Issues in Information Systems
Volume 26, Issue 4 pp. 670-685, 2026

DOI: https://doi.org/10.48009/4_iis_2026_154

From saliency to accountability: A systematic review of explainable AI validation in radiology

Vahini Atluri, *Dakota State University, vahini.atluri@trojans.dsu.edu*
Omar El-Gayar, *Dakota State University, omar.el-gayar@dsu.edu*

Abstract

Explainable artificial intelligence (XAI) is increasingly positioned as essential for clinical trust and regulatory approval of deep learning systems in radiology. This systematic literature review synthesizes 65 empirical studies published between 2020 and 2026 to characterize XAI methods and evaluate validation practices in radiological imaging. Attribution-based saliency methods, particularly Grad-CAM variants, dominated the field. However, explanation fidelity testing remained critically limited, only 11% performed localization agreement with expert annotations, 3% employed sanity checks, and 20% assessed faithfulness metrics. Perturbation testing appeared in 57% of studies. Human-centered evaluation was conducted in 28% through expert radiologist assessment, with 58% employing user studies that predominantly measured perceived trust rather than decision quality. Behavioral calibration assessment occurred in only 9% of studies. Clinical outcome validation was rare. The findings demonstrate widespread XAI adoption alongside substantial validation gaps, requiring standardized fidelity metrics, behavioral paradigms measuring calibrated reliance and clinical outcome validation.

Keywords: Explainable AI, XAI validation, Radiology, Grad CAM, Explanation fidelity

Introduction

Deep learning has transformed diagnostic radiology, achieving high performance across image classification, detection, and segmentation tasks in modalities such as chest radiography, CT, MRI, and mammography (Esteva et al., 2019; Rajpurkar et al., 2018). Despite these advances, the opacity of deep neural networks remains a major barrier to clinical integration. Unlike conventional diagnostic systems with transparent reasoning pathways, deep learning models generate predictions through complex internal representations that are often inaccessible to human interpretation (Rudin, 2019). In high-stakes clinical environments, this lack of transparency raises concerns regarding trust, accountability, regulatory oversight, and patient safety.

Explainable artificial intelligence (XAI) has emerged as a proposed solution to these concerns by generating visual or feature-based explanations intended to make model predictions more understandable (Barredo Arrieta et al., 2020). In radiology, attribution-based approaches such as saliency maps and Grad-CAM variants have become especially prevalent because they highlight image regions associated with model predictions (Selvaraju et al., 2017). Their adoption accelerated substantially between 2020 and 2026 due to growing regulatory emphasis on trustworthy AI, increased clinical deployment of deep learning systems, and the rapid expansion of imaging AI during the COVID-19 pandemic.

However, the rapid adoption of XAI has outpaced the development of rigorous validation practices. Many studies present visually plausible explanations without establishing whether those explanations faithfully represent the model's reasoning process or correspond to clinically meaningful pathology. Prior research

demonstrates that saliency methods may produce unstable or misleading explanations while still appearing convincing to users (Adebayo et al., 2018; Ghassemi et al., 2021). Consequently, explanation plausibility should not be mistaken for explanation validity.

This problem introduces evaluation requirements beyond traditional predictive performance metrics. While model evaluation focuses on accuracy and sensitivity, XAI evaluation must additionally assess whether explanations are faithful, robust, clinically meaningful, and behaviorally beneficial (Doshi-Velez & Kim, 2017). Yet several important validation gaps remain, localization agreement with expert annotations is rare. Perturbation and sanity-check testing remain inconsistent despite their importance for assessing explanation robustness(Adebayo et al., 2018; Asmita et al., 2026). Human-centered evaluation frequently measures perceived trust or satisfaction rather than calibrated reliance, decision quality, or automation bias(Kinger & Kulkarni, 2025; Singhal et al., 2024). Evidence demonstrating workflow improvement, sustained adoption, or patient-level benefit is similarly limited (Lysdahlgaard, 2023; Wu et al., 2024).

Although prior reviews (D. Muhammad & Bendechache, 2024; F. Muhammad et al., 2026) summarize XAI methods in healthcare and medical imaging, limited attention has been devoted specifically to validation practices in radiological XAI. Existing reviews primarily emphasize technical categorization of explanation methods rather than systematic evaluation of explanation fidelity, behavioral impact and clinical utility. To address this gap, this systematic literature review synthesizes empirical studies published between 2020 and 2026 that evaluate explainable AI methods in radiological deep learning systems. The study addresses the following research questions:

RQ1: What explainable AI methods are currently employed in radiological deep learning systems, and how does their adoption evolve over time?
RQ2: To what extent do current studies validate explanation quality through quantitative metrics, perturbation testing, localization agreement with expert annotations, and faithfulness assessment?
RQ3: How does human-centered evaluation proceed, and what is the balance between measuring subjective perceptions versus objective outcomes?
RQ4: What evidence exists for clinical outcome improvements, including diagnostic accuracy impact, workflow integration, and patient-level outcomes, that are attributable to XAI augmented diagnostic systems?

Methodology

Search Strategy

This systematic review was conducted following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 guidelines (Page et al., 2021). Publications were retrieved from seven major databases: ABI/INFORM, ACM Digital Library, Academic Search Premier, Business Search Premier, IEEE Xplore, PubMed, and Web of Science for studies published between 1st January 2020 and 15th April 2026. The search strategy combines four concept groups using Boolean operators. The deep learning terms include ("deep learning" OR "neural network*"). The radiology terms include ("radiology" OR "medical imaging" OR "diagnostic imaging"). The explainable AI terms include ("explainable artificial intelligence" OR "XAI" OR "Interpretable AI" OR "Grad-CAM" OR "saliency map"). The validation terms include ("validation" OR "evaluation" OR "fidelity" OR "clinical outcome" OR "human-centered"). This composite string, joining all four concept groups with the boolean and operator, was applied identically across all seven databases without platform-specific modification. Searches were limited to peer-reviewed journal articles, conference proceedings, and book chapters. Conference proceedings were included because

methodological innovation in XAI validation frequently appears first in computer science venues prior to journal publication.

Study Selection
Studies were excluded if they involved non-radiological imaging (histopathology, dermatology), lacked XAI implementation, provided only informal qualitative inspection, review articles without original data, or involved screening. Studies were included if they met all of the following criteria: application of deep learning models to radiological imaging data for clinical diagnostic purposes, explicit implementation of at least one explainable AI method generating explanations, at least one explicit validation approach beyond informal visual inspection. This criterion was operationalized as the presence of at least one of the following: a quantitative fidelity or faithfulness metric, perturbation or sanity-check testing, localization agreement analysis or a structured human-centered evaluation (expert rating or user study). Narrative description of heat maps without an accompanying metric or structured assessment did not satisfy this criterion.

Each study was coded using a structured extraction protocol capturing five domains: study design, XAI implementation, explanation fidelity validation, human-centered evaluation and reporting transparency consistent with the domains described in synthesis of validation approaches. Coding categories and operational definitions were established prior to data extraction and applied consistently across all 65 included studies.

Synthesis of Validation Approaches
Included studies were assessed in five domains: study design (clear objectives, appropriate design), XAI implementation (method description, implementation details), explanation fidelity validation (quantitative assessment, localization agreement, perturbation testing, overall rigor), human-centered evaluation (expert evaluation, decision quality, behavioral effects, overall rigor), and reporting transparency (reproducibility, limitations, data/code availability). Studies were classified as high quality (strong design with rigorous multi-domain validation), moderate quality (adequate design with validation in at least two domains), low quality (weak design or single domain validation), or very low quality (methodological concerns). Given substantial heterogeneity in imaging modalities, XAI methods, validation protocols, outcome measures, and quantitative meta-analysis was not methodologically appropriate. Descriptive statistics and a validation depth score (range: 1-7) were calculated by counting distinct validation approaches.

Results

As shown in Figure 1, the combined search across databases identified 210 records. After removing 78 duplicates, 132 unique records underwent title and abstract screening. Title and abstract screening excluded 18 articles, resulting in 114 records for full-text assessment. Of these, 38 full texts were unavailable despite direct author contact attempts, leaving 76 full texts reviewed. After applying the eligibility criteria, 65 studies met all the inclusion requirements. The 11 studies excluded at the full-text stage included 6 focusing on non-radiology imaging, 5 lacking explicit validation.

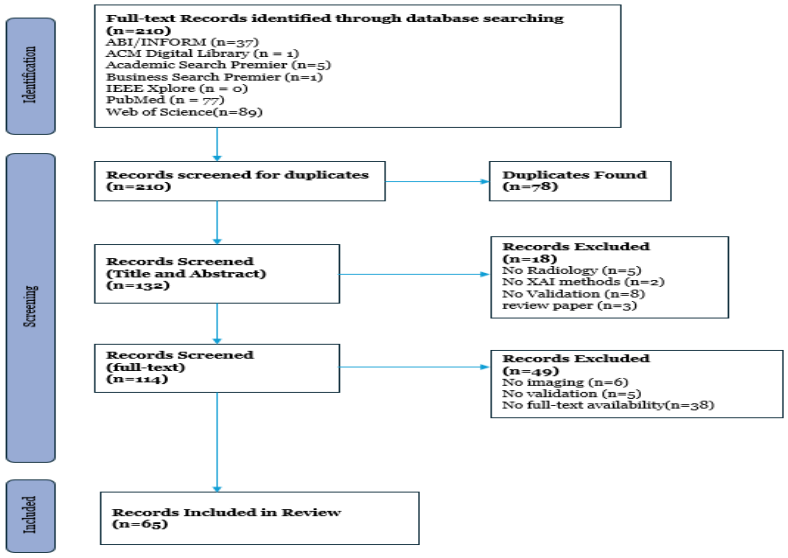


Figure 1: PRISMA flow chart

The 65 included studies span from 2021 to 2026, with marked temporal concentration: 2 studies (3.1%) from 2021, 4 (6.2%) from 2022, 5 (7.7%) from 2023, 9 (13.8%) from 2024, 27 (41.5%) from 2025 and 18 (27.7%) through partial publication year of 2026, as it only reflects four months of publications to year-over-year comparison. The increase from 2024 onward, with 45 studies (69.2%) published in 2025-2026, reflects accelerating interest in XAI validation. Studies originated from 27 countries, highest from the United States (18 studies, 27.7%), China (12, 18.5%), and India (8, 12.3%).

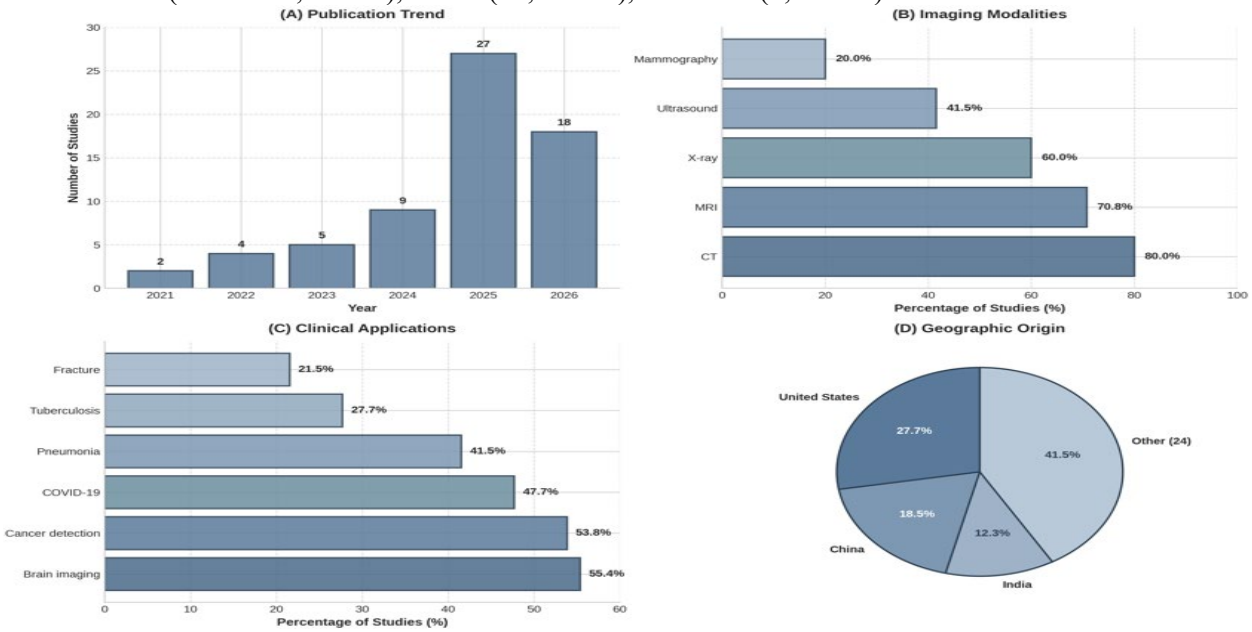


FIGURE 2: Temporal Trends in XAI Validation

Studies employed diverse imaging modalities, with the majority investigating multiple modalities (87.7%). Computed tomography was most prevalent (52 studies, 80.0%), followed by MRI (46, 70.8%) and chest X-ray (39, 60.0%). Ultrasound appeared in 27 (41.5%), mammography in 13 (20.0%). Clinical applications: brain imaging (36, 55.4%), cancer detection (35 studies, 53.8%), COVID-19 detection (31, 47.7%),

pneumonia (27, 41.5%), tuberculosis (18, 27.7%), fracture detection (14, 21.5%). The COVID-19 emphasis reflects pandemic-driven priorities (2020-2022).

Research Question 1: What explainable AI methods are currently employed in radiological deep learning systems, and how does their adoption evolve over time?

Attribution-based saliency methods dominated the field, with Grad-CAM and its variants representing the overwhelming majority. Standard Grad-CAM appeared in 47 studies (72.3%), while Grad-CAM *variants,* including Grad-CAM++, Score-CAM, XGrad-CAM, and Layer-CAM, appeared in 18 studies (27.7%), bringing the total Grad-CAM family prevalence to 59 studies (90.8%). Other attribution methods were also common: LIME in 26 studies (40.0%), SHAP in 28 studies (43.1%), and gradient-based saliency maps in 27 studies (41.5%). Attention visualization methods appeared in 35 studies (53.8%), including transformer attention weights in 22 studies (33.8%) and CNN attention modules in 18 studies (27.7%). The vast majority (58 studies, 89.2%) employed multiple XAI methods simultaneously, typically combining Grad-CAM with complementary approaches to provide multiple perspectives on model decision-making.

Explicit justification for the selection of the XAI method was provided in only 31 studies (47.7%). Common rationales included widespread use in medical imaging (23 studies), architecture-agnostic properties (12 studies), visual interpretability (18 studies), and computational efficiency (7 studies). Few studies (8, 12.3%) provided a comparative analysis of multiple XAI methods to justify their selection. Among the five studies that conducted systematic comparative evaluations, the findings suggested that Grad-CAM achieved higher localization agreement than LIME or SHAP. (Cerekci et al., 2024) reported mean IoU of 0.64 for Grad-CAM versus 0.51 for LIME and 0.48 for SHAP in mammogram analysis. *(Asmita et al., 2026)* found Grad-CAM achieved 68% expert preference versus 18% for LIME and 14% for SHAP across multiple imaging domains. These comparative studies provide empirical support for Grad-CAM's dominance beyond mere convention or familiarity.

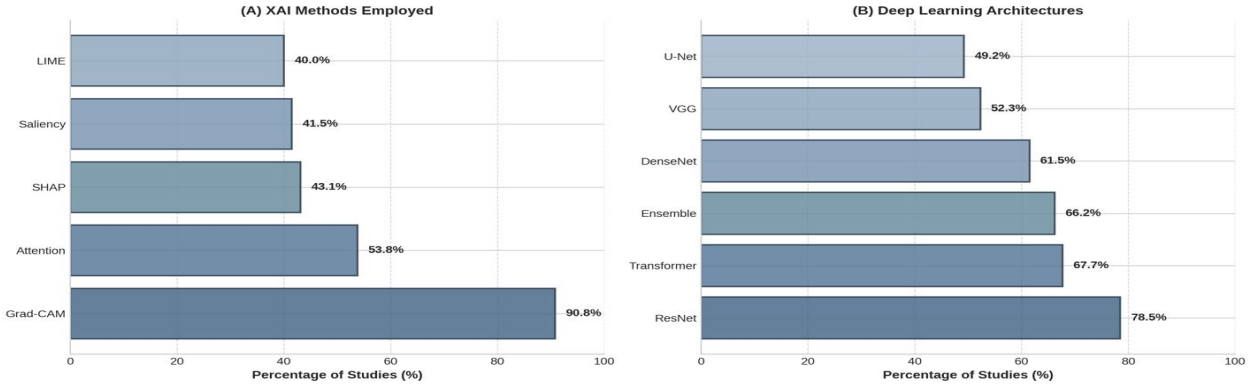


FIGURE 3: XAI Method Adoption and Architecture Distribution

Deep learning architectures employed included ResNet variants in 51 studies (78.5%), DenseNet variants in 40 studies (61.5%), EfficientNet in 26 studies (40.0%), and transformer-based models in 44 studies (67.7%). U-Net architectures for segmentation tasks appeared in 32 studies (49.2%), VGG architectures in 34 studies (52.3%), and YOLO architectures for detection in 11 studies (16.9%). Ensemble approaches combining multiple architectures appeared in 43 studies (66.2%), raising questions about explanation aggregation when models disagree. Multimodal approaches integrating explanations across multiple imaging modalities or data types appeared in 42 studies (64.6%). However, only 15 studies (23.1%) explicitly addressed cross-modal explanation consistency or aggregation strategies.

Research Question 2: To what extent do current studies validate explanation quality through quantitative metrics, perturbation testing, localization agreement with expert annotations, and faithfulness assessment?

All studies reported some form of quantitative assessment, but depth and rigor varied substantially. The most rigorous fidelity test, localization agreement with expert-annotated ground truth, appeared in only 7 studies (10.8%). These studies compared explanation heat maps to radiologist-annotated pathology using metrics including Intersection over Union, Dice coefficient and Hausdorff distance. (Cerekci et al., 2024) reported IoU values ranging from 0.32 to 0.78 across different XAI methods in mammography, with a mean of 0.58 (SD 0.15). (Buga et al., 2025) assessed localization agreement in brain metastasis monitoring, achieving Dice coefficients of 0.67-0.74 between Grad-CAM outputs and expert annotations. Inter-rater agreement among expert annotators was reported in 4 of the 7 studies, with kappa values ranging from 0.61 to 0.84, indicating substantial to almost perfect agreement.

Perturbation testing appeared in 37 studies (56.9%), representing the most common rigorous fidelity assessment approach. Specific methods included insertion and deletion curves in 28 studies (43.1%), RISE (Randomized Input Sampling for Explanation) in 8 studies (12.3%), meaningful perturbation approaches in 7 studies (10.8%), and occlusion sensitivity analysis in 12 studies (18.5%). Studies employing perturbation testing generally reported moderate correlation between saliency and predictive importance, with correlation coefficients ranging from 0.45 to 0.72. (Aravinda et al., 2026) conducted extensive perturbation analysis in fracture detection, reporting insertion curve AUC of 0.73 and deletion curve AUC of 0.69, demonstrating reasonable but imperfect correspondence.

Sanity checks, which verify that explanations change appropriately when model parameters or training data are randomized, appeared in only 2 studies (3.1%). This represents perhaps the most severe validation gap. *(Asmita et al., 2026)* performed model parameter randomization tests across multiple XAI methods, finding that some saliency techniques produced similar outputs despite complete weight randomization, indicating failure of fundamental faithfulness requirements. Data label randomization is also an additional sanity check, demonstrating that Grad-CAM appropriately changed explanations when trained on randomized labels while some other methods failed this test.

Faithfulness metrics, which directly quantify correspondence between explanation importance weights and actual feature contributions to predictions, appeared in 13 studies (20.0%). Specific approaches included pixel-flipping in 7 studies (10.8%), region perturbation in 6 studies (9.2%) and infidelity scoring in 3 studies (4.6%). *(E Ihongbe et al., 2024)* assessed multiple faithfulness metrics in chest radiology, finding correlation coefficients between 0.52 and 0.68 depending on pathology type, indicating moderate but imperfect faithfulness.

Multiple fidelity validation approaches were common: 25 studies (38.5%) employed a single fidelity method beyond basic metrics, 28 studies (43.1%) employed two methods, and 12 studies (18.5%) employed three or more. The mean number of distinct fidelity approaches per study was 2.1 (SD 1.3, range 1-5).

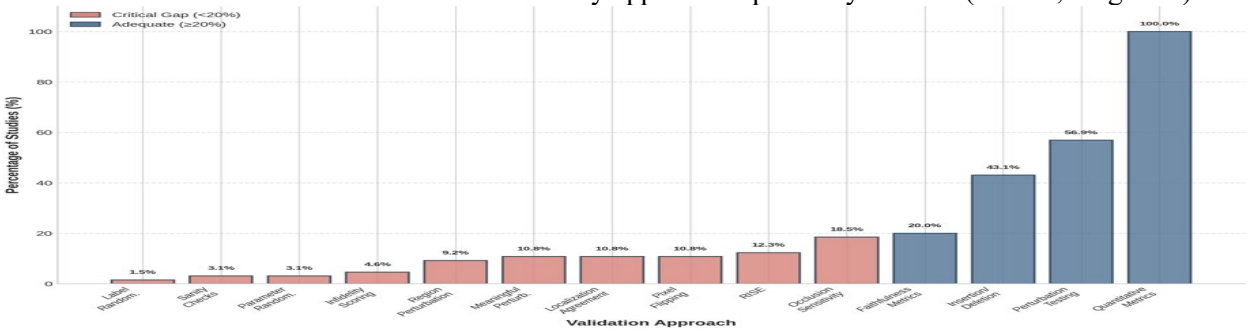


Figure 4: Fidelity Validation Approaches

Research Question 3: How does human-centered evaluation proceed, and what is the balance between measuring subjective perceptions (trust, satisfaction) versus objective outcomes (decision quality, behavioral calibration)?

Expert radiologist evaluation was conducted in 18 studies (27.7%), with the number of experts ranging from 1 to 12 (mean 3.8, SD 2.4). Evaluation focus varied substantially: comprehensibility assessment in 15 studies (23.1%), clinical utility perception in 12 studies (18.5%), localization accuracy judgment in 7 studies (10.8%) and diagnostic confidence impact in 8 studies (12.3%). Expert evaluation typically employed Likert-scale ratings of explanation quality, with mean scores ranging from 3.2 to 4.6 on 5-point scales across studies. However, evaluation protocols were rarely standardized, and only 4 studies (6.2%) reported inter-rater agreement among expert evaluators, limiting confidence in the reliability of expert judgments. The lack of protocol standardization means that an explanation rated highly by experts in one study may not receive similar ratings in another, precluding cross-study comparison of explanation quality. User studies were conducted in 38 studies (58.5%), with sample sizes ranging from 5 to 127 participants (mean 28, SD 23). Participant types included practicing radiologists in 24 studies (36.9%), radiology residents in 18 studies (27.7%), medical students in 12 studies (18.5%), and non-clinical participants in 8 studies (12.3%). The inclusion of non-clinical participants in user studies raises questions about ecological validity, as clinical expertise likely influences how individuals interact with and benefit from medical explanations. User study focus varied substantially: perceived trust measured in 32 studies (49.2%), satisfaction assessed in 22 studies (33.8%), comprehension evaluated in 16 studies (24.6%), decision quality measured in 11 studies (16.9%), and time efficiency tracked in 9 studies (13.8%).

The predominance of perceived trust assessment (49.2%) over decision quality measurement (16.9%) represents a critical gap in human-centered evaluation. Trust perception does not guarantee appropriate reliance, as users may trust incorrect predictions or distrust correct ones based on explanation plausibility rather than validity. Behavioral calibration assessment, which measures whether XAI systems improve decision quality by supporting appropriate skepticism and calibrated confidence, occurred in only 6 studies (9.2%). These studies examined appropriately calibrated reliance, where users accept correct AI predictions and reject incorrect ones at higher rates than without explanations.

Ihongbe et al. (2024) conducted controlled experiments measuring diagnostic accuracy with versus without XAI, finding +6.2% improvement on AI-correct cases but only +1.8% on AI-incorrect cases, suggesting explanations help clinicians recognize correct diagnoses but provide limited assistance identifying AI errors. This asymmetric benefit pattern raises concerns about potential automation bias, where compelling explanations lead to uncritical acceptance even when predictions are wrong. (Kinger & Kulkarni, 2025) assessed confidence calibration, finding that explanations increased clinician confidence by 12% but actual accuracy improved only 4%, indicating potential overconfidence. (Singhal et al., 2024) explicitly examined automation bias, finding that 32% of participants accepted incorrect AI predictions more readily when explanations appeared plausible, demonstrating XAI's potential to exacerbate rather than mitigate bias. These mixed findings highlight that XAI effects on human decision-making are not uniformly beneficial and require empirical validation for each application context.

Among the 11 studies measuring decision quality, improvements in diagnostic accuracy ranged from -1.2% to +12.5% (mean +4.2%, SD 3.8%). *(F. Muhammad et al., 2026)* reported the largest improvements (+12.5%) but in a small sample (n=15 radiologists), limiting generalizability. (Zaheer et al., 2025) found null effects (0.0%) in prostate cancer grading, while one study reported slight degradation (-1.2%). Four studies reported no significant improvement, highlighting that XAI benefits cannot be assumed and must be empirically demonstrated. Subgroup analysis revealed larger improvements on AI-correct cases (mean

+2.8%) than AI-incorrect cases (mean +1.4%), suggesting explanations help clinicians recognize correct diagnoses but provide less benefit in identifying AI errors.

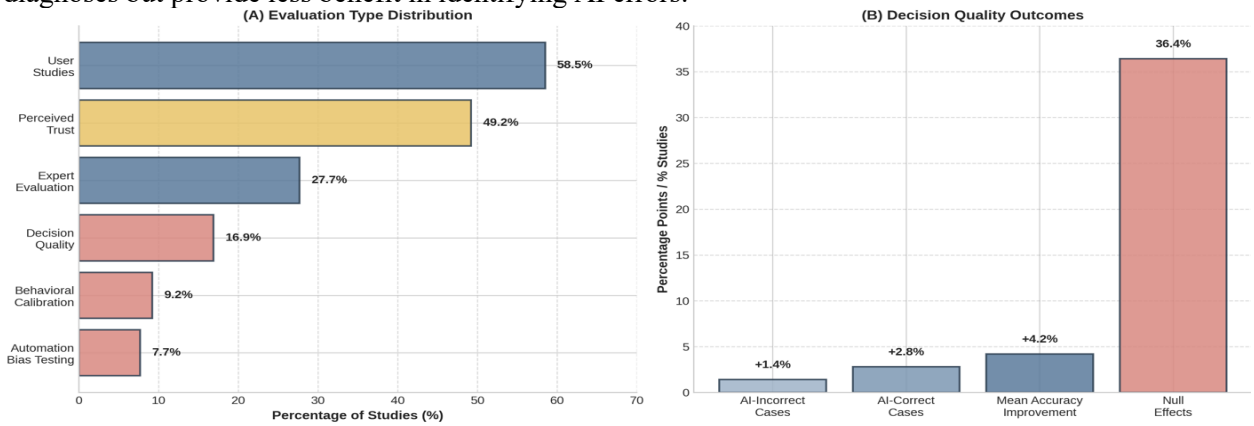


Figure 5: Human-Centered Evaluation Approaches

Research Question 4: What evidence exists for clinical outcome improvements, including diagnostic accuracy impact, workflow integration, and patient-level outcomes, that are attributable to XAI augmented diagnostic systems?

Clinical outcome assessment was reported in 23 studies (35.4%), but the depth of reporting varied substantially. Diagnostic accuracy improvement was assessed in 15 studies (23.1%), workflow efficiency in 8 studies (12.3%) and real-world deployment in 3 studies (4.6%). No studies tracked patient-level outcomes such as diagnostic error rates in clinical practice or treatment consequences.

The 15 studies assessing diagnostic accuracy reported mean improvements of 4.2 percentage points (SD 3.8%, range -1.2% to +12.5%). Four studies reported null effects. Subgroup analysis revealed larger improvements on AI-correct cases (mean +2.8%) than AI-incorrect cases (mean +1.4%). Workflow impact assessments across 8 studies reported time-to-diagnosis reductions ranging from 5% to 32% (mean 18%), though integration challenges were noted. Two studies found no significant workflow changes.

Only 3 studies (4.6%) conducted prospective evaluation in clinical settings. (O'Shea et al., 2023) deployed weakly supervised segmentation with XAI in a radiology department, reporting 72% acceptance rate. (Lysdahlgaard, 2023) conducted prospective evaluation of heat map explanations, finding initial enthusiasm but declining usage from 85% in month one to 42% at six months. (Wu et al., 2024) implemented XAI-augmented ultrasound diagnosis for necrotizing enterocolitis, achieving 78% clinician acceptance but requiring significant training and workflow modifications. These prospective studies revealed clinician skepticism (acceptance rates 65-82%), workflow integration challenges and declining usage over time.

We computed a validation depth score by counting distinct validation approaches employed in each study. The mean score was 3.29 (SD 1.52, range 1-7), indicating most studies employed multiple validation approaches but rarely achieved comprehensive assessment across all validation domains. The distribution of validation depth scores showed 17 studies (26.2%) with 1-2 approaches, 31 studies (47.7%) with 3-4 approaches, and 17 studies (26.2%) with 5 or more approaches. Studies with scores of 5 or higher, including *(Aravinda et al., 2026; Asmita et al., 2026; Cerekci et al., 2024; E Ihongbe et al., 2024)* and others demonstrated substantially more methodological rigor and were more likely to report nuanced findings, including contexts where explanations failed or provided limited utility.

Cross-domain validation patterns revealed that 22 studies (33.8%) conducted validation only in the explanation fidelity domain, 36 studies (55.4%) validated across fidelity and human-centered domains, and

only 6 studies (9.2%) validated across all three domains, including clinical outcome assessment. No studies validated human-centered or clinical outcomes without also conducting fidelity validation, suggesting a hierarchical approach where technical fidelity is considered a prerequisite to higher-level validation. However, 11 studies (16.9%) conducted no human-centered evaluation whatsoever, relying solely on technical fidelity metrics without assessing how clinicians actually interact with or benefit from explanations.

Quality appraisal yielded overall ratings: high quality for 12 studies (18.5%), moderate quality for 38 studies (58.5%), low quality for 13 studies (20.0%), and very low quality for 2 studies (3.1%). Domain-specific quality: XAI implementation quality was high (93.8% of methods were clearly described). Explanation fidelity validation achieved moderate or high rigor in 61.5% of studies. Human-centered evaluation showed the weakest performance, with moderate or high rigor in only 36.9% of studies. Reporting transparency: 89.2% acknowledged limitations, but only 47.7% provided methods for full reproducibility, and 41.5% made code or data available. Coding consistency across the 65 included studies is summarized in aggregate in figures 3, 4 & 5. A complete study-level coding matrix, recording the validation tier's achieved, validation depth score and quality rating for each included study, is available from the authors upon request.

Temporal trends showed modest improvement: mean validation depth increased from 2.3 (2021-2022) to 3.6 (2025-2026). Human-centered evaluation increased from 16.7% to 31.1%, and clinical outcome validation from 0% to 42.2%. However, substantial gaps persist even in recent literature: localization agreement (11%), sanity checks (3%), behavioral calibration (9%), and patient outcome tracking (0%) remain critically underutilized.

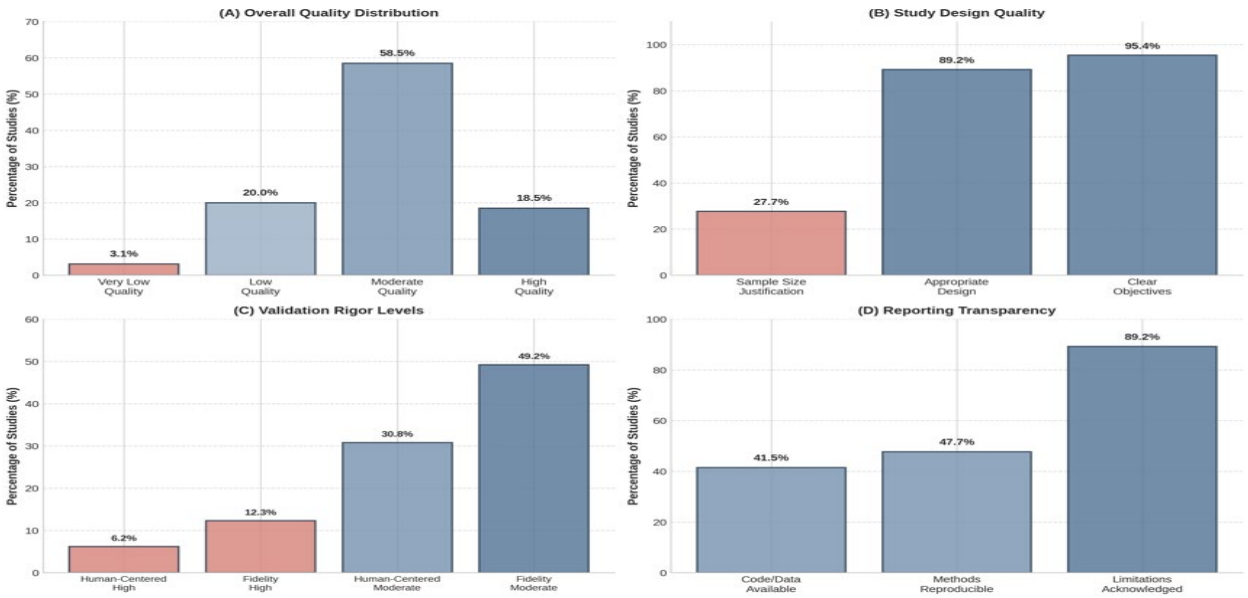


Figure 6: Quality Appraisal

Discussion

The findings reveal widespread adoption of XAI in radiology alongside substantial validation deficiencies across fidelity, behavioral, and clinical outcome assessment domains. Collectively, they suggest that

radiological XAI development has progressed more rapidly than the validation frameworks required to support trustworthy clinical deployment.

Four-Tier Framework for XAI Validation

An integrated four-tier framework is proposed that organizes XAI validation from foundational technical assessment through clinical impact. This framework emerges from observed patterns in how studies approach validation. It serves as both an understanding of current practices and as a normative guide for comprehensive validation.

Tier 1 - Algorithmic Transparency: This tier addresses what XAI methods researchers employ and how those methods technically generate explanations. This includes XAI method category (gradient-based, perturbation-based, attention-based), technical implementation details, and explanation output format. All 65 studies necessarily engage with Tier 1 through method selection and implementation. This represents the baseline entry point for XAI research.

Tier 2 - Explanation Fidelity: This tier evaluates whether explanations accurately reflect the ground-truth pathology and the model's decision processes. Fidelity validation assesses correspondence between explanations and pathological ground truth. Do highlighted regions align with expert-annotated lesions? It also assesses model decision processes. Do explanations faithfully represent how the model actually makes predictions, as opposed to post hoc rationalization? Tier 2 methods include localization agreement with expert annotations, perturbation testing, sanity checks through parameter randomization and faithfulness metrics assessing correspondence between explanations and model decision processes. All studies conduct some form of Tier 2 validation, but rigor varies substantially. Only 18.5% achieve comprehensive fidelity assessment across multiple methods.

Tier 3 - Human Centered Evaluation: This tier assesses how clinicians interact with and respond to XAI outputs. This includes comprehension assessment (do clinicians understand the explanations?), appropriate reliance measurement (do explanations improve clinician accuracy on correct AI predictions while maintaining accuracy on incorrect predictions?), behavioral calibration (do explanations reduce automation bias and support appropriate skepticism?), and diagnostic process integration (how do explanations fit into clinical workflows?). Tier 3 methods include expert evaluation, user studies examining decision quality and behavioral calibration assessment. Although human-centered evaluation appears in many studies, most focus on subjective trust or satisfaction rather than calibrated reliance and diagnostic performance. This is the weakest domain across all tiers.

Tier 4 - Clinical Outcome Validation: This tier assesses real-world impact, including diagnostic accuracy improvement in ambiguous cases (23.1%), time to diagnosis reduction (12.3%), and, when feasible, patient outcome effects. Tier 4 validation requires prospective studies, ideally with designs that compare XAI augmented versus standard diagnostic workflows. Only 35.4% of studies conduct any Tier 4 validation. Just 4.6% employ prospective deployment. Zero percent track patient-level outcomes. The near absence of Tier 4 validation represents the most severe gap. Without evidence of outcomes, the clinical utility proposition of XAI systems remains largely theoretical.



Figure 7: Four-Tier Evaluation Framework

This four-tier framework reveals that current validation practices exhibit an inverted pyramid. There is universal engagement at Tier 1 (algorithmic transparency). There is substantial but incomplete validation at Tier 2 (explanation fidelity). There is limited assessment at Tier 3 (human-centered evaluation). There is minimal evidence at Tier 4 (clinical outcomes). Comprehensive validation across all four tiers occurs in only 9% of studies. This indicates that the field has prioritized technical development and fidelity assessment over human-centered and clinical outcome validation.

Grad-CAM Dominance

Grad CAM's overwhelming prevalence (91%) likely reflects multiple factors. These include computational efficiency such as 0.3 seconds per image compared to 12 to 18 seconds for LIME or SHAP based on comparative studies). They also include architecture-agnostic properties that allow applications across diverse model types. They also include visually coherent outputs that align with clinical intuition about lesion localization. The five comparative studies provide empirical support for Grad CAM's advantages. *(Asmita et al., 2026; Cerekci et al., 2024)* find that Grad CAM achieves higher localization agreement than LIME or SHAP (mean IoU 0.64 versus 0.51 and 0.48) and higher expert preference (68% versus 18% and 14%). However, 92% of studies lack systematic comparisons of methods. This means the field has collectively selected a dominant method without rigorous comparative evaluation in most cases.

This monoculture adoption carries risks. Diverse XAI approaches, including attention mechanisms, concept-based explanations, and example-based methods, may offer complementary advantages for different tasks. These remain underexplored. Moreover, Grad CAM's dominance may reflect path dependence and network effects, such as established implementations, validation examples and reviewer familiarity, rather than genuine superiority across all contexts. The field should resist convergence on a single method. Instead, researchers should pursue systematic benchmarking that compares multiple approaches on standardized validation protocols. Such benchmarking would reveal when Grad CAM genuinely outperforms alternatives versus when method choice is arbitrary or context dependent.

Validation Inversion Problem

Current practices exhibit a pattern here termed validation inversion. This means substantial effort goes into computationally straightforward assessments, such as quantitative metrics and user surveys, while researchers neglect rigorous but resource-intensive approaches, such as localization agreement, sanity checks, and behavioral studies. All studies report quantitative assessments, but these often measure perceptual plausibility rather than ground truth fidelity. Metrics like average saliency on predicted lesion regions or spatial coverage of suspicious areas conflate explanation with prediction performance. They do not verify that highlighted regions contain actual pathology.

Localization agreement with expert annotations is widely regarded as one of the strongest available approaches for fidelity assessment, yet it remains relatively uncommon in the literature. The seven studies that conduct this validation demonstrate its feasibility but also its resource intensity. Acquiring expert annotations requires radiologist time, multiple raters for reliability assessment, and careful protocol design to ensure annotation quality. The rarity of this approach likely reflects these resource barriers rather than a lack of recognition that ground truth comparison represents the most definitive fidelity test.

The limited use of sanity checks represents an important methodological concern. These fundamental tests verify that explanation methods respond to the trained model rather than producing visualization artifacts that are independent of learned parameters. *(Asmita et al., 2026; D. Muhammad & Bendechache, 2024)* demonstrate that several popular XAI methods fail parameter randomization tests. These methods produce similar outputs regardless of trained weights. Yet 97% of studies have not conducted these basic verification

tests. This means most deployed explanations may fail fundamental faithfulness requirements without researchers even being aware.

The ratio of trust surveys to behavioral studies is 5.4 to 1. This reflects resource differentials but creates dangerous measurement gaps. Measuring subjective trust is straightforward: show explanations, then ask users to rate them. Measuring behavioral impact requires controlled experiments that compare diagnostic accuracy with and without XAI on identical cases. It requires challenging task designs to assess calibration. It requires sufficient sample sizes for statistical power. High perceived trust does not guarantee appropriate reliance. Users may overtrust incorrect predictions because explanations appear plausible. (Singhal et al., 2024) find that 32% accept incorrect predictions more readily with XAI. Users may undertrust correct predictions despite valid explanations. Users may experience miscalibrated confidence where they feel certain in decisions that remain uncertain. (Kinger & Kulkarni, 2025) finds that confidence increases 12% but accuracy only 4%. Only behavioral paradigms that measure decision quality across AI correct and AI incorrect cases can distinguish these scenarios. Yet such assessment occurs in only 9% of studies. Because perceived trust does not guarantee appropriate reliance, behavioral impact should be evaluated empirically rather then inferred from subjective trust measure alone.

Clinical Outcome Vacuum

The near absence of clinical outcome validation represents one of the most important gaps in the current radiological XAI literature. Only three studies conducted prospective deployment and none tracked patient-level outcomes. The limited prospective evidence revealed clinician skepticism, integration challenges and declining usage over time, indicating that short-term laboratory benefits do not necessarily translate into durable adoption. This gap has direct implications for all stakeholders. Researchers must prioritize localization agreement testing and sanity checks as routine validation. Human-centered evaluation should measure decision quality through controlled experiments comparing accuracy with and without XAI on identical cases, not just trust perception. Clinical outcome validation should pursue prospective deployment with longitudinal tracking. Regulatory frameworks should establish tiered requirements based on risk level, requiring localization agreement and perturbation testing for approval. Healthcare systems should require local validation before deployment, establish monitoring systems, and provide training emphasizing that explanations are tools for critical evaluation, not proof of correctness.

Future Research

A systematic analysis of the reviewed literature revals critical research questions and methodological needs across four key areas.

Explanation Fidelity Research: These include questions such as, what factors predict when explanation fidelity will be high versus low? How do different perturbation approaches relate to each other and to localization agreement? And can automated fidelity assessment reduce reliance on expensive expert annotations? Of seven localization agreement studies, pathology complexity, image quality and model confidence may moderate fidelity, but systematic investigation is needed. The 37 perturbation studies used heterogeneous methods without establishing whether they measure the same construct or provide complementary information. Only 11% achieved ground truth via localization agreement. Scalable alternatives enabling more widespread fidelity testing are urgently needed.

Human Centered Evaluation Research: Questions include what explains the discrepancy between trust perception and decision quality? Under what conditions does XAI reduce versus exacerbate automation bias? How do novice and expert clinicians use explanations differently? And what explanation properties beyond fidelity support diagnostic reasoning? 49% of user studies measure trust but only 17% measure decision quality, with variable relationships between subjective ratings and objective performance. Six

behavioral studies showed mixed evidence on automation bias (increase vs. reduction), suggesting unidentified moderators. Studies included mixed expertise levels (24 radiologists, 18 residents, 12 students) but rarely compared usage patterns or effectiveness across expertise levels.

Clinical Integration Research: Representative questions include: What explains the decline in XAI usage over time in prospective deployments? How should XAI systems adapt explanations to time pressure and clinical context? What training approaches optimize clinician calibration with XAI systems? And how can XAI support clinical education and accelerate trainee learning? One study found XAI usage declined from 85% to 42% over six months and root causes unknown. No studies examined context-adaptive explanations, though clinical practice ranges from triage to detailed review. Only 6% addressed calibration and optimal training for appropriate reliance remains unexplored.

Method Comparison and Innovation Research: Example questions include when concept-based or prototype explanations are superior to saliency maps? How should ensemble XAI aggregate conflicting explanations across models? What makes multimodal explanations effective versus confusing? What are the limits of saliency based explanations? Only 8% used non-saliency approaches and their comparative advantages remain unknown. 62% used ensembles but only 23% systematically addressed aggregation, leaving whether explanation disagreement signals uncertainty unresolved. 65% integrated multiple modalities without systematic evaluation of integration approaches or presentation formats.

Conclusion

This systematic review demonstrates that while explainable AI methods have been widely adopted in radiological deep learning, with Grad-CAM variants employed in 91% of 65 analyzed studies, validation practices remain critically incomplete across all assessment dimensions. A four-tier validation framework is proposed, organizing assessment from algorithmic transparency through explanation fidelity and human-centered evaluation to clinical outcomes, emerging from analyzing of how the reviewed studies approached validation. Comprehensive multi-tier validation occurs in only 9% of studies despite representing the minimum requirements for clinical accountability.

Explanation fidelity testing remains limited, particularly for approaches assessing correspondence with expert annotations, explanation faithfulness, and robustness to model perturbation. The few studies that conduct rigorous fidelity assessments report moderate correlation values (0.45-0.72), suggesting that many explanations provide imperfect signals of actual model reasoning.

Human-centered evaluation focuses predominantly on perceived trust rather than calibrated reliance or decision quality, leaving important questions unresolved regarding the behavioral impact of XAI on clinical decision making. The six studies assessing behavioral effects found mixed evidence, with some documenting increased automation bias, highlighting that XAI effects are not uniformly beneficial.

Clinical outcome validation is rare: existing evidence for diagnostic and workflow improvement remains limited and inconsistent across studies. The three prospective studies revealed clinician skepticism (65-82% acceptance rates), workflow integration challenges, and declining usage over time. Another study documented a decline from 85% to 42% usage over six months, suggesting that short-term laboratory benefits may not predict sustained clinical adoption.

These validation gaps mean current evidence cannot support strong claims about XAI clinical utility. Most explanations have not been subjected to basic fidelity tests, do not demonstrate behavioral benefits in rigorous evaluation and lack clinical outcome evidence. This review identifies emergent research questions

grounded in these validation gaps, providing concrete priorities for advancing the field toward more rigorous validation practices. The path forward requires transforming XAI from technical innovation into clinical accountability tool through treating explanation validation with the same rigor as model performance evaluation. Without more comprehensive validation, XAI risks providing only partial evidence of clinical accountability.

Limitations

Several limitations warrant consideration when interpreting our findings. Database coverage was limited to seven major sources, and English-only inclusion may have introduced geographic bias, particularly by underrepresenting research from non-English-speaking countries. 38 studies could not be retrieved despite direct author contact, this may introduce selection bias. Screening and data extraction were conducted by a single reviewer following the pre-specified coding protocol. Publication bias likely affects the sample, as studies with negative validation results, such as low fidelity scores or absent performance improvements, may be underrepresented in published literature. Substantial heterogeneity in XAI methods, validation approaches, imaging modalities, and clinical applications precluded meta-analysis, limiting the ability to provide quantitative effect estimates beyond descriptive statistics. Our focus on radiology means findings may not generalize to XAI applications in other medical imaging domains, such as pathology or dermatology, or to non-imaging clinical AI applications. The emphasis on validation practices means, limited analysis of XAI technical innovations is provided, and methodological advances may not yet have reached empirical validation literature, given publication lags.

Acknowledgements

The authors acknowledge the use of OpenAI's ChatGPT to enhance the efficiency of the writing process. The tool was employed to support tasks such as improving clarity, refining language, and ensuring coherence. All study ideas, definitions, findings, interpretations, and conclusions are the authors' own and remain original and independent of AI-generated content.

References

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity checks for saliency maps. *Proceedings of the 32nd International Conference on Neural Information Processing Systems, NIPS'18*, 9525–9536. https://dl.acm.org/doi/10.5555/3327546.3327621

Aravinda, C. V., Shaffi, N., Viswan, V., Nair, S. S. K., Khan, M. S., & Arjun, B. C. (2026). MoRE-XAI: Moment-Based Rigid Registration and Multi-Perspective Explainability for Fracture Localization in X-Ray Imaging. *Ieee Access*, *14*, 58635–58653. https://doi.org/10.1109/ACCESS.2026.3680031

Asmita, C., Gizem, K., & Nirvana, M. (2026). A Cross-Domain Benchmark of Intrinsic and Post Hoc Explainability for 3D Deep Learning Models. *Journal of Imaging*, *12*(2), 63–83. https://doi.org/10.3390/jimaging12020063

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion*, *58*(C), 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Buga, R., Buzea, C. G., Agop, M., Ochiuz, L., Vasincu, D., Popa, O., Rusu, D. I., Știrban, I., & Eva, L. (2025). Streamlit Application and Deep Learning Model for Brain Metastasis Monitoring After Gamma Knife Treatment. *Biomedicines*, *13*(2), 423. https://doi.org/10.3390/biomedicines13020423

Cerekci, E., Alis, D., Denizoglu, N., Camurdan, O., Ege Seker, M., Ozer, C., Hansu, M. Y., Tanyel, T., Oksuz, I., & Karaarslan, E. (2024). Quantitative evaluation of Saliency-Based Explainable artificial intelligence (XAI) methods in Deep Learning-Based mammogram analysis. *European Journal of Radiology*, *173*, 111356. https://doi.org/10.1016/j.ejrad.2024.111356

Doshi-Velez, F., & Kim, B. (2017). Towards A Rigorous Science of Interpretable Machine Learning. *arXiv: Machine Learning*. https://www.semanticscholar.org/paper/Towards-A-Rigorous-Science-of-Interpretable-Machine-Doshi-Velez-Kim/5c39e37022661f81f79e481240ed9b175dec6513

E Ihongbe, I., Fouad, S., F Mahmoud, T., Rajasekaran, A., & Bhatia, B. (2024). Evaluating Explainable Artificial Intelligence (XAI) techniques in chest radiology imaging through a human-centered Lens. *PloS One*, *19*(10), e0308758. https://doi.org/10.1371/journal.pone.0308758

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., & Dean, J. (2019). A guide to deep learning in healthcare. *Nature Medicine*, *25*(1), 24–29. https://doi.org/10.1038/s41591-018-0316-z

Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*, *3*(11), e745–e750. https://doi.org/10.1016/S2589-7500(21)00208-9

Kinger, S., & Kulkarni, V. (2025). Transparent and trustworthy interpretation of COVID-19 features in chest X-rays using explainable AI. *Multimedia Tools and Applications*, *84*(18), 19853–19881. https://doi.org/10.1007/s11042-024-19755-y

Lysdahlgaard, S. (2023). Utilizing heat maps as explainable artificial intelligence for detecting abnormalities on wrist and elbow radiographs. *Radiography*, *29*(6), 1132–1138. https://doi.org/10.1016/j.radi.2023.09.012

Muhammad, D., & Bendechache, M. (2024). Unveiling the black box: A systematic review of Explainable Artificial Intelligence in medical image analysis. *Computational and Structural Biotechnology Journal*, *24*, 542–560. https://doi.org/10.1016/j.csbj.2024.08.005

Muhammad, F., Kim, H., Sufyan, D., Minh, D. L., Abolghasem, S.-N., & Hyeonjoon, M. (2026). Demystifying Artificial Intelligence: A Systematic Review of Explainable Artificial Intelligence in Medical Imaging. *Sensors*, *26*(7), 2131–2185. https://doi.org/10.3390/s26072131

O'Shea, R., Manickavasagar, T., Horst, C., Hughes, D., Cusack, J., Tsoka, S., Cook, G., & Goh, V. (2023). Weakly supervised segmentation models as explainable radiological classifiers for lung tumour detection on CT images. *Insights into Imaging*, *14*(1), 195. https://doi.org/10.1186/s13244-023-01542-2

Page, M. J., Moher, D., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … McKenzie, J. E. (2021). PRISMA 2020 explanation and elaboration: Updated guidance and exemplars for reporting systematic reviews. *The BMJ*, *372*, n160. https://doi.org/10.1136/bmj.n160

Rajpurkar, P., Irvin, J., Ball, R. L., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C. P., Patel, B. N., Yeom, K. W., Shpanskaya, K., Blankenberg, F. G., Seekins, J., Amrhein, T. J., Mong, D. A., Halabi, S. S., Zucker, E. J., … Lungren, M. P. (2018). Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLoS Medicine*, *15*(11), e1002686. https://doi.org/10.1371/journal.pmed.1002686

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *2017 IEEE International Conference on Computer Vision (ICCV)*, 618–626. https://doi.org/10.1109/ICCV.2017.74

Singhal, A., Agrawal, K., Quezada, A., Aguiñaga, A., Jiménez, S., & Yadav, S. (2024). Explainable Artificial Intelligence (XAI) Model for Cancer Image Classification. *Computer Modeling in Engineering & Sciences*, *141*(1), 401–441. https://doi.org/10.32604/cmes.2024.051363

Wu, Z., Zhuo, R., Liu, X., Wu, B., & Wang, J. (2024). Enhancing surgical decision-making in NEC with ResNet18: A deep learning approach to predict the need for surgery through x-ray image analysis. *Frontiers in Pediatrics*, *12*, 1405780. https://doi.org/10.3389/fped.2024.1405780

Zaheer, A. N., Farhan, M., Min, G., Alotaibi, F. A., & Alnfiai, M. M. (2025). Attention-enhanced hybrid U-Net for prostate cancer grading and explainability. *Scientific Reports (Nature Publisher Group)*, *15*(1), 34038. https://doi.org/10.1038/s41598-025-13824-4