRI.	Brain tumor detection, medical diagnostics, healthcare decision support.
LRP [64]	Informs input layer about output relevance scores and visual factors influencing the choice.	MRI/CT	Alzheimer's Biomarker Detection	MRI/PET biomarker detection for Alzheimer's.
Surrogate Models [65]	Approximates complex models with simpler, interpretable models.	CT	Chest X-ray Analysis	Explaining predictions in chest X-ray analysis.

Table 9 Summary of feature attribution techniques in XAI

Method family	Strengths	Weaknesses	Typical use-cases/modalities	Reliability in high-stakes settings
Gradient-based	Speedy; visually intuitive heatmaps; CNN friendly	Noise-sensitive; Highlights superfluous textures; local explanations only	CT/MRI lesion localization; chest X-ray triage	**Medium –** effective in more superficial lesions, less solid in more subtile pathology
Perturbation/occlusion	Architecture-agnostic; less architecture-dependent explanations; intuitive what-if-what-will-be insight	Calculational expensive; unstable if traits are significantly connected	Robustness testing; radiology QA; testing or validating CNN outputs	**Medium –** good for audits, but sucks for real time
Attention-based (Transformers, ViTs)	Maintains long-term dependencies; models dial explanations	Massive data needed; attention maps may not reflect clinical relevance.	Whole-slide pathology, multimodal MR/CT	**Medium-High -** Low prospective validation, potential
Model-agnostic attribution (LIME, SHAP, surrogate models)	Any black box; quantitative attribution; tabular+image	High variance; explanations vary per run; hard to scale.	Cross-modal auditing; research prototype	**Medium** – informative but unstable in production
Rule-based/symbolic (DTs, expert systems)	Reasonable, transparent decision reasoning; easy auditing	Scalability issues; complicated nonlinear patterns	Structured Triage, structured decision support protocols,	**High (when accurate)** – preferred where accountability is critical

they can exhibit slow convergence across runs. Lastly, rule-based or symbolic systems remain the gold standard for transparency and accountability, and, like the other categories, do not scale well to highly nonlinear data.

As shown in this comparative synthesis, a structured viewpoint may provide a sense of direction for practical XAI implementation in radiology, pathology, and multimodal medical imaging, where both interpretability and reliability are essential properties for the acceptance of the technology by the clinical community.

Applications of XAI in medical imaging

Explorable methods of XAI have shown substantial promise in various domains of medical imaging, providing deeper insight into the decisions made by the model and enhancing clinician trust. This section describes the

generalized applications of XAI in radiology imaging, pathology imaging, and multimodal medical imaging, and presents specific use cases, methodologies, and clinical effects.

Radiology imaging

XAI has mainly been used to explain radiographs and CT images. These imaging modalities play a central role in diagnosing diseases such as pneumonia, lung cancer, tuberculosis, fractures, and thoracic injuries, among others [48]. DL models, specifically CNN, have demonstrated promising accuracy but lack transparency in their learning process. XAI approaches, including Grad-CAM and attention mechanisms, strongly explain CNN decision-making by highlighting abnormal regions in the images [66].

For example, Grad-CAM, when applied to chest X-rays used for pneumonia detection, locates critical pulmonary infiltrates and assists radiologists in visually interpreting the models. Likewise, guided Backpropagation has identified slight fracture areas in skeletal X-rays, minimizing diagnostic errors [67]. For CT imaging, particularly in lung cancer screening, saliency map-based methods can significantly expedite lesion detection by accurately localizing nodules, thereby increasing interpretability and facilitating clinical decision-making [68]. XAI techniques, therefore, enable the development of transparent and trusted decision-support systems, resulting in more verifiable, clinically confirmed interpretations in radiographic and tomographic analysis [69].

MRI is a crucial imaging modality for detecting and characterizing neurological, musculoskeletal, and oncological lesions [70]. The "black-box" outputs generated by DL algorithms applied for automated lesion detection or classification in MRI are standard in these situations. XAI strategies address these problems by proposing visualizations that highlight features that influence diagnostic predictions [71].

Attention-based transformer architectures, for instance, are used on MRIs to accurately detect intricate lesion borders in brain tumors, precisely visualizing the tumor's extent and shape [72]. LRP in multiple sclerosis detection has demonstrated its ability to identify specific demyelinated plaques, thereby facilitating radiologists' detection of subtle pathological changes [73]. In addition, multimodal MRI that utilizes XAI methods has been exceptionally successful in detecting Alzheimer's and dementia, providing insights into which brain regions are crucial for model analysis and associating specific brain areas with diseases [74]. XAI-enhanced MRI predictions offer radiologists transparent and validated insights, playing a crucial role in patient management strategies [50, 75].

To better demonstrate the Grad-CAM process, Figure 5 presents a step-by-step workflow on lung CT images. The CNN-based classifier operates on the input image, which is then further processed by Grad-CAM, producing class-discriminative localization maps. These maps are then superimposed in the form of heatmaps on the original CT image, showing which areas of the lesions affected the model's decision. Visual depiction empowers clinicians to accurately identify the focus areas of the model, thereby creating further transparency and confidence in the diagnosis that is made.

On a clinical note, the visualization supports the rapid and accurate recognition of nodules, enabling radiologists to correlate AI-based diagnostics with the anatomical and pathological context, thereby substantially reducing the number of false negatives and overlooked lesions.

Pathology imaging

Digital pathology refers to the digitization of histopathological slides for detailed analysis, disease classification, and prediction. Traditional pathological examinations undertaken by human pathologists are time-consuming, cost-effective, and subject to subjectivity [76]. AI-based automated diagnostic models, primarily CNN-based classifiers, have significantly enhanced the efficiency and reproducibility of pathology imaging assessment [77].

Nonetheless, their sophistication requires clear explanations to gain clinical approval. XAI methods, such as self-attention mechanisms and LIME-based local explanations, have seen significant growth in slide-level and patch-level interpretation [64]. For example, self-attention models used in analyzing breast cancer histology can reveal vital tumor regions on whole-slide images, with supposed diagnostics such as nuclear atypia or mitotic activity being verifiable. These methods illustrate to pathologists why some slides were identified as

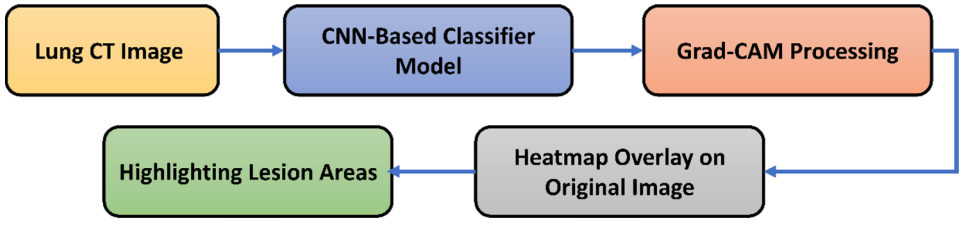

Fig. 5 Example of Grad-CAM visualization for lung lesion detection

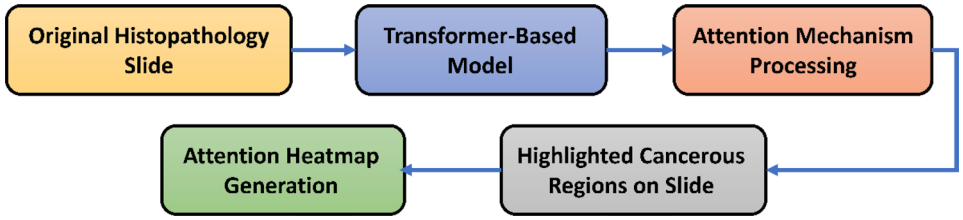

Fig. 6 Attention heatmaps in histopathological cancer detection

malignant or benign, significantly increasing diagnostic confidence [78].

The diagnosis of minimal cellular atypia highlights the importance of histopathological diagnostics in detecting early cancer or pre-cancerous changes. XAI methods, such as SHAP and attention-based models, effectively provide pixel-level or cellular-level explanations for AI-driven predictions of cellular abnormalities.

For instance, SHAP values derived from histological image analysis identify which cellular structure like size, density, or texture of nuclei plays the most significant role in determining malignant classification. DL model for classification of histopathological medical images, focusing on breast, colon, and lung cancers [79]. Attention-based methods have identified normal cell clusters in prostate and colorectal cancer slides, clearly depicting the pathological features utilized by the DL model [80]. These granular-level explanations enhance the transparency of diagnostic AI models, significantly facilitating interaction between pathologists and AI, and leading to the accurate and rapid detection of cellular abnormalities. XAI techniques like LIME and Grad-CAM in medical image classification tasks, emphasizing their importance in improving model transparency and fostering trustworthy AI systems [81].

To demonstrate the utility of attention mechanisms, Figure 6 illustrates the steps of a transformer-based model applied to histopathology slides. The first step involves an original input slide, followed by processing within the model based on attention. This creates the attention heatmap, which, when applied to the slide, displays the cancerous areas that were the major contributors to the prediction. This will enhance its interpretability, as clinicians will be able to justify the model's decision based on areas that need to be checked, informed by their visualization.

Multimodal medical imaging

Multimodal imaging is a hybrid approach combining data from multiple imaging sources, such as PET-CT, MRI-PET, and US-MRI, providing complementary details of both anatomical and functional disease states [82]. AI models with the capacity to work with multimodal data significantly improve diagnostics, but their complexity requires transparent and interpretable methods [83].

Table 10 Prominent studies applying XAI techniques in clinical imaging contexts

Citations	Imaging Modality	Clinical Application	XAI Technique	Key Findings
[73]	MRI	Multiple sclerosis diagnosis	LRP	Highlighted demyelination areas; enhanced diagnostic precision
[87]	Chest X-ray	Pneumonia detection	Grad-CAM	Localized pulmonary infiltrates; improved clinician trust
[88]	CT scan	Lung cancer detection	SHAP	Identified high-risk nodules; improved interpretability
[89]	Digital Pathology	Breast cancer detection	Attention Models	Precisely localized malignant cells; improved diagnostic accuracy
[90]	MRI-PET	Alzheimer's disease	Multimodal Attention Transformers	Improved diagnostic accuracy; precise multimodal feature attribution

XAI methods, specifically multimodal attention mechanisms, have become more conspicuous. Transformer-based architecture designs with multimodal self-attention explicitly assign weights to the importance of data from diverse imaging techniques, providing clear attribution of the contributions from each method [84]. To name a few, transformer-based multimodal models that utilize MRI and PET imaging for Alzheimer's diagnosis highlight the features derived from structural MRI (atrophy regions) and functional PET data changes in metabolic activity [85].

XAI-enhanced multimodal methods ensure the accurate fusion of features and provide transparent interpretation, enabling clinicians to confidently utilize multimodal diagnostic knowledge, thereby significantly transforming patient care and outcomes [86]. Table 10 presents studies that have applied XAI techniques in clinical imaging contexts, highlighting key methodologies and outcomes.

This table summarizes evidence-based, high-profile clinical studies that employ XAI methods across various

imaging modalities, yielding more interpretable, diagnostic, and clinically translatable outcomes.

Table 11 summarizes the distribution of the 133 included studies across medical imaging modalities, highlighting the primary clinical applications of XAI in radiology and pathology.

In the 133 studies that were included in this systematic review, the distribution of imaging modalities is a clear representation of current research priorities regarding explainable AI for medical imaging. Chest X-ray was the most studied modality in 41 studies and mostly focused on pneumonia, Covid-19, tuberculosis, and cardiomegaly. CT of different body parts were close behind with 33 studies, most in the field of lung nodules, brain tumors, liver lesions, and findings related to the Covid-19. MRI studies 26 (major use in Brain Tumor Segmentation, Alzheimer's Disease, Multiple Sclerosis, Breast & Prostate Cancer Evaluation). Digital pathology and whole slide imaging was a fast-growing area and accounted for 19 studies focused on histological analysis for breast, prostate, colon, and lung cancer. Mammography-specific research was conducted in 6 studies, all of which were associated with breast cancer screening and classification. Multimodal strategies that combined PET-CT, MRI-PET, or CT-MRI were investigated by 5 studies, focusing primarily on Alzheimer's and staging of oncology. Finally, ultrasound and fundus/retinal photography were less common with only 2 and 1 study respectively, covering thyroid nodules, breast ultrasound, and diabetic retinopathy. This distribution reflects the preeminence of thoracic radiology while demonstrating the growing importance of pathology and multimodal imaging in XAI literature.

Table 11 Summary of study counts per modality under applications of XAI

Imaging Modality	Number of Studies	Representative Examples from your paper
Chest X-ray	41	Pneumonia, COVID-19, TB, cardiomegaly
Computed Tomography (CT) – all types	33	Lung nodules, brain tumor, liver, COVID-CT
Magnetic Resonance Imaging (MRI)	26	Brain tumor, Alzheimer's, MS, breast, prostate
Digital Pathology/ Whole-Slide Images	19	Breast, prostate, colon, lung cancer histology
Mammography	6	Breast cancer screening & classification
Multimodal	5	Alzheimer's, oncology staging
Ultrasound	2	Thyroid nodules, breast US
Fundus/Retinal photography	1	Diabetic retinopathy

Clinical impact of XAI in healthcare

XAI significantly alters the way healthcare is delivered, as it enhances clinical diagnostics, strengthens clinicians' trust, and consequently improves patient outcomes. This section provides a detailed elaboration on how XAI impacts diagnostic accuracy, clinicians' decision-making, and overall transparency in healthcare practices.

Enhancing diagnostic accuracy and decision support

Recent advancements in AI and Computer Vision have increased the use of AI systems in medical imaging for classification and segmentation tasks. This article focuses on XAI, particularly how its methods, including case-based, textual, and auxiliary explanations, can enhance understanding in medical contexts. By addressing techniques beyond saliency and aiming at healthcare professionals, the article seeks to improve communication between DL developers and healthcare providers regarding XAI methods and their applications in medical imaging [91].

LIME, SHAP, and Grad-CAM are three commonly used post-hoc explainability methods for interpreting CNN models on image data. LIME highlights the regions that influence a prediction, SHAP explains feature contributions through game-theory principles, and Grad-CAM creates visual heatmaps that show which parts of the image guide the model's decision. Each method has its own strengths and limitations, helping practitioners choose the approach that fits their needs [92]. Research has shown that when clinicians are supported in their clinical decision-making by XAI-driven visualizations, there is increase in diagnostic accuracy, particularly in cases involving complex diseases such as early-stage cancers or subtle neurodegenerative diseases [93].

XAI categorizes methods into four main approaches: Functioning-based, Result-based, Conceptual, and Mixed. It highlights key aspects such as the principles and properties of explainability, the need for XAI, and challenges in its implementation. The review underscores the significance of XAI in critical areas like healthcare and finance, providing a foundation for developing new techniques and advancing the field [94]. Consequently, Decision support through transparent AI reasoning significantly reduces diagnostic diversity and improves clinical accuracy, favoring patient outcomes [95].

Increasing clinicians' trust and understanding

Trust and comprehension are fundamental prerequisites for applying AI technologies in clinical practice. Conventionally, the 'black-box' AI models have been unpopular, even though they are generally reliable, due to a lack of interpretability; hence, clinicians are reluctant to trust those AI systems fully [96]. XAI addresses this problem

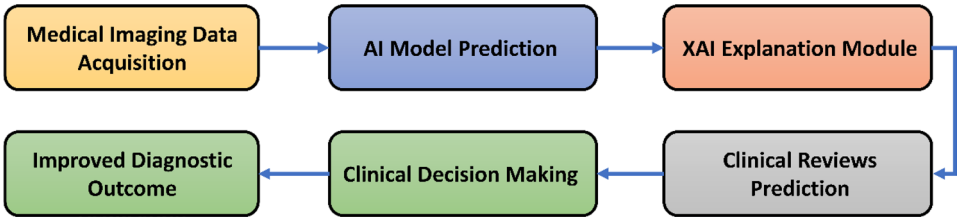

Fig. 7 A diagram illustrating clinician interaction with XAI models in diagnostic workflows

directly by simplifying model decision visualization and associated explanatory frameworks [97].

AI prediction has been shown to increase clinicians' trust significantly when accompanied by explanations in clinical validation studies. In particular, attention-based heatmaps and LIME explanation-based models significantly increased acceptance among clinicians, with trust rising compared to non-explainable models [98]. Furthermore, the level of transparency that AI techniques provide enables clinicians to follow AI recommendations confidently [99].

XAI-enabled learning and training environments facilitate the development of clinicians' perception of XAI abilities and boundaries, leading to improved communication with AI systems. This enhanced knowledge significantly accelerates the clinical integration of sophisticated AI tools, making them vital to contemporary clinical decision-making workflows [100].

Improving patient outcomes through transparent decision-making

Direct, measurable impacts on patient outcomes are realized through transparent decision-making with XAI. Armed with useful and comprehensible AI outputs, clinicians will be empowered to make well-informed, confident, and rapid decisions that minimize the time spent on diagnosis and improve diagnosis accuracy [101]. For example, XAI-driven transparent explanations in oncology imaging have enabled earlier and more precise tumor diagnoses, leading to improved prognoses and earlier interventions [102].

Moreover, transparency improves clinician-patient communication. Clinicians can utilize XAI outputs to present visually diagnostic reasoning to patients, enhancing patient engagement, understanding, and participation in shared decision-making processes. In cases of improved transparency and patient involvement, increased compliance with treatment schemes, patient satisfaction, and overall clinical outcomes have been reported [103].

Figure 7 provides a description of the clinical pathways of XAI models, which begin with the collection of medical image data and utilize an AI-based model to make predictions. The explanation module is then capable of giving transparent details that assist clinical review and validation. This loop reinforces the clinical decision-making process, thereby yielding positive diagnostic results, and illustrates how XAI connects AI prediction with trustworthiness and usability in clinical practice.

Table 12 Comparative summary of XAI techniques across clinical domains

Technique	Clinical domains	Reported strengths	Reported limitations
Grad-CAM (saliency)	Radiology (X-ray, MRI, CT)	Easy visualization; intuitive heatmaps	Sensitive to noise; limited quantitative rigor
LIME (perturbation)	Pathology, dermatology	Model-agnostic; local explanations	Computationally expensive; unstable results
SHAP (feature attribution)	Multimodal EHR + imaging	Theoretically solid; consistent	Hard to scale for large datasets
Attention maps	Histopathology, multimodal imaging	Integrates well with DL models; interpretable weights	Can be misleading ("attention ≠ explanation")
Rule-based methods	Clinical guidelines integration	Transparent; human-readable	Limited scalability; not adaptable to DL

Table 12 summarizes various XAI techniques used in different clinical domains, highlighting their strengths and limitations. Each technique is associated with specific medical fields, such as radiology, pathology, and histopathology. It provides an overview of how these techniques perform in terms of ease of use, scalability, interpretability, and their ability to handle complex data.

Challenges and limitations of XAI in medical imaging

Despite the latent benefits, significant challenges exist in regularly implementing XAI approaches in clinical practice. This section explains these challenges in detail by highlighting interpretability metrics, data quality issues, integration problems, and ethical concerns.

Lack of standardized interpretability metrics

The lack of commonly agreed-upon interpretability metrics to measure is one of the key problems at the time of XAI integration. Lacking standardized metrics, comparisons, and validations of various XAI methods, their clinical adoption is complex and constrained. The use of XAI in MI is increasing with focus on MRI, CT and

Table 13 Summary of critical challenges and potential solutions in adopting XAI

Challenge	Explanation & Implications	Impact Level	Potential Solutions	Example Tools/Standards
Lack of standardized interpretability metrics	Difficulty validating and comparing XAI models due to inconsistent interpretability evaluations.	**High**	Set global interpretability metrics, standards, and consensus-driven recommendations.	QUANTUS (quantitative XAI), IEEE P3301, ISO/IEC TR 29,119–11
Issues with data quality, bias, and generalizability	Poor or biased datasets lead to unreliable model predictions and explanations.	**High**	Careful data curation, diversified datasets, population-wide validation, and bias avoidance.	AIMI Datasets (Stanford), MONAI Label, IBM AI Fairness 360
Complexities of clinical workflow integration	Challenges in integrating XAI without disrupting current clinical practices.	**Medium**	User-centered design, clinician training, seamless technology integration, and multidisciplinary collaboration.	Epic AI Toolkit, Cerner Ignite APIs, XAI-PACS Plugins
Ethical and legal considerations regarding interpretability and patient safety	Interpretability and AI reliance raise liability, ethical, and patient safety issues.	**High**	Clear regulations, ethics, openness, and clinical responsibility.	FDA SaMD Guidance (updated in 2019 and 2023), EU AI Act (Annex III for high risk, transparency requirements) Article 10 of the EU Medical Device Regulation (MDR) and Article 22 of the GDPR (only automated decisions) AMA Principles for XAI in Medicine (2023) Ethics Framework for ACR-AI

radiography in the diagnosis of lung and brain pathologies. Most studies give visual and numerical explanations with local explanations being the common ones. However, a lack of standardization of terminology exists, as terms such as "explainable" and "interpretable" are typically used interchangeably. Future XAI developments should align better with user needs and views [104].

It is extremely valuable to develop standard quantitative metrics like interpretability accuracy scores or standards for saliency metrics that significantly enhance the clinical validation of XAI. To date, research is being pursued to define, with a high degree of consistency, criteria for measuring and benchmarking interpretability across model types, which is necessary to gain clinical confidence and facilitate the widespread adoption of XAI approaches [105].

Issues with data quality, bias, and generalizability

Attitudes towards AI prediction and explanation are very much dependent on the data quality and the underlying biases in the data. Medical imaging datasets typically suffer from class imbalance, inconsistency in annotation and a lack of representative diversity, which can result in biased and suboptimal AI results.

Biases that are introduced in the training data can be propagated through the XAI explanations, which will potentially lead clinicians into a false direction if the explanations do not correctly represent the pathological features of the data. The solutions to this issue are carefully managed datasets balanced across demos and clinical presentation. In addition, systematic validation in a variety of different patient populations needs to be prioritized to ensure generalizability of AI models and equitable clinical outcome.

Complexities of clinical workflow integration

Deploying XAI in established clinical workflows is challenging due to technical and organizational complexities. The clinical workflows involved are highly structured, extremely time-sensitive, and require the smooth integration of AI with minimal interference to current processes. Clinicians may be hesitant to adopt XAI tools that appear cumbersome or time-consuming.

Thus, XAI systems must be designed in close collaboration with clinicians to incorporate intuitive usability, minimal disruption to the workflow, and clinical significance. Effective integration also demands intensive clinician training, multidisciplinary collaboration, and well-defined operational guidelines to ensure a smooth transition and continued use in clinical practice.

Ethical and legal considerations regarding interpretability and patient safety

Issues of ethics and the law form a critical barrier to the implementation of XAI in clinical environments. Ethics prescribe transparent decision-making, but a problematic and unclear view of accountability and liability in AI-assisted misdiagnoses. Interpretability itself may create liability issues, especially if clinicians can't read or overrely on AI explanations.

Patient safety must remain of the utmost importance. XAI models must be properly clinically validated and more heavily regulated to ensure that their robust interpretability supports robust, ethical medical practice. To address these concerns, we need transparent regulations, transparent regulations of access, transparent standards of access, principles of transparency, liberal rules of access, and transparent regulations with careful regulations on the part of AI providers to ensure that patient safety is maintained and that clinicians are protected if

the AI gives an erroneous prediction or explanation that is misleading. Table 13 summarizes the critical challenges and potential solutions in adopting XAI, outlining key barriers and strategies for overcoming them.

Several limitations are evident in this review. To begin with, only the English-language publications have been considered, which can be considered a language bias. Second, the reviews targeted peer-reviewed journal papers and conference papers, excluding gray literature and preprints, which could have influenced the comprehensiveness. Third, the search period (2015–2025) restricts generalization to earlier works of the foundation. Lastly, even though the PRISMA approach was adhered to, there may have been a subjective bias during the screening, leading to the inclusion of certain articles. Such limitations are to be taken into account when interpreting the findings.

Role of XAI in medical imaging

AI can be used to improve medical image identification, categorization, diagnosis, segmentation, prediction, and quality improvement. ML helps computers to get better on predictions and actions based on data. Most medical image analysis algorithms of recent times are based on DL, which is a form of ML and relies on artificial neural networks. Operations of feature extraction, learning, and other operations of ML models can be performed directly on raw data. When ML is used in more complex models or pipelines such as an 'ensemble' of simpler models voting on a prediction or deep neural networks with millions of weights, data, and task insights being hidden [106]. AI systems are being implemented into applications of increasing complexity; hence accuracy and metrics are no longer enough. Safety, nondiscrimination, uncertainty quantification, explanation of the model findings and features are now prioritized. The EU considers that algorithms that take decisions based on user-level predictions and have a significant impact on individuals must explain the operation of the algorithm [107].

AI has progressed through 50 years, and ML and DL play a major role in sectors such as recommendations, computer vision and diagnoses. Despite good prediction performance, AI models such as ensembles and neural networks are not transparent and pose a problem in the process of decision-making. XAI and ML interpretability are important, but the systematic methodologies of XAI are underdeveloped [108]. Some argue that DL methods are explainable but not interpretable; therefore, humans require additional tools to comprehend complex activities. Oversimplified, explainability and interpretability should be considered as a spectrum. Easy-to-understand predictions make a model more interpretable [109]. XAI may explain decisions locally or the model's logic and shapes across inputs globally. This difference aids model decision-making and result analysis.

The post-hoc or an ante-hoc local XAI methods explain the performance before or after training. Methods such as Bayesian DL use the uncertainty of the prediction and provide an explanation of performance with regard to particular inputs. The most typical XAI technique is post-hoc and gradient-based, and occlusion-based methods are especially popular. Gradient-based XAI techniques yield the derivatives of model outputs with respect to each of the input dimensions, like a picture pixel. A sharp slope indicates the model's sensitivity to even slight dimensional changes, highlighting its significance. Heatmaps (or attention maps) can be generated for a model and an input image [110]. Gradient-based methods, such as Grad-CAM or Integrated Gradients, are capable of producing such maps by selecting parts of a model or areas of an input image that are most influential in making predictions [110]. Furthermore, gradient-based approaches have also been coupled with numerical attribution techniques, such as Shapley values, to provide more scientifically grounded, rigorous feature significance scores. Skin diseases are a global health concern and require thorough examinations, as well as reviews of medical history and laboratory tests. The study [111] suggests a DL model that can rapidly identify and forecast skin diseases, such as chickenpox, measles, and monkeypox. Pre-trained VGG16 determines patterns of symptoms and can diagnose and predict an illness with 93.29% accuracy. Nevertheless, its opacity does not render the model applicable in the healthcare sector as a whole. Layer-wise relevance pooling is embedded in the system, which finds input relevance scores, suggesting useful knowledge for timely diagnosis and treatment decisions [111].

The performance of XAI techniques, specifically Grad-CAM and SHAP, is evaluated in the context of national demand forecasting. Both methods enhance forecasting accuracy and model efficiency by addressing irrelevant features. Grad-CAM is noted for its computational efficiency, while SHAP excels in identifying features that diminish forecast accuracy. Nevertheless, Grad-CAM's characteristics can compromise model stability, and SHAP may misrank significant features. Future research should focus on refining these techniques and exploring additional XAI methods for time series forecasting. XAI holds considerable promise for improving load forecasting and thereby enhancing model accuracy and efficiency [112].

Transfer Learning (TL) model using a pre-trained VGG16 model with XAI to accurately classify Kidney-Ureter-Bladder (KUB) X-ray images. The model achieves 97.41% accuracy in distinguishing between kidney stones and normal cases. However, the model lacks fairness and

explainability. The study employs the LRP technique to enhance transparency and effectiveness, enabling doctors to accurately identify kidney stones and implement effective treatment strategies [113].

AI should be compared to drug approval processes in other healthcare sectors. Many common medications are safe even without a biological reason. The key is trustworthy, verifiable efficacy data from randomized controlled trials. Should we expect more from AI if clinical trials are sufficient for drug approval? Do safety and technical concerns drive explainability, or do societal considerations promote doctor and patient acceptance of AI?

Discussion

The systematic literature review highlighted medical imaging XAI developments, clinical importance, and challenges. We thoroughly analyze XAI trends and technology, suggest standardized reporting systems, and clearly define future research to solve the foregoing constraints in this part. AI and healthcare professionals will explore this topic in a webinar, sharing insights and advances. Some research suggests XAI can improve model understanding but reduce accuracy. Explainability-based decision help may overuse incomprehensible or incorrect models, reducing decision performance [114]. XAI debugging tools might potentially cause data scientists and other professionals to overestimate AI tool deployment readiness [115]. Other AI concerns like biases and domain changes affect XAI-based models. Thus, strict verification and robust design are needed to assure therapeutic relevance.

Emerging trends and technologies in XAI

XAI is growing quickly because to new computational methods, algorithmic constructs, and advanced technology. The trend toward transformer-based designs and multimodal self-attention mechanisms, which outperform CNN-based models, is promising. Transformers' attention mechanisms reveal the model's decision-making process directly and intuitively, generating huge advances in the field. Federated Learning (FL) and XAI offer safe analysis of dispersed medical datasets, another promising trend. Despite data sparsity and bias in AI models, Federated XAI allows hospitals to collaborate securely without compromising patient privacy. Blockchain technology may improve AI model transparency, traceability, and auditability, improving clinical decision-making data provenance and accountability. In addition, causal inference and symbolic thinking help create trustworthy and interpretable AI systems. Causal XAI frameworks accurately explain the relationship between imaging features and diagnostic outcomes, improving clinical insights and AI-assisted diagnostics trust. Traditional symbolic AI combined with DL is producing highly interpretable models that can reason for themselves directly using human-speakable rules and logic, which is crucial for regulatory approval and clinical application.

Recommendations for standardized reporting frameworks for XAI studies

Comparability, repeatability, and clinical validation have suffered in the XAI professional arena, particularly in medical imaging research, due to the lack of established reporting standards. XAI is becoming more common in healthcare, therefore universal reporting standards like those for clinical trials (CONSORT), diagnostic studies (STARD), and predictive modeling are needed. It improves XAI research by making results comparable, verifiable, and clinically useful. Comprehensive and standardized XAI research in medical imaging is needed. It should reveal various crucial elements that ensure reportage clarity and transparency, such as:

Details of Used Datasets: The reporting framework should describe the datasets, patient demographics, and image acquisition techniques. It is significant since the dataset's patient age, gender, medical history, and imaging techniques affect XAI model performance and generalization. Diverse datasets are essential for constructing strong and relevant models for different patient populations. Transparency regarding the dataset helps future studies replicate and compare findings.

Unambiguous AI Model Architecture, Training Method, and Optimization Strategies: Whether the AI model is CNN, Transformer, or other DL, the framework should describe its design. It should also cover training, data pre-processing, augmentation, loss functions, and hyperparameter optimization. Replicating this and understanding why architecture was chosen might help others comprehend its performance. Details about XAI Methods: The standardized reporting framework requires a clear description of study XAI procedures. Grad-CAM, SHAP, LIME, and attention mechanisms have different use cases depending on the AI prediction explanation approach. Researchers must record their methods, crucial factors, and changes. This transparency will affect interpretability strategies and how they can be examined to guarantee they solve real-world problems.

Measures of Quantitative Interpretability The absence of quantitative measurements for model explanation quality and dependability is a major difficulty in XAI. Standardized reporting should include widely accepted interpretability criteria to help researchers and physicians move beyond XAI model effectiveness discussions. Metrics should quantify explanation fidelity, human understandability, and cross-dataset consistency across clinical situations. Quantitative measurements will simplify XAI technique evaluation and clinical validation of AI-based diagnostic tools. Clinical Validation findings:

Publishing clinical validation findings show XAI tools' practicality. The suggested framework shall report XAI approaches' effects on physician trust, diagnostic accuracy, and patient outcomes. Comparing normative diagnostic methods should show how XAI has improved clinical workflows. Researchers can justify XAI tool use in clinical settings by submitting validation results.

An ethical perspective: Ethics should be part of the standardized framework, especially in healthcare. The framework should explain how the study managed data privacy, patient-informed consent, and confidentiality. A bias evaluation should also be done to see if certain patient groups are disproportionately affected. Report ethical issues such model accountability, decision-making openness, and AI overuse. This will ensure ethical and safe healthcare XAI applications. Technology-neutral regulations can control AI. EU GDPR and US HIPAA govern patient data consent, protection, and access. As data-intensive technologies like AI grow more common, healthcare 'model-as-a-service' companies increase the potential of privacy violations and patient record misuse. State and federal supervision by the Federal Trade Commission promotes compliance and demands the removal of illegally acquired goods. Universal Declaration of Human Rights prioritizes human rights and data protection. Equality, non-discrimination, and wellness can remove injustice. Governance frameworks limit AI transparency, scalability, and autonomy, necessitating adaptive modifications for end-to-end governance. AI in healthcare may be prejudiced, compromising fairness and accuracy. Historical healthcare data may bias models to unequal outcomes for ethnic minorities, women, and low-income populations [116]. Data bias worsens inequality because diagnostic algorithms trained on underrepresented groups may fail other groups [117]. Data-collection attitudes such coding and recording techniques that favor one category, like scoring whites vs. blacks in healthcare distribution constitute measurement bias [118, 119]. Finally, labelling bias causes supervised learning to overlay diagnoses or outcomes depending on human decision-making or structural prejudice, creating models with biases [120, 121].

For clinical use and marketing, they must meet with local regulations: EU Conformity Europeans marking, US FDA approval, and UK Conformity Assessment mark after Brexit, with a transition period until mid-2023. However, working across such diverse systems is difficult, hence several countries have eased acceptance by adopting FDA-approved devices. AI-SaMDs use black-box models that don't explain how they generate outputs, making validation harder and malpractice liability issues worse. Risk mitigation relies on interpretability and explainability: White-box models like linear and DTs explain. XAI then explains black-box output with secondary models. Present methods struggle to describe complex, patient-specific healthcare decisions [122, 124].

Access to large datasets like EHRs is difficult due to data protection laws like the GDPR in Europe, which require patient consent [124], and the lack of public validation/testing data makes it difficult to assess generalizability and identify biases [125]. Data sharing on open platforms like OpenNeuro (https://openneuro.org/) raises ownership and privacy concerns. Healthcare facilities own imaging data in the US, although patients have HIPAA privacy protections [126]. Most medical photographs are CC-0, therefore patient privacy is the key problem in medical imaging sharing [127, 128]. Future XAI medical research will be more transparent, reproducible, and comparable with this standardization. This will clinically validate XAI tools and speed their inclusion into healthcare systems, possibly leading in regulatory approval.

Prospective research directions for overcoming current limitations

Implementing XAI approaches in medical imaging is difficult due to the complexity of AI models, data quality, and clinical procedures. To overcome these challenges and integrate XAI into clinical practice, future research should aim to quantify interpretability, mitigate bias, integrate clinical workflow, develop causal and symbolic reasoning, and establish ethical and regulatory frameworks.

Formulating broadly agreed quantitative criteria for XAI model interpretability is a research priority. Like doctor satisfaction or expert review, qualitative explanation ratings are subjective and might vary greatly between physicians. Future work should build defined standards to make AI descriptions objective regarding fidelity and transparency. These metrics could assess consistency, reliability, and reasoning process compliance. These criteria will help researchers subjectively compare XAI approaches and give doctors practical insights.

Training dataset biases cause AI model predictions to diverge across patient categories, making forecasts unreliable. This is a major issue in medical imaging, as ethnicity, age, and comorbidities can affect model performance. Federated learning, data augmentation, and TL studies reduce biases and improve XAI model generalizability. FL can be used to construct an AI model without exchanging sensitive data between organizations while protecting data privacy and exploiting varied datasets. Adding more clinical cases to training data will also increase model generalization to real-world contexts.

Future research should focus on seamlessly integrating XAI models into healthcare workflows. Even if uninterpretable, AI models must work well in clinical practice's fast-paced environment. Research should focus on

developing XAI systems that provide simple model outputs in real time without interfering with clinician workflow. Clinicians, AI researchers, and user experience designers must collaborate to create systems with user-friendly interfaces, fast decision assistance, and intuitive explanations. Understanding how clinicians trust AI systems will improve their usability.

Future research should examine XAI causality and symbolic reasoning. Grad-CAM and SHAP, classic XAI methods, investigate image attributes and predictions but sometimes miss causality. Causal XAI models can explain the cause-and-effect links contributing to a diagnosis, revealing why it was chosen. But symbolic reasoning approaches try to encapsulate AI reasoning in a human-readable, rule-based manner so physicians may better understand and assess it. Causal reasoning and symbolic AI may help create more understandable, robust, and transparent AI systems that explain connections and causal mechanisms.

Development of ELR XAI is increasingly used in medical imaging and clinical decision-making, requiring simple ethical, legal, and regulatory frameworks. Data privacy, model accountability, and clinical duty for AI tools should be researched. This means AI-generated explanations must be transparent, verifiable, and clinically consistent. Researchers must also examine the ethical consequences of overusing AI, where physicians may overemphasize AI models and fail to critically evaluate their results. Research on XAI regulatory requirements and ethical application in healthcare should improve patient safety and physician confidence in these technologies.

Regulatory Implications of XAI in medical imaging

Implementing XAI in medical imaging technology is facing growing regulatory scrutiny to guarantee clinical safety, transparency, and accountability. Although XAI improves interpretability, e.g. using Grad-CAM heatmaps to show tumor locations in CT scans, it is one of the main considerations in approvals, which balances the black-box and regulatory views of human interaction.

As of October 2023, the FDA has approved many AI/ML-enabled medical devices. These machines are primarily used in radiology and are designed to scan lesions and reconstruct images. GE Healthcare is a market leader in the field and holds various certifications. This contains algorithms for lung detection and MRI enhancement, used by clinicians to assess XAI outputs, such as saliency maps. A Class II-level XAI-based whole-slide imaging system, the Philips IntelliSite Pathology Solution, was cleared by the FDA a few years ago. The GMLP of the FDA indicates the importance of explainability in premarket validation. Nevertheless, the examination of different devices showed that very little explicit XAI was used, suggesting that human-centered design is required to meet the so-called right to explanation [127–129].

The international AI medical devices regulator is the European Medicines Agency (EMA) which regulates medical devices subject to the regulation of medical devices i.e. the Medical Device Regulation (MDR) and the EU AI Act (effective 2024) and which treats high-risk imaging AI, including automated diagnostics, as requiring transparency and risk management. Contrary to the FDA device-specific list, the EMA approvals are less centralized, however, with XAI as an accountability measure, which is in line with the GDPR safeguard of Article 22 of the partly automated decisions. Oncology tools that are approved by EMA should be explainable, like with the help of attention mechanisms that encourage clinician intervention, which in meta-analyses a small share of XAI studies included regulatory-compliant assessments. New directions put more emphasis on multimodal XAI of CE marking, so the results such as a textual rationale decrease bias in different patient data [130–132]. Medical image analysis A novel XAI approach involves the combination of statistical, visual, and rule-based explanations to increase medical image analysis model transparency. It combines statistical characteristics, decision trees and rule-based models to produce visualizations that look like heatmaps and human-readable rules, which enhance the interpretability of different medical imaging datasets [133].

Conclusion and future directions

Conclusion

This systematic review of the literature on XAI for medical imaging provides essential, up-to-date information on methodologies, clinical consequences, limitations, and future research. The final section summarizes major discoveries, emphasizes current research, and emphasizes XAI's healthcare system potential. The review found numerous important medical imaging XAI discoveries. Grad-CAM, heatmap-based approaches, and saliency maps helped clinicians understand diagnostic details graphically. Attention-based models, including transformer topologies, are better for complicated imaging instances in interpretability and clinical efficacy. Model-agnostic techniques like SHAP and LIME gave flexible, theory-based explanations for AI model applicability. ET complained that XAI boosted clinician accuracy and diagnostic trust. The transparent model's predictions helped clinicians and AI collaborate, improving clinical decision-making. The lack of established interpretability criteria, data biases, clinical integration challenges, and legal and ethical concerns remain. Addressing present limits, improving interpretability, and increasing clinical trust in AI technology require ongoing XAI research. The rising use of AI models in diagnostics requires

transparency and interpretability to improve patient safety, ethics, and regulatory compliance. Research will also build strong and uniform interpretability criteria to eliminate biases and improve generalizability across varied populations. XAI adoption requires clinical workflow integration research to help healthcare workers use these technologies smoothly in clinical practice. The systematic use of XAI methods in medical imaging could alter clinical practice and healthcare systems. Clinically, XAI techniques may increase diagnostic accuracy, efficacy, and reliability. Improved interpretability improves clinical decision-making, which improves patient outcomes, satisfaction, and healthcare effectiveness. Better resource allocation, diagnostic variance reduction, and clinical standardization are possible using XAI in healthcare. High-quality healthcare delivery requires meaningful clinical training and persistent continuing medical education, which transparent AI technologies enable. To conclude, XAI technologies offer innovative clinical, functional, and educational benefits that strengthen and transparent patient-centered treatment worldwide.

Future directions

The next step in the work should be to change the current research-based XAI environment into solutions that are clinically viable, acceptable to regulators, and widely used. First, they should work on making a set of internationally recognized quantitative indicators of interpretability. Then, they should do large prospective multicenter clinical studies that quantitatively measure the actual effect on diagnostic accuracy and clinician behavior. Finally, they should work with radiologists and pathologists to design human-interactive explanation interfaces. Future research endeavors should transcend correlation-based heatmaps to embrace causal, counterfactual XAI, integrate bias detection and mitigation within distributed pipelines, and cultivate federated, privacy-preserving XAI models that facilitate secure collaboration among institutions. It is equally important to extend explanations to genuinely multimodal (imaging + EHR + genomics) and longitudinal contexts, and to establish official regulatory-grade standards for XAI validation in collaboration with the FDA and EMA. Additionally, it is essential to publish open benchmark datasets containing ground-truth clinical rationales and to incorporate XAI literacy into medical education programs. It is through these interdisciplinary collaborations over the next 3–5 years that XAI in medical imaging will evolve into an academic imperative, a clinical essential capable of demonstrably enhancing patient outcomes, and will gradually earn the trust of the medical community.

Appendix A

$$\alpha_k^c = \frac{1}{Z}\sum_i\sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{A1}$$

$$L_{Grad-CAM}^c = ReLU\left(\sum_k \alpha_k^c A^k\right) \tag{A2}$$

$$\alpha_k = \sum_{i=1}^{Z}\sum_j \frac{\partial A_{kij}}{\partial y_c} \tag{A3}$$

$$X_{PCA} = XW \tag{A4}$$

$$f(S) = \sum_{i=1}^{N} accuracy(S_i) \tag{A5}$$

$$R^{(l)} = \left(f^{(l)} > 0\right) \odot \left(\frac{\partial L}{\partial f^{(l)}} > 0\right) \odot \frac{\partial L}{\partial f^{(l)}} \tag{A6}$$

$$R_i^{(l)} = \sum_j \left(\frac{z_i j}{\sum(i') z_{(i'j)} + \varepsilon}\right) R_j^{(l+1)} \tag{A7}$$

$$\Delta y_i = \sum_{j=1}^{n} (f_i(x_j) - f_i(x_0)) \times \Delta f_j \tag{A8}$$

$$G_i(x) = (x_i - x_i') \times \int_0^1 \frac{\partial f(x' + \alpha(x - x'))}{\partial x_i}\, d \tag{A9}$$

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) V \tag{A10}$$

$$MultiHead(Q, K, V) = Concat(head_1, \ldots, head_h) W^O \tag{A11}$$

$$head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right) \tag{A12}$$

$$\xi(x) = argargL(f, g, \pi_x) + \Omega(g) \tag{A13}$$

$$\phi = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(M - |S| - 1)!}{M!} \left[f_x(S \cup \{i\}) - f_x(S)\right] \tag{A14}$$

$$P(A) \approx P\left(\hat{A}\right) \tag{A15}$$

$$If X_1 \langle t_1 \wedge X_2 \rangle t_2, then predict class C \tag{A16}$$

$$(P_1 \wedge P_2 \wedge \ldots \wedge P_n) \rightarrow Class \tag{A17}$$

$$h_v^{(k+1)} = \sigma\left(\sum_{u \in N(v)} W^{(k)} h_u^{(k)} + b^{(k)}\right) \tag{A18}$$

$$p\left(M\right) = \frac{p\left(M\right) p\left(F\right)}{p\left(F\right)} \tag{A19}$$

$$L_{SSL} = \sum_{j=1}^{d} \left(x_{ij} - \widehat{x_{ij}}\right)^2 \tag{A20}$$

$$InformationGain = Entropy\left(S\right) - \sum_{v \in V} \frac{|S_v|}{|S|} \cdot Entropy\left(S_v\right) \tag{A21}$$

Acknowledgements
Not applicable.

Author contributions
F.A., N.S.N., and S.K. collected data from different sources. F.A., N.S.N., S.K., and M.A.K contributed to writing—original draft preparation; F.A., S.K., and A.U.R. contributed to writing—review and editing. N.S.N., A.U.R., and M.A.K. supervised the paper. F.A., N.S.N., W.M.I., and A.U.R. drafted pictures and tables. W.M.I., M.A.K., S.K., A.U.R., and N.S.N. performed revisions and improved the quality of the draft. All authors have read and agreed to the published version of the manuscript.

Funding
Funding is not applicable.

Data availability
This paper contains all the data.

Declarations
The authors declare that they have no conflicts of interest regarding the publication of this work.

Ethics approval
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 22 June 2025 / Accepted: 9 December 2025
Published online: 05 January 2026

References

1. Ranjbarzadeh R, Caputo A, Tirkolaee EB, Ghoushchi SJ, Bendechache M. Brain tumor segmentation of MRI images: a comprehensive review on the application of artificial intelligence tools. Comput In Biol Med. 2023;152:106405. https://doi.org/10.1016/j.compbiomed.2022.106405.
2. Bai J, Posner R, Wang T, Yang C, Nabavi S. Applying deep learning in digital breast tomosynthesis for automatic breast cancer detection: a review. Med Image Anal. 2021;71:102049. https://doi.org/10.1016/j.media.2021.102049.
3. Sengupta S, Singh A, Leopold HA, Gulati T, Lakshminarayanan V. Ophthalmic diagnosis using deep learning with fundus images-A critical review. Artif Intel in Med. 2020;102:101758. https://doi.org/10.1016/j.artmed.2019.101758.
4. Janik A, Dodd J, Ifrim G, Sankaran K, Curran K. Interpretability of a deep learning model in the application of cardiac MRI segmentation with an ACDC challenge dataset. In: Medical imaging 2021: image processing. Vol. 11596. SPIE; 2021. p. 861–72. https://doi.org/10.1117/12.2582227.
5. Mahner FP, Muttenthaler L, Güçlü U, et al. Dimensions underlying the representational alignment of deep neural networks with humans. Nat Mach Intell. 2025;7:848–59. https://doi.org/10.1038/s42256-025-01041-7.
6. Minh D, Wang HX, Li YF, et al. Explainable artificial intelligence: a comprehensive review. Artif Intell Rev. 2022;55:3503–68. https://doi.org/10.1007/s10462-021-10088-y.
7. Goodman B, Flaxman S. European Union regulations on algorithmic decision-making and a "right to explanation. AI Mag. 2017;38(3):50–57. https://doi.org/10.1609/aimag.v38i3.2741.
8. Chakraborty C, Bhattacharya M, Pal S, Lee S-S. From machine learning to deep learning: advances of the recent data-driven paradigm shift in medicine and healthcare. Curr Res In Biotechnol. 2024;7:100164. https://doi.org/10.1016/j.crbiot.2023.100164.
9. Kaul V, Enslin S, Gross SA. History of artificial intelligence in medicine. Gastrointestinal Endoscopy. 2020;92(4):807–12. https://doi.org/10.1016/j.gie.2020.06.040.
10. Pinto-Coelho L. How artificial intelligence is shaping medical imaging technology: a survey of innovations and applications. Bioengineering. 2023;10(12):1435. https://doi.org/10.3390/bioengineering10121435.
11. Ullah N, Khan JA, Falco ID, Sannino G. Explainable artificial intelligence: importance, use domains, stages, output shapes, and challenges. ACM Comput Surv. 2024;57(4):1–36. https://doi.org/10.1145/3705724.
12. Tjoa E, Guan C. A survey on explainable artificial intelligence (xai): toward medical xai. IEEE Trans On Neural Networks And Learn Syst. 2020;32(11):4793–813. https://doi.org/10.1109/TNNLS.2020.3027314.
13. Linardatos P, Papastefanopoulos V, Kotsiantis S. Explainable ai: a review of machine learning interpretability methods. Entropy. 2020;23(1):18. https://doi.org/10.3390/e23010018.
14. Varnosfaderani SM, Forouzanfar M. The role of AI in hospitals and clinics: transforming healthcare in the 21st century. Bioengineering. 2024, doi: 11(4):337. https://doi.org/10.3390/bioengineering11040337.
15. Bhati D, Neha F, Amiruzzaman M. A survey on explainable artificial intelligence (xai) techniques for visualizing deep learning models in medical imaging. J Imag. 2024;10(10):239. https://doi.org/10.3390/jimaging10100239.
16. Shobayo O, Saatchi R. Developments in deep learning artificial neural Network techniques for medical image analysis and interpretation. Diagnostics. 2025;15(9):1072. https://doi.org/10.3390/diagnostics15091072.
17. Hassija V, Chamola V, Mahapatra A, et al. Interpreting black-box models: a review on Explainable artificial intelligence. Cogn Comput. 2024;16:45–74. https://doi.org/10.1007/s12559-023-10179-8.
18. Mienye ID, et al. A survey of explainable artificial intelligence in healthcare: concepts, applications, and challenges. Inf In Med Unlocked. 2024;101587. https://doi.org/10.1016/j.imu.2024.101587.
19. Sheu RK, Pardeshi MS. A survey on medical Explainable AI (XAI): recent progress, explainability approach, human interaction and scoring System. Sensors (basel, Switz). 2022;22(20):8068. https://doi.org/10.3390/s22208068.
20. [Chaddad A, Peng J, Xu J, & Bouridane A. Survey of Explainable AI techniques in healthcare. Sensors. 2023;23(2):634. https://doi.org/10.3390/s23020634.
21. Turner C, Okorie O, Oyekan J. XAI sustainable human in the loop maintenance. IFAC-Papersonline. 2022;55(19):67–72. https://doi.org/10.1016/j.ifacol.2022.09.185.
22. Van der Velden BH, Kuijf HJ, Gilhuijs KG, Viergever MA. Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. Med Image Anal. 2022;79:102470. https://doi.org/10.1016/j.media.2022.102470.
23. de Vries BM, Zwezerijnen GJC, Burchell GL, van Velden FHP, Menke-van der Houven van Oordt CW, Boellaard R. Explainable artificial intelligence (XAI) in radiology and nuclear medicine: a literature review. Front Med. 2023;10:1180773. https://doi.org/10.3389/fmed.2023.1180773.
24. Chen H, Gomez C, Huang CM, et al. Explainable medical imaging AI needs human-centered design: guidelines and evidence from a systematic review. npj Digit. Med. 2022;5:156. https://doi.org/10.1038/s41746-022-00699-2.
25. Borys K, Schmitt YA, Nauta M, Seifert C, Krämer N, Friedrich CM, Nensa F. Explainable AI in medical imaging: an overview for clinical practitioners-saliency-based XAI approaches. Eur J Radiol. 2023;162:110787. https://doi.org/10.1016/j.ejrad.2023.110786.
26. Li X, Zhang L, Yang J, et al. Role of artificial intelligence in medical image analysis: a review of Current trends and future directions. J. Med. Biol. Eng. 2024;44:231–43. https://doi.org/10.1007/s40846-024-00863-x.

27. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA, 2020 statement: an updated guideline for reporting systematic reviews. Syst Rev. 2021;10(89). https://doi.org/10.1186/s13643-021-01626-4.
28. Garg P, Sharma MK, Kumar P. Transparency in diagnosis: unveiling the Power of deep learning and Explainable AI for medical image interpretation. Arab J Sci Eng. 2025. https://doi.org/10.1007/s13369-024-09896-5.
29. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016, pp. 2921–29. https://doi.org/10.1109/CVPR.2016.319.
30. Ali S, et al. Explainable artificial intelligence (XAI): what we know and what is left to attain trustworthy artificial intelligence. Inf Fusion. 2023;99:101805. https://doi.org/10.1016/j.inffus.2023.101805.
31. Mersha M, Lam K, Wood J, Alshami AK, Kalita J. Explainable artificial intelligence: a survey of needs, techniques, applications, and future direction. Neurocomputing. 2024;599:128111. https://doi.org/10.1016/j.neucom.2024.128111.
32. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE international conference on computer vision. 2017, pp. 618–26. https://doi.org/10.1109/ICCV.2017.74.
33. Nagarajan G, Minu RI, Muthukumar B, Vedanarayanan V, Sundarsingh SD. Hybrid genetic algorithm for medical image feature extraction and selection. Procedia Comput Sci. 2016;85:455–62. https://doi.org/10.1016/j.procs.2016.05.192.
34. Sindiramutty SR, Tee WJ, Balakrishnan S, Kaur S, Thangaveloo R, Jazri H, … & Manchuri AR. Explainable AI in healthcare application. In: Advances in explainable AI applications for smart cities. IGI Global Scientific Publishing; 2024. p. 123–76. https://doi.org/10.4018/978-1-6684-6361-1.ch005.
35. Eitel F, Ritter K, for the Alzheimer's Disease Neuroimaging Initiative (ADNI). Testing the robustness of attribution methods for convolutional neural networks in MRI-Based Alzheimer's disease classification. in: et al. et al. Interpretability of machine intelligence in medical image Computing and multimodal learning for clinical decision support. ML-CDS IMIMIC, 2019, 2019. In: Lecture notes in computer science(). Vol. 11797. Cham: Springer; 2019. https://doi.org/10.1007/978-3-030-33850-3_1.
36. Bach S, Binder A, Montavon G, Klauschen F, Müller K-R, Samek W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS One. 2015;10(7):e0130140. https://doi.org/10.1371/journal.pone.0130140.
37. Shrikumar A, Greenside P, Kundaje A. Learning important features through propagating activation differences. International conference on machine learning. 2017:PMlR, pp. 3145–3153.
38. Singh A, Sengupta S, Lakshminarayanan V. Explainable deep learning models in medical image analysis. J Imag. 2020;6(6):52. https://doi.org/10.3390/jimaging6060052.
39. Huff DT, Weisman AJ, Jeraj R. Interpretation and visualization techniques for deep learning models in medical imaging. Phys In Med Biol. 2021;66(4):04TR01. https://doi.org/10.1088/1361-6560/abcd17.
40. Wang Y, Wu B, Zhang N, Liu J, Ren F, Zhao L. Research progress of computer aided diagnosis system for pulmonary nodules in ct images. J X-Ray Sci Technol. 2020;28(1):1–16. https://doi.org/10.3233/XST-190581.
41. Hamamci A, Kucuk N, Karaman K, Engin K, Unal G. Tumor-cut: segmentation of brain tumors on contrast enhanced MR images for radiosurgery applications. IEEE Trans On Med Imag. 2011;31(3):790–804. https://doi.org/10.1109/TMI.2011.2181857.
42. Khojaste-Sarakhsi M, Haghighi SS, Ghomi SF, Marchiori E. Deep learning for Alzheimer's disease diagnosis: a survey. Artif Intel in Med. 2022;130:102332. https://doi.org/10.1016/j.artmed.2022.102332.
43. Zhang J, Chen X, Yang B, Guan Q, Chen J, Wu Q, Xie Y, Xia Y. Advances in attention mechanisms for medical image segmentation. Comput Sci Rev. 2025;56(100721). https://doi.org/10.1016/j.cosrev.2024.100721.
44. Ribeiro MT, Singh S, Guestrin C.""Why should i trust you?" Explaining the predictions of any classifier," in Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016, pp. 1135–44. https://doi.org/10.1145/2939672.2939778.
45. Haq I, Gong Z, Liang H, Zhang W, Khan R, Gu L, … & Huang B. A review of breast cancer histopathology image analysis with deep learning: challenges, innovations, and clinical integration. Image Vision Comput. 2025;105708. https://doi.org/10.1016/j.imavis.2025.105708.
46. Remeseiro B, Bolon-Canedo V. A review of feature selection methods in medical applications. Comput In Biol Med. 2019;112:103375. https://doi.org/10.1016/j.compbiomed.2019.103375.
47. Arora R, Lee J. Understanding impact of textural changes for mammogram analysis using counterfactuals. In: Medical imaging, 2024: computer-aided diagnosis. Vol. 12927. SPIE; 2024. p. 487–93. https://doi.org/10.1117/12.3008526.
48. Adeniran OT, Ojeme B, Ajibola TE, Peter OOE, Ajala AO, Rahman MM, Khalifa F. Explainable MRI-Based ensemble learnable architecture for Alzheimer's disease detection. Algorithms. 2025;18(3):163. https://doi.org/10.3390/a18030163.
49. Natekar P, Kori A, Krishnamurthi G. Demystifying brain tumor segmentation networks: interpretability and uncertainty analysis. Front Comput Neurosci. 2020;14:6. https://doi.org/10.3389/fncom.2020.00006.
50. Singh SK, Virdee BS, Aggarwal S, Maroju A. Incorporation of XAI and deep learning in biomedical imaging: a review. Polytechnic J. 2025;15(1):1–15. https://doi.org/10.59341/2707-7799.1845.
51. Sattarzadeh S, et al. Explaining convolutional neural networks through attribution-based input sampling and block-wise feature aggregation. Proc AAAI Conf On Artif Intel. 2021;35(13):11639–47. https://doi.org/10.1609/aaai.v35i13.17384.
52. Fong R, Vedaldi A. Explanations for attributing deep neural Network predictions. In: Samek W, Montavon G, Vedaldi A, Hansen L, Müller K, editors. Explainable AI: interpreting, explaining and visualizing deep learning. Lecture notes in computer science(). Vol. 11700. Cham: Springer; 2019. https://doi.org/10.1007/978-3-030-28954-6_8.
53. Montavon G, Binder A, Lapuschkin S, Samek W, Müller KR. Layer-wise relevance propagation: an overview. In: Samek W, Montavon G, Vedaldi A, Hansen L, Müller K, editors. Explainable AI: interpreting, explaining and visualizing deep learning. Lecture notes in computer science(). Vol. 11700. Cham: Springer; 2019. https://doi.org/10.1007/978-3-030-28954-6_10.
54. Mane D, Magar A, Khode O, Koli S, Bhat K, Korade P. Unlocking machine learning model decisions: a comparative analysis of LIME and SHAP for enhanced interpretability. J. Electr. Syst. 2024;20(2):1252–67. https://doi.org/10.52783/jes.1768.
55. Gu X, Angelov PP. Highly interpretable hierarchical deep rule-based classifier. Appl Soft Comput. 2020;92:106310. https://doi.org/10.1016/j.asoc.2020.106310.
56. Wang Z, Mardziel P, Datta A, Fredrikson M. Interpreting interpretations: organizing attribution methods by criteria. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2020, pp. 10–11. https://doi.org/10.1109/CVPRW50498.2020.00013.
57. Liang Z, Zhao K, Liang G, Li S, Wu Y, Zhou Y. Maxformer: enhanced transformer for medical image segmentation with multi-attention and multi-scale features fusion. Knowl-Based Syst. 2023;280:110987. https://doi.org/10.1016/j.knosys.2023.110987.
58. Zhou J, et al. Graph neural networks: a review of methods and applications. AI Open. 2020;1:57–81. https://doi.org/10.1016/j.aiopen.2021.01.001.
59. Bhandari M, Shahi TB, Neupane A. Evaluating retinal disease diagnosis with an interpretable lightweight CNN model resistant to adversarial attacks. J Imag. 2023;9(10):219. https://doi.org/10.3390/jimaging9100219.
60. Li M, Sun H, Huang Y, Chen H. Shapley value: from cooperative game to explainable artificial intelligence. Auton Intell Syst. 2024;4(1):2. https://doi.org/10.1007/s43684-023-00060-8.
61. Abdullakutty F, Akbari Y, Al-Maadeed S, Bouridane A, Talaat IM, Hamoudi R. Histopathology in focus: a review on explainable multi-modal approaches for breast cancer diagnosis. Front Med. 2024;11:1450103. https://doi.org/10.3389/fmed.2024.1450103.
62. Moujahid H, et al. Combining CNN and Grad-cam for COVID-19 disease prediction and visual explanation. Intell Automation Soft Comput. 2022;32(2). https://doi.org/10.32604/iasc.2022.022179.
63. M MM, R T, M V, VK, et al. Enhancing brain tumor detection in MRI images through explainable AI using Grad-CAM with Resnet 50. BMC Med Imag. 2024;24:107. https://doi.org/10.1186/s12880-024-01292-7.
64. Sadeghi Z, et al. A review of Explainable artificial intelligence in healthcare. Comput Electr Eng. 2024;118:109370. https://doi.org/10.1016/j.compeleceng.2024.109370.
65. Sigut J, Fumero F, Arnay R, Estévez J, Díaz-Alemán T. Interpretable surrogate models to approximate the predictions of convolutional neural networks in glaucoma diagnosis. Mach Learn: Sci Technol. 2023;4(4):045024. https://doi.org/10.1088/2632-2153/ad0798.

66. Hosain MT, Jim JR, Mridha M, Kabir MM. Explainable AI approaches in deep learning: advancements, applications and challenges. Comput Electr Eng. 2024;117:109246. https://doi.org/10.1016/j.compeleceng.2024.109246.
67. Jitha K, Navas K, Ahmed CN, Nasrin CN. Deep learning and grad-cam for the diagnosis of pneumonia based on x-rays. 2023 International Conference on Innovations in Engineering and Technology (ICIET). IEEE; 2023, pp. 1–5. https://doi.org/10.1109/ICIET57285.2023.10220839.
68. Gandomkar Z, Khong PL, Punch A, Lewis S. Using occlusion-based saliency maps to explain an artificial intelligence tool in lung cancer screening: agreement between radiologists, labels, and visual prompts. J Digit Imag. 2022;35(5):1164–75. https://doi.org/10.1007/s10278-022-00631-w.
69. Thulasiram PP. Explainable artificial intelligence (Xai): enhancing transparency and trust in machine learning models. 2025. https://doi.org/10.2139/ssrn.5057400.
70. Raghavan M. Conventional modalities and novel, emerging imaging techniques for musculoskeletal tumors. Cancer Control. 2017;24(2):161–71. https://doi.org/10.1177/107327481702400208.
71.] Muhammad D, Bendechache M. Unveiling the black box: a systematic review of Explainable artificial intelligence in medical image analysis. Comput Struct Biotechnol J. 2024. https://doi.org/10.1016/j.csbj.2024.08.005.
72. Hekmat A, Zhang Z, Khan SUR, Shad I, Bilal O. An attention-fused architecture for brain tumor diagnosis. Biomed Signal Process Control. 2025;101:107221. https://doi.org/10.1016/j.bspc.2024.107221.
73. Eitel F, et al. Uncovering convolutional neural network decisions for diagnosing multiple sclerosis on conventional MRI using layer-wise relevance propagation. neuroimage: Clin. 2019;24:102003. https://doi.org/10.1016/j.nicl.2019.102003.
74. Yao Z, et al. Artificial intelligence-based diagnosis of Alzheimer's disease with brain MRI images. Eur J Radiol. 2023;165:110934. https://doi.org/10.1016/j.ejrad.2023.110934.
75. Singh R, Gupta S, Mohamed HG, et al. Advancing prenatal healthcare by explainable AI enhanced fetal ultrasound image segmentation using U-Net++ with attention mechanisms. Sci Rep. 2025;15:19612. https://doi.org/10.1038/s41598-025-04631-y.
76. Jahn SW, Plass M, Moinfar F. Digital pathology: advantages, limitations and emerging perspectives. J Clin Med. 2020, doi: 9(11):3697. https://doi.org/10.3390/jcm9113697.
77. Kumar Y, Koul A, Singla R, Ijaz MF. Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda. J Ambient Intel Humanized Comput. 2023;14(7):8459–86. https://doi.org/10.1007/s12652-021-03612-z.
78. Aldakhil LA, Alharbi SS, Aloraini A, Alhasson HF. Leveraging attention-based deep learning in binary classification for early-Stage breast cancer diagnosis. Diagnostics. 2025, doi: 15(6):718. https://doi.org/10.3390/diagnostics15060718.
79. Ukwuoma CC, et al. Enhancing histopathological medical image classification for early cancer diagnosis using deep learning and explainable AI-LIME & shap. Biomed Signal Process Control. 2025;100:107014. https://doi.org/10.1016/j.bspc.2024.107014.
80. Tomita N, Abdollahi B, Wei J, Ren B, Suriawinata A, Hassanpour S. Attention-based deep neural networks for detection of cancerous and precancerous esophagus tissue on histopathological slides. JAMA Network Open. 2019;2(11):e1914645. https://doi.org/10.1001/jamanetworkopen.2019.14645.
81. Garg P, Sharma M, Kumar P. Transparency in diagnosis: unveiling the power of deep learning and Explainable AI for medical image interpretation. Arabian J For Sci Eng. 2025;1–17. https://doi.org/10.1007/s13369-024-09896-5.
82. Tirupal T, Mohan BC, Kumar SS. Multimodal medical image fusion techniques-a review. Curr Signal Transduct Ther. 2021, doi: 16(2):142–63. https://doi.org/10.2174/1574362415666200226103116.
83. Acosta JN, Falcone GJ, Rajpurkar P, et al. Multimodal biomedical AI. Nat Med. 2022;28:1773–84. https://doi.org/10.1038/s41591-022-01981-2.
84. Rodis N, Sardianos C, Radoglou-Grammatikis P, Sarigiannidis P, Varlamis I, Papadopoulos GT. Multimodal explainable artificial intelligence: a comprehensive review of methodological advances and future research directions. IEEE Access. 2024. https://doi.org/10.1109/ACCESS.2024.3467062.
85. Abdelaziz M, Wang T, Anwaar W, Elazab A. Multi-scale multimodal deep learning framework for Alzheimer's disease diagnosis. Comput in Biol Med. 2025;184:109438. https://doi.org/10.1016/j.compbiomed.2024.109438.
86. de Mortanges AP, et al. Orchestrating explainable artificial intelligence for multimodal and longitudinal data in medical imaging. NPJ Digit Med. 2024;7(1):195. https://doi.org/10.1038/s41746-024-01190-w.
87. Seerala PK, Krishnan S. Grad-CAM-based classification of chest X-Ray images of pneumonia patients. In: In *advances in signal processing and intelligent recognition Systems: 6th International symposium, sirs 2020, Chennai, India, October 14-17, 2020, revised selected papers 6*. Springer; 2021. p. 161–74. https://doi.org/10.1007/978-981-16-0425-6_13.
88. Nahiduzzaman M, Abdulrazak LF, Ayari MA, Khandakar A, Islam SR. A novel framework for lung cancer classification using lightweight convolutional neural networks and ridge extreme learning machine model with SHapley Additive exPlanations (shap). Expert Syst With Appl. 2024;248:123392. https://doi.org/10.1016/j.eswa.2024.123392.
89. Vanitha K, et al. Attention-based feature fusion with external attention Transformers for breast cancer histopathology analysis. IEEE Access. 2024. https://doi.org/10.1109/ACCESS.2024.3443126.
90. Yu Q, et al. A transformer-based unified multimodal framework for Alzheimer's disease assessment. J Comput in Biol Med. 2024;180:108979. https://doi.org/10.1016/j.compbiomed.2024.108979.
91. Borys K, et al. Explainable AI in medical imaging: an overview for clinical practitioners-beyond saliency-based XAI approaches. Eur J Radiol. 2023;162:110786. https://doi.org/10.1016/j.ejrad.2023.110786.
92. Narkhede J. Comparative evaluation of post-hoc explainability methods in AI: lime, shap, and Grad-CAM. 2024 4th International Conference on Sustainable Expert Systems (ICSES). IEEE; 2024, pp. 826–30. https://doi.org/10.1109/ICSES63445.2024.10762963.
93. Miller KE, Singh H, Arnold R, Klein G. Clinical decision-making in complex healthcare delivery systems. In: Clinical engineering handbook. Elsevier; 2020. p. 858–64. https://doi.org/10.1016/B978-0-12-813467-2.00123-1.
94. Saranya A, Subhashini R. A systematic review of Explainable artificial intelligence models and applications: recent developments and future trends. Decis Analytics J. 2023;7:100230. https://doi.org/10.1016/j.dajour.2023.100230.
95. Khalifa M, Albadawy M, Iqbal U. Advancing clinical decision support: the role of artificial intelligence across six domains. Comput Methods Programs In Biomed Update. 2024;5:100142. https://doi.org/10.1016/j.cmpbup.2024.100142.
96. Afroogh S, Akbari A, Malone E, Kargar M, Alambeigi H. Trust in AI: progress, challenges, and future directions. Humanit Soc Sci Commun. 2024;11(1):1–30. https://doi.org/10.1057/s41599-024-04044-8.
97. Dwivedi R, et al. Explainable AI (XAI): core ideas, techniques, and solutions. ACM Comput Surv. 2023;55(9):1–33. https://doi.org/10.1145/3561048.
98. Saw SN, Yan YY, Ng KH. Current status and future directions of explainable artificial intelligence in medical imaging. Eur J Radiol. 2024;111884. https://doi.org/10.1016/j.ejrad.2024.111884.
99. Labkoff S, et al. Toward a responsible future: recommendations for AI-enabled clinical decision support. J Am Med Inf Assoc. 2024;31(11):2730–39. https://doi.org/10.1093/jamia/ocae209.
100. Kim SY, Kim DH, Kim MJ, Ko HJ, Jeong OR. XAI-Based clinical decision support systems: a systematic review. Appl Sci. 2024, doi: 14(15):6638. https://doi.org/10.3390/app14156638.
101. Bruch JD, Khazen M, Mahmic-Kaknjo M, Légaré F, e MEJP. Ellen, and counseling, "the effects of shared decision making on health outcomes, health care quality, cost, and consultation time: an umbrella review. 2024. p. 108408. https://doi.org/10.1016/j.pec.2024.108408.
102. Singhal A, Agrawal KK, Quezada A, Aguiñaga AR, Jiménez S, Yadav SP. Explainable artificial intelligence (XAI) model for cancer image classification. CMES-Comput Modeling In Eng Sci. 2024;141(1). https://doi.org/10.32604/cmes.2024.051363.
103. Danaher TS, Berry LL, Howard C, Moore SG, Attai DJ. Improving how clinicians communicate with patients: an integrative review and framework. J Service Res. 2023;26(4):493–510. https://doi.org/10.1177/10946705231190018.
104. 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. 2023;169:111159. https://doi.org/10.1016/j.ejrad.2023.111159.
105. Zhang Y, Weng Y, Lund J. Applications of explainable artificial intelligence in diagnosis and surgery. Diagnostics. 2022, doi: 12(2):237. https://doi.org/10.3390/diagnostics12020237.
106. Watson DS, et al. Clinical applications of machine learning algorithms: beyond the black box. BMJ. 2019;364. https://doi.org/10.1136/bmj.l886.
107. Deac A. Regulation (eu) 2016/679 of the European Parliament and of the council on the protection of individuals with regard to the processing of personal data and the free movement of these data. Erspectives of Law Public Adm. 2018;7(2):151–56.

108. Jung JY, Park D. Are AI models explainable, interpretable, and understandable? In: Human-centered artificial intelligence. Academic Press; 2022. p. 3–16. https://doi.org/10.1016/B978-0-323-85648-5.00003-7.
109. Abrantes J, Rouzrokh P. Explaining explainability: the role of XAI in medical imaging. Eur J Radiol. 2024;173. https://doi.org/10.1016/j.ejrad.2024.111389.
110. Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. International conference on machine learning. PMLR; 2017, pp. 3319–28.
111. Abbas S, Ahmed F, Khan WA, et al. Intelligent skin disease prediction system using transfer learning and explainable artificial intelligence. Sci Rep. 2025;15:1746. https://doi.org/10.1038/s41598-024-83966-4.
112. Van Zyl C, Ye X, Naidoo R. Harnessing eXplainable artificial intelligence for feature selection in time series energy forecasting: a comparative analysis of Grad-CAM and SHAP. Appl Energy. 2024;353:122079. https://doi.org/10.1016/j.apenergy.2023.122079.
113. Ahmed F, Abbas S, Athar A, et al. Identification of kidney stones in KUB X-ray images using VGG16 empowered with explainable artificial intelligence. Sci Rep. 2024;14:6173. https://doi.org/10.1038/s41598-024-56478-4.
114. Zhang Y, Liao QV, Bellamy RK. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. Proceedings of the 2020 conference on fairness, accountability, and transparency. 2020, pp. 295–305. https://doi.org/10.1145/3351095.3372852.
115. Kaur H, Nori H, Jenkins S, Caruana R, Wallach H, Vaughan JW. Interpreting interpretability: understanding data scientists' use of interpretability tools for machine learning. Proceedings of the 2020 CHI conference on human factors in computing systems. 2020, pp. 1–14. https://doi.org/10.1145/3313831.3376219.
116. Leslie D, Mazumder A, Peppin A, Wolters MK, Hagerty A. Does "AI" stand for augmenting inequality in the era of covid-19 healthcare? BMJ. 2021;372. https://doi.org/10.1136/bmj.n304.
117. Seyyed-Kalantari L, Zhang H, McDermott MBA, et al. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nat Med. 2021;27:2176–82. https://doi.org/10.1038/s41591-021-01595-0.
118. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019;366(6464):447–53. https://doi.org/10.1126/science.aax2342.
119. Norori N, Hu Q, Aellen FM, Faraci FD, Tzovara A. Addressing bias in big data and AI for health care: a call for open science. Patterns. 2021;2(10). https://doi.org/10.1016/j.patter.2021.100347.
120. Yang J, Soltan AAS, Eyre DW, et al. Algorithmic fairness and bias mitigation for clinical machine learning with deep reinforcement learning. Nat Mach Intell. 2023;5:884–94. https://doi.org/10.1038/s42256-023-00697-3.
121. Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential biases in machine learning algorithms using electronic health record data. JAMA Intern Med. 2018;178(11):1544–47. https://doi.org/10.1001/jamainternmed.2018.3763.
122. Babic B, Gerke S, Evgeniou T, Cohen IG. Beware explanations from AI in health care. Science. 2021;373(6552):284–86. https://doi.org/10.1126/science.abg1834.
123. Gagliardi J. Medical Device Software: verification, validation and compliance. Biomed Instrum Technol. 2011;45(2):95–95. https://doi.org/10.2345/0899-8205-45.2.95.
124. Angehrn Z, Haldna L, Zandvliet AS, Gil Berglund E, Zeeuw J, Amzal B, … & Heckman NM. Artificial intelligence and machine learning applied at the point of care. Front Pharmacol. 2020;11:759. https://doi.org/10.3389/fphar.2020.00759.
125. Giansanti D. The regulation of artificial intelligence in digital radiology in the scientific literature: a narrative review of reviews. In: Healthcare. Vol. 10(10). MDPI; 2022 September. p. 1824. https://doi.org/10.3390/healthcare10101824.
126. White T, Blok E, Calhoun VD. Data sharing and privacy issues in neuroimaging research: opportunities, obstacles, challenges, and monsters under the bed. Hum Brain Mapp. 2022;43(1):278–91. https://doi.org/10.1002/hbm.25120.
127. Singh R, Bapna M, Diab AR, et al. How AI is used in FDA-authorized medical devices: a taxonomy across 1, 016 authorizations. NPJ Digit. Med. 2025;8:388. https://doi.org/10.1038/s41746-025-01800-1.
128. Tosun AB, Pullara F, Becich MJ, Taylor DL, Fine JL, Chennubhotla SC. Explainable AI (xAI) for anatomic pathology. Adv Anat Pathol. 2020, Jul;27(4):241–50. https://doi.org/10.1097/PAP.0000000000000264. PMID: 32541594.
129. McNamara SL, Yi PH, Lotter W. The clinician-AI interface: intended use and explainability in FDA-cleared AI devices for medical image interpretation. NPJ Digit. Med. 2024;7(80). https://doi.org/10.1038/s41746-024-01080-1.
130. Amann J, Blasimme A, Vayena E, et al. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Med Inf Decis Mak. 2020;20:310. https://doi.org/10.1186/s12911-020-01332-6.
131. Abbas Q, Jeong W, Lee SW. Explainable AI in clinical decision support systems: a meta-analysis of methods, applications, and usability challenges. Healthcare (Basel, Switz). 2025;13(17):2154. https://doi.org/10.3390/healthcare13172154.
132. Houssein EH, Gamal AM, Younis EMG, et al. Explainable artificial intelligence for medical imaging systems using deep learning: a comprehensive review. Cluster Comput. 2025;28:469. https://doi.org/10.1007/s10586-025-05281-5.
133. Ullah N, Guzmán-Aroca F, Martínez-Álvarez F, De Falco I, Sannino G. A novel explainable AI framework for medical image classification integrating statistical, visual, and rule-based methods. Med Image Anal. 2025;103665. https://doi.org/10.1016/j.media.2025.103665.

Publisher's Note

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